Accuracy of ChatGPT-4o in Text and Video Analysis of Laryngeal Malignant and Premalignant Diseases

[*,†,§§,¶¶]Carlos M. Chiesa-Estomba, [*,†]Maider Andueza-Guembe, [‡,§§,¶¶]Antonino Maniaci, [§,§§,¶¶]Miguel Mayo-Yanez, [¶,¶¶]Frank Betances-Reinoso, [‖,**,§§]Luigi A. Vaira, [††,¶¶]Alberto Maria Saibene, and [‡‡,§§,¶¶]Jerome R. Lechien, *San Sebastian, †Bilbao, §A Coruña, ¶Lugo, Spain, ‡Enna, ‖**Sassari, ††Milan, Italy, ‡‡Mons, Belgium, and §§¶¶Paris, France

Summary: Introduction. Chatbot Generative Pretrained Transformer (ChatGPT), a multimodal generative AI, has been studied for potential applications in healthcare, including otolaryngology-head and neck surgery. In this study, authors investigates the consistency of ChatGPT-4o in analyzing clinical fiberoptic videos of suspected laryngeal malignancies compared to expert clinicians.

Methods. This experimental study involved twenty patients with primary laryngeal disease consulting at a tertiary academic center. Data, including laryngeal fiberoptic video examinations, were retrospectively analyzed using the ChatGPT-4o application programming interface. Responses were assessed for diagnostic accuracy, consistency, and clinical recommendations. Three otolaryngology-head and neck consultants independently evaluated ChatGPT-4o's performance using the Artificial Intelligence Performance Instrument and a five-point Likert scale for complexity and consistency.

Results. ChatGPT-4o identified malignant diagnoses as the primary diagnosis in 30% of cases, while proposing malignancies as one of the top three diagnoses in 90% of cases. Despite high sensitivity, specificity was limited. The mean consistency score for image analysis was 2.36 ± 1.13, with an intraclass correlation coefficient of 0.890 ($P = 0.03$). The model showed a tendency to prioritize text over visual data, limiting the improvement in diagnostic accuracy from video input.

Conclusion. While ChatGPT-4o demonstrates potential in analyzing laryngeal pathologies through multimodal data, current limitations in specificity and image interpretation indicate the need for further refinement. Ongoing advancements could enhance its integration into clinical workflows, supporting accurate diagnoses and decision-making in otolaryngology.

Key Words: Laryngeal cancer—ChatGPT—Fiberoptic video analysis—Multimodal AI—Otolaryngology.

INTRODUCTION

The Chatbot Generative Pretrained Transformer (ChatGPT) was launched on November 20, 2022, by OpenAI (OpenAI) based on a transformer algorithm architecture, looking to respond to simple-to-complicated questions.[1] In an evolving field like health care, multimodal generative artificial intelligence systems, such as ChatGPT-4o, represent a significant advancement in comparison with previous Natural Processing Language (NPL) tools, as they can integrate visual data with text data.

Since the release of the application programming interface (API) to the public, many hypothetical applications have been studied in Otolaryngology-Head and Neck Surgery (OHNS) related to clinical and basic science research,[2,3] referencing,[4] medical examinations,[5] and editing scientific reports through spelling correction.[6]

The accessibility and popularity of ChatGPT encourage patients to use them for self-education before the clinical visit.[7] Medical students, residents, or fellow-in-training considered ChatGPT an adjunctive clinical tool for improving their practice.[8]

Although the last version of ChatGPT can perform video analysis, according to the best of our knowledge, to date, no studies have investigated the consistency between practitioners and ChatGPT in analyzing clinical laryngeal fiberoptic videos in malignant laryngeal disease.

This experimental study aimed to investigate the consistency of ChatGPT-4o in the analysis of clinical fiberoptic videos of patients with suspicions of laryngological malignancies.

MATERIAL AND METHODS

Consecutive patients consulting at the Department of Otorhinolaryngology-Head and Neck Surgery of a Spanish

Accepted for publication March 4, 2025.

From the *Department of Otorhinolaryngology, Osakidetza, Donostia University Hospital, Biodonostia Research Institute, 20014 San Sebastian, Spain; †Otorhinolaryngology Department, Faculty of Medicine, Deusto University, Bilbao, Spain; ‡Faculty of Medicine and Surgery, University of Enna"Kore", 94100 Enna, Italy; §Otorhinolaryngology - Head and Neck Surgery Department, Complexo Hospitalario Universitario A Coruña (CHUAC), 15006, A Coruña, Galicia, Spain; ¶Department of Otorhinolaryngology, Hospital Lucus Augusti, Lugo, Spain; ‖Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy; **Ph.D. School of Biomedical Sciences, Biomedical Sciences Department, University of Sassari, Sassari, Italy; ††Otolaryngology Unit, Santi Paolo E Carlo Hospital, Department of Health Sciences, Università DegliStudi di Milano, Milan, Italy; ‡‡Department of Otolaryngology and Head and Neck Surgery, Division of Laryngology and Broncho-Esophagology, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons, Mons, Belgium; §§Head and Neck Study Group of Young-Otolaryngologists of the International Federations of Oto-rhino-laryngological Societies (YO-IFOS), Paris, France; and the ¶¶Young-Otolaryngologists of the International Federation of Oto-Rhino-Laryngological Societies (YO-IFOS) Study Group, Paris, France.

Address correspondence and reprint requests to: Carlos M. Chiesa-Estomba, Servicio de Otorrinolaringología – Cirugía de Cabeza y Cuello, Hospital Universitario Donostia, Paseo Dr. Begiristain #1, 20014, San Sebastian-Donosti, Spain. E-mail: chiesaestomba86@gmail.com

Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997

https://doi.org/10.1016/j.jvoice.2025.03.006

Tertiary Academic Centre for primary laryngeal disease (malignant or premalignant) were recruited from January 1, 2024, to February 1, 2024. Patient data and laryngeal fiberoptic video examinations were retrospectively entered into the API of ChatGPT-4o. Patients were included if complete information related to their medical records, clinical examinations, and at least a computed tomography (CT) scan of the neck were available. Patients with incomplete data, without malignant-suspicious disorders, or lacking fiberoptic video findings were excluded. The clinical decision regarding treatment was unrelated to ChatGPT's recommendations during the study period.

The chatbot was systematically queried to analyze the clinical cases with and without the related videolaryngoscopic images using standardized questions (Figure 1). For each case, the laryngeal configuration included a video of 8 seconds, no larger than 20 MB, with vocal fold movements performing abduction and adduction. In cases of vocal fold paresis or paralysis, the movement of the laryngeal structures was analyzed with and without movement data in the prompt strategy. The study was approved by the IRB (n°CH000341), and the patients consented to participate.

The diagnosis of laryngeal conditions was discussed before analysis by three board-certified Otolaryngology-Head and Neck consultants, considering patient history, symptoms, laryngeal fiberoptic examination findings, and histopathological findings. The responses from ChatGPT-4o were then independently evaluated by the same three Otolaryngology—Head and Neck surgeons for consistency and performance analysis. The complexity of clinical images was assessed using a five-point Likert scale, ranging from 1 (very low complexity) to 5 (very high complexity). Researchers also independently rated the consistency of ChatGPT-4o's analysis of clinical images using a five-point Likert scale, from 1 (very low consistency) to 5 (very high consistency). The performance of ChatGPT-4o in providing accurate primary and differential diagnoses, suggesting additional examinations, and recommending treatments was assessed using the Artificial Intelligence Performance Instrument (AIPI).[9]

Statistical analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (SPSS version 26.0; IBM Corp - USA). The interrater reliability for the AIPI score assigned by the three Otolaryngology-Head and Neck surgeons was evaluated with the intraclass correlation coefficient (ICC) consistency. The relationship between case difficulty levels, the five-point consistency scores, and the AIPI scores from the judges was analyzed using the Spearman correlation coefficient. A level of significance of $P < 0.05$ was used.

RESULTS

Twenty patients completed the evaluations. The cohort comprised 18 males (90%) and two females (10%), with a mean age of 66 years (SD = 6.48). All patients were smokers, and 40% ($n = 8$) reported alcohol consumption. The frames selected by the API and the CT-Scan inputted into ChatGPT-4o are available in Figure 2. The mean complexity score was 1.80 ± 0.98.

The lesions were predominantly located in the glottis (60%, $n = 12$), with the remaining 40% ($n = 8$) in the supraglottic. Laryngeal mobility was preserved in 90% ($n = 18$) of the cases, while one patient exhibited limited mobility. Histopathological analysis revealed that 70% ($n = 14$) had squamous cell carcinoma, with varying degrees of differentiation: moderate (71.4%, $n = 10$), mild (20%, $n = 21.4$), and basaloid (7.2%, $n = 1$). The remaining 30% ($n = 6$) had moderate dysplasia.

Tumor staging varied, with cT1 in 40% ($n = 8$), cT2 in 20% ($n = 4$), and cT3 in 10% ($n = 2$) of the patients. Lymph node involvement was absent (N0) in 65% ($n = 13$), and one

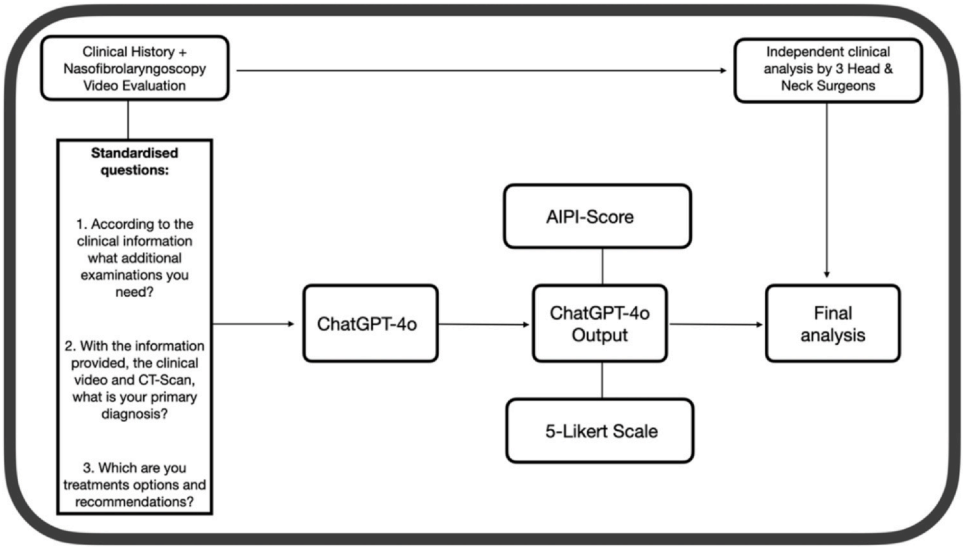

FIGURE 1. Study workflow.

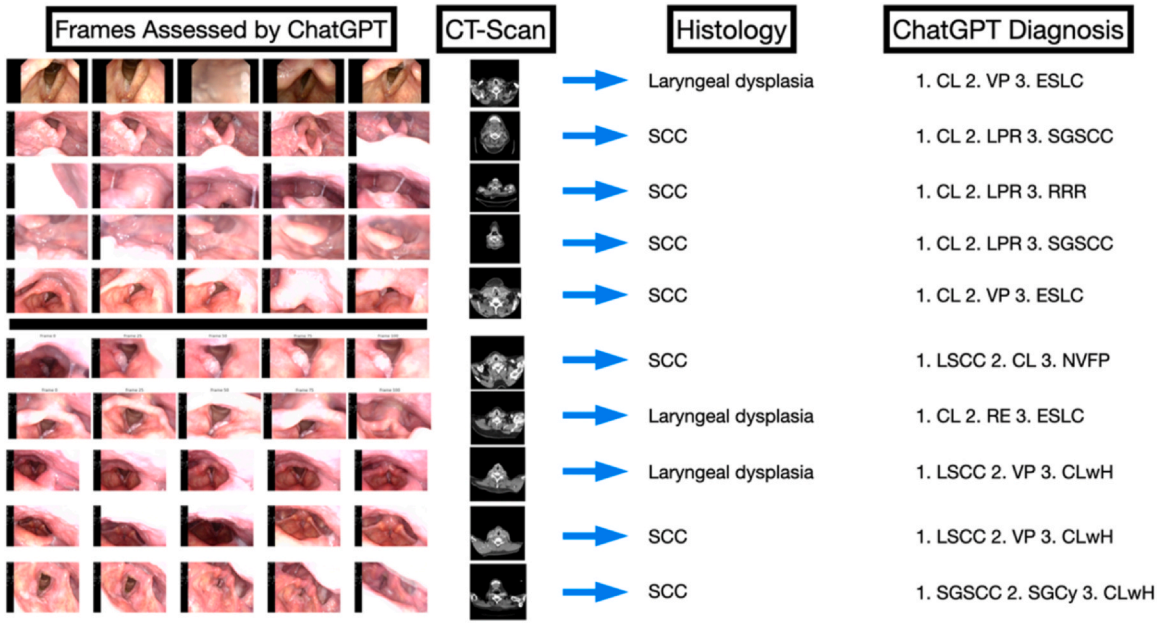

FIGURE 2. Information assessed by Chat-GPT4o (videolaryngoscopic frames images + CT Scan), blinded histology and the three most probable diagnosis proposed by Chat-GPT4o. CL, chronic laryngitis; VP, vocal polyp; LPR, laryngopharyngeal reflux; SGSCC, supraglottic squamous cell carcinoma; LSCC, laryngeal squamous cell carcinoma; RRR, recurrent respiratory papillomatosis; ESLC, recurrent respiratory papillomatosis; NVCP, neurological vocal cord paralysis; CLwH, chronic laryngitis with hyperplasia; SGCy, supraglottic cyst; RE, reinke edema.

patient had N2c in the group of malignancies. All patients with recorded metastasis status were M0. Surgical intervention was performed in 70% ($n = 14$) of the cases, with cordectomy being the most common procedure (60%, $n = 12$). Exclusive radiotherapy was administered to 30% ($n = 6$), and one patient received chemoradiotherapy.

Regarding ChatGPT-4o's suggestions, the API proposed preoperative studies, including CT scans and laryngeal fiberoptic examination in all cases. As treatment options, direct laryngoscopy and biopsy were universally recommended, along with smoking cessation, demonstrating high performance in considering the medical history and symptoms for case management. Additionally, in some cases, other treatments such as proton pump inhibitors, dietary and lifestyle modifications, and hydration were advised.

Potential primary diagnoses identified by ChatGPT included chronic laryngitis, a common consideration. Other potential diagnoses encompassed vocal cord nodules or polyps, early-stage laryngeal cancer, laryngeal cancer (T3), neurological vocal cord paralysis, Reinke edema, chronic laryngitis with hyperplasia, supraglottic cancer, and supraglottic cysts. In this subset, ChatGPT-4o made a correct malignant diagnosis in 6 (30%) of cases. However, a potential malignant diagnosis was proposed in up to 90% of cases as one of the three potential diagnoses. The mean consistency score for image analysis was 2.36 ± 1.13, with an intraclass correlation coefficient of 0.890 (P = 0.03)..

Regarding the performance of image analysis, the mean consistency score of judges for ChatGPT-4o's ability to interpret images was 2.36 ± 1.13. The ICC regarding the consistency score was 0.890 ($P = 0.03$). However, the rates of final diagnoses within the differential diagnosis lists generated by ChatGPT-4o tended to be inferior to those proposed by clinicians and did not show improvement over those generated without video or image. This suggests that ChatGPT-4o appears to prioritize text data over image input despite its multimodal data processing capabilities.

DISCUSSION

The primary objective of this study was to investigate the consistency and performance of ChatGPT-4o in the analysis of clinical fiberoptic videos of patients with suspected laryngeal malignancies.

Our results suggest that ChatGPT-4o can correctly identify malignant diagnoses and propose as the most probable diagnosis in up to 30% of the cases. However, it proposed a potential malignant diagnosis in up to 90% of cases as one of the three most potential diagnoses. This high rate of proposing malignancies suggests that while ChatGPT-4o is highly sensitive in detecting possible cancerous conditions, it may lack the specificity necessary to differentiate effectively between benign and malignant lesions. It was being this over-sensitivity, a potential cofounder, that could lead to over-diagnosis and unnecessary concern for patients and clinicians.

In this regard, we need to discuss why is relevant to consider the potential of an API in the clinical context? An API enables seamless communication between different software systems, ensuring standardized and automated data handling. In healthcare, APIs can integrate textual

and video inputs into ChatGPT-4o, allowing for efficient processing of clinical data. For example, a DICOM APIs standardize patient records and medical imaging, ensuring interoperability across electronic health records and diagnostic tools. For textual inputs, APIs extract and preprocess data from clinical notes, converting speech-to-text and structuring unstructured records. Natural language processing APIs analyze symptoms, diagnoses, and medical terminology before feeding the information into the ChatGPT-4° architecture for clinical decision support and documentation automation. For video inputs, APIs capture and process diagnostic recordings, improving hypothetically the clinical decision-making.[10]

By automating data exchange, APIs eliminate manual workflows, enhancing interoperability, and enable AI-driven medical insights. Moreover, they ensure that ChatGPT-4o can analyze and generate responses based on structured clinical information, optimizing patient care, improving diagnostics, and streamlining healthcare operations.[10]

The performance assessment using the AIPI showed that ChatGPT-4o demonstrated high performance in considering medical history and symptoms for case management. This suggests that the model is adept at integrating and analyzing textual patient data to provide relevant clinical recommendations. However, its performance in interpreting clinical images was rated lower, with a mean consistency score of 2.36 ± 1.13, and the interrater reliability for the consistency score was 0.890 ($P = 0.03$), indicating moderate agreement among the evaluators.

These results indicate that the model prioritizes text data over image input when generating differential diagnoses. This finding is critical as it highlights a limitation in the current multimodal capabilities of ChatGPT-4o. While including video data did not significantly improve the diagnostic accuracy over text-only data, it suggests that further development and refinement are needed to enhance the model's ability to utilize visual data effectively.

These results are similar to those obtained by Hirosawa et al, which try to analyze the impact of adding image data on ChatGPT-4's regarding diagnostic accuracy and provide insights into how image data integration can enhance the accuracy of multimodal AI in medical diagnostics. For this purpose, the authors analyze 557 case reports from the American Journal of Case Reports and demonstrated that ChatGPT predominantly relies on textual data.[11] Wu et al also try to assess the performance of ChatGPT-4, specifically in the realm of multimodal medical diagnosis, evaluating human body systems, including Central Nervous System, Head and Neck, Cardiac, Chest, Hematology, Hepatobiliary, Gastrointestinal, Urogenital, Gynecology, Obstetrics, Breast, Musculoskeletal, Spine, Vascular, Oncology, Trauma, Pediatrics, with images taken from eight modalities used in daily clinic routine, eg, X-ray, CT, magnetic resonance imaging, positron emission tomography, digital subtraction angiography, mammography, ultrasound, and pathology. However, the authors described that the API faces significant challenges in disease diagnosis and generating comprehensive reports.[12]

Regarding laryngeal image analysis with ChatGPT, in a study published by Maniaci et al looking to establish the consistency of ChatGPT analyzing clinical pictures of common laryngeal disorders, the authors found that ChatGPT was more efficient in primary diagnosis rather than in the image analysis, selecting the most adequate additional examinations and treatments.[13]

Looking forward, future advancements in AI models, such as the upcoming ChatGPT-O3 model, may provide the added benefit of articulating their reasoning during response generation. This "chain-of-thought" approach can improve transparency and clarity in the AI's decision-making process, fostering greater understanding and trust in clinical applications. Highlighting this prospective development could emphasize the continuous progress toward more interpretable and clinically relevant AI systems.

In this vein, the development of AI tools in OHNS reflects a broader trend in healthcare, where the potential of AI for clinical applications far exceeds its current implementation. This disparity, often termed the "AI chasm," arises from limitations in model reliability, fairness, and external validation. Despite advancements in machine learning models, deep learning, or NPL, many still a of lack comprehensive reporting, with critical gaps in transparency and clinical integration. Bridging this gap requires adherence to robust reporting standards and external validation efforts to ensure AI-driven diagnostics, such as this proposed in laryngeal image disease analysis, are both reliable and clinically meaningful. Addressing these limitations will be essential for the effective deployment of AI in OHNS, ultimately improving patient outcomes and clinical decision-making.[14]

Finally, despite this being the first study about video image analysis in laryngeal cancer, we must highlight some limitations of our study. In this vein, one of the key findings of this study is the need for further refinement of ChatGPT-4o's algorithms to improve the specificity of its diagnoses. We need to consider that ChatGPT is an open software trained with the current information available on the internet, but is not an artificial intelligence interphase specifically trained in Laryngeal disorders or Head and Neck Malignancies. Moreover, the model is based in the ability to predict potential grammar correlations. By incorporating nuanced clinical data and refining the model's training parameters, we will probably see an improvement in its ability to distinguish between benign and malignant conditions. Also, we need to highlight another limitation, related with the absence of voice analysis. Additionally, the possibility of training the model and integrate it into a collaborative clinical framework that complements healthcare professionals' expertise could optimize patient outcomes by providing a second opinion or supporting clinical decision-making. Also, the low number of patients included in our study can be considered a limitation.

CONCLUSION

Although ChatGPT could potentially be useful analyzing images from laryngeal malignancies, particularly through

its integration of text and video data, its current iteration exhibits specificity and diagnostic accuracy limitations. Ongoing advancements and careful implementation of AI tools like ChatGPT-4o can become valuable assets in clinical practice, supporting healthcare professionals in making more informed and accurate diagnoses. Further research and development are necessary to enhance the model's capabilities and ensure effective integration into clinical workflows.

Declaration of Competing Interest

Authors declare they don't have any conflict of interest.

References

1. Lechien JR, Rameau A. Applications of ChatGPT in otolaryngology-head neck surgery: a systematic review. *Otolaryngol Head Neck Surg.* 2024;171(3):667–677.In press. doi:10.1002/ohn.807.
2. Chiesa-Estomba CM, Lechien JR, Vaira LA, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur Arch Otorhinolaryngol.* 2024;281:2081–2086.
3. Nachalon Y, Broer M, Nativ-Zeltzer N. Using ChatGPT to generate research ideas in dysphagia: a pilot study. *Dysphagia.* 2023;39:407–411. https://doi.org/10.1007/s00455-023-10623-9.
4. 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;280:5129–5133. https://doi.org/10.1007/s00405-023-08205-4.
5. 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. https://doi.org/10.1002/ohn.489.
6. 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. https://doi.org/10.1002/ohn.526.
7. Lechien JR, Naunheim MR, Maniaci A, et al. Performance an consistency of ChatGPT-4 versus otolaryngologists: a clinica case-series. *Otolaryngol Head Neck Surg.* 2024;170:1519–1526.
8. Mat Q, Briganti G, Maniaci A, Lelubre C. Will ChatGPT soon replace otolaryngologists? *Eur Arch Otrhinolaryngol.* 2024;281:3303–3304. https://doi.org/10.1007/s00405-024-08543-x.
9. Lechien JR, Maniaci A, Gengler I, et al. 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. https://doi.org/10.1007/s00405-023-08219-y.
10. Fávero LP, Belfiore P, de Freitas Souza R. *Using application programming interfaces to collect data. Data Science, Analytics and Machine Learning With R.* Cambridge, MA: Academic Press,; 2023:157–167.
11. Hirosawa T, Harada Y, Tokumasu K, et al. Evaluating ChatGPT-4's diagnostic accuracy: impact of visual data integration. *JMIR Med Inform.* 2024;12:e55627. https://doi.org/10.2196/55627.
12. Wu C, Lei J, Zheng Q, et al. Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. *ArXiv.* 2023. https://doi.org/10.48550/arXiv.2310.09909. Preprint posted online on December 04.
13. Maniaci A, Chiesa-Estomba CM, Lechien JR. ChatGPT-4 consistency in interpreting laryngeal clinical images of common lesions and disorders. *Otolaryngol Head Neck Surg.* 2024;171(4):1106–1113. https://doi.org/10.1002/ohn.897.Epub ahead of print. PMID: 39045737.
14. Lu JH, Callahan A, Patel BS, et al. Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: a systematic review. *JAMA Netw Open.* 2022;5:e2227779 .