ORIGINAL ARTICLE

DOI: 10.4274/csmedj.galenos.2026.2026-9-7

Cam and Sakura Med J

CSMJ

Comparing ChatGPT's Performance in Osteoporotic Fracture Risk Assessment with FRAX

Gonca Yazıcı Saçaklıdır[1], Rekib Saçaklıdır[2]

[1]University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Physical Medicine and Rehabilitation, İstanbul, Türkiye

[2]University of Health Sciences Türkiye, Şişli Hamidiye Etfal Training and Research Hospital, Clinic of Physical Medicine and Rehabilitation, Pain Management Section, İstanbul, Türkiye

What is known on this subject?

Osteoporotic fracture risk assessment is a significant health issue in elderly patients. The risk of hip and major osteoporotic fractures can be calculated using Fracture Risk Assessment Tool (FRAX).

What this study adds?

A moderate level of agreement was observed between ChatGPT and FRAX regarding the risk of major osteoporotic fractures; however, ChatGPT did not show full agreement with FRAX data concerning hip fracture risk. These findings do not support the direct use of ChatGPT as a substitute for FRAX.

ABSTRACT

Objective: Fracture Risk Assessment Tool (FRAX) is frequently used for osteoporotic fracture risk assessments. The aim of this study is to evaluate the correlation between ChatGPT and FRAX fracture risk measurements.

Material and Methods: In June 2026, 150 synthetic patient profiles were generated under the supervision of two physicians. These patient profiles were evaluated for major osteoporotic fracture risk (MOFR) and hip fracture risk using FRAX and ChatGPT 5.5. Furthermore, the correlation and intraclass correlation coefficient (ICC) between FRAX and ChatGPT were calculated.

Results: No statistically significant difference was found between FRAX and ChatGPT in terms of 10-year MOFR ($p=0.150$). However, ChatGPT estimated significantly lower 10-year hip fracture risk than FRAX ($p<0.001$). Pearson correlation analysis showed a moderate positive correlation between ChatGPT and FRAX for MOFR ($r=0.649$, $p<0.001$), and a strong positive correlation for hip fracture risk ($r=0.701$, $p<0.001$). The ICC showed moderate agreement for MOFR [ICC: 0.687, 95% confidence interval (CI): 0.569–0.773] and for hip fracture risk (ICC: 0.509, 95% CI: 0.322–0.644).

Conclusion: A moderate level of agreement was found between ChatGPT and FRAX for MOFR; however, ChatGPT did not fully agree with FRAX regarding hip fracture risk. While these data do not support the use of ChatGPT instead of FRAX, evaluating them with real patient data is appropriate.

Keywords: Osteoporosis, bone fractures, risk assessment, artificial intelligence, ChatGPT

Introduction

Osteoporosis is a chronic skeletal disorder in which compromised bone strength results from reduced bone mass and impaired bone tissue quality; this condition increases susceptibility to bone fractures (1). With the increasing aging population worldwide, osteoporosis and related

Corresponding Author: Prof. Rekib Saçaklıdır, MD, University of Health Sciences Türkiye, Şişli Hamidiye Etfal Training and Research Hospital, Clinic of Physical Medicine and Rehabilitation, Pain Management Section, İstanbul, Türkiye

E-mail: rakipsacakli@hotmail.com **ORCID ID:** orcid.org/0000-0001-6786-5306

Received: 02.10.2026 **Accepted:** 03.11.2026 **Epub:** 23.11.2026

Cite this article as: Yazıcı Saçaklıdır G, Saçaklıdır R. Comparing ChatGPT's performance in osteoporotic fracture risk assessment with FRAX. Cam and Sakura Med J. [Epub Ahead of Print]

Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of the Basaksehir Cam & Sakura City Hospital.
This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

fragility fractures have become a significant public health problem (2). Fractures caused by osteoporosis lead to problems such as immobility, pain, and social isolation. Therefore, predicting and preventing fractures that may develop due to osteoporosis is crucial (3).

Bone mineral density (BMD) serves as an important clinical measure for evaluating skeletal status and estimating an individual's susceptibility to fractures (4). However, the patient's age, history of smoking and alcohol use, family history of fractures, prior fractures, medications, and illnesses are also important factors in assessing fracture risk. Therefore, calculation methods that include clinical risk factors are used to evaluate fracture risk. The Fracture Risk Assessment Tool (FRAX) is one of the most commonly used assessment methods (5,6). The FRAX algorithm combines patient-specific demographic data, body mass index (BMI), clinical factors, and femoral neck BMD to provide an estimate of the 10-year risk for major osteoporotic fracture risk (MOFR) and hip fractures (7).

Artificial intelligence tools are also frequently used for osteoporosis assessment (8,9,10). Ghanem et al. (9) stated that ChatGPT-3.5 offers accurate and high-quality educational content consistent with expert opinions on osteoporosis. Conversely, Erden et al. (10) stated that ChatGPT's current quality and readability are insufficient to meet comprehensive healthcare standards. To our knowledge, no studies have directly compared the risk of osteoporotic fractures estimated by ChatGPT and FRAX. Therefore, the aim of the present study was to compare fracture-risk predictions generated by ChatGPT with those from the FRAX calculator and to evaluate the level of agreement between FRAX and ChatGPT.

Material and Methods

Study Design

The present study was conducted in June 2026 by comparing FRAX and ChatGPT 5.5 data from a simulation dataset consisting of 150 synthetic patient profiles (11,12). Synthetic patient profiles were created by randomly selecting values from predefined ranges of variables that might be encountered under the supervision of two physicians. The profiles created were included in the study after their clinical consistency was reviewed by the researchers. Parameters considered in preparing the patient profiles included: age (50-90 years), gender, height (145-185 cm), body weight (45-110 kg), history of fragility fracture, history of hip fracture in a first-degree relative, current tobacco use, exposure to glucocorticoids, rheumatoid arthritis, secondary causes of osteoporosis, regular alcohol intake, and the femoral neck T-score. Since it does not contain human- or patient-identifying data, ethics committee approval and informed consent are not required.

FRAX Risk Assessment

The FRAX calculator for Turkey was used to calculate the 10-year risks of MOFR and hip fracture for each model.

ChatGPT Risk Assessment

All synthetic patient profiles were evaluated on the same day and within the same session, using the same standardized prompt and ChatGPT model version. "Based only on the information provided below, estimate the likelihood of sustaining a major osteoporotic fracture and a hip fracture within the subsequent 10 years, expressed as percentages. Do not provide explanations, recommendations or additional comments. Report only two risk values as percentages."

Statistical Analysis

Statistical analyses were performed using SPSS Version 22.0 (IBM Corp., Armonk, NY, USA). The normality of continuous variables was assessed using the Shapiro–Wilk test. Data are presented as mean ± standard deviation or median. Categorical variables were expressed as numbers (n) and percentages (%). The relationships between fracture risk and ChatGPT and FRAX were evaluated using Pearson's correlation analysis. Intraclass correlation coefficient (ICC) analysis was used to evaluate the absolute agreement between FRAX and ChatGPT. The mean difference between the risk percentages obtained from ChatGPT and FRAX was evaluated using a paired-sample t-test. The statistical significance level was set at two-sided $p<0.05$.

Results

A total of 150 synthetic patient profiles were created for the study. The mean age and BMI of these models were 69.56 ± 11.93 and 28.70±8.07, respectively. The mean femoral neck BMD (T-score) was -2.21±0.98. Women comprised the majority of participants in the study: 79 (52.7%) (Table 1). No significant difference was found between the FRAX-derived and ChatGPT-derived estimates with respect to 10-year MOFR (%) ($p=0.150$). However, the FRAX-derived estimates had a significantly higher 10-year hip fracture risk ($p<0.001$) (Table 2).

Estimates of MOFR from ChatGPT and FRAX showed a significant positive association ($r=0.649$, $p<0.001$). Significant positive associations were also identified between ChatGPT-derived MOFR and FRAX-derived hip fracture risk ($r=0.666$,

p<0.001) and between ChatGPT-derived hip fracture risk and FRAX-derived MOFR (r=0.658, p<0.001). For hip fracture risk, the corresponding estimates from ChatGPT and FRAX demonstrated a statistically significant correlation (r=0.701, p<0.001). Within the ChatGPT estimates, MOFR and hip fracture risks were significantly correlated (r=0.862, p<0.001). The correlation was strongest between the two fracture-risk estimates generated by FRAX, with MOFR and hip fracture risks showing a correlation coefficient of 0.902 (p<0.001) (Table 3) (13).

The ICC between ChatGPT and FRAX was moderate: 0.687 (0.569–0.773) for MOFR and 0.509 (0.322–0.644) for hip fracture risk (Table 4).

Table 1. Baseline characteristics of the 150 synthetic patient profiles

Variable	Value (n=150)
Age (years)	69.55 (50-90)
BMI (kg/m^2)	28.70±8.07
Femoral neck (T-score)	-2.21±0.98
Fracture history (n)	89 (59.3 %)
Parental hip fracture history (n)	75 (50.0 %)
Smoking (n)	66 (44.0 %)
Glucocorticoid use (n)	74 (49.3 %)
Rheumatoid arthritis (n)	84 (56.0 %)
Secondary osteoporosis (n)	78 (52.0 %)
Alcohol 3 or more units/day (n)	68 (45.3 %)
Sex (n)	
Male	71 (47.3 %)
Female	79 (52.7 %)

BMI: Body mass index

Discussion

The FRAX calculator has been a frequently used and easy method for osteoporosis assessment from the past to the present (5,14,15). ChatGPT, on the other hand, has become popular in recent years and has been used for osteoporosis assessment (10,16,17). A moderate agreement was observed between ChatGPT and FRAX regarding fracture risk. However, ChatGPT was associated with a significantly lower risk of hip fractures. These data do not support the direct replacement of FRAX with ChatGPT. Nevertheless, ChatGPT can be considered a supportive tool in clinical decision-making.

ChatGPT-3.5 has been shown to exhibit a high degree of accuracy in addressing questions related to osteoporosis and

Table 2. Comparison of fracture risk assessment results between FRAX and ChatGPT

	FRAX (n=150)	ChatGPT (n=150)	p value
Major osteoporotic fracture risk (%)	25.16±16.08	26.62±8.19	0.150
Hip fracture risk (%)	15.49±6.95	9.47±3.81	<0.001

Data are presented as mean ± standard deviation
FRAX: Fracture Risk Assessment Tool

Table 3. Pearson correlation analysis between ChatGPT and FRAX fracture risk estimates

	ChatGPT major osteoporotic fracture risk	ChatGPT hip fracture risk	FRAX major osteoporotic fracture risk	FRAX hip fracture risk
ChatGPT major osteoporotic fracture risk	1.000	0.862*	0.649*	0.666*
ChatGPT hip fracture risk	0.862*	1.000	0.658*	0.701*
FRAX major osteoporotic fracture risk	0.649*	0.658*	1.000	0.902*
FRAX hip fracture risk	0.666*	0.701*	0.902*	1.000

*: indicates p values <0.05
FRAX: Fracture Risk Assessment Tool

Table 4. Intraclass correlation coefficients between ChatGPT and FRAX fracture risk estimates

Variable	ICC (95% confidence interval)	p value
Major osteoporotic fracture risk	0.687 (0.569-0.773)	<0.001
Hip fracture risk	0.509 (0.322-0.644)	0.001

ICC: Intraclass correlation coefficient, FRAX: Fracture Risk Assessment Tool

to provide high-quality educational content (9). Güleç et al. (18) stated that although AI chatbots like GPT provide accurate and complete answers regarding osteoporosis, the readability of their content is insufficient. However, Erden et al. (10) stated that ChatGPT is insufficient for comprehensive healthcare delivery in terms of information quality and readability. While previous studies have evaluated the quality of information, the present study investigates the extent to which ChatGPT can approximate a validated fracture risk calculation. FRAX is an internationally validated assessment tool that estimates 10-year fracture risk by combining variables such as age, BMI, gender, clinical factors, and femoral neck BMD using specific mathematical algorithms. In contrast, ChatGPT is not an algorithm developed for mathematical calculation but rather a generative artificial intelligence system based on a large language model. Although ChatGPT can recognize relationships between clinical risk factors, it does not apply the computational model underlying the FRAX algorithm precisely. The moderate level of agreement observed in our study, especially the statistically lower risk of hip fractures, may stem from this fundamental difference.

No statistically significant difference was detected between FRAX-derived and ChatGPT-derived estimates for MOFR. However, ChatGPT-derived estimates of hip fracture risk were significantly lower than FRAX-derived estimates. This difference in the hip may be due to the complex interaction of age, BMD, and other risk factors. Hip fractures represent a major public health concern because they can result in substantial morbidity and mortality while imposing a considerable economic burden on healthcare systems (19,20). The low agreement in this area does not support the use of ChatGPT alone in clinical decision making. Underestimation of hip fracture risk may result in high-risk patients being incorrectly classified as low-risk, potentially delaying appropriate evaluation and preventive interventions. Although our experiment used simulated data and did not assess clinical outcomes, the results demonstrate the potential risks of using ChatGPT estimates in clinical settings. Hence, ChatGPT is not an adequate substitute for validated risk assessments.

Study Limitations

First, real-world patient data were not used. Second, FRAX is not a gold-standard measurement method for osteoporosis. Third, only one AI model and only one version were used, and there was no a priori sample size calculation. Despite these limitations, a strength of our study is that it is the first to compare ChatGPT with a powerful computational method such as FRAX.

Conclusion

Moderate agreement was found between ChatGPT and FRAX for fracture risk; however, ChatGPT did not fully agree with FRAX regarding hip fracture risk. While these data do not support the use of ChatGPT instead of FRAX, it is appropriate to evaluate ChatGPT's performance with real patient data.

Ethics

Ethics Committee Approval: Since it does not contain human- or patient-identifying data, ethics committee approval is not required.

Informed Consent: This study did not involve human participants or animals; therefore, informed consent was not required.

Footnotes

Authorship Contributions

Surgical and Medical Practices: G.Y.S., Concept: G.Y.S., Design: G.Y.S., R.S., Data Collection or Processing: G.Y.S., R.S., Analysis or Interpretation: G.Y.S., R.S., Literature Search: G.Y.S., R.S., Writing: G.Y.S., R.S.

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. Pouresmaeili F, Kamalidehghan B, Kamarehei M, Goh YM. A comprehensive overview on osteoporosis and its risk factors. Ther Clin Risk Manag. 2018;14:2029-2049.
2. Morin SN, Leslie WD, Schousboe JT. Osteoporosis: a review. JAMA. 2025;334:894-907.
3. Nuti R, Brandi ML, Checchia G, et al. Guidelines for the management of osteoporosis and fragility fractures. Intern Emerg Med. 2019;14:85-102.
4. Kanis JA, Johansson H, Harvey NC, McCloskey EV. A brief history of FRAX. Arch Osteoporos. 2018;13:118.
5. Schini M, Johansson H, Harvey NC, Lorentzon M, Kanis JA, McCloskey EV. An overview of the use of the Fracture Risk Assessment Tool (FRAX) in osteoporosis. J Endocrinol Invest. 2024;47:501-511.
6. Adami G, Biffi A, Porcu G, et al. A systematic review on the performance of Fracture Risk Assessment Tools: FRAX, DeFRA, FRA-HS. J Endocrinol Invest. 2023;46:2287-2297.
7. Kanis JA, McCloskey EV, Johansson H, et al. Predictive value of BMD for hip and other fractures: a meta-analysis to update FRAX. Osteoporos Int. 2026;37:1067-1085. Erratum in: Osteoporos Int. 2026;37:1439-1444.
8. Paderno A, Ataide Gomes EJ, Gilberg L, et al. Artificial intelligence-enhanced opportunistic screening of osteoporosis in CT scan: a scoping

review. Osteoporos Int. 2024;35:1681-1692. Erratum in: Osteoporos Int. 2024;35:1877.

9. 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;34:2757-2765.
10. Erden Y, Temel MH, Bağcıer F. Artificial intelligence insights into osteoporosis: assessing ChatGPT's information quality and readability. Arch Osteoporos. 2024;19:17.
11. Jeanson F, Farkouh ME, Godoy LC, Minha S, Tzuman O, Marcus G. Medical calculators derived synthetic cohorts: a novel method for generating synthetic patient data. Sci Rep. 2024;14:11437.
12. Pilgram L, El Kababji S, Liu D, El Emam K. Should we synthesize more than we need: impact of synthetic data generation for high-dimensional cross-sectional medical data. J Am Med Inform Assoc. 2025;32:1843-1854.
13. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24:69-71.
14. Kanis JA, Harvey NC, Johansson H, et al. A decade of FRAX: how has it changed the management of osteoporosis? Aging Clin Exp Res. 2020;32:187-196.
15. Kanis JA, McCloskey EV, Johansson H, Oden A, Ström O, Borgström F. Development and use of FRAX in osteoporosis. Osteoporos Int. 2010;21 Suppl 2:S407-S413.
16. Bucak ÖF, Cinar C. The role of ChatGPT in osteoporosis management: a comparative analysis with clinical expertise. Arch Osteoporos. 2025;20:51.
17. Choi JY, Han E, Yoo TK. Application of ChatGPT-4 to oculomics: a cost-effective osteoporosis risk assessment to enhance management as a proof-of-principles model in 3PM. EPMA J. 2024;15:659-676.
18. Güleç GG, Bulut ÖÖ, Bağıcer 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;31:162-168.
19. Veronese N, Maggi S. Epidemiology and social costs of hip fracture. Injury. 2018;49:1458-1460.
20. Leal J, Gray AM, Prieto-Alhambra D, et al. Impact of hip fracture on hospital care costs: a population-based study. Osteoporos Int. 2016;27:549-558.