Cumhuriyet Dental Journal, 28(3): 390-396, 2025
DOI: https://doi.org/10.7126/cumudj.1673333

Cumhuriyet Dental Journal

| cdj.cumhuriyet.edu.tr | Founded: 1998 Available online, ISSN: 1302-5805 Publisher: Sivas Cumhuriyet Üniversitesi

Validity and Reliability of Responses to Periodontology Questions by 4 Different Artificial Intelligence Chatbots as Public Information Sources

Mahmure Ayşe Tayman[1,a,*]

[1]*Department of Periodontology, Faculty of Dentistry, Ankara Yıldırım Beyazıt University, Ankara, Türkiye.*
**Corresponding author*

Research Article

History

Received: 10/04/2025
Accepted: 16/06/2025

ABSTRACT

Objectives: To assess and check the validity and reliability of the answers given by ChatGpt-4o mini, Deepseek, Copilot and Gemini 1.5 flash daily chatbots to often seeked queries in the area of periodontology.

Materials and Methods: Questions were selected from the most frequently asked patient questions by a periodontologist. Each question was asked to the chatbots three times. The answers (n = 240) were independently evaluated by two periodontologists on a Likert scale (5 = violently agree; 4 = agree; 3 = neutral; 2 = disagree; 1 = violently disagree). Disputes in scoring were removed through evidence-based negotiations. In evaluating the validity of the answers: Low threshold was determined as a score ≥4 for whole three answers; high threshold was determined as a score 5 for whole three answers. Fisher's exact test was performed to compare the validity of the answers among the chatbots. Cronbach's alpha was computed to evaluate the consistency and reliability of recurrent answers for each chatbot.

Results: All four chatbots answered the questions. In the low-threshold validity test, ChatGpt had 100%, Deepseek and Copilot had 95%, Gemini had 65%. Gemini was significantly different from the others ($p < 0.05$). In the high-threshold validity test, ChatGpt had 80%, Deepseek had 75%, Copilot and Gemini were significantly lower at 5%. While there was no significant difference between ChatGpt and Deepseek ($p > 0.05$), both were significantly higher than Copilot and Gemini ($p < 0.05$). All four chatbots reached an acceptable level of reliability (Cronbach's alpha >0.7).

Conclusions: ChatGpt and Deepseek provided more reliable information on periodontology-related topics than Copilot and Gemini.

Keywords: Chatbots, artificial intelligence, deep learning, dentistry, periodontology.

Kamuya Açık Bilgi Kaynağı Olan 4 Farklı Yapay Zeka Sohbet Robotunun Periodontoloji Sorularına Cevaplarının Geçerliliği ve Güvenilirliği

Araştırma Makalesi

Süreç

Geliş: 10/04/2025
Kabul: 16/06/2025

ÖZ

Amaç: Periodontoloji alanında sık sorulan sorulara ChatGpt-4o mini, Deepseek, Copilot ve Gemini 1.5 flash günlük sohbet robotları tarafından verilen cevapların geçerliliğini ve güvenilirliğini değerlendirmek ve karşılaştırmaktır.

Gereç ve Yöntemler: Sorular uzman bir periodontolog tarafından en çok sorulan hasta soruları arasından seçildi. Her soru, sohbet robotlarına üçer kez soruldu. Cevaplar (n = 240), iki periodontolog tarafından 5 puanlık Likert ölçeğiyle (5 = kesinlikle katılıyorum; 4 = katılıyorum; 3 = nötr; 2 = katılmıyorum; 1 = kesinlikle katılmıyorum) bağımsız olarak değerlendirildi. Puanlamadaki anlaşmazlıklar, kanıta dayalı tartışmalar yoluyla çözüldü. Yanıtların geçerliliği değerlendirilirken: Düşük eşik, üç cevabın tümü için puan ≥4 olarak belirlenirken; yüksek eşik, üç cevabın tümü için puan 5 olarak belirlendi. Sohbet robotları arasındaki cevapların geçerliliğini karşılaştırmak için Fisher'ın kesin testi yapıldı. Her sohbet robotu için tekrarlanan cevapların tutarlılığını ve güvenilirliği değerlendirmek için Cronbach'ın alfa değeri hesaplandı.

Bulgular: Dört sohbet robotu da tüm sorulara cevap verdi. Düşük eşikli geçerlilik testinde ChatGpt %100, Deepseek ve Copilot %95, Gemini %65 geçerliliğe sahipti. Gemini diğerlerinden anlamlı farklıydı ($p < 0.05$). Yüksek eşikli geçerlilik testinde ChatGpt %80, Deepseek %75'di, Copilot ve Gemini %5 olarak önemli ölçüde düşüktü. ChatGpt ve Deepseek arasında anlamlı fark yokken ($p > 0.05$) her ikisi de Copilot ve Geminiden önemli ölçüde daha yüksekti ($p < 0.05$). Dört sohbet robotu da kabul edilebilir bir güvenilirlik düzeyine ulaştı (Cronbach'ın alfa değeri >0,7).

Sonuçlar: ChatGpt ve Deepseek, Copilot ve Gemini'ye göre periodontoloji ile ilgili konularda daha güvenilir bilgiler sağladı.

***Anahtar Kelimeler*:** Sohbet robotu, yapay zekâ, derin öğrenme, diş hekimliği, periodontoloji.

[a] *ayseatay06@hotmail.com* *0000-0001-8924-6725*

How to Cite: Tayman MA. (2025) Validity and Reliability of Responses to Periodontology Questions by 4 Different Artificial Intelligence Chatbots as Public Information Sources, Cumhuriyet Dental Journal, 28(3): 390-396.

Introduction

In the digital communication era, information can be easily disseminated to large audiences in a short time. Artificial intelligence (AI), which imitates human intelligence, solves problems by analyzing them through

deep learning, and these abilities have been tested many times.[1] Chatbots with deep learning algorithms can get to know information from data sources by imitating human neural networks and can provide focused answers to personalized questions.[2,3] Chatbots provide information on a wide range of topics in many areas. In the field of health, they inform and guide patients by answering their questions about treatments, side effects, prognosis, results and complications.[4,5] While the person's awareness of their diseases and treatments increases, it is important that this easily accessible information is accurate, evidence-based, valid and reliable. Thus, patients' expectations become more realistic and chatbots can be of critical importance in many issues such as approving treatment, making informed decisions, and complying with recommendations.

Designed and developed by OpenAI in 2018 and later, the ChatGpt series was designed with large language models and natural language processing systems to establish text-based interactions, answer questions, and generate text. Gpt-3 and later Gpt-4 models were developed.[6] Deepseek, designed and developed in China, was designed to be used in tasks such as natural language processing and text understanding.[7] Copilot was designed by Microsoft and developed specifically to help users in Microsoft 365 (Office) products. Gemini, on the other hand, was designed by Google, focused on multiple models, was developed for use in both corporate and consumer applications, and the latest Gemini 2.0 flash was produced.[8] Chatbots are becoming increasingly smarter and more useful with the advancement of technology, and it is expected to take place in every area of our lives in the future.

Periodontology is the field of dentistry that studies the health, diseases, and treatment of the tissues surrounding and supporting the teeth, such as gingival disease, gingival recession, gingival hyperplasia, periodontal surgery, implantology, and aesthetic dentistry. The performance of AI-based chatbots as publicly available information supplies in different disciplines of medicine and dentistry has been evaluated.[9-11] The purpose of this study is to assess and check the validity and reliability of the answers given by Gpt 4o mini, Deepseek, Copilot, and Gemini 1.5 flash daily help chatbot systems to frequently asked questions in the field of periodontology.

Materials and Methods

20 regularly asked questions in the scope of periodontology including periodontal disease, dental calculus, gingival recession, oral hygiene and implant were asked 3 times to 4 different chatbots. These chatbots were ChatGpt 4o mini, Deepseek, Copilot and Gemini 1.5 flash daily help. The chatbots were accessed using the addresses https://chatgpt.com/?model=gpt-4o-mini, https://deepseek-r1.com,https://copilot.microsoft.com, https://deepmind.google/technologies/gemini/flash/respectively. The questions were selected from the most frequently asked questions during a patient consultation by a periodontologist (M.A.T.). On the same day (30.01.2025), a new chat was generated to each question and asked. The questions cover various areas such as terminology, diagnosis, treatment, procedure, outcome, prognosis, treatment risks, side effects and prevention. The questions are given in Table 1.

The answers were scored by two periodontologists using a 5-score Likert scale, taking into account the context (correctness of the answer) and content (incompleteness and correctness of the information).[12]

Accordingly: 5 points; “I strongly agree” because the answers are true and the meaning is detailed. 4 points; the response is true and mainly the meaning is true, however there is a lack of information or it includes false information, so I “agree”. 3 points; the answer is partially true, but the details are mostly false, incomplete or unrelated, so I “undecided”. 2 points; the answer is false, but the content contains some true components, so I “disagree”. 1 point; “I strongly disagree” because the response and the whole meaning are false or unrelated.

The different scores in the independently completed scores were determined to through evidence-based negotiations and at last prepared by giving a single score for all answers.

Statistical Analysis

Statistical Package for Social Sciences analysis program (SPSS Inc, Chicago, IL, USA, version 20.0) was implemented for the variables. Shapiro-Wilk test was used to define whether the word count information of the groups were not normally distributed. Fisher's Exact test was used in the analysis of the categorical variables. Man Whitney U test was used to compare non-normally distributed and ordinal data of two independent groups, and Kruskal vallis test was used to compare more than two groups. Non-normally distributed data and ordinal variables were presented as median and interquartile range (IQR) and min-max. The results of categorical variables were gived as frequency. P value <0.05 was defined as significant.

To evaluate the consistency of assessments made by the two periodontists, inter-rater reliability analysis was performed by calculating Cohen’s Kappa coefficient.

Validity

The degree to which the chatbot correctly captured the answer to a question. Scores were categorized as “valid” and “invalid.” Two validity tests were used:

Low threshold test: The value of 4 was the threshold. If every 3 answers got ≥4, the chatbot’s response was regarded as valid. If any answer got <4, the chatbot’s response was noted invalid.

High threshold test: The value of 5 was the threshold. If every 3 answers got 5 points, the chatbot’s response was regarded as valid. If any answer scored <5, it was noted invalid. Fisher’s exact test was performed to check the validity of answers among chatbots. The significance level was set at $p < 0.05$.

Reliability (consistency)

The scale to which a chatbot gives closer answers when used again and again under similar situations. The

5-score Likert scale determined to the responses given to the questions when repeated three times were analyzed. To assess the reliability of the answers, Cronbach's alpha was evaluated for all 3 answers in all 20 questions. Cronbach's alpha measures the level of consistency on a normalized measurement between 0 and 1. (0 = no consistency, 1 = perfect consistency). A superior alpha coefficient indicates that the chatbot answers the similar question consistently, and the scale is reliable. An inferior alpha coefficient indicates that the chatbot does not answer the similar question consistently, and the scale is not so great reliable. A Cronbach's alpha coefficient of ≥ 0.70 defines suitable reliability in medical and health-associated studies.[13]

Table 1. Questions asked to artificial intelligence chatbots

1	What is periodontal disease?
2	What are the symptoms of periodontal disease?
3	How is periodontal disease treated?
4	What is the treatment period for periodontal disease?
5	Is periodontal treatment painful?
6	Is there a relationship between periodontal disease and systemic diseases?
7	Does smoking affect periodontal health?
8	What causes gingival bleeding? Is it normal?
9	What is calculus, how does it occur?
10	Is calculus cleaning harmful?
11	What is gingival recession?
12	What are the causes of gingival recession?
13	Is gingival recession treated and how long does it take?
14	Does periodontal disease cause gingival recession?
15	How can we protect our oral and dental health?
16	What are the features of the ideal toothbrush?
17	How should we brush our teeth?
18	What is dental implant treatment?
19	Who can get dental implant treatment?
20	How long does dental implant treatment take to complete?

Results

Four chatbots answered all 20 questions performed 3 sets, resulting in a total of 240 responses. Comparisons of chatbot word counts are provided in Table 2 and Figure 2. In general, ChatGpt 4o mini and Deepseek gave longer answers with more detail (average word count: 643 (402.5) (range: 118-1034) and 523.5 (192) (range: 189-849), respectively). Responses from Copilot (average word count: 235.5 (98) (range: 133-398) and Gemini 1.5 flash daily help (average word count: 289 (138.5) (range: 127-508) were shorter and less detailed. When the questions were repeated, it was seen that all chatbots pursued correlative by repeated responses with little or no variations in the wording. When evaluated in terms of word count, it was determined that there was no difference ($p = 0.996$) between ChatGpt 4o mini and Deepseek, but both had more words ($p = 0.000$) than Copilot ($p = 0.000$) and Gemini 1.5 flash daily ($p = 0.000$). The word counts of Copilot and Gemini were not different ($p = 0.805$).

All four chatbots many times urged on their users to consult a dentist or a specialist for more particulars regarding their questions. No difference was found in the four chatbots in terms of recommending a visit to a dentist ($p < 0.05$) (Table 2). ChatGpt made recommendations 45/60 times; Deepseek 54/60 times; and Copilot 48/60 times, while Gemini 1.5 flash daily help made recommendations 51/60 times.

Details of all scores for the 240 responses after compromise among the raters are ensured in Table 2. The average score for each of the 3 responses to all question is showed in Figure 1. Scores range from 3 to 5. Copilot scored 3 on two questions (Questions 7 and 13) and Gemini 1.5 flash daily help scored 3 on six questions (Questions 3, 4, 6, 7, 9, and 14).

Low threshold validity

The whole of the answers from the four chatbots achieved a relatively superior level of validity, as shown in Table 2 and Figure 3. ChatGpt had the highest validity with all 20 responses (100%) classified as valid. Deepseek had 19 (95%) of the questions, Copilot 18 (90%), and Gemini 13 (65%) of the questions categorized as valid, respectively, and had lower validity. When the difference between the four chatbots was analyzed using the low threshold validity test, it was found that there was no difference between Deepseek ($p = 0.311$) and Copilot ($p = 0.147$) compared to ChatGpt, but ChatGpt was stronger than Gemini in terms of low threshold validity ($p = 0.004$). While there was no significant difference between Deepseek and Copilot ($p = 0.548$), Deepseek was found to be stronger than Gemini in terms of low threshold validity ($p = 0.018$). In addition, Copilot was also found to be stronger than Gemini in terms of low threshold validity ($p = 0.012$).

Table 2. Word counts, low and high validity, consultation suggestion numbers and scores of questions asked to chatbots

Variables	ChatGpt 4o mini	Deepseek	Copilot	Gemini 1.5 flash daily help
Word count median (IQR)	643 (402.5)	523.5 (192)	235.5 (98)	289 (138.5)
Min-max	(118-1034)	(189-849)	(133-398)	(127-508)
Score	5 (0)	5 (0)	4 (0)	4 (0)
	(4-5)	(3-5)	(3-5)	(3-5)
Low threshold validity test (score ≥ 4) n / %	20 / 100%	19 / 95%	18 / 90%	13 / 65%
High threshold validity test (score = 5) n / %	16 / 80%	15 / 75%	1 / 5%	1 / 5%
Consultation recommendation	45 / 75%	54 / 90%	48 / 80%	51 / 85%

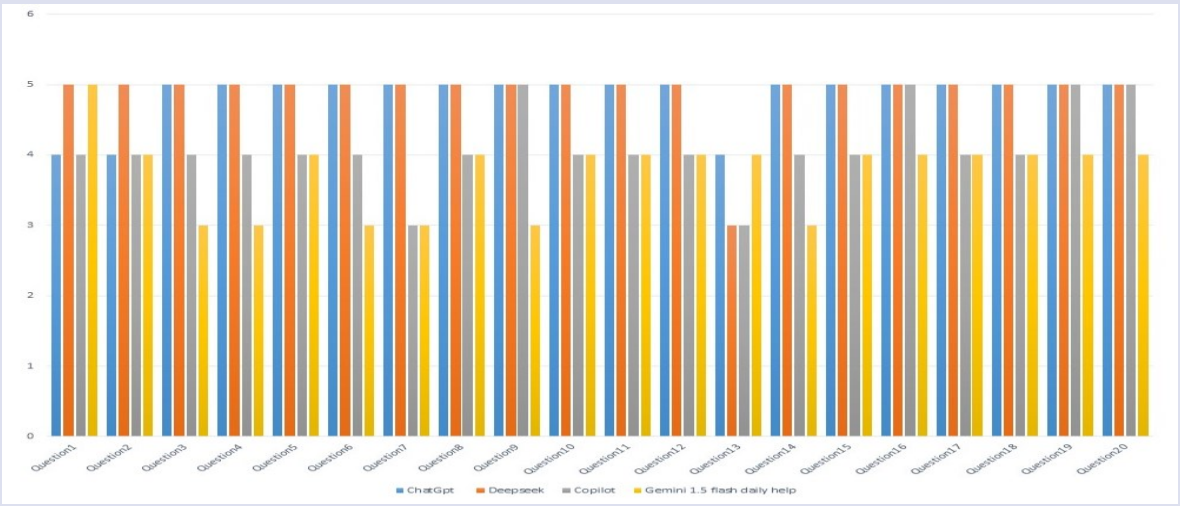

Figure 1. Distribution of average scores of chatbots' answers to 20 frequently asked questions. Each question was asked 3 times.

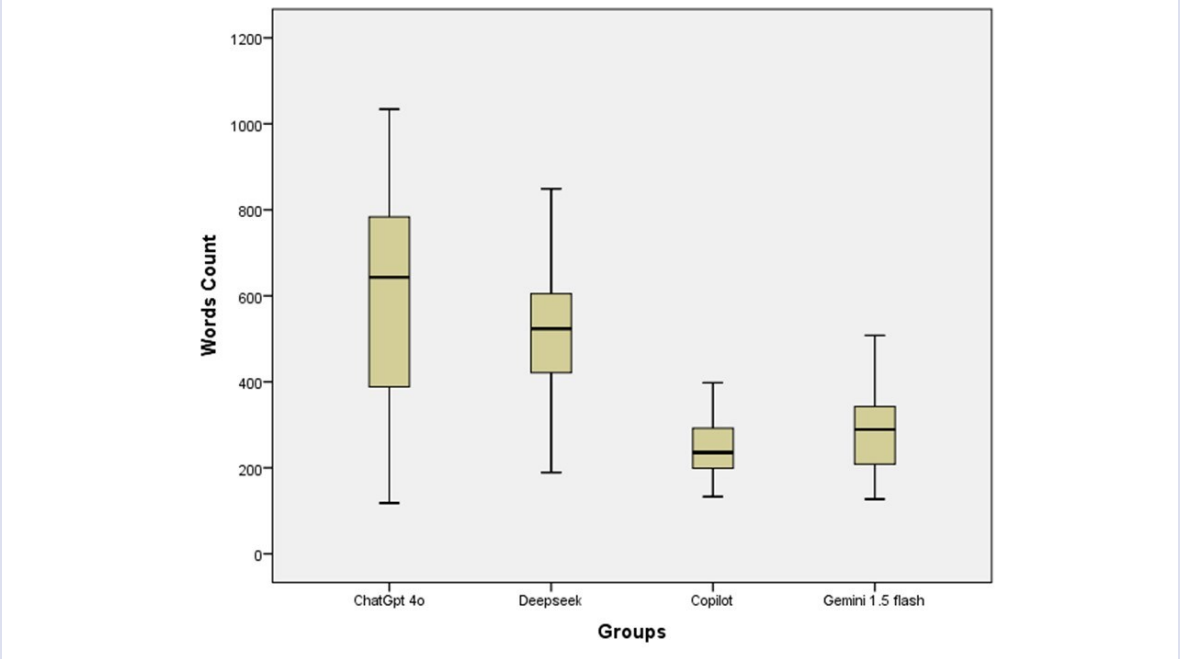

Figure 2. Comparison of the words count of chatbots' answers to questions.

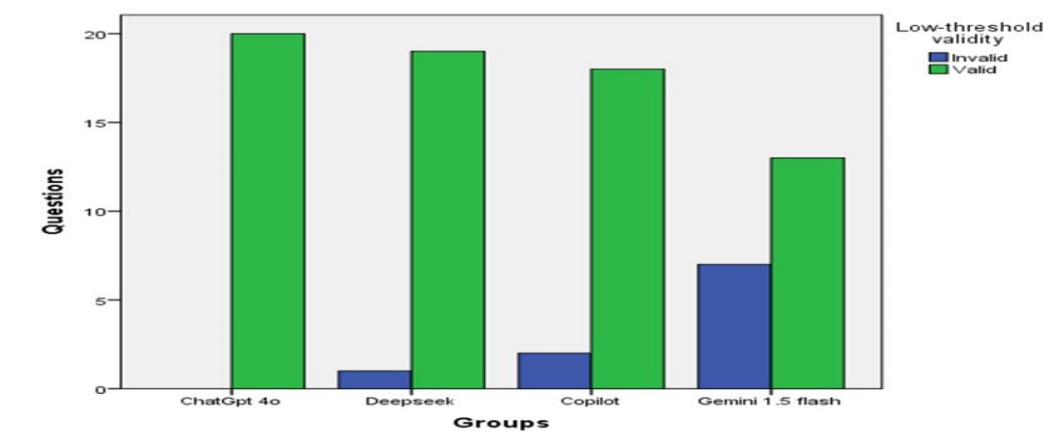

Figure 3. Low-threshold validity test: Results of a pairwise comparison of four chatbots on the validity of their answers to 20 questions repeated 3 times (Fisher's exact test).

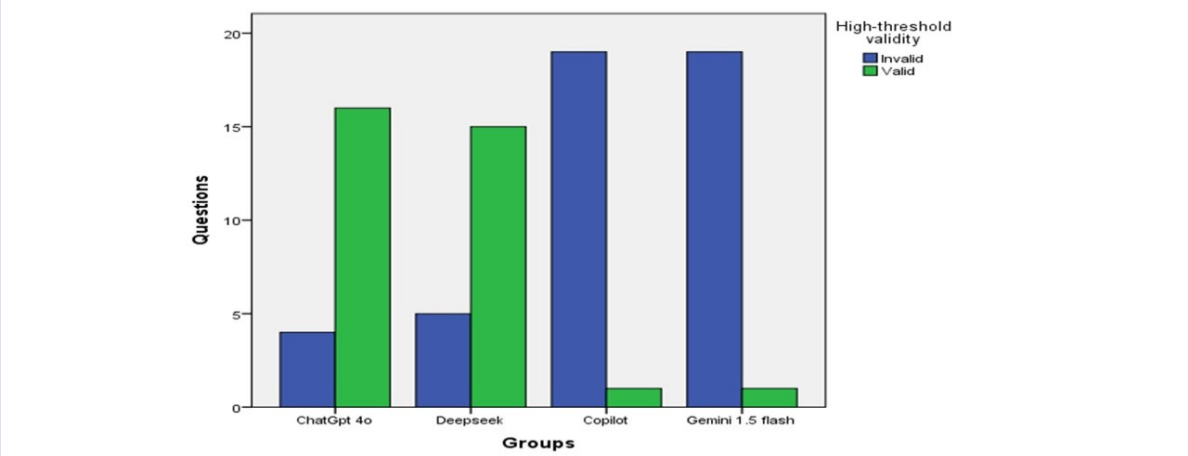

Figure 4. High-threshold validity test: Results of a pairwise comparison of four chatbots on the validity of their answers to 20 questions repeated 3 times (Fisher's exact test).

High threshold validity

All answers from the four chatbots achieved a relatively low level of validity when examined with the high threshold test. As shown in Table 2 and Figure 4. ChatGpt had the highest validity with 16 out of 20 responses (80%) classified as valid. ChatGpt was followed by Deepseek with 15 out of 20 responses (75%) classified as valid. Copilot and Gemini had lower validity with 1 out of 20 valid responses (5%) for both. No significant difference was found between ChatGpt and Deepseek ($p = 0.705$). ChatGpt achieved a significantly higher validity compared to both Copilot ($p = 0.000$) and Gemini ($p = 0.000$). Deepseek also had higher validity than both Copilot ($p = 0.000$) and Gemini ($p = 0.000$). However, there was no difference between Copilot and Gemini ($p = 1$).

Reliability

All four chatbots had an acceptable level of reliability. Among the four chatbots, ChatGpt had the highest consistency with a Cronbach's alpha of 0.962, followed by Deepseek with a Cronbach's alpha of 0.912, Copilot with a Cronbach's alpha of 0.875, and Gemini with a Cronbach's alpha of 0.842.

The inter-rater agreement between the two periodontists was evaluated using Cohen's Kappa coefficient. The level of agreement was found to be very high, with a Kappa value of 0.83 ($p < 0.001$), indicating a statistically significant concordance beyond chance.

Discussion

Artificial intelligence chatbots have frequently appeared in recent years as easily accessible sources of information for those seeking information about care, diseases, treatment, and treatment results in the field of dentistry.[14] However, the accuracy, deficiencies, benefits, or harms of the information provided to the public by chatbots should be taken into consideration and evaluated critically. In fact, the public should be informed about the validity and reliability of these data sources. It does not seem possible to monitor artificial intelligence chatbots and prevent the spread of false information today. However, these robots can help guide patients correctly with easy-to-access, accurate information, make

patient expectations realistic, increase patient satisfaction, ensure effective and appropriate use of health services, and reduce service costs. This is the first study to evaluate the validity and reliability of the answers given by artificial intelligence chatbots to constantly asked questions by patients in the field of periodontics.

This article evaluates the answers given by chatbots to the most frequently asked questions of patients about diseases, treatments and treatment results in the field of periodontics. The questions were synthesized in a representative and comprehensive way from frequently asked questions in real patient consultations and asked to artificial intelligence chatbots in English. The source and language of the questions are important when testing chatbots. Because chatbots use a trained language model that can synthesize content on numerous topics from very large data sets while giving answers. In the future, chatbots can be tested by preparing questions in different languages from questions of individuals from different geographies, cultures and races on a global scale. In this study, it would be ideal for the questions evaluated by two experts to be scored by a larger expert team. In addition, it is the duty of researchers to continuously test the quality of the information ensured by chatbots that develop and update over time, publish it and report it to the public.

While the validity of chatbots is high according to low threshold validity, the validities are low at high threshold. Nevertheless, ChatGpt 4o and Deepseek showed close and high validity, while Copilot and Gemini 1.5 flash could not give valid answers according to high threshold validity. Therefore, getting misinformed from these platforms, which the society can easily access and obtain information in the field of periodontics, can be dangerous for health. The spread of misinformation can be dangerous for human health of these chatbots. This current study found Gpt 4o as the most valid and safe chatbot, in line with the studies in endodontics,[10] radiology[15] and public health[16] that found Gpt-3.5 to be the chatbot that provides the best performance. Technically, ChatGpt 4o understands the contexts in the language with its deep learning architecture called transformer and is quite effective in producing text. It establishes meaningful dialogues by giving logical answers with its high language skills[6]. While ChatGpt focuses on text understanding and text generation, Deepseek is a data search and discovery application. It analyzes databases by searching in depth, extracts the most meaningful result, and reaches the most appropriate information.[17] Launched in January 2025, Deepseek combines machine learning and natural language processing technologies. In this study, both chatbots provided highly valid and reliable answers to questions asked in the field of periodontics, and the answers were detailed in content and almost equally long. For patient care, both increase reproducibility by standardizing data interpretation and supporting evidence-based decisions, but neither ethically replaces human judgment in complex scenarios.[17]

Both Gemini, developed by Google, and Copilot, which works integrated with Microsoft 365 (Word, Excel, PowerPoint, Outlook, etc.) applications, provided relatively shorter and less detailed answers and were not found to be as valid as other chatbots at the high validity threshold in this study. However, all chatbots had an acceptable level of reliability. Another study emphasized that Gemini 1.5 Flash was slightly less capable than ChatGpt 4o-mini in predicting the research quality of an article, but found small differences.[18] There may be differences in the answers of chatbots depending on the changing design of AI companies and the algorithms used[19]. Another article investigating the validity and reliability of chatbots by addressing frequently asked questions about dental trauma[11] reported that Claude AI chatbot showed superior validity and reliability compared to ChatGpt and Google Gemini, while Bing was less reliable. The article also highlighted that authorities should establish strict guidelines to ensure the accuracy of medical information provided by chatbots.[11]

Despite chatbots' high validity scores, some critical errors in answers may be due to their lack of depth of understanding and limited ability to handle complex queries. Chatbots may synthesize information or pull from low-quality sources. This may result in responses that are biased, outdated, deficient, or incorrect. A significant limitation is that they cannot give references or citations for the data they produce. This increases concerns about trusting them for health information.[20]

This study has some limitations. The questions were selected from the most frequently asked questions during a patient consultation by a periodontologist. Further studies can be conducted with the list of questions determined by a chatbot to cover the topics that a wider audience is curious about. In addition, for wider generalizability and methodological consistency, the Global Quality Scale (GQS) or other validated tools can be included to ensure greater comparability.

Conclusions

In conclusion, although all four chatbots that are publicly available information sources were found to have high validity in this study, ChatGpt 4o and Deepseek had promising results. Although chatbots have aspects that need to be improved, it is important for researchers to provide regular information about valid information and missing information.

Conflicts of Interest Statement

No financial support was received for the current study. The author declares no conflict of interest.

References

1. Bayrakdar İ, Çelik Ö, Orhan K, Bilgir E, Odabaş A, Aslan A. Success of artificial intelligence system in determining alveolar bone loss from dental panoramic radiography images. Cumhuriyet Dent J 2020;23(4):318-324.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;28;521(7553):436-444.

3. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015; 61:85-117.
4. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med 2023;183(6):589-596.
5. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review. J Med Internet Res 2020;18;22(12):e19127.
6. Burisch C, Bellary A, Breuckmann F, Ehlers J, Thal SC, Sellmann T, Gödde D. ChatGPT-4 Performance on German Continuing Medical Education-Friend or Foe (Trick or Treat)? Protocol for a Randomized Controlled Trial. JMIR Res Protoc 2025;14:e63887.
7. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, Temsah MH. DeepSeek in Healthcare: Revealing Opportunities and Steering Challenges of a New Open-Source Artificial Intelligence Frontier. Cureus 2025;17(2):e79221.
8. Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine (Baltimore). 2024;103(33):e39305.
9. Wang S, Wang Y, Jiang L, Chang Y, Zhang S, Zhao K, Chen L, Gao C. Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation. Eur J Med Res 2025;30(1):45.
10. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J 2024;57(3):305-314.
11. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, Mohan N. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. Dent Traumatol 2025;41(2):187-193.
12. 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(9):2070-7.
13. Bland JM, Altman DG. Cronbach's alpha. BMJ. 1997;314(7080):572. doi: 10.1136/bmj.314.7080.572.
14. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res 2023;25:e47479.
15. Doshi RH, Amin K, Khosla P, Bajaj S, Chheang S, Forman H. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google Bard, and Microsoft Bing. MedRxiv. 2023; 06. 04.23290786.
16. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, Smith DM. Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Netw Open 2023;6(6):e2317517.
17. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-Redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc 2025;33(5):1553-1556.
18. Thelwall M. Is Google Gemini better than ChatGPT at evaluating research quality?. Journal of Data and Information Science 2025;10(2):1-5.
19. Bhardwaz S, Kumar J. An extensive comparative analysis of chatbot technologies -ChatGPT, Google BARD and Microsoft Bing. In: 2023 2nd international conference on applied artificial intelligence and computing (ICAAIC). Salem, India: IEEE 2023:673-679.
20. Kerbage A, Burke CA, Rouphael C. Artificial Intelligence Chatbots in Healthcare: Navigating Accuracy, Privacy, and Global Applicability. Clin Gastroenterol Hepatol 2024;22(10):2158-2159.