Generative AI and Otolaryngology—Head & Neck Surgery

Jérôme R. Lechien, MD, PhD, MS, AFACS[a,b,c,d]

KEYWORDS

• Otorhinolaryngology • Otolaryngology • Head neck • Surgery • Artificial intelligence • ChatGPT • GPT • Generative

KEY POINTS

- The current literature on generative artificial intelligence (AI) has been booming since the launch of AI-powered language models, such as Chatbot Generative Pre-trained Transformer (ChatGPT).
- Most studies investigated the accuracy of ChatGPT in providing general information on disease basic science and clinical research, clinical vignette management, scientific paper referencing, and improvement.
- ChatGPT may provide accurate theoretic information on otolaryngologic disorders commonly found in general otolaryngology, head and neck surgery, oncology, and sleep practices.
- The performance of ChatGPT as an adjunctive clinical tool for managing clinical vignettes or true clinical cases may be limited, especially in providing the most adequate additional examinations.
- The AI may revolutionize the otolaryngology—head and neck surgery field, which should lead to the improvement of patient care. The next few years will be decisive for applying the new AI technologies in the office-based practice.

INTRODUCTION

Artificial intelligence (AI) can revolutionize many fields in medicine and surgery. The mediatization related to the launch of Chatbot Generative Pre-trained Transformer

[a] Research Committee of Young Otolaryngologists of the International Federation of Otorhinolaryngological Societies (IFOS), Paris, France; [b] 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; [c] Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, Paris Saclay University, Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Paris, France; [d] Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium
E-mail address: Jerome.Lechien@umons.ac.be

Otolaryngol Clin N Am ■ (2024) ■–■
https://doi.org/10.1016/j.otc.2024.04.006 **oto.theclinics.com**

(ChatGPT) in November 2022 has led to a significant increase in public and practitioner interest in generative AI, and, particularly, artificial intelligence-powered language models (AILM).[1] Indeed, 2023 was the year associated with the highest number of publications dedicated to AI in otolaryngology—head and neck surgery and available on PubMED central (**Fig. 1**). Many AILM are available online for patients or practitioners, including ChatGPT (OpenAI, San Francisco, USA), Large Language Model Meta AI (Lamaa; MetaAI Palo Alto, CA, USA), Google Bidirectional Encoder Representations from Transformer (BERT; Mountain View, CA, USA), or Google DeepMind's Gopher.[2,3] AILM have been found to respond to simple-to-complicated questions related to clinical and basic science research,[4] referencing,[5,6] medical examinations,[7] clinical vignettes,[8] and they may improve scientific reports through spelling correction.[9,10] The mediatization, accessibility, and popularity of AILM may encourage patients to use them for medical and surgical education, while some young practitioners may consider AILM as adjunctive clinical tools to improve their knowledge and practice.[11] This article summarizes the application of AILM in Otolaryngology—Head and Neck Surgery.

HISTORY

The intelligent machine concept was born in Greek mythology where Hephaestus crafted golden robot-like statues to serve him.[12] In the Renaissance, Leonardo da Vinci imagined automatons capable of mimicking human actions. At the same time, in the eighteenth century, Gottfried Wilhelm Leibniz discussed a "universal language" and a machine that could reason, like human logical reasoning.[13] The technological revolution of the twentieth century made possible the development of AI-based processes in economic, law, or medical fields. Yet, in 1943, McCulloch and Pitts developed a computer model able to learn through a process that was comparable to human neurons.[14] A few years after this publication (1950), Alan Turing developed a test for model intelligence, which consisted of a blinded human interrogator questioning human and machine respondents (Turing test).[15,16] In 1956, AI was officially recognized at a Dartmouth College (Hanover, USA) conference in which researchers proposed the following statement "every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.[15,17]" Since then, more studies have been conducted in medicine and surgery to simulate, supplement, or efficiently augment human intelligence and skills in improving patient care.[15] The AI field currently involves machine learning and natural language processing subfields.[12] Machine learning consists of algorithms that learn from simple-to-complicated tasks for developing predictive models.[15]

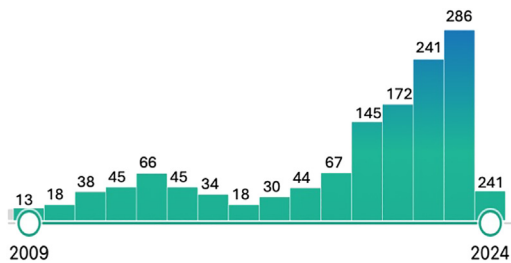

Fig. 1. Evolution of publications dedicated to artificial intelligence in otolaryngology in the past 15 years. (*Source*: figure was generated through PubMED (January 19, 2024) with the following key words: Artificial Intelligence Otolaryngology.)

Deep learning, which can be considered a subfield of machine learning, is based on artificial neural networks that can perform computations like the human brain. Recent developments in deep learning have led to many machine learning applications in medicine or surgery, which may analyze numerical data, clinical images, or videos through various databases.[12,15] AILM may consist of an association between machine learning and natural language processing and, consequently, may support practitioners in clinical decision-making, the proposition of additional examinations, or treatments.

BRIEF FUNCTIONING OF GENERATIVE ARTIFICIAL INTELLIGENCE

Generative artificial intelligence (AI) models are based on language processing, machine learning, and deep learning.[12,18–20] They include neural networks (mathematical models mimicking human brain functioning) and variational autoencoders, which rely on machine learning algorithms training on large databases, allowing them to recognize some specific patterns, understand complex relationships, and generate new outputs. In summary, the neural network is composed of many units (like neurons) connected by connections (weights). They communicate through inputs from the other units or the outside world, and may generate outputs for the others.[18,19] According to the received and potentially repeated inputs, units may learn from each other and adjust the related weights. AILM, such as GPT, are trained through a large corpus of text data used as input to a neural network composed of up to 175 billion parameters.[18,20] The weights between units may be improved with human use and related corrections. In practice, the input text is tokenized into individual words embedded into a vector space using the same embedding matrix used during training.[18] The embedded input text is subsequently passed via the encoder and decoder components. Then, AILM can generate creative, coherent, and contextually relevant sentences, making it a valuable tool for patient engagement, medical education, and clinical decision support.[20] It is important to note that the functioning of AILM is influenced by hyperparameters, which consist of settings controlling how the model learns from data, such as the learning rate, the batch size, the number of layers, and the activation function.[18,19] They are important because they influence the AILM performance, speed, and accuracy. In practice, they influence the content of the responses, leading to more coherent responses and better management of different inputs and outputs. For example, hyperparameters of ChatGPT-3.5 differ from hyperparameters of ChatGPT-4, which is the most recent model release. Importantly, all AILM have limitations, such as hallucination of facts (false positive), lack of common-sense knowledge, restricted context window, and potential privacy concerns.[18,20] All of them may be corrected by human feedback.

APPLICATIONS OF ARTIFICIAL INTELLIGENCE-POWERED LANGUAGE MODELS IN OTOLARYNGOLOGY

Patients have access to AILM to get information about symptoms of ear, nose, and throat conditions, surgery risks, and alternatives. In that way, several otolaryngologists have investigated the accuracy, precision, and performance of AILM, especially ChatGPT, in providing patient information.[11,18] Moreover, the accuracy of ChatGPT was similarly investigated in the management of theoretic or true clinical cases in the different subspecialties of otolaryngology head and neck surgery.[5–9] The current applications of ChatGPT in otolaryngology—head and neck surgery are summarized in **Fig. 2**.

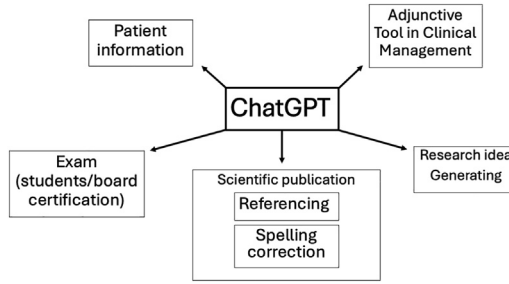

Fig. 2. Current applications of Chatbot Generative Pre-trained Transformer in otolaryngology—head and neck surgery.

Applications in General Otolaryngology

Student and resident board certification examinations

The ChatGPT accuracy in medical, surgical, and otolaryngologic questions was investigated from students or resident in-service examinations.[21–23] The first study was conducted by Hoch and colleagues to evaluate the accuracy of ChatGPT-3 on 2576 practice quiz questions designed for German otolaryngology board certification.[21] The authors reported an overall accuracy rate of 57% and observed that ChatGPT-3 responded better to single-choice questions than multiple-choice questions (34% vs 63%), while the performance of ChatGPT-3 was particularly high in allergology (72%), and low in the legal field (30%), respectively.[21] In the same vein, Mahajan and colleagues investigated the performance of ChatGPT-3.5 in responding to practice examination questions in otolaryngology head and neck surgery. The comparison of outputs from ChatGPT-3.5 with the benchmark of answers and explanations reported that ChatGPT-3.5 correctly answered 53% of the questions and provided correct explanations in 54% of the cases, respectively.[22] Long and colleagues submitted to ChatGPT-4.0 21 common questions of the licensing examination in otolaryngology, which were analyzed by 2 independent practitioners with the Concordance, Validity, Safety, and Competency model.[23] ChatGPT-4 scored 23.5/34 (accurate rate: 69.1%) but did not reach the minimum passing score for the examination (70%). However, after providing further queries that explicitly focus on otolaryngology, ChatGPT-4 improved its score to reach an accurate rate of 75%, demonstrating the ChatGPT performance improvement after human feedback.[23] The accuracy of ChatGPT-3.5 was similarly observed in responding to questions related to tympanostomy.[24] Twenty responses from ChatGPT-3.5 matched with the recommendations of the American Academy of Otolaryngology—Head and Neck Surgery, which consisted of an accuracy of 95.7%.[24]

Patient information

Several studies have been conducted to evaluate the accuracy of AILM in providing patient information related to otolaryngologic diseases or surgeries. In the study of Zalzal and colleagues, 2 sets of 30 text-based questions related to surgical anatomy, otology, head and neck surgery, oncology, laryngology, rhinology, and fundamentals were input into the ChatGPT-3.5 API.[25] Two board-certified otolaryngologists independently rated the chatbot's responses and observed total and partial response accuracy in 56.7%, and 86.7% of the cases, respectively. Interestingly, the authors observed that the repeated inputs led to an improvement of total and partial accurate responses to 73.3%, and 96.7%, respectively, which corroborated the findings of

Long and colleagues who observed improvement of performance through regenerated questions and feedback.[23] In the study of Langlie and colleagues, 2 independent practitioners assessed the capability and accuracy of ChatGPT-3.5 in providing indications, procedures, and alternative therapeutic options for adenotonsillectomy, tympanoplasty, endoscopic sinus surgery, parotidectomy, and total laryngectomy.[26] To achieve its goal, the authors interrogated ChatGPT-3.5 with standardized questions (*How do I know if I need [procedure]; What are treatment alternatives to [procedure]; What are the risks of [procedure]; How is a [procedure] performed; and What is the recovery process for [procedure]?*) and they did not observe major errors in ChatGPT-3.5 responses. However, the chatbot reported difficulties when it needed to provide precision and details in the surgery steps, forgetting key surgical steps and major risks associated with several surgeries.[26] The high accuracy of ChatGPT in providing health information in otolaryngology was tempered by Nielsen and colleagues who reported an overall 5-point Likert scale score of 3.41 for the ChatGPT-4 information related to otitis, hearing impairment, vertigo, epistaxis, rhinosinusitis, pharyngitis, dysphonia, globus sensation, and conjunctivitis.[27] To date, most studies investigating the AILM potential in providing otolaryngology information for patients focused on ChatGPT. Only 2 investigations compared ChatGPT performance with other AILM or databases.[28,29] The first one was conducted by Ayoub and colleagues who evaluated the accuracy of the outputs of Google Search and ChatGPT-4 with several recommendations from clinical practice guidelines.[28] The authors reported that the mean patient education material assessment tool scores for medical advice were 68.2% versus 89.4% for ChatGPT-4 and Google Search, respectively, meaning that Google Search scored better than ChatGPT-4 for providing readable information. The findings of Ayoub and colleagues corroborated those of Bellinger and colleagues who showed that ChatGPT-4 and Google Search similarly scored for treating the urgency of some clinical situations.[29] Note that the only field where ChatGPT-4 scored better than Google Search was the patient education questions (general medical knowledge; 87% vs 78%) according to the patient education material assessment tool.

Clinical vignettes

One of the first studies assessing the performance of ChatGPT in the management of clinical cases was conducted to validate an instrument dedicated to evaluating the ChatGPT performance (artificial intelligence performance instrument; AIPI).[7] In this study, ChatGPT-4 was accurate in proposing adequate additional examinations, treatments, and diagnoses in 29%, 22%, and 56% to 71%, respectively.[7] However, the chatbot proposed a significantly higher number of additional examinations than practitioners and did not select the most appropriate ones.[7] The performance of ChatGPT-3.5 in managing clinical cases was similarly studied by Dallari and colleagues who presented 10 clinical theoretic vignettes of common otolaryngologic symptoms with 2 different scenarios per case to ChatGPT-3.5.[30] Five otolaryngologists rated the responses of ChatGPT-3.5 using a 5-point Likert scale for difficulty, correctness, and consistency outcomes. The ChatGPT-3.5 scores of correctness and consistency were 3.80 and 2.89, respectively, without being influenced by the difficulty of the clinical cases.[30] The lack of influence of case difficulty on the ChatGPT performances was similarly observed by Lechien and colleagues in laryngology and head and neck surgery clinical cases,[31] which corroborated the findings of Dallari. The lack of influence of the level of difficulty of clinical cases on the accuracy of ChatGPT-4 was similarly supported in 2 other studies including general otolaryngology[8] and laryngology clinical cases.[31] In the study of Qu and colleagues, the authors investigated the accuracy of ChatGPT-4 on 20 clinical theoretic vignettes in

general otolaryngology[8] and they observed high and significant agreements between ChatGPT-4 and attending physicians in the propositions of adequate differential diagnoses and treatment plans.[8] To date, only Karimov and colleagues compared the accuracy of ChatGPT-3.5 with another AILM (UpToDate) in providing management information and references for 25 clinical cases.[32] The authors observed that ChatGPT-3.5 did not give references in some clinical questions in contrast to UpToDate that supported the information with subheadings, tables, figures, and algorithms from scientific papers. According to the assessment of experts, UpToDate was found to be more useful and reliable than ChatGPT-3.5.[32]

Applications in Head and Neck Surgery

General information and knowledge

One hundred and fifty-four questions related to all head and neck cancer basic knowledge, diagnosis, and treatments were input into the application programming interface (API) of ChatGPT-4 by Kuscu and colleagues who reported correct, partially correct, and incorrect GPT-4 responses in 86.4%, 11%, and 2.6% of the cases, respectively.[33] ChatGPT-4 reported highest accuracy for prevention (100%), diagnosis (92.6%), treatment (88.9%), while the questions related to surgical management, for example, recovery, risks, complications and follow-up, reached 80% of accurate outputs. The authors observed a high stability of ChatGPT-4 throughout regenerated questions with 94.1% of response stability, which corroborated findings of other studies.[30,31,34] The accuracy of ChatGPT-4 was similarly investigated for 144 theoretic questions encompassing different subspecialties of head and neck or maxillofacial surgery in a cross-sectional study involving 18 experts subdivided into 8 working groups for the output analysis.[5] The authors reported an overall accuracy score of 5.43 (6-point Likert scale) and noted that there were no significant differences between the several subspecialty scores in terms of completeness and accuracy scores.[5]

Patient information

Chiesa-Estomba and colleagues interrogated ChatGPT-3.5 with the 5 most common questions of patients toward head and neck cancer and asked patients to compare the outputs of ChatGPT-3.5 versus practitioners.[35] In laryngeal and oropharyngeal cancers, patients reported significantly preferring the responses of practitioners compared to the ChatGPT-3.5 outputs, while there were no significant differences for salivary gland cancers.[35]

The accuracy of ChatGPT-3.5 in providing information on oropharyngeal cancer information was similarly evaluated by Davis and colleagues who introduced 15 common questions into the API, whereas the outputs of ChatGPT-3.5 were analyzed by 4 independent head and neck surgeons using a 5-point Likert scale.[36] Thus, experts reported average ChatGPT-3.5 accuracy, comprehensiveness, and similarity scores ranging from 3.67 to 3.88, corresponding to somewhat accurate responses.[36] The ChatGPT-3.5 responses were particularly accurate in post-treatment information and less accurate in diagnosis-related information.[36] Contrary to other studies, the findings of this study suggested that ChatGPT-3.5 could outright misinform patients and read at a more difficult grade level than is recommended for patient material. Lee and colleagues collected presurgical educational information (indications, risks, and recovery time) for 5 common head and neck surgeries from ChatGPT-3.5 and 5 experienced head and neck surgeons compared ChatGPT-3.5 outputs with the information available on the first publicly available website.[4] The authors reported that ChatGPT-generated pre-surgical information was comparable to websites in terms of readability, content of knowledge, accuracy, thoroughness, and numbers of medical errors, which corroborated the

findings of some aforementioned studies.[4,24–26] In the field of endocrine surgery, Campbell and colleagues interrogated ChatGPT-3 for 30 questions related to thyroid nodule information and management.[37] Note that they input questions throughout 4 repeated sessions to assess the stability of the model. The authors observed an accurate rate of 69.2%, whereas 87.5% of the references provided by ChatGPT-3 were judged as legitimate citations, and 72.5% provided accurately reported information from the referenced publication.[37] The study, which was unique in the field of thyroid nodule information, supported a moderate-to-high accuracy of ChatGPT-3. Another original study was conducted in the sialendoscopy field by Chiesa-Estomba and colleagues who evaluated the accuracy of GPT-3.5 responses for providing clinical management of 6 salivary gland disorders and compared the ChatGPT-3.5 responses with those of 10 expert sialendoscopists.[38] The mean agreement score of experts was significantly higher compared to the ChatGPT-3.5 score (4.1 vs 3.4, 5-point Likert scale), while ChatGPT-3.5 and experts reported a comparable number of therapeutic alternatives. In this study, the authors observed that the expert treatment suggestions were rated higher than ChatGPT-3.5 suggestions in half of the clinical scenarios, whereas they were equal in the other half. Overall, the information provided by ChatGPT-3.5 was considered comprehensive and accurate.[38]

Clinical vignettes

A few studies investigated the accuracy of ChatGPT in managing clinical vignettes in head and neck surgery.[39–41] The accuracy of ChatGPT-3.5 was evaluated in recommending treatments for 727 head and neck cancer clinical vignettes through a comparison between ChatGPT-3.5 responses and the National Comprehensive Cancer Network Guidelines.[39] In this European cross-sectional study, the sensitivity and accuracy of ChatGPT-3.5 for primary treatments were 100% and 85.3%, respectively. The sensitivities for adjuvant treatments and follow-up indications wereboth 100%, whereas the accuracies were 95.6% and 94.1%, respectively. This study supported very high accuracy and sensitivity of ChatGPT-3.5 toward clinical vignettes of patients with head and neck cancer.[39] The findings of this study were corroborated in another European study where the findings of 20 medical records discussed in a multidisciplinary oncological board were input into the API to obtain cTNM explanations, and management propositions.[40] ChatGPT-4 was found to provide perfect explanations for cTNM staging in 19 cases (95%). In addition, ChatGPT-4 similarly indicated endoscopy-biopsy, human papilloma virus (HPV) research, ultrasonography, and PET-computed tomography (CT) to the oncological board. The therapeutic propositions of ChatGPT-4 were accurate in 13 cases (65%) but the number of proposed additional examinations was significantly higher compared to head and neck surgeons.[40] As found in other studies,[30,31,34] most additional examinations and primary treatment propositions were consistent throughout regenerated response process, which confirmed the high stability of ChatGPT-4. The last investigation conducted in head and neck oncology explored the ability of ChatGPT-4 to interpret confocal laser endomicroscopy images of normal versus cancerous oropharyngeal tissues.[41] In this study, the accuracy of ChatGPT-4 reached 71.2%, while the accuracy of the 3 experts, including 2 surgeons and 1 pathologist, was 88.5%.[41] This study is the only one that investigates the accuracy of ChatGPT-4 for analyzing clinical or histopathological images.

Laryngology and Broncho-Esophagology

Only 2 studies investigated the usefulness of ChatGPT in the field of laryngology and broncho-esophagology.[31,42] Then, the potential of ChatGPT-3.5 was investigated in

dysphagia for generating research ideas in swallowing science. The study protocol involved 26 swallowing experts who rated a list of study ideas generated by ChatGPT-3.5 with a Likert-scale ranging from 1 to 5 according to feasibility, novelty, clinical implications, and relevance to current practice. Experts reported a mean rate of rankings of research ideas (/5) of 4.03 $\pm$ 0.17 for feasibility, 3.5 $\pm$ 0.17 for potential impact on the field, 3.84 $\pm$ 0.12 for clinical relevance, and 3.08 $\pm$ 0.36 for novelty and innovation, respectively.[42] Authors concluded that ChatGPT-3.5 offered promising findings in generating research ideas in swallowing, but it is still limited in innovation. From a clinical standpoint, s team explored the accuracy and performance of ChatGPT-4 in the management of clinical cases from the laryngology clinic.[31] In sum, the accuracy of ChatGPT-4 for indicating adequate additional examinations was lower (10% to 33%) compared to its accuracy for providing primary diagnosis (65%) and the most adequate treatment (60% to 79%). Similarly to other studies,[7,34] ChatGPT-4 proposed a significantly higher number of additional examinations per patient compared to practitioners. Interestingly, the accuracy of ChatGPT did not vary according to the level of clinical cases, which corroborated the findings of other studies.[7,30]

Rhinology, Allergy, and Facial and Plastic Surgery

Patient information

Two studies conducted in the facial and plastic surgery field are available in the literature.[43,44] The first study was conducted to investigate the accuracy of ChatGPT information in facial and plastic surgery.[43] In this study, Capelleras and colleagues focused on the information provided by ChatGPT (unspecified version) in postoperative guidance during rhinoplasty recovery, for example, pain management, swelling, bruising, or asymmetries. The authors reported high performance of ChatGPT in patient education, especially in general information related to the recovery step and reassurance.[43] Another team compared the performance of ChatGPT-3.5 versus that of an experienced surgeon in providing patient education in rhinoplasty.[44] In this cross-sectional study, 7 facial and plastic surgeons used a 5-point Likert scale assessment to show that ChatGPT-3.5 outperformed surgeon responses in 75% performance areas, earning significantly higher ratings in accuracy, completeness, and overall quality. Experts preferred the ChatGPT-3.5 responses to those of the practitioner in 80.95% of instances.[44]

Clinical vignettes

Radulesco and colleagues investigated the accuracy of ChatGPT-4 in the management of 40 rhinologic and allergic cases.[34] According to the artificial intelligence performance instrument (AIPI) scores, 3 blinded rhinologists judged the ChatGPT-4 performance as higher for primary and differential diagnosis propositions (63.3%) than in indicating pertinent and necessary additional examinations (15.8%) or pertinent and necessary treatments (16.7%).[34] The authors regenerated 5 times the outputs of ChatGPT-4 and reported significantly high stability, especially in proposing therapeutic approaches. Interestingly, some differential diagnoses changed from the first output to the regenerated second, making the stability variable. In this study, the authors observed that some additional examinations were found to be more variable from one output to another, that is, psychophysical olfactory testing, which corroborated the finding of other clinical studies.[7,34] The inability of ChatGPT-4 to reliably propose some unusual additional examinations such as psychophysical olfactory testing, and impedance-pH monitoring was similarly observed in other otolaryngologic subspecialties, such as laryngology.[31] Saibene and colleagues investigated the accuracy of

ChatGPT-3.5 and ChatGPT-4 in the management of theoretic odontogenic chronic rhinosinusitis clinical vignettes.[45] The analyses of the 8 experts involved in this study confirmed the better performance of ChatGPT-4 over ChatGPT-3.5 but there were substantial disagreements between experts and ChatGPT in the management of 91.3% of the cases.[45]

Sleep Disorders

Several studies have been conducted to evaluate the accuracy of ChatGPT in general sleep knowledge or information for patients.[46–49] Cheong and colleagues compared the accuracy of ChatGPT-3.5, ChatGPT-4.0, and Google Bard in responding to 301 text-based single-best-answer multiple choice questions (10 examination categories) from the American Sleep Medicine Certification Board Examination.[46] Considering a pass mark of 80% for the examination, the authors found that ChatGPT-4 successfully achieved the pass mark with 80% or above in 5/10 examination categories. In this study, ChatGPT-4 demonstrated superior performance in all examination categories (68.1%) compared to ChatGPT-3.5 (46.8%), and Google Bard (45.5%), respectively.[46] ChatGPT-3.5 and Google Bard similarly scored in this study. This high accuracy of ChatGPT in responding to theoretic sleep questions was similarly supported by Mira and colleagues who found that ChatGPT-3.5 shared 75% of the responses of 97 sleep practitioners in a virtual examination.[47] Another study of Cheong and colleagues reported that the understandability and actionability scores of the Patient Education Materials Assessment Tool—Printable for ChatgGPT-3.5 and Google Bard ranged from 46% to 92% and 20% to 80%, respectively,[48] concluding that, as for the American Sleep Medicine Certification Board Exam, ChatGPT-3.5 scored better than Google Bard in providing understandable and actionable information related to sleep disorders.[48] The quality of ChatGPT's sleep apnea syndrome outputs for patient education was similarly evaluated by Campbell and colleagues who introduced 24 questions into the API, which were regenerated 4 times.[49] The authors observed that 69 ChatGPT-3.5 responses were at least correct, which corresponded to an accurate rate of 71.9%, while ChatGPT-3.5 provided adequate outputs in 96.1% of the questions.[49]

Otology and Vertigo

Bellinger and colleagues collected the 5 most common patient questions about benign paroxysmal positional vertigo from Google. They introduced them in the API to assess the readability, quality, understandability, and actionability of ChatGPT responses (unspecified version).[29] They reported that ChatGPT had higher Flesch-Kincaid Grade Level and lower Flesch Reading Ease scores than Google, indicating lower readability. Similar findings were found for the quality of responses, which led the authors to conclude that Google information is still superior to those provided by ChatGPT.[29] From a clinical performance standpoint, Chee and colleagues input into the API 8 theoretic vignettes of vertigo including medical history, types of prompts, or clinical pictures and reported that ChatGPT (unspecified version) succeeded in the diagnosis of 6/8 cases and differentiated well between vestibular and non-vestibular causes of dizziness.[50]

SUMMARY

The application of generative AI, particularly AILM, is booming in otolaryngology—head and neck surgery with ChatGPT as the main investigated model. The accuracy of ChatGPT appears high in providing general knowledge related to common

otolaryngologic conditions for patients, students, and practitioners. However, its accuracy should be moderate overall as an adjunctive tool in managing clinical vignettes. The rapid evolution of models and the increasing number of published studies each week make the future unpredictable and exciting in the revolution of otolaryngology—head and neck surgery practices.

ACKNOWLEDGMENTS

None.

DISCLOSURE

Competing interest: None. Sponsorships: None. Funding source: None.

REFERENCES

1. Kedia N, Sanjeev S, Ong J, et al. ChatGPT and Beyond: An overview of the growing field of large language models and their use in ophthalmology. Eye 2024. https://doi.org/10.1038/s41433-023-02915-z.
2. Tahayori B, Chini-Foroush N, Akhlaghi H. Advanced natural language processing technique to predict patient disposition based on emergency triage notes. Emerg Med Australasia (EMA) 2021;33(3):480–4.
3. Venerito V, Bilgin E, Iannone F, et al. AI am a rheumatologist: a practical primer to large language models for rheumatologists. Rheumatology 2023;62(10):3256–60.
4. Lee JC, Hamill CS, Shnayder Y, et al. Exploring the Role of Artificial Intelligence Chatbots in Preoperative Counseling for Head and Neck Cancer Surgery. Laryngoscope 2023. https://doi.org/10.1002/lary.31243.
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. https://doi.org/10.1002/ohn.489.
6. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. Eur Arch Oto-Rhino-Laryngol 2024;281(4):2159–65.
7. 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 Oto-Rhino-Laryngol 2024;281(4):2063–79.
8. Qu RW, Qureshi U, Petersen G, et al. Diagnostic and Management Applications of ChatGPT in Structured Otolaryngology Clinical Scenarios. OTO Open 2023; 7(3):e67.
9. Lechien JR, Gorton A, Robertson J, et al. Is ChatGPT-4 Accurate in Proofread a Manuscript in Otolaryngology-Head and Neck Surgery? Otolaryngol Head Neck Surg 2023. https://doi.org/10.1002/ohn.526.
10. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care 2023;27(1):75.
11. Chiesa-Estomba CM, Speth MM, Mayo-Yanez M, et al. Is the evolving role of artificial intelligence and chatbots in the field of otolaryngology embracing the future? Eur Arch Oto-Rhino-Laryngol 2024;281(4):2179–80.
12. Bur AM, Shew M, New J. Artificial Intelligence for the Otolaryngologist: A State of the Art Review. Otolaryngol Head Neck Surg 2019;160(4):603–11.
13. Panovski A. How Did Philosophy Help Develop Artificial Intelligence? The Collector 2023.

14. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33.
15. Muthukrishnan N, Maleki F, Ovens K, et al. Brief History of Artificial Intelligence. Neuroimaging Clin N Am 2020;30(4):393–9.
16. Turing AMI. Computing machinery and intelli- gence. Mind 1950;LIX(236): 433–60.
17. McCorduck P. Machines who think. 2nd edition. Natick (MA): A K Peters, Ltd; 2004.
18. Briganti G. How ChatGPT works: a mini review. Eur Arch Oto-Rhino-Laryngol 2024;281(3):1565–9.
19. Tolsgaard MG, Boscardin CK, Park YS, et al. The role of data science and machine learning in Health Professions Education: practical applications, theoretical contributions, and epistemic beliefs. Adv Health Sci Educ Theory Pract 2020; 25(5):1057–86.
20. Alter IL, Chan K, Lechien JR, et al. ChatGPT, ENT, and Me: An Introduction to Artificial Intelligence and Machine Learning for Otolaryngologists. Eur Arch Oto-Rhino-Laryngol 2024;281(5):2723–31.
21. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Oto-Rhino-Laryngol 2023; 280(9):4271–8.
22. Mahajan AP, Shabet CL, Smith J, et al. Assessment of Artificial Intelligence Performance on the OtolaryngologyResidency In-Service Exam. OTO Open 2023; 7(4):e98.
23. Long C, Lowe K, Zhang J, et al. A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology-Head and Neck Surgery Certification Examinations: Performance Study. JMIR Med Educ 2024;10:e49970.
24. Moise A, Centomo-Bozzo A, Orishchak O, et al. Can ChatGPT Guide Parents on Tympanostomy Tube Insertion? Children 2023;10(10):1634.
25. Zalzal HG, Cheng J, Shah RK. Evaluating the Current Ability of ChatGPT to Assist in Professional Otolaryngology Education. OTO Open 2023;7(4):e94.
26. Langlie J, Kamrava B, Pasick LJ, et al. Artificial intelligence and ChatGPT: An otolaryngology patient's ally or foe? Am J Otolaryngol 2024;45(3):104220.
27. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol 2023; 143(9):779–82.
28. Ayoub NF, Lee YJ, Grimm D, et al. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. Otolaryngol Head Neck Surg 2023. https://doi.org/10.1002/ohn.465.
29. Bellinger JR, De La Chapa JS, Kwak MW, et al. BPPV Information on Google Versus AI (ChatGPT). Otolaryngol Head Neck Surg 2023. https://doi.org/10.1002/ohn.506.
30. Dallari V, Sacchetto A, Saetti R, et al. Is artificial intelligence ready to replace specialist doctors entirely? ENT specialists vs ChatGPT: 1-0, ball at the center. Eur Arch Oto-Rhino-Laryngol 2024;281(2):995–1023.
31. Lechien JR, Georgescu BM, Hans S, et al. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. Eur Arch Oto-Rhino-Laryngol 2024;281(1):319–33.
32. Karimov Z, Allahverdiyev I, Agayarov OY, et al. ChatGPT vs UpToDate: comparative study of usefulness and reliability of Chatbot in common clinical presentations of

otorhinolaryngology-head and neck surgery. Eur Arch Oto-Rhino-Laryngol 2024; 281(4):2145–51.
33. Kuşcu O, Pamuk AE, Sütay Süslü N, et al. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol 2023;13: 1256459.
34. Radulesco T, Saibene AM, Michel J, et al. ChatGPT-4 performance in rhinology: A clinical case series. Int Forum Allergy Rhinol 2024. https://doi.org/10.1002/alr.23323.
35. Chiesa-Estomba CM, Urazan JD, Andueza M, et al. Comparative analysis of patient's perception between medical expert Vs chat-GPT advice for laryngeal, oropharyngeal, and salivary gland tumors. 2024. Scientific Presentation, San Sebastian Universty Hospital, Department of Otolaryngology, 2023.
36. Davis RJ, Ayo-Ajibola O, Lin ME, et al. Evaluation of Oropharyngeal Cancer Information from Revolutionary Artificial Intelligence Chatbot. Laryngoscope 2024; 134(5):2252–7.
37. Campbell DJ, Estephan LE, Sina E, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid 2024;34(3):371–7.
38. 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 Oto-Rhino-Laryngol 2024;281(4):2081–6.
39. Marchi F, Bellini E, Iandelli A, et al. Exploring the Landscape of AI-Assisted Decision-Making in Head and Neck Cancer Treatment: A Comparative Analysis of NCCN Guidelines and ChatGPT Responses. Eur Arch Oto-Rhino-Laryngol 2024; 281(4):2123–36.
40. Lechien JR, Chiesa-Estomba CM, Baudouin R, et al. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. Eur Arch Oto-Rhino-Laryngol 2024;281(4):2105–14.
41. Sievert M, Aubreville M, Muller S, et al. Confocal laser endomicroscopy, oropharyngeal squamous cell carcinoma, GPT, head and neck malignancies. Eur Arch Oto-Rhino-Laryngol 2024.
42. Nachalon Y, Broer M, Nativ-Zeltzer N. Using ChatGPT to Generate Research Ideas in Dysphagia: A Pilot Study. Dysphagia 2023. https://doi.org/10.1007/s00455-023-10623-9.
43. Capelleras M, Soto-Galindo GA, Cruellas M, et al. ChatGPT and Rhinoplasty Recovery: An Exploration of AI's Role in Postoperative Guidance. Facial Plast Surg 2024. https://doi.org/10.1055/a-2219-4901.
44. Durairaj KK, Baker O, Bertossi D, et al. Artificial Intelligence Versus Expert Plastic Surgeon: Comparative Study Shows ChatGPT "Wins" Rhinoplasty Consultations: Should We Be Worried? Facial Plast Surg Aesthet Med 2023. https://doi.org/10.1089/fpsam.2023.0224.
45. Saibene AM, Allevi F, Calvo-Henriquez C, et al. Reliability of large language models in managing odontogenic sinusitis clinical scenarios: a preliminary multidisciplinary evaluation. Eur Arch Oto-Rhino-Laryngol 2024;281(4):1835–41.
46. Cheong RCT, Pang KP, Unadkat S, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Oto-Rhino-Laryngol 2024;281(4):2137–43.
47. Mira FA, Favier V, Dos Santos Sobreira Nunes H, et al. Chat GPT for the management of obstructive sleep apnea: do we have a polar star? Eur Arch Oto-Rhino-Laryngol 2024;281(4):2087–93.

48. Cheong RCT, Unadkat S, Mcneillis V, et al. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. Eur Arch Oto-Rhino-Laryngol 2024;281(2):985–93.
49. Campbell DJ, Estephan LE, Mastrolonardo EV, et al. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med 2023;19(12):1989–95.
50. Chee J, Kwa ED, Goh X. "Vertigo, likely peripheral": the dizzying rise of ChatGPT. Eur Arch Oto-Rhino-Laryngol 2023;280(10):4687–9.