



Artificial intelligence vs human opportunistic detection of vertebral fractures in routine CT scans: results of a pilot study

Polyzois Makras^{1,2} · Eleftherios Chatzellis¹ · Christos Gravvanis¹ · Georgia Kanti¹ · Maria P. Yavropoulou^{2,3} · Konstantinos Papadimitropoulos⁴ · Nikolaos Kyriakopoulos⁵ · the Radiology Department reporting group

Received: 29 December 2025 / Accepted: 30 March 2026
© The Author(s), under exclusive licence to Hellenic Endocrine Society 2026

Abstract

Purpose Vertebral fractures (VFs) are the most common osteoporotic fractures and a hallmark of bone fragility, yet most morphometric VFs remain undiagnosed in routine imaging not focused on the spine. This pilot study evaluated the diagnostic performance of an artificial intelligence (AI) system for opportunistic vertebral fracture detection in chest and abdominal computer tomography (CT) scans compared with human readers in a real-world hospital setting.

Methods Over 2 months, all thoracic and abdominal CTs performed for any indication in patients aged ≥ 50 years at a tertiary hospital were analyzed by the Bone Solution HealthOST AI platform (Nanox AI Ltd.). Routine radiology reports and targeted reviews by endocrinologists were compared to AI outputs using a gold-standard adjudication by an expert panel. Given that all readers assessed the same cases, paired comparisons between AI and human readers were performed using McNemar's test, with chi-square analyses conducted as secondary comparisons.

Results Among 427 eligible CT scans, AI detected vertebral fractures with significantly higher sensitivity (86.3%) than radiologists (50.0%) or endocrinologists (68.8%) ($p < 0.001$ for both comparisons). Compared with radiologists, AI had 6.3-fold greater odds of correctly identifying true fractures (OR = 6.30; 95% CI: 2.90–13.58), while against targeted review by endocrinologists, AI achieved a 2.85-fold advantage (OR = 2.85; 95% CI: 1.29–6.30). Extrapolated annually, AI integration could uncover around 150 new, otherwise undiagnosed patients eligible for osteoporosis treatment.

Conclusion Automated vertebral fracture detection in routine CT scans significantly enhances diagnostic yield versus human interpretation. Its integration into clinical workflows offers a low-cost, high-impact strategy to improve early osteoporosis diagnosis and treatment within the healthcare system.

Keywords Artificial intelligence · Vertebral fractures · Computed tomography · Opportunistic screening · Osteoporosis · Diagnostic accuracy · Endocrinology · Preventive medicine

✉ Polyzois Makras
pmakras@gmail.com; polyzois.makras@haf.gr

¹ Department of Endocrinology & Diabetes and Department of Medical Research, 251 Hellenic Air Force & VA General Hospital, 3 Kanellopoulou St, 11525 Athens, Greece

² Department of Medical Research, 251 Hellenic Air Force & VA General Hospital, 11525 Athens, Greece

³ First Department of Propaedeutic and Internal Medicine, Endocrinology Unit, Medical School, National and Kapodistrian University of Athens, 11527 Athens, Greece

⁴ Hospital Command, 251 Hellenic Air Force & VA General Hospital, 11525 Athens, Greece

⁵ Department of Radiology, Section of Magnetic Resonance Imaging, 251 Hellenic Air Force & VA General Hospital, 11525 Athens, Greece

Introduction

Osteoporosis is a silent, progressive disease affecting an estimated 200 million people worldwide [1]. It remains underdiagnosed and undertreated, particularly when fragility fractures are not clinically apparent [2, 3]. Vertebral fractures (VFs) are the most common osteoporotic fractures and represent unequivocal evidence of bone fragility. Despite their prognostic importance, they are often incidental findings and are frequently missed in routine imaging performed for non-musculoskeletal indications. Studies have demonstrated that as many as 60–70% of VFs visible on chest or abdominal imaging are not reported by radiologists [4–6].

This diagnostic oversight has profound implications. Each missed VF represents a missed opportunity for

intervention, increasing the risk of subsequent hip or VFs and associated mortality [7]. Recent international and Greek osteoporosis guidelines emphasize that even a single low-energy morphometric VF constitutes a definitive diagnosis of osteoporosis, mandating immediate initiation of pharmacological treatment regardless of bone mineral density [8–10]. Identifying such fractures opportunistically could thus shift the paradigm from reactive to preventive osteoporosis care in Greece, a country with a grossly estimated osteoporosis treatment gap of 60% [11].

Recent years have witnessed an exponential rise in the number of radiological examinations per capita, amplifying the need for automated tools that can assist human readers and standardize reporting accuracy [12–15].

Artificial intelligence (AI)-based initial investigations have demonstrated comparable or superior diagnostic accuracy to that of experienced radiologists in fracture detection, particularly when AI acts as an assistive “second reader” [16–18].

Beyond accuracy, AI introduces a paradigm of opportunistic screening, leveraging existing imaging data acquired for other indications to detect conditions such as osteoporosis [19–21], sarcopenia [22–24], or cardiovascular calcification [23, 25]. This approach allows clinicians to extract additional diagnostic information from routine studies, enhancing value without increasing radiation exposure or cost. Such a model of preventive and predictive medicine radiates the vision that AI and wellbeing medicine are defining elements of twenty-first century endocrinology and healthcare delivery [26].

This is the first pilot study in Greece to evaluate the diagnostic performance of an AI system for opportunistic vertebral fracture detection in routine computer tomography (CT) scans, comparing it directly with both general radiologists and endocrinologists. The study also explores the implications of AI integration for early osteoporosis diagnosis and treatment initiation within the Greek national healthcare framework.

Methods

Study design and setting

This was a prospective, single-center pilot study conducted at the 251 Hellenic Air Force General Hospital in Athens, Greece, between April 14 and June 16, 2025, by the Department of Medical Research. The institution provides tertiary-level services to both military personnel and civilians and includes an established bone metabolism clinic specializing in osteoporosis.

The study assessed the performance of the Bone Solution HealthOST (Nanox AI Ltd., Tel Aviv, Israel), a CE-marked

AI platform developed for automated detection of vertebral height loss (or compression fractures) in chest and abdominal CT scans. The software was provided free of charge for the duration of the pilot through a bilateral institutional agreement. The AI model was trained on large-scale pseudoanonymized datasets encompassing over 100,000 CT slices, enabling recognition of vertebral deformities based on morphometric and textural features.

Study objectives

The primary objective of the study was to compare the overall ability of AI, radiologists, and endocrinologists to correctly discriminate CT examinations with or without fractures.

Secondary objectives were the following:

1. To compare the ability of the AI system to “flag” VFs on routine CT examinations performed for other than osteoporosis reasons with the corresponding routine radiology reports produced by general radiologists in their everyday clinical practice.
2. To compare AI-based fracture flagging with targeted vertebral fracture assessment performed by endocrinologists evaluating and treating patients with osteoporosis in a designated outpatient clinic.
3. To evaluate the clinical translation of AI findings by determining how many patients with AI-detected previously unreported vertebral fractures would qualify—based on the Greek Osteoporosis Guidelines—for immediate osteoporosis treatment irrespective of BMD testing.

Study population and data processing

All thoracic and abdominal CT scans of patients aged ≥ 50 years performed during the 2-month period were retrieved from the hospital picture archiving communication system (PACS). Both contrast and non-contrast scans were eligible. When repeated CT scans were performed during a patient’s hospital admission, only the first scan was evaluated (i.e., up to one abdominal and/or chest CT per patient). CT scans obtained for the same patient and anatomical site more than once during the study period were considered as separate examinations only if the time interval between them exceeded 30 days. Exclusion criteria included suboptimal image quality, missing radiology reports, or incomplete spine visualization.

All DICOM files were anonymized and uploaded to the cloud-based AI platform. Following automated analysis, AI-generated reports were returned and integrated into the local database. Each AI report included a classification of “fracture” or “no fracture” along with localization on sagittal

reformats and the density of the vertebrae in Hounsfield units (HU) (Fig. 1); included fractures were either Grade 2 or Grade 3 according to the semi-quantitative visual grading of vertebral fractures [27].

During the study period, 427 CT examinations corresponding to 331 individual patients were included in the analysis. Some patients contributed more than one CT examination, most commonly when both thoracic and abdominal CT scans were performed during the same clinical evaluation. In accordance with the study protocol, repeated CT scans of the same anatomical region were considered separate examinations only when the time interval between them exceeded 30 days. Because the primary objective of the study was to evaluate the diagnostic performance of the AI system on individual CT examinations, statistical analyses were conducted at the per-examination level.

Human readers and reference standard

Three diagnostic modalities were compared, as follows:

1. Routine radiology reports – generated by general radiologists unaware of the actual study period.
2. Endocrinologist review – conducted by endocrinologists evaluating and treating osteoporosis patients who systematically examined each scan solely for vertebral deformities.

3. AI system output – the automated classification produced by the Bone Solution HealthOST platform.

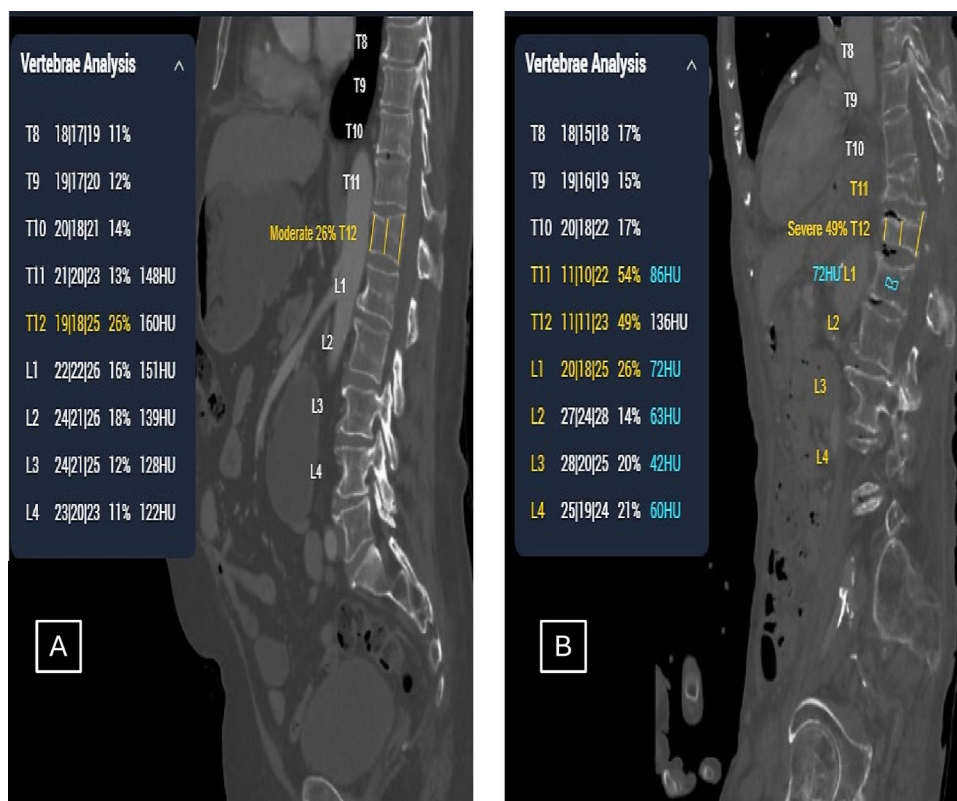
The verification of the fracture diagnosis which served as the “gold standard” was established through consensus between a senior radiologist (NK) and an endocrinologist (PM) specialized in bone diseases, both blinded to the AI and human readings, during joint review to produce the final ground-truth dataset.

The study did not include an absolute quality control of the diagnostic classifications. In particular, it was not assessed whether, for example, the AI and/or the radiologists or endocrinologists correctly graded a fracture as Grade 2 or Grade 3, or whether they accurately identified all existing fractures in each CT scan—such as, for example, all three present fractures rather than only one or two. Therefore, the study’s design allowed assessment of AI as a platform for opportunistic osteoporosis case-finding in routine clinical practice rather than strictly as a diagnostic tool.

Statistical analysis

Statistical analyses were conducted to evaluate and compare the diagnostic performance of the AI system with that of human readers (routine radiology reports and endocrinologist review) in identifying vertebral fractures on CT examinations. Diagnostic performance metrics were calculated

Fig. 1 Snapshots from AI output. Panel A: Identification of a Th12 moderate fracture. Panel B: identification of a Th12 severe fracture in a patient also hosting a Th11 fracture that was also evident in the specific AI output (but not in this snapshot)



using 2×2 contingency tables derived from the reference standard, including sensitivity and specificity. In addition, overall accuracy, balanced accuracy, F1-score, and Youden's J index were calculated to provide a more comprehensive evaluation of diagnostic performance. These measures were calculated separately for the AI system, radiologists, and endocrinologists. Ninety-five percent confidence intervals (95% CIs) for diagnostic performance metrics were calculated using exact binomial (Clopper-Pearson) methods. As all readers evaluated the same CT examinations, diagnostic outcomes represent paired observations. Therefore, comparisons of diagnostic performance between readers (AI vs radiologists and AI vs endocrinologists) were performed using McNemar's test for paired proportions. For completeness and consistency with the original analysis, chi-square tests were performed as secondary comparisons. P-values were derived from McNemar's test, which was considered the primary method for statistical comparison given the paired study design. To quantify the magnitude of differences between readers, odds ratios (ORs) were calculated from contingency tables comparing correct fracture identification relative to the reference standard and are reported as measures of effect size. The contingency tables used for these calculations are explicitly provided in a supplementary file to ensure transparency of the analysis. Additional prevalence-independent diagnostic performance measures were also calculated, including the positive likelihood ratio (LR+) and negative likelihood ratio (LR−). In instances where contingency tables contained zero cells (e.g., endocrinologist specificity), exact statistical methods were applied to avoid instability of conventional Wald estimates. Since some patients contributed more than one CT examination during the study period, the analysis was conducted at the per-examination level, reflecting the study objective of evaluating AI performance on individual CT scans. However, to account for potential clustering due to multiple CT examinations per patient, a sensitivity analysis was performed including only one CT scan per patient (the earliest available examination). Diagnostic performance metrics were recalculated in this restricted dataset to assess the robustness of the primary findings. All statistical analyses were performed using Python (SciPy, NumPy, and Matplotlib). Statistical significance was defined as $p < 0.050$ and all p-values are reported to three decimal places (or as $p < 0.001$ when appropriate).

Ethical considerations

The study was approved by the Institutional Review Board of the 251 Hellenic Air Force General Hospital (IRB No: F.076/6828/S.3128/4–6-24). All patient data were anonymized before analysis and the IRB waived the requirement of written informed consent. The study was conducted in accordance with the Declaration of Helsinki (1975, revised 2013). No external funding was received for this project.

Results

Study population

In total, 699 CTs were initially screened; 427 scans met the inclusion criteria and were analyzed, corresponding to 331 patients (136 females, mean age: 72.9 ± 13.5 years). Of the 427 included CT scans, 52% were thoracic and 48% abdominal. Indications for imaging included oncologic investigation and/or follow-up, known or suspected pulmonary disease, and chronic or acute abdominal pathology. None of the CTs were performed specifically for evaluation of clinically suspected fragility fractures.

AI vs. human readers

With respect to the primary objective—comparing the ability to correctly discriminate CT examinations with or without fractures (AI vs. radiologists and AI vs. endocrinologists)—these analyses aimed to assess the sum of both “positive” and “negative” findings regarding the presence of fracture(s). In this per-examination analysis (detection of existing fractures in the whole cohort), AI detected vertebral fractures with higher sensitivity (86.3% [95% CI: 76.7–92.9]) than radiologists (50.0% [95% CI: 38.6–61.4]) and endocrinologists (68.8% [95% CI: 57.4–78.7]) (Table 1, Fig. 2, Supplementary Table 1). Using McNemar's test for paired comparisons of overall classification performance (correct vs incorrect identification of fractures), AI significantly outperformed both radiologists and endocrinologists (both $p < 0.001$; Supplementary Table 2).

Specificity was high across all readers, reaching 98.3% (95% CI: 96.3–99.4) for AI, 99.7% (95% CI: 98.4–100.0) for radiologists, and 100.0% (95% CI: 98.9–100.0) for

Table 1 Diagnostic performance of AI and human readers

Group	Sensitivity % (95% CI)	Specificity % (95% CI)	Accuracy % (95% CI)	Balanced accuracy	F1-score	Youden's J
AI	86.3 (76.7–92.9)	98.3 (96.3–99.4)	96.0 (93.7–97.7)	0.92	0.86	0.83
Radiologists	50.0 (38.6–61.4)	99.7 (98.4–100.0)	90.4 (87.2–93.0)	0.70	0.57	0.40
Endocrinologists	68.8 (57.4–78.7)	100.0 (98.9–100.0)	94.1 (91.5–96.2)	0.81	0.73	0.62

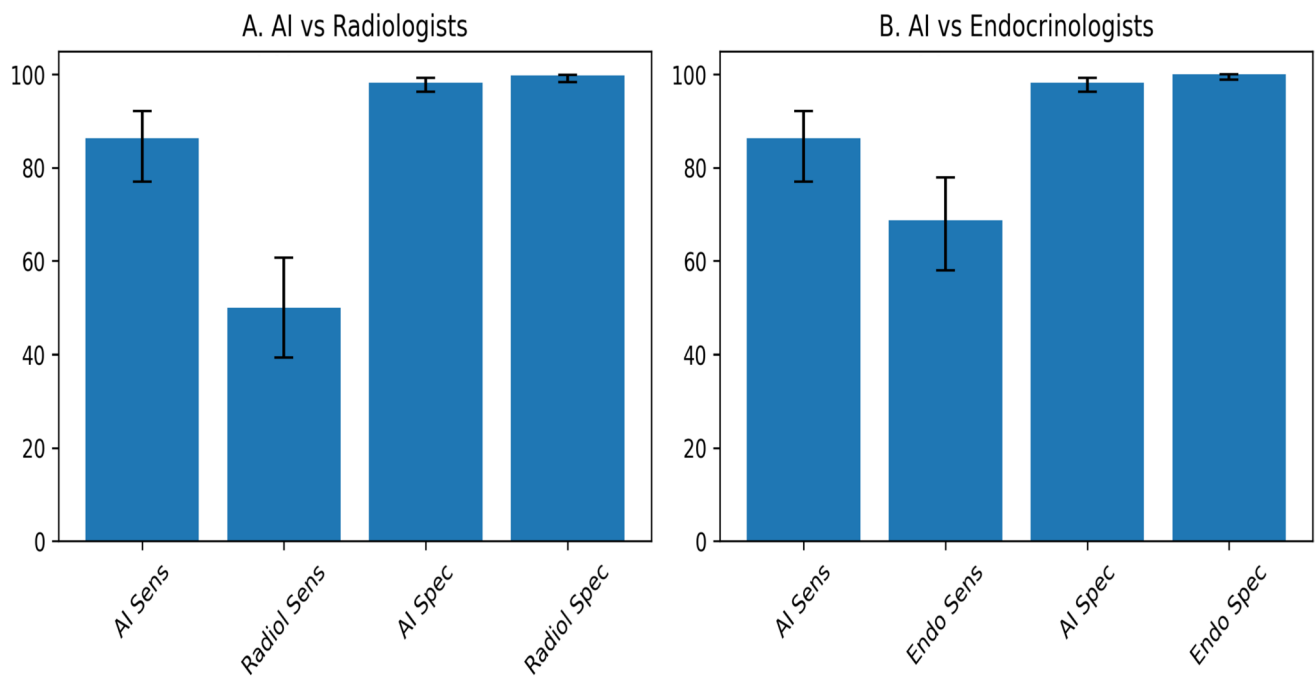


Fig. 2 Diagnostic performance of AI compared with human readers. Bars represent sensitivity and specificity with exact 95% confidence intervals. Paired comparisons of overall classification performance (correct vs incorrect identification of fractures) using McNemar's test

endocrinologists. Positive predictive value was 92.0% (95% CI: 83.4–97.0) for AI, 97.6% (95% CI: 87.1–99.9) for radiologists, and 100.0% (95% CI: 93.5–100.0) for endocrinologists, while negative predictive value was 96.9% (95% CI: 94.5–98.4), 89.6% (95% CI: 86.2–92.5), and 93.3% (95% CI: 90.2–95.6), respectively. Compared with radiologists, AI showed 6.3-fold higher odds of correctly identifying true fractures (OR=6.30; 95% CI: 2.90–13.58) (Supplementary Table 3). In the analogous per-examination comparison between the AI system and endocrinologists, the odds of the AI system identifying existing vertebral fractures were 2.85-fold higher (OR=2.85; 95% CI: 1.29–6.30) (Supplementary Table 4).

In absolute terms, the AI system detected 36.3% more vertebral fractures than routine radiology reports and 17.5% more fractures than endocrinologist review.

Across additional performance metrics, AI demonstrated overall higher accuracy (96%), balanced accuracy (92%), F1-score (0.86), and Youden's J index (0.83) compared with radiologists (90.4%, 70%, 0.57, and 0.40, respectively) and endocrinologists (94.1%, 81%, 0.73, and 0.62, respectively) (Table 1).

Additional prevalence-independent diagnostic metrics were also calculated. The AI system showed a positive likelihood ratio (LR+) of approximately 49.8 and a negative likelihood ratio (LR-) of 0.14, indicating strong diagnostic discrimination. In comparison, radiologists demonstrated

showed that AI significantly outperformed both radiologists and endocrinologists (both $p < 0.001$). Abbreviations: Spec: specificity; Sens: sensitivity; Endo: endocrinologists; Radiol: radiologists; AI: artificial intelligence

an LR+ of approximately 172.4 and LR- of 0.50, while endocrinologists showed a very high LR+ due to their absence of false-positive classifications and an LR- of approximately 0.31.

The analysis of the study's secondary endpoint of flagging performance was essentially designed to evaluate the ability to detect (flag) the presence of one or more VF in each CT examination that was ultimately confirmed by the gold standard evaluation (case finding). Accordingly, this analysis focused only on the assessment of "positive" findings related to the existence of a fracture and did not include the negative diagnoses.

A total of 80 examinations were flagged as having ≥ 1 vertebral fracture either by the AI system and/or by radiologists in their routine clinical reports. In the evaluation of flagging performance among CT examinations with true vertebral fractures as assessed by the expert panel, the AI system demonstrated significantly higher paired case-finding sensitivity than radiologists (McNemar's test, $p < 0.001$), with an OR of 14.23 (95% CI: 4.7–43.1) (Fig. 3a) (Supplementary Table 5).

The same comparative approach was applied to the fracture-flagging performance between the AI system and endocrinologists. In this analysis, the AI system again showed significantly higher paired case-finding sensitivity than endocrinologists (McNemar's test, $p < 0.001$), with an OR of 4.77 (95% CI: 1.6–13.6) (Fig. 3b) (Supplementary Table 6).

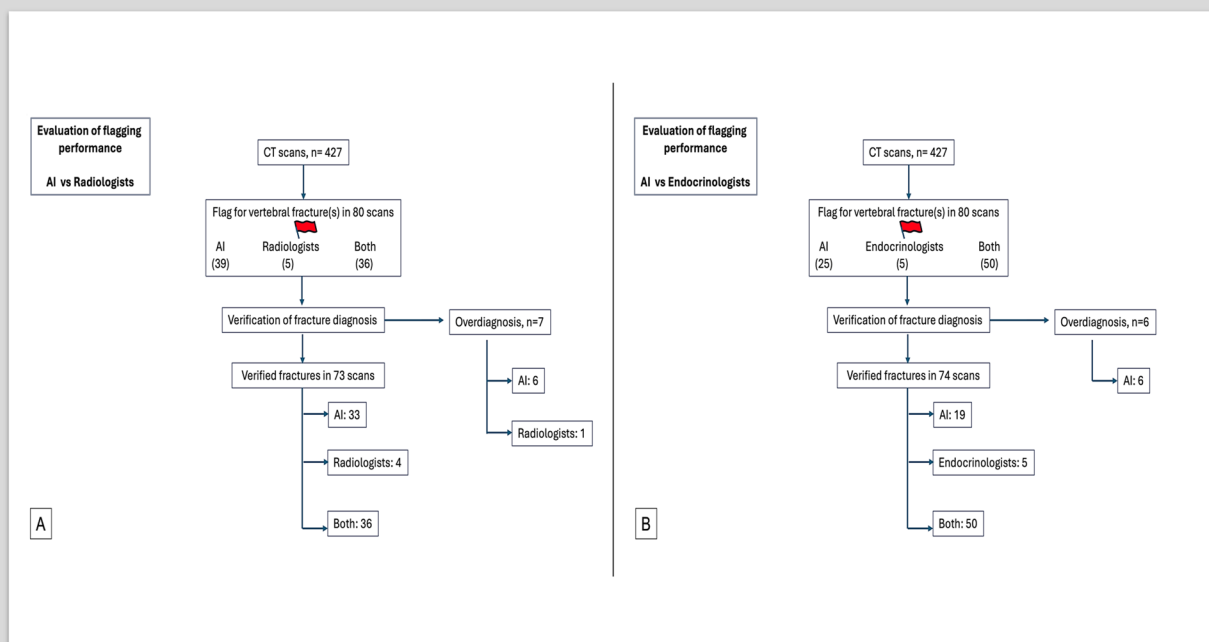


Fig. 3 Evaluation of flagging performance: comparison AI Vs Radiologists (Panel A) and comparison AI Vs Endocrinologists (Panel B). Note: The scans flagged for fracture(s) from radiologists were not

Finally, a sensitivity analysis restricted to one CT examination per patient yielded consistent results. In this per-patient analysis, AI maintained higher sensitivity (88.9%) compared with radiologists (52.4%) and endocrinologists (71.4%), with no material change in the overall pattern of results.

Missed fractures and clinical translation

AI uniquely detected 33 CT scans with true fractures missed by the routine radiology reports, which corresponded to 30 individual patients. Among these patients, 23 had never received osteoporosis therapy, four patients were currently under treatment, and three had discontinued treatment more than 3 years earlier. According to the Greek guidelines [10], these 26 individuals required immediate pharmacological therapy independent of BMD testing.

Extrapolating this 2-month detection rate across 1 year suggested that routine AI deployment in this single hospital setting could identify approximately 156 new treatable osteoporosis cases annually (95% CI: 102–229)—patients who would otherwise remain undiagnosed and at elevated risk of future fractures. However, this figure is based on the general assumption of a constant CT volume, a constant prevalence of osteoporotic vertebral fractures, and, certainly

identical with those flagged by the endocrinologists. The number of 80 scans in both instances is simply a coincidence

no seasonal variation and, therefore, needs to be proved in a relevant dedicated annual evaluation.

Anatomical distribution and error analysis

The majority of fractures were located between Th7 and L4, while the vertebrae with the highest fracture frequency were Th12 and L1 consistent with mechanical stress points.

A subgroup analysis according to CT examination type was also performed. The diagnostic advantage of the AI system over routine radiology reporting was observed in both thoracic and abdominal CT examinations. In abdominal CT scans, AI demonstrated a sensitivity of 84.4% compared with 53.1% for radiologists, while in thoracic CT scans the corresponding sensitivities were 85.7% and 47.6%, respectively. These findings suggest that the improved fracture detection associated with AI-assisted analysis was consistent across different CT imaging contexts (Supplementary Table 7).

AI missed a few true fractures in cases with severe scoliosis while occasionally overcalled mild degenerative deformities in older patients, explaining the small number of false positives (Fig. 2). However, the majority of discordant readings between AI and radiologists represented genuine missed fractures rather than overdiagnosis.

Discussion

This pilot study demonstrates that a commercially available AI system can markedly outperform both general radiologists and endocrinologists in the opportunistic detection of vertebral fractures. The magnitude of improvement in case finding—over 14-fold higher odds compared with standard reporting—highlights a significant and actionable diagnostic gap in current clinical practice.

In the present study, routine radiology reports missed approximately 40% of vertebral fractures identified by the reference standard. This observation is consistent with previous reports indicating that 60–70% of incidental vertebral fractures visible on routine chest or abdominal imaging remain unreported in clinical practice [4–6]. The higher detection rate observed among endocrinologists compared with radiologists likely reflects the focused expertise of physicians working in a specialized bone clinic, where vertebral fracture recognition is part of routine clinical assessment, and their dedicated involvement in this specific study which was solely the evaluation of CTs for vertebral fractures. In contrast, general radiologists often interpret CT examinations primarily in relation to the clinical indication for imaging and may not systematically evaluate vertebral morphology when osteoporosis or fracture is not part of the referral question. Another potential contributor to missed vertebral fractures is the clinical context in which many CT scans are interpreted. Although exact data were not systematically recorded, we estimate that approximately 30% of the examinations were performed in an emergency or urgent care setting. In such circumstances, radiologists are primarily focused on identifying acute or life-threatening findings related to the clinical indication for imaging, while chronic conditions such as vertebral fractures may receive less attention. This workflow characteristic is an important factor contributing to underreporting incidental vertebral fractures.

Every missed vertebral fracture represents a missed opportunity for secondary prevention. In most osteoporosis guidelines, the identification of a low-energy vertebral fracture establishes a diagnosis of osteoporosis requiring treatment initiation [8–10]. Early recognition of non-clinical (morphometric) vertebral fractures allows clinicians to intervene before subsequent fractures occur, improving patient outcomes and reducing healthcare costs. This principle aligns with broader international recommendations advocating for opportunistic case finding through routine imaging [28, 29]. Recent reports and meta-analyses demonstrated that AI tools consistently achieve pooled sensitivities of around 90% across anatomical regions and yield maximal benefit when combined with human oversight

[4, 16–18]. Additionally, explainable foundation models capable of opportunistic osteoporosis screening from chest X-rays have recently been developed, achieving high area under the curve (AUC) values without reliance on bone density data [30]. Our findings reinforce these results in a real-world setting, further supporting the reliability of CE-marked AI systems when integrated into clinical workflows.

In practice, the integration of AI into PACS could enable automatic flagging of suspected vertebral fractures, prompting radiologists or clinicians to review the flagged images. Such an approach would require minimal human input and no additional imaging. Moreover, automated alerts could be linked to hospital electronic medical records, facilitating direct referral to osteoporosis clinics for therapy initiation transforming the CT imaging into an active screening tool rather than merely a targeted diagnostic test. Future iterations of systems like the one we used in our pilot study could adopt similar visual interpretability modules, displaying heatmaps over detected vertebral deformities to further aid user verification.

From a population perspective, opportunistic vertebral fracture detection could dramatically reduce the “osteoporosis care gap” [31]. In Greece, where adherence to screening and treatment remains suboptimal [11], automatic identification of high-risk individuals during routine CT imaging offers a sustainable, cost-neutral intervention. Even modest improvements in detection rates could translate into hundreds of prevented fractures annually, reducing morbidity, mortality, and healthcare expenditure. In this regard and specifically for the 251 Hellenic Air Force General Hospital in which the first Greek Fracture Liaison Service (FLS) was implemented, a total of 213 incident fragility fractures were identified during the first year period, comprising approximately 26 vertebral, 51 hip, and 136 non-vertebral fractures [32]. In this context, the projected annual detection of approximately 150 previously unrecognized, treatment-eligible patients through routine AI-assisted CT analysis represents a clinically meaningful expansion of case-finding capacity. Within an established FLS framework, such an increase could nearly double the number of vertebral and other major osteoporotic fractures captured and appropriately referred for secondary fracture prevention, thereby significantly amplifying the clinical impact and effectiveness of the FLS model.

Beyond diagnostic efficiency, the implications of AI extend to the evolving concept of precision and wellbeing medicine. AI appears to be not only a technological novelty but a fundamental enabler of predictive endocrinology where data-driven insights enhance disease prevention and health maintenance [26]. In this context, the integration of

AI into bone health management represents an important transition toward holistic, anticipatory care.

This study has several limitations that should be acknowledged. First, it was conducted in a single tertiary-care hospital over a relatively short 2-month observation period and therefore the findings may not be fully generalizable to other healthcare settings or populations. Second, the analysis was performed at the per-examination level and some patients contributed more than one CT scan to the dataset. Although this approach reflects the objective of evaluating AI performance on individual CT examinations, it may introduce a degree of statistical dependence between observations originating from the same patient. Third, the reference standard was established through consensus evaluation between two experienced specialists rather than through independent adjudication by multiple readers; however, this approach reflects common practice in diagnostic accuracy studies. Fourth, the clinical context in which CT scans were interpreted may also have influenced fracture detection rates. Although exact data were not systematically recorded, we estimate that approximately 30% of examinations were performed in an emergency or urgent care setting where radiologists are typically focused on identifying acute or life-threatening findings rather than chronic conditions such as vertebral fractures. Finally, the extrapolation of the 2-month case-finding rate to an annual estimate of newly identified treatable patients assumes relatively stable CT volumes and fracture prevalence over time and therefore should be interpreted cautiously. Larger, multicenter studies with longer observation periods are needed to confirm the generalizability and long-term clinical impact of AI-assisted vertebral fracture detection.

Future research should prioritize multicenter validation of AI-assisted vertebral fracture detection across diverse patient populations to ensure generalizability and robustness. Equally important is the integration of AI outputs with established clinical tools, such as FRAX scores and BMD measurements, to develop hybrid risk-stratification models with superior predictive performance. Rigorous health-economic analyses are also needed to quantify the potential cost savings arising from earlier diagnosis and subsequent fracture prevention. In parallel, the development of explainable AI frameworks will be essential to enhance algorithmic transparency and strengthen clinician trust. Finally, evidence generated from these efforts should inform national-level policy frameworks that support the widespread implementation of AI-assisted opportunistic screening within routine clinical practice.

In conclusion, in this first Greek pilot study, AI achieved significantly higher sensitivity and diagnostic yield for vertebral fracture detection than both radiologists and

endocrinologists. Given the central role of radiologists in the diagnostic procedure, AI integration into radiology workflows could convert routine CT scans into powerful, low-cost screening instruments for osteoporosis thereby facilitating early detection and treatment.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42000-026-00784-1>.

Acknowledgements The contributors of the 251 Hellenic Air Force & VA General Hospital Radiology Department reporting group are as follows: Christos Tekedis, Vasilliki Mparmparousi, Ioannis Liakopoulos, Dimitrios Matsaridis, Emmanouil Skoulidakis, Maria Varela, Harikleia Athanasiou, Stella Koropouli, and Nikolaos Gidararakos.

Author contributions Conceptualization: Polyzois Makras; Methodology: Polyzois Makras and Nikolaos Kyriakopoulos; Formal analysis and investigation: Polyzois Makras, Eleftherios Chatzellis, Christos Gravvanis, Georgia Kanti, and Nikolaos Kyriakopoulos; Data analysis and statistics: Polyzois Makras and Maria P. Yavropoulou; Writing—original draft preparation: Polyzois Makras; Writing—review and editing: Polyzois Makras, Eleftherios Chatzellis, Christos Gravvanis, Georgia Kanti, Maria P. Yavropoulou, Konstantinos Papadimitropoulos, and Nikolaos Kyriakopoulos; Supervision: Polyzois Makras and Konstantinos Papadimitropoulos.

CT reviews in routine clinical practice: Radiology Department reporting group.

Funding No funding was received for this study.

Data Availability The study's data are available upon reasonable and legitimate request.

Declarations

Ethical approval The study was approved by the Institutional Review Board of the 251 Hellenic Air Force General Hospital (IRB No: F.076/6828/S.3128/4–6–24).

Informed consent All patient data were anonymized before analysis and the IRB waived the requirement of written informed consent.

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C (2020) The epidemiology of osteoporosis. *Br Med Bull* 133(1):105–117. <https://doi.org/10.1093/bmb/ldaa005>
2. Curtis EM, Dennison EM, Cooper C, Harvey NC (2022) Osteoporosis in 2022: care gaps to screening and personalised medicine. *Best Pract Res Clin Rheumatol* 36(3):101754. <https://doi.org/10.1016/j.berh.2022.101754>
3. Kanis JA, Norton N, Harvey NC, Jacobson T, Johansson H, Lorentzon M, McCloskey EV, Willers C, Borgstrom F (2021) SCOPE 2021: a new scorecard for osteoporosis in Europe. *Arch Osteoporos* 16(1):82. <https://doi.org/10.1007/s11657-020-00871-9>

4. Husarek J, Hess S, Razaiean S, Ruder TD, Sehmisch S, Muller M, Liodakis E (2024) Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy. *Sci Rep* 14(1):23053. <https://doi.org/10.1038/s41598-024-73058-8>
5. Muller M, Schechter CB, Hautz WE, Sauter TC, Exadaktylos AK, Stock S, Birrenbach T (2021) The development and validation of a resource consumption score of an emergency department consultation. *PLoS One* 16(2):e0247244. <https://doi.org/10.1371/journal.pone.0247244>
6. Bruls RJM, Kwee RM (2020) Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights Imaging* 11(1):121. <https://doi.org/10.1186/s13244-020-00925-z>
7. Toth E, Banefelt J, Akesson K, Spangeus A, Ortsater G, Libanati C (2020) History of previous fracture and imminent fracture risk in Swedish women aged 55 to 90 years presenting with a fragility fracture. *J Bone Miner Res* 35(5):861–868. <https://doi.org/10.1002/jbmr.3953>
8. Kanis JA, Cooper C, Rizzoli R, Reginster JY, Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis, Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (2019) European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporos Int* 30(1):3–44. <https://doi.org/10.1007/s00198-018-4704-5>
9. Shoback D, Rosen CJ, Black DM, Cheung AM, Murad MH, Eastell R (2020) Pharmacological management of osteoporosis in postmenopausal women: an Endocrine Society guideline update. *J Clin Endocrinol Metab*. <https://doi.org/10.1210/clinem/dgaa048>
10. Makras P, Anastasilakis AD, Antypas G, Chronopoulos E, Kaskanis EG, Matsouka A, Patrikos DK, Stathopoulos KD, Tournis S, Trovas G, Kosmidis C (2019) The 2018 guidelines for the diagnosis and treatment of osteoporosis in Greece. *Arch Osteoporos* 14(1):39. <https://doi.org/10.1007/s11657-019-0584-3>
11. Yavropoulou MP, Makras P, Athanasakis K, Bournia VK, Mathioudakis K, Tsolakidis A, Kassi E, Kaltsas G, Mitrou P, Sfrikakis PP (2022) Prevalence and patterns of anti-osteoporotic drug use based on 2019 real-world nationwide data in Greece. *Arch Osteoporos* 17(1):86. <https://doi.org/10.1007/s11657-022-01126-5>
12. Di Somma S, Paladino L, Vaughan L, Lalle I, Magrini L, Magranti M (2015) Overcrowding in emergency department: an international issue. *Intern Emerg Med* 10(2):171–175. <https://doi.org/10.1007/s11739-014-1154-8>
13. Lantsman CD, Barash Y, Klang E, Guranda L, Konen E, Tau N (2022) Trend in radiologist workload compared to number of admissions in the emergency department. *Eur J Radiol* 149:110195. <https://doi.org/10.1016/j.ejrad.2022.110195>
14. Smith E, Dasan S (2018) A system under pressure. *Br J Hosp Med (Lond)* 79(9):495–499. <https://doi.org/10.12968/hmed.2018.79.9.495>
15. Mdodo R, Frazier EL, Dube SR, Mattson CL, Sutton MY, Brooks JT, Skarbinski J (2015) Cigarette smoking prevalence among adults with HIV compared with the general adult population in the United States: cross-sectional surveys. *Ann Intern Med* 162(5):335–344. <https://doi.org/10.7326/M14-0954>
16. Canoni-Meynet L, Verdout P, Danner A, Calame P, Aubry S (2022) Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. *Diagn Interv Imaging* 103(12):594–600. <https://doi.org/10.1016/j.diii.2022.06.004>
17. Yang S, Yin B, Cao W, Feng C, Fan G, He S (2020) Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. *Clin Radiol* 75(9):713.e17–713.e28. <https://doi.org/10.1016/j.crad.2020.05.021>
18. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, Stewart M, Collins GS, Furniss D (2022) Artificial Intelligence in fracture detection: a systematic review and meta-analysis. *Radiology* 304(1):50–62. <https://doi.org/10.1148/radiol.211785>
19. Yang J, Liao M, Wang Y, Chen L, He L, Ji Y, Xiao Y, Lu Y, Fan W, Nie Z, Wang R, Qi B, Yang F (2022) Opportunistic osteoporosis screening using chest CT with artificial intelligence. *Osteoporos Int* 33(12):2547–2561. <https://doi.org/10.1007/s00198-022-06491-y>
20. Wu Y, Yang X, Wang M, Lian Y, Hou P, Chai X, Dai Q, Qian B, Jiang Y, Gao J (2025) Artificial intelligence assisted automatic screening of opportunistic osteoporosis in computed tomography images from different scanners. *Eur Radiol* 35(4):2287–2295. <https://doi.org/10.1007/s00330-024-11046-2>
21. Behanova M, Sokhan A, Haschka J, Zandieh S, Salzlechner C, Ljuhar R, Zwerina J, Kocijan R (2026) AI-supported opportunistic detection of vertebral fractures on routine CT scans: diagnostic performance and clinical relevance. *Bone* 203:117735. <https://doi.org/10.1016/j.bone.2025.117735>
22. Urooj B, Ko Y, Na S, Kim IO, Lee EH, Cho S, Jeong H, Khang S, Lee J, Kim KW (2025) Implementation of fully automated AI-integrated system for body composition assessment on computed tomography for opportunistic sarcopenia screening: multicenter prospective study. *JMIR Form Res* 9:e69940. <https://doi.org/10.2196/69940>
23. Pickhardt PJ, Correale L, Hassan C (2023) AI-based opportunistic CT screening of incidental cardiovascular disease, osteoporosis, and sarcopenia: cost-effectiveness analysis. *Abdom Radiol (NY)* 48(3):1181–1198. <https://doi.org/10.1007/s00261-023-03800-9>
24. Ryu J, Eom S, Kim HC, Kim CO, Rhee Y, You SC, Hong N (2023) Chest X-ray-based opportunistic screening of sarcopenia using deep learning. *J Cachexia Sarcopenia Muscle* 14(1):418–428. <https://doi.org/10.1002/jcsm.13144>
25. Jeong J, Chao CJ, Arsanjani R, Ayoub C, Lester SJ, Pereyra M, Said EF, Roarke M, Tagle-Cornell C, Koepke LM, Tsai YL, Jung-Hsuan C, Chang CC, Farina JM, Trivedi H, Patel BN, Banerjee I (2025) Artificial intelligence chest X-ray opportunistic screening model for coronary artery calcium deposition: a multi-objective model with multimodal data fusion. *Mayo Clin Proc Digit Health* 3(4):100300. <https://doi.org/10.1016/j.mcpdig.2025.100300>
26. Stratakis CA (2025) Artificial intelligence and quality of life are two leading themes of contemporary medicine. *Hormones (Athens)* 24(3):589–591. <https://doi.org/10.1007/s42000-025-00723-6>
27. Genant HK, Wu CY, van Kuijk C, Nevitt MC (1993) Vertebral fracture assessment using a semiquantitative technique. *J Bone Miner Res* 8(9):1137–1148. <https://doi.org/10.1002/jbmr.5650080915>
28. Aggarwal V, Maslen C, Abel RL, Bhattacharya P, Bromiley PA, Clark EM, Compston JE, Crabtree N, Gregory JS, Kariki EP, Harvey NC, Ward KA, Poole KES (2021) Opportunistic diagnosis of osteoporosis, fragile bone strength and vertebral fractures from routine CT scans; a review of approved technology systems and pathways to implementation. *Ther Adv Musculoskelet Dis* 13:1759720X211024029. <https://doi.org/10.1177/1759720X211024029>
29. Reginster JY, Schmidmaier R, Alokail M, Hilgsmann M (2025) Cost-effectiveness of opportunistic osteoporosis screening using chest radiographs with deep learning in Germany. *Aging Clin Exp Res* 37(1):149. <https://doi.org/10.1007/s40520-025-03048-x>
30. Kim J, Kwak S, Lee H, Chang J, Park SM (2025) Explainable opportunistic osteoporosis screening from chest X-rays: a retrospective comparison of foundation models. *Osteoporos Int*. <https://doi.org/10.1007/s00198-025-07727-3>

31. Willers C, Norton N, Harvey NC, Jacobson T, Johansson H, Lorentzon M, McCloskey EV, Borgstrom F, Kanis JA, Scientific Report Panel of the IOF (2022) Osteoporosis in Europe: a compendium of country-specific reports. *Arch Osteoporos* 17(1):23. <https://doi.org/10.1007/s11657-021-00969-8>
32. Makras P, Panagoulia M, Mari A, Rizou S, Lyritis GP (2017) Evaluation of the first fracture liaison service in the Greek health-care setting. *Arch Osteoporos* 12(1):3. <https://doi.org/10.1007/s11657-016-0299-7>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Supplementary Table 1. Confusion matrices describing diagnostic performance of AI, radiologists, and endocrinologists relative to the reference standard.

Reader	True Positives	False Positives	True Negatives	False Negatives
AI	69	6	341	11
Radiologists	40	1	346	40
Endocrinologists	55	0	347	25

Supplementary Tables 2. Paired discordance tables used for McNemar tests comparing AI with radiologists and endocrinologists.

AI vs Radiologists

	Radiologist Positive	Radiologist Negative
AI Positive	40	35
AI Negative	1	351

AI vs Endocrinologists

	Endocrinologist Positive	Endocrinologist Negative
AI Positive	55	20
AI Negative	0	352

Supplementary Table 3. Contingency table for diagnostic performance of AI versus radiologists (used for odds ratio calculation).

	Positive exams		Negative exams	
	True	False	True	False
AI	69	6	341	11
Radiologists	40	1	346	40

Supplementary Table 4. Contingency table for diagnostic performance of AI versus endocrinologists (used for odds ratio calculation).

	Positive exams		Negative exams	
	True	False	True	False
AI	69	6	341	11
Endocrinologists	55	0	347	25

Supplementary Table 5. Contingency table for “flagging” performance of AI versus radiologists (used for odds ratio calculation).

Group	Flagged Real Fractures (TP)	Missed Real Fractures (FN)	Total Real Fractures
AI	69	4	73
Radiologists	40	33	73

TP: true positives

FN: False negative

Supplementary Table 6. Contingency table for “flagging” performance of AI versus endocrinologists (used for odds ratio calculation).

Group	Flagged Real Fractures (TP)	Missed Real Fractures (FN)	Total Real Fractures
AI	69	5	74
Endocrinologist	55	24	74

TP: true positives

FN: False negative

Supplementary Table 7. Diagnostic performance of AI, radiologists, and endocrinologists stratified by CT examination type.

CT Type	Reader	TP	FP	TN	FN	Sensitivity (%)	Specificity (%)
Abdominal CT	AI	27	2	145	5	84.4	98.6
Abdominal CT	Radiologists	17	2	145	15	53.1	98.6
Abdominal CT	Endocrinologists	23	3	144	9	71.9	98.0
Thoracic CT	AI	36	4	175	6	85.7	97.2
Thoracic CT	Radiologists	20	0	180	22	47.6	100
Thoracic CT	Endocrinologists	27	4	176	15	64.3	97.8
Thoracic + Abdominal CT	AI	6	0	21	0	100	100
Thoracic + Abdominal CT	Radiologists	3	0	15	1	75.0	100
Thoracic + Abdominal CT	Endocrinologists	4	0	15	0	100	100

TP: True positives, FP: False positives

TN: True negatives, FN: False negative