Archives of Osteoporosis (2025) 20:131
https://doi.org/10.1007/s11657-025-01609-1

LETTER TO THE EDITOR

Opportunistic screening of low bone mass using knowledge distillation-based deep learning in chest X-rays with external validations

Junhyeok Park[1] · Nha-Young Kim[1,2] · Hyun-Jin Bae[1] · Jinhoon Jeong[3] · Miso Jang[3,4] · Sung Jin Bae[5] · Jung-Min Koh[6] · Seung Hun Lee[6] · Joo Hee Yoon[7] · Chang Hoon Lee[7] · Namkug Kim[3,8]

Received: 28 March 2025 / Accepted: 5 September 2025

Abstract

Summary Low bone mass (LBM), which can lead to osteoporosis, is often undetected and increases the risk of bone fractures. This study presents OsPenScreen, a deep learning model that can identify low bone mass early using standard chest X-rays (CXRs). By detecting low bone mass sooner, this tool helps prevent the disease progression to osteoporosis, potentially reducing health complications and treatment costs. OsPenScreen was validated across four external datasets and consistently performed well, showing its potential as a reliable, cost-effective solution for opportunistic early screening in CXR.

Purpose Low bone mass, an often-undiagnosed precursor to osteoporosis, significantly increases fracture risk and poses a substantial public health challenge. This study aimed to develop and validate a deep learning model, OsPenScreen, for the opportunistic detection of low bone mass using routine chest X-rays (CXRs).

Methods OsPenScreen, a convolutional neural network-based model, was trained on 77,812 paired CXR and dual-energy X-ray absorptiometry (DXA) datasets using knowledge distillation techniques. Validation was performed across four independent datasets (5,935 images) from diverse institutions. The model's performance was assessed using area under the curve (AUC), accuracy, sensitivity, and specificity. Grad-CAM visualizations were employed to analyze model decision-making. Osteoporosis cases were pre-excluded by a separate model; OsPenScreen was applied only to non-osteoporotic cases.

Results Our model achieved an AUC of 0.95 (95% CI: 0.94–0.97) on the external test datasets, with consistent performance across sex and age subgroups. The model demonstrated superior accuracy in detecting cases with significantly reduced bone mass and showed focused attention on weight-bearing bones in normal cases versus non-weight-bearing bones in low bone mass cases.

Conclusion OsPenScreen represents a scalable and effective tool for opportunistic low bone mass screening, utilizing routine CXRs without additional healthcare burdens. Its robust performance across diverse datasets highlights its potential to enhance early detection, preventing progression to osteoporosis and reducing associated healthcare costs.

Keywords Artificial intelligence · Chest X-ray · Knowledge distillation · Low bone mass · Deep learning

Junhyeok Park and Nha-Young Kim contributed equally to this work.

✉ Namkug Kim
namkugkim@gmail.com

[1] Promedius Inc., Seoul, Republic of Korea

[2] Department of Biomedical Regulatory Affairs, School of Pharmacy, University of Washington, Seattle, USA

[3] Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, 88, Olympic-Ro 43-Gil, Songpa-Gu, Seoul 05505, Republic of Korea

[4] Department of Health Screening and Promotion Center, Seoul Chuk Hospital, Seoul, Republic of Korea

[5] Department of Health Screening and Promotion Center, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

[6] Division of Endocrinology and Metabolism, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

[7] Department of Internal Medicine, Division of Cardiology, Veterans Health Service Medical Center, Seoul, South Korea

[8] Department of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

Published online: 08 October 2025

Abbreviations

AUC	Area under the receiver operating curve
AMC	Asan Medical Center
BMD	Bone mineral density
LBM	Low bone mass
CXRs	Chest X-rays
DL	Deep learning
DXA	Dual-energy X-ray absorptiometry
Grad-CAM	Gradient-weighted class activation mapping
HSPC	Health Screening and Promotion Center
IRB	Institutional Review Board
NIH	National Institutes of Health
ROC	Receiver operating characteristic curve
SCH	Seoul Chuck Hospital
VHSMC	Veterans' Health System Medical Center
WHO	World Health Organization

Introduction

Osteoporosis affects over 200 million people globally and is the leading cause of fragility fractures, which are associated with chronic pain, disability, and a threefold increase in mortality [1]. Although osteoporosis presents a higher fracture risk than low bone mass, most fractures occur in subjects with low bone mass due to its higher prevalence [2, 3]. Advanced low bone mass has a high likelihood of progressing to osteoporosis within a year [4], underscoring the importance of early detection crucial for fracture prevention and improved patient outcomes.

Dual-energy X-ray absorptiometry (DXA) remains the gold standard for measuring bone mineral density (BMD) [5], but its limited accessibility, high cost, and requirement for trained operators result in low screening rates, particularly in high-risk populations [6]. Chest X-rays (CXRs) present a more accessible alternative for opportunistic screening with recent advances in deep learning (DL) and convolutional neural networks enhancing their diagnostic potential in medical imaging [7]. Previously, we developed OsPorScreen, a DL model that effectively identified individuals with a high probability of osteoporosis using CXRs for opportunistic screening [8]. However, the model had a clinical limitation in that it could not classify low bone mass, thus missing an early stage crucial for effective disease management.

To overcome this limitation, we developed OsPenScreen, a novel DL model tailored to detect low bone mass using CXRs. We validated the model across multicenter cohorts, examining its performance across varying degrees of low bone mass (mild, moderate, and advanced). In this study, "low bone mass" is used to describe subclinical reductions in BMD (T-score between − 1.0 and − 2.5) that precede osteoporosis.

Material and methods

This retrospective study was conducted in accordance with the principles of the Declaration of Helsinki and current scientific guidelines. The research protocol for the Asan Medical Center (AMC) data was approved by the Institutional Review Board (IRB) of the University of Ulsan College of Medicine, Asan Medical Center (IRB No. 2019–1226). External test datasets from Seoul Chuck Hospital (SCH), Public IRB (IRB No. 2024–0256-001), Veterans' Health System Medical Center (VHSMC) (IRB No. 2022–10-003–001) were approved by their respective review boards. Informed consent was waived for all datasets due to the retrospective nature of the study and the use of de-identified patient data.

Acquisition of datasets

Asan medical center

The dataset of AMC consists of health examination data from men and women aged ≥ 50 years or older who visited the Health Screening and Promotion Center (HSPC) of AMC in Seoul, South Korea, between January 2012 and February 2019. This dataset includes CXRs with normal findings paired with same-day DXA scans. Areal BMD measurements (g/cm^2) were obtained using DXA (Lunar Prodigy, software version 9.30.044; GE Healthcare, Madison).

The external test dataset was compiled from patients who received care at AMC, across various departments—including outpatient clinics, inpatient wards, emergency departments, and the HSPC —between June 2006 and July 2019. This dataset is part of the Asan osteoporosis cohort, which includes consecutive ambulatory men and postmenopausal women who visited the AMC osteoporosis clinic from 2010 to 2017. Inclusion criteria were individuals aged ≥ 50 years who underwent both CXR and DXA within a three-month interval, with posteroanterior CXRs showing normal findings; for images from the HSPC, only those taken before 2012 were included. Exclusion criteria comprised patients with implanted medical devices (e.g., electrocardiographic lines, pacemakers, implantable defibrillators), those who had undergone surgeries like internal fixation or bone cement augmentation, individuals with abnormal CXR findings (such as pneumonia, postoperative lung lesions, or any abnormalities reported by radiologists), and images of low quality.

Seoul chuck hospital (SCH)

Between August 2022 and July 2023, we collected a SCH dataset of 8,392 paired CXRs and DXA results from individuals aged 40 to 87 years who underwent both

examinations on the same day. The CXRs were acquired using the DK radiography machine (DR system, DK Medical Solutions, Korea), and the DXA scans were performed with the Osteosys Dexxum T (OsteoSys, Korea). Notably, no exclusion criteria were applied to this dataset, ensuring a broad representation of the patient population.

Veterans' health system medical center (VHSMC)

From October 2012 to October 2022, we collected a VHSMC dataset of 2,433 paired CXRs and DXA results from individuals aged 50 to 98 years. These participants underwent both CXR and DXA examinations within a 180-day interval. The CXRs were acquired using various radiography machines, and the DXA scans were performed using the GE Lunar Prodigy Advance system. In cases where multiple CXRs were available for a single individual, we selected the posteroanterior CXRs taken on the date closest to the DXA examination.

ChestX-ray8 of national institutes of health (NIH)

The publicly available ChestX-ray8 dataset, provided by the National Institutes of Health, comprises 108,948 CXR images from 30,805 patients and was used in this study [9]. This dataset includes multiple labels for various thoracic diseases, as well as normal cases. For this study, we specifically filtered cases labeled as normal to ensure the dataset was representative of our target population,we further refined the selection by including only individuals aged between 50 and 100 years. This carefully selected subset was subsequently incorporated into the learning pipeline using semi-supervised learning, enhancing the model's ability to generalize across diverse CXR cases.

CheXpert of stanford health care

The publicly available CheXpert dataset, developed by Stanford Health Care, contains 224,316 chest X-ray (CXR) images from 65,240 patients, was utilized in this study [10]. This dataset provides detailed annotations for a wide range of thoracic abnormalities, including labels for normal cases. For this study, we focused on cases labeled as normal and further refined the selection to align with our target population. Specifically, we selected images from individuals aged between 50 and 100 years to create a tailored subset of the CheXpert dataset. This refined dataset was incorporated into the model training pipeline to improve its generalizability and performance in distinguishing between Normal and Low bone mass cases on CXR.

Gradient health

GH (Gradient Health) External dataset, a provider of global medical imaging data, maintains a vast collection of 5 million de-identified medical images, including chest radiographs paired with T-scores from DXA scans. Using their subscription-based service, we retrieved chest radiographs taken within six months before or after DXA examinations, with the data primarily originating from North and Latin America. This dataset comprises 987 paired CXR and DXA records obtained through a paid service, though details regarding the specific DXA machines used were unavailable.

Construction of training and validation datasets

Training dataset from AMC with DXA-paired CXRs

The OsPenScreen model was initially trained using CXRs paired with DXA scans obtained from the AMC dataset. The dataset was split into 70% for training, 10% for tuning, and 20% for internal testing. Poor-quality and misclassified CXR images ($n = 5{,}963$) as well as images paired with osteoporosis cases (T-score ≤ -2.5, $n = 2{,}732$) were excluded from training. This curated dataset, consisting exclusively of normal and low bone mass cases with paired DXA labels, served as the basis for supervised training of the initial model.

Training dataset from NIH/CheXpert datasets with proxy labels

To expand the training dataset and address the limitations of single-institution data, we incorporated two publicly available chest radiograph datasets: NIH ChestX-ray8 and Stanford CheXpert. These datasets contain CXR images with a wide range of medical conditions but lack DXA-BMD measurements. Only images labeled as normal were considered. A total of 96,763 CXR images were initially obtained. Among these, 43,396 images were excluded due to patient age < 50 or ≥ 100 years.

Since the public datasets did not include DXA-based labels, we compensated for this limitation by training a model on the AMC dataset, which contained paired DXA and CXR data. This model served as a pre-trained model and was used to infer proxy labels for the NIH and CheXpert datasets. During inference, the model generated a binary classification (normal vs. Low bone mass) along with a corresponding confidence score for each image. Based on the AMC training class distribution, we set confidence thresholds of ≥ 0.99 for normal and ≥ 0.75 for low

bone mass, and applied these thresholds to the NIH and CheXpert datasets. An additional 37,102 cases that did not meet these thresholds were excluded. The remaining 16,265 high-confidence, proxy-labeled cases were combined with the AMC dataset to construct an expanded training set. This semi-supervised learning strategy improved data diversity and contributed to enhanced model generalizability across external cohorts. Details of the threshold experiments for proxy-labeling of the low bone mass class are provided in Supplementary Material (Table S3). Specifically, we conducted experiments across multiple thresholds and selected the operating point that yielded the best overall performance. The corresponding optimal thresholds were then applied for proxy-labeling, and the resulting data were incorporated into the training set.

Validation datasets

Model performance was evaluated using one internal test set from AMC and three external datasets from independent institutions, as illustrated in Fig. 1. The AMC internal test set included 1,326 CXRs, balanced by sex and class (normal vs. low bone mass), with osteoporosis cases ($n = 663$) excluded. Initially, 749 DXA-labeled cases were selected from the AMC osteoporosis cohort, balanced across the three BMD categories. Subsequently, 318 osteoporosis cases were removed, and 340 additional normal cases were randomly selected from the AMC HSPC dataset to complete the internal validation set.

For external validation, the SCH dataset comprised 3,133 CXRs after excluding patients aged ≤ 50 years ($n = 5{,}054$) and osteoporosis cases ($n = 205$). The VHSMC dataset included 705 images after excluding lateral views and 233 osteoporosis cases; although T-score information was unavailable, institutional DXA-based diagnostic labels for normal, low bone mass, and osteoporosis were provided, making it suitable for overall validation. The GH dataset consisted of 987 paired CXR–DXA records obtained within six months before or after DXA examinations, primarily from North and Latin America, and similarly contained institutional DXA-based diagnostic labels for the three BMD categories, but details on the specific DXA manufactures used were not available.

DXA measurement and LBM classification

BMD was measured at three standard anatomical sites: the lumbar spine, femoral neck, and total hip. According to World Health Organization (WHO) criteria, which are also adopted as the clinical standard in the Korean osteoporosis guidelines [28], participants were classified based on the lowest T-score among these sites. Normal BMD was defined as a T-score of −1.00 or higher. To enable stratified analysis of low bone mass, we further divided into three subcategories: mild (T-score between −1.01 and −1.49), moderate (−1.50 to −1.99), and advanced low bone mass (−2.00 to −2.49) [4].

BMD values were primarily obtained using GE Lunar DXA systems for the AMC internal/external and VHSMC external datasets. Only the SCH external test set used an Osteosys DXA device, which has demonstrated high concordance with GE and Hologic systems in prior validation studies [11].

Model development and activation map

For the OsPenScreen model, grayscale truncation [12] was specifically applied before inputting the DICOM images into the model. This preprocessing technique was designed to suppress extreme intensity values and enhance the contrast of diagnostically relevant regions. Given a grayscale image I (x, y) with spatial dimensions H × W, we first defined a central rectangular region R that spans 50% of the image height and width:

$$R = \left\{ (x, y) | x \in \left[\frac{W}{4}, \frac{3W}{4} \right], y \in \left[\frac{H}{4}, \frac{3H}{4} \right] \right\}$$

We then calculated the minimum and maximum intensity values within this region, denoted as V_{min} and V_{max} respectively:

$$V_{min} = \min_{(x,y) \in R} I(x, y),\ V_{max} = \max_{(x,y) \in R} I(x, y)$$

Using these bounds, we clipped all pixel values in the full image I to the range $[V_{min}, V_{max}]$, and then rescaled the resulting image linearly to the 0–255 range for model input. This process effectively suppresses outlier intensity values while preserving clinically meaningful contrast in the central anatomical region. The preprocessed images were subsequently resized to 512 × 512 pixels. A ConvNeXt-small architecture pretrained on ImageNet was then used for model training through knowledge distillation. The model was designed to estimate a confidence score to classify the input images as either low bone mass or normal.

The model, initially trained on the AMC training dataset, was further refined through semi-supervised learning [13] by performing inference on public datasets such as ChestX-ray8 dataset and CheXpert dataset. For the inferred data that exceeded the confidence score threshold applied to each class (0.99 or higher for the normal class and 0.75 or higher for the low bone mass class), proxy labels were assigned. These proxy-labeled data were then added to the original AMC training data to form a new dataset, which was used for additional training. This approach allowed for the supplementation of previously limited data and enabled the model to learn from a broader range of data distributions,

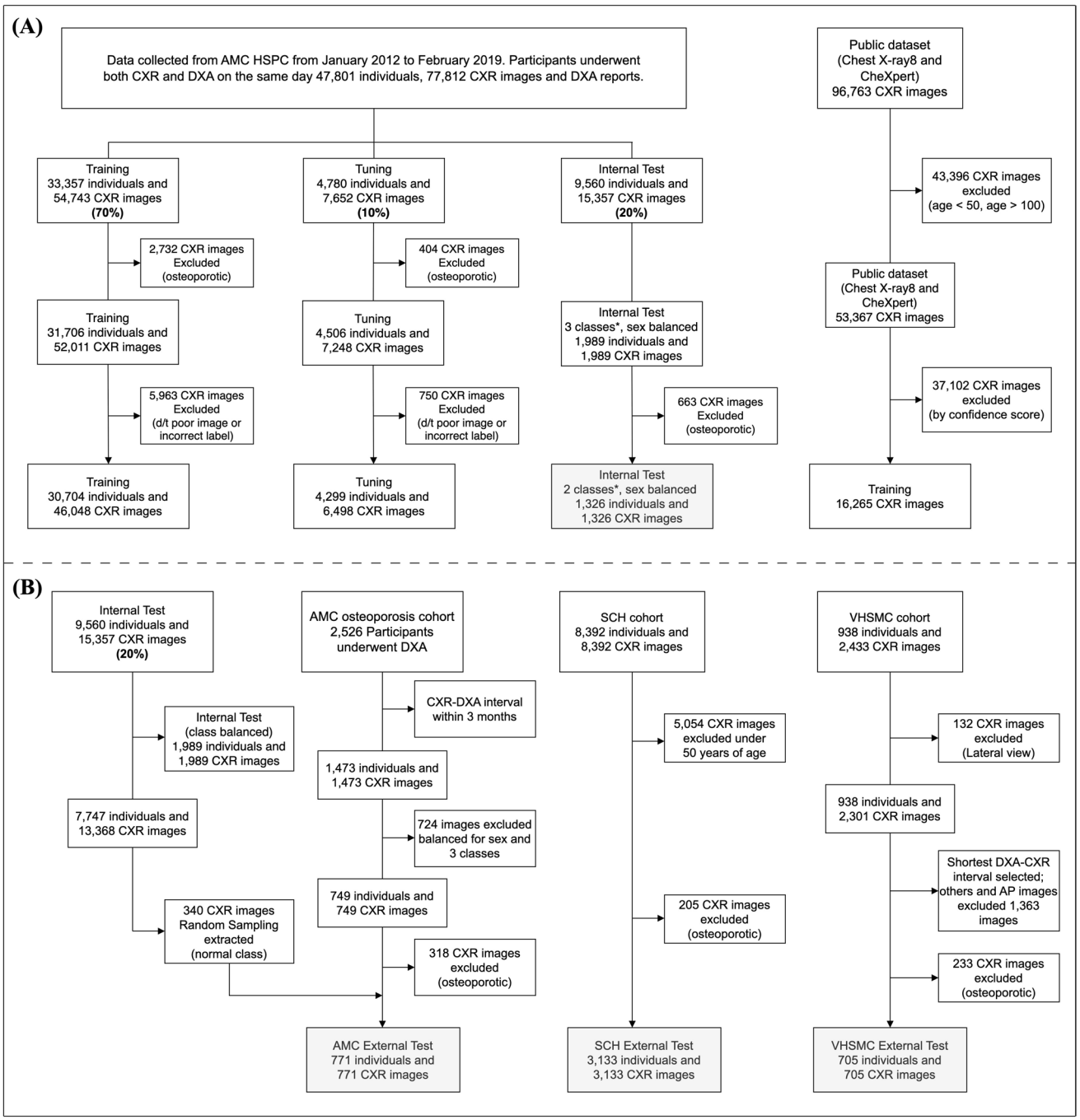

Fig. 1 Construction of training and validation datasets. **A** Training datasets; **B** validation datasets

thereby enhancing its robustness when applied to various external datasets.

The OsPenScreen model was further trained using the knowledge distillation method, employing both cross-entropy loss and Kullback–Leibler (KL)-divergence loss, with a batch size of 64 over 100 epochs. To enhance model robustness, augmentation techniques were categorized into weak augmentation and strong augmentation. Weak augmentation, including horizontal flip,and normalize, preserved essential image characteristics. Strong augmentations, such as Blur, MotionBlur, MedianBlur, ShiftScaleRotate, GridDistortion, ElasticTransform, OpticalDistortion, RandomGamma, and GridDropout, introduced substantial variations to diversify the training data. To understand the model's decision-making and identify critical regions for low bone mass screening in CXRs, this study applied Grad-CAM

[18], highlighting key areas for low bone mass detection with a red overlay. For further details, refer to Supplementary Material (Fig. S1).

Performance evaluation and statistical analysis

To evaluate the screening performance across the four test datasets, accuracy, sensitivity, specificity, and AUC were calculated, along with their corresponding 95% confidence intervals (CIs). The confusion matrix was presented as a 2×2 table, detailing the counts of true positives, false positives, false negatives, and true negatives. The 95% CIs for AUC were estimated using bootstrapping ($n = 1{,}000$) with the roc_auc_score from scikit-learn [14], and the CIs for accuracy, sensitivity, and specificity were computed using the binconf function from the Hmisc R package, assuming a binomial distribution. A 0.5 classification cut-off was applied across all datasets.

Results

Dataset characteristics

A total of 68,811 CXR images from the AMC internal training dataset, ChestX-ray8 dataset, and CheXpert datasets were used for model training, with 5,935 CXR images from one internal and three external datasets included for validation. The training dataset demonstrated a balanced sex distribution, whereas significant differences were observed in the AMC internal test dataset, AMC external test dataset, and VHSMC test dataset ($p < 0.001$), while the SCH test dataset showed no significant difference ($p = 0.075$). The training dataset had a mean age of 53.5 ± 7.0 years for female and 54.0 ± 7.2 years for male which was comparable to the AMC internal test, AMC external test, and SCH test dataset, while the VHSMC test dataset showed a significantly higher mean age ($p < 0.001$).

Regarding the distribution of low bone mass, the training dataset showed a prevalence of 52.8% among females and 25.7% among males. For females, the prevalence was significantly different from the AMC internal test dataset, AMC external test dataset, and SCH test dataset, and significantly higher in the VHSMC test dataset ($p < 0.05$). For males, the prevalence was comparable only to the SCH test dataset (25.3%, $p = 0.737$), but significantly higher in the AMC internal test dataset, AMC external test dataset, and VHSMC test dataset ($p < 0.001$). The Gradient Health (GH) external test dataset comprised 787 CXR–DXA pairs from North and Latin America, with an older patient population and higher LBMD prevalence than other cohorts. The distribution of low bone mass subcategories in the training dataset and four test datasets showed no statistically significant differences, and no single subcategory was found to be dominant for either females or males ($p > 0.05$). Detailed information is described in Table 1.

Performances of the OsPenScreen models

The model demonstrated consistently high performance across all datasets, with AUC values above 0.85. The validation of AMC external test dataset yielded the best results, achieving an AUC of 0.95 (95% CI: 0.94–0.97), a sensitivity of 0.97, and an accuracy of 0.84, underscoring the model's strong ability to detect low bone mass. In the SCH test dataset, the model maintained robust performance with an AUC of 0.90 (95% CI: 0.89–0.91) and a sensitivity of 0.83. By comparison, the VHSMC test dataset showed a slightly lower AUC of 0.85 and a specificity of 0.70, indicating some variability across datasets. The GH external dataset demonstrated the highest PPV (0.91) and specificity (0.84) among all datasets. Nevertheless, accuracy remained consistently high, ranging from 0.78 to 0.84 across all datasets. Detailed results are provided in Table 2.

To further address class imbalance and the varying prevalence of low bone mass across the test sets, we additionally analyzed the model's performance using Precision-Recall (PR) Curves and corresponding Average Precision (AP) scores. The PR analysis confirmed the model's robust performance, with AP scores of 0.91 for AMC internal, 0.95 for AMC external, 0.85 for SCH external, 0.90 for VHSMC, and 0.92 for GH test datasets (Supplementary Fig. S3).

Models trained with pseudo-labeled public data at different confidence thresholds for the low bone mass class (0.65, 0.75, 0.85) were compared to a model trained without public data. Inclusion of pseudo-labeled data improved performance across all thresholds, with the highest AUC and sensitivity observed at the 0.75 threshold (GH AUC: 0.85; AMC internal AUC: 0.86) (Supplementary Table S3).

Subgroup analysis in the OsPenScreen model

The model demonstrated improved accuracy with increasing low bone mass severity across both the AMC external and SCH test datasets. The VHSMC and GH test dataset, however, was excluded from this analysis as it provides only categorical labels (normal, low bone mass, osteoporosis) without T-score values, precluding subgroup classification by severity. Specifically, accuracy for mild low bone mass was lower and more variable, particularly in the SCH test dataset, but significantly improved for moderate and advanced stages, where near-perfect accuracy was observed (Fig. 2). This trend was consistent across three datasets—AMC internal, AMC external, and SCH test dataset -where accuracy improved progressively with low

Table 1 Characteristics of training dataset and internal and external validation datasets

	AMC internal training ($n=46,048$)		AMC internal tuning ($n=6498$)		CheXpert and Chest-Xray8 ($n=16,265$)		AMC internal test ($n=1326$)	
	Female ($n=24,281$)	Male ($n=21,767$)	Female ($n=3425$)	Male ($n=3073$)	Female ($n=6565$)	Male ($n=9700$)	Female ($n=1112$)	Male ($n=214$)
Age, years (mean ± SD)	53.4 ± 6.9	57.1 ± 7.1	53.5 ± 7.0	57.3 ± 7.0	61.6 ± 8.9	62.5 ± 8.6	57.3 ± 6.1	57.9 ± 6.6
BMD categories								
Normal, *n* (%)	12,294 (50.6%)	16,387 (75.3%)	1721 (50.3%)	2319 (75.5%)	2156 (32.8%)	6941 (71.5%)	565 (50.8%)	109 (50.9%)
Low bone mass, *n* (%)	11,987 (49.4%)	5380 (24.7%)	1704 (49.7%)	754 (24.5%)	4409 (67.2%)	2759 (28.5%)	547 (49.2%)	105 (49.1%)
Mild (−1.0 < T-score < 1.5)	4634 (19.1%)	2486 (11.4%)	681 (19.9%)	362 (11.8%)	NA	NA	153 (13.8%)	41 (19.2%)
Moderate (−1.5 ≤ T-score < 2.0)	4169 (17.2%)	1927 (8.9%)	590 (17.2%)	250 (8.1%)	NA	NA	218 (19.6%)	39 (18.2%)
Advanced (−2.0 ≤ T-score < 2.5)	3184 (13.1%)	967 (4.4%)	433 (12.6%)	142 (4.6%)	NA	NA	176 (15.8%)	25 (11.7%)
CXR manufacture	GE healthcare		GE healthcare		Unknown		GE healthcare	
DXA manufacture	GE lunar		GE lunar		Unknown		GE lunar	
	AMC external test ($n=771$)		SCH external test ($n=3133$)		VHSMC external test ($n=705$)		Gradient Health external test ($n=787$)	
	Female ($n=678$)	Male ($n=93$)	Female ($n=1703$)	Male ($n=1430$)	Female ($n=251$)	Male ($n=454$)	Female ($n=752$)	Male ($n=35$)
Age, years (mean ± SD)	57.4 ± 6.1	60.0 ± 7.1	59.2 ± 7.2	59.1 ± 7.4	70.4 ± 7.4	74.3 ± 5.0	66.7 ± 8.9	70.2 ± 8.2
BMD categories								
Normal, *n* (%)	348 (51.3%)	45 (48.4%)	867 (50.9%)	1068 (74.7%)	47 (18.7%)	225 (49.6%)	229 (30.4%)	11 (31.4%)
Low bone mass, *n* (%)	330 (48.7%)	48 (51.6%)	836 (49.1%)	362 (25.3%)	204 (81.3%)	229 (50.4%)	523 (69.6%)	24 (68.6%)
Mild (−1.0 < T-score < 1.5)	30 (4.3%)	5 (5.3%)	317 (18.7%)	172 (12.0%)	NA	NA	NA	NA
Moderate (−1.5 ≤ T-score < 2.0)	121 (17.8%)	21 (22.6%)	285 (16.7%)	122 (8.5%)	NA	NA	NA	NA
Advanced (−2.0 ≤ T-score < 2.5)	179 (26.4%)	22 (23.7%)	234 (13.7%)	68 (4.8%)	NA	NA	NA	NA
CXR manufacture	GE Healthcare (93.57%); Canon Inc. (4.96%); others (1.47%)		DK		Samsung (50.32%); GE healthcare (41.79%); DK (7.68%); others (0.21%)		Carestream (31.77%); Konica Minolta (23.00%); SIEMENS (16.90%); Samsung (14.99%); others (13.34%)	
DXA manufacture	GE lunar		Osteosys		GE lunar		Unknown	

AMC Asan Medical Center, *BMD* bone mineral density, *CXR* chest X-ray, *DXA* dual-energy X-ray absorptiometry, *SD* standard deviation, *SCH* Seoul Chuck Hospital, *SD* standard deviation, *VHSMC* Veterans' Health System Medical Center, *GH* Gradient Health

bone mass severity, showing lower accuracy and higher variability in mild cases, and near-perfect accuracy with minimal variability in moderate and advanced cases.

In the subgroup analysis by age, the OsPenScreen model demonstrated consistently robust performance across all datasets. AUC values remained high across most age groups, particularly in the AMC external and SCH test datasets, where they reached 0.97 and 0.91, respectively (Table 3).

In the sex subgroup analysis, the model demonstrated slight variations in performance across males and females in all test datasets. AUC values were higher for females, with the highest AUC of 0.96 observed in females from the AMC external test dataset, compared to 0.92 for males. Sensitivity was also consistently higher in females, reaching 0.97 in the VHSMC external test dataset, whereas males had a lower sensitivity of 0.76 in the same dataset. In contrast, specificity

Table 2 Performance of the OsPenScreen model in the validation datasets

	Internal dataset	External dataset			
	AMC internal	AMC external	SCH external	VHSMC external	GH external
TP, *n* (%)	592 (44.6%)	372 (48.2%)	989 (27.4%)	371 (52.6%)	409 (52.0%)
TN, *n* (%)	501 (37.8%)	277 (35.9%)	1582 (53.9%)	174 (24.7%)	201 (25.5%)
FP, *n* (%)	162 (12.2%)	111 (14.4%)	351 (11.3%)	98 (13.9%)	39 (5.0%)
FN, *n* (%)	71 (5.4%)	11 (1.4%)	211 (7.3%)	62 (8.8%)	138 (17.5%)
AUC (95% CI)	0.91 (0.89–0.93)	0.95 (0.94–0.97)	0.90 (0.89–0.91)	0.85 (0.82–0.87)	0.85 (0.82–0.88)
Sensitivity (95% CI)	0.92 (0.90–0.94)	0.97 (0.95–0.99)	0.83 (0.81–0.85)	0.86 (0.82–0.89)	0.75 (0.71–0.79)
Specificity (95% CI)	0.72 (0.68–0.76)	0.71 (0.67–0.76)	0.81 (0.80–0.84)	0.70 (0.59–0.70)	0.84 (0.79–0.89)
PPV (95% CI)	0.77 (0.73–0.80)	0.77 (0.73–0.81)	0.74 (0.71–0.76)	0.82 (0.76–0.83)	0.91 (0.89–0.94)
NPV (95% CI)	0.90 (0.87–0.93)	0.96 (0.94–0.98)	0.88 (0.87–0.90)	0.75 (0.68–0.79)	0.59 (0.54–0.65)
Accuracy (95% CI)	0.82 (0.80–0.84)	0.84 (0.82–0.87)	0.82 (0.81–0.83)	0.80 (0.74–0.80)	0.78 (0.75–0.80)
F1 score (95% CI)	0.84 (0.82–0.86)	0.86 (0.83–0.88)	0.78 (0.76–0.80)	0.84 (0.80–0.85)	0.82 (0.80–0.85)

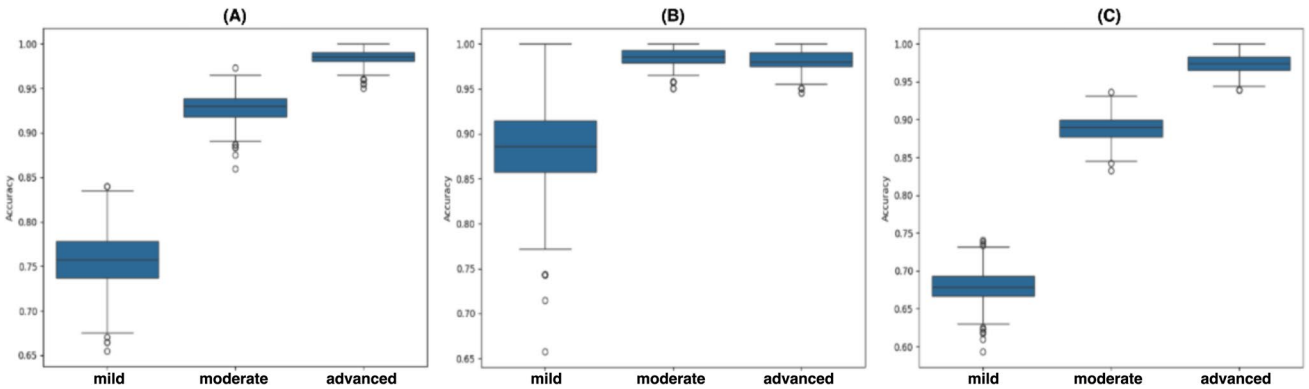

Fig. 2 Accuracies across subgroups. **A** AMC internal test dataset, **B** AMC external test dataset, and **C** SCH external test dataset

was higher in males across most test datasets, with males in the AMC internal test dataset achieving a specificity of 0.93, compared to 0.72 for females. Accuracy was balanced between the sexes, with both showing robust performance, particularly in the AMC external test dataset, where accuracy for females reached 0.86 (Table 3).

Visualization of the model predictions for normal and low bone mass cases across four validations

In the Grad-CAM overlays applied to four independent test datasets, consistent patterns emerged across normal and low bone mass cases. In normal cases, the model consistently highlighted the thoracic and lumbar spine regions, reflecting a focus on weight-bearing bones. Conversely, in low bone mass cases, the model primarily highlighted the bilateral upper arms and clavicle regions, focusing on non-weight-bearing areas. This differentiation in Grad-CAM visualization underscores distinct regions associated with normal versus low bone mass characteristics (Fig. 3). To extend these findings, Supplementary Fig. S2 presents Grad-CAM overlays from five representative cases spanning the T-score continuum (from + 2.0 to –3.0). The visualizations reveal a gradual shift in model attention from weight-bearing regions, such as the thoracic and lumbar spine, toward non-weight-bearing areas, including the clavicle and scapula, as bone mineral density declines.

Discussion

In this study, we developed and validated the OsPenScreen model, a DL-based approach utilizing CXRs for the early detection of low bone mass. The model demonstrated consistent performance across multiple external datasets, achieving an AUC over 0.85. These results highlight the potential of using CXRs for opportunistic low bone mass screening, allowing early detection and management without the need for additional healthcare burden, addressing a significant population-level fracture risk.

The OsPenScreen model was trained on a diverse dataset of 68,811 CXRs from the AMC HSPC cohort in Korea and

Table 3 Performance of the OsPenScreen model by sex and age subgroups

Dataset	Sub groups (sex, age)	Number of images (%)		AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Accuracy (95% CI)	*p*-value
		Normal	Low bone mass							
AMC internal	Female	556 (50.0%)	556 (50.0%)	0.92 (0.90–0.93)	0.92 (0.90–0.94)	0.72 (0.68–0.76)	0.77 (0.74–0.80)	0.90 (0.87–0.90)	0.82 (0.80–0.84)	$p < 0.001$
	Male	107 (50.0%)	107 (50.0%)	0.94 (0.90–0.97)	0.76 (0.67–0.84)	0.93 (0.88–0.98)	0.92 (0.86–0.97)	0.79 (0.72–0.86)	0.85 (0.79–0.90)	$p < 0.001$
	50–54	309 (60.1%)	205 (39.9%)	0.92 (0.90–0.94)	0.82 (0.77–0.87)	0.86 (0.82–0.90)	0.80 (0.74–0.85)	0.88 (0.84–0.91)	0.85 (0.82–0.88)	$p < 0.001$
	55–59	216 (51.1%)	207 (38.9%)	0.90 (0.87–0.92)	0.91 (0.87–0.94)	0.72 (0.66–0.79)	0.78 (0.72–0.83)	0.87 (0.78–0.94)	0.82 (0.78–0.85)	$p < 0.001$
	60–64	128 (57.9%)	93 (42.1%)	0.92 (0.88–0.95)	0.93 (0.88–0.97)	0.64 (0.54–0.74)	0.78 (0.71–0.84)	0.87 (0.78–0.94)	0.81 (0.75–0.86)	$p < 0.001$
	≥65	114 (67.8%)	54 (32.1%)	0.87 (0.80–0.92)	0.95 (0.90–0.98)	0.48 (0.35–0.62)	0.80 (0.73–0.86)	0.82 (0.66–0.94)	0.80 (0.74–0.86)	$p < 0.001$
AMC external	Female	342 (50.4%)	336 (49.6%)	0.96 (0.95–0.97)	0.98 (0.97–0.99)	0.70 (0.65–0.75)	0.76 (0.72–0.80)	0.98 (0.96–0.99)	0.84 (0.81–0.87)	$p < 0.001$
	Male	45 (48.4%)	48 (51.6%)	0.92 (0.85–0.98)	0.90 (0.81–0.98)	0.82 (0.70–0.92)	0.84 (0.74–0.93)	0.88 (0.78–0.97)	0.86 (0.78–0.92)	$p < 0.001$
	50–54	193 (61.1%)	123 (38.9%)	0.96 (0.94–0.98)	0.96 (0.91–0.99)	0.78 (0.72–0.83)	0.69 (0.61–0.76)	0.97 (0.94–0.99)	0.84 (0.80–0.88)	$p < 0.001$
	55–59	104 (51.2%)	99 (48.8%)	0.97 (0.95–0.98)	0.98 (0.95–1.00)	0.71 (0.62–0.79)	0.80 (0.73–0.86)	0.96 (0.92–1.00)	0.85 (0.81–0.90)	$p < 0.001$
	60–64	56 (42.7%)	75 (57.3%)	0.94 (0.90–0.98)	0.99 (0.96–1.00)	0.66 (0.55–0.78)	0.80 (0.73–0.86)	0.96 (0.92–1.00)	0.85 (0.81–0.90)	$p < 0.001$
	≥65	34 (28.1%)	87 (71.9%)	0.90 (0.82–0.96)	0.97 (0.92–1.00)	0.44 (0.29–0.62)	0.82 (0.74–0.89)	0.84 (0.65–1.00)	0.82 (0.74–0.88)	$p < 0.001$
SCH external	Female	867 (50.9%)	836 (49.1%)	0.90 (0.88–0.91)	0.87 (0.85–0.90)	0.74 (0.71–0.77)	0.76 (0.74–0.79)	0.86 (0.83–0.88)	0.81 (0.79–0.82)	$p < 0.001$
	Male	1,068 (74.7%)	362 (25.3%)	0.89 (0.86–0.91)	0.71 (0.67–0.76)	0.88 (0.86–0.90)	0.67 (0.63–0.72)	0.90 (0.88–0.92)	0.84 (0.82–0.86)	$p < 0.001$
	50–54	741 (71.1%)	301 (28.9%)	0.90 (0.87–0.92)	0.73 (0.68–0.77)	0.88 (0.86–0.90)	0.71 (0.67–0.76)	0.89 (0.87–0.91)	0.84 (0.82–0.86)	$p < 0.001$
	55–59	505 (62.1%)	308 (37.9%)	0.89 (0.87–0.92)	0.77 (0.73–0.82)	0.83 (0.79–0.86)	0.73 (0.68–0.78)	0.86 (0.82–0.89)	0.81 (0.77–0.83)	$p < 0.001$
	60–64	310 (56.8%)	236 (43.2%)	0.91 (0.88–0.93)	0.90 (0.86–0.93)	0.77 (0.73–0.82)	0.75 (0.70–0.80)	0.91 (0.87–0.94)	0.83 (0.79–0.86)	$p < 0.001$
	≥65	373 (51.4%)	353 (48.6%)	0.90 (0.88–0.92)	0.91 (0.88–0.94)	0.71 (0.67–0.76)	0.75 (0.70–0.80)	0.89 (0.85–0.93)	0.81 (0.78–0.84)	$p < 0.001$
VHSMC external	Female	47 (18.7%)	204 (81.3%)	0.83 (0.76–0.89)	0.97 (0.94–0.99)	0.36 (0.22–0.50)	0.87 (0.82–0.91)	0.74 (0.53–0.92)	0.86 (0.81–0.90)	$p < 0.001$
	Male	225 (49.5%)	229 (50.5%)	0.81 (0.77–0.85)	0.76 (0.70–0.81)	0.70 (0.64–0.76)	0.72 (0.66–0.77)	0.74 (0.68–0.79)	0.73 (0.69–0.77)	$p < 0.001$
	50–64	18 (31.5%)	39 (68.5%)	0.84 (0.74–0.94)	0.89 (0.76–0.96)	0.61 (0.39–0.80)	0.83 (0.69–0.92)	0.73 (0.48–0.89)	0.80 (0.69–0.89)	$p < 0.001$
	≥65	254 (39.2%)	394 (60.8%)	0.85 (0.82–0.88)	0.85 (0.82–0.89)	0.64 (0.58–0.70)	0.79 (0.75–0.82)	0.74 (0.68–0.80)	0.77 (0.74–0.80)	$p < 0.001$
GH external	50–54	43 (53.7%)	37 (46.3%)	0.82 (0.72–0.91)	0.59 (0.44–0.74)	0.91 (0.78–0.96)	0.85 (0.67–0.94)	0.72 (0.59–0.82)	0.76 (0.66–0.84)	$p < 0.001$
	55–59	32 (34.4%)	61 (65.6%)	0.87 (0.78–0.94)	0.74 (0.62–0.83)	0.94 (0.80–0.94)	0.96 (0.86–0.99)	0.65 (0.50–0.77)	0.81 (0.71–0.87)	$p < 0.001$
	60–64	38 (29.7%)	90 (70.3%)	0.84 (0.71–0.91)	0.73 (0.63–0.81)	0.84 (0.70–0.93)	0.92 (0.83–0.96)	0.57 (0.44–0.69)	0.77 (0.69–0.83)	$p < 0.001$
	≥65	117 (25.3%)	345 (74.7%)	0.84 (0.80–0.88)	0.79 (0.74–0.83)	0.78 (0.69–0.84)	0.91 (0.88–0.94)	0.55 (0.48–0.62)	0.78 (0.74–0.82)	$p < 0.001$

p-values were calculated using the Bonferroni method to adjust for multiple comparisons

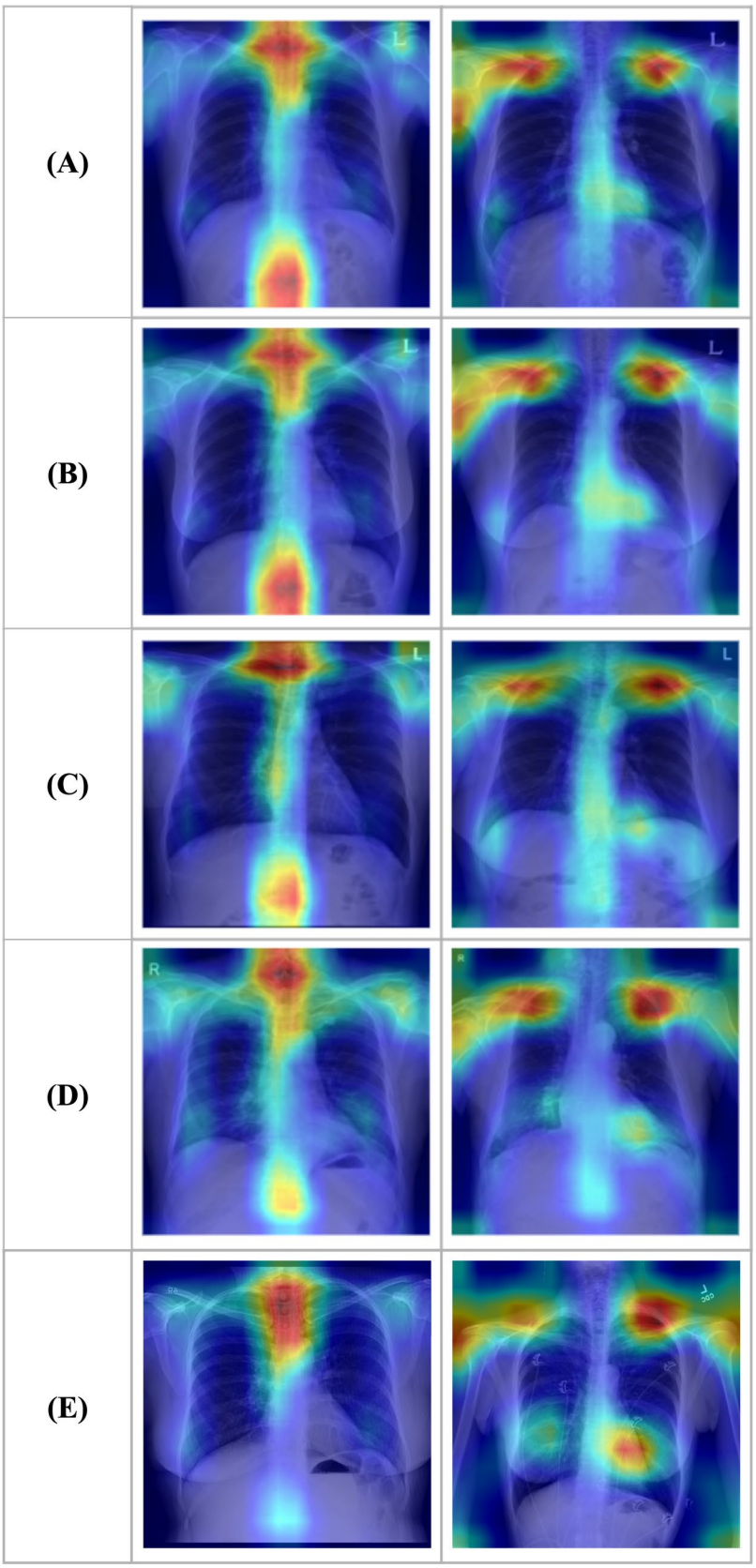

Fig. 3 Grad-CAM images for normal and low bone mass cases across five datasets. **A** AMC internal test dataset, **B** AMC external test dataset, **C** SCH test dataset, **D** VHSMC test dataset, and **E** Gradient Health test dataset. (Left, normal cases; Right, low bone mass cases.) Grad-CAM highlights the regions of the image that the model considers most important for its decision-making process, with warmer colors (e.g., red) indicating areas of higher attention or focus by the model

global datasets from NIH and Stanford Health Care, providing extensive representation of population and medical conditions, and was validated across four distinct datasets with varied cohort characteristics. While the model showed consistently strong performance overall, a decline was observed in the VHSMC test dataset. The most distinguishing feature of the VHSMC test dataset, compared to the other datasets, is that 92% of the individuals are over 65 years of age, which can be classified as elderly [15]. Aging induces widespread anatomical changes, such as costochondral calcification of the ribs, deformation of vertebral bodies, calcification of the tracheobronchial cartilage, and sarcopenia, which are visible on CXRs and present challenges to AI accuracy [16]. Similar patterns were observed in the VHSMC test dataset, where performance particularly declined in cases involving implantable devices or comorbidities associated with aging. Nevertheless, the model achieved an AUC of 0.85 in this dataset, which exceeds the threshold for clinical applicability according to the Standards for Reporting Diagnostic Accuracy Studies guidelines [17], indicating its suitability for clinical use despite the challenges presented by an aging population. In the sex-specific analysis, sensitivity was consistently higher in females across all datasets compared to males. This discrepancy may be attributed to the consistently higher prevalence of low bone mass in females across training datasets, leading to greater model exposure to low bone mass patterns in females. Consequently, this imbalance in training data may have enhanced the model's ability to recognize low bone mass-related features in females, resulting in higher sensitivity. Additionally, model accuracy improved as low bone mass severity increased, with higher accuracy in the moderate and advanced subgroups compared to mild cases. Margaret L. et al. reported that progression to osteoporosis takes 17.3 years for mild low bone mass cases, 4.7 years for moderate, and 1.1 years for advanced, highlighting that frequent DXA scans may not be necessary for mild low bone mass cases [4]. Given that most of our model's false negatives are concentrated in the mild low bone mass group around T-scores of −1, and the model's accuracy improves significantly as the severity of low bone mass increases, our model could offer valuable clinical utility in guiding follow-up intervals for DXA testing, particularly when used opportunistically with CXRs taken for other medical reasons.

Another distinctive feature of our study is the higher performance observed in the AMC external test dataset compared to the AMC internal test dataset, despite the latter sharing the same configuration as the training dataset. This superior performance in the AMC external test dataset likely stems from differences in data source, clinical characteristics of participants, labeling standards, and data distribution. The AMC internal test dataset had a higher proportion of mild low bone mass cases (14.6%) compared to the AMC external test dataset (4.5%), while advanced low bone mass cases were more prevalent in the AMC external test dataset (26.1%) than in the AMC internal test dataset (15.2%). This discrepancy can be attributed to the fact that the AMC internal test dataset primarily consists of data collected from a health screening center, leading to a higher proportion of mild low bone mass cases, whereas the external dataset comprises data labeled in clinical settings, such as outpatient clinics, inpatient wards, and emergency departments, resulting in a higher prevalence of advanced low bone mass cases. Given that our model exhibits superior differentiation for advanced low bone mass compared to mild low bone mass, this difference in data distribution likely contributes to the higher performance observed in the external dataset. Additionally, the high proportion of female participants in the AMC external dataset (87.9%), consistent with the slightly female-dominant composition of the training data, may have contributed to enhanced model consistency, given the higher prevalence of low bone mass among women. Finally, both datasets were collected from the same hospital, ensuring high image quality with consistent equipment and imaging protocols, which supported stable performance across both datasets and contributed to the observed strengths of the AMC external dataset.

In addition to these domestic evaluations, we incorporated the Gradient Health (GH) external dataset to assess the model's applicability across ethnically and institutionally diverse populations. The GH dataset comprises CXR–DXA pairs collected from multiple institutions in the United States and Latin America. Despite substantial differences in demographic composition and clinical settings compared with Korean datasets, the model demonstrated comparable performance, achieving an AUC of 0.85. These findings suggest that OsPenScreen may maintain robust diagnostic accuracy across geographically and ethnically heterogeneous environments, supporting its potential for broader international deployment in opportunistic low bone mass screening.

Our Grad-CAM visualizations across four test datasets that the model that the model consistently highlighted weight-bearing bones, particularly the thoracic and lumbar spine, in normal cases, while focusing on non-weight-bearing bones, such as the clavicles and upper arms, in low bone mass cases. While DXA scans generally assess BMD in the hip and spine due to the high risk of fragility fractures in these areas, which greatly impacts mortality, the clavicle and other non-weight-bearing bones are not typically included in routine BMD assessments [18]. However, studies suggest that generalized bone loss from aging may reduce BMD in regions like the clavicle and upper arms [19]. Although bone loss occurs systemically, weight-bearing status significantly influences BMD, often resulting in higher density in bones that bear weight [20]. Accordingly, in normal cases, the model may focus on weight-bearing bones like the

thoracic, and lumbar spine. Conversely, in low bone mass cases, non-weight-bearing bones, such as the clavicle and scapula, might present early signs of bone density reduction, potentially drawing the model's focus to these regions. These interpretations are exploratory in nature and were not supported by standardized anatomical labeling or radiologic validation, which will be addressed in future work.

As the global population continues to age, the prevalence of low bone mass is rising, especially among older adults. Low bone mass and osteoporosis are more widespread than many other diseases that receive greater public attention. For example, while the lifetime risk of breast cancer in white women is one in nine, the risk of experiencing a hip fracture is one in six [21]. Among women over 45, osteoporosis results in more hospitalizations than major conditions like diabetes, heart attacks, and breast cancer [22]. In Europe, osteoporosis-related disability surpasses that of most cancers (except lung cancer) and rivals or exceeds the burden of chronic diseases like rheumatoid arthritis, asthma, and heart disease [23]. Despite these realities, low bone mass remains largely under-recognized. A survey by the International Osteoporosis Foundation (IOF) across 11 countries revealed that many postmenopausal women underestimate their personal risk, have poor communication with their healthcare providers about osteoporosis, and often lack access to diagnosis and treatment before their first fracture [24]. This contributes to the underdiagnosis and undertreatment of low bone mass.

To address this, several studies have explored opportunistic screening using X-rays acquired for unrelated clinical indications. A recent systematic review reported a growing number of AI-based osteoporosis prediction models, driven by the availability of imaging data and computational advances. Among the 26 studies reviewed, structured clinical data were the most common input (49%), followed by X-ray (27%), CT (15%), MRI (9%), DXA (5%), and QUS (2%). Models using clinical data generally outperformed image-only models, with AUCs ranging from 0.75 to 0.98. X-ray–only models, primarily based on lumbar spine and hip radiographs, showed moderate and consistent performance (AUC 0.78–0.83), likely due to anatomical correspondence with DXA sites [24].

Several CXR-only models have recently been developed, though most focus solely on binary classification of osteoporosis. Tsai et al. reported internal and external AUCs of 0.930 and 0.892, respectively [25]. Sato et al. achieved AUCs of 0.84 for osteoporosis, 0.70 for low bone mass, and 0.89 for normal vs low BMD [26]. Asamoto et al. reported an accuracy of 79.7%, sensitivity of 77.1%, and specificity of 80.4% [27]. While these models showed high performance for osteoporosis, their utility in detecting earlier-stage bone loss remains limited. The OsPenScreen model was developed to address this gap and achieved an AUC of 0.85 for low bone mass classification using a single CXR. The model can be integrated with existing osteoporosis classification models to enable extension of clinical screening from osteoporosis to low bone mass within a unified framework.

However, several limitations exist in this study. First, all DXA-labeled datasets used for model development and validation were collected exclusively from medical institutions in South Korea, limiting diversity in terms of ethnicity, body habitus, and imaging equipment. Such homogeneity may restrict the model's generalizability, as anatomical structure and radiographic appearance can vary across populations and clinical settings. To partially address this limitation, we evaluated model performance using the GH external dataset, which comprises CXR–DXA pairs collected from multiple institutions in the United States and Latin America. The model maintained strong performance in this multiethnic external cohort (AUC = 0.85), suggesting potential robustness across demographically diverse populations. Nevertheless, further large-scale, prospective validation in international and multi-ethnic cohorts, along with methodological adaptations to account for domain shift, is warranted to more definitively establish the model's generalizability. Second, we deliberately excluded patients with osteoporosis (T-score ≤ -2.5) to focus specifically on low bone mass, which may not fully replicate real-world clinical conditions. Nonetheless, our previous research successfully developed and validated an osteoporosis classification model [8], and this study builds on that by successfully classifying low bone mass with high performance, enabling future expansion to a three-class classification model: normal, low bone mass, osteoporosis. Specifically, this study presents only the second-stage model (OsPenScreen) of a two-stage classification framework. The first-stage model identifies patients with osteoporosis, and OsPenScreen is applied subsequently to classify the remaining non-osteoporotic population into normal and low bone mass. Due to submission policies and peer review status, we were unable to include the first-stage model in this manuscript. However, integration into a unified three-class model is currently planned. Third, the model frequently misclassified cases near a T-score of -1.0, leading to confusion between normal and low bone mass, which could present challenges in clinical settings. However, given that mild low bone mass typically progresses to osteoporosis over a span of more than 15 years, these misclassifications are unlikely to have a significant clinical impact. Addressing this issue may require expanding the dataset and applying data preprocessing or normalization techniques to ensure consistent performance across institutions and improve accuracy in boundary regions. Fourth, due to the retrospective nature of this study, we were unable to assess the clinical consequences of a positive screen, including whether and how follow-up diagnostic or therapeutic actions would be initiated. Furthermore, we did not compare the model's

utility against existing risk stratification methods, such as FRAX or age-based DXA referral strategies, nor did we evaluate clinical decision metrics such as decision curve analysis or net reclassification improvement. Future prospective studies are necessary to determine how this model can be integrated into routine clinical workflows and to establish its added value over existing screening tools.

In conclusion, the OsPenScreen model provides a promising solution for opportunistic low bone mass screening using CXRs. By identifying at-risk patients without additional radiation exposure or healthcare costs, this model has the potential to significantly reduce the incidence of osteoporosis-related fractures, hospitalizations, and mortality, while providing a practical and scalable screening tool in routine clinical settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11657-025-01609-1.

Funding This work was supported by the Ministry of SMEs and Startups of Korea through the R&D project under the Startup Growth Technology Development Program (Project No. RS-2024–00446243).

Data availability Except for publicly available datasets, such as Chest X-ray 8 by NIH dataset and CheXpert dataset, the data used in this study cannot be made publicly available.

Declarations

Conflicts of interest Namkug Kim and Hyun-Jin Bae declare that they are the stakeholders of Promedius Inc.

References

1. Bliuc D, Nguyen ND, Milch VE, Nguyen TV, Eisman JA, Center JR (2009) Mortality risk associated with low-trauma osteoporotic fracture and subsequent fracture in men and women. JAMA 301(5):513–521. https://doi.org/10.1001/jama.2009.50
2. Ahn SH, Park SM, Park SY et al (2020) Osteoporosis and osteoporotic fracture fact sheet in Korea. J Bone Metab 27(4):281–290. https://doi.org/10.11005/jbm.2020.27.4.281
3. Fan Y, Li Q, Liu Y et al (2024) Sex- and age-specific prevalence of osteopenia and osteoporosis: sampling survey. JMIR Public Health Surveill 10:e48947. https://doi.org/10.2196/48947
4. Gourlay ML, Fine JP, Preisser JS et al (2012) Bone-density testing interval and transition to osteoporosis in older women. N Engl J Med 366(3):225–233. https://doi.org/10.1056/NEJMoa1107142
5. Dimai HP (2017) Use of dual-energy X-ray absorptiometry (DXA) for diagnosis and fracture risk assessment; WHO-criteria, T- and Z-score, and reference databases. Bone 104:39–43
6. Mithal A, Bansal B, Kyer CS, Ebeling P (2014) The Asia-Pacific regional audit—epidemiology, costs, and burden of osteoporosis in India 2013: a report of the International Osteoporosis Foundation. Indian J Endocrinol Metab 18(4):449–454
7. Ohta Y, Yamamoto K, Matsuzawa H, Kobayashi T (2020) Development of a fast screening method for osteoporosis using chest X-ray images and machine learning. Can J Biomed Res Technol 3:3–9
8. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N (2022) Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. J Bone Miner Res 37(2):369–377. https://doi.org/10.1002/jbmr.4477
9. Wang X, Peng Y, Lu L, et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017:3462–3471
10. Irvin J, Rajpurkar P, Ko M et al (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell 33(01):590–597
11. Ha YC, Yoo JI (2021) Cross-calibration of bone mineral densities and body composition between GE-lunar prodigy and Osteosys primus. J Bone Metab 28(3):215–222. https://doi.org/10.11005/jbm.2021.28.3.215
12. Chen L, Yu Z, Huang J et al (2023) Development of lung segmentation method in x-ray images of children based on TransResUNet. Front Radiol 3:1190745
13. Lee D (2013) Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. Presented at: ICML Workshop: Challenges in Representation Learning (WREPL)
14. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
15. Singh S, Bajorek B (2014) Defining, "elderly" in clinical practice guidelines for pharmacotherapy. Pharm Pract (Granada) 12(4):489. https://doi.org/10.4321/S1886-36552014000400007
16. Baratella E, Fiorese I, Minelli P et al (2023) Aging-related findings of the respiratory system in chest imaging: pearls and pitfalls. Curr Radiol Rep 11(1):1–11. https://doi.org/10.1007/s40134-022-00405-w
17. Cohen JF, Korevaar DA, Altman DG et al (2016) STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 6(11):e012799. https://doi.org/10.1136/bmjopen-2016-012799
18. Warriner AH, Patkar NM, Curtis JR et al (2011) Which fractures are most attributable to osteoporosis? J Clin Epidemiol 64(1):46–53. https://doi.org/10.1016/j.jclinepi.2010.07.007
19. Gehlbach S, Saag KG, Adachi JD et al (2012) Previous fractures at multiple sites increase the risk for subsequent fractures: the global longitudinal study of osteoporosis in women. J Bone Miner Res 27(3):645–653. https://doi.org/10.1002/jbmr.1476
20. Sundh D, Nilsson M, Zoulakis M et al (2018) High-impact mechanical loading increases bone material strength in postmenopausal women: a 3-month intervention study. J Bone Miner Res 33(7):1242–1251. https://doi.org/10.1002/jbmr.3431
21. Van Staa TP, Dennison EM, Leufkens HG, Cooper C (2001) Epidemiology of fractures in England and Wales. Bone 29(6):517–522. https://doi.org/10.1016/S8756-3282(01)00614-7
22. O'Neill TW, Felsenberg D, Varlow J et al (1996) The prevalence of vertebral deformity in European men and women: the European Vertebral Osteoporosis Study. J Bone Miner Res 11(7):1010–1018. https://doi.org/10.1002/jbmr.5650110719
23. Johnell O, Kanis JA (2006) An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 17(12):1726–1733. https://doi.org/10.1007/s00198-006-0172-4
24. International Osteoporosis Foundation. Epidemiology. https://www.osteoporosis.foundation/health-professionals/about-osteoporosis/epidemiology
25. 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
26. Sato Y, Yamamoto N, Inagaki N et al (2022) Deep learning for bone mineral density and T-score prediction from chest x-rays: a

multicenter study. Biomedicines 10(9):2323. https://doi.org/10.3390/biomedicines10092323
27. Mao L, Xia Z, Pan L et al (2022) Deep learning for screening primary osteopenia and osteoporosis using spine radiographs and patient clinical covariates in a Chinese population. Front Endocrinol (Lausanne) 13:971877. https://doi.org/10.3389/fendo.2022.971877
28. Korean Society for Bone and Mineral Research. Physician's guide for diagnosis & treatment of osteoporosis. 2024. https://www.ksbmr.org/bbs/?code=guideline