Automated and Precise Bone Mineral Density Prediction and Fracture Risk Assessment using Hip/Lumbar Spine Plain Radiographs via Learning Deep Image Signatures and Correlations

Abstract

Dual-energy X-ray absorptiometry (DXA) and the Fracture Risk Assessment Tool are recommended tools for osteoporotic fracture risk evaluation, but are underutilized. We present a novel and fully-automated tool to identify fractures, predict bone mineral density (BMD), and evaluate fracture risk using plain pelvis and lumbar spine radiographs. The performance of this tool were evaluated in 1639 and 11908 patients with pelvis or lumbar spine radiographs and DXA, respectively. The model was well calibrated for hip and spine BMD assessments with minimal or no bias. The area under the curve and accuracy were 0.89 and 92.4% for hip osteoporosis, 0.87 and 86.8% for spine osteoporosis, 0.92 and 94.6% for high 10-year major fracture risk, and 0.92 and 92.2% for high hip fracture risk, respectively. The success rates of our automated algorithm a real-world test were 85.3% and 90.4% for hip and spine, respectively. The clinical use of this automated tool may increase the likelihood of identifying high-risk patients in previously unscreened populations.

Introduction

Osteoporosis is a common bone disease[1] that poses an increasing global health burden.[2] All major types of osteoporosis-related fragility fractures are associated with chronic pain, disability, functional dependence,[3] and enhanced morbidity.[4] In addition, patients with fragility fractures have a two- to threefold increase in mortality,[5] despite the availability of effective anti-osteoporotic drugs.[6] Dual-energy X-ray absorptiometry (DXA) is the current preferred modality for measurement of bone mineral density (BMD) in the human hip or lumbar spine, which is the essential component of the fracture risk assessment tool (FRAX™) used to estimate the 10-year risk of hip or major osteoporotic fracture.[7] Currently, both DXA and FRAX™ are underutilized, despite their usefulness in the identification of patients at risk and in treatment decision-making processes.[8] Among Medicare beneficiaries ≥ 65 years of age, only 30% of women and 4% of men were tested for BMD with DXA.[9] In addition, only 10.2% of female[10] and 6% of male patients[11] with fragility fractures have undergone BMD testing before the index fracture. Opportunistic screening for osteoporosis using imaging modalities other than DXA is a potential strategy to effectively and feasibly stratify the unscreened population into distinct risk groups regarding osteoporosis and fragility fractures. For example, several studies used computed tomography (CT)-based matrices to classify osteoporosis,[12] simulate DXA T-scores,[13] and predict fracture outcomes.[14] However, the performance, radiation dose, and population coverage of CT-based screening strategies are barriers to their use in clinical settings.

Unlike DXA and CT, plain radiography has greater availability, broader indications, lower radiation dose, and lower overall costs. The spatial resolution of radiographs is excellent, allowing the visualization of fine bone texture, which is correlated with bone density[15] and can distinguish patients with osteoporotic fractures from controls.[15, 16, 17] Therefore, an automated tool based on hip or spine radiographs for identifying hip fracture and vertebral compression fracture (VCF), predicting BMD, and evaluating fracture risk can help identify patients with greater fracture risk among individuals undergoing radiography of the hip or spine for other reasons. Deep learning algorithms have achieved performance superior to traditional methods in many visual recognition tasks, including object detection, localization, and classification.[18] The abilities of deep neural networks to learn, identify and optimize essential task-effective image features and decision-making functions from large amounts of images are essential to their success in terms of fracture detection,[19] retinopathy grading,[20] and lung nodule identification.[21]

On the basis of our clinical experience, we hypothesized that changes in fine bone microarchitecture associated with the process of age-related bone loss could be visualized on radiographs and used for reliable prediction of BMD by visual recognition models. To test this hypothesis, we proposed and validated a novel deep learning method to predict BMD using hip and spine radiographs on paired data collected from radiographs and DXA. Our deep learning models have a three-step workflow: extraction of the region of interest (ROI) enabled by a highly robust deep adaptive graph landmark detection algorithm; automated quality control to exclude unqualified images from BMD estimation; and deep neural network joint processing of the ROI and the patient’s clinical information to calculate the BMD. We calculated the predicted BMD-based 10-year risks of major osteoporotic fracture and hip fracture, then compared these risk estimates with risks determined using DXA-based BMD measurements. To fully automate the process with regard to reproducibility, robustness, and performance reliability, we created a set of additional algorithms to localize the ROIs (proximal femur or L1–4 lumbar vertebrae), identify hip fracture or VCF, and check the radiograph quality to ensure that implants and foreign bodies were absent from the ROIs (figure 1). The automated precise ROI localization, hip and L1–4 vertebrae segmentation, detection of hip fracture and VCF, quality check for the images, inference of BMD, and FRAX risk reporting were packaged into a single tool, which was implemented on the inference server for clinical application.

Results

From 2006 to 2017, 25960 patients with paired DXA-pelvis radiographs (17.0% of patients with pelvis radiographs) and 72059 patients with paired DXA-lateral radiographs of the lumbar spine (16.8% of patients with lumbar radiographs) were screened to identify hip and spine cohorts for analysis. The first data pairs from patients with DXA and radiographs performed within 180 days were included. For patients with multiple DXA examinations, the earliest examination was used as the index DXA. For each index DXA, the radiographs with the shortest interval to DXA were chosen. After the exclusion of patients without complete data, patients with data obtained using a GE DXA scanner, and patients with radiographs of inadequate quality, 3295 patients in the hip cohort (training set: 1602; testing set: 1693 patients), and 16908 patients in the spine cohort (training set: 5000; testing set: 11908 patients) were included in the analysis (figure 2). No patient was included in more than one group.

Table 1 presents patient characteristics for the training and testing sets. The final hip testing set consisted of 1693 patients (1313 women, 77.6%; mean age, 69.6 ± 11.8 years). The median time between DXA and hip radiographs was 17 days (interquartile range, 2–56 days). Mean DXA-measured BMD was 0.692 ± 0.154 g/cm^2, which did not significantly differ from the predicted value (0.691 ± 0.144 g/cm^2, P = 0.35). In the hip testing set, 346 patients (20.4%) had osteoporosis with T-score ≤ −2.5. In the final spinal testing set, 11908 patients (9387 women, 78.8%) were included in the analysis. The mean age was 66.6 ± 10.8 years, and the median time between DXA and spine radiographs was 15 days (interquartile range, 5–43 days). After quality assessment to exclude unsuitable vertebrae, 33299 lumbar vertebrae were included in the analysis (70.0%). The mean BMD per vertebra was 0.852 ± 0.191 g/cm^2, which was significantly greater than the mean predicted value (0.846 ± 0.172 g/cm^2; $P < 0.001$); however, this difference was trival and not clinically meaningful. These trends were similar across L1–L4 and both age and sex strata, but the differences were not statistically significant. In the spine testing set, 5747 patients (48.3%) had osteoporosis (T-score ≤ −2.5 in the vertebra with the lowest T-score).

Table 2 summarizes the model performance to predict BMD using hip or lumbar spine radiographs. Pearson's correlation coefficients between DXA-measured and model-predicted BMD were 0.93 for the hip and 0.92 for the lumbar spine, suggesting excellent correlations. The linear regression model showed excellent predictive performance of predicted BMD with regard to measured BMD (hip: R^2 = 0.87, root mean square error = 0.056; spine: R^2 = 0.86, root mean square error = 0.065). The model was well calibrated in the hip (slope = 0.998, calibration-in-the-large = 0.001), as shown in the calibration plot (Figure 3a). For the lumbar spine BMD, model prediction tended to slightly underestimate BMD, although the difference was trival and not clinically significant (Figure 3b). Bland–Altman analysis of BMD indicated no significant differences between predicted and measured hip BMD (bias of −0.001 g/cm^2; 95% confidence interval, −0.004 to 0.001). A small bias of −0.005 g/cm^2 (95% confidence interval, −0.006 to −0.004) was noted for lumbar spine BMD prediction. As shown in Table 2, the model performance remained consistent across various age and sex strata, demonstrating that the algorithm was robust.

Table 3 illustrates the discriminatory performance of the model to classify hip or spine osteoporosis and identify patients with greater 10-year risks of major osteoporotic fractures (≥ 20%) and hip fractures (≥ 3%). The algorithm provided a high degree of discrimination for osteoporosis (area under the receiver operating characteristic curve [AUC], 0.89 for the hip and 0.87 for the spine). The overall accuracies were 92.4% for hip osteoporosis and 86.8% for lumbar spine osteoporosis. The median FRAX 10-year major fracture and hip fracture risks did not significantly differ when scores were based on the predicted BMD (10.81% and 2.81%, respectively; $P = 0.79$) and when scores were based on the measured BMD (10.68% and 2.78%, respectively; $P = 0.74$). The classification performances regarding patients with high 10-year risks of major osteoporotic fractures and hip fractures were better than the osteoporosis classification performance, with AUCs of 0.92 and 0.92, and accuracies of 94.6% and 92.2%, respectively. As shown in Table 4, the network performances for classification of osteoporosis and identification of patients with high risks of major and hip fractures were robust across all age and sex groups, despite significant differences in the association strength ($P < 0.001$).

Next, we packaged the ROI localization, fracture detection, image quality check, BMD estimation, osteoporosis detection, and FRAX risk evaluation into two standalone tools for the hip and spine, respectively. We implemented the tools in the central inference platform connected to the picture archiving and communication system (PACS) in the Chang Gung Memorial Hospital (Linko branch). The hospital PACS transferred all newly acquired images to the inference platform on a daily basis.

In total, 7353 consecutive pelvis radiography examinations were conducted from March 2020 to November 2020. The tool identified 1013 radiographs that had bilateral total hip replacement, hip fractures, or the presence of other image quality issues that may impede BMD estimation. The remaining 6271 (85.3%) images were successful in predicting BMD. From November 2020 to January 2021, we collected 11291 consecutive lateral radiographs of the lumbar spine. The tool identified 1084 radiographs with VCFs, implants, vertebroplasty, or other features that may impede BMD estimation. The success rate to produce predicted BMD for a single spine radiograph was 90.4% (10208 radiographs).

Discussion

Osteoporosis is a silent disease before fragility fractures, which often leads to multiple morbidities and increased mortality in affected patients.[4] Previous studies estimated that one in three women and one in five men aged > 50 years will experience fragility fractures in their lifetime.[22, 23] There is increasing evidence regarding the effectiveness and cost-effectiveness of therapeutic agents in the prevention of fragility fractures.[24, 25] Therefore, population-based screening is imperative for the identification of at-risk patients and implementation of preventive measures. However, current DXA-based programs screen fewer than one-third of eligible women and one-tenth of eligible men.[9] Therefore, osteoporosis screening based on DXA seems inadequate. In CGMH, approximately 17% of patients with pelvis or spine radiographs in our hospital previously underwent DXA-based assessment of BMD. This study developed an automated, reliable tool to evaluate fracture risk using hip or spine radiographs to effectively broaden the screening population and increase the number of identifiable high-risk patients.

The performance of the tool is robust with DXA as reference and compared favorably with non-DXA modalities, such as quantitative bone ultrasound (AUC, 0.762),[26] CT-based opportunistic screening using CT attenuation of the spine (AUC, 0.83),[12] and machine-learning-based T-score simulation (accuracy, 82%)[13] to classify osteoporosis. In addition to effective identification of patients with osteoporosis, the tool accurately predicted FRAX risk and identified patients with high risks of major osteoporotic (AUC 0.92; accuracy, 94.6%) or hip fractures (AUC 0.92; accuracy, 92.2%). Our real-world clinical assessment using consecutive pelvis radiographs also demonstrated that 85.3% of patients with pelvis radiographs and 90.4% of patients with spine radiographs could be automatically screened for osteoporosis and evaluated for future fracture risk. Importantly, most such patients had never been screened by DXA. Taken together, the results of this study demonstrated that the radiograph-based screening tool could accurately identify patients with osteoporosis and high fracture risk from previously unscreened population.

BMD is not the only determinant of fracture risk. The National Osteoporosis Risk Assessment study found that 82% of osteoporotic fractures occurred in women with T-score > −2.5, and 67% occurred in women with T-score > −2.0.[27] Other risk factors (e.g., history of osteoporotic fracture) are essential for the identification of high-risk patients. However, many patients with occult hip fractures and VCFs are asymptomatic, and are often diagnosed with other imaging modalities.[12, 28] We exploited the excellent spatial resolution of radiographs to identify unsuspected fragility fractures during the preprocessing and quality control process, prior to estimation of BMD. For hip fracture detection, we incorporated our previously published PelviXNet algorithm[29] to detect hip fracture. We also developed a vertebral fracture assessment algorithm based on a Deep Adaptive Graph network, which determines anatomical landmarks for standard six-point vertebral morphometry that facilitates VCF detection using the widely accepted semiquantitative Genant visual method.[30, 31] The overall model performance improved after the exclusion of hip or vertebral fractures. The integrated process automated the identification of hip and vertebral fractures, providing initial quality control for BMD estimation. This process also identified clinical risk factors for fragility fractures, without a requirement for clinical input. Therefore, our tool could evaluate fragility fracture risk based on a single radiograph (existing fractures and predicted BMD) and its age and sex metadata. However, other patient-related clinical risk factors (e.g., history of hip fracture, comorbidity, medication, and lifestyle) require input from electronic medical records.

Opportunistic screening for osteoporosis using other imaging modalities has been assessed previously. The best studied strategy is the use of abdominal CT to predict BMD;[13, 32, 33] classify osteoporosis based on CT attenuation,[12] simulated BMD,[32, 33] T-score,[13] or detection of osteoporotic fractures;[34] or use imaging biomarkers to predict the risk of fractures.[14] An earlier study compared the CT Hounsfield units over a manually annotated ROI involving vertebral body trabecular bone with its paired DXA T-score; this approach for detection of osteoporosis yielded an AUC of 0.83.[12] A deep learning-based model provided a better correlation between predicted and reference values, but its validation included only small datasets.[13, 32, 33] A larger study testing the performance of simulated T-scores on a larger dataset of 1843 CT-DXA pairs achieved an accuracy of 82% to detect osteoporosis.[13] This algorithm was integrated with VCF identification and CT trabecular density as biomarkers, and its performance for the prediction of 5-year fracture risks was compared with the performance of FRAX alone (i.e., without BMD input). This CT-based predictor provided automated risk evaluation using CT-derived metrics and compared favorably with FRAX prediction.[14] Osteoporosis and fragility fracture risk have also been assessed on dental,[35, 36] hip,[37] and spine radiographs,[36] as well as magnetic resonance imaging.[38] These studies demonstrated the feasibility of using non-DXA modalities to expand opportunistic screening to a broader population at risk, although the applicability and usability of such tools in real clinical settings are questionable.

In contrast, the present study provided a fully automated tool enabling opportunistic screening for osteoporosis and evaluation of fragility fracture risk using plain radiographs of the hip and spine. Our tool utilizes ubiquitous, low-cost radiographs that involve substantially lower radiation exposure than CT-based tools, thus maximizing the likelihood that eligible populations will be screened, regardless of DXA or CT scan availability. Our tool can assess both the hip and spine, and is therefore not limited to the spine alone (e.g., during evaluation with CT-based tools). Furthermore, we envision that other musculoskeletal radiographs may also be used to predict bone density and risks of fracture, regardless of the original purpose of the images. This strategy requires no additional patient time or radiation exposure and involves minimal costs, but may substantially improve the risk profiling for fragility fractures.

This study had several limitations. First, Chang Gung Memorial Hospital is a medical center in which the patients tend to have more severe disease. A large proportion of patients have fractures or implantations. Our study population may have not represented the healthier population, which is the target of osteoporosis screening. However, because the tool was developed based on this more complex population, the ROI localization, quality check, and BMD prediction processes can presumably be readily adapted to populations with fewer complications. Second, the calculation of FRAX in this study did not consider past medical history, medication use, family medical history, and lifestyle (e.g., alcohol consumption and smoking status) because this information requires input from the hospital information system. However, the performance assessment should not change because these parameters are identical for FRAX based on the DXA-measured or model-predicted BMD. For clinical implementation, the tool can be modified to report full FRAX results when digital data are available. Third, the tool was created using the reference BMD values reported by Hologic DXA scanners alone, although both Hologic and GE DXA scanners are actively used at Chang Gung Memorial Hospital. Systematic differences in BMD measurement and reporting between DXA manufacturers hampered the tool's performance in our early experiments. Manufacturer-specific models may be needed in some clinical settings. Fourth, the performance of the prediction tool is influenced by radiograph image quality. In addition to existing fractures, accurate BMD prediction may be impeded by foreign bodies, implants, bowel gas, and bone pathologies (e.g., avascular necrosis or severe osteoarthritis). The actual rate of radiographs that could be evaluated for BMD and fracture risk surpassed 85% in our real-world test. Depending on a patient's specific indications, radiographs are often examined repeatedly. Therefore, the per-patient success rate will potentially increase as more radiographs become available over time.

This study demonstrated that a robust opportunistic screening tool for osteoporosis and fracture risk assessment, based on conventional radiographs obtained for various indications, was able to provide VCF detection, BMD, and fracture risk estimation in a fully automated process. This tool leveraged state-of-the-art deep learning algorithms to provide a more efficient strategy for population-based opportunistic screening with minimal or no additional cost. The integration of this automated tool into the hospital information system may increase the likelihood of identifying high-risk patients in previously unscreened populations.

Methods

Hypothesis and study design

This retrospective cohort study was performed to test the hypothesis that an automated deep neural network-based tool could effectively predict BMD and risk of fragility fractures using plain radiographs of pelvis and lumbar spine. This tool is a collection of algorithms to identify and segment regions of interest (hip or lumbar spine), check for factors that would influence BMD prediction (e.g., image quality/positioning, existing hip or vertebral fractures, implants, and/or foreign bodies), and subsequently predict hip and vertebral BMD and fracture risk. We compared the predicted BMD with the BMD measured by central DXA. We also calculated the risks of 10-year hip and major osteoporotic fractures using FRAX tools (https://www.sheffield.ac.uk/FRAX/). The fracture risk prediction performance was compared between algorithm-predicted BMD and DXA-measured BMD.

Setting

This study was approved by the Institutional Review Board at the Chang Gung Memorial Hospital (Taiwan) and was conducted in accordance with the tenets of the Declaration of Helsinki. The requirement for informed consent was waived because the data presented in this paper were fully de-identified to protect patient confidentiality. This study was performed using data from Chang Gung Memorial Hospital, the largest private hospital system in Taiwan, which includes seven acute hospitals with 10050 beds, that received 8.2 million outpatient visits and 2.4 million inpatient care visits. The study was conducted in collaboration between the Chang Gung Memorial Hospital and PAII Inc., a research subsidiary of Ping-An Technology that focuses on state-of-the-art computer vision algorithm development. PAII Inc. used clinical images and clinical data from Chang Gung Memorial Hospital to create automated BMD and fracture risk estimation tools. The provided data were fully encrypted to prevent patient confidentiality leaks. With the exception of the training and validation data, PAII Inc. remained blinded to other clinical and testing datasets.

The study population consisted of patients who had at least one central DXA during January 2006 to December 2020 and were aged 40–90 years on the DXA index date. From the central database of Chang Gung Memorial Hospital, 154332 patients aged 40–90 years with at least one central DXA examination between January 2006 and December 2017 were identified. The study population was also required to have adequate radiographs of the hip or lumbar spine within 180 days. For patients with multiple DXA and plain film radiographs, the earliest pair was used. We performed a quality check for plain films to ensure that these images were suitable for BMD prediction; after the exclusion of inadequate plain films, model building and testing were performed based on a cohort of 3241 patients with at least one DXA-pelvis radiograph pair and 16908 patients with at least one lateral radiograph of the lumbar spine–DXA pair (Figure 1).

We also tested the algorithms in a clinical setting to ascertain the proportions of patients with hip or spine radiographs who may benefit from the tool. The algorithms were packaged in docker containers and implemented on the PACS-linked inference platform of Chang Gung Memorial Hospital, based on the Nvidia Triton architecture. We collected consecutive pelvis radiographs conducted between March 2020 and November 2020 and spine radiographs conducted between November 2020 and February 2021.

BMD measurement

Proximal femoral and lumbar spine DXA scans were performed using a Hologic QDR-4500A fan-beam densitometer (Hologic, Inc., Bedford, MA, USA) during the period 2005–2010 and a Hologic Discovery model A densitometer during the period 2011–2021. The scans were analyzed in accordance with the manufacturer's recommendations. Hip T-scores were calculated using the revised NHANES III white female reference values.[39, 40] Because there is no international reference standard for the lumbar spine BMD, lumbar T-scores were calculated using the manufacturer's reference values. For each patient, the lowest T-score of the femoral neck or lumbar vertebrae was used to categorize osteoporosis or calculate FRAX risk.

Acquisition and preprocessing of radiographs

The radiographs were collected from the PACS and anonymized before the study procedure. The images were converted to grayscale and resized to a resolution of 0.15 mm × 0.15 mm pixel spacing, then stored as 12-bit images. A deep adaptive graph (DAG) landmark detection method was developed to formulate the anatomical landmarks of the pelvis and spine as graphs, and to robustly and accurately detect these landmarks.[41] We detected 16 anatomical landmarks on hip radiographs, including 12 landmarks on the pelvic boundary and four landmarks on the femoral head and trochanter. We detected six anatomical landmarks for each of the lumbar vertebrae on spine radiographs from L1 to L4. Based on the detected anatomical landmarks, ROIs were extracted from the radiographs and used as input for the BMD prediction model. For hip radiographs, ROIs of the left and right hips were extracted. For the lumbar spine, ROIs were extracted for each vertebra from L1 to L4. Examples of the detected anatomical landmarks and ROIs are shown in Figure 1. The ROIs were used as input for the BMD prediction model. A schematic representation of the pipeline and models used to predict BMD is shown in Figure 1.

Anatomical landmark detection via Deep Adaptive Graph

The anatomical landmarks were detected using DAG, a method introduced in our previous publication.[41] In DAG, the anatomical landmarks are formulated as a graph, $G = (V, E, F)$, where the vertices V represent the landmarks, the edges E represent the relationships between them, and the features F encode visual patterns in the neighborhoods of the vertices. For a specific input image, the graph vertices are first initialized in the image using the mean shape of the anatomy, and a neural network is used to displace the vertices from the initial position to the target anatomy in the image. By formulating the anatomical landmarks as a graph and modeling their displacements by convolutional neural network–graph convolutional network (GCN), DAG can effectively exploit the structural information and shape prior to the anatomical landmarks. Therefore, DAG provides robust and accurate anatomical landmark detection on both hip and spine radiographs.

The neural network consists of a convolutional neural network to encode the input image to produce graph features F, and a GCN to process the graph to locate its vertices. Specifically, the GCN consists of two parts: a global transformation GCN and multiple local refinement GCNs. The global transformation GCN produces an affine transformation matrix, M, which brings the initial graph vertices closer to the target. The transformed vertices are written as follows:

$$V^1 = \{{v_i}^1\} = \{M{v_i}^0\},$$

where $V^0 = \{{v_i}^0\}$ and $V^1 = \{{v_i}^1\}$ denote the initial graph vertices before and after the estimated affine transformation, respectively. The local refinement GCNs then iteratively estimate the displacements of the graph vertices V^1. In each iteration, the vertices are displaced as follows:

$${v_i}^{t+1} = {v_i}^t + \mathit{\Delta}{v_i}^t,$$

where $\mathit{\Delta}{v_i}^t$ is the displacement estimated by the local refinement GCN at the t-th step.

During training, the training loss is calculated for both the global transformation

GCN and the local refinement GCNs. Because the goal of global transformation GCN is to locate the anatomy coarsely, the following margin loss is used:

$$L_{global} = \left[\frac{1}{N} \sum_{i \in N} |{v_i}^1 - v_i| - m \right]_+ ,$$

where $[u]_+ = max(0, u)$; ${v_i}^1$ and v_i denote the globally transformed and ground truth vertices, respectively; and m is a hyperparameter representing a margin that aims to achieve high robustness for coarse landmark detection and forgive small errors. To encourage the local refinement GCNs to learn a precise localization, L1 loss is directly applied to all vertices after the refinements, written as follows:

$$L_{local} = \frac{1}{N} \sum_{i \in N} |{v_i}^T - v_i|,$$

where ${v_i}^T$ denotes the vertices after the last local refinement GCN. The graph edge weights are treated as learnable parameters, which are initialized randomly at the beginning of training and updated via back-propagation during training. In our experiment, the hip and spine DAG models were trained using 3306 and 1076 pelvic and spine radiographs with expert annotations.

Automated radiograph quality assessment procedure

Some medical conditions may affect the hip and vertebra anatomy, making plain films unsuitable for BMD estimation. The most common conditions include implantation (e.g., total hip replacement or spine fusion) and fracture. Therefore, we conducted an automated quality assessment to exclude hips and vertebrae with implants or fractures that were unsuitable for BMD prediction.

Quality assessment of hip radiographs: We used an existing model, PelviXNet,[29] to detect hip fracture. PelviXNet consists of a DensetNet-121 backbone neural network and a Feature Pyramid Network, and was trained on 5204 pelvic radiographs that had been annotated by experienced physicians using an efficient and flexible point-based annotation scheme. For each input hip plain film, PelviXNet produces a dense probability map indicating the likelihood of fracture sites. The maximum response in any hip ROI is regarded as the classification score for hip fracture. In this study, hips with classification scores > 0.5 were identified as fracture cases and excluded from downstream processing.

Automated quality assessment procedure for spine radiographs: We performed quality assessment in two steps: implant and VCF detection, and six-point morphology analysis. The implant/VCF detection model had a DenseNet and feature pyramid network architecture identical to that of PelviXNet, and was trained on 1485 expert-annotated lateral spine radiographs to produce probability maps for implant and VCF. The L1 to L4 vertebrae were classified as normal, VCF, and implant by the annotator. A supervision mask was then generated by filling the vertebra polygons produced by DAG using the annotated label. Specifically, pixels in the background and normal vertebrae were assigned the background class label (i.e., 0). Pixels in the vertebrae with VCF and implant were assigned two distinct foreground class labels (i.e., 1 and 2). During training, pixel-wise categorical cross entropy was calculated between the produced probability maps and the supervision mask, written as follows:

$$CE = -\sum_{x,y}\sum_{i}^{C} t_i(x,y) log(p_i(x,y)),$$

where i denotes the class index, and $t_i(x, y)$ and $p_i(x, y)$ denote the label from the supervision mask and the predicted probability at pixel (x, y), respectively.

Using the predicted implant and VCF probability maps, the maximum responses in the vertebrae polygons were regarded as the classification scores. Vertebrae with a positive implant or VCF detection results were excluded, and the remaining vertebrae were analyzed by six-point morphology. Specifically, six landmarks were detected for each vertebra, including two anterior points, two posterior points, and two middle points of the top and bottom vertebral plates. Four distances were calculated from these six points: anterior height h_a, posterior height h_p, middle height h_m, and vertebra width w. The three heights were calculated as the pairwise distances between the two anterior, posterior, and middle points. The vertebra width was calculated as the mean distance between the anterior and posterior points. Three criteria were used to identify vertebrae with abnormal deformity, in accordance with the widely accepted Genant visual semiquantitative method,[31] with modifications to facilitate automated measurement and fracture detection:

$$\frac{\min(h_a, h_p)}{\max(h_a, h_p)} < 0.8,$$

$$\frac{h_m}{\max(h_a, h_p)} < 0.7,$$

$$\frac{\max\,(h_a, h_p)}{w} < 0.55.$$

The first criterion aimed to detect wedge and crush fractures, where the anterior and posterior heights were reduced. The second criterion aimed to detect a biconcave fracture, where the middle height was reduced. The last criterion aimed to detect severe VCF cases where the overall height of the vertebra was significantly reduced.

If a vertebra met any of the three criteria, it was considered abnormal and excluded from downstream processing. These criteria only detected apparent moderate to severe compression fractures to avoid ambiguity in determining mild or borderline deformities.

Algorithm development and training procedure for BMD prediction

We developed a deep learning algorithm to estimate the hip/spine BMD from each corresponding ROI. The neural network used a VGG-16 with batch normalization and squeeze-and-excitation block as the backbone to encode the input image. Compared with deeper and more complex backbone networks (e.g., ResNet and DenseNet), our empirical results indicated that a VGG-16 block with a shallower architecture achieved better BMD prediction performance. We hypothesized that the visual patterns correlated with the BMD were at lower levels (e.g., texture and cortical bone structure), which could be effectively modeled by shallow networks (i.e., no greater object-level abstraction is needed). Because patient age and sex were correlated with BMD, we added this information to the neural network to assist in BMD prediction. In particular, features extracted by the VGG-16 block were first flattened and processed by a fully connected layer to obtain a feature vector of length 512. The patient age and sex information were represented by two values and concatenated with the feature vector. Because L1–L4 vertebrae have distinct BMD statistics, the vertebra index information was required by the model to accurately predict the BMD. Therefore, in the spine model, the vertebra index (from L1 to L4), encoded by a one-hot vector of length 4, was also concatenated with the feature vector (in addition to the encoded patient age and sex information). During training, ROIs were augmented by random affine transformation and subsequently resized to 512 × 512 pixels, then augmented by intensity jittering (varying the brightness from

−0.2 to +0.2 and the contrast from −0.2 to 0.2). The L1 distance between the predicted BMD and the ground truth BMD obtained from DXA was regarded as the training loss. A fourfold cross-validation procedure was conducted, and ensemble learning was adopted to combine the predictions of the four trained models during inference.

Implementation details

Deep learning models were developed on a workstation with a single Intel Xeon E5-2650 v4 CPU @ 2.2 GHz, 128 GB RAM, and 4 NVIDIA TITAN V GPUs running Ubuntu 18.04 LTS. All code used in this study was developed in Python v3.6, and deep learning models were implemented using PyTorch v1.3. Image preprocessing was performed using the Python Imaging Library. ImageNet pre-trained weights were used to initialize the backbone network VGG-16 block. The Adam optimizer was used to train the model for 200 epochs with a batch size of 8, a starting learning rate of $1e^{-4}$, and a weight decay of $1e^{-4}$. The learning rate was reduced to $1e^{-5}$ after the first 100 training epochs. The trained model was evaluated on the validation set after each training epoch, and the model with the highest validation correlation coefficient r-value was selected as the best model.

Evaluation of BMD prediction performance

Evaluation of all performance measures was performed only on the test datasets. The Bland–Altman plot visualized the agreement between predicted and measured BMD scores, and Pearson’s correlation coefficient was calculated. Calibration of the tool was evaluated by comparison between the mean risk calculated based on predicted BMD and the mean risk based on DXA-measured BMD. The following measures were calculated to evaluate the overall calibration: calibration slope and calibration-in-the-large. Osteoporosis results were considered positive when T-score ≤ −2.5. Ten-year probabilities of major fracture and hip fracture with femoral neck BMD were calculated for each patient using the FRAX tool with risk estimators specific to the Taiwanese population (https://www.sheffield.ac.uk/FRAX/; FRAX® Desktop Multi-Patient Entry, version 4.0). For each patient, the lowest BMD was used to calculate the T-score and FRAX risk. Patients were also categorized in accordance with the Taiwan Osteoporosis Practice Guidelines. Ten-year risk scores of ≥ 3% for hip fracture and ≥ 20% for major osteoporotic fracture were considered high-risk, based on the intervention threshold established in the Taiwan Osteoporosis Practice Guidelines[42] and the recommendations of the National Osteoporosis Foundation.[43] The overall discriminative abilities to discern osteoporosis and high-risk patients were evaluated using the AUC. Other measures were also calculated, including sensitivity, specificity, positive predictive value, and negative predictive value. Analyses were conducted using Stata software, version 16 (StataCorp, College Station, TX, USA).

References

1. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos Int* **17**, 1726-1733 (2006).

2. Sanchez-Riera L*, et al.* The global burden attributable to low bone mineral density. *Ann Rheum Dis* **73**, 1635-1645 (2014).

3. Cree M, Carriere KC, Soskolne CL, Suarez-Almazor M. Functional dependence after hip fracture. *Am J Phys Med Rehabil* **80**, 736-743 (2001).

4. Nazrun AS, Tzar MN, Mokhtar SA, Mohamed IN. A systematic review of the outcomes of osteoporotic fracture patients after hospital discharge: morbidity, subsequent fractures, and mortality. *Ther Clin Risk Manag* **10**, 937-948 (2014).

5. Bliuc D, Nguyen ND, Milch VE, Nguyen TV, Eisman JA, Center JR. Mortality risk associated with low-trauma osteoporotic fracture and subsequent fracture in men and women. *JAMA* **301**, 513-521 (2009).

6. Saito T, Sterbenz JM, Malay S, Zhong L, MacEachern MP, Chung KC. Effectiveness of anti-osteoporotic drugs to prevent secondary fragility fractures: systematic review and meta-analysis. *Osteoporos Int* **28**, 3289-3300 (2017).

7. Kanis JA, McCloskey EV, Johansson H, Oden A, Strom O, Borgstrom F. Development and use of FRAX in osteoporosis. *Osteoporos Int* **21 Suppl 2**, S407-413 (2010).

8. Compston JE, McClung MR, Leslie WD. Osteoporosis. *Lancet* **393**, 364-376 (2019).

9. Curtis JR*, et al.* Longitudinal trends in use of bone mass measurement among older americans, 1999-2005. *J Bone Miner Res* **23**, 1061-1067 (2008).

10. E. Michael Lewiecki AJS, Pallavi B. Rane , Anne Shah , Allison A. Petrilla , Kris Norris4, Michele McDermott , Shravanthi R. Gandra. Geographic Variation in Prevalence of Osteoporosis Diagnosis and Utilization of Anti-Osteoporosis Therapies in United States Female Medicare Fee-for-Service Beneficiaries With Fragility Fractures. In: *The American Society for Bone and Mineral Research Auunal Meeting*) (2020).

11. Williams S DS, Weiss R, Wang Y, Arora T, Curtis J. Characterization of Older Male Patients with a Fragility Fracture [abstract]. *Arthritis Rheumatol* **72 (supplement 10)**, 1082 (2020).

12. Pickhardt PJ, Pooler BD, Lauder T, del Rio AM, Bruce RJ, Binkley N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Ann Intern Med* **158**, 588-595 (2013).

13. Krishnaraj A*, et al.* Simulating Dual-Energy X-Ray Absorptiometry in CT Using Deep-Learning Segmentation Cascade. *J Am Coll Radiol* **16**, 1473-1479 (2019).

14. Dagan N*, et al.* Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. *Nat Med* **26**, 77-82 (2020).

15. Benhamou CL*, et al.* Fractal analysis of radiographic trabecular bone texture and bone mineral density: two complementary parameters related to osteoporotic fractures. *J Bone Miner Res* **16**, 697-704 (2001).

16. Pothuaud L*, et al.* Fractal analysis of trabecular bone texture on radiographs: discriminant value in postmenopausal osteoporosis. *Osteoporos Int* **8**, 618-625 (1998).

17. Touvier J*, et al.* Fracture discrimination by combined bone mineral density (BMD) and microarchitectural texture analysis. *Calcif Tissue Int* **96**, 274-283 (2015).

18. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* **521**, 436-444 (2015).

19. Lindsey R*, et al.* Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A* **115**, 11591-11596 (2018).

20. Gulshan V*, et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402-2410 (2016).

21. Ardila D*, et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* **25**, 954-961 (2019).

22. Kanis JA*, et al.* Long-term risk of osteoporotic fracture in Malmo. *Osteoporos Int* **11**, 669-674 (2000).

23. Curtis EM*, et al.* Epidemiology of fractures in the United Kingdom 1988-2012: Variation with age, sex, geography, ethnicity and socioeconomic status. *Bone* **87**, 19-26 (2016).

24. Crandall CJ*, et al.* Comparative effectiveness of pharmacologic treatments to prevent fractures: an updated systematic review. *Ann Intern Med* **161**, 711-723 (2014).

25. Li N*, et al.* An Updated Systematic Review of Cost-Effectiveness Analyses of Drugs for Osteoporosis. *Pharmacoeconomics* **39**, 181-209 (2021).

26. Zha XY, Hu Y, Pang XN, Chang GL, Li L. Diagnostic value of osteoporosis self-assessment tool for Asians (OSTA) and quantitative bone ultrasound (QUS) in detecting high-risk populations for osteoporosis among elderly Chinese men. *J Bone Miner Metab* **33**, 230-238 (2015).

27. Siris ES*, et al.* Bone mineral density thresholds for pharmacological intervention to prevent fractures. *Arch Intern Med* **164**, 1108-1112 (2004).

28. Siris ES*, et al.* The effect of age and bone mineral density on the absolute, excess, and relative risk of fracture in postmenopausal women aged 50-99: results from the National Osteoporosis Risk Assessment (NORA). *Osteoporos Int* **17**, 565-574 (2006).

29. Cheng CT*, et al.* A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. *Nat Commun* **12**, 1066 (2021).

30. Guglielmi G*, et al.* Vertebral morphometry: current methods and recent advances. *Eur Radiol* **18**, 1484-1496 (2008).

31. Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. *J Bone Miner Res* **8**, 1137-1148 (1993).

32. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. *Eur Radiol* **30**, 3549-3557 (2020).

33. Fang Y*, et al.* Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. *Eur Radiol*, (2020).

34. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med* **98**, 8-15 (2018).

35. Kavitha MS, Asano A, Taguchi A, Kurita T, Sanada M. Diagnosis of osteoporosis from dental panoramic radiographs using the support vector machine method in a computer-aided system. *BMC Med Imaging* **12**, 1 (2012).

36. Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J. Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs. *J Clin Med* **9**, (2020).

37. Sapthagirivasan V, Anburajan M. Diagnosis of osteoporosis by extraction of trabecular features from hip radiographs using support vector machine: an investigation panorama with DXA. *Comput Biol Med* **43**, 1910-1919 (2013).

38. Ferizi U*, et al.* Artificial Intelligence Applied to Osteoporosis: A Performance

Comparison of Machine Learning Algorithms in Predicting Fragility Fractures From MRI Data. *J Magn Reson Imaging* **49**, 1029-1038 (2019).

39. Kanis JA, McCloskey EV, Johansson H, Oden A, Melton LJ, 3rd, Khaltaev N. A reference standard for the description of osteoporosis. *Bone* **42**, 467-475 (2008).

40. Binkley N*, et al.* Recalculation of the NHANES database SD improves T-score agreement and reduces osteoporosis prevalence. *J Bone Miner Res* **20**, 195-201 (2005).

41. Li W, Liao H, Miao S, Lu L, Luo J. Unsupervised Learning of Landmarks based on Inter-Intra Subject Consistencies. *arXiv preprint arXiv:200407936 (2020)*, (2020).

42. Health Promotion Administration MoHaW, Taiwan. Taiwan Osteoporosis Practice Guidelines.) (2018).

43. Dawson-Hughes B*, et al.* Implications of absolute fracture risk assessment for osteoporosis practice guidelines in the USA. *Osteoporos Int* **19**, 449-458 (2008).

Figure legend

Figure 1. Schematic representation of the workflow for hip and spine BMD estimation.

Figure 2. Flowchart of the study population.

Figure 3. The calibration plots for (a) 1693 pairs of predicted-measured hip BMD (1693 patients) and (b) 33299 pairs of predicted-measured lumbar vertebral BMD (11908 patients) (b). Each point represents a data pair of predicted and measure BMD. The points close to the diagonal line suggests good calibration. Histograms of the predicted and measure BMD were exhibited on the side of the relevant axis.

Table 1. Patient characteristics

	Hip testing set			Spine testing set		
	Training	Testing	p	Training	testing	p
Number	1602	1693		5000	11908	
Female, n (%)	1262 (78.8)	1313 (77.6)	0.42	3887 (77.7)	9387 (78.8)	0.12
Mean age (sd), years	70.3 (10.9)	69.6 (11.8)	0.06	66.4 (12.2)	66.6 (10.8)	0.17
Median time (IQR) between DXA and radiographs	17 (5, 51)	17 (2, 56)	0.11	16 (5, 45)	15 (5, 43)	0.27
Mean BMI (sd), kg/m^2	23.9 (3.9)	24.2 (4.2)	0.53	24.3 (3.8)*	24.3 (3.8)*	0.39
Mean BMD (g/cm^2)	0.683 (0.159)	0.692 (0.154)	0.11	0.776 (0.185)*	0.839 (0.173)*	<0.001
median T-score (IQR)	-1.5 (-2.3, -0.6)	-1.4 (-2.2, -0.6)	0.10	-2.4 (-3.4, -1.3)*	-2.4 (-3.3, -1.5)*	0.68
Osteoporosis, n (%)	357 (22.3)	346 (20.4)	0.20	2459 (49.2)*	5747 (48.3)*	0.28

* Calculated per eligible vertebrae

Table 2. Summary of performance matrices of predictive model for BMD

Patient strata	Number of ROIs	Predicted vs. measured mean BMD (g/cm^2; sd); p	Correlation coefficient	Linear regression R^2, RMSE	Calibration slop, CITL	Bland-Altman bias (g/cm^2; sd)
The hip testing set						
Overall	1693	0.691 (0.144) vs. 0.692 (0.154); p = 0.35	0.93	0.87, 0.056	0.998, 0.001	-0.001 (0.056)
Female	1313	0.664 (0.133) vs. 0.665 (0.145); p = 0.76	0.92	0.85, 0.056	1.005, 0.000	-0.001 (0.056)
Male	380	0.781 (0.143) vs. 0.785 (0.147); p = 0.17	0.93	0.86, 0.056	0.962, 0.004	-0.004 (0.056)
40-59 years	403	0.781 (0.129) vs 0.778 (0.146); p = 0.33	0.91	0.83, 0.061	1.027, -0.003	0.003 (0.061)
60-74 years	682	0.699 (0.134) vs. 0.700 (0.146); p = 0.61	0.93	0.86, 0.055	1.006, 0.001	0.001 (0.055)
75-90 years	608	0.622 (0.127) vs. 0.626 (0.137); p = 0.06	0.92	0.84, 0.045	0.996, 0.004	-0.004 (0.054)
The spine testing set*						
Overall	33299	0.846 (0.172) vs. 0.852 (0.191); p < 0.001	0.92	0.86, 0.065	1.021, 0.005	-0.005 (0.075)
Female	26167	0.820 (0.163) vs. 0.826 (0.181); p < 0.001	0.92	0.85, 0.063	1.020, 0.005	-0.005 (0.072)
Male	7132	0.943 (0.171) vs. 0.948 (0.196); p < 0.001	0.91	0.82, 0.072	1.035, 0.004	-0.004 (0.083)
40-59 years	11269	0.909 (0.158) vs. 0.910 (0.174); p = 0.21	0.91	0.83, 0.065	1.003, 0.001	-0.001 (0.072)

60-74 years	15214	0.824 (0.168) vs. 0.829 (0.189); p <0.001	0.92	0.86, 0.064	1.036, 0.005	-0.005 (0.074)
75-90 years	6816	0.793 (0.174) vs. 0.805 (0.199); p < 0.001	0.92	0.85, 0.067	1.049, 0.012	--0.012 (0.080)
L1	6435	0.745 (0.153) vs. 0.746 (0.173); p = 0.37	0.90	0.82, 0.065	1.015, 0.001	-0.001 (0.076)
L2	8611	0.824 (0.163) vs. 0.828 (0.180); p <0.001	0.92	0.85, 0.064	1.012, 0.004	-0.004 (0.072)
L3	9079	0.878 (0.172) vs. 0.882 (0.181); p <0.001	0.92	0.85, 0.063	1.016, 0.004	-0.004 (0.073)
L4	9174	0.908 (0.165) vs. 0.918 (0.186); p <0.001	0.91	0.84, 0.066	1.026, 0.010	-0.010 (0.077)

* Calculated per eligible vertebrae

Table 3 Discriminatory performance (%) of the predicted BMD to classify osteoporosis, low bone mass and osteoporosis, and high-risk group for major osteoporotic or hip fractures.

Discriminatory measures	Hip osteoporosis (T-score <= -2.5	10-year risk of major osteoporotic fracture >=20%	10-year risk of hip fracture >=3%	Lumbar vertebral osteoporosis (vertebrae with the lowest T-score<= 2.5)
Number of patients, %	346, 20.4	372, 22.0	812, 48.0	5747, 48.3
OR (95% CI)	76.83 (52.23–113.00)	147.68 (94.09–231.79)	85.98 (59.04–125.20)	40.56 (36.32–45.30)
AUC (95% CI)	0.89 (0.87–0.91)	0.92 (0.90–0.94)	0.92 (0.91–0.94)	0.87 (0.86–0.87)
Accuracy (%; 95% CI)	92.4 (91.4–93.9)	94.6 (93.4–95.6)	92.2 (90.8–93.4)	86.8 (86.2–87.4)
Sensitivity (%; 95% CI)	80.8 (76.2–84.8)	87.2 (83.4–91.4)	91.3 (89.2–93.2)	83.1 (82.1–84.1)
Specificity (%; 95% CI)	95.4 (94.2–96.5)	96.7 (95.6–97.6)	93.0 (91.1–94.6)	90.3 (89.5–91.0
PPV (%; 95% CI)	81.7 (77.7–85.1)	88.1 (84.7–90.9)	92.5 (90.6–94.0)	88.8 (84.4–85.7)
NPV (%; 95% CI)	95.1 (94.0–96.0)	96.4 (93.5–95.7)	92.0 (90.2–93.5)	85.2 (84.4–85.9)

Table 4 Network performance across age and sex subsets

Discriminatory measures	Hip osteoporosis		10-year risk of major osteoporotic fracture >=20%		10-year risk of hip fracture >=3%		Lumbar vertebral osteoporosis	
	AUC (95% CI)	OR (95% CI)	AUC (95% CI)	OR (95% CI)	AUC (95% CI)	OR (95% CI)	AUC (95% CI)	OR (95% CI)
Female	0.88 (0.86–0.90)	60.21 (40.29–89.97)	0.92 (0.90–0.94)	137.32 (96.21–218.73)	0.92 (0.91–0.94)	86.23 (56.25–132.19)	0.87 (0.86–0.88)	41.79 (36.87–47.36)
Male	0.95 (0.90–1.00)	888.82 (169.84–4651.29)	0.90 (0.77–0.99)	1773.75 (88.87–35401.73)	0.92 (0.88–0.95)	89.45 (39.36–203.29)	0.86 (0.84–0.87)	34.92 (27.55–44.27)
Age: <60 years	0.94 (0.88–1.00)	607.50 (100.37–3676.92)	0.80 (0.69–0.93)	446.22 (47.94–4153.14)	0.86 (0.80–0.93)	399.97 (84.76–1887.41)	0.85 (0.83–0.86)	36.57 (29.28–45.67)
Age: 60-74 years	0.86 (0.82–0.90)	61.68 (34.12–111.52)	0.92 (0.90–0.95)	249.75 (116.11–537.23)	0.90 (0.87–0.92)	57.86 (35.22–95.04)	0.87 (0.86–0.88)	41.36 (35.24–48.53)
Age: 75-90 years	0.90 (0.86–0.92)	71.40 (40.82–124.89)	0.91 (0.89–0.94)	94.75 (50.44–178.00)	0.88 (0.85–0.92)	108.73 (54.91–215.30)	0.86 (0.85–0.88)	41.25 (33.33–51.07)