Original Investigation / Orijinal Araştırma

DOI: 10.4274/tod.galenos.2025.01488
Turk J Osteoporos

A New Era in Digital Health: Evaluation of Artificial Intelligence Supported Chatbots (ChatGPT-4, BingAI, and Gemini) Responses on Osteoporosis

Dijital Sağlıkta Yeni Bir Dönem: Yapay Zeka Destekli Sohbet Robotlarının (ChatGPT-4, BingAI ve Gemini) Osteoporozla İlgili Yanıtlarının Değerlendirilmesi

Gamze Gül Güleç[1], Özge Özpolat Bulut[2], Fatih Bağıcer[3]

[1]Başkent University Adana Dr. Turgut Noyan Education and Research Center, Department of Physical Medicine and Rehabilitation, Adana, Türkiye
[2]Viranşehir State Hospital, Clinic of Physical Medicine and Rehabilitation, Şanlıurfa, Türkiye
[3]University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Department of Physical Medicine and Rehabilitation, İstanbul, Türkiye

Abstract

Objective: This study aimed to evaluate and compare the quality and readability of osteoporosis-related information generated by three artificial intelligence (AI) chatbots: ChatGPT-4, BingAI, and Gemini.
Materials and Methods: 25 frequently asked questions about osteoporosis (obtained via Google Trends through Gemini) were submitted to each chatbot on December 23, 2024. The first responses were evaluated for readability [Flesch-Kincaid Reading Ease (FKRE) and Flesch-Kincaid Grade Level (FKGL)] and quality (EQIP tool). Two experienced clinicians assessed the accuracy and completeness using Likert scales.
Results: The mean FKRE scores were 34.5±12.9 (ChatGPT-4), 33.8±14.3 (BingAI), and 36.1±10.9 (Gemini), indicating difficulty in reading the texts. The FKGL scores ranged from 11.2 to 12.5, suggesting that college-level reading ability was required. However, BingAI (EQIP: 55.4±7.9) and Gemini (54.4±8.8) outperformed ChatGPT-4 (48.6±6.3) in terms of quality (p=0.005). Accuracy and completeness were high across all models, with mean scores exceeding 4.3/5 for each.
Conclusion: While all three AI chatbots delivered accurate and complete answers on osteoporosis, their content readability remained suboptimal. BingAI and Gemini provide higher-quality information, possibly due to real-time web integration. Future chatbot development should focus on enhancing readability and real-time data access to support effective health communication, particularly in conditions such as osteoporosis, where patient understanding is crucial.
Keywords: Osteoporosis, artificial intelligence, chatbot, ChatGPT, readability, health communication

Öz

Amaç: Bu çalışmanın amacı, yapay zeka (AI) destekli sohbet robotlarının (ChatGPT-4, BingAI ve Gemini) osteoporoz ile ilgili verdiği bilgilerin kalitesini ve okunabilirliğini değerlendirmek ve karşılaştırmaktır.
Gereç ve Yöntem: Google Trends üzerinden osteoporoz hakkında en sık sorulan 25 soru belirlendi ve her bir sohbet robotuna ayrı ayrı soruldu. İlk verilen yanıtlar okunabilirlik [Flesch-Kincaid Okuma Kolaylığı (FKRE) ve Flesch-Kincaid Sınıf Düzeyi (FKGL)] ve bilgi ve yazım kalitesi (EQIP aracı) açısından değerlendirildi. Yanıtların doğruluğu ve yeterliliği iki deneyimli klinisyen tarafından Likert ölçeğiyle değerlendirdi.
Bulgular: Ortalama FKRE skorları ChatGPT-4, BingAI ve Gemini için sırasıyla 34,5, 33,8 ve 36,1 idi. FKGL puanları 11,2 ile 12,5 arasında değişmekteydi. Bu skorlar metinlerin okunmasının zor olduğunu ve üniversite düzeyinde okuma becerisi gerektirdiğini ortaya koydu. Kalite açısından BingAI (EQIP: 55,4±7,9) ve Gemini (54,4±8,8), ChatGPT-4'ten (48,6±6,3) anlamlı şekilde daha iyi performans gösterdi (p=0,005). Tüm modellerde doğruluk ve yeterlilik yüksek olup, ortalama puanlar 5 üzerinden 4,3'ün üzerindeydi.
Sonuç: Üç yapay zeka sohbet robotu da osteoporoz hakkında doğru ve yeterli yanıtlar üretse de içeriklerinin okunabilirliği hala istenilen seviyede değildir. BingAI ve Gemini, muhtemelen anlık veri kullandığından daha yüksek kaliteli bilgiler sunmaktadır. Sohbet robotlarının güncellemelerinde okunabilirliğin artırılması ve güncel veri erişiminin sağlanması, osteoporoz gibi anlaşılması önem arzeden konularda sağlık iletişimini güçlendirebilir.
Anahtar kelimeler: Osteoporoz, yapay zeka, sohbet robotu, ChatGPT, okunabilirlik, sağlık iletişimi

Corresponding Author/Sorumlu Yazar: Gamze Gül Güleç MD, Başkent University Adana Dr. Turgut Noyan Education and Research Center, Deparmet of Physical Medicine and Rehabilitation, Adana, Türkiye
E-mail: gamzegulgulec@gmail.com **ORCID ID:** orcid.org/0000-0003-2020-1507
Received/Geliş Tarihi: 12.04.2025 **Accepted/Kabul Tarihi:** 23.06.2025 **Epub:** 24.07.2025

Cite this article as/Atıf: Güleç GG, Özpolat Bulut Ö, Bağcıer F. A new era in digital health: evaluation of artificial intelligence supported chatbots (ChatGPT-4, BingAI, and Gemini) responses on osteoporosis. Turk J Osteoporos. [Epub Ahead of Print]

Introduction

Osteoporosis is a widespread metabolic bone disorder that has a profound impact on the health burden of the aging population and affects millions of individuals globally. This disease is characterized by a decrease in the mineral density of bone tissue, leading to brittle bones (1). This means that there is a risk of more serious fractures, particularly in the spine, hip and wrists (2). In the US, approximately 2 million osteoporotic fractures occur annually, impairing individuals' quality of life and imposing a high economic burden on the healthcare system (3). Osteoporosis is often under-recognized or diagnosed late (4). As a result, asignificant majority of patients remain undiagnosed and do not receive treatment until a fracture develops. Research has revealed that only 25% of patients with osteoporosis even know that they have it (1). This lack of awareness is especially evident in areas with limited healthcare access and among populations with limited health literacy (5).

Today, digital health solutions, particularly innovative technologies such as artificial intelligence (AI)-powered chatbots, have begun to play a crucial role in the management of chronic diseases, such as osteoporosis. There are different types of AI, such as machine learning and natural language processing, large language models (LLMs) (6). These technologies facilitate health management by offering services such as information dissemination, symptom tracking, and treatment recommendations. However, online health information often lacks adequate moderation, leading to significant variability in the quality and reliability of information (7).

In particular, LLMs have become a major focus of research and development because their ability to process and generate human-like text, given their training on large datasets, has generated significant interest. Among them are ChatGPT-4, BingAI, and Gemini, to name a few, each boasting particular characteristics and capabilities (8).

OpenAI's ChatGPT-4 is one of the Generative Pre-trained Transformer series and is recognized for its advanced natural language understanding and generation capabilities. To create a more well-behaved model, it was fine-tuned with both supervised learning and reinforcement learning, resulting in highly fluid and contextually appropriate answers on a wider array of subjects (9). For instance, BingAI, an LLM integrated with the Microsoft Bing Search Engine, supports a variant of the GPT model and provides a version optimized for real-time information retrieval along with research benefits, which improves the accuracy and relevance of the result segments. BingAI's integration with a search engine allows it to provide up-to-date information, making it a valuable tool for accessing current medical guidelines and studies (9). Gemini, developed by Google, is based on the Language Model for Dialogue Applications and is designed to create informative and conversational content, continuously updating its knowledge base with the latest web information to ensure that its responses are both current and contextually relevant (10).

The quality of health-related information on osteoporosis found through AI chatbots has been examined in the literature (11). Previous comparative studies of different chatbots have assessed the differences in readability and quality between responses to the same theme generated by different chatbots, but not for osteoporosis. The present study aimed to evaluate and compare the quality and readability of information provided by three different AI chatbots for the most common questions on osteoporosis.

Materials and Methods

The study was conducted on December 23, 2024, at the Clinic of Physical Therapy and Rehabilitation at Viranşehir State Hospital State Hospital. This study did not involve any processes with live animals or human participants; therefore, institutional ethical approval was not required. To avoid any potential bias, all personal data from the browser were cleared before the searches. Additionally, all chat sessions were initiated in clean browsers with cleared cookies and no previous prompt history to eliminate prior interaction effects.

Three separate chatbots (ChatGPT-4, BingAI and Gemini) were posed 25 of the most frequently asked questions about osteoporosis on 23.12.2024. A prompt was submitted to Gemini to retrieve the 25 most frequently asked questions about osteoporosis based on Google Trends data. Gemini served solely as an interface to access publicly available search query data, without generating or altering the content. This approach was chosen to reflect real-world public interest in a neutral and reproducible manner and has been validated by similar studies in the literature (7,8,11). The prompt given to Gemini was, "Can you write the 25 most frequently asked questions about osteoporosis according to Google Trends?" Since LLMs can produce different answers to the same question, only the first answers were considered for each question. This is because the first answers tend to reflect what LLMs consider to be the most likely and correct responses. The word count was not limited, which allowed for extensive explanations. Each question was entered into the chatbots individually on a separate page.

The answers obtained from the chatbots about osteoporosis were obtained by a researcher. The responses were then evaluated by 2 different clinicians, each with at least 5 years of experience in the diagnosis and management of OP. If there were differences between the clinicians' evaluations, they were evaluated by a third independent clinician and a joint decision was made. If a response included a reference or DOI, it was manually verified via academic databases such as PubMed and CrossRef. Inter-rater agreement for the 5-point accuracy ratings was calculated using Cohen's kappa (κ=0.92), indicating excellent reliability. The clinicians were blind to which LLM the texts belonged to.

The accuracy and adequacy of the texts obtained from the LLMs were evaluated according to the Likert scale, based on previous studies (12). The accuracy of the texts was assessed according to a 5-point Likert scale (1: very poor accuracy or unacceptable inaccuracies with high risk of harm; 2: poor accuracy or potentially harmful errors; 3: negligible moderate inaccuracies; 4: good level of accuracy with minor inaccuracies; 5: very good level of accuracy, no risk of harm). The adequacy of the texts was assessed according to a 3-point Likert scale (1: incomplete presentation of important parts of information addressing some aspects of the problem; 2: adequate presentation of information addressing all aspects of the problem; 3: more information than expected addressing all aspects of the problem).

To evaluate the quality of the text generated by large language models, we used the ensuring quality information for patients (EQIP) tool. This assessment tool evaluates the content in thirty-two different ways, including whether the information is consistent and whether the writing is appropriate (13). The tool consists of 20 questions answered "yes," "somewhat," "no" or "not applicable." The scoring is done by multiplying the number of "yes" by 1 (so the more of these you have, the better), the number of "partially" by 0.5 and the number of "no" by 0. These are summed, with the total number of "does not apply" responses subtracted from 20 total items, then divided by the new total number of items. The final value is multiplied by 100, to obtain the EQIP score which is expressed as a percentage. EQIP are classified as follows: 76-100%: Well-written, great quality; 51-75%: Good quality, minor issues; 26-50%: Serious quality issues; 0-25%: Severe quality issues.

Flesch-Kincaid Reading Ease (FKRE) and Flesch- Kincaid Grade Level (FKGL) scores were used to evaluate the readability of the texts from the LLMs. The FKRE score, which ranges from 0 to 100, is a widely used readability score tool, and a higher score corresponds to improved readability. FKRE score =206.835-1.015 × (average sentence length) + 84.6 × (average word length) The FKGL score is a modified version of the FKRE score, which denotes the average US school grade level that is capable of understanding the text, with a lower score signifying an increase in readability (14).

Statistical Analysis

All statistical analyses were conducted with IBM SPSS version 22.0 software (IBM Corp., Armonk, NY, USA). Normality of the data distribution was assessed by Kurtosis-Skewness values and the Kolmogorov-Smirnov/Shapiro-Wilk test. Mean, standard deviation and median were calculated to describe the study variables. Group differences were evaluated by ANOVA or, as appropriate, the Kruskal-Wallis test. If significant differences were found, pairwise comparisons were performed using the t-test or Mann-Whitney U test. Statistical significance was defined as $p<0.05$.

Results

In this study, Gemini was used to retrieve a cumulative total of 25 frequently asked questions relating to osteoporosis based on Google Trends data. The questions cover various aspects of osteoporosis, including its meaning, its warning signs, prevention, risks, and treatment. The top five questions looked to learn more about the condition itself ("What is osteoporosis?"), determining its symptoms ("What are the symptoms of osteoporosis?"), and ways to prevent the disease ("What are the best ways to prevent osteoporosis?") (Table 1).

Geographic analysis showed higher search interest in osteoporosis in Puerto Rico, Ecuador, and Bolivia (Figure 1). The popularity of osteoporosis over time (the analysis was conducted on a time range starting from 2004) is presented according to the Google Trends analysis (Figure 2). Responses generated by all three AI chatbots—ChatGPT-4, BingAI and Gemini—were subjected to readability analysis using the FKRE and FKGL metrics. The readability scores for all models suggested that the outputted information was hard for the average population to read (all models $p>0.05$). Mean FKRE scores obtained with ChatGPT-4, BingAI, and Gemini were 34.5, 33.8, and 36.1, respectively, all classified as "difficult" (FKRE) scoring. In a similar manner, FKGL scores from 11.2 to 12.5 were noted, suggesting a need for a college-level education to comprehend the content (Table 2).

Responses were analyzed for quality using the EQIP tool. Chatbots produced significantly different quality scores ($p=0.005$). The mean EQIP score for ChatGPT-4 was a mere 48.6 points, significantly lower than that of either BingAI (55.4) or Gemini (54.4).

For both the completeness and correctness of responses, all three models performed similarly without any statistically significant differences among them ($p>0.05$). On a 5-point Likert scale of accuracy, the mean accuracy scores were 4.3, 4.4, and 4.3 for ChatGPT-4, BingAI, and Gemini, respectively, indicating that the model-generated responses were generally accurate but less so in some instances. All answers from the models had no major mistakes (1 or 2 points).

Discussion

In this study, we assessed the responses of three distinct Big Language Models: ChatGPT-4, BingAI, and Gemini. Responses were examined with regard to multiple dimensions, including readability, information quality, completeness and accuracy. Although no statistically significant differences were observed between the models for most parameters, statistically significant differences were observed between the models for the EQIP scores. These results show that although the general performance of the models was similar, the quality of information provided by ChatGPT was worse than that of the other models.

Table 1. Top 25 questions searched about osteoporosis across countries: 2004-2024 (based on Google Trends data)		
Rank	**Question**	**Category of the topic based on EQIP**
1	What is osteoporosis?	Condition or illness
2	What are the symptoms of osteoporosis?	Condition or illness
3	What are the most effective methods to prevent osteoporosis?	Prevention or after care
4	Which age groups are at risk of osteoporosis?	Condition or illness
5	Which drugs are used in the treatment of osteoporosis?	Medication or product
6	What is the relationship between osteoporosis and nutrition?	Condition or illness
7	What is the role of physical activity in osteoporosis?	Miscellaneous
8	How is osteoporosis diagnosed?	Test, operation, investigation, or procedure
9	What are the genetic factors of osteoporosis?	Condition or illness
10	What factors increase the risk of osteoporosis in women?	Condition or illness
11	What are the best dietary recommendations for people with osteoporosis?	Miscellaneous
12	Which vitamins and minerals are effective in preventing osteoporosis?	Prevention or after care
13	What is the relationship between osteoporosis and menopause?	Condition or illness
14	What kind of exercise should be done to reduce the risk of osteoporosis?	Prevention or after care
15	What are the psychological effects of osteoporosis?	Miscellaneous
16	What is the relationship between osteoporosis and fractures?	Condition or illness
17	Are there natural methods to treat osteoporosis?	Prevention or after care
18	What are the most common misconceptions about osteoporosis?	Condition or illness
19	What are the latest technologies in the treatment of osteoporosis?	Medication or product
20	What is the impact of regular screenings on osteoporosis?	Miscellaneous
21	What are the differences between osteoporosis and other bone diseases?	Condition or illness
22	What role does physical therapy play in the treatment of osteoporosis?	Services
23	What tests should be done for osteoporosis?	Test, operation, investigation, or procedure
24	What should people with osteoporosis pay attention to in their daily life?	Prevention or after care
25	Common myths about osteoporosis	Miscellaneous
EQIP: Ensuring quality information for patients		

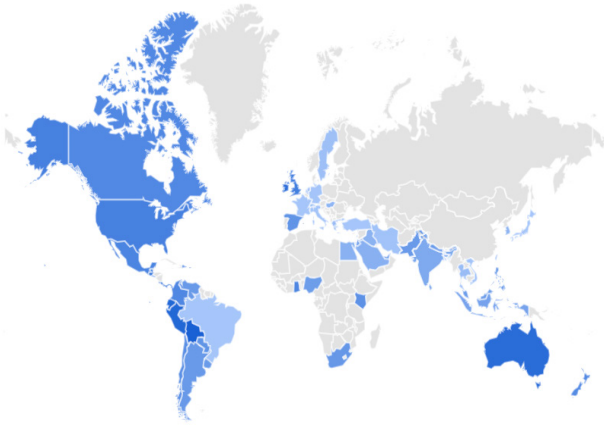

Figure 1. World map showing the relative search interest for the term "osteoporosis" by country based on Google Trends data. Darker shades of blue indicate higher levels of interest, whereas grey indicates regions with insufficient data

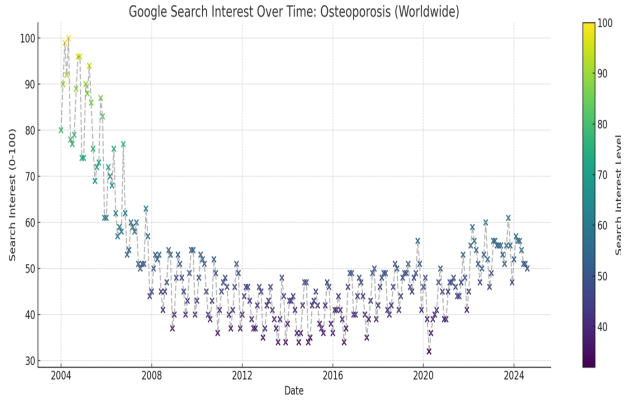

Figure 2. Timeline of global Google search interest for the term "osteoporosis" from January 2004 onwards. The color intensity of the data points corresponds to the level of search interest, with brighter colors indicating greater public attention. This trend reflects temporal changes in public awareness and potential influences such as awareness campaigns, research publications, or media coverage

Table 2. Comparison of large language models in terms of readability, quality, completeness, and accuracy

	ChatGPT	BingAI	Gemini	p
FKRE Mean ± SD Median (min-max)	 34.5±12.9 36.5 (16.0-61.2)	 33.8±14.3 34.8 (11.2-66.7)	 36.1±10.9 33.7 (23.0-67.3)	0.789
FKGL Mean ± SD Median (min-max)	 11.3±2.3 11.0 (7.1-15.6)	 12.5±2.7 12.1 (7.1-17.1)	 11.2±1.8 11.1 (8.1- 15.0)	0.127
EQIP Mean ± SD Medain (min-max)	 48.6±6.3*+ 50.0 (34.2-60.0)	 55.4±7.9 55.6 (40.7-69.4)	 54.4±8.8 55.5 (36.5-68.7)	**0.005**
Completeness Mean ± SD Median (min-max)	 2.1±0.2 2.0 (2.0-3.0)	 2.0±0.7 2.0 (1.0-3.0)	 2.0±0.2 2.0 (2.0-3.0)	0.553
Accuracy Mean ± SD Median (min-max)	 4.3±0.5 4.0 (3.0-5.0)	 4.4±0.5 4.0 (4.0-5.0)	 4.3±0.6 4.0 (3.0-5.0)	0.907

FKRE: Flesch-Kincaid Reading Ease, SD: Standard deviation, EQIP: Ensuring quality information for patients

There was no statistically significant difference between all three models in terms of FKRE and FKGL scores with all models producing content that required a higher education level to understand. These findings are consistent with previous studies reviewing AI-driven chatbots in health, for instance, those concerning retinopathy of prematurity or erectile dysfunction, which noted similar readability issues (9,15). Despite its long-established theory, the limited spread between the FKRE and FKGL scores invites novel questions about the generative mechanisms that govern these metrics across varying models. Another possible reason for the consistency in readability is the common use of large-scale datasets with a lot of technical and specialty medical information. These models are trained on large corpora of text, many of which are sourced from academic texts, clinical guidelines, and research papers, producing outputs that naturally mirror the complexity of their sources. This may lead to high and consistent FKGL scores, as rewriting complex medical acronyms in a simple manner while maintaining the same level of information is a complex task for LLMs.

Additionally, the small differences in readability scores could imply that current AI models prioritize writing accurate and complete information over writing accessible information. This is consistent with earlier research finding that LLMs can provide purportedly detailed and contextually accurate information without presenting it in an easily understandable manner for the general population. A study comparing the readability of online health information on stuttering has shown that even widely utilized resources frequently fall short of the recommended readability levels for medical literature (16). Therefore, it appears that readability, especially for medical information, remains a wider problem not limited to AI models.

Even if the FKRE and FKGL scores were similar, it might be worth speculating about the lack of a model that outperformed the others in terms of readability. Therefore, a possible explanation could be the training methodologies used in different LLMs. ChatGPT-4, BingAI, and Gemini are all from different companies, yet they may share similar pre-processing methods for medical terminology that normalize language complexity. Alternatively, all of the models may be limited by the trade-offs that come with readability versus accuracy—using basic language risks sacrificing the impact of the medical content, but the more detail you give, the more complicated it becomes.

Notably, despite no major differences in readability scores, the EQIP tool noted clear differences in the quality of information between the two models. ChatGPT-4 was greatly outperformed in quality by the production versions of BingAI and Gemini, both of which are powered by real-time information retrieval systems. This means that while all models might have difficulty with readability, having up-to-date information might result in more relevant and higher-quality content. In contrast, ChatGPT-4, which is more reliant on its pre-trained dataset with no real-time data, might be slower in providing the most up-to-date health information, which may negatively affect its EQIP score.

This finding contrasts with studies that reported more consistent EQIP scores across different AI models in other medical contexts, such as a study evaluating responses to spinal cord injury-related questions (17). One possible reason for this discrepancy is the nature of the conditions being discussed. Osteoporosis, as a chronic and evolving disease, requires up-to-date knowledge of recent clinical guidelines, medications, and prevention strategies. In conditions where real-time information plays a critical role, the advantages of models with real-time data retrieval, such as BingAI and Gemini, become more pronounced. This raises the possibility that the performance gap observed in this study could widen further in rapidly evolving fields of medicine, where new treatments and guidelines frequently emerge.

The other aspect of the question is how much their responses depend on the user interaction. BingAI and Gemini are likely

to be related to search engines and are trained on real-time data; therefore, there is a high chance that they were fine-tuned on real user queries and data. This real-time feedback loop may help improve response quality over time, unlike ChatGPT-4's static model, which would not see such incremental learning adaptively.

From a completeness and accuracy standpoint, there were no noteworthy differences between the models, which all performed well in answering the questions. This finding indicates that LLMs of all types and training data have comparable proficiency in processing all core medical concepts relevant to osteoporosis. Some responses contained small inaccuracies, highlighting the need for improvements, especially in more nuanced medical topics. The results are consistent with the existing literature, in that AI models have shown very high accuracy whenever presented with general medical knowledge but not with more specialized or context-driven information (18). Future advancements in LLMs may include improving both content readability and quality by implementing intelligent algorithms that adjust the complexity of the language to suit the individual reader's capacity to understand. Such technology could tailor the content in real time according to the user's previous encounters with content or health literacy. At the moment, human supervision is required for maintaining the readability and accuracy of AI-generated health information (19). Until now, AI and human experts have been working together in such a way that people can more easily get AI-driven health information in a hybrid style.

One limitation of this study is the relatively small number of questions used to evaluate the models. A larger and more diverse dataset could provide a more comprehensive comparison. Additionally, while readability and quality were measured, user satisfaction and engagement were not, which could be important metrics for evaluating the practical utility of these models.

Conclusion

This study underlines the strengths and weaknesses of ChatGPT-4, BingAI and Gemini for osteoporosis-related health information. While the output was similar in terms of readability levels across the models, BingAI and Gemini produced superior responses, in part due to their access to real-time data. As these AI tools become more prominent in health communication, future use should also focus on accessibility, the provision of real-time updates, and human oversight to mitigate their inaccurate use.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT-4o to obtain background information and assist with language editing. All outputs generated by the AI tool were carefully reviewed, revised, and verified by the authors. The authors take full responsibility for the accuracy and integrity of this final manuscript. The use of AI was limited to information retrieval and language refinement and did not influence the study's data, analysis, or conclusions.

Ethics

Ethics Committee Approval: This study did not require an ethics committee certificate as it was not conducted on humans.
Informed Consent: Since this study was not conducted on human subjects, patient consent was not required.

Footnotes

Authorship Contributions

Concept: Ö.Ö.B., F.B., Design: F.B., Data Collection or Processing: Ö.Ö.B., Analysis or Interpretation: G.G.G., Ö.Ö.B., Literature Search: G.G.G., F.B., Writing: G.G.G., Ö.Ö.B., F.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Sözen T, Özışık L, Başaran NÇ. An overview and management of osteoporosis. Eur J Rheumatol. 2017;4:46-56.
2. Pouresmaeili F, Kamalidehghan B, Kamarehei M, Goh YM. A comprehensive overview on osteoporosis and its risk factors. Ther Clin Risk Manag. 2018;14:2029-49.
3. Rashki Kemmak A, Rezapour A, Jahangiri R, Nikjoo S, Farabi H, Soleimanpour S. Economic burden of osteoporosis in the world: a systematic review. Med J Islam Repub Iran. 2020;34:154.
4. LeBoff MS, Greenspan SL, Insogna KL, Lewiecki EM, Saag KG, Singer AJ, et al. The clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int. 2022;33:2049-102. Erratum in: Osteoporos Int. 2022;33:2243.
5. Allen-Meares P, Lowry B, Estrella ML, Mansuri S. Health literacy barriers in the health care system: barriers and opportunities for the profession. Health Soc Work. 2020;45:62-4.
6. Singareddy S, Sn VP, Jaramillo AP, Yasir M, Iyer N, Hussein S, et al. Artificial intelligence and its role in the management of chronic medical conditions: a systematic Review. Cureus. 2023;15:e46066.
7. Battineni G, Baldoni S, Chintalapudi N, Sagaro GG, Pallotta G, Nittari G, et al. Factors affecting the quality and reliability of online health information. Digit Health. 2020;6:2055207620948996.
8. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11:1166120.
9. Durmaz Engin C, Karatas E, Ozturk T. Exploring the role of ChatGPT-4, BingAI, and gemini as virtual consultants to educate families about retinopathy of prematurity. Children (Basel). 2024;11:750.
10. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. Healthcare (Basel). 2023;11:2776.
11. Erden Y, Temel MH, Bağcıer F. Artificial intelligence insights into osteoporosis: assessing ChatGPT's information quality and readability. Arch Osteoporos. 2024;19:17.
12. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large language model's responses to questions and cases about glaucoma and retina management. JAMA

Ophthalmol. 2024;142:371-5. Erratum in: JAMA Ophthalmol. 2024;142:393.

13. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7:165-75.
14. Alas AN, Bergman J, Dunivan GC, Rashid R, Morrisroe SN, Rogers RG, et al. Readability of common health-related quality-of-life instruments in female pelvic medicine. Female Pelvic Med Reconstr Surg. 2013;19:293-7.
15. Şahin MF, Ateş H, Keleş A, Özcan R, Doğan Ç, Akgül M, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: a comparative analysis. J Med Syst. 2024;48:38.
16. Tsiamtsiouris J, Kollia B, Kamowski-Shakibai MT, Garcia P, Basch, CH. Applying tests of readability to online stuttering information. Adv Neurodev Disord. 2020;4:279-83.
17. Temel MH, Erden Y, Bağcıer F. Information quality and readability: ChatGPT's responses to the most common questions about spinal cord injury. World Neurosurg. 2024;181:e1138-44.
18. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg. 2023;33:1790-6.
19. Akkara JD, Kuriakose A. Commentary: is human supervision needed for artificial intelligence? Indian J Ophthalmol. 2022;70:1138-9.