

Generative Artificial Intelligence for Opportunistic Osteoporosis Screening Using Lumbar Spine Imaging: Diagnostic Performance, Reproducibility, Effects of Few-Shot Prompting, and Comparison with Physicians

Chikara Hayakawa

Showa Medical University School of Medicine

Koki Tsuchiya

Showa Medical University School of Medicine

Yusuke Dodo

Showa Medical University School of Medicine

Yoshifumi Kudo

Showa Medical University School of Medicine

Ichiro Okano

`okano.i@med.showa-u.ac.jp`

Showa Medical University School of Medicine

Research Article

Keywords:

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11044030/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

Purpose

Osteoporosis remains underdiagnosed despite widespread availability of lumbar radiographs and computed tomography (CT). We compared physicians and multimodal generative artificial intelligence (AI) for opportunistic osteoporosis screening, focusing on diagnostic accuracy, reproducibility, and effects of few-shot prompting.

Methods

Forty-five lateral lumbar radiographs and 45 sagittal CT images were classified using DXA T-score thresholds of -1.0 and -2.5 as reference standards. Performance was compared among 42 board-certified orthopedic surgeons, 50 non-specialists, and ChatGPT, Gemini 3.1 Pro, and Claude 4 Opus. Each AI model evaluated all images five times under zero-shot and few-shot conditions, with 20 labeled reference cases provided for few-shot prompting. An exploratory analysis evaluated combined specialist-AI classifications using OR and AND rules.

Results

At the -1.0 cutoff, specialists outperformed non-specialists on radiographs ($P < 0.001$). Claude achieved the highest radiographic accuracy (73.3%), followed by Gemini (64.4%), specialists (57.4%), ChatGPT (55.6%), and non-specialists (44.0%). Educational examples did not improve physician accuracy, and few-shot prompting did not consistently improve AI accuracy. However, reproducibility improved in several models, most notably Claude. An OR rule combining specialist and Claude classifications increased sensitivity to 80.0% on radiographs and 90.0% on CT, with reduced specificity.

Conclusion

Generative AI achieved physician-comparable performance under selected conditions but showed model-dependent variability. Few-shot prompting may improve consistency without improving accuracy. Exploratory combined assessment suggested complementary physician-AI errors may reduce missed cases at the cost of false-positive classifications. Generative AI may therefore serve as a decision-support tool for opportunistic osteoporosis screening but should complement rather than replace physician judgment.

Introduction

Osteoporosis is a major public health concern associated with fragility fractures, disability, and mortality and remains substantially underdiagnosed and undertreated despite effective therapies[1, 2]. It is also relevant to spinal surgery because poor bone quality is associated with implant loosening, fixation failure, proximal junctional failure, and revision surgery[3, 4]. Early identification of patients at risk is therefore important for both fracture prevention and surgical management.

Dual-energy X-ray absorptiometry (DXA) is the current gold standard for diagnosing osteoporosis[5]. However, many patients do not undergo bone mineral density (BMD) assessment during routine care, prompting interest in opportunistic screening using images obtained for other indications[6–8]. Physicians routinely form an intuitive impression of bone quality from radiographs and computed tomography (CT) images, although a previous study found that the accuracy of such visual assessment was only moderate, even among experienced orthopedic surgeons[9].

Deep learning algorithms can identify osteoporosis from radiographs and CT images with promising performance[10, 11], but these approaches require task-specific development. In contrast, multimodal generative artificial intelligence (AI) is widely accessible and can interpret medical images without dedicated model development. Its ability to visually identify osteoporosis from routine spinal images remains poorly understood.

Multimodal generative AI can also perform in-context learning using instructions and example images without additional model training. Zero-shot prompting provides instructions without task-specific examples, whereas few-shot prompting provides a limited number of examples to guide subsequent responses[12]. Whether such examples improve diagnostic accuracy or mainly affect response consistency remains unclear. We therefore compared physicians and multimodal generative AI for opportunistic osteoporosis screening using lumbar radiographs and sagittal CT images, focusing on diagnostic performance, reproducibility, and the effects of few-shot prompting. We additionally explored whether combining specialist and AI classifications could improve screening performance through complementary diagnostic errors.

Materials and Methods

Study Design

We hypothesized that multimodal generative AI would demonstrate diagnostic performance comparable to that of physicians for opportunistic osteoporosis screening using lumbar spine imaging, while showing model-dependent variability in reproducibility. We further hypothesized that few-shot prompting would alter AI diagnostic performance and reproducibility compared with zero-shot prompting.

This image-based comparative study evaluated physicians and generative AI systems using previously acquired lumbar spine radiographs and sagittal CT images. Physician assessment data were obtained from a previous study[9], and three generative AI models were evaluated using the same image dataset and classification task. The overall study design and evaluation framework are summarized in Table 1 and illustrated in Figure 1. The study was approved by the Institutional Review Board of the authors' affiliated institution. Informed consent was obtained through an opt-out approach because of the retrospective design and use of anonymized imaging data. This study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.

Image Dataset

Patients who underwent DXA, lumbar spine radiography, and lumbar CT were retrospectively identified. Osteoporosis was defined using the World Health Organization DXA T-score threshold of -2.5 [13]; an additional threshold of -1.0 was used to identify low bone mass. The dataset consisted of 45 lateral lumbar radiographs and 45 sagittal reconstructed lumbar CT images. Patients with obvious osteoporotic vertebral fractures, severe spinal deformity, previous spinal surgery, or focal osseous lesions were excluded. All images were anonymized before evaluation.

Physician Visual Assessment

The physician assessment protocol has been reported previously[9]. Forty-two board-certified orthopedic surgeons and 50 non-specialist physicians independently reviewed the images. Without access to demographics, clinical information, DXA results, or quantitative imaging measurements, participants classified each image as either normal or osteoporotic/weak bone quality based solely on visual appearance.

Generative Artificial Intelligence Assessment

Three publicly available multimodal generative AI systems were evaluated: ChatGPT (OpenAI, GPT-5, accessed January 2026), Claude 4 Opus (Anthropic), and Gemini 3.1 Pro (Google).

For each evaluation, a new chat session was initiated and all 45 images of a given modality were uploaded simultaneously. An identical standardized prompt instructed each model to classify every image as normal bone quality (A) or osteoporotic/weak bone quality (B) based solely on image appearance (Online Resource 1). No clinical information or DXA measurements were provided.

Each model evaluated the dataset independently five times under identical conditions, with each evaluation conducted in a new independent chat session without access to the conversation history of previous runs. Majority voting (≥ 3 of 5 responses) determined the final classification for the primary analysis.

For the few-shot condition, a reference file containing 20 labeled example cases was first presented to each model. Each case included an image, its classification as normal or osteoporosis, and lumbar and femoral young adult mean (YAM) values. After being provided with the reference file, each model classified the test images using the same task instructions. The few-shot evaluation was repeated five times under the same independent-session conditions.

Outcome Measures

The primary outcome was diagnostic accuracy using DXA as the reference standard, evaluated separately for radiographs and CT images at T-score thresholds of -2.5 and -1.0 . Secondary outcomes included sensitivity, specificity, and evaluator agreement. For AI models, majority-vote classifications across five repeated evaluations were used for the primary diagnostic performance analysis. Accuracy for each individual run and mean accuracy across the five runs were additionally summarized descriptively (Online Resource 1). Repeated evaluations were used to assess intra-model reproducibility rather than as independent observations for inferential analysis. As an exploratory post hoc analysis of human-AI complementarity, majority-vote classifications of the 42 board-certified orthopedic surgeons were combined with those of Claude, which was selected because it demonstrated the highest overall radiographic performance among the evaluated AI models. Under the OR rule, a case was classified as positive when either the specialist majority vote or Claude was positive; under the AND rule, a case was classified as positive only when both were positive. Sensitivity, specificity, positive predictive value

(PPV), and negative predictive value (NPV) were calculated for each strategy. Mutual error rescue was also assessed by determining the proportion of specialist false-negative and false-positive classifications correctly classified by Claude, and vice versa.

Statistical Analysis

Continuous variables are presented as mean $\pm$ SD and categorical variables as counts and percentages. Diagnostic performance between board-certified orthopedic surgeons and non-specialists was compared using the Mann–Whitney U test, and changes after educational images were analyzed using the Wilcoxon signed-rank test. AI accuracy across repeated evaluations was analyzed descriptively because the five evaluations were not independent. Intra-model reproducibility and inter-model agreement were assessed using Cohen's κ coefficients. The exploratory combined-decision analyses were summarized descriptively without inferential testing. A two-sided P value <0.05 was considered statistically significant. Analyses were performed using JMP version 18 (SAS Institute Inc., Cary, NC, USA).

Results

A total of 92 physicians participated, including 42 board-certified orthopedic surgeons and 50 non-specialists. Three multimodal generative AI models evaluated the identical image dataset five times, with majority-vote classifications used for the primary analysis.

For the T-score cutoff of -1.0 , board-certified orthopedic surgeons demonstrated significantly higher radiographic diagnostic accuracy than non-specialists ($P<0.001$). Claude achieved the highest accuracy (73.3%), followed by Gemini (64.4%), board-certified orthopedic surgeons (57.4%), ChatGPT (55.6%), and non-specialists (44.0%) (Figure 2). For CT images, board-certified orthopedic surgeons achieved the highest accuracy. AI results were evaluated descriptively and were not included in inferential comparisons (Table 2).

Using the conventional osteoporosis threshold of -2.5 , no significant difference in radiographic accuracy was observed between board-certified orthopedic surgeons and non-specialists ($P=0.098$). Specialists demonstrated significantly higher CT accuracy than non-specialists ($P<0.001$), whereas AI models showed broadly similar accuracy in descriptive comparisons.

Exposure to 20 educational images did not significantly improve physician accuracy for either modality. Similarly, few-shot prompting did not consistently improve majority-vote accuracy among the AI models. Changes varied substantially by model and condition, ranging from a 24.4-percentage-point decrease for Gemini on radiographs at the -1.0 cutoff to an 11.1-percentage-point increase for ChatGPT on CT images (Table 3).

AI reproducibility differed substantially among models. Under zero-shot prompting, ChatGPT showed the highest CT reproducibility ($\kappa=0.743$), whereas Gemini showed poor reproducibility ($\kappa=0.100$ – 0.203). Claude demonstrated the greatest improvement after few-shot prompting, with κ increasing from 0.543 to 0.855 for radiographs and from 0.378 to 0.732 for CT, without a corresponding consistent improvement in majority-vote accuracy (Table 4).

In the exploratory human-AI combination analysis, the OR rule increased sensitivity compared with specialist majority voting alone. At the -1.0 cutoff, sensitivity increased from 66.7% to 80.0% for radiographs and from 83.3% to 90.0% for CT, while specificity decreased from 73.3% to 66.7% and from 80.0% to 46.7%, respectively. At the -2.5 cutoff, radiographic sensitivity increased from 73.3% to 86.7%; for CT, specialist sensitivity was already 100% and remained unchanged with the OR rule. Conversely, the AND rule generally increased specificity at the expense of sensitivity (Table 5). Error-rescue analysis demonstrated partially non-overlapping errors between specialists and Claude. For radiographs at the -1.0 cutoff, Claude correctly classified 4 of 10 specialist false-negative cases (40.0%) and 2 of 4 specialist false-positive cases (50.0%). At the -2.5 cutoff, Claude rescued 2 of 4 specialist false negatives (50.0%) and 4 of 13 false positives (30.8%). Similar bidirectional error rescue was observed for CT images (Table 5).

Representative cases also demonstrated discordant classifications between physicians and AI (Figure 3). In one osteopenic case, all three AI models correctly identified low bone density, whereas only 26.2% of specialists did so. Conversely, in another case, 71.4% of specialists were correct, whereas ChatGPT and Gemini misclassified the image, consistent with the partially non-overlapping errors observed in the exploratory analysis.

Discussion

This study yielded five principal findings. First, multimodal generative AI achieved diagnostic performance comparable to physicians under several conditions, although performance varied among models and imaging modalities. Second, educational examples for physicians and few-shot prompting for AI did not consistently improve diagnostic accuracy in either physicians or AI. Third, few-shot prompting improved intra-model reproducibility in some AI models, particularly Claude, despite no consistent improvement in diagnostic accuracy. Fourth,

exploratory combination analysis showed that an OR rule increased sensitivity, although at the cost of reduced specificity. Finally, specialists and AI made partially non-overlapping false-negative and false-positive errors, providing quantitative evidence of potential diagnostic complementarity.

Visual assessment of bone quality remained challenging for both physicians and AI. DXA may overestimate lumbar BMD in the presence of degenerative changes[14], while visually distinguishing normal from mildly reduced bone mass is inherently difficult. These factors may partly explain the moderate performance observed across evaluators.

Generative AI performance also varied substantially among models, consistent with previous multimodel comparisons[15, 16]. Although these systems have similar user-facing capabilities, they should not be considered interchangeable diagnostic tools. Unlike dedicated osteoporosis-detection algorithms, current generative AI models are general-purpose systems not optimized for quantitative bone-density assessment. A recent meta-analysis similarly found generative AI comparable to non-expert physicians but inferior to expert physicians overall[17].

The clinical value of generative AI may therefore depend more on its integration into clinical workflows than on whether it independently outperforms physicians. Simply providing physicians access to a large language model does not necessarily improve diagnostic reasoning[18], and concerns remain regarding errors from autonomous use[19]. Conversely, appropriately designed human–AI collaboration may improve diagnostic performance[20]. Diagnostic complementarity has been described as arising when human and AI readers have different strengths and make errors on different cases, allowing one reader to catch errors made by the other[21]. In our exploratory analysis, Claude corrected both false-negative and false-positive classifications made by specialists, while specialists also corrected errors made by Claude. Importantly, the OR rule increased sensitivity, potentially useful for opportunistic screening, in which the goal is to identify patients who should undergo formal BMD assessment; however, this improvement came at the cost of lower specificity. Thus, these findings support potential human-AI complementarity rather than a simple improvement in overall diagnostic performance.

Another important finding was that few-shot prompting did not consistently improve AI diagnostic accuracy. Few-shot prompting does not explicitly retrain model parameters but instead provides examples in context to guide subsequent responses. Agbareia et al. reported improved retinal disease classification with few-shot prompting[22], whereas our findings suggest that its effectiveness may depend on the task, imaging modality, and AI model. Thus, providing a limited number of labeled images should not be assumed to improve diagnostic accuracy.

In contrast, few-shot prompting improved reproducibility in several models, most notably Claude. This dissociation between accuracy and reproducibility is clinically relevant: an AI system that provides inconsistent classifications for the same image may be difficult to integrate reliably into clinical workflows[23, 24], whereas greater consistency does not necessarily indicate greater diagnostic validity[23]. Therefore, accuracy and reproducibility should be evaluated as distinct dimensions of generative AI performance. Few studies have specifically examined the effect of example-based learning on both measures, making this distinction an important contribution of the present study.

Current generative AI should therefore be considered a potential decision-support tool rather than an autonomous system for opportunistic osteoporosis screening. The goal of opportunistic screening is not to establish osteoporosis from radiographs or CT alone but to identify patients who may benefit from formal BMD assessment[6]. Our exploratory results suggest that combining physician and AI judgments may reduce missed cases, but simple OR-based combination also increases false-positive referrals. Future studies should therefore investigate prospective human-AI workflows that preserve gains in sensitivity while limiting unnecessary BMD assessment.

This study has several limitations. First, commercial generative AI models evolve rapidly, and these results reflect only the versions evaluated[25]. Second, this single-center study used a relatively small image dataset. Third, evaluations were based on imaging alone without clinical or demographic information. Fourth, physician data were obtained from a previous study, whereas AI evaluations were performed subsequently, although identical images and tasks were used. Fifth, majority voting across five AI responses may not reflect routine practice, in which a single response would generally be used. Finally, the human-AI combination analysis was exploratory and post hoc, and Claude was selected on the basis of its performance in the present dataset. Moreover, the combination analysis used aggregate specialist majority voting rather than individual physician-AI interaction and therefore does not represent a real-time clinical workflow. The apparent complementary effects require prospective validation in independent cohorts.

Nevertheless, this study directly compared specialists, non-specialists, and multiple generative AI models using identical spinal images and evaluated both diagnostic accuracy and repeated-response reproducibility. The additional observation of partially non-overlapping physician and AI errors provides a basis for future evaluation of structured human-AI screening strategies.

Conclusion

Physicians demonstrated only moderate ability to visually identify osteoporosis from lumbar radiographs and CT images. Selected generative AI models achieved comparable or superior performance under specific conditions, but accuracy and reproducibility varied substantially among models. Importantly, improved reproducibility after few-shot prompting did not necessarily correspond to improved accuracy. Exploratory combination analysis suggested that complementary physician and AI judgments may reduce missed cases, although simple combination rules involve a trade-off between sensitivity and specificity. Generative AI demonstrated complementary diagnostic performance to physicians in this experimental setting, although its accuracy and reproducibility varied across models and conditions.

Declarations

Author Contribution

C. H. contributed to the study conception and design, data collection, analysis and interpretation of data, and drafting of the manuscript. K.T. and Y. D. contributed to data collection and critically revised the manuscript. Y. K. contributed to supervision and critical revision of the manuscript for important intellectual content. I. O.k contributed to the study conception and design, interpretation of data, supervision, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

The authors would like to thank all physicians who participated in the image evaluation for their valuable contributions to this study.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available because they contain clinical and imaging data and are subject to institutional and ethical restrictions. Requests for access to de-identified data may be directed to the corresponding author and will be considered in accordance with institutional and ethical requirements.

References

1. Solomon DH, Johnston SS, Boytsov NN et al (2014) Osteoporosis medication use after hip fracture in U.S. patients between 2002 and 2011. *J Bone Min Res* 29:1929–1937. <https://doi.org/10.1002/jbmr.2202>
2. Khosla S, Shane E (2016) A crisis in the treatment of osteoporosis. *J Bone Min Res* 31:1485–1487. <https://doi.org/10.1002/jbmr.2888>
3. Bjerke BT, Zarrabian M, Aleem IS et al (2018) Incidence of osteoporosis-related complications following posterior lumbar fusion. *Global Spine J* 8:563–569. <https://doi.org/10.1177/2192568217743727>
4. Anderson PA, Kadri A, Hare KJ, Binkley N (2020) Preoperative bone health assessment and optimization in spine surgery. *Neurosurg Focus* 49:E2. <https://doi.org/10.3171/2020.5.FOCUS20255>
5. LeBoff MS, Greenspan SL, Insogna KL et al (2022) The clinician's guide to prevention and treatment of osteoporosis. *Osteoporos Int* 33:2049–2102. <https://doi.org/10.1007/s00198-021-05900-y>
6. Gausden EB, Nwachukwu BU, Schreiber JJ et al (2017) Opportunistic use of CT imaging for osteoporosis screening and bone density assessment: a qualitative systematic review. *J Bone Joint Surg Am* 99:1580–1590. <https://doi.org/10.2106/JBJS.16.00749>
7. Yang J, Liao M, Wang Y et al (2022) Opportunistic osteoporosis screening using chest CT with artificial intelligence. *Osteoporos Int* 33:2547–2561. <https://doi.org/10.1007/s00198-022-06491-y>
8. Kadri A, Binkley N, Hernando D, Anderson PA (2022) Opportunistic use of lumbar magnetic resonance imaging for osteoporosis screening. *Osteoporos Int* 33:861–869. <https://doi.org/10.1007/s00198-021-06129-5>
9. Hayakawa C, Dodo Y, Tsuchiya K et al (2026) The accuracy of surgeons' intuitive diagnosis of osteoporosis using radiographs and computed tomography images of the lumbar spine. *Spine Open* 2(1). <https://doi.org/10.1097/BN9.0000000000000064>
10. Zhang B, Yu K, Ning Z et al (2020) Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: a multicenter retrospective cohort study. *Bone* 140:115561. <https://doi.org/10.1016/j.bone.2020.115561>
11. Hong N, Cho SW, Shin S et al (2023) Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography. *J Bone Min Res* 38:887–895. <https://doi.org/10.1002/jbmr.4814>

12. Ferber D, Wölflein G, Wiest IC et al (2024) In-context learning enables multimodal large language models to classify cancer pathology images. *Nat Commun* 15:10104. <https://doi.org/10.1038/s41467-024-51465-9>
13. Kanis JA (1994) Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. WHO Study Group. *Osteoporos Int* 4:368–381. <https://doi.org/10.1007/BF01622200>
14. Rand T, Seidl G, Kainberger F et al (1997) Impact of spinal degenerative changes on the evaluation of bone mineral density with dual energy X-ray absorptiometry (DXA). *Calcif Tissue Int* 60:430–433. <https://doi.org/10.1007/s002239900258>
15. Sonoda Y, Kurokawa R, Nakamura Y et al (2024) Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in Diagnosis Please cases. *Jpn J Radiol* 42:1231–1235. <https://doi.org/10.1007/s11604-024-01619-y>
16. von der Stüek MS, Vuskov R, Westfechtel S et al (2026) Visual large language models in radiology: a systematic multimodel evaluation of diagnostic accuracy and hallucinations. *Life (Basel)* 16:66. <https://doi.org/10.3390/life16010066>
17. Takita H, Kabata D, Walston SL et al (2025) A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *NPJ Digit Med* 8:175. <https://doi.org/10.1038/s41746-025-01543-z>
18. Goh E, Gallo R, Hom J et al (2024) Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 7:e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>
19. Barile J, Margolis A, Cason G et al (2024) Diagnostic accuracy of a large language model in pediatric case studies. *JAMA Pediatr* 178:313–315. <https://doi.org/10.1001/jamapediatrics.2023.5750>
20. McDuff D, Schaekermann M, Tu T et al (2025) Towards accurate differential diagnosis with large language models. *Nature* 642:451–457. <https://doi.org/10.1038/s41586-025-08869-4>
21. Wagner I, Chakradeo K (2025) Human-AI complementarity in diagnostic radiology: the case of double reading. *Philos Technol* 38:57. <https://doi.org/10.1007/s13347-025-00886-5>
22. Agbareia R, Omar M, Zloto O et al (2025) Multimodal LLMs for retinal disease diagnosis via OCT: few-shot versus single-shot learning. *Ther Adv Ophthalmol* 17:25158414251340569. <https://doi.org/10.1177/25158414251340569>
23. Güler I, Grieb G, Kraus A et al (2026) Diagnostic accuracy and stability of multimodal large language models for hand fracture detection: a multi-run evaluation on plain radiographs. *Diagnostics (Basel)* 16:424. <https://doi.org/10.3390/diagnostics16030424>
24. Krishna S, Bhambra N, Bleakney R, Bhayana R (2024) Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. *Radiology* 311:e232715. <https://doi.org/10.1148/radiol.232715>
25. Maddox TM, Embí P, Gerhart J et al (2025) Generative AI in medicine—evaluating progress and challenges. *N Engl J Med* 392:2479–2483. <https://doi.org/10.1056/NEJMs2503956>

Tables

Table 1. Study Design and Evaluation Framework		
Characteristic	Physicians	Generative AI
Evaluators	42 board-certified orthopedic surgeons + 50 non-specialists	ChatGPT (GPT-5), Gemini 3.1 Pro, Claude 4 Opus
Test images	45 radiographs + 45 CT images	Same
Classification	Normal / Osteoporosis	Same
Ground truth	DXA T-score	Same
Cutoff values	-1.0, -2.5	Same
Repeated inference	Once	Five independent evaluations
Example-based condition	20 educational images	20 reference examples provided in context
AI, artificial intelligence; CT, computed tomography; DXA, dual-energy X-ray absorptiometry.		

Table 2. Diagnostic performance of physicians and generative AI for osteoporosis screening using radiographs and CT images

Group	X-ray Accuracy (%)	X-ray Sensitivity (%)	X-ray Specificity (%)	CT Accuracy (%)	CT Sensitivity (%)	CT Specificity (%)
T-score cut-off: -1.0						
Specialists	57.4 ± 11.3	56.7 ± 17.6	60.3 ± 15.1	63.7 ± 11.7	64.6 ± 16.8	61.3 ± 20.2
Non-specialists	44.0 ± 9.9	41.1 ± 13.9	49.6 ± 15.0	45.7 ± 11.0	46.9 ± 15.9	43.3 ± 20.8
ChatGPT	25/45 (55.6)	56.7	53.3	18/45 (40.0)	30.0	60.0
Gemini	29/45 (64.4)	66.7	60.0	21/45 (46.7)	56.7	40.0
Claude	33/45 (73.3)	70.0	80.0	22/45 (48.9)	46.7	53.3
P value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
T-score cut-off: -2.5						
Specialists	55.2 ± 8.0	60.0 ± 19.8	53.4 ± 14.2	59.2 ± 9.0	73.0 ± 19.4	52.6 ± 15.0
Non-specialists	51.9 ± 8.8	44.1 ± 17.2	55.7 ± 12.3	47.7 ± 9.7	46.7 ± 19.2	48.1 ± 15.4
ChatGPT	24/45 (53.3)	60.0	50.0	21/45 (46.7)	20.0	60.0
Gemini	25/45 (55.6)	73.3	46.7	22/45 (48.9)	60.0	43.3
Claude	30/45 (66.7)	80.0	60.0	23/45 (51.1)	46.7	53.3
P value	0.098	<0.001	0.496	<0.001	<0.001	0.131

AI, artificial intelligence; CT, computed tomography.

P values represent comparisons between board-certified orthopedic surgeons and non-specialists using the Mann-Whitney U test. AI results were calculated from majority-vote classifications across five repeated inferences and were not included in the inferential group comparison.

Table 3. Effect of educational examples in physicians and few-shot prompting in generative AI on diagnostic accuracy.

Evaluator	n	X-ray Baseline (%)	X-ray After examples (%)	Δ Accuracy (percentage points)	P value	CT Baseline (%)	CT After examples (%)	Δ Accuracy (percentage points)	P value
T-score cut-off: -1.0									
Specialists	17	62.2 $\pm$ 10.5	60.0 $\pm$ 12.4	-2.2	0.484	66.9 $\pm$ 11.2	68.2 $\pm$ 11.7	+1.3	0.469
Non-specialists	7	48.6 $\pm$ 5.8	43.8 $\pm$ 10.3	-4.8	0.219	52.7 $\pm$ 9.4	54.3 $\pm$ 11.5	+1.6	0.953
ChatGPT	5 runs	55.6	51.1	-4.5	-	40.0	51.1	+11.1	-
Gemini	5 runs	64.4	40.0	-24.4	-	46.7	44.4	-2.3	-
Claude	5 runs	73.3	71.1	-2.2	-	48.9	48.9	0.0	-
T-score cut-off: -2.5									
Specialists	17	57.9 $\pm$ 8.7	57.3 $\pm$ 7.7	-0.6	1.000	60.9 $\pm$ 8.8	60.4 $\pm$ 8.8	-0.5	0.730
Non-specialists	7	48.9 $\pm$ 7.4	49.2 $\pm$ 6.6	+0.3	0.938	47.0 $\pm$ 11.4	54.0 $\pm$ 7.3	+7.0	0.094
ChatGPT	5 runs	53.3	48.9	-4.4	-	46.7	44.4	-2.3	-
Gemini	5 runs	55.6	46.7	-8.9	-	48.9	46.7	-2.2	-
Claude	5 runs	66.7	64.4	-2.3	-	51.1	55.6	+4.5	-

Physician data are presented as mean $\pm$ SD. For physicians, "after examples" indicates reassessment after viewing 20 educational images; for AI models, it indicates few-shot prompting with 20 reference examples provided in context. For AI models, n refers to five independent inference runs rather than independent evaluators, and accuracy represents majority-vote classification across those runs. P values for physicians were calculated using the Wilcoxon signed-rank test. No inferential statistical testing was performed for AI models because repeated inference runs were not considered independent observations.

Table 4. Intra-model reproducibility under zero-shot and few-shot prompting across five repeated inferences (pairwise Cohen's κ)

Modality	Model	Zero-shot κ (mean $\pm$ SD)	Few-shot κ (mean $\pm$ SD)	$\Delta\kappa$
Radiograph	ChatGPT	0.627 $\pm$ 0.131	0.699 $\pm$ 0.155	+0.072
	Gemini	0.100 $\pm$ 0.125	0.188 $\pm$ 0.089	+0.088
	Claude	0.543 $\pm$ 0.184	0.855 $\pm$ 0.090	+0.312
CT	ChatGPT	0.743 $\pm$ 0.101	0.674 $\pm$ 0.208	-0.069
	Gemini	0.203 $\pm$ 0.186	0.245 $\pm$ 0.099	+0.042
	Claude	0.378 $\pm$ 0.285	0.732 $\pm$ 0.100	+0.354

Values are presented as the mean $\pm$ SD of the 10 pairwise Cohen's κ coefficients among five independent inference runs. $\Delta\kappa$ was calculated as few-shot κ minus zero-shot κ . No inferential statistical testing was performed.

Table 5. Exploratory diagnostic performance of combined specialist and Claude classifications

T-score cutoff	Modality	Strategy	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
-1.0	Radiograph	Specialist majority	66.7	73.3	83.3	52.4
		Claude	70.0	80.0	87.5	57.1
		OR rule	80.0	66.7	82.8	62.5
		AND rule	56.7	86.7	89.5	50.0
-1.0	CT	Specialist majority	83.3	80.0	89.3	70.6
		Claude	46.7	53.3	66.7	33.3
		OR rule	90.0	46.7	77.1	70.0
		AND rule	40.0	86.7	85.7	41.9
-2.5	Radiograph	Specialist majority	73.3	56.7	45.8	81.0
		Claude	80.0	60.0	50.0	85.7
		OR rule	86.7	46.7	44.8	87.5
		AND rule	66.7	70.0	52.6	80.8
-2.5	CT	Specialist majority	100.0	56.7	53.6	100.0
		Claude	46.7	53.3	33.3	66.7
		OR rule	100.0	33.3	42.9	100.0
		AND rule	46.7	76.7	50.0	74.2

Error-rescue data: At the -1.0 cutoff, Claude rescued 4/10 (40.0%) specialist false negatives and 2/4 (50.0%) specialist false positives on radiographs, and 2/5 (40.0%) false negatives and 1/3 (33.3%) false positives on CT. At the -2.5 cutoff, corresponding rescue rates were 2/4 (50.0%) and 4/13 (30.8%) on radiographs; no specialist false negatives occurred on CT, while Claude rescued 6/13 (46.2%) specialist false positives.

OR rule: positive if either the specialist majority vote or Claude was positive. AND rule: positive only if both were positive. PPV, positive predictive value; NPV, negative predictive value.

Figures

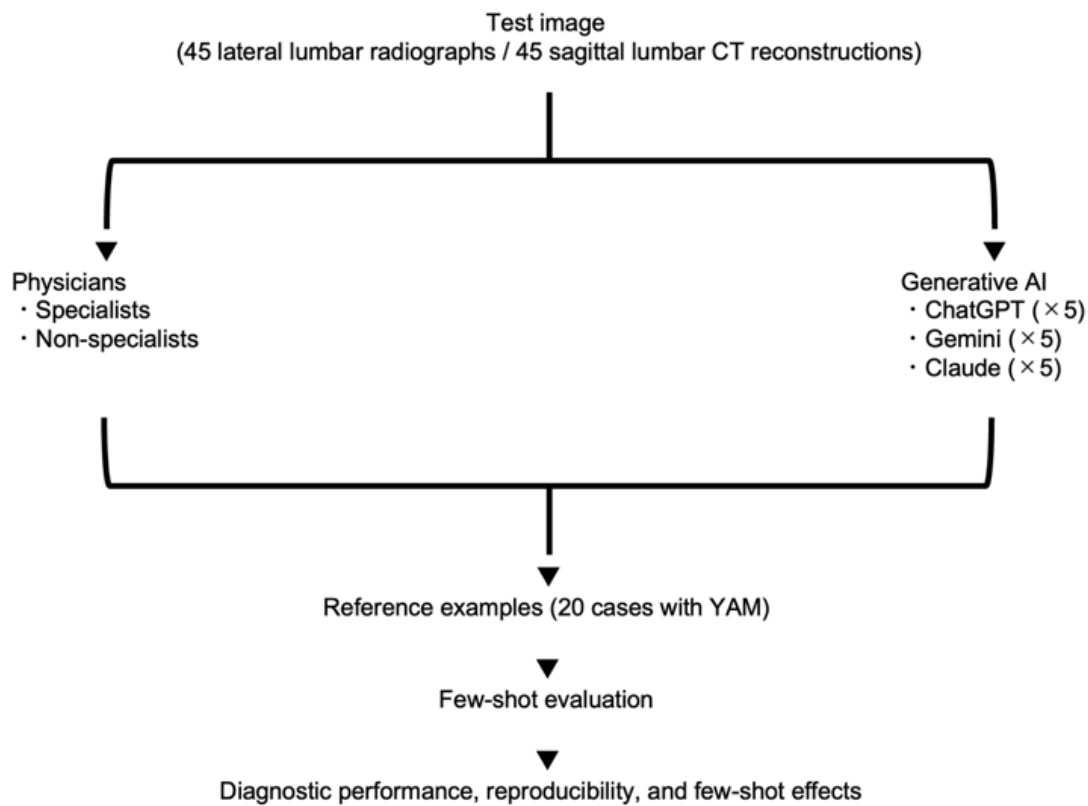


Figure 1

Study workflow for physician and generative AI evaluation of lumbar radiographs and sagittal CT images.

Physicians performed a single evaluation, whereas each generative AI model performed five independent evaluations. After the initial evaluation, educational images were provided to physicians, and reference examples were provided to AI models for few-shot evaluation. CT, computed tomography; YAM, young adult mean

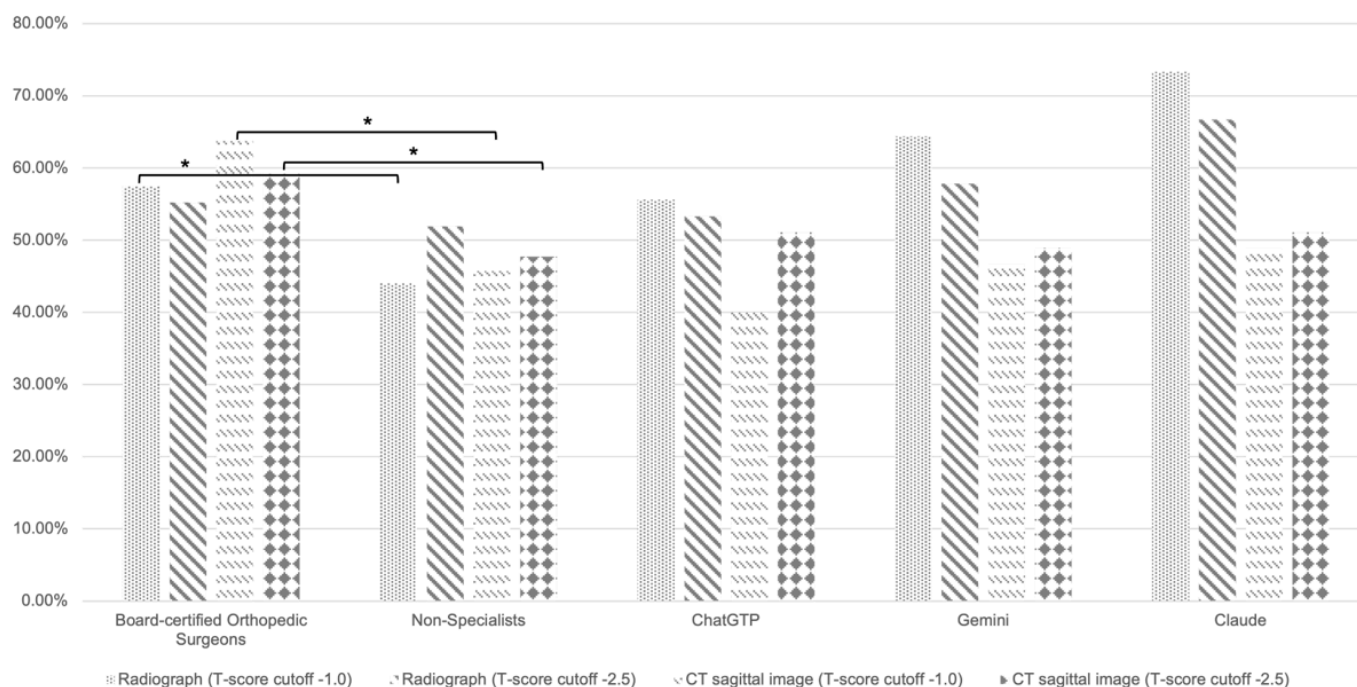


Figure 2

Diagnostic accuracy of physicians and generative AI for osteoporosis screening using lumbar radiographs and CT images at T-score cutoffs of -1.0 and -2.5.

Statistically significant differences were observed only between board-certified orthopedic surgeons and non-specialists. *P < 0.05



Ground truth: Low bone mass (T-score cutoff: -1.0)



Ground truth: Normal bone mass (T-score cutoff: -1.0)

Figure 3

Representative lumbar radiographs illustrating discordant classifications between physicians and generative AI.

a A low-bone-mass case at the T-score cutoff of -1.0 was correctly classified by all three AI models, whereas only 26.2% of specialists and 26.0% of non-specialists classified the case correctly.

b A normal-bone-mass case at the T-score cutoff of -1.0 was correctly classified by 71.4% of specialists and 58.0% of non-specialists; Claude classified the case correctly, whereas ChatGPT and Gemini misclassified it

Supplementary Files

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

- [ESM1.pdf](#)