

Blinded Expert Evaluation of Large Language Models for Osteoporosis Case Management in Older Adults: Impact of Direct Guideline Integration

Gokalp Kurthan AVLAGI

gokalpkurthan@gmail.com

Istanbul University

Seyda BILGIN

Istanbul University

Duygu OZATA

Istanbul University

Kubra CINGAR ALPAY

Istanbul University

Tugba KANDEMIR

Istanbul University

Tugce EMIROGLU GEDIK

Istanbul University

Deniz SUNA ERDINCER

Istanbul University

Research Article

Keywords: osteoporosis, artificial intelligence, older adults, guideline, large language model, clinical decision support

Posted Date: July 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10193882/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Osteoporosis management in older adults requires individualized decisions encompassing fracture risk, therapy choice, renal function, contraindications, polypharmacy, fall risk, and follow-up strategies. Although large language models may provide decision support, their reliability, safety, and guideline compliance in geriatric osteoporosis scenarios remain unclear. This study compared AI panels with and without guideline integration using cases and blinded assessments.

Methods

In this cross-sectional study, 25 hypothetical osteoporosis cases in adults aged ≥ 65 years were evaluated using ChatGPT 4.5, Gemini 3, guideline-loaded ChatGPT 4.5, and NotebookLM. From cases in five domains, 100 responses were obtained, with first answers recorded in sessions. Responses were assessed by two geriatrics specialists using a 5-point Likert scale for accuracy, guideline adherence, clarity, comprehensive geriatric approach, and safety/contraindication awareness. Hallucinations were recorded. Inter-rater agreement and panel comparisons were analyzed using weighted Cohen's kappa, ICC(2,1), Friedman, and Bonferroni-corrected Wilcoxon tests.

Results

The ChatGPT 4.5 base model achieved the highest quality score (22.76/25; 95% CI: 22.30–23.22), followed by Gemini 3 (21.80), NotebookLM + guideline (21.56), and ChatGPT 4.5 + guideline (21.42). Scores differed across panels (Friedman $\chi^2=18.377$; $p = 0.0004$). Bonferroni-corrected comparisons showed ChatGPT 4.5 base model was significantly superior to three panels. Analyses found differences only in scientific accuracy ($p = 0.0006$) and guideline compliance ($p < 0.0001$), not in clarity, holistic geriatric approach, or safety/contraindications awareness. Safety/contraindications awareness was lowest-rated criterion in base models. Hallucinations were detected in one Gemini 3 base-model case (4%).

Conclusions

This study demonstrates that although large language models support managing geriatric osteoporosis, they remain limited in safety and context-sensitive reasoning. While ChatGPT 4.5's performance aligns with findings, the lack of benefit from guideline uploading suggests unstructured presentation may be inadequate for integration. The models' limitations in safety, contraindication awareness, and holistic geriatric approaches underscore the need for clinical supervision. Although artificial intelligence can support decisions, safety, contraindication recognition, and expert supervision in individualized geriatric

management are necessary. Future studies should evaluate physician-supervised systems in clinical settings.

BACKGROUND

Osteoporosis is a systemic skeletal disease characterized by reduced bone mineral density and microarchitectural deterioration, leading to an increased risk of fractures. Systematic reviews based on WHO criteria estimate its global prevalence at approximately 19.7%, affecting nearly 200 million individuals worldwide [1]. The clinical burden is particularly pronounced in older adults, in whom osteoporotic fractures, especially hip fractures, are associated with morbidity, mortality, functional decline, dependency, and high healthcare costs. One-year mortality after hip fracture has been reported to range between 20% and 30% [2]. Consistent with global data, the FRACTURK study demonstrated that osteoporosis and hip fractures are major public health concerns among individuals aged > 50 years in Turkey [3].

Osteoporosis management in older adults requires individualized and multidimensional decision making. Clinicians must consider fracture risk, pharmacological treatment choice, treatment duration, renal function, secondary causes, contraindications, polypharmacy, fall risk, and follow-up strategies when prescribing bisphosphonates. Although clinical guidelines provide a structured framework, their complexity, periodic updates, and need for patient specific interpretation may limit their direct application in routine practice [4]. Therefore, geriatric osteoporosis represents a clinically relevant model for evaluating artificial intelligence based clinical decision support systems, as it requires the simultaneous integration of fracture prevention, medication safety, fall risk reduction, and functional preservation.

Large language models are increasingly used for medical information retrieval, clinical support, and patient education [5, 6]. Systems such as ChatGPT 4.5, Gemini 3, and NotebookLM may aid in clinical reasoning, but their outputs require systematic evaluation because of the risk of hallucinations [7, 8]. Hallucinations, defined as plausible but incorrect or unverifiable outputs, may lead to clinically relevant errors [9]. In older adults with osteoporosis, such errors may result in delayed treatment for high risk patients, inappropriate drug holidays, overlooked renal impairment or contraindications, missed medication related risks, and follow-up recommendations that are not supported by guidelines.

However, evidence regarding the performance of artificial intelligence in osteoporosis care remains limited [10]. Previous studies have evaluated the quality and readability of ChatGPT generated osteoporosis information and compared different large language models in the context of postmenopausal osteoporosis, glucocorticoid-induced osteoporosis, and treatment management [11–14]. These studies suggest that language models have the potential to generate osteoporosis related knowledge and aid clinical recommendation processes. However, significant methodological limitations persist, including the lack of evaluator blinding, insufficiently detailed independent assessment

procedures, and a predominant focus on general knowledge questions rather than case based scenarios that require complex geriatric clinical reasoning [15, 16].

Recently, retrieval augmented generation and document based guideline integration have been proposed as strategies to improve the clinical accuracy and guideline adherence of large language models [17]. However, the effects of these approaches may vary depending on the model, clinical context, integration method, and evaluation protocol [18]. A pragmatic and accessible approach is to directly upload relevant guideline documents into the chat interface. However, systematic comparisons of baseline models and guideline integrated models in older adult populations under blinded expert assessment remain scarce.

This study aimed to address this gap by comparing the performance of language models with and without guideline integration in geriatric osteoporosis management using structured case scenarios and blinded independent expert evaluations.

METHODS

Study design and reporting

This cross-sectional, methodological, and comparative study evaluated the performance of four artificial intelligence based approaches in managing osteoporosis cases using blinded expert assessment. No real patient data were used in this study. All cases were standardized hypothetical clinical scenarios developed by the researchers to assess clinical decision-making skills in the management of geriatric osteoporosis.

Case scenario development

Twenty-five clinical case scenarios related to osteoporosis were developed by two geriatric specialists who were not involved in the blinded scoring process. The scenarios were divided into five predefined groups based on clinical complexity and decision making difficulty (Table 1). All scenarios were structured to represent the clinical decision making processes involved in managing osteoporosis in adults aged ≥ 65 years. Variables such as age, comorbidities, concomitant medication use, renal function, fracture risk, functional status, and geriatric syndromes were designed to reflect the profiles of older adults commonly encountered in tertiary geriatric outpatient clinics. Each group consisted of five cases. No personally identifiable patient information was used in any of the scenarios. Each case was designed to cover the key decision areas of osteoporosis management, including diagnostic reasoning, treatment selection, guideline adherence, safety assessment, contraindication awareness, secondary causes of osteoporosis, comprehensive geriatric evaluation, fall risk assessment, and follow-up planning.

Table 1
Definition of the clinical case groups. All cases were designed for adults aged ≥ 65 years.

Group	Cases	Clinical Domain
Group 1	Cases 1–5	Classic postmenopausal osteoporosis (age ≥ 65 y) with borderline or gray-zone diagnostic and therapeutic decisions
Group 2	Cases 6–10	Geriatric challenges: polypharmacy management, fall risk, functional status, and individualized care
Group 3	Cases 11–15	Drug management problems: treatment switching, adverse effect management, treatment duration, and drug holiday decisions
Group 4	Cases 16–20	Secondary osteoporosis and comorbidity-related scenarios where underlying diseases complicate management
Group 5	Cases 21–25	Complications and clinical pitfalls including atypical fractures, treatment failure, and rare adverse events

Artificial intelligence panels and response generation

Each clinical case scenario was presented separately to four artificial intelligence panels (Table 2). The same standardized prompt structure and identical case content were used for all four panels in this study. To ensure comparability, prompts were submitted to all the panels during the same response generation period. The model names, versions, and access dates were recorded. To minimize contextual transfer and potential bias related to session memory, each case was evaluated in a new independent session. For each case–panel combination, only the first response was recorded. Responses were not regenerated, selected from among alternative responses, or subjected to any manual editing before expert evaluation. For the guideline integrated panels, the full Turkish text of the Turkish Society of Endocrinology and Metabolism Osteoporosis Guidelines 2025 was uploaded as a source document. These models were instructed to base their responses on the uploaded guidelines. The non-guideline integrated panels received the same clinical case prompts but without any uploaded guideline documents.

Table 2
Definition of artificial intelligence (AI) panels.

Panel	AI Approach	Description
Panel 1	ChatGPT 4.5 (baseline, no guideline)	Cases presented to ChatGPT 4.5 (OpenAI, February 2026) without any external source upload
Panel 2	Gemini 3 (baseline, no guideline)	Cases presented to Gemini 3 (Google DeepMind, February 2026) without any external source upload
Panel 3	ChatGPT 4.5 + Guideline	Full TES Osteoporosis Guideline uploaded to ChatGPT 4.5 (OpenAI, February 2026) before cases were presented
Panel 4	NotebookLM + Guideline	Full TES Osteoporosis Guideline uploaded to Google NotebookLM (Google, February 2026) before cases were presented

Workflow

The study workflow consisted of four stages. In the first stage, 25 hypothetical osteoporosis case scenarios were developed and divided into five clinical groups, with five cases in each. In the second stage, each case was presented to four artificial intelligence panels, yielding 100 artificial intelligence responses. In the third stage, all responses were anonymized and assigned random identification codes prior to expert evaluation. In the fourth stage, 100 anonymized responses were evaluated in a blinded manner by two independent geriatric specialists using a standardized scoring form, resulting in 200 expert evaluations. No artificial intelligence responses or expert evaluations were excluded from the analyses.

Blinded and independent expert assessment

All assessments were conducted by two geriatric specialists who were not involved in the development of the case scenarios. Before expert evaluation, all artificial intelligence responses were anonymized, randomly coded, and stripped of any information that could reveal the identity of the artificial intelligence panel. Throughout the evaluation process, the assessors were blinded to the model that generated each response.

Two geriatric specialists independently evaluated all artificial intelligence responses. No communication occurred between the experts during the evaluation process, and the scores were not shared before the completion of the assessment. Both experts used the same standardized evaluation form with a predefined scoring criteria.

The five assessment dimensions used in this study were not directly taken from a single existing scale. Instead, they were developed in accordance with existing frameworks for the multidimensional human evaluation of large language models in health-related contexts [19, 20]. Scientific accuracy, clarity and

comprehensibility, and safety/contraindication awareness overlapped with the domains of information quality, expression quality, and safety/harm, respectively. Guideline compliance reflects the need for guideline based decision making in AI-supported clinical recommendations. A holistic geriatric approach was added as a context specific domain to evaluate multidimensional clinical reasoning in older adults with osteoporosis.

Each artificial intelligence response was evaluated using a 5-point Likert scale based on five quality criteria: scientific accuracy, guideline compliance, clarity and comprehensibility, holistic geriatric approach, and awareness of safety/contraindications. Each criterion was scored from 1 to 5, where 1 indicated “very poor” and 5 indicated “very good.” The total quality score was calculated as the sum of the five criteria and ranged from 5–25. Higher scores indicate better response quality.

Hallucination was operationally defined as any factually incorrect, fabricated, internally inconsistent, or unsupported clinical statement, according to a prespecified reference standard. The prespecified reference standard was the Turkish Society of Endocrinology and Metabolism Osteoporosis Guidelines 2025. Examples of hallucinations include incorrect drug dosing, fabricated diagnostic threshold values, incorrect contraindications, internally inconsistent treatment recommendations, and non-existent or non-guideline supported recommendations. The hallucination status was recorded as a binary variable for each case panel combination, with 1 indicating presence and 0 indicating absence.

Sample size considerations

The sample size was pragmatic and consistent with previous studies evaluating large language models in osteoporosis-related contexts [12–14]. The study included 25 clinical case scenarios, each answered by four artificial intelligence panels, resulting in 100 case panel combinations. As two independent raters evaluated each response, 200 individual expert evaluations were generated. This design provided 25 repeated observations per panel for within case comparisons among the four artificial intelligence panels. Panel level comparisons were planned using Friedman’s test, and Kendall’s W was used to quantify effect size. A Kendall’s W value of ≥ 0.25 was considered to indicate a moderate effect size. The repeated measures structure also supported the calculation of inter-rater agreement using ICC(2,1) for the total quality scores and linearly weighted Cohen’s kappa coefficients for individual Likert scale criteria.

Statistical analysis

The unit of analysis was defined as the artificial intelligence response generated for each case panel combination. The mean scores of the two independent evaluators were used for panel comparisons. Descriptive statistics are presented as mean $\pm$ standard deviation or median with interquartile range, as appropriate. Inter-rater agreement was assessed using linearly weighted Cohen’s kappa coefficients for individual Likert scale criteria and the ICC(2,1) model for total quality scores. The ICC was interpreted as a two-way random effects, single measure, and absolute agreement model. Agreement levels were interpreted according to the Landis–Koch classification [21]. Comparisons among the four artificial

intelligence panels were performed using Friedman’s test. For variables with significant overall differences, pairwise comparisons were conducted using the Bonferroni-corrected Wilcoxon signed-rank test. Kendall’s W coefficient was calculated as the effect size for the Friedman test. Hallucination frequencies were reported descriptively, and Cochran’s Q test was used to compare hallucination frequencies across the panels. For continuous summary scores, 95% confidence intervals were calculated using the t-distribution method. All statistical tests were two-tailed, and $p < 0.05$ was considered statistically significant. Analyses were performed using SPSS version 31.0.

RESULTS

Overall performance

The total Likert scores across the 25 cases for the four artificial intelligence panels are presented in Table 3. The ChatGPT 4.5 base model achieved the highest mean total score (22.76/25; 95% CI: 22.30–23.22), followed by the Gemini 3 base model (21.80; 95% CI: 21.11–22.49), NotebookLM + guideline approach (21.56; 95% CI: 20.81–22.31), and ChatGPT 4.5 + guideline approach (21.42; 95% CI: 20.86–21.98) models. In the criterion based assessment, the ChatGPT 4.5 base model received higher scores than the other panels in the domains of scientific accuracy and guideline adherence, with mean scores of 4.86 and 4.94, respectively. In contrast, for the criterion of safety and contraindication awareness, the mean scores across all panels ranged from 4.12 to 4.40, making this the lowest scoring area in the base (unguided) models. Hallucinations were detected in only one case in the Gemini 3 base model (4%), whereas no hallucinations were observed in the other three panels.

Table 3
Mean criterion scores and hallucination rates of the four AI panels.

Panel	Sci. Acc.	Guid. Conc.	Clarity	Geriatric	Safety	Total	Halluc. %
ChatGPT 4.5 (baseline)	4.86	4.94	4.50	4.34	4.12	22.76	0%
Gemini 3 (baseline)	4.44	4.54	4.36	4.26	4.20	21.80	4%
ChatGPT 4.5 + Guideline	4.54	4.06	4.22	4.20	4.40	21.42	0%
NotebookLM + Guideline	4.44	4.32	4.28	4.14	4.38	21.56	0%
Sci. Acc.=Scientific Accuracy; Guid. Conc.=Guideline Concordance; Geriatric=Holistic Geriatric Approach; Safety=Safety and Contraindication Awareness; CI= Confidence Interval. The maximum total score was 25. Values represent the mean of two expert scores (n = 25 cases). 95% CI: ChatGPT 4.5 (baseline): 22.30–23.22; Gemini 3 (baseline): 21.11–22.49; ChatGPT 4.5 + Guideline: 20.86–21.98; NotebookLM + Guideline: 20.81–22.31.							

Performance by case group

The performance of the artificial intelligence panels across the five clinical groups is summarized in Table 4. The ChatGPT 4.5 base model achieved the highest total score in all clinical groups. In the geriatric challenges group (G2), the ChatGPT 4.5 base model (mean = 23.00) and Gemini 3 base model (mean = 22.80) received the highest scores, whereas the guideline-integrated models performed poorly. In this group, the mean total scores were 21.50 for ChatGPT 4.5 + guideline and 21.20 for NotebookLM + guideline condition. In the complications and clinical pitfalls group (G5), the NotebookLM + guideline approach exhibited higher performance compared with the ChatGPT 4.5 + guideline approach; the mean total scores were 22.80 and 20.40, respectively.

Table 4
Mean total scores of the four AI panels in the case group

Case Group	ChatGPT4.5	Gemini 3	ChatGPT 4.5 + Guid.	NotebookLM + Guid.
G1:Postmenopausal Gray Zone	22.70	21.80	22.10	21.70
G2:Geriatric Challenges	23.00	22.80	21.50	21.20
G3:Drug Management	22.30	20.90	21.70	20.60
G4:Secondary Osteoporosis/ Comorbidity	22.90	21.30	22.40	21.50
G5:Complications / Pitfalls	22.90	22.20	20.40	22.80

Inter-rater agreement

The ICC(2,1) values calculated at the overall quality score level ranged from 0.310 to 0.638 across all panels (Table 5). The highest interrater agreement was observed in the NotebookLM + guideline panel ($\kappa = 0.638$), which was considered to be a good level of agreement. In contrast, the ChatGPT 4.5 base model showed the lowest total score agreement ($\kappa = 0.310$) and was classified as having poor agreement. In the criterion based analyses, excellent agreement was achieved only for the scientific accuracy criterion of the ChatGPT 4.5 base model ($\kappa = 0.834$). In the same panel, the interrater agreement for the guideline adherence criterion was good ($\kappa = 0.658$). Conversely, in the NotebookLM + guideline panel, agreement for the guideline adherence criterion was markedly low ($\kappa = 0.013$) and was considered very poor. For the holistic geriatric approach and safety/contraindication awareness criteria, the inter-rater agreement across all panels ranged from very poor to low.

Table 5
Inter-rater agreement

Criterion / Measure	ChatGPT4.5	Gemini3	ChatGPT 4.5 + Guid.	NotebookLM + Guid.
ICC(2,1) Overall (n = 25)	0.310	0.531	0.400	0.638
Scientific Accuracy (κ)	0.834	0.466	0.437	0.412
Guideline Concordance (κ)	0.658	0.422	0.501	0.013
Clarity & Comprehensibility (κ)	0.263	0.375	0.340	0.461
Holistic Geriatric Approach (κ)	0.017	0.084	0.237	0.352
Safety & Contraindication (κ)	0.300	0.138	0.215	0.046
ICC(2,1) and criterion-based weighted Cohen's κ . ICC: <0.40 poor, 0.40–0.59 moderate, 0.60–0.74 good, ≥ 0.75 excellent. κ : <0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, ≥ 0.80 excellent (Landis–Koch classification).				

Statistical comparisons between panels

The Friedman test showed a statistically significant difference among the inter-rater intelligence panels in terms of the total quality scores ($\chi^2=18.377$; $df = 3$; $p = 0.0004$). The Kendall's W value was 0.245, indicating a moderate effect size. The pairwise comparisons with Bonferroni correction are shown in Table 6. The ChatGPT 4.5 base model achieved significantly higher total quality scores than those of the other three panels. In contrast, no statistically significant difference was observed between Gemini 3 inter-rater and ChatGPT 4.5 + guide and NotebookLM + guide panels.

Table 6
Pairwise comparisons among the four AI panels.

Comparison	Mean Scores	W	p	Result
ChatGPT 4.5 vs Gemini 3	22.76 vs 21.80	34.0	0.0076	ChatGPT 4.5 superior
ChatGPT 4.5 vs ChatGPT 4.5 + Guide	22.76 vs 21.42	20.0	0.0003	ChatGPT 4.5 superior
ChatGPT 4.5 vs NotebookLM + Guide	22.76 vs 21.56	37.5	0.0021	ChatGPT 4.5 superior
Gemini 3 vs ChatGPT 4.5 + Guide	21.80 vs 21.42	85.0	0.453	No significant difference
Gemini 3 vs NotebookLM + Guide	21.80 vs 21.56	121.5	0.407	No significant difference
ChatGPT 4.5 + Guideline vs NotebookLM + Guide	21.42 vs 21.56	116.5	0.743	No significant difference
Bonferroni-corrected Wilcoxon signed-rank test ($\alpha = 0.05/6 = 0.0083$). Effect size was calculated using Kendall's W.				

Friedman analyses conducted at the criterion level showed statistically significant differences between the panels in terms of scientific accuracy ($\chi^2=17.192$; $p = 0.0006$) and guideline adherence ($\chi^2=38.257$; $p < 0.0001$). In contrast, no significant differences were identified between the panels regarding clarity and comprehensibility, holistic geriatric approach, or safety/contraindication awareness (Table 7).

Table 7
Criterion-based Friedman test results.

Criterion	Friedman χ^2	p-value	Result
Scientific Accuracy	17.192	0.0006	Significant
Guideline Concordance	38.257	< 0.0001	Significant
Clarity & Comprehensibility	4.071	0.254	ns
Holistic Geriatric Approach	1.800	0.615	ns
Safety & Contraindication	7.523	0.057	ns
ns:Non-significant			

Post-hoc analyses conducted on a per-criterion basis clearly demonstrated that inter-panel differences were concentrated solely in knowledge-based criteria, while all models exhibited similar performance on criteria requiring clinical judgment (Table 8). For the scientific accuracy criterion, ChatGPT was found to be significantly superior to both models with integrated guidelines (ChatGPT vs. ChatGPT+Guideline, $p =$

0.002; ChatGPT vs. NotebookLM+Guideline, $p = 0.0008$). The guideline adherence criterion emerged as the area where the differences between the models were most pronounced. Based on this criterion, ChatGPT received significantly higher scores than the other three models. Additionally, Gemini was significantly superior to ChatGPT+Guideline ($p = 0.006$). However, no significant differences were found in the pairwise comparisons for clarity and explicitness, the geriatric comprehensive approach, and safety/contraindication awareness.

Table 8
Criterion-based comparison of the four artificial intelligence panels.

Criterion	Comparison	Mean Scores	p	Superior Model
Scientific Accuracy	ChatGPT vs ChatGPT+Guideline	4.86 vs 4.54	0.002	ChatGPT
Scientific Accuracy	ChatGPT vs NotebookLM+Guideline	4.86 vs 4.44	0.0008	ChatGPT
Guideline Adherence	ChatGPT vs Gemini	4.94 vs 4.54	0.002	ChatGPT
Guideline Adherence	ChatGPT vs ChatGPT+Guideline	4.94 vs 4.06	0.00003	ChatGPT
Guideline Adherence	ChatGPT vs NotebookLM+Guideline	4.94 vs 4.32	0.00004	ChatGPT
Guideline Adherence	Gemini vs ChatGPT+Guideline	4.54 vs 4.06	0.006	Gemini
Bonferroni-corrected Wilcoxon signed-rank test ($\alpha = 0.05/6 = 0.0083$).				

DISCUSSION

This study is among the first to compare different artificial intelligence approaches for managing osteoporosis in older adults based on a national osteoporosis guideline using a blinded and independent expert evaluation protocol. The main findings of this study can be summarized as follows: First, the ChatGPT 4.5 base model demonstrated the highest overall performance, without external guideline support. Second, directly integrating the guidelines into the model via document upload did not improve performance. Third, the differences between the panels were primarily concentrated in knowledge based criteria, such as scientific accuracy and guideline adherence, while no significant differences were observed in criteria requiring clinical reasoning, a holistic geriatric approach, and safety/contraindication awareness. Finally, the hallucination rates were generally low.

The main methodological contribution of this study was the use of a blinded and independent expert evaluation protocol. Although blinding is known to reduce assessment bias, this approach is still applied to a limited extent in artificial intelligence model comparisons [22]. In a significant portion of current

studies, evaluators are aware of the model identity [12,1315]. This may lead to evaluation bias based on brand perception and prior expectations, especially for widely known models such as ChatGPT 4.5. In this study, having experts evaluate the model's identity independently strengthens the interpretation that the observed performance differences largely reflect the response quality. Additionally, the fact that two evaluators scored independently of each other reduced the risk of consensus bias and increased the methodological reliability of the evaluation process [19, 23].

The superior performance of the ChatGPT 4.5 base model aligns with the findings reported in the current literature. In a comparative study on postmenopausal osteoporosis, ChatGPT-4o generated more appropriate responses than other large language models [12]. Similarly, an evaluation comparing the ChatGPT and Gemini models in the field of glucocorticoid-induced osteoporosis reported differences in response quality and formats [13]. Recent comparative studies on the treatment of postmenopausal osteoporosis have also shown that ChatGPT based models can exhibit strong performance in terms of the quality of clinical response [14]. The present study verified this trend under conditions of blinded model identity and independent expert evaluation, thus contributing to the literature. The relative superiority of ChatGPT 4.5 may be related to the scope of biomedical content in the model training data, contextual consistency in clinical text generation, and capacity for structured response generation in complex clinical scenarios. However, because the training datasets, weighting processes, and optimization strategies of commercial large language models are not transparently disclosed, it is not possible to draw definitive causal inferences regarding this superiority [6].

One of the most striking findings of this study was that guideline integration did not improve the model performance. Although this result may initially seem unexpected, it can be explained by several mechanisms, as follows. First, lengthy guideline texts uploaded directly may have made it difficult for the model to select, prioritize, and use relevant information appropriately in clinical scenarios. Second, contextual mismatches may have arisen between the model's pretrained medical knowledge and the uploaded guideline content. Third, directly uploading guideline text to the chat interface does not offer the technical advantages of structured retrieval augmented generation (RAG) systems. In RAG systems, relevant information is presented to the model by selecting it from a large document pool using methods such as vector indexing, targeted retrieval and reranking. In contrast, the direct text uploading approach used in the present study represents a more accessible but technically limited form of guideline integration [17, 24, 25].

RAG is an artificial intelligence method that enables the large language model to scan relevant information from external sources before generating a response and to incorporate this information into the response generation process. This approach aims to improve the accuracy and reliability of responses, particularly in fields that require up-to-date, verifiable, and context-specific information. This finding may superficially appear contradictory to a previous review highlighting the advantages of structured RAG approaches in rheumatology [26]. However, the methodological differences between these two approaches explain this apparent contradiction. While the approach emphasized in the cited review is based on a vector indexed and structured retrieval architecture, in the present study, the

guideline content was uploaded directly into the chat interface of the model. Therefore, the present findings do not indicate that the RAG approach is ineffective, but rather suggest that unstructured guideline uploading alone may not be sufficient to enhance the model performance.

When evaluated based on the assessment criteria, direct guideline upload did not achieve the expected improvement in guideline adherence, which was anticipated to be the most impacted area. In fact, panels with uploaded guidelines scored lower on this criterion than the base ChatGPT 4.5 model did. This suggests that simply providing guideline information to the model is insufficient; the information must be appropriately selected, structured in a contextually relevant manner, and processed according to the clinical priorities.

The criterion level findings support this interpretation: while significant differences were observed in knowledge based domains (scientific accuracy and guideline adherence), no significant difference was found between the two groups in the criteria requiring more clinical judgment (clarity, holistic geriatric approach, and safety/contraindication awareness). This suggests that LLMs may be strong in knowledge generation but share similar limitations in patient centered and context sensitive clinical reasoning. Similar limitations have been reported in different clinical areas, such as polycystic ovary syndrome and sepsis management [15, 27].

From a geriatric care perspective, the relatively low scores for the safety and contraindication awareness criterion across all panels indicate that clinical safety is a common and critical weakness of artificial intelligence models. It is not sufficient for models to generate only accurate diagnostic or treatment recommendations; they must also systematically address safety components such as renal function, drug interactions, contraindications, fall risk, frailty, and the potential for harm in vulnerable older patients [23]. Notably, within the geriatric challenges group (G2), basic models without guideline integration demonstrated a higher performance. This finding suggests that general purpose large language models, when not constrained by domain specific comprehensive guideline documents, can process geriatric concepts such as polypharmacy, fall risk, functional status, and individualized care with greater flexibility based on their broad training data sets.

The interrater reliability analysis provided significant methodological depth to the study. The excellent level of agreement achieved by the ChatGPT 4.5 base model on the scientific accuracy criterion ($\kappa = 0.834$) was consistent with the relatively objective and verifiable nature of biomedical factual knowledge. In contrast, the very low kappa value calculated for guideline concordance in the NotebookLM + guideline panel ($\kappa = 0.013$) may reflect a statistical artifact related to a ceiling effect rather than a genuine clinical disagreement. However, the low kappa values observed across all panels for the holistic geriatric approach and safety/contraindication awareness criteria indicate that these domains are subjective and context sensitive.

A low rate of hallucinations is a positive finding; however, the low frequency of hallucinations observed under controlled study conditions does not guarantee the same level of safety in real world clinical practice [9]. Therefore, the rarity of hallucinations should not be interpreted as meaning that artificial

intelligence outputs can be used without clinical oversight, especially in the geriatric population, where frailty, multimorbidity, and polypharmacy are prevalent conditions. It should be considered that even a single inappropriate recommendation may amplify clinical outcomes such as falls, fractures, drug side effects, or treatment delays in older adults. Overall, these findings indicate that artificial intelligence models can provide high level informational support in osteoporosis case management among older adults. However, human clinical judgment remains indispensable for safety, evaluation of contraindications, a holistic geriatric approach, and complex clinical prioritization of treatment. The limited impact of direct guideline integration through document upload demonstrates the need for hybrid clinical decision support approaches supported by structured information retrieval augmented generation systems, domain specific knowledge graphs, and geriatric assessment modules.

LIMITATIONS

This study has limitations concerning interrater reliability and subgroup analyses because of the limited number of raters and cases. The findings reflect the performance of continuously updated commercial artificial intelligence systems as of February 2026. Guideline integration was evaluated only through the direct document upload method; structured RAG or similar optimization methods were not used. Additionally, reliance on a single national osteoporosis guideline may have limited the generalizability of the findings.

CONCLUSIONS

In this blinded expert evaluation, the base ChatGPT 4.5 model demonstrated the highest performance in managing osteoporosis cases among adults aged ≥ 65 years, with direct guideline upload providing no additional benefits. These findings suggest that artificial intelligence systems may support clinical decision making in geriatric osteoporosis; however, they cannot replace expert supervision, particularly in areas such as safety assessment, recognition of contraindications, a holistic geriatric approach, and prioritization of individualized treatment. Future studies should evaluate AI assisted clinical decision systems that use guideline knowledge in a structured manner, consider geriatric risks, and operate under the supervision of physicians in real world clinical settings.

Abbreviations

AI: Artificial intelligence

CI: Confidence Interval

FRACTURK: Fracture Risk Assessment in Turkey

ICC: Intraclass correlation coefficient

LLM: Large language model

RAG: Retrieval augmented generation

SPSS: Statistical Package for the Social Sciences

WHO: World Health Organization

Declarations

Ethics approval and consent to participate

This study did not include human participants, identifiable patient data, or animal experiments. All clinical cases evaluated consisted of hypothetical scenarios created by the researchers and did not contain real patient data. Therefore, as this study did not involve research on human participants, institutional ethics committee approval was not required. Since no human participants were included in the study and no personal health data were used, obtaining informed consent was not necessary for this study.

Consent for publication

Not applicable.

Availability of data and materials

The hypothetical clinical case scenarios, raw expert scoring datasets, and supplementary evaluation forms generated and analyzed during this study are available from the corresponding author upon reasonable request. No real patient data were used in this study.

Competing interests

The authors declare no conflict of interest.

Funding

No funding was received for this study.

Authors' contributions

Author Contributions: **GKA:** study design, preparation of the clinical case scenarios, conduct of AI queries, data acquisition, statistical analyses, analysis and interpretation of the results, writing original draft preparation, writing review and editing, and validation. **SB:** participation in the expert evaluation process, contribution to the clinical interpretation of the results, statistical analyses, writing critical review and editing, and validation. **KCA:** concept and design, methodology, writing review and editing. **DO:** supervision, concept and design, methodology, oversight of the study process, writing review and editing, and validation. **TK:** writing review and editing. **TEG:** writing review and editing. **DSE:** supervision,

concept and design, methodology, clinical interpretation of the results, writing review and editing, and validation. All authors have read and approved the final version of the manuscript.

References

1. Xiao PL, Cui AY, Hsu CJ, Peng R, Jiang N, Xu XH, et al. Global, regional prevalence, and risk factors of osteoporosis according to the World Health Organization diagnostic criteria: a systematic review and meta-analysis. *Osteoporos Int.* 2022;33(10):2137–53. <https://doi.org/10.1007/s00198-022-06454-3>.
2. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C. The epidemiology of osteoporosis. *Br Med Bull.* 2020;133(1):105–17. <https://doi.org/10.1093/bmb/ldaa005>.
3. Tuzun S, Eskiurt N, Akarirmak U, Saridogan M, Senocak M, Johansson H, et al. Turkish Osteoporosis Society. Incidence of hip fracture and prevalence of osteoporosis in Turkey: the FRACTinter-raterOsteoporosis. *International.* 2012;23(3):949–55. <https://doi.org/10.1007/s00198-011-1655-5>.
4. Camacho PM, Petak SM, Binkley N, Diab DL, Eldeiry LS, Farooki A, et al. American Association of Clinical Endocrinologists/American College of Endocrinology clinical practice guidelines for the diagnosis and treatment of postmenopausal osteoporosis – 2020 update. *Endocr Pract.* 2020;26(Suppl 1):1–46. <https://doi.org/10.4158/GL-2020-0524SUPPL>.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930–40. <https://doi.org/10.1038/s41591-023-02448-8>.
6. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. *iScience.* 2024;27(5):109713. <https://doi.org/10.1016/j.isci.2024.109713>.
7. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care.* 2023;27(1):120. <https://doi.org/10.1186/s13054-023-04393-x>.
8. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res.* 2025;14:e59823. <https://doi.org/10.2196/59823>.
9. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med.* 2025;8:274. [10.1038/s41746-025-01670-7](https://doi.org/10.1038/s41746-025-01670-7).
10. Gatineau G, Shevroja E, Vendrami C, Gonzalez-Rodriguez E, Leslie WD, Lamy O, et al. Development and reporting of artificial intelligence in osteoporosis management. *J Bone Miner Res.* 2024;39(11):1553–73. <https://doi.org/10.1093/jbmr/zjae131>.
11. Erden Y, Temel MH, Bagcier F. Artificial intelligence insights into osteoporosis: assessing ChatGPT's information quality and readability. *Archives Osteoporos.* 2024;19(1):17. <https://doi.org/10.1007/s11657-024-01376-5>.

12. Lin CR, Chen Y, Hsieh WC, et al. Multiple large language models versus clinical guidelines for postmenopausal osteoporosis: a comparative study of ChatGPT-3.5, ChatGPT-4.0, ChatGPT-4o, Google Gemini, Google Gemini Advanced, and Microsoft Copilot. *Arch Osteoporos*. 2025;20(1):120. 10.1007/s11657-025-01587-4.
13. 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. <https://doi.org/10.1186/s13018-024-04996-2>.
14. 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. <https://doi.org/10.1186/s12891-025-08601-3>.
15. Gunesli I, Aksun S, Fathelbab J, Yildiz BO. Comparative evaluation of ChatGPT-4, ChatGPT-3.5 and Google Gemini on PCOS assessment and management. *Endocrine*. 2025;88(1):315–22. <https://doi.org/10.1007/s12020-024-04121-7>.
16. Hager P, Jungmann F, Holland R, Bhatt DL, Krumholz HM, Caraballo C, et al. Evaluating and mitigating limitations of large language models in clinical decision making. *Nat Med*. 2024;30(9):2613–22. <https://doi.org/10.1038/s41591-024-03097-1>.
17. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2020. pp. 9459–9474. Preprint at <https://arxiv.org/abs/2005.11401>
18. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inform Assoc*. 2025;32(4):605–15. <https://doi.org/10.1093/jamia/ocaf008>.
19. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319–28. <https://doi.org/10.1001/jama.2024.21700>.
20. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inf Decis Mak*. 2025;25(1):117. <https://doi.org/10.1186/s12911-025-02954-4>.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74. <https://doi.org/10.2307/2529310>.
22. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet*. 2002;359(9307):696–700. [https://doi.org/10.1016/S0140-6736\(02\)07816-9](https://doi.org/10.1016/S0140-6736(02)07816-9).
23. Wang S, Tang Z, Yang H, Gong Q, Gu T, Ma H, et al. A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains. *npj Digit Med*. 2025;9:91. <https://doi.org/10.1038/s41746-025-02277-8>.
24. Kresevic S, Giuffrè M, Ajčević M, et al. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med*. 2024;7(1):102. 10.1038/s41746-024-01091-y.

25. Ke YH, Jin L, Elangovan K, Abdullah HR, Chua WL, Moochhala SM, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit Med*. 2025;8:187. <https://doi.org/10.1038/s41746-025-01519-z>.
26. Benavent D, Venerito V, Michelena X. RAGing ahead in rheumatology: new language model architectures to tame artificial intelligence. *Therapeutic Adv Musculoskelet Disease*. 2025;17:1759720X251331529. <https://doi.org/10.1177/1759720X251331529>.
27. Kutbi D, Abou-Bakr E, Haidar HM. Evaluating the accuracy of medical information generated by ChatGPT and Gemini and its alignment with international clinical guidelines from the Surviving Sepsis Campaign. *JMIR Formative Res*. 2025;9:e84251. <https://doi.org/10.2196/84251>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationEN.docx](#)