Can Artificial Intelligence Chatbots (ChatGPT, DeepSeek, Copilot, and Claude) Accurately Inform Patients About Subperiostal Jaw Implants?

Selin Gaş, DDS, PhD, Tuğçe Paksoy, DDS,† Cemre Çeçen, DDS,* Sevda Altınay Uncu, DDS,† and Fırat Selvi, DDS, PhD‡*

Background: Severe jawbone atrophy often limits the use of conventional dental implants. Modern 3D imaging and CAD/CAM technologies have revitalized subperiosteal implants as a customized alternative for these challenging cases. As AI chatbots (ChatGPT, Claude, DeepSeek, and Copilot) increasingly serve as patient information tools, evaluating the accuracy and clarity of their explanations about such complex procedures has become essential.

Objective: This study aimed to evaluate the accuracy, reliability, readability, actionability, understandability, and practical usefulness of responses provided by AI chatbots to patient questions about subperiosteal jaw implants.

Methods: The authors evaluated 4 AI-based chatbots (ChatGPT, DeepSeek, Copilot, and Claude) by submitting frequently asked questions on subperiosteal jaw implants in independent sessions to avoid data leakage. Responses were compiled, duplicates removed, and refined for clarity. Two independent experts assessed the outputs using validated tools: accuracy (5-point Likert), reliability (CLEAR criteria), quality (mGQS), readability (FRE, FKGL), usefulness (4-point scale), and understandability/actionability (PEMAT).

Results: DeepSeek, Claude, and ChatGPT produced more understandable, actionable, and higher-quality responses than CoPilot, with DeepSeek performing the best overall. Across all models, clarity, mGQS, and accuracy were strongly aligned, while usefulness was inversely related. Readability and actionability-understandability correlations showed consistent patterns, with the strongest positive link observed in DeepSeek.

Conclusion: AI chatbots, such as DeepSeek, Claude, and ChatGPT, can provide accurate and understandable information about the subperiosteal jaw implants, though practical guidance and readability remain limited. Domain-specific training and integration with authoritative dental resources may enhance their clinical utility and patient education potential.

Key Words: AI chatbots, artificial intelligence, chatGPT, DeepSeek, subperiosteal implants

(*J Craniofac Surg* 2026;37: 1648–1651)

Dental implants have become a popular and effective solution for restoring missing teeth over the past 50 years, with endosseous implants being the most commonly used type. However, their effectiveness is limited by surrounding anatomic structures and challenges such as severe bone atrophy or bone resection.[1] While recent advancements in implant design and various surgical techniques like sinus lifting have provided solutions for maxillary atrophy, the mandible lacks similar options for regional anchorage, making bone augmentation technically challenging and unpredictable.[2–8] Subperiosteal jaw implants have emerged as a contemporary solution to avoid these complications. Modern approaches using 3D radiographs and CAD/CAM-supported manufacturing have addressed traditional fabrication limitations, offering advantages such as the elimination of donor site morbidity, reduced surgical time, and suitability for patients with significant bone defects due to trauma or oncological treatments.[9–12]

Artificial intelligence (AI) has evolved rapidly, with breakthroughs in deep learning presenting transformative opportunities to revolutionize health care practices.[13–15] AI chatbots are gaining popularity among health care professionals and patients for accessing medical and dental information, offering 24/7 availability and quick responses.[16,17] However, concerns about incorrect answers and misleading information raise questions about their reliability in health care settings.[18,19]

While patients are beginning to understand endosseous implants, comprehending the complex design of subperiosteal jaw implants may be challenging. The purpose of this article is to examine the questions patients most frequently ask AI chatbots (ChatGPT, Claude, DeepSeek, and Copilot) about subperiosteal implants and investigate the accuracy of the responses.

METHODS

This study was conducted in accordance with the principles outlined in the Declaration of Helsinki and did not require approval from an ethics committee.

From the *Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Istanbul Gelisim University; †Department of Periodontology, Hamidiye Faculty of Dentistry, University of Health Sciences; and ‡Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Istanbul University, Istanbul, Turkey.

Received December 15, 2025.

Accepted for publication December 16, 2025.

Address correspondence and reprint requests to Selin Gaş, DDS, PhD, Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Istanbul Gelisim University, Istanbul 34320, Turkey; e-mail: selineren104@gmail.com.

The authors report no conflicts of interest.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.jcraniofacialsurgery.com.

ISSN: 1536-3732

DOI: 10.1097/SCS.0000000000012417

We asked ChatGPT 4 (https://chat.openai.com/), DeepSeek (https://chat.deepseek.com/), Copilot (https://copilot.microsoft.com/), and Claude (https://claude.ai/new), 4 different AI-based chatbots, what are the most frequently asked questions about subperiosteal implants? Each question was submitted independently through the user interface of the respective chatbot. To mitigate the risk of data leakage and to ensure that prior queries and responses did not influence subsequent outputs, a new conversation window was initiated for each question. This procedure was used to closely approximate the conditions of real-world patient inquiries. The responses were recorded and evaluated in terms of accuracy, quality, reliability, readability, and usefulness.

All 24 questions derived from the online tool and the expert panel were consolidated into a single Excel spreadsheet. Two authors independently reviewed the items to identify and remove duplicates as well as semantically equivalent questions. The remaining items were subsequently assessed for grammatical accuracy and clarity, and any linguistic or stylistic deficiencies were corrected.

Thereafter, 2 independent experts evaluated all AI-generated responses. The evaluation used the CLEAR criteria, assessments of usefulness, the modified Global Quality Score (mGQS), accuracy ratings, the Likert scale, 2 standard readability indices, which are the Flesch Reading Ease (FRE), and the Flesch-Kincaid Grade Level (FKGL), and the Patient Education Material Assessment Tool (PEMAT).

PEMAT was used to evaluate the understandability and actionability of chatbot responses. The PEMAT provides a systematic framework for assessing and comparing patient education materials.[20] Materials are considered understandable when individuals from diverse backgrounds and with varying levels of health literacy can process and explain the key messages. Materials are considered actionable when such individuals can identify specific actions to take based on the information provided. The PEMAT consists of 2 domains. The understandability domain includes 19 items, each scored as 1 (agree), 0 (disagree), or NA (not applicable). The actionability domain comprises 7 items, scored in the same manner. For both domains, scores are calculated as percentages ranging from 0% to 100%, with higher values indicating greater understandability or actionability.[20] The chatbot's answers are completely incorrect; Score 2 means the chatbot's answers contain more incorrect items than correct items; Score 3 means the chatbot's answers contain an equal balance of correct and incorrect items; Score 4 means the chatbot's answers contain more correct items than incorrect items, and Score 5 means the chatbot's answers are completely correct.

Accuracy was evaluated using a 5-point Likert scale.[21,22] On this scale score 1: the chatbot's answers are completely incorrect; score 2: the chatbot's answers contain more incorrect items than correct items; score 3: the chatbot's answers contain an equal balance of correct and incorrect items; score 4: the chatbot's answers contain more correct items than incorrect items; and score 5: the chatbot's answers are completely correct.

Reliability was operationalized through the CLEAR criteria, which encompass 5 evaluative domains: completeness of content, absence of false information, evidence-based support, appropriateness of information, and relevance to the topic.[23] Each domain was rated on a scale from 1 to 5, with 1 representing "very poor" and 5 representing "excellent." The total CLEAR score ranged from 5 to 25, where scores between 5 and 11 indicated low quality, 12 and 18 moderate quality, and 19 and 25 high quality.

The quality of the responses was assessed using the Global Quality Scale (GQS), a 5-point scoring system that has been widely applied in previous research.[24,25]

Readability was evaluated using 2 widely recognized measures commonly used in the literature: the Flesch Reading Ease (FRE) and the Flesch-Kincaid Grade Level (FKGL), both calculated via an online tool (Readable Pro: https://app.readable.com/text/).[26] The FRE produces scores ranging from 0 to 100, with lower values reflecting greater textual complexity and reduced readability.[27] The FKGL, in contrast, assigns scores from 0 to 18, each score indicating the number of years of formal education required to understand the material.[28] According to the recommendations of the American Medical Association (AMA) and the National Institutes of Health (NIH), patient education materials should be written at or below a sixth-grade reading level.

The usefulness of each response was rated using a 4-point scale, where 1 denoted "very useful" and 4 denoted "not useful," based on the extent to which the content contributed to understanding by patients or health care professionals.[2] In this study, the usefulness of AI-generated responses was evaluated using a 4-point scale, which classifies responses as: (1) very useful (comprehensive and correct information comparable to that provided by a subject specialist); (2) useful (accurate but lacking some essential details); (3) partially useful (includes some correct content alongside incorrect or misleading elements); and (4) not useful (contains only incorrect or misleading information without any accurate content).

The DISCERN scale, a 3-component instrument, has been used in prior research to assess the reliability and quality of online health information. The 8 questions of the modified DISCERN scale assess the following:

1. Clarity of objectives.
2. Achievement of objectives.
3. Level of relevance.
4. Which information sources were used?
5. When the information used or reported was produced.
6. Balance and impartiality.
7. Details of additional support and information sources.
8. References to areas of uncertainty.

For each question in the modified DISCERN scale, no is scored as 1; partial answer as 2, 3, 4; and yes as 5, and the total score is thus calculated.[29]

Statistical Analysis

Descriptive statistics (mean, SD, median, and interquartile range) were calculated. Normality was assessed using the Shapiro-Wilk test. For comparisons among 3 or more non-normally distributed independent groups, the Kruskal-Wallis test was applied, followed by Bonferroni-adjusted post hoc analyses. Associations between non-normally distributed continuous variables were examined using the Spearman rank correlation. All analyses were conducted with IBM SPSS version 27.

RESULTS

The distributions of response characteristics by AI type are summarized in Table 1, Supplemental Digital Content 1, http://links.lww.com/SCS/J22. Kruskal-Wallis tests revealed statistically significant differences among AI types in Understandability, actionability, and DISCERN scores ($P < 0.05$). Post hoc Bonferroni analyses revealed that for Understandability, DeepSeek achieved significantly higher scores

than both ChatGPT ($P = 0.003$) and Copilot ($P < 0.001$). For Actionability, ChatGPT, Claude, and DeepSeek each scored significantly higher than Copilot ($P = 0.011$, 0.015, and 0.020, respectively). Regarding DISCERN scores, both ChatGPT and DeepSeek outperformed Copilot ($P = 0.001$ and 0.041, respectively).

Spearman correlation analyses for all AI chatbots (Tables 2–5, Supplemental Digital Content 1, http://links.lww.com/SCS/J22) revealed consistent statistically significant patterns ($P < 0.05$) across ChatGPT, Claude, Copilot, and DeepSeek. All AI types demonstrated very strong to perfect positive correlations between the CLEAR score and both GQS and Accuracy scores, alongside strong to perfect negative associations between these measures and the Usefulness score. Moderate to strong positive correlations were consistently observed between quality measures (CLEAR, GQS, and Accuracy) and DISCERN score, while Usefulness score showed moderate negative associations with DISCERN. Flesch Reading Ease Score was negatively correlated with Flesch Reading Grade Level across all AI types. In addition, strong to very strong positive associations were found between Understandability percentage and Actionability percentage for Claude and DeepSeek, with a moderate positive association for ChatGPT.

DISCUSSION

This study evaluated ChatGPT, DeepSeek, Claude, and Copilot in providing information on subperiosteal jaw implants. Ensuring accurate and understandable communication is essential for clinicians and patients, and AI chatbots—accessible through simple web interfaces—offer quick and convenient information, potentially improving health care accessibility and patient engagement.

The CLEAR tool effectively assessed the reliability of AI-generated responses, showing strong correlations with GQS and accuracy across all models, consistent with previous findings.[30,31] DeepSeek and Claude achieved the highest correlations ($\rho = 0.986$–0.988), followed by ChatGPT, while Copilot performed the weakest ($\rho = 0.959$). The inverse relationship between reliability and usefulness suggests that highly accurate content may not always translate into greater practicality for users.

In terms of GQS, Claude and ChatGPT achieved the best results, followed by DeepSeek, while Copilot performed the poorest—aligning with most prior studies.[23,30,32] Contradictory results by Dursun et al[33] may reflect topic-specific variation. Accuracy followed a similar pattern: ChatGPT and Claude scored highest (4.48 ± 0.71), DeepSeek slightly lower (4.32 ± 0.95), and Copilot lowest (3.92 ± 0.91). Consistent with earlier work, domain-specific models such as Dental GPT and ChatGPT-o1 outperformed general-purpose LLMs.[34–36] However, fabricated references remain a recurring limitation.[37,38]

Usefulness ratings were generally low, with Copilot rated most practical but least reliable. Claude's responses were most readable (grade 6–7), while ChatGPT's were most complex (postgraduate level), echoing earlier findings that AI-generated medical content often exceeds recommended readability for patient education.[30,39,40]

DeepSeek achieved the highest DISCERN scores (28.52 ± 18.82), followed by ChatGPT and Claude, confirming its reliability. Consistent with previous research,[24,30,34] all chatbots demonstrated high understandability but limited actionability, highlighting that while AI-generated responses are informative, they often lack practical guidance for patients.

Overall, ChatGPT, Claude, and DeepSeek provided accurate and reliable dental information, while Copilot underperformed across most metrics. Despite strong results, limitations such as language restriction, readability, and reference fabrication remain. AI chatbots should therefore complement, not replace, professional expertise. Continued refinement and domain-specific development are essential to improve their accuracy, transparency, and usability in dental education and patient communication.

CONCLUSION

AI chatbots demonstrate considerable potential in providing information about subperiosteal implants, a complex dental treatment for patients with severe alveolar bone loss. DeepSeek and Claude consistently produced the most reliable and understandable outputs, while ChatGPT delivered highly accurate information with moderate actionability, albeit inconsistently. CoPilot generally underperformed across reliability, quality, readability, and actionability metrics.

Given the intricate nature of subperiosteal jaw implants, accurate and understandable patient education is critical. While AI chatbots can support learning and decision-making, limitations in readability, actionability, and citation reliability indicate that human oversight remains essential.

Future improvements should focus on enhancing the practical guidance of AI-generated content, simplifying language for patient comprehension, integrating domain-specific data sets, and linking outputs to authoritative dental and medical sources. Such developments will maximize the clinical utility, safety, and accessibility of AI tools in educating patients and supporting clinicians regarding subperiosteal implant therapy.

REFERENCES

1. Goh R, Vaquette C, Breik O, et al. Subperiosteal implants: a lost art worth revisiting? *Clin Implant Dent Relat Res* 2025;27:e70025
2. Anitua E, Eguia A, Staudigl C, et al. Clinical performance of additively manufactured subperiosteal implants: a systematic review. *Int J Implant Dent* 2024;10:4
3. Antonoglou GN, Stavropoulos A, Samara MD, et al. Clinical performance of dental implants following sinus floor augmentation: a systematic review and meta-analysis of clinical trials with at least 3 years of follow-up. *Int J Oral Maxillofac Implants* 2018;33:e45–e65
4. Robert L, Aloy-Prósper A, Arias-Herrera S, et al. Vertical augmentation of the atrophic posterior mandibular ridges with onlay grafts: intraoral blocks vs. guided bone regeneration. Systematic review. *J Clin Exp Dent* 2023;15:e357–e365
5. Hameed MH, Gul M, Ghafoor R, et al. Vertical ridge gain with various bone augmentation techniques: a systematic review and meta-analysis. *J Prosthodont* 2019;28:421–427
6. Anitua E, Anitua B, Alkhraisat MH, et al. Dental implants survival after nasal floor elevation: a systematic review. *J Oral Implantol* 2022;48:595–603
7. Plonka AB, Urban IA, Wang HL, et al. Decision tree for vertical ridge augmentation. *Int J Periodontics Restor Dent* 2018;38:269–275
8. Yu SH, Wang HL. An updated decision tree for horizontal ridge augmentation: a narrative review. *Int J Periodontics Restor Dent* 2022;42:341–349
9. Garefis PN. Complete mandibular subperiosteal implants for edentulous mandibles. *J Prosthet Dent* 1978;39:670–677
10. Carnicero A, Peláez A, Restoy-Lozano A, et al. Improvement of an additively manufactured subperiosteal implant structure design by finite elements based topological optimization. *Sci Rep* 2021;11:15390
11. Cebrián Carretero JL, Del Castillo Pardo de Vera JL, Montesdeoca García N, et al. Virtual surgical planning and

customized subperiosteal titanium maxillary implant (CSTMI) for three dimensional reconstruction and dental implants of maxillary defects after oncological resection: case series. *J Clin Med* 2022;11:4594
12. Jehn P, Spalthoff S, Korn P, et al. Oral health-related quality of life in tumour patients treated with patient-specific dental implants. *Int J Oral Maxillofac Surg* 2020;49:1067–1072
13. Koç M. Artificial intelligence in geriatrics. *Turkish J Geriatr* 2023;26:352–360
14. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;11:887
15. Giansanti D. Precision Medicine 2.0: how digital health and AI are changing the game. *J Pers Med* 2023;13:1057
16. Abu Arqub S, Al-Moghrabi D, Allareddy V, et al. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. *Angle Orthod* 2024
17. Acar AH. Can natural language processing serve as a consultant in oral surgery? *J Stomatol Oral Maxillofac Surg* 2023;125:101724
18. Kilinc DD, Mansiz D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofac Orthop* 2024;165:546–555
19. Eggmann F, Weiger R, Zitzmann NU, et al. Implications of large language models such as ChatGPT for dental medicine. *J Esthet Restor Dent* 2023;35:1098–1102
20. Shoemaker SJ, Wolf MS, Brach C, et al. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns* 2014;96:395–403
21. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. *J Clin Med* 2024;13:735
22. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Res Square* 2023;28:rs.3.rs-2566942
23. Vaishya R, Misra A, Vaish A, et al. ChatGPT: Is this version good for healthcare and research? *Diabetes Metab Syndr Clin Res Rev* 2023;17:102744
24. Guven Y, Ozdemir OT, Kavan MY, et al. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. *Dental Traumatol* 2024;41:338–347
25. Bernard A, Langille M, Hughes S, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol* 2007;102:2070–2077
26. Daraz L, Morrow AS, Ponce OJ, et al. Readability of online health information: a meta-narrative systematic review. *Am J Med Qual* 2018;33:487–492
27. Flesch R. A new readability yardstick. *J Appl Psychol* 1948;32:221–233
28. Kincaid JP, Fishburne RP Jr, Rogers RL, et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Defense Technical Information Center* 1975;1:51
29. DISCERN online quality criteria for consumer health information. Accessed September 2, 2025. http://www.discern.org.uk
30. Hassona Y, Alqaisi D, AL-Haddad A, et al. How good is ChatGPT at answering patients' questions related to early detection of oral (mouth) cancer? *Oral Surg Oral Med Oral Pathol Oral Radiol* 2024;138:269–278
31. Sallam M, Barakat M, Sallam M, et al. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. *Cureus* 2023;15:e49373
32. Esmailpour H, Rasaie V, Babaee Hemmati Y, et al. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health* 2025;25:574
33. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak* 2024;24:211
34. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. *BMC Oral Health* 2025;25:871
35. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. *Cureus* 2025;17:e77292
36. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F, et al. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health* 2025;25:573
37. Çitir M. ChatGPT and oral cancer: a study on informational reliability. *BMC Oral Health* 2025;25:86
38. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg* 2023;124:101471
39. Sahin S, Erkmen B, Duymaz YK, et al. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. *Front Surg* 2024;11:1373843
40. Helvacioglu-Yigit D, Demirturk H, Ali K, et al. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2025;139:750–759

