BMC Oral Health

RESEARCH

Open Access

Artificial intelligence-based chatbot assistance in clinical decision-making for medically complex patients in oral surgery: a comparative study

Alanur Çiftçi Şişman[1*] and Ahmet Hüseyin Acar[2]

Abstract

Aim This study aims to evaluate the potential of AI-based chatbots in assisting with clinical decision-making in the management of medically complex patients in oral surgery.

Materials and methods A team of oral and maxillofacial surgeons developed a pool of open-ended questions *de novo*. The validity of the questions was assessed using Lawshe's Content Validity Index. The questions, which focused on systemic diseases and common conditions that may raise concerns during oral surgery, were presented to ChatGPT 3.5 and Claude-instant in two separate sessions, spaced one week apart. Two experienced maxillofacial surgeons, blinded to the chatbots, assessed the responses for quality, accuracy, and completeness using a modified DISCERN tool and Likert scale. Intraclass correlation, Mann-Whitney U test, skewness, and kurtosis coefficients were employed to compare the performances of the chatbots.

Results Most responses were high quality: 86% and 79.6% for ChatGPT, and 81.25% and 89% for Claude-instant in sessions 1 and 2, respectively. In terms of accuracy, ChatGPT had 92% and 93.4% of its responses rated as completely correct in sessions 1 and 2, respectively, while Claude-instant had 95.2% and 89%. For completeness, ChatGPT had 88.5% and 86.8% of its responses rated as adequate or comprehensive in sessions 1 and 2, respectively, while Claude-instant had 95.2% and 86%.

Conclusion Ongoing software developments and the increasing acceptance of chatbots among healthcare professionals hold promise that these tools can provide rapid solutions to the high demand for medical care, ease professionals' workload, reduce costs, and save time.

Keywords Artificial intelligence, Chatbot, Clinical decision-making, ChatGPT, Digital health, Maxillofacial surgery, Oral surgery

*Correspondence:
Alanur Çiftçi Şişman
alanurciftci@gmail.com
[1]Hamidiye Faculty of Dental Medicine, Department of Oral and Maxillofacial Surgery, University of Health Sciences, Istanbul, Türkiye
[2]Faculty of Dentistry, Department of Oral and Maxillofacial Surgery, Istanbul Medeniyet University, Istanbul, Türkiye

Introduction

The field of oral surgery presents unique challenges, particularly when managing medically complex patients [1]. Effective treatment of these cases requires a comprehensive understanding of patients' medical histories to ensure safe surgical procedures, appropriate management of complications,, and a smooth recovery. Medical consultations among healthcare providers are essential for achieving optimal patient outcomes, as effective communication between professionals plays a critical role [2]. However, research indicates that interprofessional communication is often suboptimal, leading to disruptions in care continuity, diagnostic delays, excessive medication use, unnecessary testing, reduced healthcare quality, wasted time, and increased financial costs [3, 4].

The healthcare industry is undergoing a significant transformation driven by the rapid advancement of artificial intelligence (AI) technologies [5]. Traditional, time-consuming, and observer-dependent tasks are increasingly being replaced by AI-based approaches, which can match or even exceed human accuracy [6].

Artificial intelligence (AI)-based chatbots, also known as large language models (LLMs), are advanced software applications that rely on several key technologies to function effectively. Natural Language Processing (NLP) enables chatbots to understand and interpret human language, while Machine Learning (ML) allows them to improve their responses over time by learning from interactions [7, 8]. Using NLP algorithms, chatbots engage in human-like conversations, interpret user queries, and provide immediate responses [9, 10].

The capacity of AI-based chatbots to provide valuable medical information has made them increasingly appealing to both patients and physicians. By delivering precise, real-time answers, they offer a significant advantage over traditional online resources, enhancing their popularity and fostering user trust [10]. Researchers suggest that chatbots could become valuable tools for medical professionals in the future and help alleviate the burden on healthcare systems [10–13].

While AI chatbots have been studied in the context of patient education [9, 12, 14–16] their effectiveness in assisting healthcare professionals in clinical decision-making —particularly in oral and maxillofacial surgery (OMFS) —remains underexplored. It is still uncertain whether these chatbots can consistently offer reliable information to healthcare professionals and assist them in making informed clinical decisions [10, 16, 17]. Given the complexity of managing medically compromised patients in oral surgery, it is crucial to assess whether AI-based chatbots can provide reliable, evidence-based guidance for clinicians.

This study aimed to investigate whether ChatGPT-3.5 and Claude-instant can serve as reliable sources of medical information and also explores their potential to assist professionals in clinical decision-making. By addressing the gap in the literature, this research contributes to the ongoing discussion about the role of AI-driven tools in clinical practice. The null hypothesis was that the chatbots would perform comparably in terms of accuracy, completeness, and quality when providing information on oral surgery for medically complex patients.

Methods

Ethical approval

This study did not involve human or animal subjects; therefore, ethical approval was not required, consistent with previous studies [9, 12, 14, 18].

Sample size and study design

The study was designed as an analytical cross-sectional observational study, following the STROBE checklist, similar to previous research [11, 19].

The sample size estimation was performed using the G*Power 3.1.9.2 software *(University of Düsseldorf, Düsseldorf, NRW, Germany).* The following parameters were considered: (a) test power of 0.8, (b) significance level of 0.05, and (c) effect size of 0.25. Based on these standards, the minimum sample size required was 34 for reliability analyses and 47 per group for difference analyses.

Question development

The study aimed to evaluate the reliability of chatbots and assess the quality, accuracy, and completeness of their responses to specific medical questions. To achieve this, a pool of questions was created, similar to previous studies [20, 21]. Three experienced volunteer oral and maxillofacial surgeons (Surgeons A, B, and C, with 10, 12, and 17 years of experience, respectively), acting as content experts developed the questions *de novo*. The developers were instructed to ensure that the questions met the following criteria: they should be single-focused, clear, and easy to understand; reflect real-world situations; be written in a scientific manner; and be relevant to the field of OMFS.

Relevant literature was identified through a comprehensive search process focusing on systemic diseases and common conditions that typically require professional consultation or may raise concerns during oral surgery. The search terms included specific keywords related to oral and maxillofacial surgery, systemic diseases, and common conditions encountered in this field. These terms included but not limited to: 'oral surgery,' 'systemic diseases and oral surgery,' 'prevalence of systemic diseases in oral surgery,' 'oral health and systemic conditions,' 'dental management considerations,' 'oral surgery complications,' 'oral surgery risk factors,' and 'oral surgery patient management.' The terms were used in various

combinations across databases such as PubMed, Scopus, and Google Scholar to ensure the selection of evidence-based and clinically relevant topics for the development of the questions.

A total of 89 questions were developed. To assess the validity of these questions, each one was evaluated by 10 volunteer oral and maxillofacial surgeons using Lawshe's Content Validity Index (CVI), a widely recognized method for establishing content validity [22]. This method helps determine whether to retain or reject individual items. Experts rated each question as "essential," "useful but not essential," or "not necessary." These ratings were then converted into a quantitative ratio known as the Content Validity Ratio (CVR), using the formula:

$$\mathrm{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

where n_e is the number of experts who rated the item as 'essential' and N is the total number of experts. The critical CVR value is 0.62 for 10 raters at a 0.05 significance level [22]. Hence, questions with a CVR value of ≥ 0.62 were selected for inclusion. As a result, 64 open-ended, clinically relevant questions that requires text-based responses were included. Each question was framed with the prompt: "How would you respond to the following question if you were a doctor?" The examples of the questions are shown in Table 1.

Data collection.

Two chatbots were selected for evaluation: ChatGPT 3.5 (Open AI, San Francisco, USA) and Claude-instant (Anthropic, San Francisco, USA). Access to chatbots was provided online [23, 24] with a new account created in February 2024 for the study. A new chat window was opened for each question to minimize the influence of prior responses. All questions and responses were in English. The identical set of questions was administered in two sessions, one week apart. In each session, the questions were asked consecutively to the chatbots, and their responses were recorded simultaneously. The content generated by the chatbots was used solely for research purposes. The questions and responses are provided as supplementary data. [Supplementary materials]

Chatbot evaluation

The raters for this study were two oral and maxillofacial surgeons, each with 12 and 15 years of experience, respectively. They were not involved in developing the questions, ensuring unbiased evaluations. The number of raters was determined to maximize reliability and agreement levels [25]. To minimize potential bias, inter-rater agreement was assessed using Cohen's kappa coefficient. The kappa values were interpreted as follows: ≤0 indicates no agreement; 0.01–0.20 indicates none to slight agreement; 0.21–0.39 indicates fair agreement; 0.40–0.59 indicates weak agreement; 0.60–0.79 indicates moderate agreement; 0.80–0.90 indicates strong agreement; and >0.90 indicates almost perfect agreement [26, 27]. The raters were blinded to the identity of the chatbots. Consensus scores for each answer were determined based on practical clinical knowledge and PubMed, as described in a previous study [19].

The quality was assessed using a modified DISCERN tool (mDISCERN) as in previous studies [12, 19, 28]. DISCERN is a validated tool for assessing the quality of written consumer health information on treatment options [19, 29]. Details of the mDISCERN scoring are provided in Table 2. Accuracy and completeness were assessed using a Likert scale (Table 3). Through an internal validation process, answers that received lower ratings were subjected to retesting after 12–14 days. Answers rated below 3 points for accuracy were not evaluated for completeness. To assess each chatbot's reliability, the Intraclass Correlation Coefficient (ICC) was used, consistent with a previous study [30]. The flowchart of the methodology is shown in Fig. 1.

Statistical analysis

Skewness and kurtosis coefficients were calculated to examine the normal distribution of the data. Mann-Whitney U Test was used to compare the chatbots chosen due to the non-normal distribution of data. Intraclass correlation (ICC, 95CI%) was used to evaluate intrarater agreement of the chatbots. All analyses were performed using SPSS for Windows (release 21.0, SPSS Inc.), with a 5% significance level.

Table 1 Examples of questions posed to the chatbots

Questions (All starting with the same statement: How would you respond to the following question if you were a doctor?"
Q1. Can oral surgery be performed on patients taking anticoagulants?
Q4. Can oral surgery be safely performed on patients on patients at risk for infective endocarditis?
Q7. Can oral surgery be performed on patients with an INR of 3.5?
Q24. Can oral surgery be performed on patients receiving treatment for leukemia?
Q30. Can oral surgery be safely performed on patients who have previously received chemotherapy?

Table 2 mDISCERN scale

Identification	mDISCERN question
Q1	Is the content relevant, clear and understandable?
Q2	Does the content achieve its aims?
Q3	Does the content describe the risks?
Q4	Is it clear that there may be more than one possible treatment choice?
Q5	Does the content describe what precautions should be taken?
Q6	Does the content describe what would happen if no precaution is taken?
Q7	Is it clear what sources of information were used to compile the content?
Q8	Does the content provide details of additional sources of support and information?
Q9	Does the content refer to areas of uncertainty?
Q10	Is the information presented in the content balanced and unbiased?
Q11	Is it clear when the information used or reported in the information was produced?
Q12	Does the content provide support for shared decision-making?
Overall Rating	
Based on the answers to all of the above questions, rate the overall quality of the content:	
Low (Serious or extensive shortcomings): 1–2	
Moderate (Potentially important but not serious shortcomings): 3–4	
High (Minimal shortcomings): 5	

Table 3 Accuracy and completeness scales

Accuracy scale: 6 point Likert scale	
1	Completely incorrect information
2	More incorrect information than correct information
3	Approximately equal correct and incorrect information
4	More correct than incorrect information
5	Nearly all correct information
6	Completely correct information
Completeness scale: 3-point Likert scale	
1	Incomplete: Addresses some aspects of the question, but significant parts are missing or incomplete.
2	Adequate: Addresses all aspects of the question and provides the minimum amount of information required to be considered complete.
3	Comprehensive: Addresses all aspects of the question and provides additional information or context beyond what was expected.

Results

The chatbots provided one response to each question. Each question was administered to 2 chatbots across 2 separate sessions, 1 week apart, resulting in a total of 128 responses (64 questions × 2 sessions × 1 response per session per chatbot).

The majority of answers were rated as high quality, with 86% ($n = 55/64$) and 79.6% ($n = 51/64$) of responses from ChatGPT in sessions 1 and 2, respectively, receiving scores of 5. For Claude-instant, 81.25% ($n = 52/64$) and 89% ($n = 57/64$) of responses were rated as high quality in sessions 1 and 2, respectively. In terms of accuracy, most answers were rated as completely correct (scores of 4 or above). ChatGPT had 92% ($n = 56/61$) and 93.4% ($n = 57/61$) of responses rated as completely correct in sessions 1 and 2, respectively. Claude-instant had 95.2% ($n = 60/63$) and 89% ($n = 57/64$) of responses rated as completely correct in sessions 1 and 2, respectively. Regarding completeness, most answers were rated as adequate or comprehensive (scores of 2 or above). ChatGPT had 88.5% ($n = 54/61$) and 86.8% ($n = 53/61$) of responses rated as adequate or comprehensive in sessions 1 and 2, respectively. Claude-instant had 95.2% ($n = 60/63$) and 86% ($n = 55/64$) of responses rated as adequate or comprehensive in sessions 1 and 2, respectively. Responses to medication-related osteonecrosis of the jaws (MRONJ)-related questions (Q62-64) received the lowest accuracy scores. The scores for quality, accuracy, and completeness from both chatbots across two sessions are summarized as mean [SD] and median [IQR] in Table 4. Both chatbots showed high consistency in quality across both sessions. In terms of completeness, they exhibited moderate consistency in each session (Table 4). When comparing the chatbots, no statistically significant differences were found in accuracy and completeness. However, ChatGPT showed significantly higher performance in terms of quality in the first session (Table 5). The inter-rater agreement was assessed using Cohen's kappa test, yielding a kappa coefficient (95% CI) of 0.736, indicating a good level of agreement between the two raters.

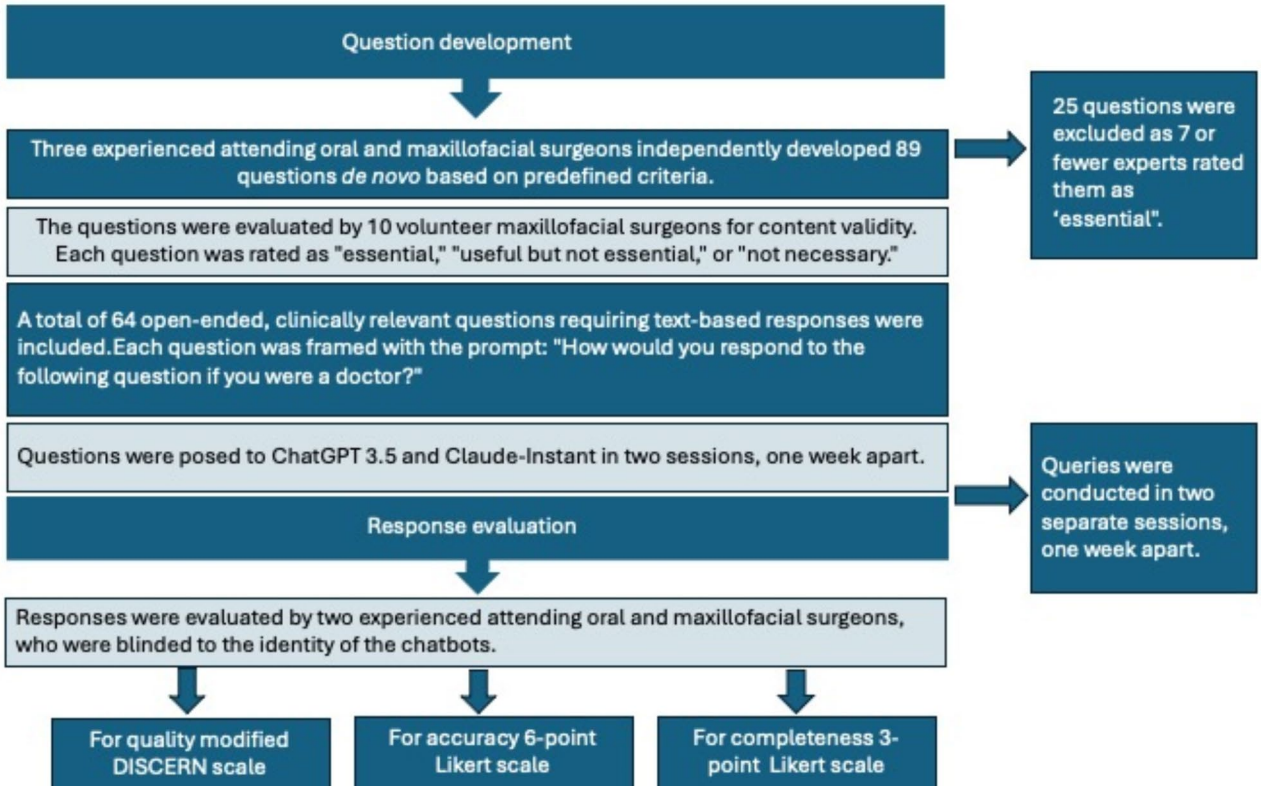

Fig. 1 Flowchart of the study

Table 4 Assesment of reliability of each chatbot in terms of quality, accuracy and completeness

Assessment	Chatbot	Session	*N*	Median [IQR]	Mean [SD]	Intraclass correlation* (ICC) (95%CI)
Quality	ChatGPT	1st session	64	3.00 [2.0–4.0]	2.95 [0.653]	0.906 (0.844-0.943)
		2nd session	64	3.000 [2.0–4.0]	2.88 [0.604]	
	Claude-instant	1st session	64	3.00 [2.0–4.0]	3.02 [0.577]	0.890 (0.820-0.933)
		2nd session	64	3.00 [2.0–4.0]	3.03 [0.616]	
Accuracy	ChatGPT	1st session	61	5.00 [2.0–5.0]	4.88 [0.542]	0.801 (0.672-0.879)
		2nd session	61	5.00 [2.0–5.0]	4.91 [0.357]	
	Claude-instant	1st session	63	5.00 [2.0–5.0]	4.94 [0.393]	0.480 (0.392-0.646)
		2nd session	64	5.00 [2.0–5.0]	4.91 [0.294]	
Completeness	ChatGPT	1st session	61	2.00 [1.0–3.0]	2.24 [0.546]	0.429 (0.061-0.653)
		2nd session	61	2.00 [1.0–3.0]	2.09 [0.334]	
	Claude-instant	1st session	63	2.00 [1.0–3.0]	2.30 [0.558]	0.418 (0.043-0.647)
		2nd session	64	2.00 [1.0–3.0]	2.11 [0.362]	

*Interpretation of ICC Intraclass Corelation Coefficient: 1.0: Perfect agreement. 0.99 to 0.81: Almost perfect agreement. 0.80 to 0.61: Substantial agreement. 0.60 to 0.41: Moderate agreement. 0.20 to 0.01: Slight agreement. 0.0 to 0.1: Poor agreement

Discussion

This study assessed the potential of two AI-based chatbots to assist professionals in clinical decision-making for medically complex oral surgery patients. AI chatbots can vary widely in their performance due to differences in algorithms, datasets, training methods, and design objectives [31]. In this study, the chatbots were selected based on specific criteria, prioritizing ease of access and free subscription. ChatGPT 3.5 was chosen as a pioneering large language model (LLM) with over 100 million users since its release in November 2022 [19, 32]. For comparison, Claude-instant, introduced in March 2023, was selected as a representative of Constitutional AI—a novel alignment strategy focused on context-aware responses aligned with human values [19]. The findings indicate that both chatbots performed similarly in terms of accuracy and completeness. However, ChatGPT received significantly higher quality scores than Claude-instant ($p < .001$), leading to the rejection of the null hypothesis.

Table 5 Comparison of chatbots with Mann-Whitney U test

Assesment	Sessions	Chatbot	*N*	Mean ranks	Z	U	*P*
Quality	1st session	ChatGPT	64	75.59	-3.69	1338.50	$p<.001$
		Claude-instant	64	53.41			
	2nd session	ChatGPT	64	76.53	-3.88	1278.00	0.31
		Claude-instant	64	52.47			
Accuracy	1st session	ChatGPT	61	62.99	-1.17	1951.50	0.24
		Claude-instant	61	66.01			
	2nd session	ChatGPT	63	65.43	-0.98	1926.00	0.32
		Claude-instant	64	62.59			
Completeness	1st session	ChatGPT	61	58.32	-1.19	1666.50	0.23
		Claude-instant	61	64.68			
	2nd session	ChatGPT	63	62.56	-0.79	1925.00	0.43
		Claude-instant	64	65.42			

One of the study's main strengths is that, to the best of our knowledge, it is the first in the field of oral surgery to compare the performance of two different AI-based chatbots across two separate sessions. Additionaly, existing literature primarily focuses on chatbots responding to patient queries, with evaluations centered on patient needs. However, in these studies the assessment were made by professionals. In contrast, this study involves professionals assessing the chatbots' performance specifically for professional use, providing early evidence on the reliability of chatbots in delivering qualified, accurate, and comprehensive information for clinical decision-making while highlighting potential limitations in AI-generated medical content.

The practical application of AI-based chatbots in clinical settings is diverse and can enhance multiple aspects of healthcare delivery. Several studies explore the potential of chatbots to seamlessly integrate into existing workflows, enhancing patient interactions, streamlining administrative tasks, and supporting clinical decision-making. For example, if chatbots can serve as supplementary tools in clinical settings to triage patients and conduct initial assessments [33] provide personalized health advice [34] and function as auxiliary assistants in clinical environments [35]. AI chatbots, such as ChatGPT, can extract information from unstructured data sources like electronic health records, identify patterns and recurring symptoms, and generate diagnostic reports. By automating these tasks, they have the potential to reduce the workload of frontline healthcare workers during routine medical checks. This could, in turn, help alleviate healthcare worker shortages and improve overall efficiency in clinical settings [10].

Despite these promising applications, regulatory considerations for AI-based chatbots in patient care are essential. Ensuring patient safety, data protection, and transparency is critical. There is also a need for clear guidelines regarding liability and accountability, especially in cases where erroneous or harmful advice is provided. Furthermore, continuous monitoring and quality assurance are necessary to ensure that these AI systems remain effective and up-to-date with evolving clinical standards.

Several issues need to be addressed regarding the use of chatbots as a source of medical information. One important factor to consider is the formulation of questions, which significantly affects chatbot responses [32]. Studies have employed various formats, including multiple-choice questions [36], open-ended questions [12, 14, 19, 37] as used in this study or a combination of both [19]. Open-ended better capture the nuances of medical decision-making [11]. However, to ensure standardization, it is essential to structure them consistently. For instance, Wilhelm et al. employed a straightforward pattern, framing questions as "How to treat.?" [19], while Azadi et al. [20] prefaced them with, "What would your response be to the following question if you were an oral and maxillofacial surgeon?". Building on previous studies [20, 21], a pool of open-ended questions was developed *de novo* in this study, and a structured framework was applied to ensure they accurately reflected the complexities physicians face in clinical practice. Care was taken to maintain consistency in question development, with the goal of standardizing the evaluation process and eliciting detailed responses.

Another significant concern about AI-driven information is the variability in chatbot responses to identical questions. Sanmarchi et al. reported that the responses can vary when the same questions is repeated, reflecting the nature of ML algorithms [31]. Most studies [12, 18, 20] have posed each question only once. To address this variety, questions were posed in two separate sessions in this study. One-week duration was selected to minimize memory bias for both raters and chatbots. This waiting period was intended to better simulate a real-life scenario, where patients typically experience some time between consultations. Furthermore, through an internal validation process, the answers that received lower

ratings were rated after 12–14 days. In this regard, the current study aligns with existing research. Several studies in the literature have re-evaluated chatbot-generated responses. Onder et al. [28] tested each question twice on different days for variation in answers but did not specify the duration between tests. In studies by Wilhelm et al. [19] and Goodman et al. [11] re-evaluations occurred between 8 to 17 days.

In this study, internal consistency showed almost perfect agreement in quality for both chatbots, though completeness exhibited moderate agreement. In addition, all responses obtained were relevant and generated within seconds, but some were broad or non-specific, while others were detailed. For example, in Q9: Is it safe to perform oral surgery on patients taking Coumadin?, ChatGPT included the International Normalized Ratio (INR) value in its response. In Q34:Can oral surgery be performed on patients receiving radiotherapy to the head and neck region?, Both chatbots provided a detailed answer, mentioning hyperbaric oxygen therapy and antibody prophylaxis. In Q22:Is hematocrit level important for performing oral surgery?, Claude-instant provided a specific numeric value for the hematocrit level, whereas ChatGPT did not. Chatbots' responses generally indicated that oral surgery in patients at risk for infective endocarditis should be approached with caution, recommending prophylactic antibiotic use in line with established guidelines [38]. However, upon closer examination, while the chatbots acknowledged risk factors, they occasionally oversimplified the decision-making process. This simplification may have overlooked critical nuances, such as the specific dental procedure being performed or the presence of patient comorbidities, both of which are key considerations when making informed clinical decisions. In another example, chatbot responses on the management of leukemia patients emphasized the need for multidisciplinary care, including attention to immunosuppression and bleeding risks. This aligns with existing literature that underscores the complexity of surgical interventions in immunocompromised patients [39, 40]. However, the chatbots' responses sometimes lacked depth, particularly regarding the importance of preoperative hematological assessments or the specific timing of surgical interventions in relation to chemotherapy cycles. These factors are essential in clinical decision-making and were not sufficiently addressed in the chatbot-generated responses, highlighting a gap in their clinical applicability.

Notably, responses with the lowest accuracy were specifically related to MRONJ, likely due to its evolving status in OMFS. We observed that AI chatbots struggle to provide accurate interpretations without specialized training, particularly in areas where personalized information and human judgment are essential. This result is consistent with the study by Suárez et al., [35] which was also conducted in oral surgery and shares a similar methodology with this study. The researchers reported that ChatGPT, by its nature, does not specify the sources of its information and cannot access recently updated documents. This finding underscores the current limitations of AI-based chatbots in handling specific medical topics and highlights the need for continuous updates and training to improve their reliability.

The ethical risks associated with AI-generated medical content must be carefully considered, as misinformation could compromise patient safety. The use of non-specialized training data, the potential for outdated information, and ethical and legal concerns regarding patient confidentiality necessitate thorough evaluation [35]. Goodman et al. evaluated ChatGPT's responses to medical queries from 33 physicians across 17 specialties and found that, while ChatGPT generally provided accurate information, it occasionally made unwarranted assumptions [11]. This phenomenon, known as "hallucination," refers to the generation of scientifically incorrect content. It occurs when a chatbot provides seemingly reliable but inaccurate answers, posing a serious concern due to the potential for misinformation in clinical settings. The real danger of these "made up facts" is that they often appear scientifically plausible, making them particularly misleading [13, 41]. Chow et al. suggested that if ChatGPT were professionally trained, it could operate more efficiently, access larger datasets, and help reduce medical errors [10]. However, the dynamic and continuously evolving nature of AI learning makes it challenging to ensure the credibility of the information generated by AI models [42]. Accurate medical information is critical for patient health, and medicine cannot rely on tools that occasionally provide incorrect answers, even if such instances are infrequent [11, 43]. In this study, no instances of hallucination were observed; however, this finding should be interpreted with caution. The controlled study design may have played a role in the absence of hallucinations.

In OMFS, several studies have investigated the information provided by chatbots. Balel conducted a study evaluating the usability of ChatGPT in OMFS by assessing the quality of patient information and educational content produced by ChatGPT. Commonly asked patient-questions about OMFS procedures, as well as technical questions for training purposes, were posed to the chatbot. The responses were evaluated by 33 academic maxillofacial surgeons. The study reported that, despite concerns about its safety in educational contexts, ChatGPT demonstrates significant potential as a valuable tool for patient information in OMFS [18]. Similarly, Acar compared the effectiveness of 3 AI-based chatbots (ChatGPT, Microsoft Bing, Google Bard) regarding the information they provide to patients. Twenty questions related to oral

surgery complications were posed to each chatbot, and 10 oral surgeons evaluated the responses for accuracy and completeness. ChatGPT provided both more accurate and understandable answers compared to the other two platforms [14]. Jacobs et al. evaluated the accuracy and readability of AI-generated responses to common patient questions regarding third molar extraction, specifically using ChatGPT. They reported that ChatGPT provided largely accurate information, though with some minor inaccuracies [37]. The present study yielded similar results, with ChatGPT receiving higher quality scores. This may be attributed to ChatGPT being the large language model (LLM) with the largest user base worldwide. The advancement of LLMs in generating knowledge is largely due to their continuous training on extensive text data and a feedback loop mechanism through which they learn from corrections and user interactions. However, these studies have primarily focused on assessing the content for patients. Consequently, the potential of AI-based chatbots to deliver valuable insights to healthcare professionals remains an underexplored area.

A study similar to the present one was conducted by Azadi et al., who evaluated the accuracy of chatbot responses to clinical decision-making questions in OMFS using the Global Quality Scale (GQS) [20]. Their study assessed Google Bard, GPT-3.5, GPT-4, Claude-instant, and Bing by presenting them with 50 case-based questions prepared by 3 oral and maxillofacial surgeons. These questions were designed in both multiple-choice and open-ended formats, specifically focusing on OMFS-related topics. While the chatbots performed relatively well in answering open-ended questions, the study concluded that they are not yet reliable advisors for clinical decision-making due to significant inaccuracies in their responses. Additionally, the researchers noted a preference for asking open-ended questions rather than multiple-choice ones when using these AI tools. Given the similarities in methodology, this study also adopted an open-ended question format to better reflect real-world usage and assess the quality of chatbot-generated responses in a clinical context. In comparison to existing literature on AI in clinical decision-making, these findings suggest that while chatbots may be able to provide useful guidance in general terms, they often fall short in capturing the full complexity of clinical scenarios. This limitation is important to consider when evaluating the potential for chatbots to be integrated into real-world clinical settings, where nuanced decision-making is frequently required.

The study has several limitations. First, it was conducted at a single center and evaluated by only two experts. Although this approach was chosen to ensure maximum reliability and consensus, as supported by the literature [25], using only two expert raters may introduce confirmation bias, as their assessments might be influenced by preexisting expectations or familiarity with clinical guidelines. Furthermore, the single-center design limits the diversity of expert opinions, and multicenter studies with a larger number of evaluators could provide more comprehensive insights. In this context, potential biases related to the evaluation process should be acknowledged. Another concern is the inherent risk of bias arising from the chatbot training data, which may lead to systemic biases in the generated responses. The data used to train the chatbot may contain biases or imbalances that reflect the views, demographics, or limitations present in the original sources. Since the chatbot learns from this training data, any existing bias—such as the underrepresentation of certain groups, stereotypes, or outdated information—can be incorporated into and reproduced in its responses. This can result in systematic errors or skewed information, which may affect the quality and fairness of the chatbot's output in real-world applications.

The present study, like most chatbot studies, was conducted in English and yielded similar results [14, 18, 20]. To our knowledge, only one study by Soto-Chávez et al., has evaluated ChatGPT's performance in Spanish, and reported that while ChatGPT can be a reliable source of information for Spanish-speaking patients, its readability and accuracy vary across languages [44].

Numerous AI-based chatbots are available today, including those specifically designed for medical purposes. However, this study focused solely on two general-purpose chatbots, chosen for their free accessibility, ease of use, and widespread recognition. This selection may restrict the broader applicability of the findings and does not account for potential variability among other AI models.

The field of AI is rapidly evolving, and the quality, accuracy, and completeness of chatbot responses may improve with subsequent model updates. At the time of the study, the most advanced iterations of these chatbots were available only in a limited number of countries and required a paid subscription. However, evidence has been presented suggesting that there is no significant difference in the quality of medical content generated by ChatGPT 3.5 and ChatGPT 4 [20]. Based on this evidence, it was decided to proceed with ChatGPT 3.5 for this study. Further research exploring the performance of newer iterations of the model could make a valuable contribution to the literature.

The limited number of questions included in the study may not fully capture the breadth of clinical scenarios encountered in practice. Future research that incorporate a broader and more discriminating set of questions can better assess the capabilities and limitations of AI tools in diverse clinical contexts.

Despite these limitations, this study is novel in both its aim and methodology. It enhances the existing literature on AI-based chatbots in oral and maxillofacial surgery by evaluating the quality of medical content intended for professionals.

Conclusion

In conclusion, this study underscores the potential of AI-based chatbots to support professionals in clinical decision-making for medically complex patients undergoing oral surgery. It also highlights the necessity for ongoing advancements in AI-generated content to ensure patient safety and deliver high-quality, reliable, and accurate information. Further research is needed to assess the evolution of these tools over time, addressing the dynamic nature of machine learning algorithms and their limitations. Although they are currently insufficient as a sole source of information, AI-based chatbots continue to develop and offer a promising solution to the growing demand for medical care. Their potential to enhance the efficiency and effectiveness of healthcare, could help alleviate the workload of healthcare professionals, reduce costs, and save time.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-05732-w.

Supplementary Material 1

Acknowledgements
The authors sincerely appreciate the reviewers for their time and effort in evaluating the manuscript. Their thoughtful and constructive feedback greatly contributed to enhancing the quality of the final version.

Author contributions
AÇŞ and AHA were involved in conceptualization, data collection and processing, and methodology. AÇŞ carried out the analysis and interpretation of the data and literature review. All the authors contributed to the writing of the original draft and commented on previous versions of the manuscript. All the authors read and approved the final manuscript.

Funding
The authors did not receive support from any organization for the submitted work.

Data availability
The datasets used and/or analyzed during the current study are available in the Supplementary Materials.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 20 December 2024 / Accepted: 27 February 2025
Published online: 07 March 2025

References

1. Mahtani A, Santhosh MP. Prevalence of systemic diseases in patients undergoing minor oral surgeries. Bioinformation. 2020;16(12):1051–9. PMID: 34938005; PMCID: PMC8600205.Mahtani A, Santhosh MP. Prevalence of systemic diseases in patients undergoing minor oral surgeries. Bioinformation. 2020. https://doi.org/10.6026/973206300161051.
2. Kessler CS, Tadisina KK, Saks M, Franzen D, Woods R, Banh KV, Bounds R, Smith M, Deiorio N, Schwartz A. The 5Cs of Consultation: Training Medical Students to Communicate Effectively in the Emergency Department. J Emerg Med. 2016. https://doi.org/10.1016/j.jemermed.2016.08.002.
3. Kömerik N, Çadır B. The analysis of referral letters requested from the oral surgery department: Is the communication between medical and dental professionals a neglected issue. AOT. 2004;21(3):205–208 (Article In Turkish).
4. Gandhi TK, Sittig DF, Franklin M, Sussman AJ, Fairchild DG, Bates DW. Communication breakdown in the outpatient referral process. J Gen Intern Med. 2000. https://doi.org/10.1046/j.1525-1497.2000.91119.x.
5. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. Digit Health. 2019. https://doi.org/10.1177/2055207619871808.
6. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019. https://doi.org/10.7861/futurehosp.6-2-94.
7. Krishnan C, Gupta A, Gupta A, Singh G. Impact of artificial intelligence-based Chatbots on customer engagement and business growth. Deep Learn Soc Media Data Anal 2022. https://doi.org/10.1007/978-3-031-10869-3_11.
8. Sidlauskiene J, Joye Y, Auruskeviciene V. AI based chatbots inconversational commerce and their effects on product and priceperceptions. Electron Mark 2023. https://doi.org/10.1007/s12525-023-00633-8.
9. Perez-Pino A, Yadav S, Upadhyay M, Cardarelli L, Tadinada A. The accuracy of artificial intelligence-based virtual assistants in responding to routinely asked questions about orthodontics. Angle Orthod. 2023. https://doi.org/10.2319/100922-691.1.
10. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell. 2023. https://doi.org/10.3389/frai.2023.1166014.
11. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023. https://doi.org/10.1001/jamanetworkopen.2023.36483.
12. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024. https://doi.org/10.1186/s12911-024-02619-8.
13. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023. https://doi.org/10.3390/healthcare11060887.
14. Acar AH. Can natural language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg. 2024. https://doi.org/10.1016/j.jormas.2023.101724.
15. Ferrari-Light D, Merritt RE, D'Souza D, Ferguson MK, Harrison S, Madariaga ML, Lee BE, Moffatt-Bruce SD, Kneuertz PJ. Evaluating ChatGPT as a patient resource for frequently asked questions about lung cancer surgery-a pilot study. J Thorac Cardiovasc Surg. 2024. https://doi.org/10.1016/j.jtcvs.2024.09.030.
16. Hosseini M, Gao CA, Liebovitz DM, Carvalho AM, Ahmad FS, Luo Y, MacDonald N, Holmes KL, Kho A. An exploratory survey about using ChatGPT in education, healthcare, and research. PLoS One. 2023. https://doi.org/10.1371/journal.pone.0292216.
17. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer KJ, Succi MD. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. medRxiv [Preprint]. 2023 Feb 26:2023.02.21.23285886. https://doi.org/10.1101/2023.02.21.23285886. Update in: J Med Internet Res. 2023 Aug 22;25:e48659. https://doi.org/10.2196/48659. PMID: 36865204; PMCID: PMC9980239.
18. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023. https://doi.org/10.1016/j.jormas.2023.101471.
19. Wilhelm TI, Roos J, Kaczmarczyk R. Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study. J Med Internet Res. 2023. https://doi.org/10.2196/49324.

20. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024. https://doi.org/10.1016/j.oooo.2024.02.018.
21. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023. https://doi.org/10.2196/51580.
22. Wilson FR, Pan W, Schumsky DA. Recalculation of the Critical Values for Lawshe's Content Validity Ratio. Measurement and Evaluation in Counseling and Development 2012. https://doi.org/10.1177/0748175612440286.
23. https://chatgpt.com.
24. https://poe.com.
25. Bikmaz Bilgen O, Doğan N. The comparison of interrater reliability estimating techniques. JMEEP. 2017;8(1):63–78.
26. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276–282.
27. Özkan E, Acar AH. YouTube™ as a Source of Information for Patients Regarding Dental Implant Failure: A Content Analysis. J Craniofac Surg. 2022. https://doi.org/10.1097/SCS.0000000000008609.
28. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep. 2024. https://doi.org/10.1038/s41598-023-50884-w.
29. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999. https://doi.org/10.1136/jech.53.2.105.
30. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation - A discussion and demonstration of basic features. PLoS One. 2019. https://doi.org/10.1371/journal.pone.0219854.
31. Sanmarchi F, Bucci A, Nuzzolese AG, Carullo G, Toscano F, Nante N, Golinelli D. A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: an exploratory analysis of ChatGPT using the STROBE checklist for observational studies. Z Gesundh Wiss. 2023. https://doi.org/10.1007/s10389-023-01936-y.
32. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in Dentistry: A Comprehensive Review. Cureus. 2023. https://doi.org/10.7759/cureus.38317.
33. Liu X, Lai R, Wu C, Yan C, Gan Z, Yang Y, Zeng X, Liu J, Liao L, Lin Y, Jing H, Zhang W. Assessing the utility of artificial intelligence throughout the triage outpatients: a prospective randomized controlled clinical study. Front Public Health. 2024. https://doi.org/10.3389/fpubh.2024.1391906.
34. Akarsu K, Er O. Artificial Intelligence Based Chatbot in E-Health System. Artificial Intelligence Theory and Applications. 2023;3(2):113–122.
35. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, Freire Y. Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. 2023. https://doi.org/10.1016/j.csbj.2023.11.058.
36. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthcare (Basel). 2023. https://doi.org/10.3390/healthcare11142046.
37. Jacobs T, Shaari A, Gazonas CB, Ziccardi VB. Is ChatGPT an Accurate and Readable Patient Aid for Third Molar Extractions? J Oral Maxillofac Surg. 2024. https://doi.org/10.1016/j.joms.2024.06.177.
38. https://www.ada.org/resources/ada-library/oral-health-topics/antibiotic-prophylaxis
39. O'Rourke K. Study examines oral health of patients with leukemia. Cancer. 2022. https://doi.org/10.1002/cncr.34081.
40. Ptasiewicz M, Maksymiuk P, Chałas R. Changes of Dentition State in Leukemic Patients during Chemotherapy. Int J Environ Res Public Health. 2021. https://doi.org/10.3390/ijerph18158193.
41. Erren TC, Lewis P, Shaw DM. Brave (in a) new world: an ethical perspective on chatbots for medical advice. Front Public Health. 2023. https://doi.org/10.3389/fpubh.2023.1254334.
42. 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. https://doi.org/10.7759/cureus.37432.
43. Kessels RP. Patients' memory for medical information. J R Soc Med. 2003;96(5):219– 22. PMID: 12724430; PMCID: PMC539473.Kessels RP. Patients' memory for medical information. J R Soc Med. 2003. https://doi.org/10.1177/014107680309600504.
44. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish Language. Digit Health. 2024;10:20552076231224603. PMID: 38188865; PMCID: PMC10768597.Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digit Health. 2024. https://doi.org/10.1177/20552076231224603.

Publisher's note