JOURNAL OF DENTAL HEALTH & ORAL RESEARCH

e-ISSN 3050-9432

Journal of Dental Health and Oral Research

ATHENAEUM SCIENTIFIC PUBLISHERS

Research Article

Assessing the Usefulness and Quality of Artificial Intelligence-Generated Content on Bone Graft Materials in Dentistry Based on Patient Questions

Zeynep Hazan Yildiz[1*], Rumeysa Nur Kayaci[1], İrem Sezen[1], Ayşin İlayda Yildirim[1], Zeynep Cansu Güzel[1]

[1]Department of Periodontology, Gulhane Faculty of Dentistry, University of Health Sciences, Ankara, Turkey

***Correspondence author:** Zeynep Hazan Yildiz, Department of Periodontology, Gulhane Faculty of Dentistry, University of Health Sciences, Emrah Mah. Etlik, Kecioren, Ankara, 06010, Turkey; E-mail: zeynephazan.yildiz@sbu.edu.tr

Citation: Yildiz ZH, et al. Assessing the Usefulness and Quality of Artificial Intelligence-Generated Content on Bone Graft Materials in Dentistry Based on Patient Questions. J Dental Health Oral Res. 2025;6(3):1-10.

https://doi.org/10.46889/JDHOR.2025.6308

Received Date: 21-09-2025

Accepted Date: 08-10-2025

Published Date: 16-10-2025

Abstract

Background: This study evaluated the performance of three Artificial Intelligence (AI) conversational agents-ChatGPT-4, DeepSeek and Copilot-in providing clinically relevant information on bone graft materials in dentistry.

Methodology: A standardized set of questions related to bone graft types, indications, effectiveness, risks and patient comfort was posed to each model and these questions were derived from commonly asked patient inquiries. Responses were independently assessed by calibrated reviewers using validated tools: CLEAR criteria, a modified Global Quality Score (mGQS), a 5-point Likert scale for accuracy, a 4-point usefulness scale and readability metrics (Flesch Reading Ease and Flesch-Kincaid Grade Level).

Results: ChatGPT-4 outperformed Copilot in CLEAR scores ($p = 0.030$) and exceeded both DeepSeek and Copilot in mGQS ($p = 0.022$ and $p = 0.017$, respectively). However, no significant differences were observed in accuracy or readability ($p > 0.05$). ChatGPT-4 scored significantly lower in usefulness compared to DeepSeek and Copilot ($p < 0.05$). Negative correlations were found between mGQS and usefulness ($r = -0.807$) and between readability and grade level ($r = -0.938$).

Conclusion: ChatGPT-4 provided higher-quality and more comprehensive answers, but all models showed limitations in readability and usefulness. Improving the clarity and practical relevance of AI-generated content is essential to support dental education and enhance patient communication.

Keywords: Artificial Intelligence; ChatGPT-4; Copilot; DeepSeek; Graft; Patient Questions

Abbreviations

AI: Artificial Intelligence; Q and A: Question and Answer; mGQS: Modified Global Quality Score; Likert scale: A psychometric scale commonly used for measuring attitudes or responses; CLEAR: Completeness, Lack of false information, Evidence-based support, Appropriateness of formation and Relevance to the topic; FRE: Flesch Reading Ease; FKGL: Flesch-Kincaid Grade Level; IBM SPSS: International Business Machines Statistical Package for the Social Sciences

Introduction

Bone graft materials were initially developed as passive, biocompatible scaffolds aimed at providing structural support for bone regeneration [1]. With the advent of tissue engineering and regenerative medicine, these materials have undergone significant refinement, resulting in a wide range of graft types, each with distinct characteristics, advantages and limitations [2]. In the context of dentistry, alveolar bone loss represents a major clinical concern, most frequently arising from periodontal disease but also triggered by tooth extraction and other factors. These conditions can initiate a cascade of biological events that lead to considerable dimensional changes in the alveolar ridge [3,4]. Effective management of hard tissue loss is critical not only for maintaining long-term tooth stability and functional integrity but also for achieving optimal aesthetic outcomes, especially in the anterior maxillary region and for ensuring successful dental implant placement [5]. Consequently, bone graft materials have

been extensively studied and applied in dental practice.

Clinical indications for bone grafting range from localized intraosseous defects to complex full-arch rehabilitations. An ideal bone graft material must meet a wide range of clinical and biological criteria, including excellent biocompatibility, mechanical stability, favorable surface properties, optimal porosity and geometry and ease of handling during surgical procedures. Bone graft materials are generally classified into four main types: autografts, allografts, xenografts and alloplastic (synthetic) grafts [6]. Each type varies in its source, biological behavior and clinical application potential. In addition to their origin, bone graft materials are evaluated based on their osteogenic, osteoinductive and osteoconductive properties. An optimal graft should: contain viable osteogenic progenitor cells capable of forming new bone matrix, demonstrate osteoinductive potential by promoting the differentiation of mesenchymal stem cells into osteoblasts; and provide a robust osteoconductive scaffold that facilitates three-dimensional tissue ingrowth and vascularization [7].

The term Artificial Intelligence (AI), defined as the simulation of human cognitive functions by computer systems, was first introduced in 1956 and has since evolved into a transformative technological domain [8]. AI refers to machines or software systems designed to perform tasks that normally require human intelligence, such as learning, reasoning, problem-solving and decision-making. In recent years, rapid progress in computational power, algorithm design and big data analytics has significantly accelerated AI development. This evolution has catalyzed the integration of AI technologies into various sectors, including health sciences. In dentistry, AI holds increasing promises for enhancing diagnostic accuracy, optimizing treatment planning and supporting educational endeavors [9,10]. Concurrently, patients increasingly seek medical information online to address concerns regarding diagnoses, procedures, risks and expected outcomes. The rapid maturation of AI-driven conversational agents has further lowered barriers to access by providing prompt, on-demand explanations tailored to lay users, thereby potentially enhancing health literacy and supporting shared decision making. Nevertheless, the clinical value of such tools depends on the accuracy, clarity and contextual appropriateness of their outputs. Systematic evaluation of AI-generated content particularly for specialized topics such as dental bone graft materials is therefore essential to ensure that patient facing information is both reliable and comprehensible and to guide safe integration of these technologies into patient education and clinician-patient communication.

Recent research has explored the capabilities of advanced AI models-such as ChatGPT-4, DeepSeek and Copilot-in generating responses to a broad range of queries, including those in the fields of medicine and dentistry [11-13]. As reliance on AI-driven conversational agents grows, assessing the relevance, accuracy and clinical value of their responses to complex dental questions become imperative. Evaluating the ability of these systems to address topics such as bone graft materials, their biological integration and related clinical outcomes may enhance their utility as adjunctive tools in clinical education, decision-making and patient communication. This study aimed to evaluate the accuracy, quality, readability, reliability and usefulness of responses generated by AI-based conversational agents-specifically ChatGPT-4, DeepSeek and Copilot-regarding bone graft materials used in dentistry. By systematically analyzing the responses to clinically relevant questions, the study seeks to determine the potential of these AI models as educational and decision-support tools in clinical dental practice.

Material and Methods

This study was conducted in accordance with the Declaration of Helsinki and did not require approval from an ethics committee. Questions related to graft materials-including general information, clinical effectiveness, pain, patient comfort, treatment process, clinical indications, potential risks and limitations- were posed to AI-based conversational chatbots, specifically ChatGPT-4, DeepSeek and Copilot.

Sample Size Calculation

In this study, the "G*Power 3.1.9.2" program was used [14]. At a 95% confidence level ($\alpha = 0.05$), the standardized effect size was taken as 0.46 from a similar study (Table 4, three Ai comparison for FRE) and with a theoretical power of 0.80 the minimum sample size was calculated as 12 [15].

Two researchers refined the questions through the removal of duplicates and semantic overlaps and standardized for grammar and clarity. Access to each of the artificial intelligence systems was provided on July 21, 2025 and a new account was created for this study. Before starting the question-and-answer (Q and A), all search history and cookies on the computer were cleared. To

minimize the influence of previously given answers, a new conversation window was opened for each question asked and responses were recorded for later analysis. To ensure the specificity and relevance of the generated responses, the following prompt was entered: Please answer the frequently asked questions about dental graft materials. All responses were independently evaluated by two other authors using a set of validated scoring tools. Accuracy was assessed using a five-point Likert scale, where 1 indicated completely incorrect responses and 5 indicated entirely correct responses, with intermediate scores reflecting varying proportions of correct and incorrect information [16].

Three independent periodontology specialists, with no prior knowledge of the question formulation team, were selected to evaluate all AI responses for the CLEAR criteria, modified Global Quality Score (GQS), accuracy and usefulness. Each expert has at least 10 years of experience in clinical periodontology practice and possesses advanced knowledge and expertise in the use of various bone graft materials and regenerative techniques in periodontal therapy. All of them have advanced training in the selection and application of grafting materials, regularly employ these biomaterials in clinical practice and have contributed to continuing education courses in periodontal regeneration and grafting procedures.

For investigating whether repeated values of the same variables are similar, the Kappa test was applied for categorical variables as a test retest measure. The inter-observer analysis revealed a Kappa statistic of 0.850 (95% CI: 0.827-0.870) for inter observer agreement. It was determined that the reliability between measurements was statistically significant and demonstrated a high level of agreement ($p < 0.05$).

A modified version of the Global Quality Score (GQS) was used to assign scores based on the "context" and "content" of the responses:

- Score 5 (Strongly Agree): The answer is correct and the content is comprehensive
- Score 4 (Agree): The answer is correct and most of the content is correct, but it lacks information or contains incorrect information
- Score 3 (Neutral): The answer is somewhat correct, but details are primarily incorrect, missing or irrelevant
- Score 2 (Disagree): The answer is incorrect, but the content includes some correct elements
- Score 1 (Strongly Disagree): The answer and the entire content are incorrect or irrelevant [17]

Reliability was evaluated using the CLEAR criteria, which assess five domains: Completeness of content, Lack of false information, Evidence-based support, Appropriateness of information and Relevance to the topic. 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 5-11 indicated low quality, 12-18 moderate quality and 19-25 high quality[18].

The readability of AI-generated responses was evaluated using two validated metrics: the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) scores, calculated via online tools (Readable Pro: https://app.readable.com/text/). FRE assigns a score between 0 and 100, with higher values indicating easier readability; scores above 60 are generally considered acceptable, while scores below 30 suggest high difficulty. FKGL estimates the U.S. school grade level required to comprehend a text, with lower values representing greater accessibility. Both indices are based on average sentence length and word complexity (syllables per word). Health information materials are recommended to be written at or below an 8th-grade reading level to ensure optimal public comprehension [19,20].

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) [21].

Statistical Analysis

Descriptive statistics, including frequency, percentage, mean, standard deviation, median and interquartile range (25^{th} and 75^{th} percentiles), were calculated. The normality of data distribution was assessed using the Shapiro-Wilk test. For comparisons involving three or more independent groups with non-normally distributed variables, the Kruskal-Wallis test was applied. When

statistically significant differences were observed, post hoc pairwise comparisons were conducted using the Bonferroni correction. For categorical variables with expected frequencies below five, Fisher's exact test was employed. Associations between non-normally distributed continuous variables were examined using Spearman's rank correlation coefficients. All statistical analyses were performed using IBM SPSS Statistics version 27.

Results

The distribution of response characteristics across different AI models is presented in Table 1. Group comparisons were conducted using the Kruskal-Wallis test, which revealed statistically significant differences among AI types in terms of CLEAR, mGQS and Usefulness ($p < 0.05$). Bonferroni adjusted pairwise comparisons showed a significant difference in CLEAR scores between the ChatGPT and CoPilot groups ($p = 0.030$), with ChatGPT achieving higher scores. For CLEAR scores, no significant differences were detected between DeepSeek and ChatGPT or between DeepSeek and CoPilot ($p = 0.227$ and $p = 1.000$, respectively). The CLEAR scores indicated that ChatGPT-4 and DeepSeek performances were high compared to the moderate quality of Copilot. Regarding mGQS, ChatGPT-4 demonstrated significantly higher scores than both CoPilot ($p = 0.017$) and DeepSeek ($p = 0.022$), while no significant difference was found between CoPilot and DeepSeek ($p = 0.922$). In terms of Usefulness, significant differences were observed between ChatGPT-4 and both CoPilot ($p = 0.019$) and DeepSeek ($p = 0.023$); however, CoPilot and DeepSeek scored higher than ChatGPT-4. No significant difference was found between DeepSeek and CoPilot ($p = 0.939$).

No statistically significant differences were observed among the AI models in terms of Accuracy, Flesch Reading Ease Score or Flesch-Kincaid Grade Level ($p > 0.05$). The distribution of mGQS, Accuracy and Usefulness scores by AI model is summarized in Table 2. Fisher's exact test was used to evaluate the associations among these variables. The results indicated no statistically significant associations between AI type and mGQS, Accuracy or Usefulness ($p > 0.05$). Pearson and Spearman correlation analyses were conducted to investigate the relationships among evaluation metrics within the ChatGPT group (Table 3). A statistically significant, strong negative correlation was found between CLEAR and Usefulness ($r = -0.837$, $p < 0.001$) and between mGQS and Usefulness ($r = -0.807$, $p < 0.001$). Additionally, a strong, negative correlation was observed between the Flesch Reading Ease Score and the Flesch-Kincaid Grade Level ($r = -0.938$, $p < 0.001$).

Spearman correlation analyses were performed for the DeepSeek group (Table 4). The results demonstrated a strong, positive correlation between CLEAR and mGQS ($r = 0.900$, $p < 0.001$) and a strong, negative correlation between CLEAR and Usefulness ($r = -0.939$, $p < 0.001$). mGQS also showed a statistically significant, strong, positive correlation with Accuracy ($r = 0.839$, $p < 0.001$) and a strong, negative correlation with Usefulness ($r = -0.948$, $p < 0.001$). Furthermore, a strong, negative correlation was found between the Flesch Reading Ease Score and the Flesch-Kincaid Grade Level ($r = -0.860$, $p < 0.001$). In the CoPilot group (Table 5), Pearson correlation analyses revealed statistically significant, strong positive correlations between CLEAR and mGQS ($r = 0.867$, $p < 0.001$) and CLEAR and Accuracy ($r = 0.751$, $p < 0.001$), as well as a strong, negative correlation between CLEAR and Usefulness ($r = -0.905$, $p < 0.001$). Similarly, mGQS was strongly and positively correlated with Accuracy ($r = 0.866$, $p < 0.001$) and strongly and negatively correlated with Usefulness ($r = -0.941$, $p < 0.001$). A statistically significant, strong, negative correlation was also observed between Accuracy and Usefulness ($r = -0.754$, $p < 0.001$). Additionally, the Flesch Reading Ease Score was strongly and negatively correlated with the Flesch-Kincaid Grade Level ($r = -0.854$, $p < 0.001$).

	ChatGPT-4		**DeepSeek**		**CoPilot**			
	Mean ± S.D.	**Median (%25-%75P.)**	**Mean. ± S.D.**	**Median (%25-%75P.)**	**Mean. ± S.D.**	**Median (%25-%75P.)**	**Test Statistic**	**p**
CLEAR	22.4 ± 3.03[b]	24(20.5-25)	19.35 ± 5.53	20(15-24.5)	18.1 ± 5.88	20(13-24)	6.95	0.031*
mGQS	4.2 ± 0.62[a,b]	4(4-5)	3.55 ± 0.94	4(3-4)	3.5 ± 1.05	4(3-4)	7.27	0.026*
Accuracy	3.6 ± 0.75	3(3-4)	3.15 ± 0.93	3(2.5-4)	3 ± 0.97	3(2-4)	3.749	0.153
Usefulness	1.65 ± 0.67 [a,b]	2(1-2)	2.35 ± 1.04	2(2-3)	2.4 ± 1.1	2(2-3)	7.129	0.028*

Flesch Reading Ease Score	43.03 ± 23.62	45.1(29.3-59.7)	44.77 ± 18.29	47(31.85-54.9)	41.5 ± 20.29	42.65(30.65-58.75)	0.14	0.932
Flesch Reading Grade Level	10.33 ± 4.11	9.25(7.65-12.65)	11.19 ± 2.93	11.1(9.85-12.2)	11.64 ± 2.99	11.15(9.2-13.5)	2.626	0.269
*: Statistically significant differences (p<0.05) a p<0.05 ChatGPT-4 versus Deepseek group, b p<0.05 ChatGPT-4 versus CoPilot group. SD: Standard Deviation.								

Table 1: Distribution and comparison of response characteristics by AI type.

		ChatGPT-4			DeepSeek			CoPilot				
		n	**%**	**%Ai.**	**n**	**%**	**%Ai.**	**n**	**%**	**%Ai.**	**Test Statistic**	**p**
mGQS	Poor	0	0	0	1	50	5	1	50	5	8.389	0.332
	Generally poor	0	0	0	1	33.3	5	2	66.7	10		
	Moderate	2	14.3	10	6	42.9	30	6	42.9	30		
	Good	12	40	60	10	33.3	50	8	26.7	40		
	Excellent	6	54.5	30	2	18.2	10	3	27.3	15		
Accuracy	1	0	0	0	1	50	5	1	50	5	12.129	0.084
	2	0	0	0	4	44.4	20	5	55.6	25		
	3	11	44	55	6	24	30	8	32	40		
	4	6	30	30	9	45	45	5	25	25		
	5	3	75	15	0	0	0	1	25	5		
Usefulness	Very useful	9	52.9	45	4	23.5	20	4	23.5	20	7.569	0.05
	Useful	9	36	45	8	32	40	8	32	40		
	Partially useful	2	15.4	10	6	46.2	30	5	38.5	25		
	Not useful	0	0	0	2	40	10	3	60	15		
%: Row percentage. %Ai: Column percentage for AI types												

Table 2: Distribution and relationships of mGQS, accuracy and usefulness scores by AI type.

		mGQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level
CLEAR	r	0.629	0.266	-0.837	-0.044	-0.048
	p	0.003	0.256	<0.001*	0.855	0.84
mGQS	r		0.624	-0.807	0.311	-0.332
	p		0.003	<0.001*	0.182	0.153
Accuracy	r			-0.35	-0.035	0.034
	p			0.13	0.883	0.886
Usefulness	r				-0.195	0.267
	p				0.411	0.255
Flesch Reading Ease Score	r					-0.938
	p					<0.001*
*: Statistically significant differences (p<0.05). r: Correlation Coefficient						

Table 3: Correlations among measurements for ChatGPT-4.

		mGQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level
CLEAR	r	0.9	0.695	-0.939	-0.383	0.113
	p	<0.001*	0.001	<0.001*	0.095	0.636
mGQS	r		0.839	-0.948	-0.202	-0.095
	p		<0.001*	<0.001*	0.394	0.69
Accuracy	r			-0.698	-0.016	-0.263
	p			0.001	0.946	0.262
Usefulness	r				0.242	0.044
	p				0.303	0.853
Flesch Reading Ease Score	r					-0.86
	p					<0.001*
*: Statistically significant differences (p<0.05), r: Correlation Coefficient						

Table 4: Correlations among measurements for DeepSeek.

		mGQS	Accuracy	Usefulness	Flesch Reading Ease Score	Flesch Reading Grade Level
CLEAR	r	0.867	0.751	-0.905	-0.054	-0.004
	p	<0.001*	<0.001*	<0.001*	0.82	0.987
mGQS	r		0.866	-0.941	-0.167	0.102
	p		<0.001*	<0.001*	0.481	0.668
Accuracy	r			-0.754	-0.186	0.101
	p			<0.001*	0.431	0.673
Usefulness	r				0.177	-0.091
	p				0.456	0.704
Flesch Reading Ease Score	r					-0.854
	p					<0.001*

Table 5: Correlations among measurements for Copilot.

Discussion

This study evaluated the performance of three prominent AI conversational agents-ChatGPT-4, DeepSeek and Copilot-in providing responses related to bone graft materials in dentistry. Given the clinical complexity and biological nuances involved in bone grafting procedures, ensuring accurate, reliable and comprehensible information is critical for both clinicians and patients. Our findings reveal notable differences among these AI platforms in terms of response quality, usefulness and readability, reflecting varying capabilities in addressing specialized dental topics. While AI technologies hold considerable promise as adjunctive educational and decision-support tools, the present results underscore that current models still face challenges in delivering consistently clear and clinically actionable content. These observations highlight the ongoing need to refine AI outputs, particularly with respect to linguistic accessibility and evidence-based accuracy, to better support informed decision-making and optimize patient communication in dental practice. The CLEAR tool provides a concise yet effective framework for evaluating the quality of AI-generated health information. In this study, it revealed that ChatGPT produced significantly higher CLEAR scores than CoPilot, with no significant differences between DeepSeek and the other models. The CLEAR scores indicated that Chatgpt-4 and DeepSeek performances were high quality compared to the moderate quality of Copilot. Similarly, Sallam, et al., reported that ChatGPT-4 delivered superior content quality, achieving "Excellent" ratings across key domains, while other models were rated as "Above average" [22]. A recent study evaluating ChatGPT-3.5 and ChatGPT-4 using the CLEAR tool demonstrated that both models effectively deliver primary prevention information for common musculoskeletal disorders, with ChatGPT-4 significantly outperforming its predecessor in completeness, appropriateness and relevance [23]. These findings highlighted the growing potential of advanced AI models, particularly ChatGPT-4, in supporting public health education and preventive strategies.

In the study by Yagcı, et al., it was initially highlighted that widely used AI chatbots in implant dentistry demonstrated comparable accuracy and reliability, though their overall performance was deemed suboptimal for clinical decision-making [24]. However, our current findings reveal significant differences among AI platforms, particularly in terms of GQS and usefulness. ChatGPT-4 outperformed both DeepSeek and Copilot, showing statistically higher scores in these domains ($p < 0.05$), whereas no significant difference was found between DeepSeek and Copilot. Specifically, ChatGPT-4 achieved the highest mean scores in completeness, accuracy, clarity and relevance, suggesting notable advantages for clinical support in implant dentistry. Related to pediatric dental trauma, Gökcek Taraç and Nale reported no statistically significant differences between ChatGPT-3.5 and Google Gemini in answering parental questions, indicating comparable reliability between platforms. However, Gemini's slightly higher mean scores, although not significant, may reflect a tendency to deliver more actionable responses [25]. In the field of orthodontics, Dursun and Geçer evaluated ChatGPT-3.5, ChatGPT-4, Google Gemini and Microsoft Copilot in responding to frequently asked questions about clear aligners. They found all models provided generally accurate, moderately reliable and moderate to good quality answers. However, they emphasized that readability was low overall, with Gemini producing the most readable responses, although still below recommended standards for patient education [26]. Similarly, Tokgöz Kaplan and Cankar compared ChatGPT and Gemini for dental avulsion questions based on IADT guidelines. Their results showed Gemini provided significantly more accurate answers overall, while ChatGPT performed better on open-ended and true/false questions [27]. Regarding oral cancer education, Hassona, et al., found that although ChatGPT offers moderately high-quality responses about early detection, its practical usefulness was limited [11]. Only a minority of answers were fully useful and most exceeded recommended readability levels, highlighting the need to improve clarity and accessibility in AI-generated patient content. Taken together, these studies suggest that while AI tools may demonstrate similar baseline reliability, their clinical applicability and user-friendliness vary significantly across platforms and medical contexts. Therefore, careful platform selection and efforts to improve readability and practical usefulness are essential for optimizing AI's role in patient education and clinical decision support.

A study by Ekmekci and Durmazpinar revealed substantial disparities among AI platforms in addressing questions related to regenerative endodontic procedures [28]. ChatGPT-4, equipped with a PDF plugin, outperformed other tools by providing 98.1% correct responses and no incorrect answers. In contrast, Gemini showed limited reliability with a high rate of insufficient and incorrect responses, as well as notable inconsistency throughout the day. These findings underscore the importance of stability and the integration of evidence-based guidance in AI systems used in clinical dentistry. Consistent with these results, this study found that ChatGPT-4 significantly outperformed both DeepSeek and Copilot in Global Quality Score (GQS) and perceived usefulness ($p < 0.05$), while no significant difference was observed between DeepSeek and Copilot in these metrics. Additionally, ChatGPT-4 achieved higher scores in clarity, completeness and accuracy, suggesting a more robust and reliable content generation capability. Despite these advantages, even the best-performing models demonstrated moderate Flesch Reading Ease scores and elevated Flesch-Kincaid Grade Levels, indicating that readability remains a key area for improvement to better serve diverse users, including patients and practitioners.

The study by Helvacioglu-Yigit, et al., highlighted that although AI chatbots generate scientifically accurate patient education content, their readability generally falls below recommended standards, limiting accessibility [19]. Similarly, Esmailpour, et al., emphasized that despite producing high-quality and clinically relevant responses, AI-generated language complexity often reduces practical utility for patients [15]. Jacobs, et al., supported this concern, noting that ChatGPT's responses, while accurate, have readability levels exceeding ideal patient comprehension and lack proper citations, raising transparency issues [29]. Hunter, et al., showed that tailoring ChatGPT outputs to an eighth-grade reading level can improve accessibility, though essential authorship and attribution remain absent [30]. These findings align with our results, where ChatGPT-4, DeepSeek and Copilot all produced responses with low readability, indicating that AI-generated content, though accurate and informative, remains challenging for many patients to understand.

Güven, et al., evaluated ChatGPT 3.5, ChatGPT 4.0 and Google Gemini for traumatic dental injury questions, finding all responses difficult to read; ChatGPT 3.5 had lower accuracy and understandability, while ChatGPT 4.0 and Gemini delivered more accurate and comprehensive answers [31]. However, none can replace dentists for diagnosis and treatment in complex cases. Therefore, improving linguistic accessibility through health literacy frameworks or simplification algorithms is crucial to

enhance AI's role in patient education and clinical communication. These findings align with the readability results of our current study, where ChatGPT-4, DeepSeek and Copilot all produced outputs with Flesch Reading Ease scores below 45 and Flesch-Kincaid Grade Levels above 10, reinforcing the notion that AI-generated content, while increasingly accurate and informative, remains difficult for the average patient to understand. Therefore, optimizing the linguistic accessibility of chatbot responses through integration with health literacy standards or algorithmic simplification tools is essential to enhance their utility in clinical communication and patient education.

Despite providing valuable insights, this study has several limitations. First, the evaluation was limited to responses generated by only three AI platforms, which may not represent the full spectrum of available or emerging AI technologies in dentistry. Second, the assessment focused on predefined clinical questions related to bone graft materials, potentially limiting the generalizability of findings to other dental topics or real-time clinical scenarios. Third, although multiple evaluators were involved to reduce subjective bias, the scoring of response quality, accuracy and readability inherently contains an element of subjectivity. Additionally, the readability metrics used primarily assess linguistic complexity but may not fully capture patient comprehension or health literacy variations. Finally, AI models continuously evolve through updates and the results reflect the performance of these platforms at a specific point in time, which may change with future improvements. Further longitudinal and broader investigations are needed to validate and extend these findings.

Conclusion

This study provides evidence that ChatGPT-4 outperforms DeepSeek and Copilot in delivering clinically relevant, accurate and high-quality responses regarding bone graft materials in dentistry. Despite these advantages, the readability of all AI-generated outputs remained below recommended thresholds, which may limit their effectiveness in patient-facing communication. The observed correlations between quality indicators underscore the potential of large language models as supportive tools in clinical decision making and dental education. Nevertheless, enhancing linguistic accessibility and ensuring alignment with evidence-based standards remain critical for their safe and effective integration into dental practice.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Financial Disclosure

No funding.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial or non-profit sectors.

Author Contributions

Z.H.Y.: Supervision, Project administration, Methodology, Investigation, Data curation, Formal analysis, Conceptualization, Writing - original draft, Writing - review and editing. R.N.K.: Investigation, Methodology, Data curation, Formal analysis, Writing - original draft, Writing - review and editing. İ.S.: Investigation, Methodology, Formal analysis, Data curation, Writing - review and editing. A.İ.Y: Investigation, Conceptualization, Formal analysis, Writing - review and editing. Z.C.G: Data curation, Conceptualization, Methodology, Writing - review and editing.

References

1. Langer R, Tirrell DA. Designing materials for biology and medicine. Nature. 2004;428:487-92.
2. Giannoudis PV, Dinopoulos H, Tsiridis E. Bone substitutes: An update. Injury. 2005;36(Suppl 3):S20-7.
3. Araújo MG, Lindhe J. Dimensional ridge alterations following tooth extraction. An experimental study in the dog. J Clin Periodontol. 2005;32:212-8.
4. Tan WL, Wong TL, Wong MC, Lang NP. A systematic review of post-extractional alveolar hard and soft tissue dimensional changes in humans. Clin Oral Implants Res. 2012;23(Suppl 5):1-21.
5. Chappuis V, Engel O, Reyes M, Shahim K, Nolte LP, Buser D. Ridge alterations post-extraction in the esthetic zone: A 3D analysis with CBCT. J Dent Res. 2013;92(12 Suppl):195S-201S.

6. Miron RJ. Optimized bone grafting. Periodontol 2000. 2024;94:143-60.
7. Miron RJ, Hedbom E, Saulacic N, Zhang Y, Sculean A, Bosshardt DD, et al. Osteogenic potential of autogenous bone grafts harvested with four different surgical techniques. J Dent Res. 2011;90:1428-33.
8. Thurzo A, Urbanová W, Novák B, Czako L, Siebert T, Stano P, et al. Where is the artificial intelligence applied in dentistry? Systematic review and literature analysis. Healthcare (Basel). 2022;10(7):1269.
9. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230-43.
10. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: Chances and challenges. J Dent Res. 2020;99:769-74.
11. Hassona Y, Alqaisi D, Al-Haddad A, Georgakopoulou EA, Malamos D, Alrashdan MS, 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;138:269-78.
12. Dincer HA, Dogu D. Evaluating artificial intelligence in patient education: DeepSeek-V3 versus ChatGPT-4o in answering common questions on laparoscopic cholecystectomy. ANZ J Surg. 2025;95.
13. 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 (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.
14. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009;41:1149-60.
15. Esmailpour H, Rasaie V, Babaee Hemmati Y, Falahchai M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. BMC Oral Health. 2025;25:574.
16. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med. 2024;13(7):735.
17. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102:2070-7.
18. Sallam M, Barakat M, Sallam M. 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.
19. Helvacioglu-Yigit D, Demirturk H, Ali K, Tamimi D, Koenig L, Almashraqi A. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. Oral Surg Oral Med Oral Pathol Oral Radiol. 2025;139:750-9.
20. National Library of Medicine. How to write easy-to-read health materials. Updated November 2020. Bethesda (MD): U.S. National Library of Medicine. 2022.
21. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29:721-32.
22. Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, Bard, ChatGPT-3.5 and humans in clinical chemistry multiple-choice questions. Adv Med Educ Pract. 2024;15:857-71.
23. Yilmaz Muluk S, Olcucu N. The role of artificial intelligence in the primary prevention of common musculoskeletal diseases. Cureus. 2024;16:e65372.
24. Yagcı F, Eraslan R, Albayrak H, İpekten F. Accuracy and reliability of artificial intelligence chatbots as public information sources in implant dentistry. Int J Oral Maxillofac Implants. 2025;40:1-23.
25. Gökcek Taraç M, Nale T. Artificial intelligence in pediatric dental trauma: do artificial intelligence chatbots address parental concerns effectively? BMC Oral Health. 2025;25:736.
26. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024;24:211.
27. Tokgöz Kaplan T, Cankar M. Evidence-based potential of generative artificial intelligence large language models on dental avulsion: ChatGPT versus Gemini. Dent Traumatol. 2025;41:178-86.
28. Ekmekci E, Durmazpinar PM. Evaluation of different artificial intelligence applications in responding to regenerative endodontic procedures. BMC Oral Health. 2025;25:53.
29. 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;82:1239-45.
30. Hunter N, Allen D, Xiao D, Cox M, Jain K. Patient education resources for oral mucositis: A Google search and ChatGPT analysis. Eur Arch Otorhinolaryngol. 2025;282:1609-18.
31. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: a comparative study. Dent Traumatol. 2025;41:338-47.

Supplementary Files

The queries that were asked to ChatGPT-4, Deepseek, and Copilot

1. What is a dental graft?
2. Why is this procedure necessary?
3. What is the source of the graft material?
4. How is the dental graft procedure performed?
5. Which graft material is most suitable for my condition?
6. Will I experience pain during the grafting procedure?
7. How long does the grafting procedure take?
8. How can the success of the graft be determined?
9. Is it religiously permissible if the graft is derived from animals?
10. Is the graft permanent, or will it be resorbed over time?
11. What are the risks associated with dental grafting?
12. Can the graft become infected?
13. Is there a risk of graft rejection by the body?
14. Is there a possibility that the graft may fail to integrate?
15. How long does it take for the graft to ossify?
16. Is it mandatory to undergo grafting?
17. Is implant placement possible without grafting?
18. Is the grafting procedure painful?
19. What post-operative care is required following graft placement?
20. What is the expected healing period after the grafting procedure?