



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

Luigi Angelo Vaira^{1,2} · Jerome R. Lechien^{3,4} · Vincenzo Abbate⁵ · Fabiana Allevi⁶ · Giovanni Audino⁵ · Giada Anna Beltramini^{7,8} · Michela Bergonzani⁹ · Paolo Boscolo-Rizzo¹⁰ · Gianluigi Califano¹¹ · Giovanni Cammaroto¹² · Carlos M. Chiesa-Estomba¹³ · Umberto Committeri⁵ · Salvatore Crimi¹⁴ · Nicholas R. Curran¹⁵ · Francesco di Bello¹¹ · Arianna di Stadio¹⁶ · Andrea Frosolini¹⁷ · Guido Gabriele¹⁷ · Isabelle M. Gengler¹⁵ · Fabio Lonardi¹⁸ · Fabio Maglittero¹⁹ · Miguel Mayo-Yáñez²⁰ · Marzia Petrocelli²¹ · Resi Pucci²² · Alberto Maria Saibene²³ · Gianmarco Saponaro²⁴ · Alessandro Tel²⁵ · Franco Trabalzini²⁶ · Eleonora M. C. Trecca^{27,28} · Valentino Vellone²⁹ · Giovanni Salzano⁵ · Giacomo De Riu¹

Received: 3 March 2024 / Accepted: 27 April 2024 / Published online: 4 May 2024
© The Author(s) 2024

Abstract

Background The widespread diffusion of Artificial Intelligence (AI) platforms is revolutionizing how health-related information is disseminated, thereby highlighting the need for tools to evaluate the quality of such information. This study aimed to propose and validate the Quality Assessment of Medical Artificial Intelligence (QAMAI), a tool specifically designed to assess the quality of health information provided by AI platforms.

Methods The QAMAI tool has been developed by a panel of experts following guidelines for the development of new questionnaires. A total of 30 responses from ChatGPT4, addressing patient queries, theoretical questions, and clinical head and neck surgery scenarios were assessed by 27 reviewers from 25 academic centers worldwide. Construct validity, internal consistency, inter-rater and test–retest reliability were assessed to validate the tool.

Results The validation was conducted on the basis of 792 assessments for the 30 responses given by ChatGPT4. The results of the exploratory factor analysis revealed a unidimensional structure of the QAMAI with a single factor comprising all the items that explained 51.1% of the variance with factor loadings ranging from 0.449 to 0.856. Overall internal consistency was high (Cronbach's alpha = 0.837). The Interclass Correlation Coefficient was 0.983 (95% CI 0.973–0.991; $F(29,542) = 68.3$; $p < 0.001$), indicating excellent reliability. Test–retest reliability analysis revealed a moderate-to-strong correlation with a Pearson's coefficient of 0.876 (95% CI 0.859–0.891; $p < 0.001$).

Conclusions The QAMAI tool demonstrated significant reliability and validity in assessing the quality of health information provided by AI platforms. Such a tool might become particularly important/useful for physicians as patients increasingly seek medical information on AI platforms.

Keywords ChatGPT · Artificial intelligence · AI · Natural language processing · Neural networks · Machine learning · Health-related information quality · Maxillofacial surgery · Otorhinolaryngology · Head and neck surgery

Introduction

Artificial Intelligence (AI) has brought about a sea change in numerous fields, with healthcare standing out as one of the most significantly impacted [1, 2]. Among the myriad AI models available, OpenAI's (San Francisco, CA, USA) Chat-based Generative Pre-trained Transformer (ChatGPT) has been particularly striking in its reach and influence in just a few months [3, 4].

Luigi Angelo Vaira and Jerome R. Lechien have contributed equally to this work and should be considered as co-first authors.

Giovanni Salzano and Giacomo De Riu have contributed equally to this work and should be considered as co-senior authors.

Extended author information available on the last page of the article

Furthermore, ChatGPT could revolutionize healthcare delivery by not only offering health-related information [5–8] and decision-support to professionals but also by making the service more effective, efficient, and patient-friendly [9].

It is essential, however, to recognize that while these possibilities are exciting, they are still largely in the realm of potential and have yet to be fully studied or validated. Despite its potential benefits, the use of AI platforms like ChatGPT in healthcare also presents significant risks that must be thoroughly addressed [10–12]. Given these risks, rigorous, ongoing evaluation of the quality of health information provided by AI platforms is critical.

Despite the existence of several tools to assess the quality of online health information [13–15], these do not translate effectively to AI platforms. These tools have primarily been designed for manual, human-centric assessments and are not compatible with AI-generated outputs. To date, no validated tool exists to accurately assess the health information provided by ChatGPT, and the few clinical studies published on this topic have used non-validated instruments [16–22].

Recognizing these limitations, and acknowledging the critical gap that exists in the assessment of information quality from AI platforms like ChatGPT, this study aims to bridge this lacuna. We propose and validate the Quality Analysis of Medical AI (QAMAI), a novel tool designed specifically to assess the quality of health information offered by AI platforms regarding otorhinolaryngology, head and neck surgery.

Materials and methods

Working group

In February 2023, an international collaborative group was established, composed of maxillofacial surgeons, otorhinolaryngologists, and head and neck surgeons from the Italian Society of Maxillofacial Surgery and the young members' section of the International Federation of Otorhinolaryngology Societies. The group included researchers from different centers around the world, with the aim of studying the reliability and safety of using AI platforms within the field of head and neck surgery for education, diagnosis, therapy, patient communication, and information processes. For this study, 27 researchers from 25 centers across 5 countries (Italy, Belgium, France, Spain, and the United States) were involved.

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 principles.

Quality analysis of medical artificial intelligence tool development

The QAMAI tool has been developed based on the Modified DISCERN (mDISCERN) instrument [23, 24]. The mDISCERN is a well-validated and widely used tool for assessing the quality of health information conveyed by websites [25], social networks [26], YouTube and other multimedia platforms [27]. However, the use of mDISCERN for evaluating information provided by artificial intelligence is not possible as the tool takes into account certain human characteristics such as board certification and the reputation of the content creator, which cannot be applied to artificial intelligence.

The draft of QAMAI was drawn up in English by a group of experts consisting of a public health researcher, two head and neck surgeons, a computer engineer specializing in AI, a bioethics expert, a communications engineer specializing in health communication, a representative of general patient associations, and a native English-speaking linguist. The diverse backgrounds of this expert group ensured that the development of QAMAI was comprehensive, and its applicability in diverse contexts was taken into account. In analogy with the mDISCERN, the consensus of experts decided to elaborate a unidimensional construct of the instrument with 6 items, evaluated using Likert scales. The six domains of the information quality were hypothesized to be correlated, just as in mDISCERN, to one dimension: the quality of content of the information itself. Each parameter was evaluated by a Likert scale from 1 (strongly disagree) to 5 (strongly agree). The score was then summed into an overall score (QAMAI score) that identified the quality of the information. The streamlined structure of the tool was intentionally designed to ensure quick application and promote its widespread use, thereby broadening its potential impact.

The first draft of the tool was preliminarily tested by an international sample of five researchers from the working group, different from those who developed the first draft, who evaluated a set of responses provided by ChatGPT4. The results were reviewed by the consensus of experts, along with feedback provided by the researchers. Any areas of uncertainty or confusion regarding any item were addressed and corrected until the final version of the tool was developed (Table 1).

The QAMAI included six items: accuracy, clarity, relevance, completeness, sources, and usefulness. The QAMAI score, ranging from 6 to 30, allowed the classification of the response into five quality grades (Table 2).

Quality analysis of medical artificial intelligence tool validation process

A group of three researchers including two head and neck surgeons and a computer engineer specializing in AI

Table 1 The Quality Analysis of Medical Artificial Intelligence tool

	1 Strongly Disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly Agree
<i>Accuracy</i> : The information provided is accurate and up-to-date					
<i>Clarity</i> : The answer is clear and comprehensible in terms of language and scientific terminology					
<i>Relevance</i> : the information provided is relevant and directly answer to the question posed					
<i>Completeness</i> : the response adequately covers all aspects of the question and provides sufficient information including areas of uncertainty					
<i>Provision of sources and references</i> : the response provides reliable sources and references to support the health information presented					
<i>Usefulness</i> : the response provides to meet the user's health information needs					

Table 2 The Quality Analysis of Medical Artificial Intelligence tool scoring system

Score	Classification	Description
6–11 points	Poor quality	The AI system provides information that is largely unreliable or incomplete. Immediate improvement is required
12–17 points	Fair quality	The AI system provides some useful information, but there are significant areas for improvement
18–23 points	Good quality	The AI system provides mostly reliable and complete information, but there may be some areas for refinement
24–29 points	Very good quality	The AI system provides reliable and complete information in most areas. There are minor areas for improvement
30 points	Excellent quality	The AI system provides highly reliable and complete information

prepared a set of 30 questions covering various areas of head and neck surgery. Three types of questions were included: patient inquiries, theoretical questions, and clinical scenarios. The questions were reviewed by the research group and revised if they presented errors or areas of uncertainty until full consensus was reached. The questions were individually entered by a single researcher into the ChatGPT version 4 chatbot on May 25, 2023. The AI was asked to provide the most complete and exhaustive answers possible, including the bibliographic sources from which it drew its information. The exact prompt, used for all the questions, was: “please act as a head and neck surgeon and provide an answer to this question that is as exhaustive and precise as possible, taking into consideration the most recent guidelines and clear scientific evidence you have available and citing the bibliographical sources from which you drew your answers”. The responses were recorded by the researcher for subsequent analysis. The full set of questions and answers is reported in the Supplementary Table 1.

The set of answers was provided to a pool of 27 head and neck surgeons specializing in otolaryngology or maxillofacial surgery. The researchers were asked to independently evaluate the responses using the QAMAI tool, refraining from giving an evaluation if the subject matter was beyond their knowledge. The evaluation was repeated a second time,

10 days after the first. The responses obtained from the 27 researchers were collected and analyzed for the validation of the QAMAI tool.

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 [28]. 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 \pm standard deviation (SD). The QAMAI score differences among the three categories of questions were assessed using a one-way ANOVA; if significant differences were found, Dwass-Steel-Crichlow-Fligner test was employed for post-hoc analysis.

For the validation of the QAMAI, the number of questions and respondents was preliminarily determined with the aim of having at least 30 responses to evaluate for each item of the tool. The resulting sample size should be considered excellent [29]. Furthermore, Bartlett's test of sphericity and Kaiser–Meyer–Olkin (KMO) test were used to assess the sampling adequacy. A KMO value > 0.6 indicates adequate sampling [30].

For the construct validation of the questionnaire, an exploratory factor analysis was conducted to uncover the inter-relations between clusters of items and the number of factors assessed by the questionnaire. In order to maximize the loading of each variable on the extracted factors a minimum residual extraction method and a promax rotation with a cut-off point of 0.40 and the Kaiser's criterion of eigenvalues greater than 1 were used for the analysis. Upon identification of the number of factors through exploratory factor analysis, a confirmatory factor analysis was conducted to verify if the data would fit the specific theoretical model that had been identified. The assessment of the model's goodness-of-fit was carried out using the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI). Indicators of a good model fit were considered to be RMSEA values of less than 0.05 and CFI values greater than 0.95 [31].

Internal consistency was then assessed using Cronbach's alpha, to determine whether the tool's items were inter-correlated and consistent with the tool's construct. A Cronbach's alpha of at least 0.70 has been suggested to indicate adequate internal consistency [32].

Inter-rater reliability was assessed by comparing, for each question, the evaluations provided by the different reviewers using intraclass correlation coefficient (ICC). The values considered for ICC were as follows: ICC < 0.5 as poor reliability, ICC = 0.5–0.75 as moderate reliability, ICC = 0.75–0.9 as good reliability, ICC > 0.9 as excellent reliability [33].

Finally, the test–retest reliability between the two evaluations provided by researchers 10 days apart was assessed using Pearson's correlation coefficient. For all tests, the level of statistical significance was set at $p < 0.05$.

Results

Twenty-seven reviewers provided a total of 792 assessments for the 30 responses given by ChatGPT (18 assessments missing in the dataset). The median QAMAI scores reported for each question are displayed in Supplementary Table 1. Out of the 792 quality assessments collected, 30 were found to be excellent (i.e., QAMAI score 30), 261 very good (i.e., QAMAI score 24–29), 399 good (i.e., QAMAI score 18–23), 100 fair (i.e., QAMAI score 12–17), and 3 poor (i.e., QAMAI score 6–12). The differences in QAMAI scores across the three categories of questions (patient questions, theoretical questions, and clinical scenarios) were not found to be statistically significant (one-way ANOVA: $\chi^2 = 3.86$; $p = 0.145$).

To evaluate the adequacy of the sampling, the Bartlett's test of sphericity was initially performed, which returned a significant result ($\chi^2 = 2005$; $p < 0.001$), indicating the

appropriateness of conducting the KMO test. This latter test yielded an overall measure of sampling adequacy of 0.882, indicating an excellent sample size.

The results of the exploratory factor analysis revealed a unidimensional structure of the QAMAI with a single factor comprising all the items that explained 51.1% of the variance with factor loadings ranging from 0.449 to 0.856 [Table 3].

The evaluation of the goodness-of-fit confirmed a good fit between the one factor model and the data, revealing an RMSEA value of 0.053 (90% confidence interval 0.033–0.075) and a CFI value of 0.989.

The internal consistency of the questionnaire was evaluated using Cronbach's alpha, which yielded a value of 0.837 confirming an excellent internal consistency among the items in the questionnaire, suggesting that they measure the same underlying concept or construct (Fig. 1). Excellent inter-rater reliability was found comparing reviewers' QAMAI scores for each answer. The average ICC was 0.983 with a 95% confidence interval from 0.973 to 0.991 [$F(29,542) = 68.3$; $p < 0.001$].

In order to evaluate the test–retest reliability of the QAMAI tool, a Pearson's correlation test was performed comparing the scores given at two different times, spaced one week apart. The results showed a Pearson's correlation coefficient of 0.876 (95% confidence interval 0.859–0.891; $p < 0.001$) indicating a moderate-to-strong and significant correlation between the two sets of scores (Fig. 2).

Discussion

AI platforms, such as ChatGPT, are poised to revolutionize the way we access and interpret health-related information in the coming years. This innovative application of AI technology offers several potential advantages, but it also presents certain challenges that need to be addressed in the near future. The quality and accuracy of the information provided by these platforms is a significant concern [34, 35].

Table 3 Exploratory factor analysis results

	Factor	
	1	Uniqueness
Accuracy	0.743	0.448
Clarity	0.681	0.536
Relevance	0.772	0.403
Completeness	0.719	0.484
Sources	0.449	0.798
Usefulness	0.856	0.267

'Minimum residual' extraction method was used in combination with a 'promax' rotation

Correlation Heatmap

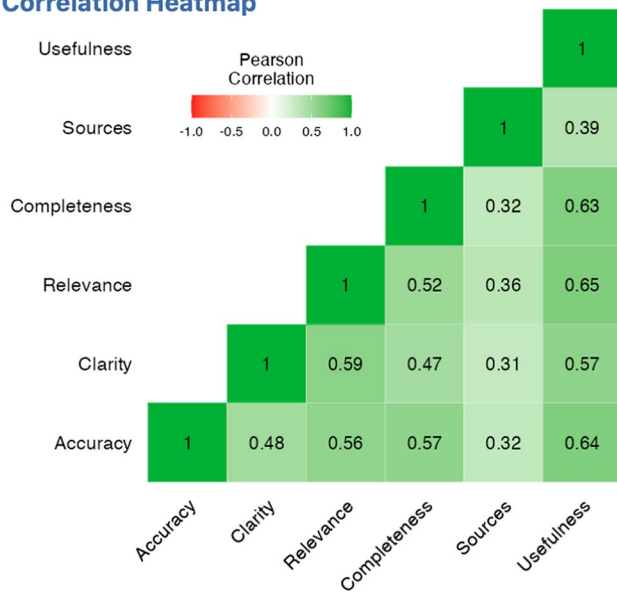


Fig. 1 Heatmap of the correlations between QAMAI items

AI platforms rely heavily on the quality and accuracy of the data they have been trained on. Misinformation or outdated data could lead to the provision of incorrect advice, which could have severe consequences in healthcare.

Given these circumstances, it becomes essential to develop and adopt tools to evaluate the quality of health information delivered by AI platforms. These evaluation tools are crucial for several reasons. They first allow for the systematic assessment of AI-generated information's quality and accuracy, providing an additional layer of scrutiny. Secondly, they can identify areas of weakness or inaccuracy within the AI platform, guiding its further development and improvement. Furthermore, these tools can aid in promoting transparency and trust in AI systems. They can ensure that the information provided meets certain standards, assuaging user concerns about the reliability of AI-generated advice. Finally, they contribute to the broader effort of promoting digital health literacy, empowering users to critically evaluate and make informed decisions about the health information they encounter online.

The QAMAI tool represents the first attempt to provide a validated, widely usable instrument for assessing health information delivered by AI platforms. The design of the QAMAI tool is lean and streamlined, specifically aimed at enabling a swift but comprehensive assessment of AI responses. This agile/specific design is crucial for several reasons. First, it ensures that the evaluation process does not become overly cumbersome or time-consuming, thereby facilitating the tool's widespread application. Secondly, while being expedient, the QAMAI tool does not compromise on thoroughness. It is designed to deliver a

comprehensive evaluation of the AI responses, looking at a range of aspects to determine the quality of the health information provided.

The exploratory factor analysis and internal consistency assessment have revealed that the QAMAI is a unidimensional tool. This unidimensional factor can be identified with the quality of information, and all instrument items are inherently tied to this core construct. On the internal consistency analysis, the item presenting the weakest correlations with others was the quality of sources (Fig. 1). This is likely tied to the propensity of ChatGPT to often provide correct or near correct answers in content, but with non-existent bibliographic references. Literature reports a rate of erroneous sources exceeding 80% for version 3.5 [36–38]. In this data series, 30.6% of the sources were non-existent, indicating an improvement in version 4 of ChatGPT. However, this rate remains unacceptably high, underscoring the need for further improvements in the AI system's ability to provide accurate and reliable source references.

Inter-rater reliability was found to be excellent, showcasing the consistency of QAMAI scores across different reviewers. This strongly supports the tool's reliability in different hands and further validates its robustness. Furthermore, test–retest reliability demonstrated a strong and statistically significant correlation between two sets of scores spaced one week apart. This indicates the stability of the QAMAI scores over time, affirming the reliability of the tool in providing consistent evaluations across different time intervals.

This study presents several limitations that should be acknowledged. The dataset of questions used in this research was limited to those pertaining solely to head and neck surgery, which may limit the generalizability and applicability of QAMAI across other medical branches. Secondly, the questionnaire was designed to evaluate solely the quality of the information provided, and not the information transmission modalities, an element that is often crucial in human communication. In patient interactions, empathy and understanding of patients' needs are nearly as important as the quality of the information itself. In the case of AI-generated health information, these human elements may not be fully captured or conveyed. Therefore, further refinement and expansion of tools like QAMAI are needed to account for these factors when assessing the quality of health information directed at patients. Finally, although all 27 reviewers possess a high level of English proficiency, they are not native speakers. This could have introduced linguistic bias, as the assessment relied solely on the English version of the QAMAI. Such bias might affect the interpretation and understanding of nuanced language-specific elements. Future studies will need to validate the QAMAI versions in other languages to comprehensively address this limitation.

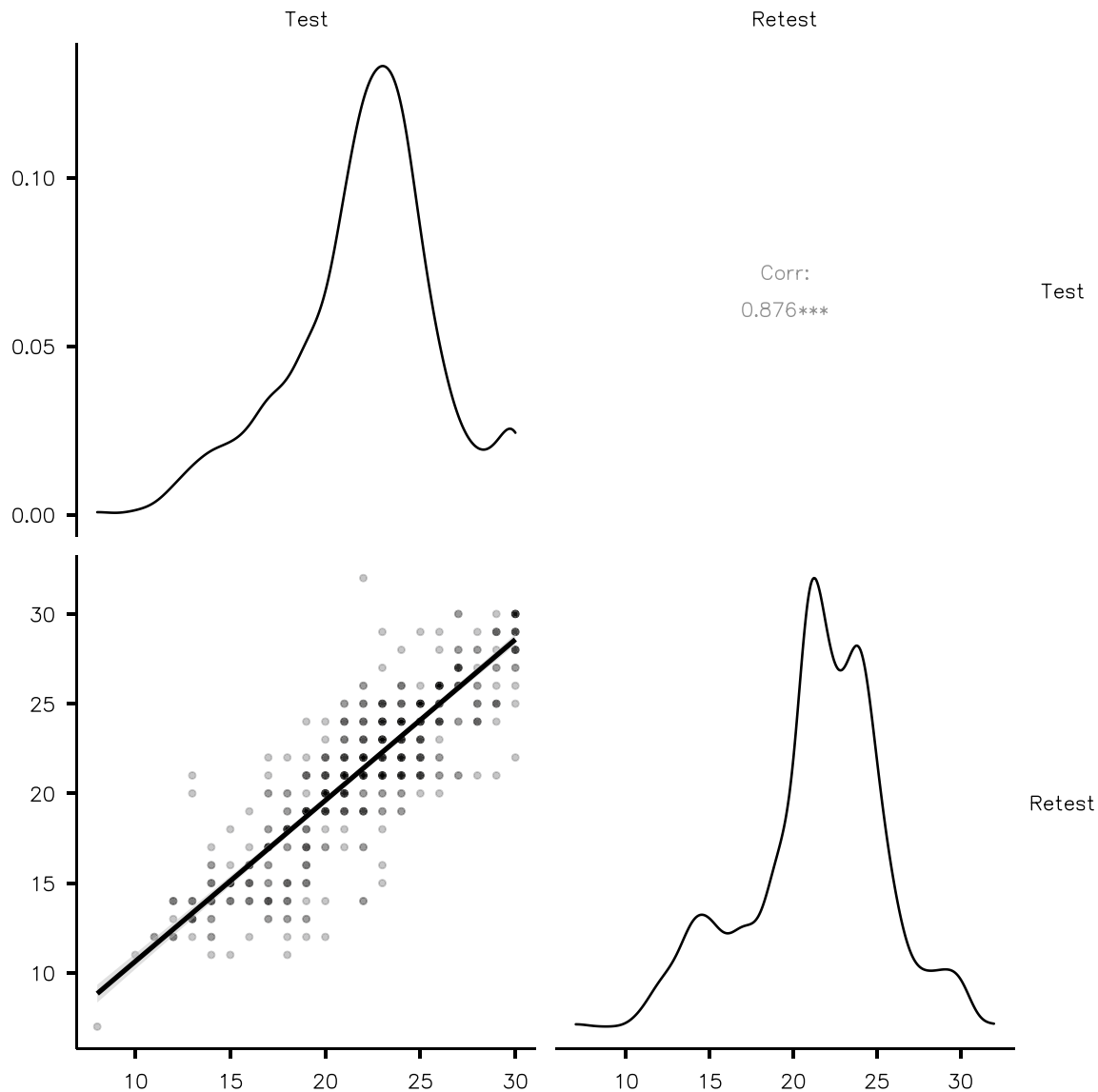


Fig. 2 Test–retest reliability analysis

Conclusions

The findings of this study demonstrate that the QAMAI tool represents a valid and reliable instrument to evaluate the quality of health information delivered by AI platforms, such as ChatGPT. The implementation and large-scale utilization of such tools are critical for monitoring the quality of this rapidly expanding source of health information, currently largely unverified. Ensuring the accuracy and reliability of AI-generated health information is of utmost importance in preventing potential harm to users independently seeking health advice or information. As AI continues to revolutionize the health information landscape, the need for robust quality control tools like QAMAI will only increase. Therefore, our findings pave the way for future research and

validation efforts, ultimately contributing to safer and more effective use of AI in health information delivery.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00405-024-08710-0>.

Authors 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 and final draft, final approval. Fabiana Allevi: data collection, data curation, revision of the original and final draft, final approval. Giovanni Audino: data collection, data curation, revision of the original and final draft, final approval. Giada Anna Beltramini: data collection, data curation, revision of the original and final draft, final approval. Michela Bergonzani: data collection, data curation, revision of the original and final draft, final approval. Paolo Boscolo-Rizzo:

data collection, data curation, revision of the original and final draft, final approval. Gianluigi Califano: data curation and analysis, revision of the original and final draft, final approval. Giovanni Cammaroto: data collection, data curation, revision of the original and final draft, final approval. Carlos Miguel Chiesa-Estomba: provision of study instrumentation, development of the methodology, review of the first and final draft, final approval. Umberto Committeri: data collection, data curation, revision of the original and final draft, final approval. Salvatore Crimi: data collection, data curation, revision of the original and final draft, final approval. Nicholas R. Curran: data collection, data curation, revision of the original and final draft, final approval. Francesco di Bello: data curation and analysis, literature review, revision of the original and final draft, final approval. Arianna di Stadio: data collection, data curation, revision of the original and final draft, final approval. Andrea Frosolini: data collection, data curation, revision of the original and final draft, final approval. Guido Gabriele: data collection, data curation, revision of the original and final draft, final approval. Isabelle M. Gengler: data collection, data curation, revision of the original and final draft, final approval. Fabio Lonardi: data collection, data curation, revision of the original and final draft, final approval. Fabio Maglito: data collection, data curation, revision of the original and final draft, final approval. Miguel Mayo-Yáñez: data collection, data curation, revision of the original and final draft, final approval. Marzia Petrocelli: data collection, data curation, revision of the original and final draft, final approval. Resi Pucci: data collection, data curation, revision of the original and final draft, final approval. Alberto Maria Saibene: data collection, data curation, revision of the original and final draft, final approval. Gianmarco Saponaro: data collection, data curation, revision of the original and final draft, final approval. Alessandro Tel: data collection, data curation, revision of the original and final draft, final approval. Franco Trabalzini: data collection, data curation, revision of the original and final draft, final approval. Eleonora M.C. Trecca: data collection, data curation, revision of the original and final draft, final approval. Valentino Vellone: data collection, data curation, revision of the original and 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, final approval.

Funding Open access funding provided by Università degli Studi di Sassari within the CRUI-CARE Agreement.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest None of the authors has a financial interest in any of the products, devices or drugs mentioned in this manuscript.

Ethical approval The author Jerome R. Lechien is also guest editor of the special issue on ‘ChatGPT and Artificial Intelligence in Otolaryngology-Head and Neck Surgery’. He was not involved with the peer review process of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in

the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

1. Aung YYM, Wong DCS, Ting DSW (2021) The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull* 139:4–15
2. Yu KH, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. *Nat Biomed Eng* 2:719–731
3. ChatGPT. Available online: <https://openai.com/blog/chatgpt>. Accessed on 19th June 2023
4. Number of ChatGPT Users (2023). Available online: <https://explodingtopics.com/blog/chatgpt-users>. Accessed on 30th June 2023
5. Barat M, Soyer P, Dohan A (2023) Appropriateness of recommendations provided by ChatGPT to interventional radiologists. *Can Assoc Radiol J* 74:758–763
6. Cheng K, Sun Z, He Y et al (2023) The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? *Int J Surg* 109:1545–1547
7. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR (2024) ChatGPT-4 performance in rhinology: a clinical case-series. *Int Forum Allergy Rhinol*. <https://doi.org/10.1002/alr.23323>
8. Lechien JR, Gorton A, Robertson J, Vaira LA (2023) Is ChatGPT-4 accurate in proofread a manuscript in otolaryngology-head and neck surgery? *Otolaryngol Head Neck Surg*. <https://doi.org/10.1002/ohn.526>
9. Hopkins AM, Logan JM, Kichenadasse G et al (2023) Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr* 7:pkad010
10. Dave T, Athaluri SA, Singh S (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 6:1169595
11. Rao A, Pang M, Kim J et al (2023) Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res* 25:e48659
12. Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 11:887
13. Lee J, Lee EH, Chae D (2021) eHealth literacy instruments: systematic review of measurement properties. *J Med Internet Res* 23:e30644
14. Bernstam EV, Shelton DM, Walji M et al (2005) Instruments to assess the quality of health information on the world wide web: what can our patients actually use? *Int J Med Inform* 74:13–19
15. Drozd B, Couvillon E, Suarez A (2018) Medical YouTube videos and methods of evaluation: literature review. *JMIR Med Educ* 4:e3
16. Vaira LA, Lechien JR, Abbate V et al (2023) Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg*. <https://doi.org/10.1002/ohn.489>
17. Deiana G, Dettori M, Arghittu A et al (2023) Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines* 11:1217
18. Chiesa-Estomba CM, Lechien JR, Vaira LA et al (2023) Exploring the potential of Chat-GPT as a supportive tool for sialoendoscopy and clinical decision making and patient information support. *Eur Arch Otolaryngol* 281:2081–2086

19. Johnson D, Goodman R, Patrinely J et al (2023) Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq*. <https://doi.org/10.21203/rs.3.rs-2566942/v1>
20. Lechien JR, Neunheim MR, Maniaci A et al (2024) Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case series. *Otolaryngol Head Neck Surg*. <https://doi.org/10.1002/ohn.759>
21. Mayo Yanez M, Gonzalez-Torres L, Saibene AM et al (2024) Application of ChatGPT as a support tool in the diagnosis and management of acute bacterial tonsillitis. *Health Technol*. <https://doi.org/10.1007/s12553-024-00858-3>
22. Saibene AM, Allevi F, Calvo-Henriquez C et al (2024) Reliability of large language models in managing odontogenic sinusitis clinical scenarios: a preliminary multidisciplinary evaluation. *Eur Arch Otorhinolaryngol* 281:1835–1841
23. Charnock D, Shepperd S, Needham G et al (1999) DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 53:105–111
24. Khazaal Y, Chatton A, Cochand S et al (2009) Brief DISCERN, six questions for the evaluation of evidence-based content of health-related websites. *Patient Educ Couns* 77:33–37
25. Olkun HK, Olkun RS (2021) Evaluation of the quality of information on the internet about 2019 coronavirus outbreak in relation to orthodontics. *Health Technol (Berl)* 11:437–441
26. Terrens AF, Soh SE, Morgan P (2022) What web-based information is available for people with Parkinson's disease interested in aquatic physiotherapy? A social listening study. *BMC Neurol* 22:170
27. Vaira LA, Sergnese S, Salzano G et al (2023) Are YouTube videos a useful and reliable source of information for patients with temporomandibular joint disorders? *J Clin Med* 12:817
28. The jamovi project (2022). Jamovi. (version 2.3) [Computer Software]. Retrieved from <https://www.jamovi.org>
29. Tsang S, Royse CF, Terkawi AS (2017) Guidelines for developing, translating, and validating a questionnaire in preoperative and pain medicine. *Saudi J Anaesth* 11:S80–S89
30. Dziuban CD, Shirkey EC (1974) When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychol Bull* 81:358–361
31. Wolf MG, McNeish D (2023) Dynamic: an R package for deriving dynamic fit index cutoffs for factor analysis. *Multivariate Behav Res* 58:189–194
32. Streiner DL (2003) Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess* 80:99–103
33. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15:155–163
34. Minssen T, Vayena E, Cohen IG (2023) The challenges for regulating medical use of ChatGPT and other large language models. *JAMA* 330:315–316
35. Marks M, Haupt CE (2023) AI chatbots, health privacy, and challenges to HIPAA compliance. *JAMA* 330:309–310
36. Frosolini A, Franz L, Benedetti S et al (2023) Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol* 280:5129–5133
37. Wagner MW, Ertl-Wagner BB (2024) Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol J* 75:69–73
38. Lechien JR, Briganti G, Vaira LA (2024) Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. *Eur Arch Otorhinolaryngol* 281:2159–2165

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Luigi Angelo Vaira^{1,2}  · Jerome R. Lechien^{3,4} · Vincenzo Abbate⁵ · Fabiana Allevi⁶ · Giovanni Audino⁵ · Giada Anna Beltramini^{7,8} · Michela Bergonzani⁹ · Paolo Boscolo-Rizzo¹⁰ · Gianluigi Califano¹¹ · Giovanni Cammaroto¹² · Carlos M. Chiesa-Estomba¹³ · Umberto Committeri⁵ · Salvatore Crimi¹⁴ · Nicholas R. Curran¹⁵ · Francesco di Bello¹¹ · Arianna di Stadio¹⁶ · Andrea Frosolini¹⁷ · Guido Gabriele¹⁷ · Isabelle M. Gengler¹⁵ · Fabio Lonardi¹⁸ · Fabio Maglito¹⁹ · Miguel Mayo-Yáñez²⁰ · Marzia Petrocelli²¹ · Resi Pucci²² · Alberto Maria Saibene²³ · Gianmarco Saponaro²⁴ · Alessandro Tel²⁵ · Franco Trabalzini²⁶ · Eleonora M. C. Trecca^{27,28} · Valentino Vellone²⁹ · Giovanni Salzano⁵ · Giacomo De Riu¹

✉ Luigi Angelo Vaira
lavaira@uniss.it

¹ Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Viale San Pietro 43/B, 07100 Sassari, Italy

² PhD School of Biomedical Science, Biomedical Sciences Department, University of Sassari, Sassari, Italy

³ Department of Laryngology and Bronchoesophagology, EpiCURA Hospital, 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, USA
- 10 Department of Medical, Surgical and Health Sciences, Section of Otolaryngology, University of Trieste, Trieste, Italy
- 11 Department of Neurosciences, Reproductive and Odontostomatological Science, Federico II University of Naples, Naples, Italy
- 12 ENT Department, Morgagni Pierantoni Hospital, AUSL Romagna, Forlì, Italy
- 13 Department of Otorhinolaryngology-Head and Neck Surgery, Hospital Universitario Donostia, San Sebastian, Spain
- 14 Operative Unit of Maxillofacial Surgery, Policlinico San Marco, University of Catania, Catania, Italy
- 15 Department of Otolaryngology-Head and Neck Surgery, University of Cincinnati Medical Center, Cincinnati, OH, USA
- 16 Otolaryngology Unit, GF Ingrassia Department, University of Catania, Catania, Italy
- 17 Department of Maxillofacial Surgery, University of Siena, Siena, Italy
- 18 Department of Maxillofacial Surgery, University of Verona, Verona, Italy
- 19 Maxillo-Facial Surgery Unit, University of Bari “Aldo Moro”, Bari, Italy
- 20 Otorhinolaryngology, Head and Neck Surgery Department, Complejo Hospitalario Universitario A Coruña (CHUAC), A Coruña, Galicia, Spain
- 21 Maxillofacial Surgery Operative Unit, Bellaria and Maggiore Hospital, Bologna, Italy
- 22 Maxillofacial Surgery Unit, San Camillo-Forlanini Hospital, Rome, Italy
- 23 Otolaryngology Unit, Santi Paolo e Carlo Hospital, Department of Health Sciences, University of Milan, Milan, Italy
- 24 Maxillo-Facial Surgery Unit, IRCSS “A. Gemelli” Foundation–Catholic University of the Sacred Heart, Rome, Italy
- 25 Clinic of Maxillofacial Surgery, Department of Head and Neck Surgery and Neuroscience, University Hospital of Udine, Udine, Italy
- 26 Department of Otorhinolaryngology, Head and Neck Surgery, Meyer Children’s Hospital, Florence, Italy
- 27 Department of Otorhinolaryngology and Maxillofacial Surgery, IRCCS Hospital Casa Sollievo Della Sofferenza, San Giovanni Rotondo, Foggia, Italy
- 28 Department of Otorhinolaryngology, University Hospital of Foggia, Foggia, Italy
- 29 Maxillofacial Surgery Unit, “S. Maria” Hospital, Terni, Italy