Quality of Information Provided by Artificial Intelligence Chatbots Surrounding the Management of Vestibular Schwannomas: A Comparative Analysis Between ChatGPT-4 and Claude 2

*Daniele Borsetto, †Egidio Sia, *Patrick Axon, *Neil Donnelly, *James R. Tysome, ‡Lukas Anschuetz, §Daniele Bernardeschi, ||Vincenzo Capriotti, ¶Per Caye-Thomasen, ¶Niels Cramer West, **Isaac D. Erbele, ††Sebastiano Franchella, †Annalisa Gatto, ‡‡Jeanette Hess-Erga, §§||||Henricus P.M. Kunst, ¶¶John P. Marinelli, ***Richard Mannion, †††Benedict Panizza, ‡‡‡Franco Trabalzini, §§§Rupert Obholzer, ||||||Luigi Angelo Vaira, ¶¶¶Jerry Polesel, ¶¶¶Fabiola Giudici, ¶¶Matthew L. Carlson, †Giancarlo Tirelli, and †Paolo Boscolo-Rizzo

**Department of Otolaryngology–Head & Neck Surgery, Cambridge University Hospitals NHS Trust, Cambridge, UK; †Department of Medical, Surgical and Health Sciences, Section of Otolaryngology, University of Trieste, 34149, Trieste, Italy; ‡Department of Otorhinolaryngology, Head and Neck Surgery, Inselspital, University Hospital and University of Bern, Bern, Switzerland; §AP-HP6, GHU Pitié-Salpêtrière, Service ORL, Unité Fonctionnelle Implants Auditifs et explorations fonctionnelles, Paris, France; ||Department of Otorhinolaryngology, Portogruaro Hospital, UlSS4 Veneto Orientale, Portogruaro, VE, Italy; ¶Department of Otorhinolaryngology, Head and Neck Surgery, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark; **Department of Otolaryngology–Head and Neck Surgery, Brooke Army Medical Center, Fort Sam Houston, Texas, USA; ††Section of Otorhinolaryngology–Head and Neck Surgery, Department of Neurosciences, University of Padova, 35128 Padova, Italy; ‡‡Department of Otorhinolaryngology, Head and Neck Surgery, Haukeland University Hospital, 5021 Bergen, Norway; §§Department of Otorhinolaryngology and Head & Neck Surgery, Academic Alliance Skullbase Pathology MUMC+-Radboudumc, Radboud University Medical Center, Nijmegen, The Netherlands; ||||School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, the Netherlands; ¶¶Department of Otolaryngology–Head and Neck Surgery, Mayo Clinic, Rochester, Minnesota, USA; ***Department of Neurosurgery, Addenbrooke's Hospital, Cambridge University Hospitals, Cambridge, UK; †††Department of Otolaryngology–Head & Neck surgery, Princess Alexandra Hospital, Brisbane, QLD, Australia; ‡‡‡Department of Otolaryngology, Ospedale Pediatrico Meyer, Firenze, Italy; §§§Victor Horsley Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK; ||||||Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy; ¶¶¶Unit of Cancer Epidemiology, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Aviano, Italy*

Objective: To examine the quality of information provided by artificial intelligence platforms ChatGPT-4 and Claude 2 surrounding the management of vestibular schwannomas.

Study design: Cross-sectional.

Setting: Skull base surgeons were involved from different centers and countries.

Intervention: Thirty-six questions regarding vestibular schwannoma management were tested. Artificial intelligence responses were subsequently evaluated by 19 lateral skull base surgeons using the Quality Assessment of Medical Artificial Intelligence (QAMAI) questionnaire, assessing "Accuracy," "Clarity," "Relevance," "Completeness," "Sources," and "Usefulness."

Main Outcome Measure: The scores of the answers from both chatbots were collected and analyzed using the Student *t* test. Analysis of responses grouped by stakeholders was performed with McNemar test. Stuart-Maxwell test was used to compare reading level among chatbots. Intraclass correlation coefficient was calculated.

Results: ChatGPT-4 demonstrated significantly improved quality over Claude 2 in 14 of 36 (38.9%) questions, whereas higher-quality scores for Claude 2 were only observed in 2 (5.6%) answers. Chatbots exhibited variation across the dimensions of "Accuracy," "Clarity," "Completeness," "Relevance," and "Usefulness," with ChatGPT-4 demonstrating a statistically significant superior performance. However, no statistically significant difference was found in the assessment of "Sources." Additionally, ChatGPT-4 provided information at a significant lower reading grade level.

Conclusions: Artificial intelligence platforms failed to consistently provide accurate information surrounding the management of vestibular schwannoma, although ChatGPT-4 achieved significantly higher scores in most analyzed parameters. These findings demonstrate the potential for significant misinformation for patients seeking information through these platforms.

Key Words: Acoustic neuroma—AI—Artificial intelligence—Chatbots—ChatGPT—Claude—GPT—QAMAI—Vestibular schwannomas—VS.

Otol Neurotol **46:**432–436, 2025.

Address correspondence and reprint requests to Egidio Sia, M.D., Address: Strada di Fiume, 447, Trieste (TS), CAP 34149, Italy. Email: sia.egidio@gmail.com

Egidio Sia, https://orcid.org/0000-0003-0343-9070

Paolo Boscolo-Rizzo, https://orcid.org/0000-0002-4635-7959

Daniele Borsetto and Egidio Sia contributed equally to the paper.

The work of Drs. Polesel and Giudici is partially supported by the Italian Ministry of Health–Ricerca Corrente (no grant number available).

The authors disclose no conflicts of interest.

Supplemental digital content is available in the text.

DOI: 10.1097/MAO.0000000000004410

INTRODUCTION

Artificial intelligence (AI) is transforming healthcare. Large Language Models (LLMs) constitute a category of AI that is engineered to process and generate text that emulates human language based on patterns and information learned from vast amounts of data. Autoregressive language models work by generating text one word or character at a time, considering the preceding words to predict the next one. This process is similar to how humans construct sentences (1). Chat Generative Pre-trained Transformer (ChatGPT, version 4) and Claude 2 have emerged as sophisticated models within the realm of natural language processing (2,3). Developed through state-of-the-art deep learning methodologies, they have undergone thorough training across diverse and expansive datasets, empowering them to generate responses in conversations with users that try to emulate human language (4).

In the medical field, although numerous papers analyze the capabilities of ChatGPT to provide answers and to assist physicians in the diagnostic and therapeutic processes, there is limited evidence on the use of Claude 2, which has the ability to handle around 75,000 words per prompt—12 times the standard amount offered by ChatGPT-4. The Claude 2 platform represents a newer AI with most recent data availability compared to ChatGPT: Claude 2 has been trained on data up to early 2023, whereas GPT's knowledge cutoff is September 2021. ChatGPT has displayed proficiency in a range of specialized fields, such as obesity, stereotactic radiosurgery, child healthcare, dermatology, radiology, and ophthalmology (5–10). Furthermore, it demonstrates notable capabilities in assisting with decision-making in medical treatment and diagnosis (11–13). Although AI may not always show a sufficient accuracy in providing medical answers, it still holds promising potential to evolve into a valuable tool for both patients and healthcare professionals and therefore needs evaluation. Furthermore, it must be considered that the implementation of ChatGPT in medical care raises concerns regarding privacy, legal, and ethical issues (14–21).

Application of these AI platforms to the study of vestibular schwannoma has received little attention to date. Vestibular schwannomas are rare tumors arising from the vestibular nerve that can lead to symptoms like hearing loss, tinnitus, dizziness, imbalance, facial numbness, and headaches. They represent a complex pathology with many points that are still debated among experts, such as choice of different managements strategies (observation, surgical resection, and radiotherapy), function preservation (hearing, balance, facial nerve), and quality of life. Notably, in the context of a disease where many patients receive multiple recommendations on the most appropriate course of treatment, patients oftentimes seek information online. With AI technology becoming ubiquitously available, patients are expected to implement this technology when attempting to learn about their vestibular schwannoma and its treatment. For these reasons, the central objective of this paper was to examine the quality of information provided by ChatGPT-4 and Claude 2 surrounding various aspects of management in vestibular schwannomas. Specifically, questions of varying difficulty levels were analyzed, simulating potential questions that may be asked by patients, general physicians, and ENT surgeons.

MATERIALS AND METHODS

Questions and Answers Generation

A team of three researchers (P.B.R., D.B., E.S.) developed a collection of 36 questions (Supplemental Digital Content 1, http://links.lww.com/MAO/C35) spanning different facets of VS management. The questions were all written in American English, reviewed by the research team, and adjusted for any inaccuracies or ambiguities until unanimous agreement was achieved, with particular attention to the fluency and clarity of the language used. All raters were experts in the field and fluent in English. Questions were equally divided into three groups, representing various research possibilities tailored to stakeholders with diverse expertise levels, thereby encompassing a range of complexities and specificities. The first group included questions that could be posed by patients. The second group consisted of questions suitable for individuals with medical knowledge but not specialized expertise in skull base surgery, such as general practitioners and other healthcare professionals. The last group addressed more in-depth questions that could be raised by surgeons with experience of managing vestibular schwannomas, reflecting higher complexity. The questions were directed to both ChatGPT-4 and Claude 2, and the answers were subsequently collected. Specifically, the questions in the first group were prompt engineered with the instruction, "Please provide patient education material in response to the following question, ensuring it is accessible at a fifth-grade reading level." Then, a second request was made after the initial answer was given, using the phrase: "Please provide bibliographic references to your sources." For each question posed, a fresh chat instance was generated to ensure that the responses remained uninfluenced by previous answers.

Evaluation

Twenty skull base surgeons were involved from different centers and countries to reflect variations in international practice. They were asked to evaluate the answers according to the Quality Assessment of Medical Artificial Intelligence (QAMAI) tool, in order to define the quality information provided by AIs (22). Thus, all answers generated by ChatGPT-4 and Claude 2 were evaluated by each expert for six parameters ("Accuracy," "Clarity," "Relevance," "Completeness," "Sources," "Usefulness"), and a score ranging from 1 (Strongly Disagree) to 5 (Strongly Agree) was assigned for each parameter.

Additionally, answers belonging to the first group were assessed using the Flesch Kincaid Calculator to ascertain if the AIs adhered to the request to answer at a fifth-grade reading level (23).

The execution of this study did not require the approval of an ethics committee as it did not involve patients or animals. The study was conducted in accordance with the Helsinki Declaration principles.

Statistical Analysis

Experts' evaluations were reported overall and by each parameter, grouping scores into three levels: strongly disagree

TABLE 1. *Overall evaluation (mean value) of ChatGPT-4 and Claude 2*

Parameter	ChatGPT-4	Claude 2	Student *t* Test
Accuracy	129.5	118.4	$p < 0.001$
Clarity	145.5	137.4	$p = 0.007$
Relevance	132.7	127.2	$p = 0.048$
Completeness	119.1	108.5	$p = 0.001$
Sources	99.5	100.3	$p = 0.847$
Usefulness	124.7	118.3	$p = 0.042$

and disagree, neutral, agree and strongly agree. Differences between ChatGPT-4 and Claude 2 were evaluated through the McNemar test (a statistical test for dependent classification factors or paired proportions). For each expert, an overall score was further calculated summing up the rates over the 36 questions; overall score was summarized as mean values, and differences between chatbots were compared through Student *t* tests. Inter-rater reliability was analyzed using intraclass correlation coefficient (ICC) considering the score as an ordinal variable. ICC estimates and their 95% confidence intervals were calculated based on individual-surgeon rating, absolute-agreement, two-way random-effects model (24). An ICC below 0.50 was considered to indicate poor; >0.50 and ≤0.75, moderate; >0.75 and ≤0.90, good; and >0.90, excellent correlations. Heatmap correlation matrix was used to visualize the strength of relationships between the six parameters. The Stuart-Maxwell marginal homogeneity test was used to compare ChatGPT-4 versus Claude 2 according to the Flesch Kincaid Calculator evaluation of answers. We used R version 4.2.3 (R Core Team 2023) and the following R packages: Likert (version 1.3.5), irr (version 0.84.1), and DescTools (version 0.99.54). All statistical tests were two-sided, and $p < 0.05$ were considered statistically significant.

RESULTS

Out of the 20 skull base surgeons contacted, 18 agreed to participate, whereas 1 more expert was recommended by a surgeon already involved. Finally, a total of 19 skull base surgeons participated in the study, evaluating all the answers according to the QAMAI questionnaire. Initial comparisons between the two chatbots were made for each question by considering the sum of the scores collected for each parameter. Sixteen out of 36 questions showed significantly different scores between ChatGPT-4 and Claude 2: specifically, ChatGPT-4 showed higher scores for 14 answers (38.9%), whereas for 2 answers (5.6%), higher scores were observed in Claude 2 (Supplemental Digital Content 2, http://links.lww.com/MAO/C35).

A further comparison among chatbots was conducted by analyzing scores across different parameters, as shown in Table 1. Significantly higher scores in "Accuracy," "Clarity," "Relevance" (bordering significance), "Completeness," and "Utility" were observed for ChatGPT-4, whereas the parameter "Sources" did not show a statistically significant difference between the two AIs.

ChatGPT-4 also exhibited a statistically significant higher rate of agreement/strong agreement than Claude 2 (Fig. 1). "Sources" was the only parameter that did not reach statistical significance between the two chatbots ($p = 0.300$). Agreement of experts with the answers by ChatGPT-4 ranged from 46 to 77% and from 35 to 69% for Claude 2.

A more detailed analysis was conducted, grouping the answers by stakeholders (Table 2). The answers generated by ChatGPT-4 and Claude 2 on questions belonging to the patients group and non-ENT group did not differ in all the considered parameters, although ChatGPT-4 showed a nonsignificant superiority on overall evaluation. On the contrary, within the ENT surgeon/resident group, ChatGPT-4 was significantly superior to Claude 2 for all considered parameters (60.9% versus 46.5%; $p < 0.001$), except for the parameter "Sources" (difference among chatbots of 0.4%; $p = 1.000$).

Inter-rater reliability, i.e., the agreement among the 19 experts, was good for most of the parameters for both chatbots, as measured as the interclass correlation coefficient (Table 3).

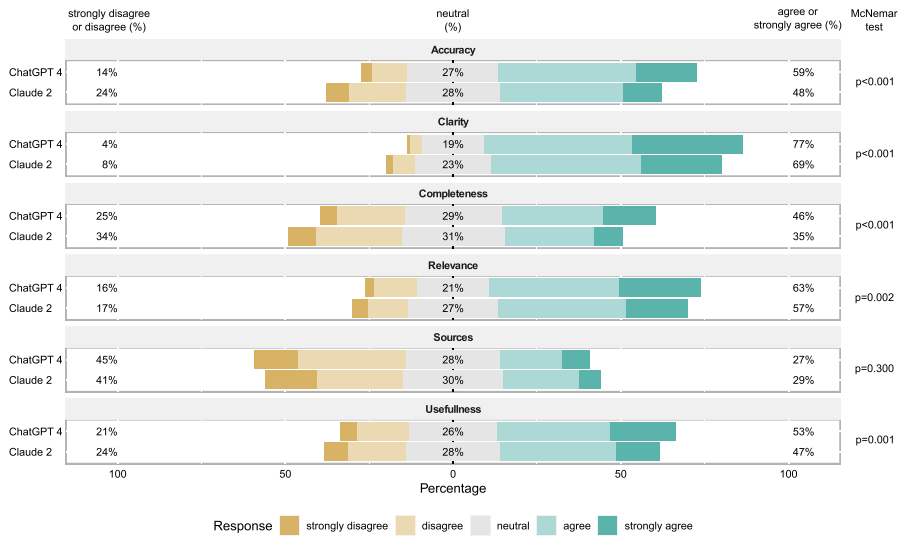

FIG. 1. Overall rating of ChatGPT-4 and Claude 2.

TABLE 2. *Percentage of raters reporting agreement/strong agreement (score 4–5) according to stakeholders and parameters*

		Parameter					
Stakeholders	Overall	Accuracy	Clarity	Relevance	Completeness	Sources	Usefulness
Patient							
ChatGPT-4	40.8%	44.7%	66.7%	51.3%	25.0%	21.1%	36.0%
Claude 2	38.2%	36.8%	60.5%	46.1%	25.4%	24.1%	36.0%
McNemar test	$p = 0.099$	$p = 0.060$	$p = 0.201$	$p = 0.201$	$p = 1.000$	$p = 0.476$	$p = 1.000$
General practitioner/non-ENT surgeon							
ChatGPT-4	60.6%	65.1%	82.7%	66.9%	55.0%	32.1%	62.0%
Claude 2	57.8%	58.0%	81.0%	65.9%	45.7%	35.6%	60.6%
McNemar test	$p = 0.055$	$p = 0.077$	$p = 0.626$	$p = 0.894$	$p = 0.018$	$p = 0.403$	$p = 0.804$
ENT resident/surgeon							
ChatGPT-4	60.9%	67.1%	81.6%	71.1%	57.5%	26.8%	61.4%
Claude 2	46.5%	49.1%	64.9%	57.5%	34.6%	27.2%	45.6%
McNemar test	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 1.000$	$p < 0.001$

Only the parameters "Clarity" for ChatGPT-4 and "Sources" for Claude 2 had moderate reliability (ICC < 0.70).

Finally, the reading level of the answers generated by ChatGPT and Claude 2 for the patient group was analyzed with the Flesch Kincaid readability calculator (Table 4). In general, ChatGPT-4 presented lower complexity than Claude 2 ($p = 0.023$), requiring an easy-fairly easy English in 75.0% of cases compared to 16.7% for Claude 2. Moreover, the average words per sentence were lower for ChatGPT-4 (mean: 9.6 words, SD = 1.8) than for Claude 2 (mean: 11.8 words, SD = 2.3; $p = 0.002$).

The heatmap correlation matrix to visualize the strength of relationships between the six parameters is included as Supplemental Digital Content 3, http://links.lww.com/MAO/C35.

DISCUSSION

This study focused on assessing the quality of responses provided by ChatGPT-4 and Claude 2 regarding the management of vestibular schwannomas. The results revealed a significant disparity in information quality, with ChatGPT-4 outperforming Claude 2 in 39% of the questions while underperforming in only 6%, and demonstrating statistically superior performance across the dimensions of "Accuracy," "Clarity," "Completeness," "Relevance," and "Usefulness". Answers generated by ChatGPT-4 showed higher levels of agreement among experts, especially when answering questions from the ENT resident/skull base surgeon perspective. Additionally, patient-oriented answers generated by ChatGPT-4 employed simpler and more easily understandable language. Despite the overall higher quality of ChatGPT-4's answers, citing appropriate sources represents a significant limitation for both chatbots.

TABLE 3. *Inter-rater reliability (intraclass coefficients)*

	Intraclass Correlation Coefficient (95% CI)	
Parameter	ChatGPT-4	Claude 2
Accuracy	0.75 (0.62–0.86)	0.87 (0.80–0.93)
Clarity	0.69 (0.53–0.82)	0.78 (0.65–0.88)
Relevance	0.76 (0.63–0.86)	0.83 (0.73–0.90)
Completeness	0.83 (0.74–0.90)	0.86 (0.79–0.92)
Sources	0.83 (0.72–0.90)	0.67 (0.49–0.81)
Utility	0.75 (0.59–0.85)	0.86 (0.77–0.92)

Even though AI is a relatively new technology, evidence of improvement of intelligent chatbots in the medical fields with newer software versions is already reported (11–13,25). In the United States, ChatGPT-3.5's performance was equivalent to that of a third-year medical student, whereas ChatGPT-4 demonstrated its reliability in clinical reasoning and medical knowledge in non-English languages by passing the Japanese Medical Licensing Examination (26,27).

In general, our data demonstrated that ChatGPT-4 showed a higher capacity in answering questions related to the management of VS than Claude 2. Given the lower rate of agreement, it could have been expected that Claude 2 would have produced more neutral responses than disagreeing ones; however, scores given from the experts showed a higher grade of disagreement for the answer generated from Claude 2. Moreover, it is interesting to note that the agreement of chatbots varied according to the complexity of the posed questions. As the complexity of the questions increased, ChatGPT-4 performed better in agreement than Claude 2. Finally, it is noteworthy that a few responses generated by AI show significant disagreement among authors, which could potentially lead to substandard patient care or even patient harm.

Given the increasing use of AI, the analysis of the quality of the answers generated by the two chatbots in the first group reflects the quality of information that patients could access through AI, which may impacts on the relationship between doctors and patients (28). Although both platforms provided answers that would likely be interpretable by the average patient, ChatGPT-4 responded with answers that were easier to comprehend while Claude 2 wrote with a higher reading grade. Generally speaking, in our experience, the Claude 2 user interface carried an advantage over ChatGPT. For instance, to obtain the list of references, a

TABLE 4. *Flesch Kincaid Calculator evaluation of question and answers by ChatGPT-4 versus Claude 2*

Reading—Grade Level	Questions	ChatGPT-4	Claude 2
Very easy	0 (0.0%)	0 (0.0%)	0 (0.0%)
Easy–fairly easy	2 (16.7%)	9 (75.0%)	2 (16.7%)
Plain English	5 (41.7%)	1 (8.3%)	7 (58.3%)
Fairly difficult or higher	5 (41.7%)	2 (16.7%)	3 (25.0%)
		Stuart-Maxwell test: $p = 0.023$	

plugin was necessary for ChatGPT-4 (which is not available for the free version, ChatGPT-3.5), whereas Claude 2 was able to provide references without any plugins. Interestingly, both platforms struggled to provide accurate references for the answers they provided. Hyperlinks within references were often nonfunctional—a finding commonly experienced in Claude 2. Similarly, references provided often included references for NF2-related schwannomatosis, which would be irrelevant to most users.

A further possible consequence of the use of AIs in everyday life could consist in influencing the questionnaires used for the evaluation of quality of life (QOL). Lately, particular attention is being paid to satisfaction/regret regarding the diagnosis and the therapeutic choice: according to this, a specific domain has been introduced in the Mayo Clinic VSQOL (29,30). Similarly, it is generally acknowledged in the literature that apprehension about the correct therapeutic choice and expectations regarding outcomes impact the quality of life of patients with vestibular schwannoma (31,32). In light of this, the easy accessibility to AI for medical information could have an impact on QOL evaluation; thus, it is important to assess appropriateness of AIs for adequate counseling (33).

This study is limited to the current state of AI, which will continue to evolve. Furthermore, this study considered questions on a rare tumor, which can have impacted the capability of both chatbots in finding appropriate scientific evidence and in formulating answers.

In conclusion, ChatGPT-4 achieved significantly higher scores in most of the analyzed items, but citing appropriate sources represented a significant limitation for both ChatGPT-4 and Claude 2. Answers given by the chatbots presented good reliability, although both software did not seem to be able to replace the role of expert clinicians in the management of VS.

REFERENCES

1. Thapa S, Adhikari S. ChatGPT, bard, and large language models for biomedical research: Opportunities and pitfalls. *Ann Biomed Eng* 2023;51:2647–51.
2. Introducing ChatGPT. Available at: https://openai.com/blog/chatgpt. Accessed August 6, 2023.
3. Anthropic\Claude 2. Available at: https://www.anthropic.com/index/claude-2. Accessed August 10, 2023.
4. Calvo IR, Jiao E, Florit-Scarvelis A, Cheung N. A Comparative Analysis of Innate Knowledge in AI and Human Language Models | OxJournal. Available at: https://www.oxjournal.org/a-comparative-analysis-of-innate-knowledge-in-ai-and-human-language-models/. Accessed March 5, 2024.
5. Bays HE, Fitch A, Cuda S, et al. Artificial intelligence and obesity management: An Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2023. *Obes Pillars* 2023;6:100065.
6. Lin YY, Guo WY, Lu CF, et al. Application of artificial intelligence to stereotactic radiosurgery for intracranial lesions: Detection, segmentation, and outcome prediction. *J Neurooncol* 2023;161:441–50.
7. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019;25:433–8.
8. Clark Lambert W, Grzybowski A. Dermatology and artificial intelligence. *Clin Dermatol* 2024;42:207–9.
9. Bizzo BC, Almeida RR, Alkasab TK. Artificial intelligence enabling radiology reporting. *Radiol Clin North Am* 2021;59:1045–52.
10. Popescu Patoni SI, Muşat AAM, Patoni C, et al. Artificial intelligence in ophthalmology. *Rom J Ophthalmol* 2023;67:207–13.
11. Perlis RH. Research Letter: Application of GPT-4 to select next-step antidepressant treatment in major depression. *medRxiv [Preprint]* 2023;2023.04.14.23288595.
12. Chiesa-Estomba CM, Lechien JR, Vaira LA, et al. Exploring the potential of chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur Arch Otorhinolaryngol* 2024;281:2081–6.
13. Xv Y, Peng C, Wei Z, Liao F, Xiao M. Can Chat-GPT a substitute for urological resident physician in diagnosing diseases?: A preliminary conclusion from an exploratory investigation. *World J Urol* 2023;41:2569–71.
14. Arslan S. Exploring the potential of Chat GPT in personalized obesity treatment. *Ann Biomed Eng* 2023;51:1887–8.
15. Dayawansa S, Mantziaris G, Sheehan J. Chat GPT versus human touch in stereotactic radiosurgery. *J Neurooncol* 2023;163:481–3.
16. Imran N, Hashmi A, Imran A. Chat-GPT: Opportunities and challenges in child mental healthcare. *Pak J Med Sci* 2023;39:1191–3.
17. Porter E, Murphy M, O'Connor C. Chat GPT in dermatology: Progressive or problematic? *J Eur Acad Dermatol Venereol* 2023;37:e943–4.
18. Ismail AMA. Chat GPT in tailoring individualized lifestyle-modification programs in metabolic syndrome: Potentials and difficulties? *Ann Biomed Eng* 2023;18:2634–5.
19. Boscolo-Rizzo P, Marcuzzo AV, Lazzarin C, et al. Quality of Information Provided by Artificial Intelligence Chatbots Surrounding the Reconstructive Surgery for Head and Neck Cancer: A Comparative Analysis Between ChatGPT4 and Claude2. *Clin Otolaryngol* 2024. doi:10.1111/coa.14261.
20. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: A comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus* 2023;15:e40822.
21. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology* 2023;307:e230171.
22. Vaira LA, Lechien JR, Abbate V, et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur Arch Oto-Rhino-Laryngol*. 2024;281:6123–6131.
23. Flesch Kincaid Calculator|Good Calculators. Available at: https://goodcalculators.com/flesch-kincaid-calculator/. Accessed February 17, 2024.
24. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
25. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare* 2023;1357633X231181922.
26. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002.
27. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
28. Fung SM, Sud C, Suchodolski M. Survey of customers requesting medical information: preferences and information needs of patients and health care professionals to support treatment decisions. *Ther Innov Regul Sci* 2020;54:75–84.
29. Carlson ML, Lohse CM, Link MJ, et al. Development and validation of a new disease-specific quality of life instrument for sporadic vestibular schwannoma: The Mayo Clinic Vestibular Schwannoma Quality of Life Index. *J Neurosurg* 2023;138:981–91.
30. Kazantzaki E, Kondylakis H, Koumakis L, et al. Psycho-emotional tools for better treatment adherence and therapeutic outcomes for cancer patients. *Stud Health Technol Inform* 2016;224:129–34.
31. Pruijn IMJ, Van Heemskerken P, Kunst HPM, Tummers M, Kievit W. Patient-preferred outcomes in patients with vestibular schwannoma: A qualitative content analysis of symptoms, side effects and their impact on health-related quality of life. *Qual Life Res* 2023;32:2887–97.
32. Hudgins WR. Patients' attitude about outcomes and the role of gamma knife radiosurgery in the treatment of vestibular Schwannomas. *Neurosurgery* 1994;34:459–65.
33. Marinelli JP, Lohse CM, Link MJ, Carlson ML. Quality of life in sporadic vestibular schwannoma. *Otolaryngol Clin North Am* 2023;56:577–86.