Check for updates

AMERICAN ACADEMY OF
OTOLARYNGOLOGY–
HEAD AND NECK SURGERY
FOUNDATION

Original Research—Patient Safety and Quality Improvement

Otolaryngology–
Head and Neck Surgery
2024, Vol. 170(6) 1492–1503
© 2023 The Authors.
Otolaryngology–Head and Neck
Surgery published by Wiley
Periodicals LLC on behalf of
American Academy of
Otolaryngology–Head and Neck
Surgery Foundation.
DOI: 10.1002/ohn.489
http://otojournal.org
WILEY

Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis

Luigi Angelo Vaira, MD[1,2]*, Jerome R. Lechien, MD, PhD[3,4]*, Vincenzo Abbate, MD[5], Fabiana Allevi, MD[6], Giovanni Audino, MD[5], Giada Anna Beltramini, MD[7,8], Michela Bergonzani, MD[9], Alessandro Bolzoni, MD[7], Umberto Committeri, MD[5], Salvatore Crimi, MD[10], Guido Gabriele, MD[11], Fabio Lonardi, MD[12], Fabio Maglitto, MD[13], Marzia Petrocelli, MD[14], Resi Pucci, MD[15], Gianmarco Saponaro, MD[16], Alessandro Tel, MD[17], Valentino Vellone, MD, PhD[18], Carlos Miguel Chiesa-Estomba, MD[19], Paolo Boscolo-Rizzo, MD[20], Giovanni Salzano, MD[5], and Giacomo De Riu, MD[1]****

Abstract

Objective. To investigate the accuracy of Chat-Based Generative Pre-trained Transformer (ChatGPT) in answering questions and solving clinical scenarios of head and neck surgery.

Study Design. Observational and valuative study.

Setting. Eighteen surgeons from 14 Italian head and neck surgery units.

---

[1] Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy
[2] Biomedical Sciences Department, PhD School of Biomedical Science, University of Sassari, Sassari, Italy
[3] Department of Anatomy and Experimental Oncology, Mons School of Medicine, UMONS, Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium
[4] Department of Otolaryngology–Head Neck Surgery, Elsan Polyclinic of Poitiers, Poitiers, France
[5] Head and Neck Section, Department of Neurosciences, Reproductive and Odontostomatological Science, Federico II University of Naples, Naples, Italy
[6] Maxillofacial Surgery Department, ASSt Santi Paolo e Carlo, University of Milan, Milan, Italy
[7] Department of Biomedical, Surgical and Dental Sciences, University of Milan, Milan, Italy
[8] Maxillofacial and Dental Unit, Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico, Milan, Italy
[9] Maxillo-Facial Surgery Division, Head and Neck Department, University Hospital of Parma, Parma, Italy
[10] Operative Unit of Maxillofacial Surgery, Policlinico San Marco, University of Catania, Catania, Italy
[11] Department of Maxillofacial Surgery, University of Siena, Siena, Italy
[12] Department of Maxillofacial Surgery, University of Verona, Verona, Italy
[13] Maxillo-Facial Surgery Unit, University of Bari "Aldo Moro", Bari, Italy
[14] Maxillofacial Surgery Operative Unit, Bellaria and Maggiore Hospital, Bologna, Italy
[15] Maxillofacial Surgery Unit, San Camillo-Forlanini Hospital, Rome, Italy
[16] Maxillo-Facial Surgery Unit, IRCSS "A. Gemelli" Foundation—Catholic, University of the Sacred Heart, Rome, Italy
[17] Department of Head and Neck Surgery and Neuroscience, Clinic of Maxillofacial Surgery, University Hospital of Udine, Udine, Italy
[18] Maxillofacial Surgery Unit, "S. Maria" Hospital, Terni, Italy
[19] Department of Otorhinolaryngology–Head and Neck Surgery, Hospital Universitario Donostia, San Sebastian, Spain
[20] Department of Medical, Surgical and Health Sciences, Section of Otolaryngology, University of Trieste, Trieste, Italy

*These authors contributed equally to this article and should be considered co-first authors.
**These authors contributed equally to this article and should be considered co-senior authors.

Corresponding Author:
Luigi Angelo Vaira, MD, Viale San Pietro 43/B, 07100 Sassari, Italy.
Email: lavaira@uniss.it

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Methods. A total of 144 clinical questions encompassing different subspecialities of head and neck surgery and 15 comprehensive clinical scenarios were developed. Questions and scenarios were inputted into ChatGPT4, and the resulting answers were evaluated by the researchers using accuracy (range 1-6), completeness (range 1-3), and references' quality Likert scales.

Results. The overall median score of open-ended questions was 6 (interquartile range[IQR]: 5-6) for accuracy and 3 (IQR: 2-3) for completeness. Overall, the reviewers rated the answer as entirely or nearly entirely correct in 87.2% of cases and as comprehensive and covering all aspects of the question in 73% of cases. The artificial intelligence (AI) model achieved a correct response in 84.7% of the closed-ended questions (11 wrong answers). As for the clinical scenarios, ChatGPT provided a fully or nearly fully correct diagnosis in 81.7% of cases. The proposed diagnostic or therapeutic procedure was judged to be complete in 56.7% of cases. The overall quality of the bibliographic references was poor, and sources were nonexistent in 46.4% of the cases.

Conclusion. The results generally demonstrate a good level of accuracy in the AI's answers. The AI's ability to resolve complex clinical scenarios is promising, but it still falls short of being considered a reliable support for the decision-making process of specialists in head-neck surgery.

Keywords

artificial intelligence, ChatGPT, maxillofacial surgery, otorhinolaryngology

Received April 27, 2023; accepted July 14, 2023.

Chat-Based Generative Pre-trained Transformer (ChatGPT)[1] is an advanced artificial intelligence (AI) language model developed by Open Artificial Intelligence. The first version was released in November 2022 and quickly became the fastest-growing application in history, counting 100 million active users as of February 2023.[2] It is based on the Generative Pre-trained Transformer 4 architecture and is designed to understand and generate human-like text responses in a conversation. ChatGPT has been trained on a wide range of data sources, making it capable of providing information, answering questions, and engaging in conversation across various topics.

The potential impact of ChatGPT on health care is substantial. First, ChatGPT could be used to assist health care professionals in diagnosing medical conditions by analyzing patient symptoms, medical history, and other relevant information. This could lead to faster and more accurate diagnoses, improving patient outcomes.[3-5] Second, ChatGPT can help synthesize and analyze vast amounts of medical literature, which could lead to the discovery of new treatments, medications, or a better understanding of diseases. The potential of ChatGPT in medical writing is currently the most studied and discussed in the literature.[6-8] Third, ChatGPT could be used as a supplementary learning tool by medical students and professionals, as it can provide explanations, answer questions, and help review medical concepts.[9-11] Finally, ChatGPT could power virtual assistants that help patients manage their health, provide information about medications and treatments, and answer common health-related questions.[12-14]

However, it is essential to consider the ethical implications and data privacy concerns related to using AI in health care, as well as to ensure that AI systems are properly validated and tested before implementation. This is important because this large amount of information is accessible to a lay public that does not have the means to criticize the information and can create unattainable expectations, misinformation, or influence the relationship between the patient and the professional.[15,16]

Despite the extensive literature discussing the potential benefits and possible negative aspects of using ChatGPT in scientific research processes,[17,18] there remains a significant gap in understanding its applications in clinical settings. In particular, the quality of information provided by ChatGPT to health care professionals and patients has not been thoroughly explored.[13,14,19-23] Additionally, there is a noticeable absence of research addressing the role of ChatGPT in head and neck pathology.

The primary objective of this study was to comprehensively assess the quality of responses generated by ChatGPT across various branches of head and neck surgery. This investigation also aims to evaluate the quality of ChatGPT's analysis of commonly encountered clinical scenarios in head and neck surgery departments, such as differential diagnoses, treatment planning, and patient counseling.

By conducting this in-depth examination, we hope to shed light on the practical applications and limitations of ChatGPT in the specialized field of head and neck surgery, providing insights into the potential use of AI in assisting health care professionals in diagnosing and treating patients, as well as enhancing patient education and understanding of their conditions. Moreover, this study seeks to contribute to the ongoing discourse on the ethical implications and responsibilities associated with utilizing AI-based tools in medical practice.

Materials and Methods

Setting

In February 2023, a research group was established to investigate the potential applications of AI platforms in head and neck surgery. This collaborative effort involved 18 surgeons from 14 different Italian head and neck surgery centers, encompassing a wide range of expertise

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by CochraneItalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and experiences. The research group aimed to explore the practicality, effectiveness, and reliability of AI platforms in various head and neck surgery-related tasks and decision-making processes. The requirement for an ethical review and approval was waived because the study did not include any analysis of humans or animals.

Working Group

For the purposes of this study, the participating researchers were strategically divided into 8 working groups based on their respective areas of expertise. This division was designed to ensure comprehensive coverage of the different areas of head and neck surgery, including: pathology of the salivary glands, oral oncology, reconstructive surgery, facial nerve reanimation; sinonasal pathology; traumatology of the mandible, the mandibular condyle and the middle and upper third of the face, malformation pathology, orthognathic surgery, pathology of the temporomandibular joint and oral surgery.

Outcomes

Each group was tasked with leveraging their specific expertise to develop a set of 6 clinical questions that would challenge the AI platform. The 6 questions were divided as follows: 3 open-ended questions and 3 closed-ended questions (true or false).

For each of the 2 categories, researchers were asked to generate 3 questions of increasing difficulty (easy, medium, and hard). The difficulty level of each question was subjectively assessed by the researchers, based on their expertise and understanding of the subject matter as previously described.[23]

Closed-ended questions were included to compel the AI system to provide a definitive answer, while open-ended questions allowed for a broader exploration of the AI system's problem-solving abilities. The researchers were instructed to propose questions that could potentially yield universally accepted answers. To ensure that the questions were well-structured and sufficiently specific, the researchers were asked to frame the questions in a comprehensive manner, providing enough context and information for the AI system to generate an absolute response. Furthermore, the researchers prepared a series of 15 comprehensive clinical scenarios, drawing inspiration from genuine cases that were assessed at their respective affiliated medical facilities. Each scenario includes detailed patient histories, along with signs and symptoms exhibited by the patients, with the primary objective of evaluating the AI platform's proficiency in accurately framing the patient's clinical context and subsequently proposing a suitable diagnostic pathway. The scenarios encompassed cases in traumatology, oral oncology, and salivary gland pathology.

Upon completion, all questions and clinical scenarios were submitted for collective review and approval by the entire research group. This process facilitated a thorough evaluation of the questions' quality, relevance, and consistency within the context of head and neck surgery applications. Any necessary revisions and refinements were made during this stage, ensuring that the final set of questions was both challenging and appropriate for testing the AI platform. Binary questions were included only if all researchers agreed on the answer, otherwise, the question was modified or elaborated on until full consensus was reached. The complete set of clinical questions and scenarios is presented in Supplemental Tables S1-S3, available online.

ChatGPT Analysis

In order to ensure consistency in the study, a single researcher inserted all the questions and clinical scenarios into the ChatGPT version 4 on March 27, 2022. For each question, the researcher instructed the AI to provide specific answers, taking into account all available guidelines before submitting the question. In cases where a dichotomous response was expected, the AI was asked to respond with either "true" or "false" exclusively. Additionally, for the clinical scenarios, the AI was requested to identify the most probable diagnosis and to outline the most appropriate diagnostic pathway for the described case. For both open-ended questions and clinical scenarios, the AI was also asked to provide all bibliographic references from which it drew information to formulate the response.

All responses obtained through this process were meticulously recorded and provided to the researchers for evaluation. In this way, each answer was subsequently evaluated by 18 different researchers. As described by Johnson et al,[23] the researchers assessed the accuracy and completeness of the open-ended answers and the clinical scenarios using 2 predefined scales. For accuracy, a 6-point Likert scale was employed, with 1 representing a completely incorrect response, 2 denoting more incorrect than correct, 3 indicating an equal balance of correct and incorrect elements, 4 signifying more correct than incorrect, 5 representing nearly all correct, and 6 being entirely correct. As for completeness, a 3-point Likert scale was used: 1 stood for an incomplete answer that only addressed some aspects of the question with significant parts missing or incomplete, 2 represented an adequate answer that addressed all aspects of the question and provided the minimum information required for completeness, and 3 denoted a comprehensive response that covered all aspects of the question and offered additional information or context beyond expectations. Researchers were asked not to express a judgment if they did not know the topic and were therefore unable to evaluate the answer. Dichotomous questions were evaluated solely on the basis of whether the responses were correct or incorrect. The bibliographic references were preliminarily evaluated by 2 independent reviewers who discharged references to nonexistent sources.

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by CochraneItalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The remaining references were then subjected to analysis by the reviewers, alongside the corresponding answer provided by the AI. The reviewers assessed the relevance of the proposed question, the quality of the bibliographic references, their timeliness, and the extent to which they justified the given answer.[24] Based on these criteria, the reviewers expressed a judgment on the quality of the bibliographic references using a Likert scale ranging from 1 (no or poor quality of bibliographic references) to 5 (excellent quality of bibliographic references). The clinical scenarios were also provided to a second-year head and neck surgery resident, who was asked to respond to the proposed questions in his native language. The resident was given the opportunity, just like the AI, to access information available on the internet and from any bibliographic source if deemed necessary. The responses provided by the AI and the surgeon were blindly evaluated by 3 independent senior reviewers using the same methodology mentioned above.

Statistical Analysis

Statistical analyses were performed using Jamovi version 2.3.18.0, a freeware and open statistical software available online at www.jamovi.org.[25] Categorical variables are reported in numerals and percentages of the total. Descriptive statistics for quantitative variables are given as the median (interquartile range [IQR]) or mean ± standard deviation. The interrater reliability of accuracy and completeness scores for open-ended questions and clinical scenarios was evaluated using Cronbach's α. Differences in accuracy and completeness scores between groups of different difficulty levels and distinct question categories were assessed using the Kruskal-Wallis test. To evaluate the differences between each subgroup, a post hoc analysis was performed with the Dwass-Steel-Critchlow-Fligner method. The differences in the proportion of correct dichotomous responses among various subgroups were assessed using Fisher's exact test. The differences in accuracy, completeness, and references' quality scores between the responses provided by the second-year resident and ChatGPT were evaluated using the Wilcoxon signed-rank test. The level of statistical significance was set at $P < 0.05$ with a 95% confidence interval.

Results

The analysis encompassed a total of 15 clinical scenarios and 144 questions (72 open-ended and 72 binary) spanning across 12 subspecialties within head and neck surgery. The complete set of questions and answers can be found in Supplemental Tables S1-S3, available online. **Table 1** presents a few examples of questions with varying levels of difficulty and the corresponding answers provided by ChatGPT.

Open-Ended Questions

A total of 1072 evaluations were collected and analyzed from the 18 reviewers, who assessed 72 open-ended questions (224 missing evaluations in the data set). The overall median score for accuracy was 6 (IQR: 5-6; mean 5.43 ± 1.03), while the median score for completeness was 3 (IQR: 2-3; mean 2.7 ± 0.55). An excellent internal consistency was observed among the reviewers' assessments for both accuracy (Cronbach's $\alpha = .974$) and completeness (Cronbach's $\alpha = .955$) scores. The differences in accuracy scores between groups of questions, subdivided by difficulty levels, were found to be statistically significant (Kruskal-Wallis $\chi^2 = 10.9$; $P = 0.004$). Post hoc analysis revealed significant differences between easy (median 6 [IQR: 5-6]; mean 5.55 ± 0.83) and intermediate (median 6 [IQR: 5-6]; mean 5.52 ± 0.84) questions compared to hard questions (median 6 [IQR: 5-6]; mean 5.22 ± 1.31), but no significant differences were observed between easy and intermediate questions themselves (**Figure 1A**). The differences in completeness scores were also statistically significant (Kruskal-Wallis $\chi^2 = 7.74$; $P = 0.021$) but only between easy (median 3 [IQR: 3-3]; mean 2.74 ± 0.48) and hard (median 3 [IQR: 2-3]; mean 2.62 ± 0.6) questions and not between easy and intermediate (median 3 [IQR: 3-3]; mean 2.72 ± 0.51) and between intermediate and hard questions (**Figure 1B**). No significant differences were found between the accuracy (Kruskal-Wallis $\chi^2 = 13.9$; $P = 0.238$) and completeness (Kruskal-Wallis $\chi^2 = 19.7$; $P = 0.177$) scores between the questions in the different subspecialties of head and neck surgery considered. Overall, the reviewers rated the answer as entirely or nearly entirely correct (ie, accuracy score 5 or 6) in 87.2% of cases and as comprehensive and covering all aspects of the question (ie, completeness score 3) in 73% of cases (**Table 2**). Considering the median of the scores reported by the reviewers for each question: all questions except 1 reported accuracy values of at least 5 and 93.1% reported a completeness score of 3 (Supplemental Table S1, available online). A total of 218 bibliographic references were provided for the 72 open-ended questions. Of these, 109 (50%) referred to nonexistent articles. Out of the remaining 109 references,[22] (10.1%), contained partial errors in the authors' list, year, volume, and page references, or in the digital object identifier. Finally, 87 references (39.9%), were found to be completely correct. Upon review, the median quality score for the sources used by ChatGPT was determined to be 1 (IQR: 1-3). Only 10 responses were supported by bibliographic sources deemed to be of good or excellent quality (eg, a quality score of 4 or 5) (Supplemental Table S1, available online).

Closed-Ended Questions

The analysis of 72 responses provided by ChatGPT to true-false forced-choice questions revealed that the AI model achieved a correct response in 84.7% of the cases (11 out of 72 wrong answers) (Supplemental Table S2,

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by Cochraneltalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Examples of Questions With Varying Levels of Difficulty and the Corresponding Answers Provided by ChatGPT

Category	Question	ChatGPT-4 answer
Open-ended Reconstruction Easy	Before performing which type of reconstructive flap is it appropriate to perform the Allen test?	Allen's test is typically performed before setting up a radial forearm free flap, which is a reconstructive procedure that uses skin, subcutaneous tissue, and fascia from the forearm to repair defects in other parts of the body, particularly the head and neck region. The radial forearm free flap is often favored for its thin and pliable tissue, making it suitable for various reconstructive purposes. Allen's test is used to assess the blood supply and patency of the ulnar artery, which is important because the radial artery is usually sacrificed during the flap harvest. The test ensures that the hand will still receive adequate blood supply after the procedure. If Allen's test indicates inadequate blood flow through the ulnar artery, the surgeon may consider an alternative flap or donor site to minimize the risk of complications.
Open-ended Reconstruction Medium	What tests should be performed to assess the feasibility of a free fibula flap?	Before performing a free fibula flap, a thorough preoperative evaluation is necessary to ensure the safety and viability of the flap. The primary goal is to confirm adequate blood supply in the lower leg and foot, as the peroneal artery, which supplies the fibula flap, will be sacrificed during the procedure. Two common tests to evaluate the vascular supply are: 1. *Ankle-Brachial Index (ABI)*: ABI is a noninvasive test that compares the blood pressure in the ankle to the blood pressure in the arm. It helps to assess the overall arterial circulation in the lower extremities and can identify any peripheral arterial disease that may contraindicate the use of a fibula flap. 2. *Doppler ultrasound*: A Doppler ultrasound can provide real-time visualization of the blood flow in the lower leg vessels, including the peroneal, anterior tibial, and posterior tibial arteries. It helps to evaluate the patency of these vessels and identify any anatomic variations that may impact the flap harvest. In addition to these tests, some surgeons may perform additional imaging studies, such as CT angiography or MR angiography, to further assess the vascular anatomy and the precise location of the perforating vessels that supply the fibula flap. Finally, clinical examination is essential, and the surgeon may perform a manual test, such as the capillary refill test or the toe pulse test, to confirm adequate blood flow in the foot before proceeding with the flap harvest.
Open-ended Reconstruction Hard	Which technical modification of the free fibula flap allows for the restoration of proper bone thickness and facilitates the placement of dental implants?	The double-barrel fibula-free flap is a technical modification that allows for the re-establishment of correct bone thickness and facilitates implant rehabilitation. In this technique, the harvested fibula bone is longitudinally split into 2 segments, which are then stacked on top of each other to create a double-layered, or "double-barrel," construct. This double-barrel construct provides increased bone thickness and height, which can better mimic the native jaw bone's dimensions. This technique is particularly useful in reconstructing the mandible or maxilla after resection due to cancer or other causes, as it offers several advantages for dental implant rehabilitation:

(*continued*)

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by CochraneItalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. (continued)

Category	Question	ChatGPT-4 answer
		1. *Enhanced bone thickness*: The double-barrel fibula-free flap provides increased bone thickness, which is crucial for the proper placement and stability of dental implants. 2. *Improved osseointegration*: The increased bone thickness can lead to better osseointegration and the long-term success of dental implants. 3. *Easier implant placement*: The double-barrel construct offers a larger surface area for implant placement, which can facilitate the surgical procedure and improve the chances of successful implant integration. 4. *Improved aesthetics and function*: With the double-barrel fibula free flap, the reconstructed jawbone can more closely resemble the native jaw's dimensions, leading to better aesthetic and functional outcomes, including improved dental occlusion and facial contour. It is essential to consider that the double-barrel fibula-free flap may not be suitable for all patients and situations, and the decision to use this technique should be made on a case-by-case basis in consultation with the patient and a multidisciplinary team of specialists. The patient's overall health, the size of the defect, and the quality of the fibula bone should be taken into consideration when planning the reconstruction.
Closed-ended Oral oncology Easy	Leukoplakia is a precancerous condition	True
Closed-ended Oral oncology Medium	Oral tongue Squamous cell carcinoma of 0.5 cm with a DOI < 3 mm is classified as T2.	False
Closed-ended Oral oncology Hard	Elective neck dissection is indicated in a cT4cN1cM0 squamous cell carcinoma of the retromolar trigone	True
Clinical scenario n°1	I saw a 43-year-old male patient in good general health. The patient reports having detected an asymptomatic rounded subcutaneous lesion of about 2 cm in diameter immediately below the left ear for about 3 months. For 10 days the patient has had difficulty moving the left side of the mouth. In your opinion what is the most probable cause? What tests do you recommend to perform?	I am not a doctor, but I can try to provide some general information that you could discuss with a health care professional. It is essential to consult with a medical professional for a thorough examination and accurate diagnosis. Based on the information provided, one potential cause to consider might be a parotid gland lesion, such as a benign or malignant tumor. Facial nerve weakness could be a sign of involvement of the facial nerve (cranial nerve VII), which runs through the parotid gland. However, there are various other potential causes for the symptoms, so it is crucial to get a professional evaluation.To determine the underlying cause, a health care professional may recommend the following tests: 1. *Physical examination*: A thorough examination of the head and neck region, including assessment of the cranial nerves, can provide essential information. 2. *Imaging studies*: Ultrasound, CT, or magnetic resonance imaging can help evaluate the lesion and surrounding structures, including the parotid gland and facial nerve.

(*continued*)

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by CochraneItalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. (continued)

Category	Question	ChatGPT-4 answer
		3. *Fine needle aspiration biopsy*: If the lesion appears suspicious, a health care professional may recommend a fine needle aspiration biopsy to collect a sample of cells or tissue for further examination under a microscope. This can help determine if the lesion is benign or malignant. 4. Blood tests: In some cases, blood tests may be useful to assess for potential infections or inflammatory conditions that might be causing the symptoms. It is important to consult with a health care professional who can evaluate the patient's specific situation and recommend the most appropriate diagnostic tests and treatment options.

Abbreviations: ChatGPT, Chat-Based Generative Pre-trained Transformer; CT, computed tomography; DOI, depth of invasion.

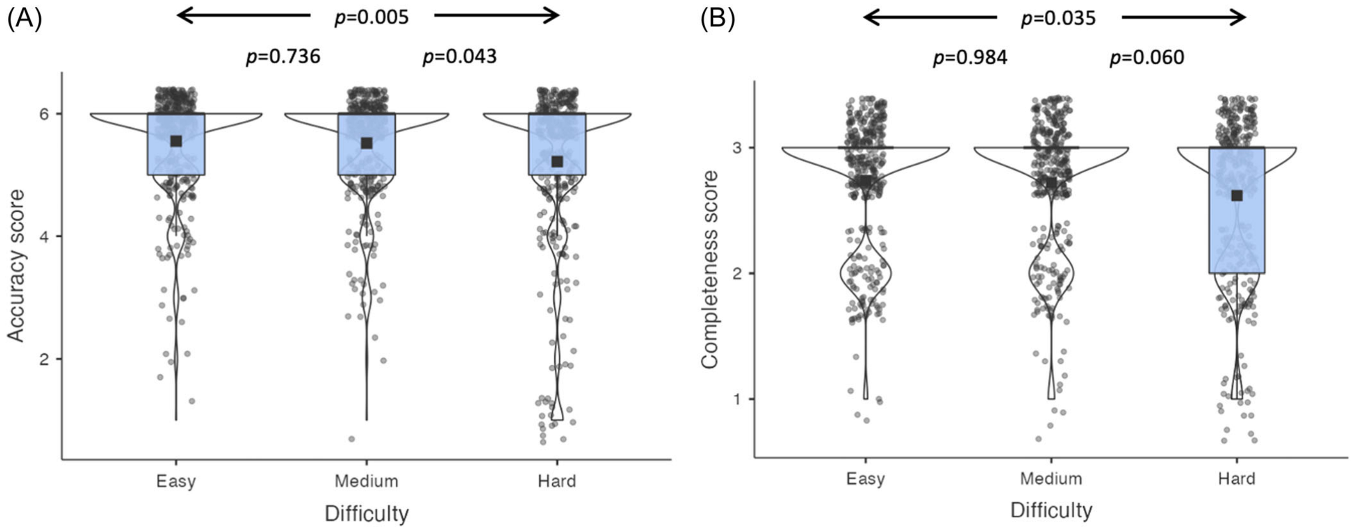

Figure 1. Accuracy (A) and completeness (B) scores of the open-ended questions divided by difficulty.

Table 2. Accuracy and Completeness Evaluation Results of Open-Ended Questions

Accuracy			Completeness		
Score	Counts	% of total	Score	Counts	% of total
1	19	1.8	1	39	3.6
2	13	1.2	2	251	23.4
3	30	2.8	3	782	73
4	75	7			
5	222	20.7			
6	713	66.5			

available online). Fisher's exact test identified significant differences in error rates depending on the subspecialty in question. Specifically, ChatGPT exhibited considerably higher error rates in the fields of malformative pathology (83.3%), reconstructive surgery (50%), and condylar traumatology (33.3%) (**Table 3**). The differences between the various subgroups of questions, identified by the level of difficulty, were not significant ($P = 1$) (**Table 3**).

Clinical Scenarios

Overall, 268 assessments of 15 clinical scenarios were included and analyzed (2 assessments missing from the

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by CochraneItalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. Results of the Analysis of the Correctness of the Closed-Ended Answers

Subspeciality	Correct answer	% of total	Difficulty	Correct answer	% of total
Facial reanimation	6	100	Easy	20	83.3
Sinonasal pathology	6	100	Medium	20	83.3
Orthognathic surgery	5	83.5	Hard	21	87.5
Temporomandibular joint pathology	6	100			
Salivary gland pathology	6	100			
Middle and upper-third traumatology	6	100			
Mandibular traumatology	6	100			
Condylar traumatology	4	66.7			
Oral oncology	6	100			
Reconstructive surgery	3	50			
Malformative pathology	1	15.3			
Oral surgery	6	100			

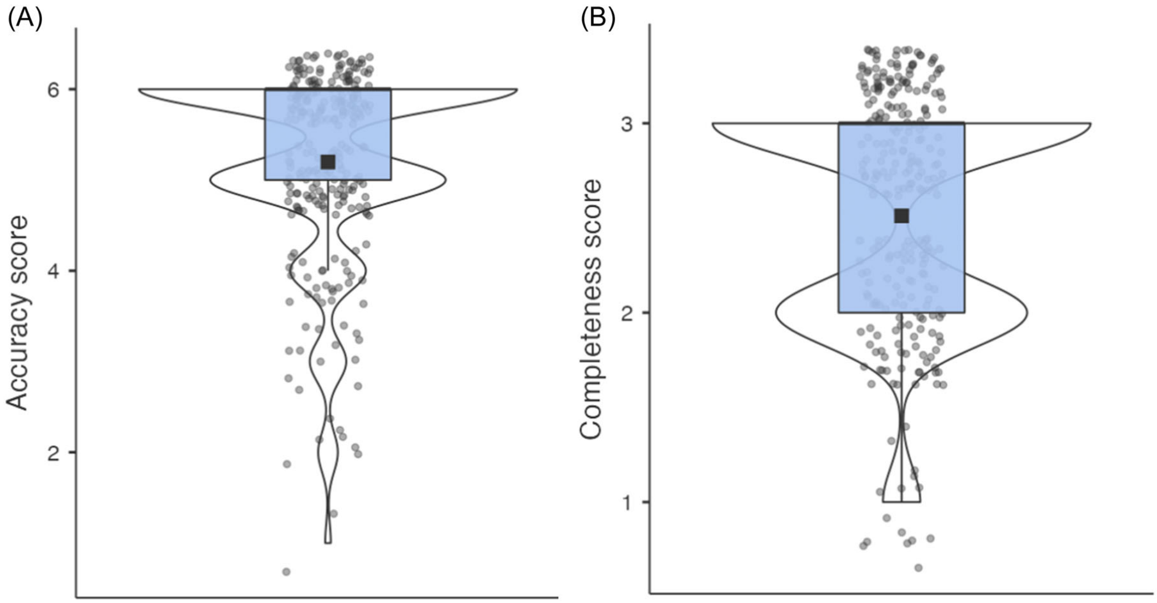

Figure 2. Accuracy (A) and completeness (B) scores of the clinical scenarios.

data set). The interrater reliability was high for both accuracy (Cronbach's $\alpha = .875$) and completeness scores (Cronbach's $\alpha = .792$). ChatGPT provided a fully or nearly fully correct diagnosis in 81.7% of cases (median accuracy score 6 [IQR: 5-6], mean 5.2 ± 1.06) (**Figure 2A**). The proposed diagnostic or therapeutic procedure was judged to be complete in 56.7% of cases (median completeness score 3 [IQR 2-3], mean 2.51 ± 0.6) (**Figure 2B**). Considering the median of the scores reported by the reviewers for each question: all questions except 1 reported accuracy values of at least 5 and 60% reported a completeness score of 3 (Supplemental Table S3, available online). In support of the 15 clinical scenario responses, ChatGPT provided a total of 49 bibliographic references. Out of these, 15 (30.6%) were found to be nonexistent. Of the remaining 34 references (69.4%) that indeed had a corresponding source, 10 (20.4%) still contained errors. Overall, the median references' quality score was determined to be 2 (IQR: 1.5-4), and only 4 responses were supported by bibliographic references evaluated as good or excellent (Supplemental Table S3, available online). The comparative analysis of the quality of responses provided by ChatGPT and the second-year resident (Supplemental Table S4, available online) revealed significant differences in terms of accuracy ($P = 0.001$) completeness ($P = 0.011$), and references' quality score ($P < 0.001$) (**Figure 3**).

Discussion

AI chatbox such as ChatGPT holds the potential to drastically transform the way we provide assistance to patients and how patients access health information.[20,26-30] However, the use of AI health care brings several ethical issues that need to be carefully considered and are not yet

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by CochraneItalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by Cochraneltalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

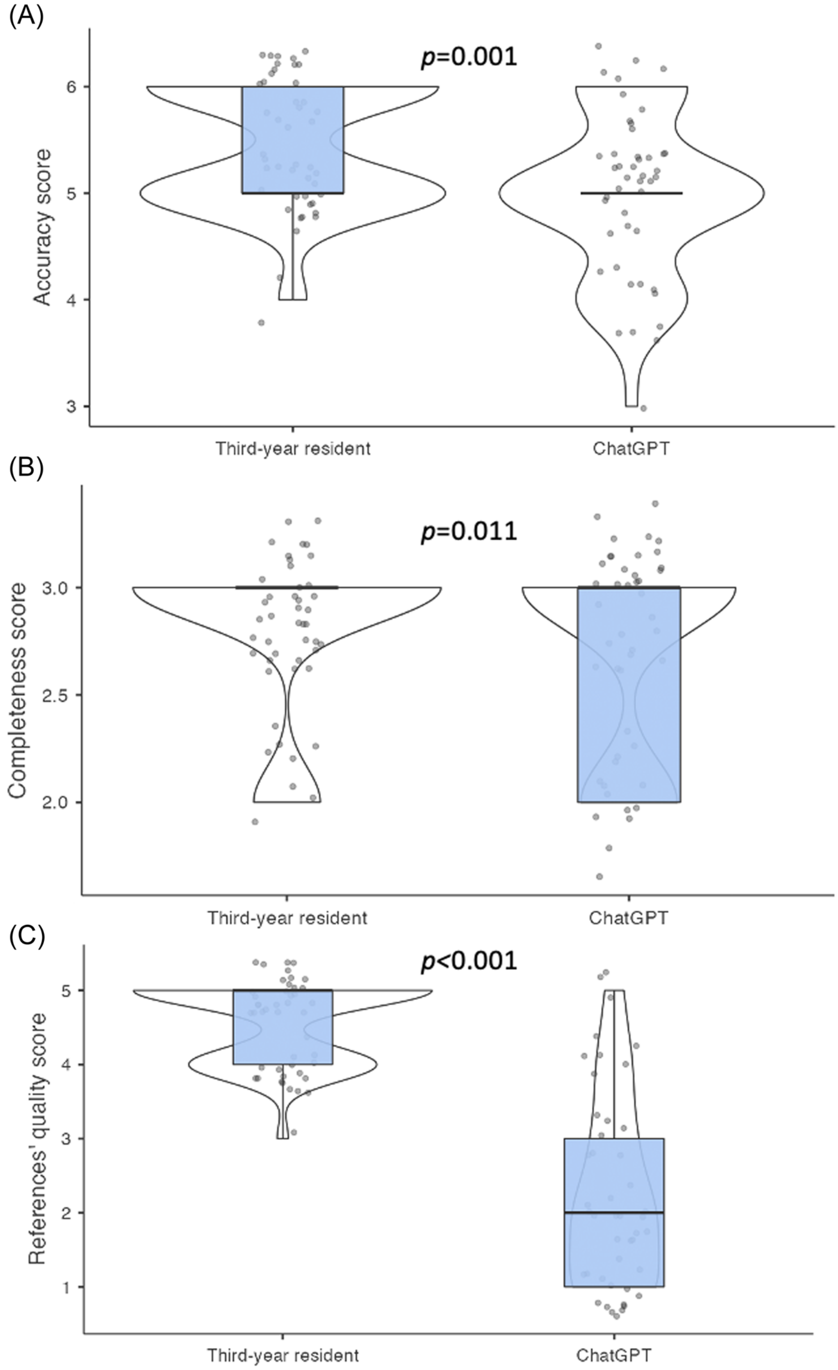

Figure 3. Comparative analysis of the differences in accuracy (A), completeness (B), and quality of reference (C) scores reported by ChatGPT and the second-year resident.

explored and clarified. First, ChatGPT processes and stores sensitive health care information, including personal details and medical records that are entered by patients in the chatbox.[31] Ensuring the privacy and security of this data is crucial to avoid unauthorized access, data breaches, or identity theft.[32] Second, potential medicolegal aspects related to the use of ChatGPT remain to be fully elucidated and will be a key focus of future investigation. These include determining accountability for incorrect or misleading information provided by the AI and establishing guidelines for its use in different clinical settings to ensure patient safety and uphold professional ethical standards.[33] Third, patients need to be aware that they are interacting with an AI chatbot and not a human health care provider and they must be provided with information about the system's capabilities and limitations. In all responses to clinical scenarios included in this study, where the AI was asked to provide diagnostic or therapeutic guidance, ChatGPT consistently began its answers by acknowledging that it is not a doctor or health care professional.

Moreover, it always concluded by advising the individual to consult with a specialist, often even suggesting the specific type of specialist to approach. This is of utmost importance, as AI should not replace medical professionals, but rather serve as a valuable tool to assist patients in navigating the appropriate diagnostic and therapeutic pathways.

Fourth, AI systems may provide incorrect or misleading information, which can have negative consequences in a health care setting. Furthermore, it is important to note that the answers provided by ChatGPT may not be up-to-date, as they are currently based on information available up until September 2021. Consequently, any advances, discoveries, or changes in medical practice and knowledge made after this date may not be reflected in the AI's responses. For these reasons, it is essential to constantly evaluate the accuracy of health information provided by AI language models such as ChatGPT, research studies in this field are of utmost importance. The present study, focusing on head and neck pathology, demonstrated good levels of accuracy and completeness in open-ended responses, as well as a strong ability of ChatGPT to synthesize answers and express forced judgments in closed-ended questions. These results outperform those reported in studies involving other types of diseases or medical branches.[20-23,34,35] A key factor in these improved results is likely related to the fact that ChatGPT version 4 was used in the present study. This latest version, released on March 13, 2023, is characterized by a significant improvement in reasoning skills and integrating information to generate answers.[13] Compared to what was reported for the previous version,[23] it was infrequent to find answers which, although well-argued, reached completely wrong conclusions (eg, Supplemental Table S1, available online, question oral oncology—hard 1). In open-ended questions, the AI demonstrated lower accuracy in responding to difficult questions as compared to easy and intermediate ones. This is likely due to the increased challenge of performing complex reasoning and assimilating a larger amount of information required to address harder questions. Similarly, ChatGPT proved to be less effective in providing forced answers to closed-ended questions on ultraspecialist topics such as malformation pathology on which it is more difficult to find information to integrate.

ChatGPT has demonstrated promising analytical capabilities in solving complex clinical scenarios based on the same clinical information that the patients provided to the specialist during the first visit. However, the unacceptable figure of 18.3% completely or partially incorrect diagnoses, coupled with the 43.3% incompleteness rate, renders ChatGPT still far from having a reliable supportive role in clinical decision-making for head-neck surgeons. The quality of responses provided by ChatGPT was inferior to that of a second-year resident in terms of both accuracy and completeness. The risk of ChatGPT providing incorrect responses holds serious implications in medical decision-making, potentially leading to erroneous actions and harmful patient outcomes. This highlights the critical importance of ensuring the reliability and accuracy of such AI tools when applied within health care contexts. Therefore, developing rigorous guidelines and protocols for assessing these systems, and improving their accuracy and reliability, becomes an urgent requirement for their future integration into health care practice.

In the present study, 46.4% of the bibliographic references provided by ChatGPT4 were found to be nonexistent. Though some researchers have hypothesized a potential for ChatGPT in conducting systematic literature reviews,[36] it has been reported by several authors that version 3.5 of ChatGPT has the issue of generating nonexistent bibliographic references in over 80% of instances.[37,38] The results obtained from this study reveal a lower frequency of nonexistent bibliographic references compared to these values, but the rate remains intolerable.

The generation of nonexistent references presents a series of problems. First, it undermines the reliability and credibility of the tool, making it unfit for serious academic or professional use. Second, it may mislead users and result in misinformation, as individuals may not verify the sources and could assume the data provided is accurate. Lastly, in the case of students or researchers, it could lead to waste of valuable time and effort in tracking down nonexistent references, and ultimately to the production of work based on inaccurate or fabricated information. Therefore, significant improvements in the accuracy of ChatGPT's bibliographic references are necessary before they can be used effectively in scholarly and professional contexts. Furthermore, the quality of the bibliographic references provided by ChatGPT was generally poor, not including systematic reviews or randomized trials. Without references to high-quality evidence, the risk of introducing bias and providing potentially misleading information increases.

The present study has several limitations that need to be acknowledged. First, the data set of questions used is limited and cannot be considered representative of the entire knowledge base surrounding head and neck pathology. The generalizability of this study is further limited due to the fact that not all areas within the otorhinolaryngology–head and neck surgery were covered. The selection of difficulty level for both open-ended and multiple-choice questions was subjective and solely based on the researcher's judgment, even though the questions were later evaluated collectively. Furthermore, the questions were restricted to topics with universally accepted answers, leaving it uncertain how the AI might perform on currently debated subjects. However, it should be noted that the AI consistently attempted to present areas of uncertainty and alternative hypotheses in its responses. Moreover, the ChatGPT AI model is constantly learning and evolving, which means that it cannot be ruled out that its responses may vary over time. It must also be acknowledged that ChatGPT4's

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by Cochraneltalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

knowledge is updated only up to September 2021, and this could change in the near future. Lastly, assessing the fidelity of ChatGPT responses comes with inherent limitations. Foremost, there's a lack of validated evaluation systems, necessitating the development of such tools in the future. Second, the evaluation can be subjective, with varying opinions among individuals regarding response accuracy; despite this, the present study maintained high interrater reliability. Lastly, as ChatGPT lacks a consistent personality, it can provide inconsistent responses to similar or identical queries. ChatGPT, like many deep learning models, is largely a black box, meaning its internal workings are not easily interpretable. This makes it difficult to understand why a given response was generated, which complicates the assessment of fidelity.

Conclusions

In conclusion, this study represents the first comprehensive evaluation of the accuracy of ChatGPT's responses in the context of head and neck surgery. The results generally demonstrate a good level of accuracy in the AI's answers. The AI's ability to resolve complex clinical scenarios is promising, but it still falls short of being considered a reliable support for the decision-making process of specialists in head-neck surgery.

Numerous ethical concerns will need to be addressed and resolved in the future, but with the proper precautions in place, ChatGPT has the potential to improve health care delivery and patient outcomes by assisting head and neck surgeons in making better-informed decisions.

Author Contributions

Luigi Angelo Vaira, conceptualization of the work, development of the methodology, data curation, writing the original draft, writing the final draft, final approval; **Jerome R. Lechien**, development of the methodology, writing the original draft, final approval; **Vincenzo Abbate**, data collection, data curation, revision of the original end final draft, final approval; **Fabiana Allevi**, data collection, data curation, revision of the original end final draft, final approval; **Giovanni Audino**, data collection, data curation, revision of the original end final draft, final approval; **Giada Anna Beltramini**, data collection, data curation, revision of the original end final draft, final approval; **Michela Bergonzani**, data collection, data curation, revision of the original end final draft, final approval; **Alessandro Bolzoni**, data collection, data curation, revision of the original end final draft, final approval; **Umberto Committeri**, data collection, data curation, revision of the original end final draft, final approval; **Salvatore Crimi**, data collection, data curation, revision of the original end final draft, final approval; **Guido Gabriele**, data collection, data curation, revision of the original end final draft, final approval; **Fabio Lonardi**, data collection, data curation, revision of the original end final draft, final approval; **Fabio Maglitto**, data collection, data curation, revision of the original end final draft, final approval; **Marzia Petrocelli**, data collection, data curation, revision of the original end final draft, final approval; **Resi Pucci**, data collection, data curation, revision of the original end final draft, final approval; **Gianmarco Saponaro**, data collection, data curation, revision of the original end final draft, final approval; **Alessandro Tel**, data collection, data curation, revision of the original end final draft, final approval; **Valentino Vellone**, data collection, data curation, revision of the original end final draft, final approval; **Carlos Miguel Chiesa-Estomba**, provision of study instrumentation, development of the methodology, supervision, review of the first and final draft, final approval; **Paolo Boscolo-Rizzo**, data collection, data curation, revision of the original end final draft, final approval; **Giovanni Salzano**, conceptualization of the work, development of the methodology, data curation, writing the original draft, writing the final draft, final approval; **Giacomo De Riu**, conceptualization of the work, development of the methodology, data curation, writing the original draft, writing the final draft, and final approval.

Disclosures

Competing interests: None.

Funding source: None.

Supplemental Material

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

ORCID iD

Luigi Angelo Vaira https://orcid.org/0000-0002-7789-145X

References

1. OpenAI. ChatGPT. 2023. Accessed March 28, 2023. https://openai.com/blog/chatgpt
2. Exploding Topics. Number of ChatGPT users 2023. 2023. Accessed March 30, 2023. https://explodingtopics.com/blog/chatgpt-users
3. Barat M, Soyer P, Dohan A. Appropriateness of recommendations provided by ChatGPT to interventional radiologists. *Can Assoc Radiol J*. Published online April 13, 2023. doi:10.1177/08465371231170133
4. Cheng K, Sun Z, He Y, Gu S, Wu H. The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? *Int J Surg*. 2023;109:1545-1547. doi:10.1097/JS9.0000000000000388
5. Strong E, DiGiammarino A, Weng Y, et al. Performance of ChatGPT on free-response, clinical reasoning exams. *medRxiv*. Published online March 29, 2023. doi:10.1101/2023.03.24.23287731
6. Zimmerman A. A Ghostwriter for the masses: ChatGPT and the future of writing. *Ann Surg Oncol*. 2023;30:3170-3173. doi:10.1245/s10434-023-13436-0
7. Gupta R, Herzog I, Weisberger J, Chao J, Chaiyasate K, Lee Es. Utilization of ChatGPT for plastic surgery research: friend or foe? *J Plast Reconstr Aesthet Surg*. 2023;80:145-147.
8. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R. A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiol*. 2023;52:1755-1758. doi:10.1007/s00256-023-04340-5

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by Cochraneltalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

9. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. Published online March 14, 2023. doi:10.1002/ase.2270
10. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9:e46885.
11. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. *Pak J Med Sci*. 2023;39:605-607.
12. Kahambing JG. ChatGPT, public health communication and 'intelligent patient companionship'. *J Public Health*. Published online April 8, 2023. doi:10.1093/pubmed/fdad028
13. Cox A, Seth I, Xie Y, Hunter-Smith DJ, Rozen WM. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. *Aesthet Surg J*. 2023;43:NP658-NP662. doi:10.1093/asj/sjad096
14. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EHC, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*. Published online March 13, 2023. doi:10.1111/aos.15661
15. Al Ghamdi KM, Moussa NA. Internet use by the public to search for health-related information. *Int J Med Inform*. 2012;81:363-373.
16. Tonsaker T, Bartlett G, Trpkov C. Health information on the internet: gold mine or minefield? *Can Fam Physician*. 2014;60:407-408.
17. Van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023; 614:224-226.
18. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379:313.
19. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv*. Published online February 7, 2023. doi:10.1101/2023.02.02.23285399
20. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47:33.
21. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;228:696-705. doi:10.1016/j.ajog.2023.03.009
22. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv*. Published online February 26, 2023. doi:10.1101/2023.02.21.23285886
23. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq*. Published online February 28, 2023. doi:10.21203/rs.3.rs-2566942/v1
24. Hooshafza S, Mc Quaid L, Stephens G, Flynn R, OConnor L. Development of a framework to assess the quality of data sources in healthcare settings. *J Am Med Inform Assoc*. 2022; 29:944-952.
25. The jamovi project. Jamovi. (version 2.3) [Computer Software]. 2022. Accessed january 24, 2023. https://www.jamovi.org
26. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' view. *J Med Internet Res*. 2020;22:e16775.
27. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25:1248-1258.
28. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11:887.
29. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectrum*. 2023;7:pkad010.
30. Goodman RS, Patrinely, Jr. JR, Osterman T, Wheless L, Johnson DB. On the cusp: considering the impact of artificial intelligence language models in healthcare. *Med*. 2023;4:139-140.
31. Zhang J, Zhang Z. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decis Mak*. 2023;23:7.
32. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. *Med Teach*. 2023;45:574-584. doi:10.1080/0142159X.2023.2186203
33. 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
34. Cheng K, Li Z, He Y, et al. Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. *Ann Biomed Eng*. 2023;51:1130-1135. doi:10.1007/s10439-023-03203-3
35. Cheng K, Li Z, Guo Q, Sun Z, Wu H, Li C. Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. *Int J Surg*. 2023;109:1816-1818. doi:10.1097/JS9.0000000000000410
36. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. 2023;12:72.
37. Frosolini A, Gennaro P, Cascino F, Gabriele G. In reference to "role of Chat GPT in public health", to highlight the AI's incorrect reference generation. *Ann Biomed Eng*. Published online 2023 May 22, 2023. doi:10.1007/s10439-023-03248-4
38. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol J*. 2023. doi:10.1177/08465371231171125

10976817, 2024, 6, Downloaded from https://aao-hnsfjournals.onlinelibrary.wiley.com/doi/10.1002/ohn.489 by Cochraneltalia, Wiley Online Library on [10/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License