AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Original Research—Laryngology

Otolaryngology–Head and Neck Surgery
2024, Vol. 171(4) 1106–1113

DOI: 10.1002/ohn.897
http://otojournal.org
WILEY

ChatGPT-4 Consistency in Interpreting Laryngeal Clinical Images of Common Lesions and Disorders

Antonino Maniaci, MD, PhD[1,2], Carlos M. Chiesa-Estomba, MD, PhD, MS[1,3,4], and Jérôme R. Lechien, MD, PhD, MS[1,5,6]

Abstract

Objective. To investigate the consistency of Chatbot Generative Pretrained Transformer (ChatGPT)-4 in the analysis of clinical pictures of common laryngological conditions.

Study Design. Prospective uncontrolled study.

Setting. Multicenter study.

Methods. Patient history and clinical videolaryngostroboscopic images were presented to ChatGPT-4 for differential diagnoses, management, and treatment(s). ChatGPT-4 responses were assessed by 3 blinded laryngologists with the artificial intelligence performance instrument (AIPI). The complexity of cases and the consistency between practitioners and ChatGPT-4 for interpreting clinical images were evaluated with a 5-point Likert Scale. The intraclass correlation coefficient (ICC) was used to measure the strength of interrater agreement.

Results. Forty patients with a mean complexity score of 2.60 ± 1.15. were included. The mean consistency score for ChatGPT-4 image interpretation was 2.46 ± 1.42. ChatGPT-4 perfectly analyzed the clinical images in 6 cases (15%; 5/5), while the consistency between GPT-4 and judges was high in 5 cases (12.5%; 4/5). Judges reported an ICC of 0.965 for the consistency score ($P = .001$). ChatGPT-4 erroneously documented vocal fold irregularity (mass or lesion), glottic insufficiency, and vocal cord paralysis in 21 (52.5%), 2 (0.05%), and 5 (12.5%) cases, respectively. ChatGPT-4 and practitioners indicated 153 and 63 additional examinations, respectively ($P = .001$). The ChatGPT-4 primary diagnosis was correct in 20.0% to 25.0% of cases. The clinical image consistency score was significantly associated with the AIPI score ($r_s = 0.830$; $P = .001$).

Conclusion. The ChatGPT-4 is more efficient in primary diagnosis, rather than in the image analysis, selecting the most adequate additional examinations and treatments.

Keywords

accuracy, artificial intelligence, ChatGPT, GPT, head neck surgery, images, laryngology, otolaryngology, picture, video

Received March 21, 2024; accepted June 9, 2024.

The Chatbot Generative Pretrained Transformer (ChatGPT) was launched on November 20, 2022 by OpenAI (OpenAI) and uses algorithms to respond to simple-to-complicated questions.[1] Since then, many applications of ChatGPT have been studied in otolaryngology–head and neck surgery related to clinical and basic science research,[2] referencing,[3] medical examinations,[4] clinical vignettes,[5,6] and editing scientific reports through spelling correction.[7] The accessibility and popularity of ChatGPT encourage patients to use them for education prior to medical consultation.[5] In the same vein, medical students, residents, and fellow-in-training can consider ChatGPT an adjunctive clinical tool for improving their practice.[8] Among the studies dedicated to clinical vignettes,[1,5,6] the authors assessed the diagnosis and treatment performance of ChatGPT by providing comprehensive information related to the clinical history and examinations. Although ChatGPT is purported to have picture analysis capabilities, to date, no studies have investigated the consistency between practitioners and ChatGPT in analyzing clinical images in otolaryngology–head and neck surgery.

[1]Research Committee of Young Otolaryngologists of the International Federation of Otorhinolaryngological Societies (IFOS), Paris, France
[2]Department of Medicine and Surgery, Kore University, Enna, Italy
[3]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
[4]Department of Otorhinolaryngology–Head and Neck Surgery, Donostia University Hospital Donosti-San, Sebastián, Spain
[5]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
[6]Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium

Corresponding Author:
Jérôme R. Lechien, MD, PhD, MS, Division of Laryngology and Broncho-esophagology, Department of Otolaryngology–Head and 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 objective of this preliminary study was to investigate the consistency of ChatGPT-4 in the analysis of clinical pictures of common laryngological conditions.

Methods

Patients and Setting

Forty patients consulting at the Division of Laryngology of CHU Saint-Pierre (Brussels, Belgium) and Dour Medical Center (Dour, Belgium) for primary voice or swallowing symptoms were consecutively recruited from November 1, 2023, to February 1, 2024. The patient data (eg, demographics, history, symptoms, medication, physical examination) and the laryngeal fiberoptic pictures were retrospectively entered into the application programming interface (API) of ChatGPT-4. Patients were included if the information related to the medical record, clinical examination, and any additional examinations were fully available. Patients with incomplete data, without identified laryngopharyngeal disorders, or lacking videolaryngo (strobo)scopic video findings were excluded.

The chatbot was systematically queried to provide an analysis of the clinical cases and the related videolaryngo (strobo)scopic images using standardized questions (**Figure 1**). For each case, the laryngeal configuration included images with vocal folds in abduction and adduction. Because ChatGPT-4 cannot receive videos, in cases of diseases with movement abnormalities, such as tremors, the movement of the laryngeal structures was described in the history. The study was approved by the CHU Saint-Pierre-IRB (n°BE0762023230708). The patient consented to participate.

ChatGPT-4 Consistency

The diagnosis of laryngeal conditions was made prior to analysis by 3 board-certified laryngologists, considering patient history, symptoms, videolaryngostroboscopy findings, and additional findings, including evaluations of voice quality, additional examinations, or histopathological findings. The responses from ChatGPT-4 were then independently evaluated by 3 laryngologists for consistency and performance analysis.

The complexity of clinical images was assessed by the laryngologists (agreement) using a 5-point Likert scale, ranging from 1 (very low complexity) to 5 (very high complexity; **Figure 1**).

Laryngologists also independently rated the consistency of ChatGPT-4's analysis of clinical images using a 5-point Likert scale, from 1 (very low consistency) to 5 (very high consistency).[9] The performance of ChatGPT-4 in providing accurate primary and differential diagnoses, suggesting additional examinations, and recommending treatments were assessed using the artificial intelligence performance instrument (AIPI; Supplemental Appendix S1, available online).[10] AIPI is a 9-item validated instrument for the performance analysis of generative AI chatbots, covering medical and surgical history; symptoms; physical examination; diagnosis; additional examinations; management plans; and treatments in the overall management of real clinical cases. AIPI is subdivided into the following subscores, each 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).[10]

Errors in ChatGPT-4 responses were documented, and correct information was subsequently entered into the API as human feedback. The images of cases that were associated with an erroneous interpretation by ChatGPT-4 were regenerated 5 days postfeedback to assess potential improvements in ChatGPT-4's responses.

Statistical Analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (SPSS version 29.0; IBM Corp).

The number of additional examinations indicated by ChatGPT-4 and by otolaryngologists were compared with the Mann-Whitney *U* test. The interrater reliability for the AIPI score assigned by the 3 laryngologists was evaluated with intraclass correlation coefficient (ICC) consistency. The relationship between case difficulty levels, the 5-point consistency scores, and the AIPI scores from the judges was analyzed using the Spearman correlation coefficient. A level of significance of $P < .05$ was used.

Results

The data for 40 patients was presented to ChatGPT-4. The mean age of patients was 54.7 ± 18.2 years. There were 21 (52.5%) females and 19 (47.5%) males (**Table 1**). The most common diagnoses included laryngopharyngeal reflux disease, glottic insufficiency, and unilateral or bilateral vocal fold paralysis. Some images inputted into ChatGPT-4 are available in **Figure 2**. The mean complexity score was 2.60 ± 1.15.

Selection of Additional Examinations

The additional examinations indicated by both practitioners and ChatGPT-4 are reported in **Table 2**. ChatGPT-4 indicated 153 additional examinations, and practitioners indicated 63, respectively ($P = .001$). Thus, 90 additional examinations were proposed by ChatGPT, while they were not necessary according to the otolaryngologists. In most cases, ChatGPT-4 indicated a significantly higher number of additional examinations than practitioners (**Table 2**). The following additional examinations were only indicated by ChatGPT-4: neck magnetic resonance imaging (N = 19); laryngeal electromyography (N = 14); single or dual-probe pH metry (N = 6); sensory laryngeal testing (N = 6); thyroid

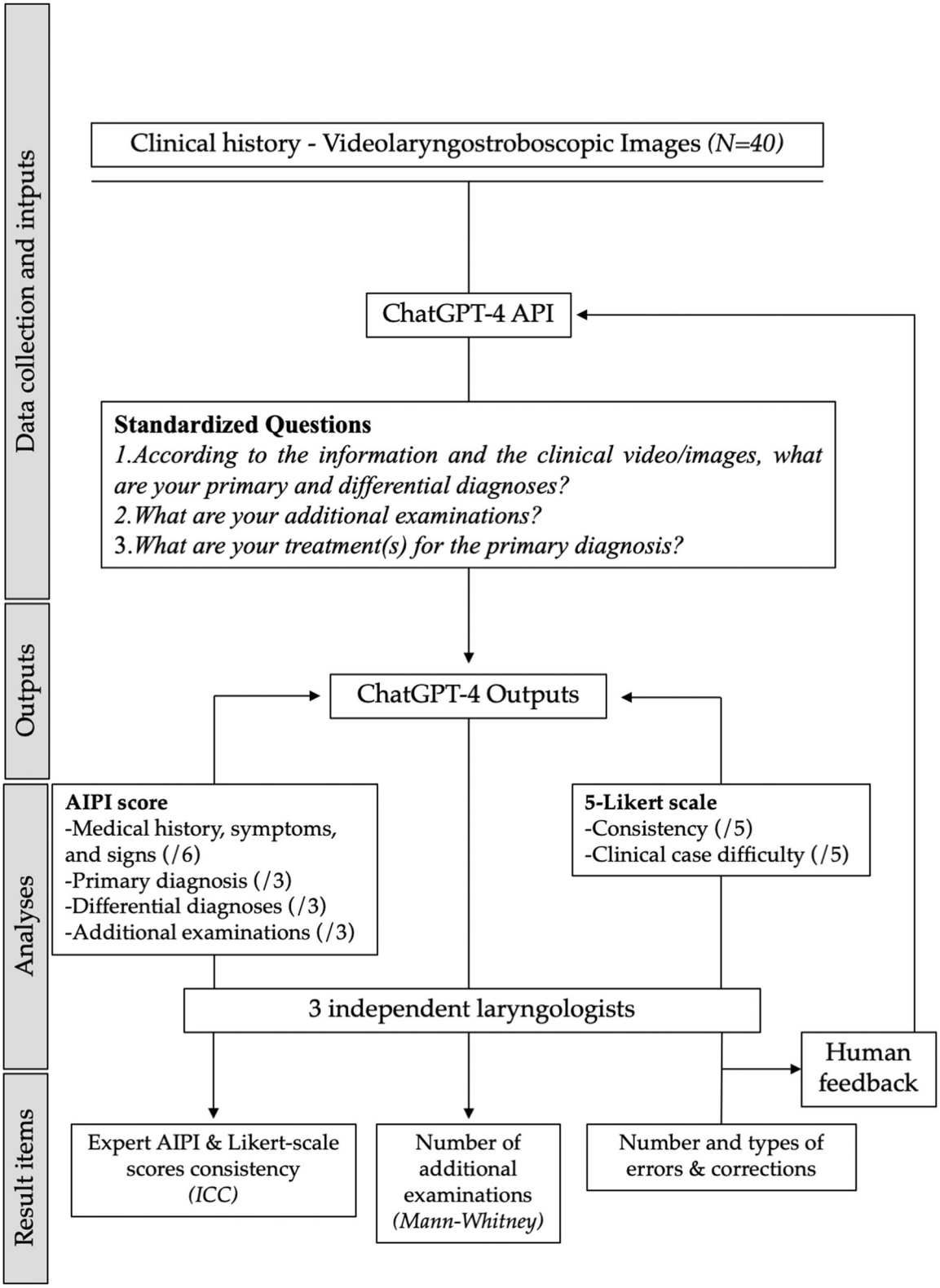

Figure 1. Chart flow. AIPI, artificial intelligence performance instrument; API, application programming interface; ChatGPT-4, Chatbot Generative Pretrained Transformer-4; ICC, intraclass correlation coefficient.

check-up (N = 3); allergy testing (N = 2); videokymography (N = 1); throat swab (N = 1); general blood biology (N = 1), and esophageal manometry (N = 1). For 2 patients with idiopathic recurrent laryngeal nerve paralysis, practitioners, but not ChatGPT-4, indicated chest tomodensitometry. In 19 cases, ChatGPT-4 recommended a stroboscopy, while the authors mentioned that the images were related to videolaryngostroboscopic

Table 1. Patient Symptoms

Outcomes	Patients (N = 40)
Age (mean, SD)	54.7 ± 18.2
Gender (N, %)	
Female	21 (52.5)
Male	19 (47.5)
Primary and secondary diagnoses	
Laryngopharyngeal reflux disease	12 (30.0)
Glottic insufficiency	5 (12.5)
Unilateral or bilateral vocal cord paralysis	5 (12.5)
Aging voice	4 (10.0)
Reinke edema	3 (7.5)
Unilateral or bilateral vocal fold sulcus	3 (7.5)
Vocal fold granuloma	2 (5.0)
Posterior glottic stenosis	2 (5.0)
Vocal cord cyst	2 (5.0)
Iatrogenic laryngitis (inhaled corticosteroids)	2 (5.0)
Vocal fold nodules	2 (5.0)
Vocal fold polyp	1 (2.5)
Vocal cord scarring	1 (2.5)
Vocal fold hemorrhage	1 (2.5)
Essential tremor	1 (2.5)
Leukoplakia	1 (2.5)
Anterior commissure synechia	1 (2.5)
Complexity score	2.60 ± 1.15

Abbreviations: N, number; SD, standard deviation.

examination. Barium swallow study and thyroid check-up were systematically indicated for patients with dysphagia or globus pharyngeus by ChatGPT-4. ChatGPT commonly indicated several imaging procedures for the same patient, for example, CT-scan, MRI, and ultrasonography, while only 1 was sufficient to investigate the diagnosis. In addition, ChatGPT-4 proposed "voice assessment" in 18 cases (45%), "acoustics" in 3 cases (7.5%), and "voice assessment and acoustics" in 1 case (2.5%) with dysphonia, respectively.

Performance in Diagnosis and Treatments

The performance of ChatGPT-4 in the management of clinical cases is described in **Tables 3** and **4**. ChatGPT-4 demonstrated high performance in considering medical history and symptoms for case management. In most cases, the performance in proposing complete or plausible primary and differential diagnoses was considered moderate. According to judges, the primary diagnosis was correct in 20.0% to 25.0% of cases. The variation in scores among judges stems from the consideration of additional primary diagnoses in 2 patients who had 2 primary diagnoses. ChatGPT-4's performance was low in proposing relevant and necessary additional examinations and treatments (**Table 3**). The mean AIPI scores are reported in **Table 4**. The ICC was excellent for the AIPI (**Table 4**).

Performance in Image Analysis

The mean consistency score of judges for the ChatGPT-4 ability to interpret images was 2.46 ± 1.42. The ICC regarding the consistency score was 0.965 ($P = .001$). Judges agreed that ChatGPT-4 perfectly analyzed the clinical images in 6 cases (15%) and demonstrated high consistency (Likert-scale score of 4/5) in 5 additional cases (12.5%). The accuracy of ChatGPT-4 in analyzing images was 27.5% (11/40). The 11 primary diagnoses correctly diagnosed by ChatGPT-4 on the laryngeal images (scores of 4/5 or 5/5) included glottic insufficiency (N = 5), laryngopharyngeal reflux disease (N = 2), vocal fold scarring (N = 1), posterior glottic stenosis with posterior synechia (N = 1), bilateral vocal cord paralysis (N = 1), and essential tremor (N = 1). For the last one, ChatGPT-4 did not detect any abnormalities on the vocal fold and suggested the essential tremor while mentioning the need to confirm the diagnosis through a videolaryngostroboscopy. In these 11 cases, 2 patients also reported signs of laryngopharyngeal reflux disease and, therefore, had 2 primary diagnoses.

The main errors of ChatGPT-4 in the analysis of images are reported in **Table 5**. Among the 29 patients with inaccurate analysis of ChatGPT-4, ChatGPT-4 documented vocal fold irregularity, mass, or lesion in 21 patients without vocal fold structure abnormalities. Glottic insufficiency and vocal fold paralysis were wrongly reported in 2 and 5 cases, respectively. ChatGPT-4 accurately described the position of vocal folds in 9 cases and inaccurately in 2 cases, respectively. Some examples of errors or regenerated responses are available in supplementary online files. Considering the initial errors (N = 29/40), ChatGPT-4 corrected only 1 error in the regenerated response (N = 1/29).

Focusing on the 11 cases where ChatGPT-4 provided adequate image analysis, the primary diagnosis score was plausible in 1 case (9.1%) and correct in 10 cases (90.9%), respectively.

The clinical image consistency score was significantly associated with the AIPI score ($r_s = 0.830$; $P = .001$) and the complexity score ($r_s = 0.350$; $P = .027$). There was no significant association between the complexity score and the AIPI ($r_s = 0.264$; $P = .100$).

Discussion

The performance, accuracy, and consistency of ChatGPT were investigated in the management of real clinical cases in a few studies.[1,4-6] However, there is no study considering the inclusion of clinical images in the API.

The primary findings of this study support a low performance of ChatGPT-4 in the analysis of laryngological images. ChatGPT-4 recognized the anatomical region (larynx) but could not detect laryngological signs or lesions in the submitted material. This observation does not corroborate the findings of Sievert et al who reported high

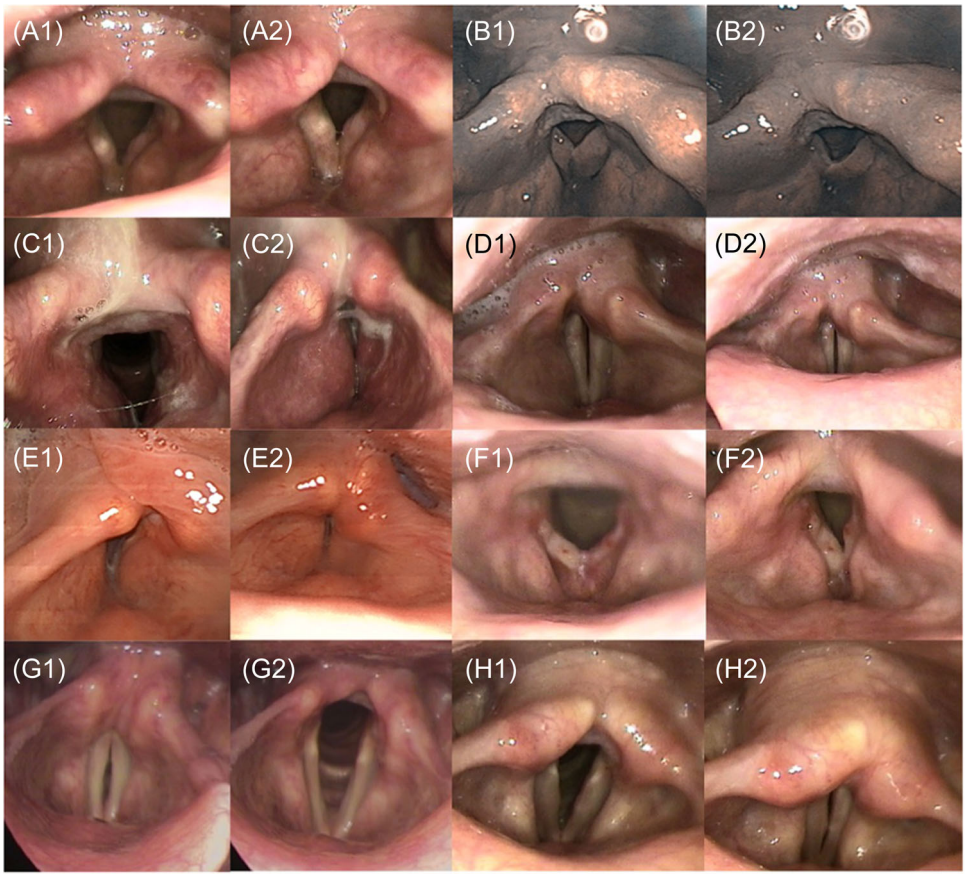

Figure 2. Examples of laryngeal images were introduced into the ChatGPT-4 API. Vocal cord cyst (A1-A2), Reinke edema (NBI, B1-B2), cordectomy and supraglottic voice (C1-C2), bilateral (D1-D2)/unilateral vocal cord paralysis (E1-E2), anterior commissure synechia (F1-F2), glottic insufficiency (G1-G2), and bilateral nodules (H1-H2). API, application programming interface; ChatGPT-4, Chatbot Generative Pretrained Transformer-4; NBI, narrow band imaging.

Table 2. Additional Examination Consistency

Additional examinations (N, %)	Laryngologists	ChatGPT-4	*P* value
HEMII-pH monitoring	13 (32.5)	1 (2.5)	.001
Neck tomodensitometry	2 (5.0)	13 (32.5)	.002
Neurological consultation	1 (2.5)	1 (2.5)	NS
In-office/operating room biopsy	1 (2.5)	14 (35.0)	.001
Voice quality assessment	38 (95.0)	22 (55.0)	.001
Suspension laryngoscopy	1 (2.5)	14 (35.0)	.001
Barium swallow study	1 (2.5)	7 (17.5)	.026
Lung assessment	2 (5.0)	16 (40.0)	.001
Swallowing testing	1 (2.5)	9 (22.5)	.007
GI endoscopy	1 (2.5)	3 (7.5)	NS

This table presents additional examinations commonly indicated by practitioners and ChatGPT-4 at least 1 time.
Abbreviations: ChatGPT-4, Chatbot Generative Pretrained Transformer-4; GI, gastrointestinal; HEMII-pH, hypopharyngeal-esophageal multichannel intraluminal impedance-pH monitoring; N, number; NS, nonsignificant.

ChatGPT-4 consistency with practitioners detecting cancer in confocal laser endomicroscopy images.[11] Thus, the authors submitted 139 images of normal multilayer epithelium and squamous cell carcinoma confocal images into the API. They observed that ChatGPT-4 and 3 experts adequately identified carcinoma in 71.2% and 88.5% of cases, respectively, which were encouraging results.[11] The comparison of our primary findings with the current otolaryngological and nonotolaryngological literature is still limited because most studies discuss the role of ChatGPT in detecting radiological[12,13] electrophysiological,[14] and histopathological[15] findings. In contrast, no study has investigated consistency in clinical picture analyses. To the best of our knowledge, this is the first study to examine the performance of ChatGPT-4 in the assessment of clinical images in otolaryngology–head and neck surgery.

In the present study, the performance of ChatGPT-4 in providing additional examinations and therapeutic options was judged as mild-to-moderate and moderate-to-high, respectively. ChatGPT-4 commonly indicated a high

Table 3. Performance Analysis of ChatGPT-4 in Diagnostic, Examinations, and Treatment

AIPI management outcomes	Judge 1 N (%)	Judge 2 N (%)	Judge 2 N (%)
1. Consideration of medical history			
Complete	34 (85.0)	34 (85.0)	32 (80.0)
Partial	3 (7.5)	3 (7.5)	5 (12.5)
No consideration	3 (7.5)	3 (7.5)	3 (7.5)
2. Consideration of symptoms			
Complete	38 (95.0)	38 (95.0)	36 (90.0)
Partial	1 (2.5)	1 (2.5)	4 (10.0)
No consideration	1 (2.5)	1 (2.5)	0 (0)
3. Consideration of physical examination findings			
Complete	14 (35.0)	12 (30.0)	24 (60.0)
Partial	10 (25.0)	12 (30.0)	13 (32.5)
No consideration	16 (40.0)	16 (40.0)	13 (32.5)
4. Differential diagnosis			
Complete and plausible	11 (27.5)	13 (32.5)	10 (25.0)
Incomplete but plausible	11 (27.5)	9 (22.5)	12 (30.0)
Incomplete and not plausible	15 (37.5)	16 (40.0)	15 (37.5)
Absent	3 (7.5)	2 (5.0)	3 (7.5)
5. Primary diagnosis			
Correct	10 (25.0)	9 (22.5)	8 (20.0)
Plausible	9 (22.5)	11 (27.5)	11 (27.5)
Not plausible	18 (45.0)	17 (42.5)	18 (45.0)
Absent	3 (7.5)	3 (7.5)	3 (7.5)
6. Additional examinations			
Pertinent and necessary	5 (12.5)	5 (12.5)	5 (12.5)
Pertinent and partially necessary	14 (35.0)	15 (37.5)	15 (37.5)
Association of pertinent, necessary, and inadequate	21 (52.5)	20 (50.0)	20 (50.0)
Association of inadequate examinations	0 (0)	0 (0)	0 (0)
7. The most relevant additional examination	19 (47.5)	18 (45.0)	21 (52.5)
8. Treatment			
Pertinent and necessary	10 (25.0)	11 (27.5)	9 (22.5)
Pertinent but incomplete	15 (37.5)	15 (37.5)	15 (37.5)
Association of pertinent, necessary, and inadequate	14 (35.0)	13 (32.5)	15 (37.5)
Inadequate	1 (2.5)	1 (2.5)	1 (2.5)

The 3 judges reported adequate ICC.
Abbreviations: AIPI, artificial intelligence performance instrument; ChatGPT-4, Chatbot Generative Pretrained Transformer-4; ICC, intraclass consistency; N, number.

Table 4. Performance of ChatGPT-4 According to Judges

AIPI outcomes	Mean score (SD) Judge 1	Judge 2	Judge 3	ICC
Patient feature score	4.65 ± 1.31	4.60 ± 1.26	4.64 ± 1.21	0.914
Diagnostic score	3.87 ± 2.15	3.98 ± 2.08	3.80 ± 2.03	0.976
Additional examination score	2.01 ± 1.05	2.08 ± 1.07	2.15 ± 1.03	0.947
Treatment	1.85 ± 0.83	4.65 ± 1.31	1.80 ± 0.82	0.880
AIPI total score	12.45 ± 4.28	12.55 ± 4.19	12.40 ± 4.01	0.968

The table described the ChatGPT-4 performance according to AIPI sub- and total scores.
Abbreviations: AIPI, artificial intelligence performance instrument; ChatGPT-4, Chatbot Generative Pretrained Transformer-4; ICC, intraclass consistency; SD, standard deviation.

Table 5. Primary Inaccuracies of ChatGPT-4 in the Image Analysis

ChatGPT-4 inaccuracies	N
Inaccurate documentation of VF irregularity or mass	21
Leukoplakia in place of cyst	1
Leukoplakia in place of Reinke edema	2
Sulcus	1
Nodules	2
Cyst	1
Polyp/nodule/cancer in place of Reinke edema	1
Nodules in place of sulcus	1
Nodules in place of Leukoplakia	1
Polyp/nodule/cancer in place of LPR	3
Hemorrhage/nodules in place of polyps	1
Polyp/nodule/cancer in place of unilateral paralysis	1
Cyst or polyps in place of granuloma	1
Inaccurate detection of laryngeal stent	1
Inaccurate detection of VF edema	2
Inaccurate detection of VF mass	2
Inaccurate documentation of VF paralysis	5
Inaccurate documentation of glottic insufficiency	2
Nondetection of lesions or conditions	3
Granuloma	1
Cordectomy	1
Anterior commissure synechia	1

ChatGPT-4 inaccurately analyzed 29 clinical images for a total of 31 inaccuracies. Indeed, some cases had several inaccuracies related to the image analysis by ChatGPT-4.
Abbreviations: ChatGPT-4, Chatbot Generative Pretrained Transformer-4; LPR, laryngopharynge; N, number; VF, vocal fold.

number of additional examinations, associating inadequate, adequate, and necessary examinations, which corroborates the observation of other studies conducted in the field of otolaryngology–head and neck surgery.[5,6,16,17] In the study by Vaira et al, ChatGPT-4 proposed a nearly or fully correct primary diagnosis in 81.7% of cases, while the treatment was accurate in 56.7% of cases.[4] The primary diagnosis accuracy of ChatGPT-4, ranging from 47% to 71% across to studies,[5,10,16,17] supports the observations made in the current study.

The low performance of ChatGPT-4 in analyzing clinical images has significant implications for both patients and

practitioners. On the one hand, there is an increasing number of patients who use ChatGPT for symptom checking, medical jargon interpreting, or medication management,[18,19] making the assessment of ChatGPT's accuracy in providing health care information important. On the other hand, the large database underlying the information supplied by ChatGPT can encourage its use by medical students, residents, or young practitioners,[20,21] who need to know the limitations of the current versions carefully. As such, ChatGPT-4's interpretation of laryngological images is inaccurate in most cases.

The primary limitations of the present preliminary study are the low number of patients and the lack of information about the hyperparameters of ChatGPT-4. Most large language models (LLMs) are nondeterministic and their outputs may vary with each run, which may be curtailed by fine-tuning specific hyperparameters. Specifically, tuning the hyperparameters may influence ChatGPT-4's task performance, especially in integrating mistakes and corrections into the algorithms. The oversight of feedback provided to ChatGPT-4 after mistakes underscores the critical role of hyperparameters in the machine learning process. Moreover, ChatGPT-4 is not a specific software for the analysis of images, meaning that specific AI software dedicated to clinical analysis could probably report a higher performance score.

The hyperparameter information could be important to better understand the relationship between the accuracies of image analysis and primary diagnosis making. Indeed, in the present study, authors observed high accuracy in primary diagnosis making in patient cases where ChatGPT-4 provided an adequate image analysis. However, it is difficult to know the influence of the clinical information (medical history) or images on the primary diagnosis accuracy. In other words, we cannot know if the image analysis influenced the primary diagnosis making or if it is the other way around. This point requires further studies to better understand the functioning of LLMs, such as ChatGPT-4.

The originality and the assessment of the machine learning process through human feedback and regenerated ChatGPT-4 responses were the primary strengths of our study. Indeed, no study in the otolaryngology literature has evaluated ChatGPT's performance in image analysis. However, its potential to improve outputs after human feedback has been theoretically suggested but not yet investigated.

Conclusion

ChatGPT-4 may become a promising adjunctive tool in clinical laryngology but its performance remains low in the analysis of clinical images. Moreover, ChatGPT-4 appears to be more efficient in providing therapeutic options based on the clinical history and description of findings rather than proposing the most adequate additional examinations.

Author Contributions

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; **Carlos M. Chiesa-Estomba**, 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; **Jerome 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.

Disclosures

Competing interests: The authors have no conflict of interest.

Funding source: None.

Supplemental Material

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

References

1. Lechien JR, Rameau A. Applications of ChatGPT in otolaryngology-head neck surgery: a systematic review. *Otolaryngol Head Neck Surg*. 2024. In press. doi:10.1002/ohn.807
2. Nachalon Y, Broer M, Nativ-Zeltzer N. Using ChatGPT to generate research ideas in dysphagia: a pilot study. *Dysphagia*. 2023;39:407-411. doi:10.1007/s00455-023-10623-9
3. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otrhinolaryngol*. 2023;280(11):5129-5133. doi:10.1007/s00405-023-08205-4
4. 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;170:1492-1503. doi:10.1002/ohn.489
5. Lechien JR, Naunheim MR, Maniaci A, et al. Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case-series. *Otolaryngol Head Neck Surg*. 2024;170:1519-1526.
6. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur Arch Otrhinolaryngol*. 2024;281(4):2105-2114. doi:10.1007/s00405-023-08326-w
7. 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;170:1527-1530. doi:10.1002/ohn.526

8. Mat Q, Briganti G, Maniaci A, Lelubre C. Will ChatGPT soon replace otolaryngologists? *Eur Arch Otrhinolaryngol*. 2024;281:3303-3304. doi:10.1007/s00405-024-08543-x
9. Davis RJ, Ayo-Ajibola O, Lin ME, et al. Evaluation of oropharyngeal cancer information from revolutionary artificial intelligence chatbot. *Laryngoscope*. 2023;134: 2252-2257. doi:10.1002/lary.31191
10. 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
11. Sievert M, Aubreville M, Mueller SK, et al. Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using GPT 4.0 with vision. *Eur Arch Otrhinolaryngol*. 2024;281:2115-2122.
12. Bajaj S, Gandhi D, Nayar D. Potential applications and impact of ChatGPT in radiology. *Acad Radiol*. 2024; 31:1256-1261. doi:10.1016/j.acra.2023.08.039
13. Currie G, Barry K. ChatGPT in nuclear medicine education. *J Nucl Med Technol*. 2023;51(3):247-254. doi:10.2967/jnmt.123.265844
14. Martínez-Sellés M, Marina-Breysse M. Current and future use of artificial intelligence in electrocardiography. *J Cardiovasc Dev Dis*. 2023;10(4):175. doi:10.3390/jcdd10040175
15. Laohawetwanit T, Namboonlue C, Apornvirat S. Accuracy of GPT-4 in histopathological image detection and classification of colorectal adenomas. *J Clin Pathol*. 2024. In press. doi:10.1136/jcp-2023-209304
16. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: a clinical case series. *Int Forum Allergy Rhinol*. 2024;14:1123-1130. doi:10.1002/alr.23323
17. 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
18. Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. *Open Heart*. 2023;10(2):e002455. doi:10.1136/openhrt-2023-002455
19. Hudgins A. How patients are using ChatGPT in healthcare Health eCareers. 2024. Accessed March 16, 2024. https://www.healthecareers.com/career-resources/industry-news/patients-gpt-healthcare
20. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595. doi:10.3389/frai.2023.1169595
21. Lechien JR. Generative artificial intelligence in otolaryngology-head and neck surgery editorial: be an actor of the future or follower. *Eur Arch Otrhinolaryngol*. 2024;281(4):2051-2053. doi:10.1007/s00405-024-08579-z