Computational and Structural Biotechnology Journal 24 (2024) 46–52

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery

Ana Suárez [a], Jaime Jiménez [b], María Llorente de Pedro [a], Cristina Andreu-Vázquez [c], Víctor Díaz-Flores García [a,*], Margarita Gómez Sánchez [a], Yolanda Freire [a]

[a] Department of Pre-Clinic Dentistry, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, 28670 Madrid, Spain
[b] Department of Clinic Dentistry, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, 28670 Madrid, Spain
[c] Department of Veterinary Medicine, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, 28670 Madrid, Spain

ARTICLE INFO

Keywords:
Oral surgery
Artificial Intelligence
ChatGPT
Open AI
Chatbot
Dentistry
Large language models
Natural generative language

ABSTRACT

AI has revolutionized the way we interact with technology. Noteworthy advances in AI algorithms and large language models (LLM) have led to the development of natural generative language (NGL) systems such as ChatGPT. Although these LLM can simulate human conversations and generate content in real time, they face challenges related to the topicality and accuracy of the information they generate. This study aimed to assess whether ChatGPT-4 could provide accurate and reliable answers to general dentists in the field of oral surgery, and thus explore its potential as an intelligent virtual assistant in clinical decision making in oral surgery.

Thirty questions related to oral surgery were posed to ChatGPT4, each question repeated 30 times. Subsequently, a total of 900 responses were obtained. Two surgeons graded the answers according to the guidelines of the Spanish Society of Oral Surgery, using a three-point Likert scale (correct, partially correct/incomplete, and incorrect). Disagreements were arbitrated by an experienced oral surgeon, who provided the final grade Accuracy was found to be 71.7%, and consistency of the experts' grading across iterations, ranged from moderate to almost perfect.

ChatGPT-4, with its potential capabilities, will inevitably be integrated into dental disciplines, including oral surgery. In the future, it could be considered as an auxiliary intelligent virtual assistant, though it would never replace oral surgery experts. Proper training and verified information by experts will remain vital to the implementation of the technology. More comprehensive research is needed to ensure the safe and successful application of AI in oral surgery.

1. Introduction

Artificial intelligence (AI) has revolutionized many fields by enabling computer systems to perform tasks that traditionally required human intervention [1]. Since the introduction of conversational agents with ELIZA in 1966 [2], they have had a remarkable impact on how humans interact with machines and seek answers [3]. Significant advances in AI algorithms and models have led to the development of natural generative language (NGL) systems, including Chatbot Generative Pre-trained Transformer (ChatGPT), launched by OpenAI (OpenAI, San Francisco, CA, USA) in late 2022 [4].

These systems, based on large language models (LLMs), have the ability to emulate human conversations and generate original content from the training data to which they have been exposed [5], providing real-time responses and covering a wide range of applications, from creative activities to solving complex problems [3,4,6–8].

However, the use of LLMs presents challenges related to the nature of their training data and the inherent limitations of each model. For example, ChatGPT's data collection only goes until 2021, restricting its capacity to generate current information [9]. Nevertheless, there is a beta version of ChatGPT-4, which works with Bing (Microsoft, Redmond, Washington, USA) conducting searches without time constraints,

* Correspondence to: Department of Pre-Clinic Dentistry, School of Biomedical Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, 28670 Madrid, Spain.
E-mail address: victor.diaz-flores@universidadeuropea.es (V. Díaz-Flores García).

https://doi.org/10.1016/j.csbj.2023.11.058
Received 3 November 2023; Received in revised form 28 November 2023; Accepted 28 November 2023
Available online 6 December 2023

but it is still under development [10]. It has also been shown that NGL models such as ChatGPT can produce convincing but completely incorrect answers, a phenomenon known as "hallucination" [11,12]. This raises questions about the reliability and accuracy of generated answers and highlights the relevance of using appropriate prompts to obtain more accurate responses [13].

In medicine, although ChatGPT is not specifically trained to answer questions related to this field, the model has been shown to outperform official exams such as the United States Medical Licensing Exam (USMLE) [14,15] or the examination for Internal Medical Residents in Spain [16]. Studies in specific areas have evaluated the performance of ChatGPT in fields such as orthopedics, urology, hepatic pathology, plastic surgery, and others, reporting mixed results in terms of accuracy and reliability [17–39].

In this context, oral surgery as a dental discipline faces many challenges due to the diversity of pathologies and treatment options. In countries such as Spain, Austria, and Luxembourg, there is no legal requirement for specialties in dentistry. Therefore, dental specialists are not officially recognized, despite being professionals trained in universities through postgraduate courses in oral surgery [40]. Thus, a general dentist will have the necessary basic skills to legally perform any dental treatment, including oral surgery [41]. This fact raises the issue of whether LLMs, such as ChatGPT, can provide accurate and reliable answers to specific questions in the field, and thereby serve as a valuable, virtual intelligent assistant for general dentists in clinical practice with respect to the decision-making process in oral surgery.

With this in mind, the goal of this study was to assess the answers provided by ChatGPT to questions regarding oral surgery. Through the evaluation of the ChatGPT answers by experts in oral surgery, we aim to highlight both the potential benefits and limitations of this emerging technology.

2. Material and methods

2.1. Ethical approval

The study adhered to the tenets of the Declaration of Helsinki. Ethical approval was not required as no human subjects were involved in the study.

2.2. Question design

For the design of the questions, the documents for oral surgery practice of the Spanish Society of Oral Surgery (Sociedad Española de Cirugía Bucal [SECIB]) [42] were used. This compendium of documents was chosen because it represents clinical practice guidelines in oral surgery developed by a multidisciplinary group of experts in teaching, research and clinical practice in the field. 82 questions were obtained from all available documents, from which, 30 questions were randomly selected (https://www.random.org, accessed: 09.09.2023). (Table 1).

Table 1
Rubric used to score the answers.

Experts' grading	Description
Incorrect (0)	The answer provided is completely incorrect or unrelated to the question. It does not show an adequate understanding or knowledge of the topic.
partially correct or Incomplete (1)	The answer shows some understanding or knowledge of the topic, but there are significant errors or missing elements. Although not entirely incorrect, the answer is not sufficiently accurate or complete to be considered confident or appropriate.
Correct (2)	The answer is completely correct and shows a sound and accurate understanding of the topic. All key elements are addressed accurately and comprehensively.

2.3. Generation of answers in ChatGPT-4

The questions were introduced into ChatGPT-4 using two different accounts (Y.F. and V.D-FG). Each question was introduced individually using the "new chat" option.

Since ChatGPT's probabilistic algorithm can generate different answers to the same question, [34], 30 answers were obtained for each question. The entire data collection process lasted 7 days (11–17 September 2023).

Due to the specific and technical nature of oral surgery, it was essential to adapt and optimize ChatGPT to respond effectively. A specific prompt was designed to simulate an interaction between an oral surgery specialist and a general dentist. It was also necessary to make the answer more focused and deterministic: *"Imagine that you are an oral surgeon and I am a general dentist. Please answer the following question accurately and directly, without rambling or creative answers: [QUESTION] "*. Finally, to avoid any memory bias, the 'new chat' mode was reset for each repetition. The answers were stored in an Excel© spreadsheet (Microsoft, Redmond, Washington, USA). (Fig. 1), (Supplementary material 1).

2.4. Evaluation of ChatGPT-4 answers by human experts

Two postgraduate dentists specialized in oral surgery (A.S. and M. LLdP) evaluated the 900 answers generated by ChatGPT-4 using a 3-point Likert scale, (Table 1), and with access the documents for oral surgery practice of the Spanish Society of Oral Surgery from the SECIB [42]. Table 1. Any discrepancies in the grading of answers by the two raters assessed independently by a third senior experienced oral surgery specialist (J.J.). (Supplementary material 1).

2.5. Statistical analysis

The 900 ratings for each answer were stored in an Excel© spreadsheet (Microsoft, Redmond, Washington, USA) and analyzed using STATA© (StataCorp, College Station, Texas, USA) statistical software, version BE 14.

For each of the 30 questions, the absolute (n) and relative frequency (%) of grade 0 (incorrect), 1 (incomplete or partially correct), and 2 (correct) assigned by the expert were described. In order to analyze the accuracy of the answers generated by ChatGPT, the proportion of questions yielding an answer with a grade of 2 (correct) was calculated for the total answers in the question set and for each individual question, along with its 95% confidence interval (Wald binomial method).

To assess repeatability, the consistency of grades across repetitions was analyzed using concordance analyses weighted for ordinal categories and multiple repetitions (including percentage agreement, Brennan and Prediger's coefficient, Conger's generalized Cohen's kappa, Fleiss' kappa, Gwet's AC and Krippendorff's alpha) along with their corresponding 95% confidence intervals.

2.6. Data availability

The data that support the findings of this study are available on reasonable request from the corresponding author, [V.D-FG].

3. Results

Table 2 shows the absolute and relative frequencies of the experts' grading for each answer. As can be seen, the proportion of correct answers varied from 0% to 100% for certain questions. The overall accuracy was 71.7%, with a 95% confidence interval of 68.9–74.6%.

Table 3 shows the repeatability of the experts' grading of the answers (consistency in the grading of the answers).

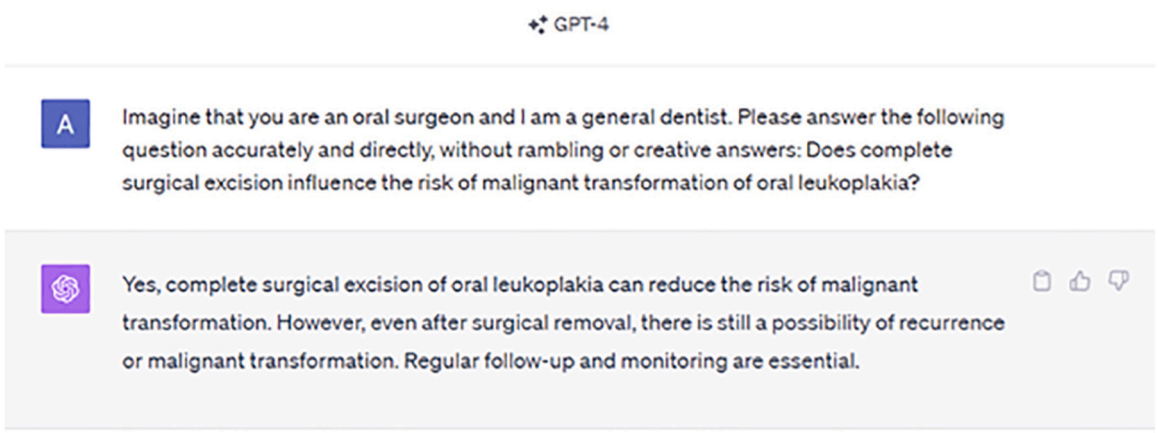

Fig. 1. Screenshot of a question and answer in ChatGPT-4.

4. Discussion

Disruptive technologies represent a new horizon in healthcare. In this context, AI-based chatbots such as ChatGPT could improve access to patient care services [43,44]. However, any emerging technology raises concerns about the potential risks inherent in its implementation. The reliance on non-expert training data, the risk of working with outdated information, as well as ethical and legal concerns related to patient privacy, require careful consideration [45–49]. Given recent advances in machine learning and the ability of these systems to process and understand vast amounts of information, we are witnessing a paradigm shift in healthcare [50].

Several studies [11,33,38,51–53] have examined ChatGPT's ability to correctly answer multiple-choice questions, with accuracy rates ranging from 59.4% to 81%. It is plausible that these differences are due to the different versions of the model used. When comparing ChatGPT-4 with previous versions, it's clear that its performance exceeds that of ChatGPT-3.5 [38,39,51,54–57]. Therefore, ChatGPT-4 was used in the present study. In addition, we chose to ask questions in an open-ended format, considering that in a real-life consultation with a specialist, questions would be presented directly, without pre-determined answer options.

It has been shown that ChatGPT can give different answers to the same question due to its unpredictable nature [58]. Some studies have assessed the consistency of ChatGPT answers by repeating the same question multiple times, with iterations ranging from 2 to 10 times [11,20,24,25,35,37]. In this study, to minimize and assess potential bias from repetition, 30 different answers were collected for each question in individualized sessions using the 'new chat' option. This ensured that answers were not influenced by previous dialogues.

However, there is a significant problem known in the literature [11,12,54,59,60] as "hallucination" or "stochastic parroting", which describes the generation of convincing but false information by artificial intelligence systems such as ChatGPT. This includes the possibility of generating fictitious bibliographic references or incorrect data, which raises ethical and legal dilemmas that can affect the integrity of clinical practice and patient safety [61–63]. To address this, our study implemented a prompt strategy that specified the roles to be played by both ChatGPT and the human interlocutor, as well as the desired type of answer: direct and free of rambling or creativity. No references to this prompt approach were found in the literature reviewed. In assessing the repeatability of experts' grading of ChatGPT responses, our study found that it ranged from moderate to almost perfect, a finding in line with other studies reporting acceptable or high levels of consistency [24,25,35].

Assessing the accuracy of answers in a medical context requires clear and defined criteria, given the direct impact on patient health and safety. In the literature published to date, there is considerable heterogeneity in the Likert scales used to assess ChatGPT answers [18–20,26,32,36,64], which may complicate the interpretation of results. Broad Likert scales can introduce a degree of ambiguity that can be critical in healthcare settings. For example, a ChatGPT answer rated as 5 on a scale of 0–10 may be considered acceptable by some standards. However, such a rating does not provide sufficient clarity about possible omissions or errors in content, which could have serious implications for clinical decision making. In this study, as in previous studies [27,30,65], we chose a 3-point Likert scale with the aim of providing a more precise and direct assessment of answers, reducing ambiguity and improving patient safety. It is important to note that although there are calibrated scales in the literature [66] specifically designed to evaluate the performance of intelligent chatbots in clinical contexts, these are intended for complex cases that require a comprehensive evaluation of history, symptoms, physical examination, additional tests, diagnosis and treatment plan. In contrast, our study focused on more specific and less complex questions, aimed at situations where a healthcare professional consults on specific doubts, away from the dynamics of a comprehensive clinical case.

To contextualize our findings, we considered previous studies by Sütcüoğlu and Güler [30], and the study by Ali [65], that evaluated the performance of ChatGPT in different knowledge domains and reported different rates of correct, partially correct and incorrect answers. The results of Sütcüoğlu and Güler [30] showed a high percentage of correct answers (76%) in their analysis of 25 answers related to premature ovarian insufficiency, which is remarkably similar to the results of our own study (71.7%). On the other hand, Ali [65] achieved a lower performance in the area of lacrimal drainage disorders, with 40% correct answers out of 21 questions. It is important to emphasize that our study was carried out using the latest version of the model, ChatGPT-4, and was based on the collection of a large and repetitive dataset, with a total of 900 answers. This methodology allowed us to assess the consistency of the model on different occasions. The variability observed in the percentages of correct, partially correct and incorrect answers between studies can be attributed to factors such as the specific nature of the questions asked, the domain of knowledge associated with the training data and the particular version of the model used in each case.

In our study, we observed that certain questions (Q03, Q06, Q07 and Q08) achieved a 100% correct answer rate. A possible explanation for this perfect rate could be the combination of clear and direct questions related to basic and well-established medical knowledge, as in the case of leukoplakia. In contrast, there are other questions (Q02, Q10, Q11, Q12, Q13 and Q14) where more incorrect or partially correct/incomplete answers were recorded. However, these questions do not seem to have a common structure and include different topics such as leukoplakia and periapical surgery. Despite this, we found no study in the literature that analyzed the reasons for these potential discrepancies

Table 2
Distribution of experts' grading for ChatGPT answers.

Question		Incorrect		Partially correct or Incomplete		Correct	
		n	%	n	%	n	%
Q01	What types of oral lesions are currently classified as leukoplakia?	1	3.33	9	30.00	20	66.67
Q02	Is a biopsy necessary to make a definitive diagnosis of oral leukoplakia?	0	0.00	23	76.67	7	23.33
Q03	Are age and/or gender significant factors in the malignant transformation of oral leukoplakia?	0	0.00	0	0.00	30	100.00
Q04	Is there a higher risk of developing oral cancer in patients with multiple leukoplakia?	1	3.33	5	16.67	24	80.00
Q05	Is the degree of epithelial dysplasia in oral leukoplakia a predictive factor for its malignant transformation?	0	0.00	1	3.33	29	96.67
Q06	Does complete surgical excision influence the risk of malignant transformation of oral leukoplakia?	0	0.00	0	0.00	30	100.00
Q07	Is there any medical treatment for oral leukoplakia that reduces the risk of malignant transformation?	0	0.00	0	0.00	30	100.00
Q08	What should be the clinical follow-up of a patient diagnosed and treated for oral leukoplakia?	0	0.00	0	0.00	30	100.00
Q09	In a patient with a tooth that has undergone failed orthograde endodontic treatment and has persistent chronic apical periodontitis, is a higher cure rate of the periapical inflammatory process achieved by periapical surgery or by orthograde retreatment?	0	0.00	2	6.67	28	93.33
Q10	In periapical surgery, does bevelling result in greater apical leakage (or lower cure rate) than not bevelling?	27	90.00	0	0.00	3	10.00
Q11	In a patient with a tooth treated by periapical surgery, is there less leakage (or a higher cure rate) when the retrograde cavity is prepared with diamond-tipped ultrasonic tips than when other techniques are used?	17	56.67	3	10.00	10	33.33
Q12	In a patient with a tooth that has undergone periapical surgery to prepare and fill the retrograde cavity, does the use of amplification and magnification devices reduce leakage (or increase the cure rate) compared to not using them?	0	0.00	30	100.00	0	0.00
Q13	Does the use of hemostatic materials significantly reduce bleeding in patients undergoing periapical surgery?	0	0.00	23	76.67	7	23.33
Q14	Does the use of bone graft substitutes and/or membranes improve the healing rate in patients undergoing periapical surgery?	0	0.00	28	93.33	2	6.67
Q15	In relation to oral cancer, in the preoperative period, when is the best time for dental treatment?	0	0.00	3	10.00	27	90.00
Q16	What actions in the preoperative period can improve quality of life in adult oral cancer patients?	1	3.33	6	20.00	23	76.67
Q17	What actions in the pre-treatment period can reduce the occurrence of osteoradionecrosis in adult oral cancer patients?	1	3.33	1	3.33	28	93.33
Q18	What actions during cancer treatment can reduce mucositis in oral cancer patients?	0	0.00	1	3.33	29	96.67
Q19	In which situations is it justified to perform dental treatment during the period of cancer treatment?	7	23.33	8	26.67	15	50.00
Q20	What information is needed in an oncological discharge summary to ensure Good postoperative dental treatment?	0	0.00	3	10.00	27	90.00
Q21	In adult oral cancer patients. In which situations is palliative dental treatment justified?	0	0.00	7	23.33	23	76.67
Q22	Following cancer treatment, what is the treatment of choice for xerostomia depending on its stage?	0	0.00	7	23.33	23	76.67
Q23	In which patients with third molars with associated pathology (such as pericoronitis, cysts, distal surface cavities of the second molar, periodontal disease of the second molar, mandibular fracture, etc.) is there a better clinical outcome (fewer complications) when extraction is performed versus a conservative therapeutic approach (clinical and	0	0.00	4	13.33	26	86.67

(*continued on next page*)

Table 2 (*continued*)

Question		Incorrect		Partially correct or Incomplete		Correct	
		n	%	n	%	n	%
	radiographic monitoring)?						
Q24	Are there preoperative clinical and radiographic criteria that correlate with the degree of surgical difficulty in patients indicated for third molar extraction (shorter operative time and lower morbidity)?	0	0.00	15	50.00	15	50.00
Q25	Do patients with a periodontal probing depth of 4 mm or more distal to the second molar who have (or have not) undergone third molar extraction have a higher incidence of generalized periodontal disease than those with a probing depth of less than 4 mm?	1	3.33	4	13.33	25	83.33
Q26	In patients without anterior crowding, does third molar extraction help to maintain the alignment of the lower anteriors?	1	3.33	6	20.00	23	76.67
Q27	Do patients with third molars without associated pathology benefit from their extraction compared with abstention?	0	0.00	4	13.33	26	86.67
Q28	What guidelines should be followed in patients with fully impacted third molars that are not extracted to avoid complications?	0	0.00	3	10.00	27	90.00
Q29	In third molar surgery, when is it recommended to perform a computed tomography (CT) scan to prevent clinical and/or surgical complications?	0	0.00	2	6.67	28	93.33
Q30	In which patients can the position of the third molar be related to the potential future occurrence of clinical symptoms or pathology compared to those who remain asymptomatic?	0	0.00	0	0.00	30	100.00

between questions within the same topic area.

In the specific field of oral surgery, limited studies have been conducted on the use of ChatGPT, and it is important to note that these studies used different methodological approaches to ours. For example, Vaira et al. [67] formulated 72 open-ended questions for ChatGPT-4, although, unlike our research, they did not use prompts or establish a protocol for generating multiple answers to the same question. Despite these differences, they achieved a remarkable 84.7% accuracy rate in their answers. On the other hand, Balel et al.[68] used ChatGPT-3.5 in their study, which focused on the maxillofacial surgery domain rather than oral surgery per se. These authors concluded that although surgeons should use it cautiously in their technical answers due to its technical imprecision (3.1 ± 1.49), it could be a very useful tool for informing patients in the field of maxillofacial surgery (4.62 ± 0.78).

Table 3
Repeatability assessment (the consistency of the experts' grading) across the 30 repetitions provided by ChatGPT for the 30 questions.

Calculations	Coefficient	Standard Error	95% Confidence Interval		Benchmark scale
Percent Agreement	0.897	0.019	0.858	0.936	Almost Perfect
Brennan and Prediger	0.722	0.052	0.615	0.828	Substantial
Cohen/Conger's Kappa	0.498	0.092	0.310	0.686	Moderate
Scott/Fleiss' Kappa	0.498	0.092	0.310	0.686	Moderate
Gwet's AC	0.825	0.044	0.736	0.915	Almost Perfect
Krippendorff's Alpha	0.498	0.092	0.310	0.686	Moderate

Benchmark scale: Poor < 0.001, Slight 0.001–0.200, Fair 0.200–0.400, Moderate 0.400–0.600, Substantial 0.600–0.800 and Almost Perfect 0.800–1.000.

Considering the diversity of methodological approaches and scales used in the aforementioned studies, it is clear that there is a need to establish a standardized approach to effectively evaluate the utility and accuracy of AI language models in specialized medical contexts. In addition, the integration of these AI tools into healthcare environments must address fundamental ethical issues, including the protection of privacy and security, as well as determining the degree of responsibility in the use of information generated by AI [69,70]. Thus, it is clear that the oversight and active involvement of healthcare professionals and other human experts is essential. While technology can be a valuable ally, it cannot replace the understanding, ethical decision-making and inherent responsibility in healthcare.

This study has several strengths. One is the use of a detailed prompt, which reduces the likelihood of rambling answers from ChatGPT and encourages more specific answers consistent with expert-level knowledge. The reference questions and answers were derived from guidelines developed by a multidisciplinary group of specialists in oral surgery education, research, and clinical practice. Nevertheless, it is important to highlight certain limitations: ChatGPT, by its nature, does not specify the sources of its information and cannot access recently updated documents. In addition, it is important to note that a validated scale was not used in this study. This limitation should be taken into account when evaluating the conclusions and practical applications derived from our study.

5. Conclusion

Given its development and capabilities, it is inevitable that ChatGPT and similar technologies will be actively integrated into the daily practice of oral surgery. However, it is important to recognize that ChatGPT in its current state should not be used indiscriminately. Despite its accessibility to dentists today and the apparent enthusiasm for its adoption, it is imperative that research continues to ensure the safe and effective implementation of AI in oral surgery, with a focus on patient safety and the integrity of medical practice.

In the future, with proper training from validated sources and monitoring by expert oral surgeons, ChatGPT has the potential to become an auxiliary intelligent virtual assistant, but it will never replace the expertise of an oral surgeon.

Contributions

A.S. and YF conceived and designed the study. A.S., Y.F., M.LLdP and J.J. performed the data collection, C.A-V. was responsible for the data analysis, V.D-FG and M.GS performed the literature review, A.S. and Y.

F. drafted the manuscript, all authors have reviewed the manuscript and have approved the final version.

Author statement

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication,with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author.

Conflict of interests

The authors declare no competing interests.

Acknowledgements

The authors have not received any grants or funding for the conduct of this study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.11.058.

References

[1] Manickam P, Mariappan SA, Murugesan SM, Hansda S, Kaushik A, Shinde R, et al. Artificial Intelligence (AI) and internet of medical things (IoMT) assisted biomedical systems for intelligent healthcare. Biosens (Basel) 2022;12:562. https://doi.org/10.3390/bios12080562.

[2] Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, acceptability, and effectiveness of web-based conversational agents to facilitate problem solving in older adults: controlled study. J Med Internet Res 2020;22:e16794. https://doi.org/10.2196/16794.

[3] Talyshinskii A, Naik N, Hameed BMZ, Juliebø-Jones P, Somani BK. Potential of AI-driven chatbots in urology: revolutionizing patient care through artificial intelligence. Curr Urol Rep 2023. https://doi.org/10.1007/s11934-023-01184-3.

[4] Cadamuro J, Cabitza F, Debeljak Z, De Bruyne S, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Medicine (CCLM) 2023;61:1158–66. https://doi.org/10.1515/cclm-2023-0355.

[5] Abd-alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291. https://doi.org/10.2196/48291.

[6] Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023;5:e333–5. https://doi.org/10.1016/S2589-7500(23)00083-3.

[7] Dahmen J, Kayaalp ME, Ollivier M, Pareek A, Hirschmann MT, Karlsson J, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surg, Sports Traumatol, Arthrosc 2023;31:1187–9. https://doi.org/10.1007/s00167-023-07355-6.

[8] Puladi B, Gsaxner C, Kleesiek J, Hölzle F, Röhrig R, Egger J. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int J Oral Maxillofac Surg 2023. https://doi.org/10.1016/j.ijom.2023.09.005.

[9] Arif TBin, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? Med Educ Online 2023;28. https://doi.org/10.1080/10872981.2023.2181052.

[10] OpenAI. Browse is rolling back out to Plus users. Https://HelpOpenaiCom/En/Articles/6825453-Chatgpt-Release-Notes 2023.

[11] Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023;9:e46599. https://doi.org/10.2196/46599.

[12] Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus 2023;15(4):e37432. https://doi.org/10.7759/cureus.37432.

[13] Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J 2023. https://doi.org/10.1111/iej.13985.

[14] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198.

[15] Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13:16492. https://doi.org/10.1038/s41598-023-43436-9.

[16] Carrasco JP, García E, Sánchez DA, Porter E, De La Puente L, Navarro J, et al. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. Rev Esp De Educ Médica 2023;4. https://doi.org/10.6018/edumed.556511.

[17] Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1. https://doi.org/10.3352/jeehp.2023.20.1.

[18] Das D, Kumar N, Longjam LA, Sinha R, Deb Roy A, Mondal H, et al. Assessing the capability of chatgpt in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus 2023. https://doi.org/10.7759/cureus.36034.

[19] Yeo YH, Samaan JS, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023. https://doi.org/10.3350/cmh.2023.0089.

[20] Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33:1790–6. https://doi.org/10.1007/s11695-023-06603-5.

[21] Montastruc F, Storck W, de Canecaude C, Victor L, Li J, Cesbron C, et al. Will artificial intelligence chatbots replace clinical pharmacologists? An exploratory study in clinical practice. Eur J Clin Pharm 2023;79:1375–84. https://doi.org/10.1007/s00228-023-03547-8.

[22] Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621. https://doi.org/10.2196/47621.

[23] Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith DJ, Rozen WM, et al. Evaluating Chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. Aesthet Surg J 2023;43:1126–35. https://doi.org/10.1093/asj/sjad140.

[24] Rawashdeh B, Kim J, AlRyalat SA, Prasad R, Cooper M. ChatGPT and artificial intelligence in transplantation research: is it always correct? Cureus 2023. https://doi.org/10.7759/cureus.42150.

[25] Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI bot has entered the patient chat: chatGPT has limitations in providing accurate urologic healthcare advice. Urology 2023. https://doi.org/10.1016/j.urology.2023.07.010.

[26] Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Oto-Rhino-Laryngol 2023. https://doi.org/10.1007/s00405-023-08104-8.

[27] Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry 2023;22:479–80. https://doi.org/10.1002/wps.21145.

[28] Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023;6:e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483.

[29] Li W, Chen J, Chen F, Liang J, Yu H. Exploring the Potential of ChatGPT-4 in responding to common questions about abdominoplasty: an ai-based case study of a plastic surgery consultation. Aesthetic Plast Surg 2023. https://doi.org/10.1007/s00266-023-03660-0.

[30] Sütcüoğlu BM, Güler M. Appropriateness of premature ovarian insufficiency recommendations provided by ChatGPT. Menopause 2023;30:1033–7. https://doi.org/10.1097/GME.0000000000002246.

[31] Hong D, Huang C, Chen X, Chen L. ChatGPT's responses to gout-related questions. Asian J Surg 2023. https://doi.org/10.1016/j.asjsur.2023.08.217.

[32] Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023;25:e49240. https://doi.org/10.2196/49240.

[33] Hofmann HL, Guerra GA, Le JL, Wong AM, Hofmann GH, Mayfield CK, et al. The rapid development of artificial intelligence: GPT-4's performance on orthopedic surgery board questions. Orthopedics 2023:1–5. https://doi.org/10.3928/01477447-20230922-05.
[34] Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg 2023. https://doi.org/10.1007/s00266-023-03338-7.
[35] Huang C, Chen L, Huang H, Cai Q, Lin R, Wu X, et al. Evaluate the accuracy of ChatGPT's responses to diabetes questions and misconceptions. J Transl Med 2023;21:502. https://doi.org/10.1186/s12967-023-04354-6.
[36] Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. Cureus 2023. https://doi.org/10.7759/cureus.42133.
[37] Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the Japanese society of hypertension guidelines. Circ J 2023;87:CJ-23-0308. https://doi.org/10.1253/circj.CJ-23-0308.
[38] 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. https://doi.org/10.7759/cureus.40822.
[39] Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol 2023. https://doi.org/10.1093/ced/llad255.
[40] Council of European Dentist. The EU Manual of Dental Practice. Https://CedentistsEu/Library/Eu-ManualHtml n.d.
[41] Boletín Oficial del Estado (BOE). Orden CIN/2136/2008. Https://WwwBoeEs/Diario_boe/TxtPhp?Id=BOE-A-2008–12390 2008:31687–92.
[42] Oral Surgery de la Spanish Society of Oral Surgery (Sociedad Española de Cirugía Bucal [SECIB]). Documents of Interest to the Practice. Https://SecibonlineCom/Documentos-de-Interes-Secib/ n.d.
[43] Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6. https://doi.org/10.3389/frai.2023.1166014.
[44] Liu G, Ma X, Zhang Y, Su B, Liu P. GPT4: the indispensable helper for neurosurgeons in the new era. Ann Biomed Eng 2023;51:2113–5. https://doi.org/10.1007/s10439-023-03241-x.
[45] Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021;7:e27850. https://doi.org/10.2196/27850.
[46] Adhikari K, Naik N, Hameed BZ, Raghunath SK, Somani BK. Exploring the ethical, legal, and social implications of ChatGPT in urology. Curr Urol Rep 2023. https://doi.org/10.1007/s11934-023-01185-2.
[47] Biswas S. ChatGPT and the future of medical writing. Radiology 2023;307. https://doi.org/10.1148/radiol.223312.
[48] Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568. https://doi.org/10.2196/48568.
[49] Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 2023;15:29. https://doi.org/10.1038/s41368-023-00239-y.
[50] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Medicine 2023;388:1233–9. https://doi.org/10.1056/NEJMsr2214184.
[51] Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol 2023;13. https://doi.org/10.3389/fonc.2023.1265024.
[52] Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online 2023;28. https://doi.org/10.1080/10872981.2023.2220920.
[53] Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology. Ophthalmol Sci 2023;3:100324. https://doi.org/10.1016/j.xops.2023.100324.
[54] Frosolini A, Franz L, Benedetti S, Vaira LA, de Filippis C, Gennaro P, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Oto-Rhino-Laryngol 2023. https://doi.org/10.1007/s00405-023-08205-4.
[55] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. https://doi.org/10.2196/45312.
[56] Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 Versus ChatGPT-4: vignette study. JMIR Ment Health 2023;10:e51232. https://doi.org/10.2196/51232.
[57] Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP exam: a comparative study with ChatGPT-3.5. MedRxiv 2023:23287957.
[58] Reiss MV. Testing the reliability of ChatGPT for text annotation and classification: a cautionary remark. ArXiv 2023:230411085.
[59] Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: lessons for editors, reviewers, and teachers. Med Teach 2023;45:673–5. https://doi.org/10.1080/0142159X.2023.2208731.
[60] Curtis N. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. 275–275 Pediatr Infect Dis J 2023;42. https://doi.org/10.1097/INF.0000000000003852.
[61] Sharun K, Banu SA, Pawde AM, Kumar R, Akash S, Dhama K, et al. ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references - a preliminary study. Ann Med Surg (Lond) 2023;85:5275–8. https://doi.org/10.1097/MS9.0000000000001228.
[62] Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Method 2023;13:170–8. https://doi.org/10.5662/wjm.v13.i4.170.
[63] Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023. https://doi.org/10.7759/cureus.35179.
[64] Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics 2023;13:1950. https://doi.org/10.3390/diagnostics13111950.
[65] Ali MJ. ChatGPT and Lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg 2023;39:221–5. https://doi.org/10.1097/IOP.0000000000002418.
[66] Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). Eur Arch Oto-Rhino-Laryngol 2023. https://doi.org/10.1007/s00405-023-08219-y.
[67] Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Neck Surg 2023. https://doi.org/10.1002/ohn.489.
[68] Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124:101471. https://doi.org/10.1016/j.jormas.2023.101471.
[69] Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era. Cardiol J 2023. https://doi.org/10.5603/cj.97515.
[70] Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: artificial intelligence in healthcare. Cureus 2023. https://doi.org/10.7759/cureus.43262.