Abdominal Radiology (2026) 51:2631–2641
https://doi.org/10.1007/s00261-025-05213-2

RESEARCH

Advancing osteoporosis opportunistic screening: multicenter validation of a deep learning algorithm using abdominal CT scans

Augusto Sarquis Serpa[1,2,3] · Marcelo Straus Takahashi[4] · Eduardo Moreno Júdice de Mattos Farina[1,5] · Sean Bennett[3] · Nishith Khandwala[3] · Melissa Major[3] · Jack Paparian[3] · Fernanda Garozzo Velloni[5] · Steven Rothenberg[6] · Leonardo Kayat Bittencourt[7,8] · Ross Warren Filice[9] · John Mongan[10] · Felipe Campos Kitamura[1,3]

Received: 11 August 2025 / Revised: 13 September 2025 / Accepted: 16 September 2025 / Published online: 31 October 2025

Abstract

Objectives To develop and do multicenter validation on an algorithm that screens for osteoporosis from abdominal CTs.

Methods This is a diagnostic accuracy study with retrospective data from January 2022 to July 2022 consisting of two steps: a segmentation step of the lumbar vertebral bodies, involving outpatient non-contrast abdominal CTs from Diagnósticos da América S/A (Dasa), and a multicenter validation step incorporating data from four additional institutions. The segmentation employed a 2D UNet with a ResNet34 backbone. We determined the Pearson correlation coefficient (r) between the mean of the slices' mean attenuations (MSMA) on CT scans against the bone mineral density (BMD) on recent DEXA scans, calculated AUCs and performance metrics for osteoporosis prediction, including 95% confidence intervals, and evaluated calibration plots.

Results The multicenter validation included 504 participants (median age, 66 years, interquartile range 56–72; 388 women). A linear regression analysis showed an r of 0.62 (95% CI, 0.56–0.67) between MSMA (HU) and BMD (g/cm^2). The AUCs (95% CI) for distinguishing between normal and osteoporosis were 0.96 (0.89, 1.0) for the internal dataset and 0.82 (0.74, 0.89) for the external dataset, and the performance metrics (95% CI), for a globally optimized threshold of 202.6 HU, were 100% sensitivity (44, 100) and 91% specificity (84, 95) for internal data and 79% sensitivity (61, 90) and 81% specificity (76, 84) for external sites. MSMA was independently associated with osteoporosis on a mixed-effects logistic regression analysis. The model showed good calibration, with Brier score of 0.054.

Conclusion We developed and performed a multicenter validation of a DL model for osteoporosis prediction on CT.

✉ Augusto Sarquis Serpa
augusto.esd5@gmail.com

Marcelo Straus Takahashi
mst@ad.unc.edu

Eduardo Moreno Júdice de Mattos Farina
eduardofarina61@gmail.com

Sean Bennett
sean@bunkerhillhealth.com

Nishith Khandwala
nishith@bunkerhillhealth.com

Melissa Major
melissa.major@bunkerhillhealth.com

Jack Paparian
jackpaparian@gmail.com

Fernanda Garozzo Velloni
fernandavelloni@gmail.com

Steven Rothenberg
sarothenberg@uabmc.edu

Leonardo Kayat Bittencourt
Leonardo.KayatBittencourt@uhhospitals.org

Ross Warren Filice
ross.w.filice@medstar.net

John Mongan
john.mongan@ucsf.edu

Felipe Campos Kitamura
kitamura.felipe@gmail.com

[1] Departamento de Diagnóstico por Imagem, Federal University of São Paulo, São Paulo, Brazil

[2] Instituto de Radiologia, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil

[3] Bunkerhill Health, San Francisco, USA

[4] Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, USA

[5] Dasa, São Paulo, Brazil

[6] Department of Radiology, University of Alabama at Birmingham, Birmingham, USA

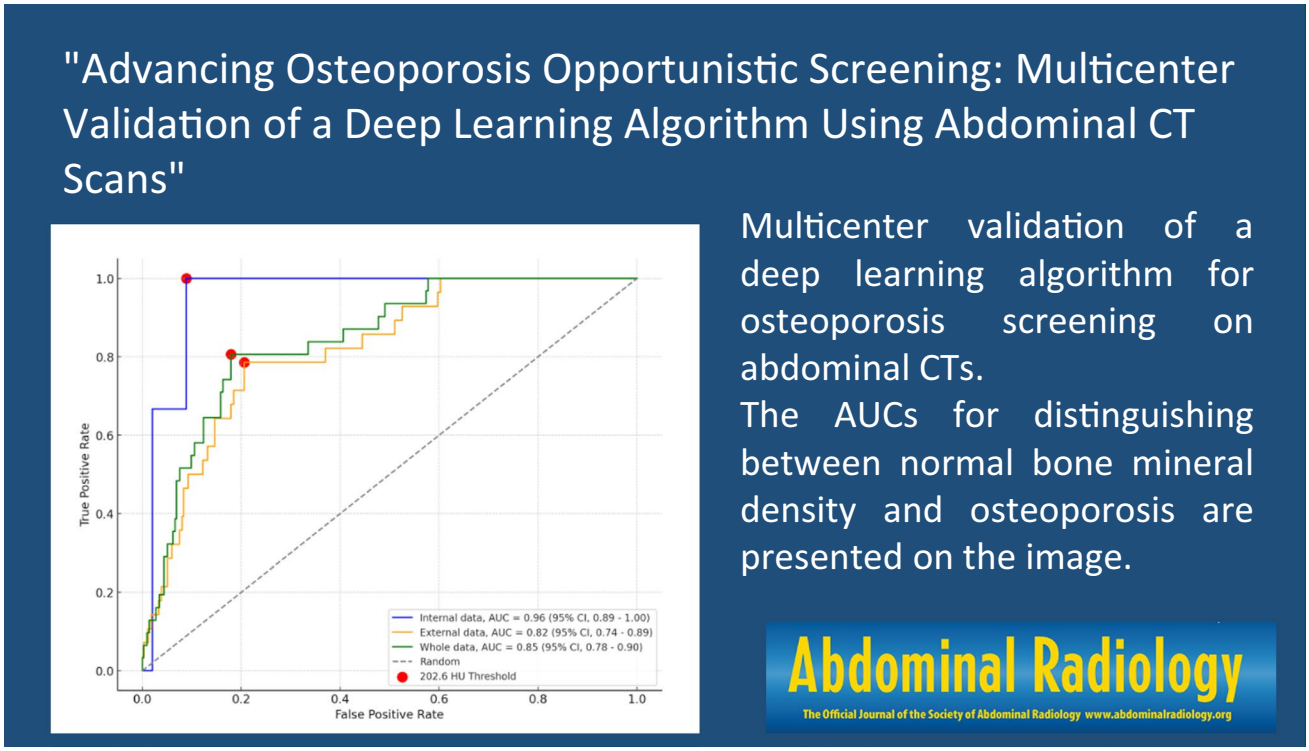
[7] Department of Radiology, University Hospitals Cleveland Medical Center, Cleveland, USA

[8] Case Western Reserve University, Cleveland, USA

[9] Department of Radiology, MedStar Georgetown University Hospital, Washington, DC, USA

[10] Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, USA

Graphical abstract



Keywords Lumbar spine · Osteoporosis · Computed tomography · Deep learning · Opportunistic screening

Introduction

Osteoporosis, a systemic skeletal disorder characterized by reduced bone mass and the microarchitectural deterioration of bone tissue [1], poses a significant public health challenge globally [2, 3]. It increasingly leads to a higher risk of fractures, particularly in the hip, spine, and wrist, significantly impacting the quality of life in the aging population worldwide. There has been an increase, from 1990 to 2019, in the total number of deaths attributable to low bone mineral density (LBMD) and LBMD-related fractures [4]. The silent nature of the disease, often remaining undetected until a fracture occurs, underlines the importance of early detection and intervention [5].

Traditional methods for diagnosing osteoporosis predominantly rely on DEXA scans to measure bone mineral density (BMD) [6]. However, the limited availability of DEXA in certain regions, coupled with its cost, poses a barrier to widespread osteoporosis screening [7].

The development of concept of opportunistic screening stems from efforts to reduce unnecessary diagnostic imaging and the pursuit of innovative strategies to enhance the utility of imaging procedures [8]. Specifically, abdominal (and thoracic) CT scans, have emerged as a promising tool for opportunistic osteoporosis screening [9–15]. Initially centered on manual measurement techniques, it has progressively evolved to embrace automated methods, particularly deep learning (DL) approaches [16–31].

Some osteoporosis screening CT methods are based on region-of-interest (ROI) placement in lumbar vertebrae, most commonly a circular or elliptical ROI within the trabecular compartment of L1 [12, 13, 18, 22]. While clinically useful, these ROI-based approaches may suffer from limited reproducibility, as they represent only a fraction of the vertebral trabecular volume and don't include cortical bone. Whole-lumbar volumetric methods might provide a more comprehensive assessment of bone attenuation and volumetric BMD [29–31]. These approaches not only improve reproducibility but also show strong correlations with fracture risk [29, 30].

Albeit these recent advances on osteoporosis detection with DL, very few algorithms have undergone external validation. Accordingly, this paper focuses on the development and multicenter validation of an algorithm designed to predict osteoporosis from abdominal CT scans.

Methods

This is a diagnostic accuracy study with retrospective data, approved by each local Institutional Review Board (IRB) with a waiver for informed consent. It was carried out through a partnership among multiple institutions centralized by the Bunkerhill Health consortium. The study is reported in accordance with the STARD 2015 guidelines [32]. A completed STARD checklist is provided in the Supplementary Material.

The methodology comprises two main components: a raw segmentation step and a multicenter validation step.

Raw segmentation step

The primary goal of this step was to delineate the vertebral bodies of the lumbar spine in non-contrast abdominal CT scans. Cortical and trabecular bone are included in the segmentations.

Sample selection, data split and annotations

We utilized a retrospective dataset of outpatient scans acquired between January and March 2022 from Diagnósticos da América S/A (Dasa), for various clinical indications. A convenience sample was used because there is no universally accepted method for sample size calculation in deep learning models. Inclusion criteria were patients aged 18 years and older. Scans showing movement artifacts were excluded. The data were divided into training, validation, and testing sets at a 70/20/10 ratio. The reference standard of the segmentations were annotated by MST, a senior radiologist with 10 years of experience.

Training details

A 2D UNet architecture [33] with a ResNet34 backbone [34] was used. The model was initialized with weights pretrained on the ImageNet database. During the training, a batch size of 32 and an initial learning rate of 10^{-3} were employed. Dice loss was used because it is well-suited when class distribution is imbalanced, as it aids in achieving a balance between precision and recall. CT images were windowed with a center of 60 HU and width of 400 HU, and then rescaled to an intensity range of [0,1] using min–max scaling.

The only augmentation technique implemented was a vertical flip, ensuring the model's effectiveness for imaging patients in different orientations, including prone positions. Model training incorporated early stopping mechanisms to prevent overfitting, with monitoring of the validation set for improvements in the Dice score. Training was halted if there was no improvement after 10 epochs.

Multicenter validation step

Our multicenter validation step utilized another retrospective, convenience sample, including new data from Dasa, distinct from that used in the initial segmentation step. This dataset was supplemented with data from four additional institutions: University of Alabama (UAB), MedStar, University of California, San Francisco (UCSF), and University Hospitals Cleveland Medical Center (UHH). Anonymization was done by each institution before sending the data to the internal institution. The scans incorporated into the study ranged from October 2021 and July 2022 and spanned a wide range of clinical scenarios, from outpatient visits and inpatient stays to emergency room presentations.

Inclusion and exclusion criteria

Patients who had undergone an abdominal CT scan, which could also include chest and/or pelvis imagery, were eligible. They should have been performed for any reason in patients aged over 18 years. An inclusion criterion was the presence of a non-contrast phase in the CT scan for algorithm inference. Because of their widespread clinical use and availability, DEXA scans, conducted for the respective patient within a maximum interval of six months before or after the CT scan, were chosen as the reference standard. DEXA scans that did not adequately depict the lumbar spine were excluded. Patients with metallic spinal implants, recent spinal surgery and/or vertebral body fractures were excluded. If the assessment was based on a single vertebral body they were also excluded, in line with the 2023 official positions from the International Society for Clinical Densitometry (ISCD) [35].

Main metric definition and formula

For the whole lumbar spine, the segmented axial slices were processed to compute the mean of segmented pixels per slice. The mean slice values were then averaged across all slices, yielding the metric we call the *mean of slices' mean attenuation (MSMA)*, expressed in HU:

$$MSMA = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M}\sum_{j=1}^{M_i} HU_{ij}\right)$$

where, N: number of slices (L1–L5), M_i: number of segmented pixels in slice i, HU_{ij}: attenuation value of pixel j in slice i

We selected MSMA as our primary metric because preliminary testing showed that it provided slightly better sensitivity and specificity, which are widely used in clinical practice, than the global voxel mean (Supplementary Table 1), while yielding similar values for the other performance metrics. It also mitigates potential biases from vertebral cross-sectional area variations, weighting each segmented slice the same.

Statistical analysis

The strength of the association between the MSMA from our segmentation and the BMD of the lumbar spine was assessed using linear regression and the Pearson correlation coefficient was calculated.

The diagnostic capability of the algorithm was further evaluated by calculating the Area Under the Curve (AUC). Site-specific diagnostic performance was examined by calculating AUCs for each institution separately, and results were summarized in a forest plot. Confidence intervals for AUCs were obtained using bootstrap resampling (2,000 iterations, stratified by outcome status).

Performance metrics were calculated using thresholds determined by the Youden J method [36] to identify global optimal cutoffs, and were also evaluated using fixed thresholds derived from internal data. The independent variable was the MSMA obtained from the segmentation, and the dependent variable was the classification of patients based on their spine BMD values. Confidence intervals for sensitivity, specificity, predictive values, and accuracy were calculated using the Wilson score interval method. We additionally constructed Precision–Recall (PR) curves and calculated the area under the PR curve (AUPRC).

To evaluate whether MSMA was independently associated with the diagnosis of osteoporosis, we performed a logistic regression with cluster-robust standard errors. The dependent variable was osteoporosis. The primary independent variable was MSMA (continuous, HU). Models were adjusted for CT acquisition parameters including kVp (continuous), slice thickness (continuous), and scanner vendor (categorical). Robust covariance estimates were clustered by institution to account for within-site correlation.

To evaluate calibration, we generated calibration plots comparing predicted probabilities against observed outcomes and quantified calibration using the Brier score.

Our diagnostic approach focused exclusively on the lumbar spine due to the absence of segmented CT data for other skeletal sites. The categorization relied on the t-score, with values below or equal to − 2.5 being considered LBMD/osteoporosis. The reference values were consistent across all institutions. In our study, the t-score calculation was based on a standardized reference sample of young caucasian women [37], aligning with the latest official positions from the ISCD [35].

All statistical analyses were conducted using Python 3.11, utilizing the latest versions of the SciPy, NumPy, Scikit-learn, Pandas, and Matplotlib libraries.

Results

Detailed information regarding the selection process of participants, including the inclusion and exclusion criteria for both the segmentation and validation steps, is systematically illustrated in the STARD (Standards for Reporting of Diagnostic Accuracy Studies) diagram (Fig. 1).

Raw segmentation

In the segmentation step, we included 105 participants after applying the exclusion criteria. Two exams were excluded because of motion artifacts.

In assessing the raw segmentation performance of our model, the Dice coefficient achieved a score of 0.998 on the validation set. When applied to the unseen test set, the Dice score remained high at 0.978. It also demonstrated robust performance across various patient orientations, including oblique and prone positions (Fig. 2) and across multiple acquisition settings (Supplemental fig.1). Typical failure examples are also depicted in Supplemental figure 1.

Multicenter validation

Demographics

In the multicenter validation step, the number of participants considered for the final sample was 504. This sample consisted of 388 (80%) female patients, with a median age of 66 years (interquartile range 56–72) (Table 1). For this step, 21 cases were excluded: six due to surgeries, six because of fractures, five due to external artifacts, and four because the DEXA scans depicted only a single vertebral body.

Data encompassing diagnostic categories, CT scan vendors, slice thickness for all participants, and segregated details for in-house (Dasa) and external datasets are also systematically listed in Table 1. The CT scans originated from various vendors and were conducted with a range of slice thicknesses. In our dataset, 88.5% of the scans were acquired using a voltage of 120 kV. As for the DEXA scan diagnosis, 35.8% had at least osteopenia, while 6% had osteoporosis. There were 16 exams acquired in the prone position.

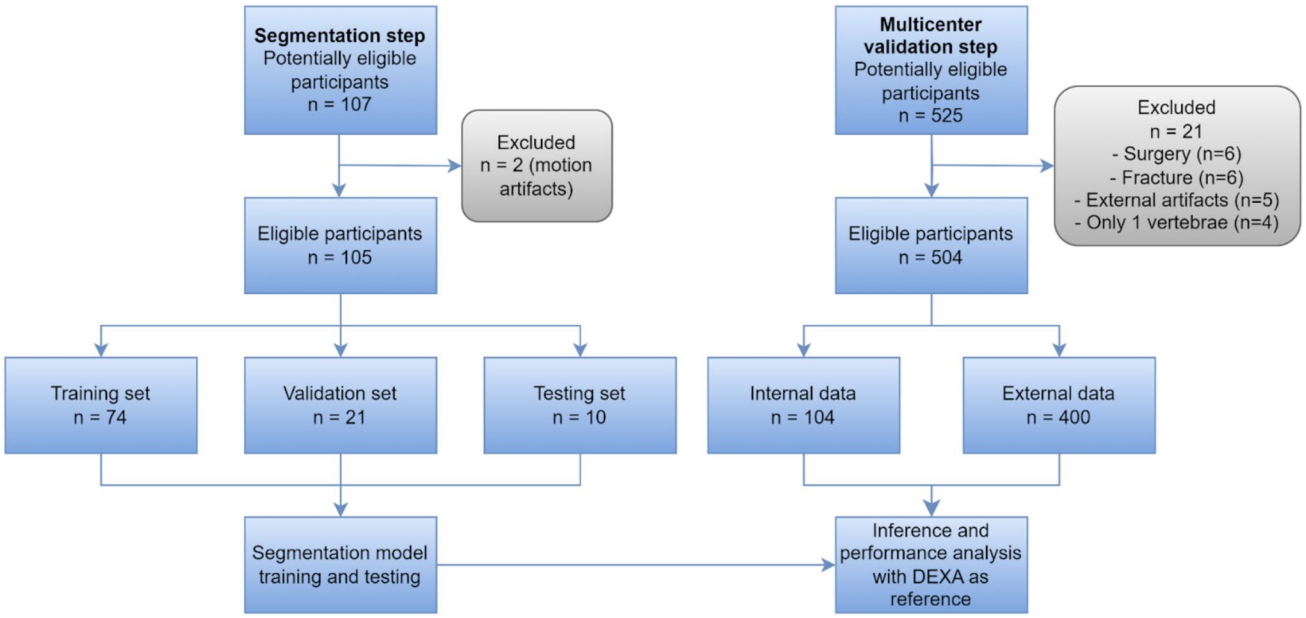


Fig. 1 STARD (Standards for Reporting of Diagnostic Accuracy Studies) diagram illustrating each phase of the study

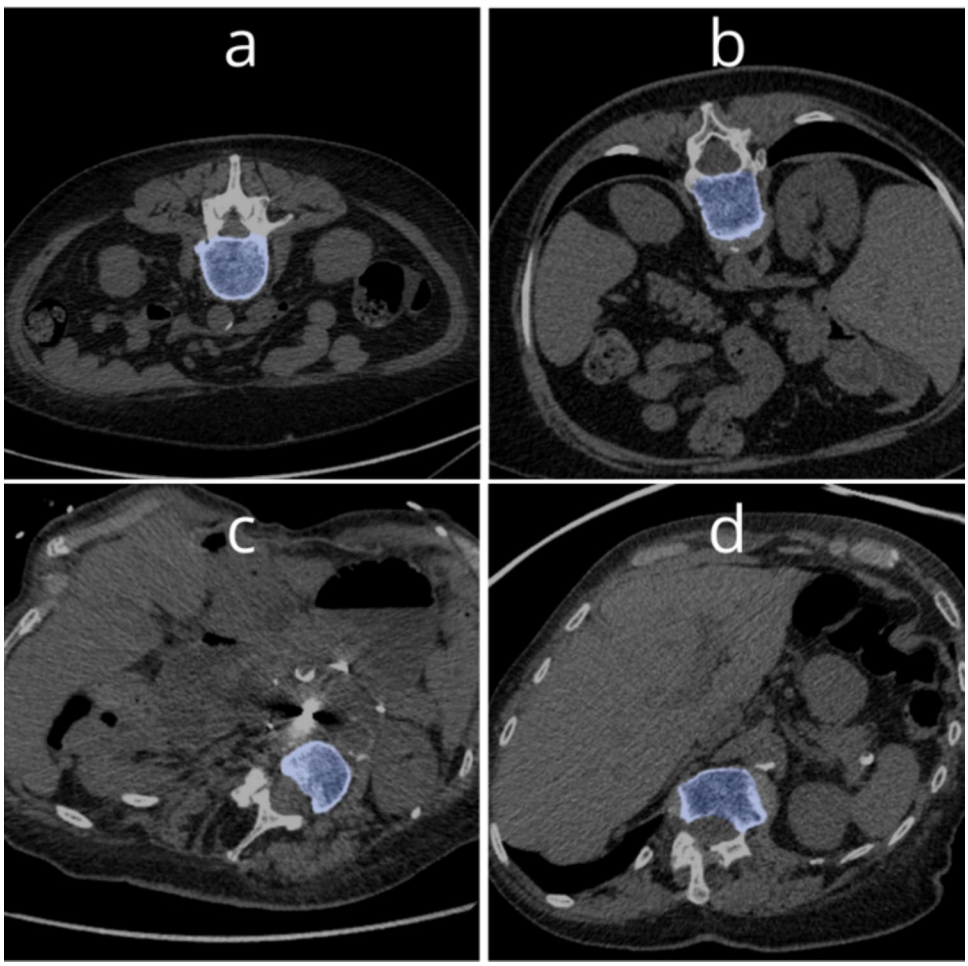


Fig. 2 Examples of segmentations in pronated (**a** and **b**) and oblique (**c** and **d**) patients

Table 1 Caracteristics of the multicenter validation sample

Variables	All data (n = 504)	Internal data (n = 104)	External data (n = 400)
Sex, n (%)			
Feminine	400 (79.4)	91 (87.5)	309 (77.2)
Masculine	104 (20.6)	13 (12.5)	91 (22.8)
Age (years), median (IQR)	66 (56–72)	62 (52.75–72.5)	66 (57–72)
Institution			
Dasa (Internal institution)	104 (20.6)		
MedStar (External institution 1)	92 (18.2)		
UAB (External institution 2)	181 (35.8)		
UCSF (External institution 3	90 (17.8)		
UHH (External institution 4)	37 (7.5)		
CT Vendor, n (%)			
General electric	277 (54.9)	68 (65.4)	209 (52.3)
Siemens healthineers	123 (24.4)	25 (24.0)	98 (24.5)
Philips	94 (18.7)	1 (1.0)	93 (23.2)
Canon	10 (2.0)	10 (9.6)	0 (0.0)
kV			
120	446 (88.5)	77 (74.0)	369 (92.3)
> 120	29 (5.8)	15 (14.4)	14 (3.5)
< 120	29 (5.7)	12 (11.6)	17 (4.2)
Slice thickness (mm), n (%)			
< = 1.5	153 (30.2)	66 (63.5)	87 (21.8)
1.5–2.5	141 (28.0)	37 (35.6)	104 (26.0)
2.5–3.5	142 (28.3)	0 (0.0)	142 (35.5)
> 3.5	68 (13.5)	1 (0.9)	67 (16.7)
Patient orientation			
Supine	488 (96.8)	104 (100.0)	384 (96.0)
Prone	16 (3.2)	0 (0.0)	16 (4.0)
BMD category, n (%)			
t-score > = − 1	324 (64.3)	83 (79.8)	241 (60.3)
−1 < t-score < −2.5	150 (29.8)	18 (17.3)	132 (33.0)
t-score < = −2.5	30 (6.0)	3 (2.9)	27 (6.8)

IQR interquartile range

Correlation analysis

A linear regression analysis demonstrated a correlation coefficient (r = 0.62, 95% CI, 0.56–0.67) between the MSMA (in HU from CT scans) and the BMD (in g/cm^2) from DEXA scans. When selecting only exams acquired with 120 kV, the coefficient did not change significantly. The equation of the best-fit line was:

$$BMD\left(\frac{g}{cm^2}\right) = 0.41 + 2.6 * MSMA\,(HU)$$

AUC calculations

The AUC of the Receiver Operating Characteristic (ROC) curves were calculated for various thresholds to predict osteoporosis, with values exceeding 0.8 for both internal and external datasets (Fig. 3). The values were similar when excluding exams acquired with voltages other than 120 kV. Site-level AUCs for MSMA ranged from 0.84 to 0.96, with overlapping 95% CIs across all institutions indicating no major site-specific outlier (Fig. 4).

Performance metrics

Optimal thresholds were also obtained using the Youden J method. The performance metrics derived from the threshold of 202.6 HU, obtained using the whole data, are detailed in Table 2. These metrics were also calculated using an internal-site derived threshold of 198 HU (Table 3).

Precision–recall analysis further illustrated the performance of the MSMA model for osteoporosis detection (Fig. 5). Despite the modest precision values, reflecting the low prevalence of osteoporosis in the study population, recall (sensitivity) remained adequate across thresholds, consistent with the model's high negative predictive value.

Mixed-effects logistic regression

Mixed-effects logistic regression showed that lower MSMA values were significantly associated with higher odds of osteoporosis, independent of vendor, kVp, and slice thickness ($p < 0.001$). The effect size for MSMA remained robust after adjustment, with an odds ratio (OR) < 1 (Table 4).

Calibration

The MSMA model also demonstrated good calibration between predicted probabilities and observed outcomes. The calibration plot showed small deviation from the reference line across most of the probability range (Supplemental figure 2). The Brier score was 0.054, indicating good agreement between predicted and actual outcomes (values closer to 0 reflect better calibration).

Discussion

General considerations

Our study examines the potential of using abdominal CT scans, a commonly performed imaging test, as an opportunistic screening tool for identifying individuals with LBMD, using a DL technique, enhanced by a multicenter validation.

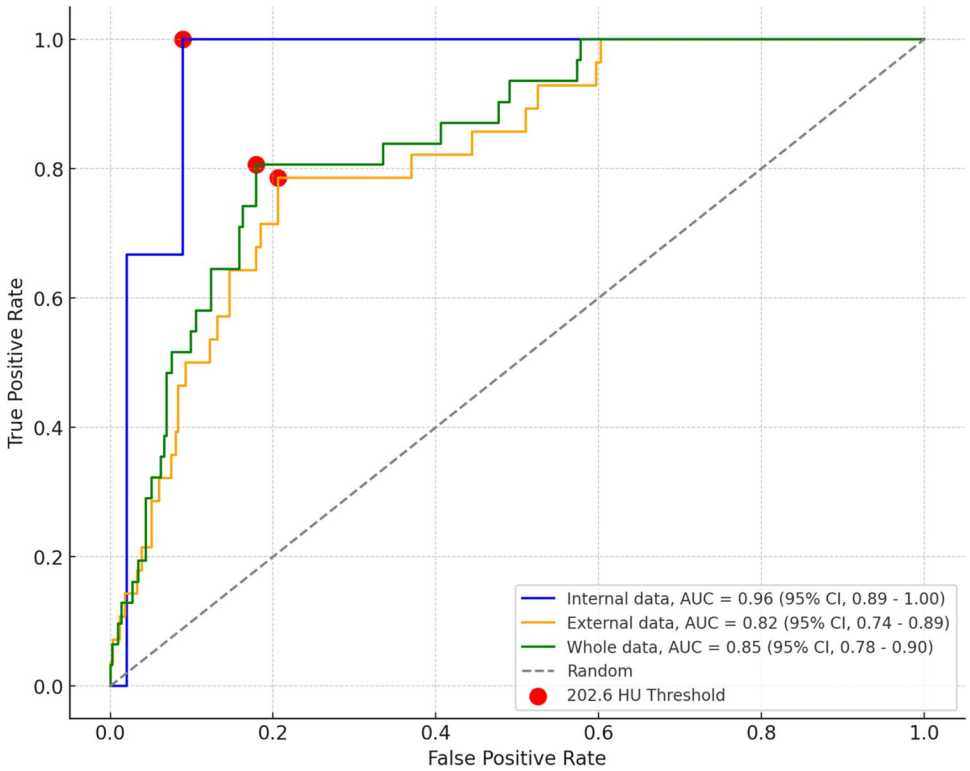


Fig. 3 ROC curve for each dataset for predicting osteoporosis. Red markings correspond to a threshold of 202.6 HU

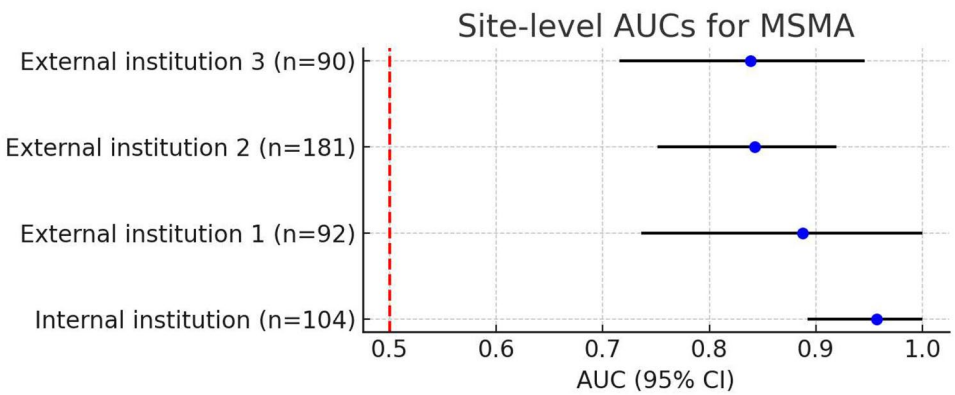


Fig. 4 Site-specific AUCs for osteoporosis prediction. External institution 4 (n = 38) is not shown because of the absence of positive cases

Table 2 Performance metrics [95% CI] for predicting osteoporosis, considering a threshold of 203 HU

	All data (n = 504)	Internal data (n = 104)	External data (n = 400)
Sensitivity	81 (25/31) [64–91]	100 (3/3) [44–100]	79 (22/28) [61–90]
Specificity	83 (393/474) [79–86]	91 (92/101) [84–95]	81 (301/373) [76–84]
Positive predictive value	24 (25/106) [17–33]	25 (3/12) [9–53]	23 (22/94) [16–32]
Negative predictive value	99 (393/399) [97–99]	100 (92/92) [96–100]	98 (301/307) [95–99]
Accuracy	83 (418/505) [79–86]	91 (95/104) [84–95]	81 (323/401) [76–84]
Positive likelihood ratio	4.83 [4.09–5.69]	11.22 [10.56–11.93]	4.17 [3.46–5.03]
Negative likelihood ratio	0.20 [0.14–0.29]	0 [95% CI not estimable]	0.23 [0.16–0.34]

Table 3 Performance metrics [95% CI] for predicting osteoporosis, considering a threshold of 198 HU

	All data (n = 504)	Internal data (n = 104)	External data (n = 400)
Sensitivity	70 (21/30) [52–83]	100 (3/3) [44–100]	67 (18/27) [48–81]
Specificity	85 (403/475) [81–88]	91 (92/101) [84–95]	83 (311/374) [79–87]
Positive Predictive Value	23 (21/93) [15–32]	25 (3/12) [9–53]	22 (18/81) [15–32]
Negative Predictive Value	98 (403/412) [96–99]	100 (92/92) [96–100]	97 (311/320) [95–99]
Accuracy	84 (424/505) [81–87]	91 (95/104) [84–95]	82 (329/401) [78–85]
Positive Likelihood Ratio	4.62 [3.64–5.86]	11.22 [10.56–11.93]	3.96 [3.02–5.19]
Negative Likelihood Ratio	0.35 [0.24–0.51]	0 [95% CI not estimable]	0.40 [0.27–0.59]

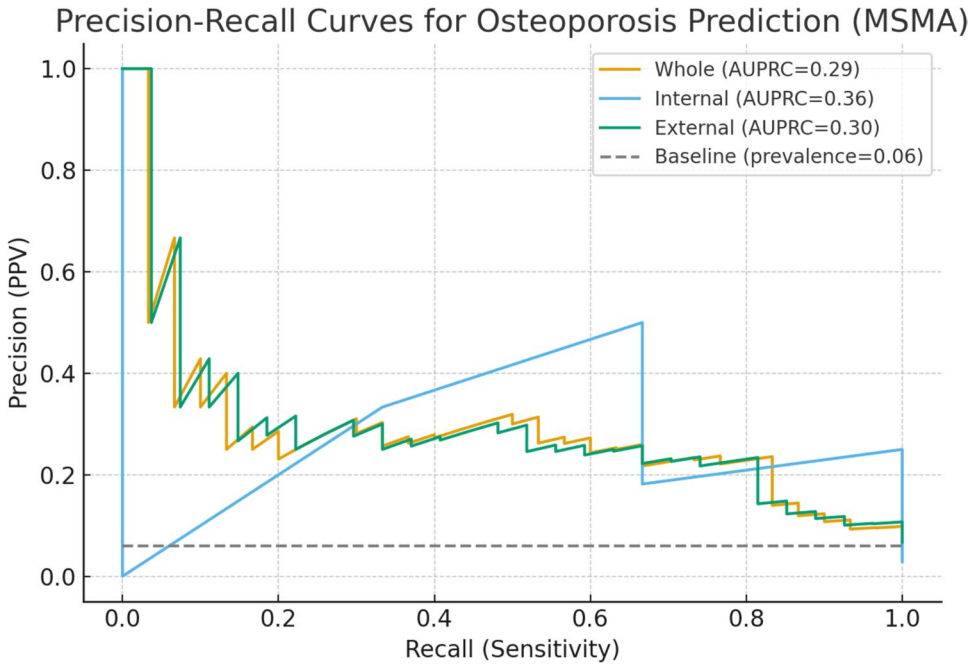


Fig. 5 Precision–recall curves. The dashed horizontal line represents the prevalence of osteoporosis in the study population (≈6%). The improvement of the model curves above this baseline illustrates added value compared with random classification

Table 4 Mixed-effects logistic regression to evaluate whether MSMA was independently associated with the diagnosis of osteoporosis

	Odds-ratio	95% CI	p-value
MSMA (HU)	0.96	0.95–0.97	<0.001
CT Vendor (reference = GE)			
Philips	1.49	0.88–2.55	0.141
Siemens	0.94	0.27–3.29	0.928
Toshiba	0.0	0.0–0.0	<0.001
kVp	0.99	0.97–1.0	0.016
Slice thickness (mm)	0.5	0.38–0.67	<0.001

To our knowledge, ours is one of the few DL algorithms to be published in the literature that has done this step. The robustness of our algorithm across different patient orientations, including pronated positions, further reinforces its potential for clinical application in diverse settings. By testing our algorithm across a variety of patient populations, we have significantly enhanced the generalizability and credibility of our findings. Furthermore, our algorithm has shown to keep good performance in a dataset comprising multiple CT vendors and varying slice thickness. This suggests that it may work also in a variety of equipment settings.

The segmentation performance, as indicated by the Dice coefficient, was consistent in both the validation and hold-out test sets. This high level of precision is crucial, as vertebrae attenuation in routine CT scans has been shown to correlate well with BMD values derived from both Quantitative Computed Tomography and DEXA scans [12, 15]. However, we acknowledge this step comprised only 105 patients, resulting in a limited sample size. We also didn't include 3D segmentation techniques in our approach. As such, future research should evaluate 3D methods and larger training datasets to further improve precision.

This potential is bolstered by the Pearson correlation coefficient (r) value of 0.62 obtained between MSMA and BMD in our study, consistent with values reported in the literature comparing trabecular bone attenuation and DEXA-based BMD, which range from 0.399 to 0.891[31], although through studies that utilized internal validation only. A recent single center study by Pickhard et al. [23], featuring another DL method, using a seven-slice median approach, with an ellipsis placed on the anterior aspect of the trabecular bone of L1, had an AUC of 0.93 with up to 94% sensitivity and 84% specificity for osteoporosis prediction, similar results to ours' internal validation performance. An advantage of using the whole lumbar spine is that fractures or other pathological processes that may affect L1 will interfere less with the results. Also, by including cortical bone in the segmentations, we take into account that fragility fractures are often a result of both cortical thinning and loss of trabecular bone [38–40].

While radiomics approaches for osteoporosis are increasingly recognized for their diagnostic performance [41–43], the strength of our study lies in demonstrating that a simple HU-based method retains significant diagnostic value in a multi-center setting. Despite the lower complexity of HU-based models, they offer key advantages: transparency, interpretability, and immediate feasibility in routine clinical workflows. These features make HU-based screening particularly attractive for broad adoption.

From a workflow perspective, a threshold of ~ 198 HU derived from the internal dataset identified most osteoporosis

cases, but a threshold of ~ 203 HU derived from the entire dataset yielded even better results overall and might be better suited for clinical decision. Future work may require site-specific recalibration.

External performance drop considerations

There was a noted drop in performance during the external testing phase, however, it is important to acknowledge that, while our internal data derives from a private practice outpatient setting, three out of the four external testing institutions, UCSF, UAB and UHH are high-complexity healthcare facilities, which handle a lower proportion of normal or near-normal exams. This is a factor that can impact the performance of our MSMA model, as these patients with multiple comorbidities often present with greater anatomical variability, are more likely to have pathological bone alterations, and their scans might be of reduced image consistency owing to higher fat-to-muscle ratios and increased susceptibility to noise and artifacts.

Differences in vendor profiles between the internal and external datasets may also have contributed to the external performance drop. However, there are some mitigation strategies, including phantom-based calibration techniques or vendor-specific calibration, that may help address inter-scanner variability in future research.

Limitations

The segmentation model was trained on a convenience sample of 105 cases, and demographic data were not available. While this may limit evaluation of its representativeness, segmentation was used solely as a preprocessing step, and the model's utility was supported by the external validation of CT-derived attenuation values against DEXA.

Our validation cohort showed a relatively low prevalence of positive cases, although the osteoporosis values were not markedly different from those reported in postmenopausal women in the NHANES cohort (2005–2018) [44]. This resulted in low PPVs, while NPVs remained high. Clinically, this pattern suggests that the algorithm is best suited as a triage or flagging tool—helping to reliably rule out normal patients and prioritize those who may require confirmatory DEXA. It is important to emphasize that, although overall PPVs were low, 73.6% of individuals with attenuation values below 203 HU had at least osteopenia (T-score < -1) and may still benefit from preventive or therapeutic interventions. In addition, the PRAUC curves demonstrated precision values well above the baseline prevalence of osteoporosis, underscoring the clinical relevance of the model.

We also utilized BMD values obtained from DEXA scans as our primary reference standard. However, it's vital to acknowledge that other reference standards, particularly those based on adverse outcomes like fractures or injurious falls, are also critical. We aim to explore these additional endpoints in our forthcoming research. Another limitation of our study is the employment of a methodology that is not as extensively validated in scientific literature compared to the ellipsis method, and is more challenging to benchmark against manual measurements. Notably, these manual measurements were not incorporated into our analysis. A less critical yet noteworthy limitation of our study is the validation of our algorithm exclusively on non-contrast enhanced scans. This focus may restrict its applicability in certain healthcare settings where scanning protocols exclusively include contrast phases, potentially limiting its broader adoption.

Conclusion

In conclusion, our study advances the field of osteoporosis screening by demonstrating the feasibility and effectiveness of using DL algorithms with abdominal CT scans. The multicenter nature of our validation provides a foundation for future research and potential clinical application. As validation studies continue to evolve, it is anticipated that such opportunistic screening methods will become increasingly integrated into routine clinical practice, aiding in the early detection and management of osteoporosis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00261-025-05213-2.

Author contributions ASS cleaned external datasets, analyzed the results, wrote and revised the main manuscript, created all the figures and is the corresponding author. MST gathered the internal dataset, trained the segmentation model and wrote part of the inference code. Bunkerhill Health promoted the cooperation of the institutions through its consortium. EMJDF, SB, NK, MM, JP and FK helped with the statistical analysis. EMJDF, FV, SR, LKB, RF, JM and FK helped providing data from their respective institutions. SR, LKB, RF, JM and FK made changes to the main manuscript. All authors reviewed the main manuscript.

Data availability If requested, a de-identified dataset containing a table of patient characteristics, algorithm results, and bone mineral density values will be made available. However, the original CT images and segmentation data will not be shared due to privacy and proprietary constraints. Access will be granted to researchers who provide a sound scientific proposal and agree to use the data only for the intended purpose outlined in their proposal, for an unlimited amount of time.

Declarations

Conflict of interest This is an unfunded research supported by the Bunkerhill Health Consortium and the algorithm is of private use of the company. Augusto Sarquis Serpa, Sean Bennett, Nishith Khandwala, Melissa Major, Jack Paparian, and Felipe Campos Kitamura are (or were) employees or partners of the Bunkerhill consortium.

References

1. Föger-Samwald U, Dovjak P, Azizi-Semrad U, Kerschan-Schindl K, Pietschmann (2020) Osteoporosis: Pathophysiology and therapeutic options. EXCLI J 19:1017–1037. https://doi.org/10.17179/excli2020-2591
2. Lane NE (2006) Epidemiology, Etiology, and Diagnosis of Osteoporosis. Am J Obstet Gynecol 194(2 Suppl):S3–11. https://doi.org/10.1016/j.ajog.2005.08.047
3. Black DM, Rosen CJ (2016) Clinical Practice. Postmenopausal Osteoporosis. New Engl J Med 374(3):254–262. https://doi.org/10.1056/NEJMcp1513724
4. Shen Y, Huang X, Wu J et al. (2022) The Global Burden of Osteoporosis, Low Bone Mass, and Its Related Fracture in 204 Countries and Territories, 1990-2019. Front Endocrinol 13:882241. https://doi.org/10.3389/fendo.2022.882241
5. Munch S, Shapiro S (2006) The silent thief: osteoporosis and women's health care across the life span. Health Soc Work 31(1):44-53. https://doi.org/10.1093/hsw/31.1.44
6. Sheu A, Diamond T (2016) Bone mineral density: testing for osteoporosis. Aust Prescr 39(2):35-39. https://doi.org/10.18773/austprescr.2016.020
7. Hayes BL, Curtis JR, Laster A et al. (2010) Osteoporosis care in the United States after declines in reimbursements for DXA. J Clin Densitom 13(4):352-360. https://doi.org/10.1016/j.jocd.2010.08.001
8. Oren O, Kebebew E, Ioannidis JPA (2019) Curbing Unnecessary and Wasted Diagnostic Imaging. JAMA 321(3):245–246. https://doi.org/10.1001/jama.2018.20295
9. Pickhardt PJ, Graffy PM, Perez AA, Lubner MG, Elton DC, Summers RM (2021) Opportunistic screening at abdominal CT: use of automated body composition biomarkers for added cardiometabolic value. RadioGraphics 41(2):524–542. https://doi.org/10.1148/rg.2021200056
10. Pickhardt PJ, Summers RM, Garrett JW (2021) Automated CT-based body composition analysis: a golden opportunity. Korean J Radiol 22(12):1934–1937. https://doi.org/10.3348/kjr.2021.0775
11. Lee SJ, Graffy PM, Zea RD, Ziemlewicz TJ, Pickhardt PJ (2018) Future osteoporotic fracture risk related to lumbar vertebral trabecular attenuation measured at routine body CT. J Bone Miner Res 33(5):860–867. https://doi.org/10.1002/jbmr.3383
12. Pickhardt PJ, Pooler BD, Lauder T, del Rio AM, Bruce RJ, Binkley N (2013). Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. Ann Intern Med. 158(8):588-595. https://doi.org/10.7326/0003-4819-158-8-201304160-00003
13. Lee SJ, Pickhardt PJ (2017) Opportunistic screening for osteoporosis using body CT scans obtained for other indications: the UW experience. Clin Rev Bone Miner Metab 15(3):128–137. https://doi.org/10.1007/s00198-015-3318-4
14. Graffy PM, Lee SJ, Ziemlewicz TJ, Pickhardt PJ (2017) Prevalence of vertebral compression fractures on routine CT scans according to L1 trabecular attenuation: determining relevant thresholds for opportunistic osteoporosis screening. AJR Am J Roentgenol 209(3):491–496. https://doi.org/10.2214/AJR.17.17853
15. Gausden EB, Nwachukwu BU, Schreiber JJ, Lorich DG, Lane JM (2017) Opportunistic use of CT imaging for osteoporosis screening and bone density assessment: a qualitative systematic review. J Bone Joint Surg Am 99(18):1580–1590. https://doi.org/10.2106/JBJS.16.00749
16. Fang Y, Li W, Chen X et al. (2021) Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. Eur Radiol 31(4):1831–1842. https://doi.org/10.1007/s00330-020-07312-8
17. Cheng X, Zhao K, Zha X et al. (2021) Opportunistic screening using low-dose CT and the prevalence of osteoporosis in China: a nationwide, multicenter study. J Bone Miner Res 36(3):427–435. https://doi.org/10.1002/jbmr.4187
18. Boutin RD, Lenchik L (2020) Value-added opportunistic CT: insights into osteoporosis and sarcopenia. AJR Am J Roentgenol 215(3):582–594. https://doi.org/10.2214/AJR.20.22874
19. Pickhardt PJ, Lee SJ, Liu J et al. (2019) Population-based opportunistic osteoporosis screening: validation of a fully automated CT tool for assessing longitudinal BMD changes. Br J Radiol 92(1094):20180726. https://doi.org/10.1259/bjr.20180726
20. Dagan N, Elnekave E, Barda N et al. (2020) Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. Nat Med 26(1):77–82. https://doi.org/10.1038/s41591-019-0720-z
21. Roux C, Rozes A, Reizine D et al. (2022) Fully automated opportunistic screening of vertebral fractures and osteoporosis on more than 150,000 routine computed tomography scans. Rheumatology (Oxford) 61(8):3269–3278. https://doi.org/10.1093/rheumatology/keab878
22. Summers RM, Baecher N, Yao J et al. (2011) Feasibility of simultaneous computed tomographic colonography and fully automated bone mineral densitometry in a single examination. J Comput Assist Tomogr 35(2):212–216. https://doi.org/10.1097/RCT.0b013e3182032537
23. Pickhardt PJ, Nguyen T, Perez AA et al. (2022) Improved CT-based Osteoporosis Assessment with a Fully Automated Deep Learning Tool. Radiol Artif Intell 4(5):e220042. https://doi.org/10.1148/ryai.220042
24. Niu X, Huang Y, Li X et al. (2023) Development and validation of a fully automated system using deep learning for opportunistic osteoporosis screening using low-dose computed tomography scans. Quant Imaging Med Surg 13(8):5294-5305. https://doi.org/10.21037/qims-22-1438
25. Schmidt D, Ulén J, Enqvist O et al. (2022) Deep learning takes the pain out of back-breaking work - Automatic vertebral segmentation and attenuation measurement for osteoporosis. Clin Imaging 81:54-59. https://doi.org/10.1016/j.clinimag.2021.08.009
26. Oh S, Kang WY, Park H et al. (2024) Evaluation of deep learning-based quantitative computed tomography for opportunistic osteoporosis screening. Sci Rep 14(1):363. https://doi.org/10.1038/s41598-023-45824-7
27. Park H, Kang WY, Woo OH, Lee J, Yang Z, Oh S. (2024) Automated deep learning-based bone mineral density assessment for opportunistic osteoporosis screening using various CT protocols with multi-vendor scanners. Sci Rep14(1):25014. https://doi.org/10.1038/s41598-024-73709-w.
28. Küçükçiloğlu Y, Şekeroğlu B, Adalı T, Şentürk N. Prediction of osteoporosis using MRI and CT scans with unimodal and multimodal deep-learning models. (2024) Diagn Interv Radiol 30(1):9-20. https://doi.org/10.4274/dir.2023.232116.
29. Löffler MT, Jacob A, Scharr A et al. (2021) Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA.

Eur Radiol 31:6069–6077. https://doi.org/10.1007/s00330-020-07655-2
30. Goller SS, Rischewski JF, Liebig T et al. (2023) Automated Opportunistic Trabecular Volumetric Bone Mineral Density Extraction Outperforms Manual Measurements for the Prediction of Vertebral Fractures in Routine CT. Diagnostics 13(12):2119. https://doi.org/10.3390/diagnostics13122119
31. Bodden J, Prucker P, Sekuboyina A et al. Reproducibility of CT-based opportunistic vertebral volumetric bone mineral density measurements from an automated segmentation framework. (2024) Eur Radiol Exp 8:86. https://doi.org/10.1186/s41747-024-00483-9
32. Bossuyt PM, Reitsma JB, Bruns DE et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. (2015) BMJ 351:h5527. https://doi.org/10.1136/bmj.h5527. PMID: 26511519; PMCID: PMC4623764
33. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv. https://doi.org/10.48550/arXiv.1505.04597
34. He K, Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition. arXiv. https://doi.org/10.48550/arXiv.1512.03385
35. 2023 International Society for Clinical Densitometry Addult Official Positions, International Society for Clinical Densitometry website (2023) Available via https://iscd.org/official-positions-2023/. Accessed 21 Jan 2024
36. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3:32–35.
37. Xue S, Zhang Y, Qiao W et al. (2021) An Updated Reference for Calculating Bone Mineral Density T-Scores. J Clin Endocrinol Metab 106(7):e2613-e2621. https://doi.org/10.1210/clinem/dgab180
38. Seeman E (2002) Pathogenesis of bone fragility in women and men. Lancet 359(9320):1841–1850. https://doi.org/10.1016/S0140-6736(02)08706-8
39. Cooper DM, Kawalilak CE, Harrison K, Johnston BD, Johnston JD (2016) Cortical Bone Porosity: What Is It, Why Is It Important, and How Can We Detect It? Curr Osteoporos Rep 14(5):187-198. https://doi.org/10.1007/s11914-016-0319-y
40. Ramchand SK, Seeman E (2018) The Influence of Cortical Porosity on the Strength of Bone During Growth and Advancing Age. Curr Osteoporos Rep 16(5):561-572. https://doi.org/10.1007/s11914-018-0478-0
41. Tong X, Wang S, Zhang J, Fan Y, Liu Y, Wei W (2024). Automatic Osteoporosis Screening System Using Radiomics and Deep Learning from Low-Dose Chest CT Images. Bioengineering (Basel) 11(1):50. https://doi.org/10.3390/bioengineering11010050
42. Liu J, Wang H, Shan X et al (2024). Hybrid transformer convolutional neural network-based radiomics models for osteoporosis screening in routine CT. BMC Med Imaging 24:62. https://doi.org/10.1186/s12880-024-01240-5
43. Chen YC, Li YT, Kuo PC et al. (2023). Automatic segmentation and radiomic texture analysis for osteoporosis screening using chest low-dose computed tomography. Eur Radiol 33:5097–5106. https://doi.org/10.1007/s00330-023-09421-6
44. Zhang X, Wang Z, Zhang D et al (2023). The prevalence and treatment rate trends of osteoporosis in postmenopausal women. PLoS One 18(9):e0290289. https://doi.org/10.1371/journal.pone.0290289