



RESEARCH ARTICLE

F.U. Med.J.Health.Sci.
2026; 40 (3): 239 - 246
<http://www.fusabil.org>

Mithat MIZRAK ^{1, a}
Sema MIZRAK ^{2, b}

¹ Fırat University,
Faculty of Medicine,
Department of
Endocrinology and
Metabolism,
Elazığ, TÜRKİYE

² Fethi Sekin City Hospital,
Department of Physical
Medicine and Rehabilitation,
Elazığ, TÜRKİYE

^a ORCID: 0000-0002-0453-6174

^b ORCID: 0000-0002-3065-4000

Evaluation of the Quality, Reliability and Readability of Artificial Intelligence Generated Information About Osteoporosis

The use of artificial intelligence (AI) chatbots, which have become increasingly prevalent in various aspects of daily life, is gradually rising as a source of information for patients. This study aimed to evaluate the quality, reliability, and readability of responses provided by ChatGPT-4o and Gemini to frequently asked questions about osteoporosis.

Forty-four questions about osteoporosis, commonly encountered in clinical practice and patient education guidelines, were posed to ChatGPT-4o and Gemini. Four experts specializing in osteoporosis evaluated the quality and reliability of the responses generated by the AI platforms. The readability of the responses was assessed using the Ateşman and Çetinkaya formulas.

Gemini scored significantly higher than ChatGPT in the global quality score (3.98 ± 0.82 vs. 3.40 ± 0.87 , $p < 0.001$). Gemini also achieved higher reliability scores compared to ChatGPT (5.85 ± 0.96 vs. 5.23 ± 1.14 , $p < 0.001$). The DISCERN scores indicated "good quality" for both platforms (ChatGPT: 51.50 ± 2.06 , Gemini: 59.75 ± 4.71 , $p = 0.152$). Regarding readability, responses from both AI platforms were classified as difficult and required at least a high school level of education.

This study presents data on the usability of artificial intelligence platforms for osteoporosis patient education. The challenging readability levels of both platforms present limitations for patients. The current findings suggest that artificial intelligence tools may serve as a complementary resource in osteoporosis patient education when used under the supervision of healthcare professionals.

Key Words: Osteoporosis, artificial intelligence, ChatGPT, Gemini, patient education

Osteoporoz Hakkında Yapay Zeka Tarafından Üretilen Bilgilerin Kalitesinin, Güvenilirliğinin ve Okunabilirliğinin Değerlendirilmesi

Hayatımızın her alanında önemli yer edinen yapay zeka destekli sohbet robotlarının günümüzde hastalar tarafından bir bilgi kaynağı olarak kullanımı giderek artmaktadır. Osteoporoz hakkında sıkça sorulan sorulara ChatGPT-4o ve Gemini tarafından verilen yanıtların kalite, güvenilirlik ve okunabilirliğini değerlendirmek amaçlanmıştır.

Osteoporoz konusunda belirlenen klinik pratikte ve hasta eğitim rehberlerinde sık rastlanılan 44 soru ChatGPT-4o ve Gemini'ye soruldu. Yapay zeka tarafından verilen yanıtlar kalite ve güvenilirlik açısından osteoporoz alanında çalışan 4 uzman tarafından değerlendirildi. Yanıtların okunabilirliği Ateşman ve Çetinkaya formüllerine göre değerlendirildi.

Gemini, global kalite skoru açısından ChatGPT'ye göre anlamlı derecede yüksek puan aldı (3.98 ± 0.82 vs. 3.40 ± 0.87 , $p < 0.001$). Güvenilirlik indeksi değerlendirmesinde de Gemini üstün bulundu (5.85 ± 0.96 vs. 5.23 ± 1.14 , $p < 0.001$). DISCERN skorları her ikisi için de "iyi kalite" kategorisinde yer aldı (ChatGPT: 51.50 ± 2.06 , Gemini: 59.75 ± 4.71 , $p = 0.152$). Okunabilirlik açısından her iki yapay zekanın verdiği yanıtlar da zor okunabilirlik düzeyinde bulundu ve en az lise seviyesinde eğitim gerektirdiği saptandı.

Bu çalışma, yapay zeka platformlarının osteoporoz hasta eğitiminde kullanılabilirliği konusunda veriler sunmaktadır. Her iki platformun da zor okunabilirlik düzeyi, hastalar için sınırlılık oluşturmaktadır. Mevcut bulgular, yapay zeka araçlarının sağlık profesyoneli gözetiminde osteoporoz konusunda hasta eğitiminde tamamlayıcı bir kaynak olarak kullanılabileceğini göstermektedir.

Anahtar Kelimeler: Osteoporoz, yapay zeka, ChatGPT, Gemini, hasta eğitimi

Received : 31.01.2026
Accepted : 01.06.2026

Correspondence

Mithat MIZRAK
Fırat University,
Faculty of Medicine,
Department of
Endocrinology and
Metabolism,
Elazığ - TÜRKİYE

mithat.mizrak@gmail.com

Introduction

Osteoporosis is a progressive metabolic bone disorder characterized by reduced bone mineral density and deterioration of bone microarchitecture. Globally, it affects approximately one-third of women and one-fifth of men aged >50 years. In Türkiye, osteopenia and osteoporosis have been identified in 50% and 25% of individuals aged ≥ 50 years, respectively (1). Osteoporosis can lead to bone structure weakening and fragility fractures. Fractures can occur with minimal or no trauma. Osteoporosis remains asymptomatic until fractures develop. A significant majority of patients remain undiagnosed and do not receive treatment until a fracture occurs (2). Each fracture is a

predictor of other new fractures that may develop in different areas, and multiple fractures can impair the quality of life and increase mortality. Osteoporosis-related fractures not only increase mortality and morbidity but also result in significant healthcare expenditures (3, 4). The prevalence of osteoporosis and related fractures has increased substantially worldwide owing to rapid population aging and changes in lifestyle habits. The increasing frequency of osteoporotic fractures further aggravates the public health problems arising from the prevalence of osteoporosis, intensifying its overall impact on public health. Treatment and care costs impose an excessive financial burden not only on individuals but also on governments' social security institutions and healthcare expenditure. Diagnosing osteoporosis, which is insufficiently recognized and generally diagnosed late, before fractures occur, and preventing health problems that would be caused by fractures through necessary precautions and medical treatments, is of critical importance.

Patient education is essential for the management of osteoporosis, as it significantly contributes to increasing disease awareness, understanding risk factors, maintaining treatment compliance, and promoting lifestyle adjustments. Informed patients have higher treatment adherence and a reduced fracture risk. In recent years, a growing number of people have been using the Internet to search for information related to health. These resources play a role in raising awareness about public health, and many patients obtain information about health problems from Internet sources such as Instagram, Twitter, and YouTube. Recently, artificial intelligence (AI) chatbots have become more popular. Many people use Large Language Models (LLMs), such as OpenAI's ChatGPT and Google's Gemini. They provide information on different areas, including health advice. Although AI is widely employed to generate health information, there is a notable lack of sufficient oversight, leading to concerns regarding the quality, accuracy, and reliability. Inaccurate or incomplete data can pose significant health risks to patients, disrupt adherence to treatment plans, and result in adverse outcomes for patients.

In recent years, there has been a notable increase in research evaluating the quality of responses generated by Generative AI platforms for medical subjects. However, studies that specifically compare the performance of various LLMs remain scarce. This study was designed to assess the quality, reliability, and readability of responses provided by the ChatGPT and Gemini AI platforms to common inquiries found in clinical settings and patient guidelines related to osteoporosis. The outcomes of this study will provide evidence-based insights into the effectiveness of AI-supported information on osteoporosis for patient education.

Materials and Methods

Research and Publication Ethics: This study was designed to compare the quality, reliability, and readability of responses provided by AI chatbots on medical topics and was conducted between September and December 2024. In this study, the responses generated by two different AI platforms (ChatGPT-4o and Gemini) to frequently asked questions about osteoporosis were evaluated by an expert panel. Since the study did not involve human or animal subjects, ethics committee approval was not required. This study is based on the evaluation of publicly available responses from AI platforms.

Patient education questions related to osteoporosis were developed using reliable websites frequently visited by patients, hospital websites, national and international osteoporosis associations (International Osteoporosis Foundation, Turkish Society of Endocrinology and Metabolism) and Google Trends data. The question pool was reviewed and finalized by two specialist physicians (an endocrinologist and a physical medicine and rehabilitation specialist) to ensure clinical relevance and comprehensiveness. Through this process, a total of 44 questions were developed covering a broad range of topics including osteoporosis pathophysiology, risk factors, diagnostic methods, treatment options, nutrition, exercise, lifestyle modifications and complications. The complete list of 44 questions is provided as Supplementary Material (Table 1).

Table 1. List of 44 questions posed to ChatGPT-4o and Gemini AI platforms regarding osteoporosis

No	Question
1	What is osteoporosis?
2	What are the risk factors for osteoporosis?
3	What are the symptoms of osteoporosis?
4	How is osteoporosis diagnosed?
5	Under what circumstances should bone mineral density be measured?
6	How often should bone mineral density measurement be performed?
7	What are the treatment options for osteoporosis?
8	What are the non-pharmacological treatment options for osteoporosis?
9	Which foods are beneficial for osteoporosis?

No	Question
10	What are the grades of osteoporosis?
11	Can osteoporosis be treated?
12	Is osteoporosis life-threatening?
13	Is osteoporosis a disease of the elderly?
14	At what age does osteoporosis begin?
15	Can osteoporosis occur in young individuals?
16	Which diseases cause osteoporosis?
17	Is osteoporosis common in men?
18	Why is osteoporosis more prevalent in women?
19	What should the diet consist of in osteoporosis?
20	What should the daily calcium and vitamin D intake be?
21	Which foods are rich in calcium?
22	Does osteoporosis cause pain?
23	What is the osteoporosis injection and what are its side effects?
24	What are the exercise recommendations for osteoporosis?
25	What lifestyle modifications are recommended to prevent osteoporosis?
26	What are the side effects of medications used in osteoporosis treatment?
27	Is there a role for orthotic bracing (corset) in osteoporosis management?
28	Is low body weight a risk factor for osteoporosis?
29	Should bone density screening be performed in underweight individuals?
30	What are the surgical treatment options in osteoporosis?
31	What fall prevention measures should a patient with osteoporosis take?
32	Which medical specialty should a patient with osteoporosis consult?
33	I became pregnant while on osteoporosis treatment. Could the medications adversely affect my pregnancy?
34	I have osteoporosis. Which foods should I avoid?
35	Should osteoporosis medications be discontinued prior to dental procedures?
36	Is osteoporosis reversible?
37	My mother has a hip fracture. What is my fracture risk?
38	Does osteoporosis cause loss of height?
39	I sustained a forearm fracture. What is the likelihood of sustaining another fracture?
40	Do proton pump inhibitors cause osteoporosis?
41	Does breastfeeding cause osteoporosis?
42	I have osteoporosis. Should I quit smoking?
43	I have osteoporosis. Should I quit alcohol?
44	I am 30 years old and have kyphosis. Should I have my bone mineral density measured?

Artificial Intelligence Chatbots and Data Collection: In this study, two different AI chatbots were used: OpenAI's ChatGPT-4o model and Google's Gemini. The 44 prepared questions were posed to both platforms in a standard format and with the same expressions. Each question was asked on both platforms once, and the first responses received were recorded. Additionally, to ensure consistency, a new chat session was initiated for each question. The responses did not contain any information indicating the AI platform from which the response was obtained. The evaluators scored by viewing only the questions and related responses.

Readability Analysis: Two different formulas exist for analyzing the readability of Turkish texts: the Ateşman readability formula developed by Ateşman in 1997 [Ateşman readability formula: $198.825 - (40.175 \times \text{average syllables per word}) - 2.610 (\text{average words per sentence})$] (5) and the Çetinkaya readability formula developed by Çetinkaya in 2010 [Çetinkaya readability formula: $118.823 - (25.987 \times \text{average word length}) - (0.971 \times \text{average sentence length})$] (6). According to the Ateşman formula, the readability categories are: 1-29: very difficult; 30-49: difficult; 50-69: moderately difficult; 70-89: easy; and 90-100: very easy. According to the Çetinkaya formula, the readability categories are: 0-34: insufficient reading level, corresponding to 10th-12th grade; 35-50: educational reading level, corresponding to 8th-9th grade; and ≥ 51 , independent reading level, corresponding to 5th-7th grade.

Expert Panel and Evaluation Scales: Experts with at least 5 years of experience in physical medicine and rehabilitation and endocrinology were selected to evaluate the responses provided by AI-powered chatbots on osteoporosis. Each expert scored all 88 responses, comprising two responses for each question.

The responses generated by the AI platforms were evaluated using three different scales: the global quality score (Table 2), which assesses the overall quality, scientific flow, and usability for patients; the reliability index (Table 3), a 7-point Likert scale that evaluates the potential to guide patients correctly or incorrectly; and the DISCERN scale, which analyzes the overall quality of health information. The global quality score has been used in various studies to evaluate the overall quality of information sources. This scoring is rated between 1 and 5 points, and as it progresses from 1 to 5, the quality of information and its usability by patients increases (7). DISCERN was developed as a tool to analyze the quality of health information and is one of the most widely used tools for evaluating written health information. It has been used in many studies to evaluate health education materials generated by AI (8, 9). It consisted of three sections and a total of 16 questions. Each question was scored on a 1-5 scale. The DISCERN scores were classified as follows: excellent (scores 64-80), good (scores 52-63), fair (scores 41-51), poor (scores 30-40), and very poor (scores 16-29) (10, 11).

Statistical Analysis: All statistical analyses were performed using the SPSS software (version 27.0; IBM, New York, USA). Descriptive statistics are presented as mean±standard deviation, median, minimum, and maximum values of the data. The normality of the data distribution was assessed using the Shapiro-Wilk test. To compare the Global Quality Score, Reliability Index, and DISCERN scores of the responses generated by the ChatGPT-4o and Gemini platforms, a paired t-test was used for normally distributed data, and the Wilcoxon signed-rank test was used for non-normally distributed data. The intraclass correlation coefficient (ICC) was calculated for the inter-rater reliability assessment. Statistical significance was set at $p < 0.05$.

Table 2. Global Quality Scale

Score	Quality Level	Description
5	Excellent Quality	Excellent quality and flow; highly comprehensive and very useful for patients.
4	Good Quality	Good quality and flow; covers most relevant headings but has some minor gaps. Suitable for patients.
3	Moderate Quality	Moderate quality with suboptimal flow; some key points are discussed sufficiently, while others are inadequate. Partially useful for patients.
2	Poor Quality	Generally low quality with weak flow; provides some information but misses many important headings. Very limited use for patients.
1	Very Poor Quality	Low quality with weak flow; most information is missing. Not useful for patients at all.

Table 3. Rating of ChatGPT and Gemini's responses using a seven-point Likert scale

Score	Reliability Level	Description
7	Completely Accurate	Guides the patient completely correctly.
6	Mostly Accurate	Guides the patient correctly to a large extent.
5	Partially Accurate	Guides the patient partially correctly.
4	Neutral	Neither guides the patient correctly nor incorrectly.
3	Partially Misleading	Guides the patient partially incorrectly.
2	Highly Misleading	Guides the patient incorrectly to a large extent.
1	Completely Misleading	Guides the patient entirely incorrectly.

Results

The study included 44 questions related to osteoporosis that were posed to two AI platforms: ChatGPT-4o and Gemini. A total of 88 responses (44 questions×2 platforms) were evaluated by three physical medicine and rehabilitation specialists and one endocrinology specialist. Evaluations were conducted using the global quality score (1-5), reliability index (1-7 Likert scale) and DISCERN scale (16-80). In terms of the global quality score, the Gemini platform received a higher mean score than ChatGPT (3.98 ± 0.82 vs. 3.40 ± 0.87). The Gemini platform also demonstrated superiority in the reliability index evaluation. Gemini's mean reliability score was 5.85 ± 0.96 , while ChatGPT's was calculated as 5.23 ± 1.14 . In the overall quality evaluation using the DISCERN scale, the mean DISCERN score for ChatGPT was 51.50 ± 2.06 , whereas Gemini's was 59.75 ± 4.71 . Although both platforms were in the "good quality" category, Gemini scored 8.25 points (16%) higher. In comparisons of the global quality score and reliability index, Gemini received significantly higher scores than ChatGPT ($p < 0.001$). In the DISCERN score comparison, although Gemini scored numerically higher than ChatGPT, this difference was not statistically significant ($p = 0.152$).

The agreement between the scores of the four expert evaluators was analyzed using the Intraclass Correlation Coefficient (ICC). The ICC value for the global quality score was 0.699, and the ICC value for the reliability index was 0.729. Inter-rater reliability was in the "good reliability" category for both scoring systems. (ICC < 0.40, poor; 0.40-0.59, fair; 0.60-0.74, good; 0.75-1.00, excellent).

Readability analysis using the Ateşman formula revealed that ChatGPT responses had a mean score of 40.45 ± 11.58 , corresponding to a difficult readability level. The Çetinkaya formula showed a mean score of 27.35 ± 5.13 , indicating a 10th-12th grade reading level requirement. For Gemini, the mean Ateşman score was 48.31 ± 5.02 (difficult readability) and the Çetinkaya score was 30.46 ± 2.71 (10th-12th grade level) (Table 4).

According to the Ateşman formula, 88.6% and 59.1% of ChatGPT and Gemini's responses, respectively, were in the difficult and very difficult readability categories. None of the responses provided by either AI platform were in the easy or very easy categories according to the Ateşman formula. Of the 44 questions, 18 (40.9%) had moderate readability levels in Gemini, whereas only five questions (11.4%) had moderate readability levels in ChatGPT (Table 5).

The Çetinkaya formula revealed that 97.7% of responses from both platforms required high school-level reading proficiency (0-34 points, 10th-12th grade), limiting accessibility for patients with lower educational attainment (Table 6).

In summary, the Gemini platform demonstrated a statistically significantly superior performance compared with ChatGPT-4o in terms of both global quality score and reliability index in response to patient questions about osteoporosis. In the overall quality evaluation using the DISCERN scale, both platforms were in the "good quality" category. In terms of readability, both AI platforms exhibited a difficult readability level, and the texts required a high school-level education for comprehension.

Table 4. Readability scores according to Ateşman and Çetinkaya formulas

Readability	ChatGPT	Category	Gemini	Category	p
Ateşman	40.45 ± 11.58	Difficult (30-49)	48.31 ± 5.02	Difficult (30-49)	<0.001
Çetinkaya	27.35 ± 5.13	Insufficient (0-34, High School)	30.46 ± 2.71	Insufficient (0-34, High School)	<0.001

Table 5. Readability categories of responses according to Ateşman formula

Category	ChatGPT (n)	ChatGPT (%)	Gemini (n)	Gemini (%)
Very Easy	0	0.0%	0	0.0%
Easy	0	0.0%	0	0.0%
Moderate	5	11.4%	18	40.9%
Difficult	35	79.5%	26	59.1%
Very Difficult	4	9.1%	0	0.0%

Table 6. Readability categories according to Çetinkaya formula

Category	ChatGPT (n)	ChatGPT (%)	Gemini (n)	Gemini (%)
Independent (5th-7th grade)	0	0.0%	0	0.0%
Educational (8th-9th grade)	1	2.3%	1	2.3%
Insufficient (10th-12th grade)	43	97.7%	43	97.7%

Discussion

The emergence of LLMs such as ChatGPT and Gemini has increased interest in their potential applications for healthcare, including patient education. While several studies have evaluated the medical information quality provided by AI chatbots, comparative studies between different platforms remain limited. To our knowledge, this is the first study to compare the quality, reliability, and readability of Turkish language responses from ChatGPT-4o and Gemini regarding osteoporosis patient education.

Artificial intelligence technologies are increasingly being utilized in healthcare, including patient education. Patients have increasingly preferred LLMs over internet sources to learn health information. Despite their significant advantages, their lack of a controlling mechanism and their ability to provide responses that appear convincing but are not based on fact (hallucinations) are among their negative features. As incorrect or incomplete information can lead to serious problems for patient health or affect decision-making processes related to health conditions, the effectiveness, accuracy, and reliability of information provided by AI must be investigated. Various results have been reported in studies evaluating AI responses on osteoporosis in the literature. While some studies find AI responses to be of quality and reliable, others report that these responses may contain incorrect or incomplete information and recommend a cautious approach. In one of these studies conducted in 2024, 20 questions about osteoporosis were posed to ChatGPT, and the responses were evaluated by different experts. In this study, 91% of the responses provided by ChatGPT were accurate, with no harmful content (12). In another comparative study evaluating responses provided by AI platforms BingAI, Gemini, and ChatGPT-4 to 25 frequently asked questions about osteoporosis, BingAI and Gemini performed significantly better than ChatGPT-4 in terms of quality. In this study, in terms of completeness and accuracy of responses, all three AI chatbots provided accurate and complete responses regarding osteoporosis and demonstrated similar performance (13). In another study related to osteoporosis, 25 questions were asked to three different AI platforms, and the responses were scored by five orthopedic specialists according to a 5-point Likert scale. The median scores of the AI platforms were 4.6 for ChatGPT-4o, 4.4 for Gemini-2.5 Pro, and 4.2 for DeepSeek-R1, emphasizing that all three AI platforms could be used for patient education (14). In a study in which experts evaluated the responses of ChatGPT-4 to 100 different clinical presentations created from 206 patients diagnosed with osteoporosis, ChatGPT demonstrated high diagnostic accuracy, with 91% of responses being completely correct and comprehensive (15). In glucocorticoid-induced osteoporosis, Gemini provided superior general responses while ChatGPT-4 excelled at guideline-specific questions (16). In a study evaluating the quality of different AI platforms for postmenopausal osteoporosis, the quality of responses provided by ChatGPT-4o was generally found to be

more successful than the other two models (ChatGPT-4o mini and Gemini) (17). Contrary to the positive studies on the use of AI in patient education, conflicting studies also exist in the literature. In one study, responses provided by ChatGPT-3.5 and ChatGPT-4 to four different clinical cases regarding osteoporosis were compared with treatment plans recommended by seven different rheumatologists based on local guidelines. While ChatGPT-4 demonstrated 80% agreement with the recommendations made by rheumatologists, ChatGPT-3.5 showed 65% agreement. This study emphasized the importance of expert opinion, particularly in complicated cases with comorbidities (18). In a study where adequate and appropriate responses were provided to questions related to osteoporosis at a rate of 80.6%, the accuracy rate of responses to questions based on the National Osteoporosis Guideline Group (NOGG) guidelines decreased to 61.3% (19). In another study conducted in 2024, related to osteoporosis, the quality of information was evaluated. In this study, the average score was 48.71%, and the quality of responses generated by ChatGPT regarding osteoporosis showed notable deficiencies. Responses were generally evaluated in the "serious problems in terms of quality" category (20). Conflicting results exist in the literature regarding the presentation of osteoporosis information using AI platforms. In addition to positive evaluations, studies have emphasized that responses may be incorrect or incomplete. In our study, responses provided regarding osteoporosis were generally found to be of quality and reliable for both platforms. Although Gemini's statistical superiority over ChatGPT-4o in terms of both global quality score (3.98 ± 0.82 vs. 3.40 ± 0.87 , $p < 0.001$) and reliability index (5.85 ± 0.96 vs. 5.23 ± 1.14 , $p < 0.001$) suggests that platform selection may be a factor in patient education, the fact that both platforms were in the "good quality" category in the evaluation using the DISCERN scale (ChatGPT: 51.50 ± 2.06 , Gemini: 59.75 ± 4.71 , $p = 0.152$) demonstrates that both tools meet acceptable standards in overall information presentation. Differences between study results may have arisen from the continuous updating and development of AI models, differences in the content of question sets used and evaluation methodologies, or hallucinations.

Readability is crucial for patient comprehension of health information and treatment adherence. It is important that medical content is written in simple language. Scales exist to evaluate the readability of medical texts. Most readability evaluations in the literature have been conducted on English texts. However, studies examining the readability of responses provided in local languages to patients' questions about their health conditions are also needed. The Ateşman and Çetinkaya formulas were used to test Turkish text readability. Due to the increasing use of Internet resources for health consultations, ensuring that AI tools such as ChatGPT and Gemini present readable and understandable information is important. In the literature, the number of studies evaluating the readability of Turkish responses in studies on the usability of AI in

patient education is limited. In a study conducted in 2025 that evaluated the responses provided by ChatGPT to the most frequently asked patient questions about colorectal cancer, the readability level of the responses was generally found to be understandable by the public (21). In another study evaluating the readability of responses to frequently asked patient questions about glaucoma, the readability status was found to be at a difficult level (22). In a study related to myasthenia gravis, the readability level was found to be moderately difficult (23). Our analysis revealed difficult readability levels for both platforms according to both the Ateşman and Çetinkaya formulas. Çetinkaya analysis revealed that 97.7% of responses from both platforms fell within the insufficient readability range (0-34 points), requiring 10th-12th grade educational attainment. The difficult readability level of both platforms demonstrates the need for improvement in content production suitable for the target audience. Optimizing AI chatbots to generate texts at the recommended 6th-grade reading level (24) and their ability to provide clear and understandable answers to questions about their diseases will increase the benefit of AI in health education.

The accessibility of AI chatbots to everyone is increasing patients' frequency of using these platforms to obtain information about their diseases. Increasing patients' knowledge of osteoporosis management and osteoporotic fractures can positively affect treatment

adherence and clinical outcomes. In our study, both LLMs generally provided responses of acceptable quality regarding osteoporosis. The high readability difficulty demonstrated by both platforms and their requirement for high school-level education emphasize the limitations of these resources. This situation may create a barrier to accessing this information, particularly for patient groups with low education levels.

This study has several limitations. First, the questions were posed to the AI platforms exclusively in Turkish; therefore, the findings may not be generalizable to responses generated in other languages. Second, both AI platforms are continuously updated, and the quality and reliability of responses may vary across different model versions; thus, the findings reflect the performance of the platforms at the time of the study. Third, although the 44-question set was designed to cover a broad range of osteoporosis-related topics, it may not encompass all relevant clinical scenarios.

In conclusion, current findings demonstrate that artificial intelligence tools can be used as a complementary resource in osteoporosis patient education; however, due to the potential for hallucination and presentation of outdated information by AI platforms, this information must be evaluated under the supervision of healthcare professionals and supported by expert opinions.

References

1. Tuzun S, Eskiurt N, Akarımak U, et al. Incidence of hip fracture and prevalence of osteoporosis in Turkey: The FRACURK study. *Osteoporos Int* 2012;23(3):949-955.
2. Cosman F, de Beur SJ, LeBoff MS, et al. Clinician's guide to prevention and treatment of osteoporosis. *Osteoporos Int* 2014;25(10):2359-2381.
3. Bianchi ML, Orsini MR, Saraifoger S, et al. Quality of life in post-menopausal osteoporosis. *Health Qual Life Outcomes* 2005;3(1):78.
4. Melton LJ, Achenbach SJ, Atkinson EJ, Thorneau TM, Amin S. Long-term mortality following fractures at different skeletal sites: A population-based cohort study. *Osteoporos Int* 2013;24(5):1689-1696.
5. Ateşman E. Türkçede okunabilirliğin ölçülmesi. *Dil Dergisi* 1997;58:71-74.
6. Çetinkaya G. Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması [dissertation]. Ankara: Ankara Üniversitesi; 2010.
7. Onder CE, Koc G, Gokbulut P, Taskaldıran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep* 2024;14(1):243.
8. Ghanem YK, Rouhi AD, Al-Houssan A, et al. Dr. Google to Dr. ChatGPT: Assessing the content and quality of artificial intelligence-generated medical information on appendicitis. *Surg Endosc* 2024;38(5):2887-2893.
9. Kane K, Rakauskas T, Eng E, et al. Quality of ChatGPT responses to common patient questions on Dupuytren contracture. *J Orthop Exp Innov* 2025;6(1).
10. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999;53(2):105-111.
11. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. *Int J Med Inform* 2025;198:105871.
12. Ghanem D, Shu H, Bergstein V, et al. Educating patients on osteoporosis and bone health: Can "ChatGPT" provide high-quality content? *Eur J Orthop Surg Traumatol* 2024;34(5):2757-2765.
13. 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* 2025;31(3):162-168.
14. Li X, Li G, Zhao Y, et al. Exploring and comparing the use of large language models in supporting osteoporosis health consultations. *Clin Interv Aging* 2025;20:2133-2143.
15. Bucak ÖF, Cinar C. The role of ChatGPT in osteoporosis management: A comparative analysis with clinical expertise. *Arch Osteoporos* 2025;20(1):51.
16. Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. *J Orthop Surg Res* 2024;19(1):574.
17. Liu R, Liu J, Yang J, Sun Z, Yan H. Comparative analysis of ChatGPT-4o mini, ChatGPT-4o and Gemini Advanced

- in the treatment of postmenopausal osteoporosis. *BMC Musculoskelet Disord* 2025;26(1):369.
18. Zniber S, Drari S, Bouia FZ, Mougui A, El Bouchti I. Therapeutic approaches to osteoporosis: A comparison between ChatGPT AI and seven rheumatologists. *Ann Rheum Dis* 2025;84:1390.
 19. Cinar C. Analyzing the performance of ChatGPT about osteoporosis. *Cureus* 2023;15(9):e45890.
 20. Erden Y, Temel MH, Bağcıer F. Artificial intelligence insights into osteoporosis: Assessing ChatGPT's information quality and readability. *Arch Osteoporos* 2024;19(1):17.
 21. Erdat EC, Yalçınır M, Utkan G. Readability and quality analysis of ChatGPT o1's responses on colorectal cancer: A study of an AI language model. *Acta Haematol Oncol Turc* 2025;58(2):74-80.
 22. Durmuş Ece BŞ, Hondur G, Yüksel E. Glokomun medikal tedavisi ile ilgili sorulara ChatGPT yanıtlarının doğruluk, kapsamlılık, okunabilirlik açısından değerlendirilmesi. *MN Oftalmoloji* 2025;32(4):239-245.
 23. İnan B, Karadaş Ö, Odabaşı Z. Evaluation of ChatGPT-4o's responses to questions about myasthenia gravis in English and Turkish. *Med J Bakirkoy* 2025;21(3):310-315.
 24. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: A 20-year analysis. *J Patient Exp* 2021;8:2374373521998847.