Vertebral fractures identified on lateral DXA images by deep learning predict incident fractures in older women

Osteoporosis International
https://doi.org/10.1007/s00198-026-08072-9

ORIGINAL ARTICLE

Vertebral fractures identified on lateral DXA images by deep learning predict incident fractures in older women

Mattias Lorentzon[1,2] · Victor Wåhlstrand[3] · Jennifer Alvén[3,4] · Ida Häggström[3,5] · Lisa Johansson[1,6]

Received: 13 January 2026 / Accepted: 2 May 2026

Abstract

Summary XVFA is an AI-based method for identifying vertebral fractures on DXA images. In 423 women followed for 8 years, vertebral fractures identified by XVFA or manual assessment were associated with a twofold increased risk of incident fractures. XVFA predicted fracture risk comparably to manual assessment, supporting automated vertebral fracture detection.

Purpose Vertebral fractures (VFs), identified by vertebral fracture assessment (VFA) using dual-energy X-ray absorptiometry (DXA), predict incident fractures independently of clinical risk factors (CRFs) and bone mineral density (BMD). Most VFs remain clinically unrecognized. This study evaluated whether VFs identified using a deep learning–based method on lateral DXA images predict incident fractures comparably to manual VFA.

Methods Associations between prevalent VFs and incident fractures were investigated in 423 women from the population-based SUPERB study who were not included in development of the explainable deep learning model (XVFA). Vertebrae were classified by manual VFA and XVFA. Incident fractures were X-ray verified. Cox proportional hazards models assessed fracture risk adjusted for CRFs and femoral neck (FN) BMD.

Results Manual VFA reading and XVFA were used on baseline lateral images and classified 4563 and 5532 vertebrae, respectively, with numerical differences partly reflecting image quality limitations. VFs were identified in 102 women by manual VFA and 187 by XVFA. During 8 years of follow-up, incident fractures occurred in 48% of women with manual VFA VFs and 43% with XVFA VFs, vs 20% and 16% of women without VFs. Women with VFs had a higher fracture risk whether identified manually (HR 2.04; 95% CI, 1.35–3.07) or by XVFA (HR 2.32; 95% CI, 1.55–3.48), compared with women without VFs. Results remained significant after adjustment for CRFs and FN BMD.

Conclusion Automated XVFA predicted incident fractures similarly to manual assessment. These findings support the clinical utility of deep learning–based VF detection, which may enhance fracture risk assessment and management in routine practice.

Keywords Deep neural networks · Dual-energy X-ray absorptiometry · Incident fracture · Older women · Vertebral fracture · Vertebral fracture assessment

✉ Lisa Johansson
lisa.s.johansson@gu.se

[1] Sahlgrenska Osteoporosis Centre, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Mölndal, Sweden

[2] Department of Internal Medicine, Geriatrics and Emergency Medicine, Sahlgrenska University Hospital, Mölndal, Sweden

[3] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

[4] Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[5] Department of Medical Radiation Sciences, University of Gothenburg, Gothenburg, Sweden

[6] Department of Orthopedics, Sahlgrenska University Hospital, Mölndal, Region Västra Götaland, Sweden

Published online: 12 May 2026

Introduction

Vertebral fracture (VF) is a strong predictor of new fractures. Both symptomatic and asymptomatic VFs increase the risk of subsequent fractures and mortality [1, 2]. As fewer than one-third of patients with VFs come to clinical attention, methods to identify VFs are needed [3]. Vertebral fracture assessment (VFA) detects VFs using lateral spine images from dual-energy X-ray absorptiometry (DXA) and identifies both symptomatic and asymptomatic VFs [4]. Although effective and inexpensive fracture prevention treatments exist, the treatment gap among high-risk individuals remains substantial and is increasing. Underdiagnosis of osteoporotic VFs is a recognized global problem [5, 6]. Even among experts, diagnosing VFs on conventional radiographs shows low interrater reliability [7], and strict quantitative criteria can lead to overdiagnosis [8], highlighting the need for objective, automated approaches. VFA is available in most DXA clinics but requires specialized expertise [9]. Artificial intelligence (AI) enables machines to perform tasks requiring human cognition. Machine learning (ML) learns patterns from data, while deep learning (DL), a branch of ML, uses multilayer neural networks to model complex imaging and clinical relationships [10]. AI has shown human-level performance in X-ray analysis [11]. Automated VF detection was first applied to CT [12], but CT is not first-line due to radiation and cost.

In a retrospective study using spinal radiographs, an ML algorithm was trained and validated on 300 patients, including 150 with VFs and 150 without [13]. The model performed as well as orthopedic surgeons, achieving an accuracy of 86.0%, a sensitivity of 84.7%, and a specificity of 87.3%. In another retrospective study, a DL classifier for osteoporotic compression fractures was developed using 15,524 spine radiographs from 4461 men [14]. VFs were categorized as moderate-to-severe versus normal/mild using an adaptation of Genant's semiquantitative method (GSQ), yielding an area under the receiver operating characteristic curve (AUC-ROC) of 0.99. We recently developed a deep-learning method, *Xplainable VFA* (XVFA), using 11,605 annotated vertebrae from the population-based cross-sectional **S**ahlgrenska **U**niversity **P**rospective **E**valuation of **R**isk of **B**one fractures (SUPERB) cohort [15, 16], achieving an AUC of 0.97 and 90% sensitivity [17]. VFs identified by VFA predict incident major osteoporotic fractures (MOF) independently of clinical risk factors (CRFs) and bone mineral density (BMD) and improve fracture prediction beyond FRAX estimates [18].

This study aimed to assess whether VFs detected on lateral DXA images by XVFA predict incident fractures as well as manual VFA, by evaluating XVFA performance in the 423 SUPERB participants who were not included in the model development.

Subjects and methods

Subjects

The SUPERB study is a population-based prospective cohort of 3028 women aged 75–80 years from the Gothenburg area, recruited between March 2013 and May 2016. Baseline assessments were cross-sectional, and participants were followed prospectively for incident fractures [19]. At baseline, 2923 women had analyzable lateral DXA spine images. XVFA was developed using 11,605 annotated vertebrae on VFAs from 2919 participants, of whom 2481 (85%) were used for training and 438 (15%) as an independent test set [17]. Of the 438 women in the test set, 423 had a manually analyzable VFA at baseline and were therefore included in the present study, allowing for an unbiased comparison between XVFA and manual assessment. Participants underwent DXA scanning, questionnaires, and anthropometric measurements at baseline, and fractures and mortality were ascertained through X-ray review and linkage with the *Västfolket* registry. Each participant was followed until the date when research nurses reviewed their X-ray reports, which occurred between November 2022 and March 2023. All participants provided informed consent, and the study was approved by the regional ethics committee.

DXA

DXA was used to assess BMD (g/cm^2) at the femoral neck (FN) and lumbar spine (LS; L1–L4), as well as trabecular bone score (TBS) at the same spinal levels (Discovery A; Hologic, Waltham, MA, USA). Lumbar spine BMD and TBS were calculated as the average of at least two evaluable vertebrae between L1 and L4, excluding any vertebrae affected by fracture or containing osteosynthesis material. The coefficients of variation (CV) for FN BMD, LS BMD, and TBS were 1.3%, 0.7%, and 2.12%, respectively.

Questionnaires

At baseline, information on lifestyle factors, medical history, medication use, family history of hip fracture, and previous fractures was obtained from all participants through a self-administered questionnaire. The 10-year fracture probability of hip or major osteoporotic fracture (MOF; spine, hip, proximal humerus, or distal forearm) was estimated using the FRAX Sweden model, based on age, sex, body mass index (BMI), and CRFs. These CRFs included previous fragility

fracture, parental hip fracture, current smoking, excessive alcohol intake, systemic glucocorticoid use, rheumatoid arthritis, and other causes of secondary osteoporosis. Fracture probability was calculated with or without FN BMD [20].

VFA

Baseline lateral spine DXA images (supine position) were obtained using *Physician's Viewer* (Hologic) to identify prevalent VFs, as described previously [19]. Vertebrae T4–L4 were manually annotated at six landmarks [21], and VFs were classified by the GSQ method as mild (grade 1, 20–25% height reduction), moderate (grade 2, 25–40%), or severe (grade 3, > 40%) [22]. Non-fracture vertebral changes (e.g., Schmorl's nodes, Scheuermann's disease, scoliosis, degenerative changes) were distinguished from true VFs. Whole-body and antero-posterior lumbar spine images were used to assess the presence of scoliosis, in order to avoid misclassification of biconcave VFs [23]. Both VFA evaluators (physicians) were blinded to participants' fracture status. Reproducibility in 50 women showed intraobserver agreement of 98.9% and 97.8% (kappa 0.85 and 0.67, respectively) and interobserver agreement of 98.6% (kappa 0.77) [19]. For severity analyses, only the most severe VF per participant was reported.

Explainable VFA (XVFA)

XVFA is a novel deep learning method for VFA in lateral spine images acquired using DXA. Details of the method have been published previously [17]. In summary, it employs deep neural networks and incorporates vertebra detection and localization of six landmarks with associated uncertainty estimates [17]. XVFA was trained on heterogeneously annotated data and employs a state-of-the-art landmark detection algorithm. Identified vertebrae are classified using a scheme based on the clinical GSQ compression criteria, providing greater explainability compared to a standard black-box classifier [24]. The model also computes the uncertainty of its landmark annotations, which can be propagated to uncertainty in fracture classification, helping to identify cases where predictions are less reliable. Specifically, XVFA regresses on the GSQ landmarks and calculates posterior, middle, and anterior ridge heights, mirroring human annotations. The GSQ compression criteria are implemented as a differentiable decision tree, enabling the model to base classifications on clinically meaningful features. This approach allows researchers and clinicians to trace the model's decision back to specific vertebral measurements and clinically interpretable criteria, improving transparency and facilitating interpretation of the results. The XVFA was developed and tested using VFA images from 2919 women in the SUPERB cohort, comprising 11,605 annotated vertebrae. The relatively low proportion of annotated vertebrae in the individuals used to develop XVFA can be explained by two factors. First, during manual review of the baseline VFA images, the two assessors initially annotated all vertebrae, including normal ones. After reviewing approximately half of the cohort, they realized that annotating clearly normal vertebrae was extremely time-consuming and unnecessary. Consequently, normal vertebrae were subsequently assessed visually according to the GSQ method, resulting in some normal vertebrae lacking annotations; however, all vertebrae with fractures were annotated. Second, vertebrae that could not be analyzed due to poor image quality were also left unannotated. Randomly, 438 (15%) participants were held out for final testing, while the remaining 2481 (85%) were divided into five folds for cross-validation during model training and tuning. The criteria from the GSQ method were used as a differentiable, rule-based approach on the six predicted landmarks, enabling classification of both VF severity and morphological characteristics. The XVFA model was trained on VFA images and was not exposed to outcome data. The training set consisted of vertebrae previously evaluated independently by two physicians. Each vertebra was classified as either having non-fracture vertebral changes or no such changes. This approach ensured that the XVFA model learned to distinguish true VFs from other vertebral abnormalities. On a vertebra-level evaluation, XVFA achieved a sensitivity of 90% (mild + normal vs. moderate + severe), a specificity of 96%, an F1-score of 65%, and a positive predictive value (PPV) of approximately 51%. Given the relatively low prevalence of VFs in the dataset, this PPV is consistent with expectations, while the high sensitivity indicates that the model successfully identifies most true VFs. The end-to-end AUC of 0.97 further demonstrates strong overall discriminative performance. One of the annotators (LJ) reanalyzed a random subset of 49 scans from the test set. The annotator had a mean intra-reader deviation of approximately ± 4.5 pixels, while the XVFA demonstrated an uncertainty of ± 5.3 pixels between the 5th and 95th percentiles, indicating a level of reliability comparable to that of human observers.

Incident fractures

Incident fractures were verified through radiographic imaging and identified using a regional radiology archive encompassing all 49 municipalities in the Västra Götaland region, covering approximately 25,000 square kilometers around Gothenburg. Five research nurses systematically reviewed all radiology reports issued from baseline through to the date of review, which fell between November 2022 and March 2023. All confirmed fractures were documented. In cases where radiology reports were unavailable or the diagnosis

was uncertain, the corresponding radiographs were manually reviewed by an orthopedic surgeon (LJ). VFs mentioned in radiology reports were compared with baseline VFA data to determine whether they were pre-existing or incident. Only fractures occurring after baseline were classified as incident.

Statistics

Baseline continuous variables were compared between women with and without VF using independent-samples *t*-tests and are presented as means ± SD; categorical variables were analyzed using χ^2 tests. A *p*-value < 0.05 was considered statistically significant. Associations between VF and fracture risk were assessed using Cox proportional hazards models with each participant's follow-up time included in the Cox model. Analyses were performed for five outcomes: any fracture, MOF (fracture of the hip, spine, forearm, or proximal humerus), hip fracture, VF, and death. Three models with increasing adjustment were used: Model 1 was adjusted for age, height, and weight; Model 2 additionally included FRAX CRFs (previous fracture, parental hip fracture, smoking, glucocorticoid use, rheumatoid arthritis, high alcohol intake, and secondary osteoporosis (including type 1 or type 2 diabetes, menopause before age 45, inflammatory bowel disease, and chronic liver disease)); and Model 3 further included FN BMD. Post hoc power analyses showed that most VF categories had > 80% power (α = 0.05) to predict incident fractures for both manually identified and XVFA-identified VFs. Proportional hazards assumptions were verified graphically. Results are presented as hazard ratios (HRs) with 95% confidence intervals (CIs). Time-to-event data were also analyzed using Kaplan–Meier curves, with participants stratified by baseline VF status (manual and XVFA) and compared using the log-rank test. Statistical analyses were performed using IBM SPSS Statistics version 25.

Results

Characteristics of the cohort

Baseline characteristics of women with or without any VF, according to the method used (manual VFA or XVFA), are presented in Table 1. In total, 423 women were included in the analysis. When VFAs were annotated manually or by XVFA, 102 (24%) and 187 (44%) women, respectively, were identified as having any VF. In both methods- manual and XVFA- women with any VF were older, had lower BMD by DXA, and higher FRAX 10-year probabilities for MOF and hip fracture compared with women without VF ($p < 0.02$). The proportions of women reporting prior fracture, osteoporosis, and use of osteoporosis medication were also higher among those with VF compared with those without.

The only differences between the two methods were that women with VFs identified by XVFA were shorter and more often had a family history of hip fracture, whereas women with VFs identified by the manual method had a higher prevalence of glaucoma compared with women without VFs ($p < 0.05$).

Associations between vertebral fractures identified by manual VFA and the risk of incident fractures

Out of 423 women, 321 (76%) had no VFs identified by VFA. Among the 102 women (24%) with any VF, 32 (8%) had a grade 1 VF, 43 (10%) had a grade 2 VF (with or without grade 1), and 27 (6%) had a grade 3 VF (with or without grade 1 or grade 2). In total, 71 women (17%) had one VF, 19 (4%) had two VFs, and 12 (3%) had three or more VFs (Tables 2 and 3). Incident fractures were categorized into four groups: any fracture, MOF, VF, and hip fracture. Associations between prevalent VFs and first incident fracture are presented in Tables 2 and 3. During a median follow-up of 7.9 years (interquartile range [IQR] 7.2–8.9) 72 women died, 138 women experienced any fracture, 102 a MOF, 44 a VF, and 30 a hip fracture. When prevalent VFs were stratified by severity, the proportion of women with any incident fracture was 47% ($n = 15$), 44% ($n = 19$), and 56% ($n = 15$) among those with grade 1, grade 2, and grade 3 VFs, respectively (Table 2). When stratified by number of prevalent VFs, the corresponding proportions were 48% ($n = 34$), 58% ($n = 11$), and 33% ($n = 4$) among women with one, two, or three or more VFs, respectively (Table 3). A similar trend was observed for incident MOF: 28% ($n = 9$), 37% ($n = 16$), and 48% ($n = 13$) among women with grade 1, grade 2, and grade 3 VFs, respectively, and 34% ($n = 24$), 53% ($n = 10$), and 33% ($n = 4$) among those with one, two, or three or more VFs, respectively. Comparable trends were also seen for incident VF and hip fracture. Incidences of fractures and death among women with any VF are presented in Supplement information 1.

A Cox regression model adjusted for age, height, and weight revealed that having any VF identified by manual VFA was associated with an increased risk of sustaining any fracture (HR 1.99 [95% CI 1.39–2.83]), MOF (HR 2.04 [95% CI 1.35–3.07]), and VF (HR 5.30 [95% CI 2.85–9.88]). These associations were independent of CRFs included in FRAX and FN BMD (Fig. 1 and Supplement information 1). The risk of incident hip fracture was not significant (HR 1.18 [95% CI 0.54–2.60]) (Fig. 2 and Supplement information 1). When VFs were stratified by severity, having a grade 3 VF was associated with nearly a threefold increased risk of any fracture (HR 2.81 [95% CI 1.61–4.90]), more than a threefold increased risk of MOF (HR 3.41 [95% CI 1.86–6.26]),

Table 1 Baseline characteristics of older women with and without any VFA-identified vertebral fracture, assessed by manual VFA or by deep learning (XVFA)

	VF identified by manual VFA			VF identified by XVFA		
	No VF $n=321$	Any VF$n=102$	p value[a,b]	No VF $n=236$	Any VF $n=187$	p value[a,b]
Age (years)	77.6 ± 1.6	78.3 ± 1.6	**< 0.001**	77.6 ± 1.5	78.0 ± 1.7	**0.007**
Height (cm)	161.8 ± 5.6	161.3 ± 6.2	0.459	162.3 ± 5.6	161.0 ± 5.9	**0.024**
Weight (kg)	67.8 ± 11.8	67.3 ± 12.2	0.745	68.0 ± 11.3	67.2 ± 12.5	0.510
Body mass index (kg/m^2)	25.8 ± 4.2	25.8 ± 4.2	0.923	25.8 ± 4.0	25.9 ± 4.4	0.815
Femoral neck BMD (g/cm^2)	0.67[i] ± 0.11	0.64[j] ± 0.10	**0.005**	0.67 ± 0.11[n]	0.65 ± 0.11[o]	**0.043**
Lumbar spine BMD (g/cm^2)	0.94 ± 0.18	0.91[k] ± 0.16	0.06	0.94 ± 0.17	0.92 ± 0.18[p]	0.182
TBS	1.21 ± 0.11	1.18[k] ± 0.10	**0.037**	1.21 ± 0.11	1.20 ± 0.11[p]	0.411
FRAX MOF w/o BMD (%)	32.5 ± 13.2	36.1 ± 13.8	**0.021**	31.7 ± 13.1	35.5 ± 13.6	**0.004**
FRAX hip fracture w/o BMD (%)	20.0 ± 13.9	22.9 ± 14.3	0.074	19.3 ± 13.8	22.5 ± 14.1	**0.018**
FRAX MOF with BMD (%)	21.4 ± 11.0	26.5 ± 12.3	**< 0.001**	21.0 ± 11.1	24.6 ± 11.8	**0.002**
FRAX hip fracture with BMD (%)	10.0 ± 10.2	13.7 ± 11.7	**0.002**	9.8 ± 10.5	12.3 ± 10.8	**0.015**
Fall accident within the last year, % (n)	28.7 (92)	30.4 (31)	0.737	28.0 (66)	30.5 (57)	0.572
Self-reported prior fracture, % (n)[c]	27.7 (89)	44.1 (45)	**0.002**	27.1 (64)	37.4 (70)	**0.024**
Family history of hip fracture, % (n)	17.8 (57)	22.5 (23)	0.282	15.3 (36)	23.5 (44)	**0.031**
Current smoking, % (n)	6.5 (21)	3.9 (4)	0.328	7.2 (17)	4.3 (8)	0.205
Excessive alcohol consumption, % (n)[d]	0.6 (2)	2.0 (2)	0.246[q]	0.4 (1)	1.6 (3)	0.326[q]
Secondary osteoporosis, % (n)[e]	27.4 (88)	18.6 (19)	0.075	26.3 (62)	24.1 (45)	0.604
Medications						
Glucocorticoid use, % (n)[f]	3.1 (10)	4.9 (5)	0.370[q]	2.1 (5)	5.3 (10)	0.075
Osteoporosis medication, % (n)[g]	7.2 (22)[l]	21.7 (20)[m]	**< 0.001**	6.6 (15)	15.9 (27)	**0.003**
Medical history						
Rheumatoid arthritis, % (n)	3.7 (12)	3.9 (4)	1.000[q]	4.7 (11)	2.7 (5)	0.287
Hyperthyroidism, % (n)	5.6 (18)	4.9 (5)	0.779	4.3 (10)	7.0 (13)	0.225
Osteoporosis, % (n)[h]	11.8 (38)	34.3 (35)	**< 0.001**	10.2 (24)	26.2 (49)	**< 0.001**
Hypertension, % (n)	55.1 (177)	57.8 (59)	0.632	56.8 (134)	54.5 (102)	0.646
Stroke, % (n)	8.7 (28)	7.8 (8)	0.782	9.7 (23)	7.0 (13)	0.306
Myocardial infarction, % (n)	4.0 (13)	3.9 (4)	1.000[q]	4.7 (11)	3.2 (6)	0.450
Angina, % (n)	5.0 (16)	4.9 (5)	0.973	4.2 (10)	5.9 (11)	0.439
Heart failure, % (n)	7.5 (24)	10.8 (11)	0.295	8.5 (20)	8.0 (15)	0.856
Diabetes, % (n)	10.6 (34)	6.9 (7)	0.267	11.4 (27)	7.5 (14)	0.172
Chronic bronchitis, asthma, emphysema, % (n)	8.7 (28)	12.7 (13)	0.232	9.3 (22)	10.2 (19)	0.772
Cancer, % (n)	16.8 (54)	20.6 (21)	0.386	19.5 (46)	15.5 (29)	0.287
Glaucoma, % (n)	10.3 (33)	3.9 (4)	**0.048**	11.0 (26)	5.9 (11)	0.063

Values are presented as mean ± standard deviation for continuous variables and as percentage and number for categorical variables. *VFA*, vertebral fracture assessment; *VF*, vertebral fracture; *XVFA*, explainable VFA (a model developed using deep learning); *BMD*, bone mineral density; *TBS*, trabecular bone score. Significance was defined by a p-value < 0.05 and significant values are presented in bold. [a] Independent samples t test for continuous variables, [b] Categorical variables χ^2 test, [c] After 50 years of age, fractures of the skull and face are excluded, [d] ≥ 21 units per week, [e] diabetes (type 1 or type 2), menopause before 45 years of age, inflammatory bowel disease, chronic kidney disease, [f] Daily oral treatment of ≥ 5 mg for ≥ 3 months ever during lifetime, [g] Current treatment with bisphosphonates, teriparatide or denosumab, [h] Self-reported from the question "Has a doctor told you that you have osteoporosis?" [i] 318, [j] 101, [k] 99, [l]304, [m]92, [n]233, [o]186, [p]184,[q] Fisher's exact test

and almost a tenfold increased risk of VF (HR 9.66 [95% CI 4.46–20.94]). Stratification by number of VFs showed that having two VFs conferred an increased risk of any fracture (HR 2.98 [95% CI 1.58–5.61]), MOF (HR 3.60 [95% CI 1.83–7.10]), and VF (HR 9.56 [95% CI 4.05–22.58]). These associations remained significant after adjustment for CRFs and FN BMD (Fig. 2, Tables 2 and 3).

Associations between vertebral fractures identified by XVFA and the risk of incident fracture

Out of 423 women, 236 (56%) did not have any VFA-identified VF. Of the 187 (44%) women having any VF, 83 (20%) had a grade 1 VF, 74 (17%) had a grade 2 VF (with or without grade 1), and 30 (7%) had a grade 3 VF (with or without

Table 2 Vertebral fracture severity and associations with subsequent fracture risk: manual VFA vs deep learning-based XVFA in older women

	Manual VFA				XVFA			
	No VF $n=321$	Grade 1 VF $n=32$	Grade 2 VF $n=43$	Grade 3 VF $n=27$	No VF $n=236$	Grade 1 VF $n=83$	Grade 2 VF $n=74$	Grade 3 VF $n=30$
Any fracture								
No. (%)	89 (27.7)	15 (46.9)	19 (44.2)	15 (55.6)	57 (24.2)	38 (45.8)	29 (39.2)	14 (46.7)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	**1.81 [1.04–3.16]**	**1.72 [1.04–2.84]**	**2.81 [1.61–4.90]**	1 [Reference]	**2.18 [1.44–3.30]**	**1.75 [1.12–2.75]**	**2.39 [1.32–4.31]**
+clinical risk factors	1 [Reference]	1.73 [0.99–3.0]	**1.77 [1.07–2.95]**	**2.32 [1.30–4.15]**	1 [Reference]	**2.38 [1.56–3.64]**	1.58 [0.98–2.53]	**2.59 [1.42—4.74]**
+FN BMD	1 [Reference]	1.48 [0.85–2.60]	1.61 [0.97–2.66]	**2.22 [1.25–3.96]**	1 [Reference]	**2.36 [1.55–3.60]**	1.49 [0.93–2.39]	**2.44 [1.34–4.44]**
MOF								
No. (%)	64 (19.9)	9 (28.1)	16 (37.2)	13 (48.1)	38 (16.1)	30 (36.1)	21 (28.4)	13 (43.3)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	1.31 [0.65–2.67]	**2.00 [1.15–3.48]**	**3.41 [1.86–6.26]**	1 [Reference]	**2.47 [1.53–4 00]**	**1.82 [1.07–3.12]**	**3.31 [1.75–6.26]**
+clinical risk factors	1 [Reference]	1.28 [0.63–2.63]	**2.07 [1.18–3.64]**	**2.89 [1.54–5.43]**	1 [Reference]	**2.67 [1.63–4.37]**	1.61 [0.92–2.83]	**3.33 [1.74–6.39]**
+FN BMD	1 [Reference]	1.05 [0.51–2.15]	**1.91 [1.09–3.34]**	**2.75 [1.47–5.14]**	1 [Reference]	**2.66 [1.63–4.33]**	1.41 [0.80–2.49]	**3.05 [1.59–5.84]**
VF								
No. (%)	17 (5.3)	4 (12.5)	12 (27.9)	11 (40.7)	10 (4.2)	13 (15.7)	11 (14.9)	10 (33.3)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	2.22 [0.73–6.70]	**5.51 [2.59–11.70]**	**9.66 [4.46–20.94]**	1 [Reference]	**3.52 [1.53–8 08]**	**3.43 [1.44–8 12]**	**8.55 [3.53–20.75]**
+clinical risk factors	1 [Reference]	2.07 [0.68–6.26]	**4.99 [2.33–10.72]**	**7.27 [3.20–16.56]**	1 [Reference]	**3.31 [1.42–7.71]**	**2.83 [1.16–6.92]**	**7.59 [3.03–19.01]**
+FN BMD	1 [Reference]	1.68 [0.56–5.06]	**4.90 [2.28–10.54]**	**7.47 [3.36–16.62]**	1 [Reference]	**3.18 [1.37–7.38]**	**2.57 [1.05–6.31]**	**7.14 [2.90–17.57]**
Hip								
No. (%)	21 (6.5)	2 (6.3)	3 (7.0)	4 (14.8)	9 (3.8)	11 (13.3)	5 (6.8)	5 (16.7)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	0.68 [0.16–2.94]	0.97 [0.29–3.27]	2.51 [0.85–7.43]	1 [Reference]	**3.36 [1.38–8.13]**	1.75 [0.58–5.26]	**4.70 [1.57–14.20]**
+clinical risk factors	1 [Reference]	0.82 [0.19–3.59]	1.05 [0.31–3.62]	1.95 [0.61–6.22]	1 [Reference]	**3.26 [1.33–8.01]**	1.87 [0.59–5.94]	**3.86 [1.23–12.11]**
+FN BMD	1 [Reference]	0.62 [0.14–2.77]	0.95 [0.28–3.24]	1.41 [0.42–4.81]	1 [Reference]	**3.54 [1.42–8.82]**	1.74 [0.54–5.64]	2.36 [0.70–7.91]

Table 2 (Continued)

	Manual VFA				XVFA			
	No VF $n=321$	Grade 1 VF $n=32$	Grade 2 VF $n=43$	Grade 3 VF $n=27$	No VF $n=236$	Grade 1 VF $n=83$	Grade 2 VF $n=74$	Grade 3 VF $n=30$
Death								
No. (%)	46 (14.3)	8 (25.0)	7 (16.3)	11 (40.7)	34 (14.4)	16 (19.3)	15 (20.3)	7 (23.3)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	1.44 [0.67–3.09]	0.99 [0.45–2.21]	**3.03 [1.55–5.93]**	1 [Reference]	1.22 [0.67–2.22]	1.29 [0.70–2.29]	1.46 [0.64–3.32]
+clinical risk factors	1 [Reference]	1.57 [0.72–3.45]	1.14 [0.51–2.55]	**3.51 [1.72–7.17]**	1 [Reference]	1.31 [0.72–2.40]	1.47 [0.78–2.77]	1.51 [0.65–3.48
+FN BMD	1 [Reference]	1.50 [0.68–3.30]	1.07 [0.48–2.41]	**3.37 [1.65–6.90]**	1 [Reference]	1.31 [0.72–2.41]	1.42 [0.75–2.68]	1.43 [0.62–3.29]

Associations were examined using Cox regression model. Hazard Ratios (HR) and 95% confidence intervals (CI) are presented. Model 1: adjusted for age, height and weight. Model 2: adjusted for age, height, weight, and the clinical risk factors used in FRAX (previous fracture, family history of hip fracture, current smoking, oral glucocorticoid use, rheumatoid arthritis, excessive alcohol intake, secondary osteoporosis) except FN BMD. Model 3: adjusted for the same as model 2 with the addition of FN BMD. *FN*, femoral neck; *BMD*, bone mineral density; *MOF*, major osteoporotic fracture; *VF*, vertebral fracture; *VFA*, vertebral fracture assessment; *XVFA*, explainable vertebral fracture assessment (a model developed by using instead of by deep learning)

grade 1 VF or grade 2 VF). A total of 123 women (29%) had one VF, 43 (10%) had two VFs, and 21 (5%) had three or more VFs (Tables 2 and 3).

Associations between prevalent VFs and first incident fractures are presented in Tables 2 and 3. When VFs were categorized by severity, the proportion of women with an incident fracture was 46% ($n=38$), 39% ($n=29$), and 47% ($n=14$) among those with grade 1, grade 2, and grade 3 VFs, respectively (Table 2). When categorized by number, the corresponding proportions were 39% ($n=48$), 49% ($n=21$), and 57% ($n=12$) in women with one, two, or three or more VFs, respectively (Table 3). A similar trend was observed for incident MOFs: 36% ($n=30$), 28% ($n=21$), and 43% ($n=13$) in women with grade 1, grade 2, and grade 3 VFs, and 29% ($n=35$), 40% ($n=17$), and 57% ($n=12$) in women with one, two, or three or more VFs, respectively. Similar patterns were also seen for incident VFs and hip fractures (Tables 2 and 3). Incidences of fractures, deaths, and time at risk among women with any VF are presented in Supplement information 1.

A Cox regression model adjusted for age, height, and weight revealed that the presence of any XVFA-identified VF was associated with an increased risk of incident any fracture (HR 2.04 [95% CI 1.44–2.87]), MOF (HR 2.32 [95% CI 1.55–3.48]), VF (HR 4.21 [95% CI 2.06–8.59]), and hip fracture (HR 2.93 [95% CI 1.33–6.44]). These associations were independent of CRFs included in FRAX and FN BMD (Fig. 1 and Supplement information 1).

When XVFA-identified VFs were stratified by severity, the presence of a grade 3 VF was associated with more than a twofold increased risk of any fracture (HR 2.39 [95% CI 1.32–4.31), a more than threefold increased risk of MOF (HR 3.31 [95% CI 1.75–6.26]), an over eightfold increased risk of VF (HR 8.55 [95% CI 3.53–20.75]), and an almost fivefold increased risk of hip fracture (HR 4.70 [95% CI 1.57–14.20]) (Table 2).

When VFs were categorized by number, the risk of any fracture, MOF, VF, and hip fracture among participants with three or more VFs was even higher (HR 3.89 [95% CI 2.06–7.34], HR 6.29 [95% CI 3.24–12.21], HR 13.49 [95% CI 5.39–33.75], and HR 6.06 [95% CI 1.83–20.14], respectively) (Table 3). These associations remained robust after further adjustment for CRFs and FN BMD (Fig. 2, Tables 2 and 3).

Incident non-vertebral fractures

Additional analyses were conducted for incident non-vertebral fractures. Overall, no consistent associations were observed between VFs, whether identified by XVFA or manual VFA, and the risk of incident non-vertebral fractures. One isolated association was observed for grade 1

Table 3 Number of vertebral fractures and associations with subsequent fracture risk: manual VFA vs deep learning-based XVFA in older women

	Manual VFA				XVFA			
	No VF $n=321$	One VF $n=71$	Two VF $n=19$	Three or more VF $n=12$	No VF $n=236$	One VF $n=123$	Two VF $n=43$	Three or more VF $n=21$
Any fracture								
No. (%)	89 (27.7)	34 (47.9)	11 (57.9)	4 (33.3)	57 (24.2)	48 (39.0)	21 (48.8)	12 (57.1)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	**1.84 [1.23–2.74]**	**2.98 [1.58–5.61]**	**1.62 [0.59–4.43]**	1 [Reference]	**1.70 [1.15–2.50]**	**2.51 [1.52–4.15]**	**3.89 [2.06–7.34]**
+ clinical risk factors	1 [Reference]	**1.71 [1.13–2.57]**	**3.04 [1.58–5.88]**	1.66 [0.60–4.57]	1 [Reference]	**1.83 [1.22–2.72]**	**2.19 [1.30–3.70]**	**3.75 [1.95–7.21]**
+ FN BMD	1 [Reference]	1.48 [0.98–2.24]	**3.06 [1.59–5.89]**	1.70 [0.61–4.73]	1 [Reference]	**1.86 [1.25–2.76]**	**1.83 [1.08–3.09]**	**3.71 [1.94–7.11]**
MOF								
No. (%)	64 (19.9)	24 (33.8)	10 (52.6)	4 (33.3)	38 (16.1)	35 (28.5)	17 (39.5)	12 (57.1)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	**1.70 [1.05–2.74]**	**3.60 [1.83–7.10]**	2.23 [0.81–6.17]	1 [Reference]	**1.80 [1.14–2.87]**	**2.80 [1.58–4.97]**	**6.29 [3.24–12.21]**
+ clinical risk factors	1 [Reference]	**1.65 [1.02–2.69]**	**3.50 [1.72–7.14]**	2.21 [0.79–6.15]	1 [Reference]	**1.90 [1.18–3.05]**	**2.37 [1.30–4.32]**	**6.14 [3.09–12.21]**
+ FN BMD	1 [Reference]	1.40 [0.86–2.28]	**3.75 [1.84–7.64]**	2.19 [0.77–6.22]	1 [Reference]	**1.95 [1.21–3.13]**	**1.84 [1.00–3.37]**	**5.90 [2.97–11.72]**
VF								
No. (%)	17 (5.3)	15 (21.1)	8 (42.1)	4 (33.3)	10 (4.2)	17 (13.8)	8 (18.6)	9 (42.9)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	**3.96 [1.95–8.06]**	**9.56 [4.05–22.58]**	**8.37 [2.78–25.18]**	1 [Reference]	**3.02 [1.37–6.66]**	**4.56 [1.79–11.63]**	**13.49 [5.39–33.75]**
+ clinical risk factors	1 [Reference]	**3.49 [1.72–7.09]**	**7.80 [3.10–19.61]**	**7.43 [2.43–22.72]**	1 [Reference]	**2.67 [1.19–6.00]**	**3.65 [1.38–9.65]**	**12.89 [4.93–33.67]**
+ FN BMD	1 [Reference]	**2.98 [1.48–6.03]**	**9.16 [3.80–22.08]**	**9.29 [2.88–29.91]**	1 [Reference]	**2.85 [1.27–6.37]**	**2.77 [1.03–7.41]**	**12.51 [4.84–32.31]**
Hip								
No. (%)	21 (6.5)	6 (8.5)	2 (10.5)	1 (8.3)	9 (3.8)	10 (8.1)	7 (16.3)	4 (19.0)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	1.07 [0.43–2.67]	1.71 [0.39–7.45]	1.18 [0.16–8.85]	1 [Reference]	1.98 [0.80–4.90]	**4.99 [1.84–13.55]**	**6.06 [1.83–20.14]**
+ clinical risk factors	1 [Reference]	1.24 [0.48–3.18]	1.49 [0.32–6.94]	0.82 [0.10–6.43]	1 [Reference]	2.03 [0.81–5.09]	**4.97 [1.73–14.31]**	**6.48 [1.75–24.06]**
+ FN BMD	1 [Reference]	0.99 [0.39–2.55]	1.24 [0.25–6.11]	0.54 [0.06–4.80]	1 [Reference]	2.09 [0.82–5.35]	**3.77 [1.28–11.14]**	**4.35 [1.09–17.32]**

Table 3 (Continued)

	Manual VFA				XVFA			
	No VF $n=321$	One VF $n=71$	Two VF $n=19$	Three or more VF $n=12$	No VF $n=236$	One VF $n=123$	Two VF $n=43$	Three or more VF $n=21$
Death								
No. (%)	46 (14.3)	15 (21.1)	7 (36.8)	4 (33.3)	34 (14.4)	19 (15.4)	10 (23.3)	9 (42.9)
HR (95% CI)								
Adjusted for age, height, weight	1 [Reference]	1.29 [0.71–2.32]	**2.36 [1.05–5.30]**	2.43 [0.87–6.79]	1 [Reference]	0.96 [0.54–1.69]	1.54 [0.76–3.14]	**2.97 [1.41–6.27]**
+ clinical risk factors	1 [Reference]	1.46 [0.79–2.68]	**2.71 [1.17–6.27]**	2.42 [0.85–6.94]	1 [Reference]	1.03 [0.58–1.82]	1.81 [0.88–3.74]	**3.44 [1.56–7.60]**
+ FN BMD	1 [Reference]	1.38 [0.75–2.55]	**2.56 [1.11–5.93]**	2.33 [0.81–6.70]	1 [Reference]	1.04 [0.59–1.84]	1.67 [0.80–3.50]	**3.23 [1.45–7.18]**

Associations were examined using cox regression model. Hazard Ratios (HR) and 95% confidence intervals (CI) are presented. Model 1: adjusted for age, height and weight. Model 2: adjusted for age, height, weight, and the clinical risk factors used in FRAX (previous fracture, family history of hip fracture, current smoking, oral glucocorticoid use, rheumatoid arthritis, excessive alcohol intake, secondary osteoporosis) except FN BMD. Model 3: adjusted for the same as model 2 with the addition of FN BMD. *FN*, femoral neck; *BMD*, bone mineral density; *MOF*, major osteoporotic fracture; *VF*, vertebral fracture; *VFA*, vertebral fracture assessment; *XVFA*, explainable vertebral fracture assessment (a model developed by using instead of by deep learning)

VFs identified by XVFA after adjustment for CRFs (HR 1.69, 95% CI 1.02–2.80), which should be interpreted with caution given the number of comparisons performed. Detailed results are presented in Supplement information 2.

Kaplan–Meier survival analysis for fracture prediction

Due to differences in the number of women with prevalent VFs identified by manual VFA and XVFA, a Kaplan–Meier survival analysis was performed for four groups based on baseline VF status: both negative ($n=217$), manual only ($n=19$), XVFA only ($n=104$), and both positive ($n=83$) (Fig. 3). The agreement between methods was fair ($\kappa=0.38$). Figure 3 shows Kaplan–Meier curves for incident fracture–free survival across these groups. Participants with VFs identified by both methods had the highest incidence of subsequent fractures, while those without VFs by either method had the lowest risk. Individuals with VFs detected by only one method showed intermediate risk levels. Censoring was appropriately accounted for in the analysis. Notably, participants with VFs detected by XVFA but not by manual VFA ($n=104$) showed divergence in fracture incidence compared with participants with no VFs. The global log-rank test across all four groups was statistically significant ($p<0.001$). In pairwise comparisons, the XVFA-only group differed significantly from the no-VF group (log-rank $p<0.001$), indicating that XVFA may capture prognostically relevant fracture risk. In multivariable Cox regression adjusted for CRFs and FN BMD, the association remained statistically significant (HR 2.23, 95% CI 1.45–3.43, $p<0.001$).

Discussion

In this population-based prospective cohort study of older Swedish women, VFs identified using the novel XVFA method, based on deep neural networks, were associated with increased risk of any fracture, MOF, VF, and hip fracture, independently of CRFs and FN BMD. Both XVFA and manual VFA showed dose–response associations, with higher VF grade and number conferring progressively higher fracture risk. Severe (grade 3) or multiple VFs carried risks consistent with previous literature, emphasizing the importance of detecting prevalent VFs as independent predictors of skeletal fragility. To our knowledge, this is the first study applying deep neural networks to DXA-derived VFA images using the GSQ method and prospectively linking them to X-ray–verified incident fractures and mortality. While overall predictive strength was comparable between XVFA and manual VFA, XVFA demonstrated more consistent associations across fracture outcomes,

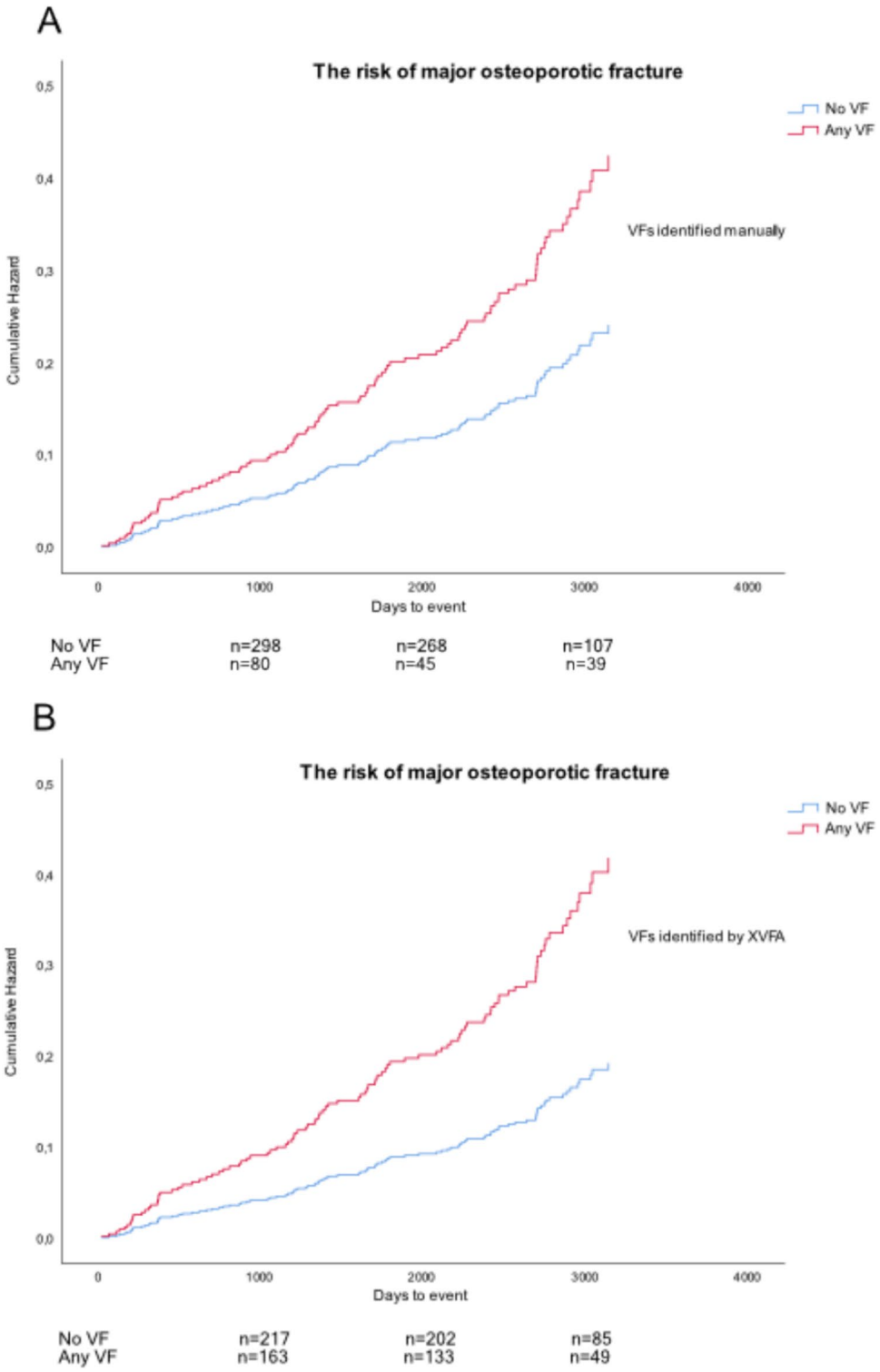

Fig. 1 Relationship between cumulative hazard for predicted major osteoporotic fracture and follow-up time (days) in older women with and without vertebral fracture (VF) at baseline, adjusted for age, height, weight, previous fracture, family history of hip fracture, current smoking, oral glucocorticoid use, rheumatoid arthritis, excessive alcohol intake, secondary osteoporosis (as used in FRAX), and femoral neck bone mineral density. In (**A**) VFs were identified by manual VFA; in (**B**) VFs were identified using the deep learning-based explainable VFA (XVFA) model

including hip fracture, suggesting that automated analysis may better capture vertebral deformities relevant to systemic skeletal weakness. The nearly tenfold risk increase for incident VF among women with severe prevalent VFs underscores the progressive nature of vertebral fragility. Additional analyses of incident non-vertebral fractures did not show any consistent associations with VFs, regardless of whether they were identified by XVFA or manual VFA (Supplement information 2). One isolated statistically significant association was observed for grade 1 VFs identified by XVFA after adjustment for CRFs; however, given the number of comparisons performed, this finding

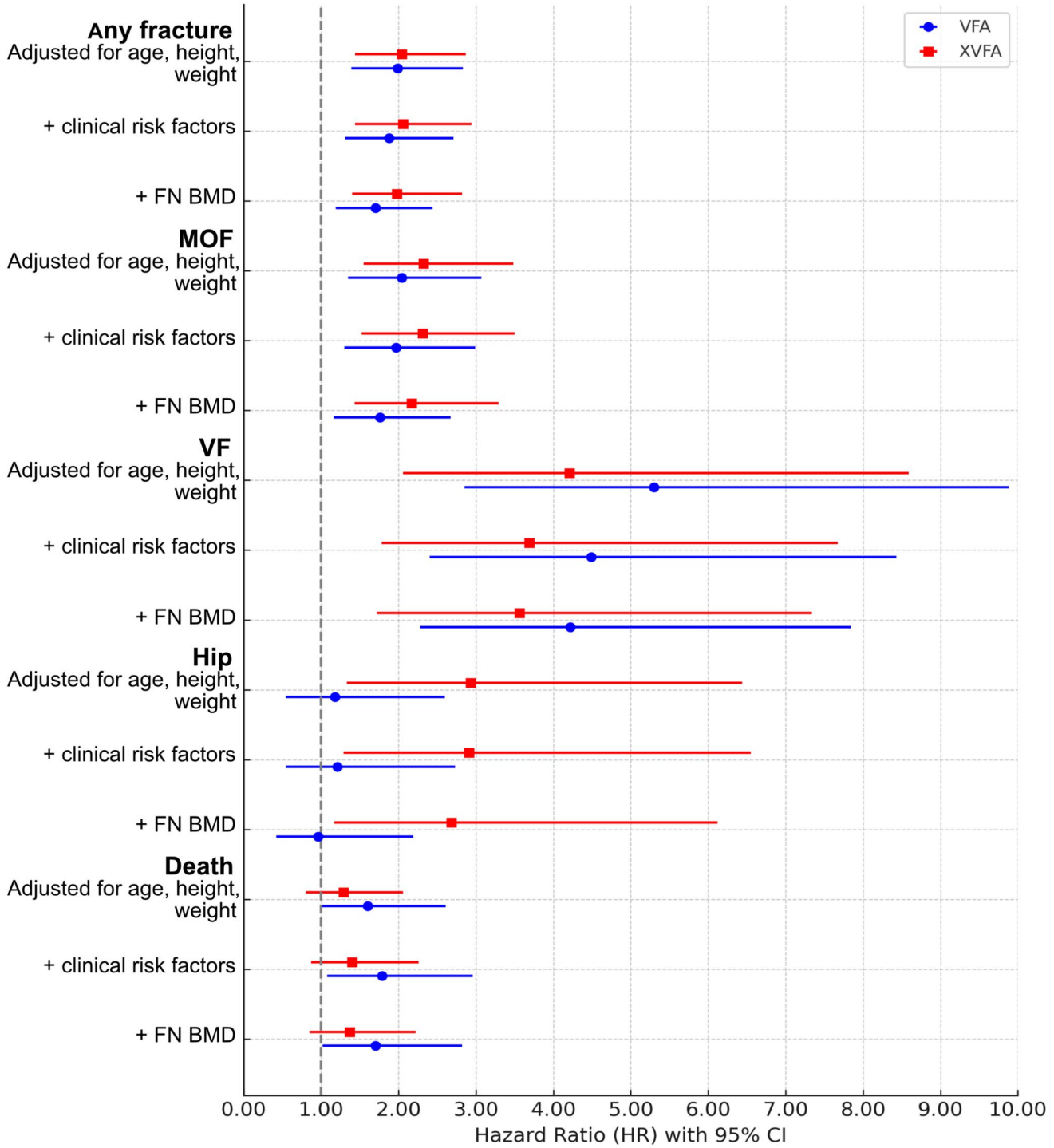

Fig. 2 Associations between vertebral fractures (VFs) identified by manual vertebral fracture assessment (VFA) or by the deep learning–based explainable VFA (XVFA) model and fracture risk in older women. Associations were examined using Cox proportional hazards regression models. Hazard ratios (HRs) and 95% confidence intervals (CIs) are presented in a forest plot. Clinical risk factors included previous fracture, family history of hip fracture, current smoking, oral glucocorticoid use, rheumatoid arthritis, excessive alcohol intake, and secondary osteoporosis (as used in FRAX)

should be interpreted with caution. The overall lack of association may reflect the heterogeneous nature of non-vertebral fracture outcomes, and/or poor statistical power due to low non-vertebral fracture events. Previous studies have shown that prevalent VFs are strong predictors of subsequent VFs and major osteoporotic fractures, whereas associations with broader categories of non-vertebral fractures tend to be weaker [1].

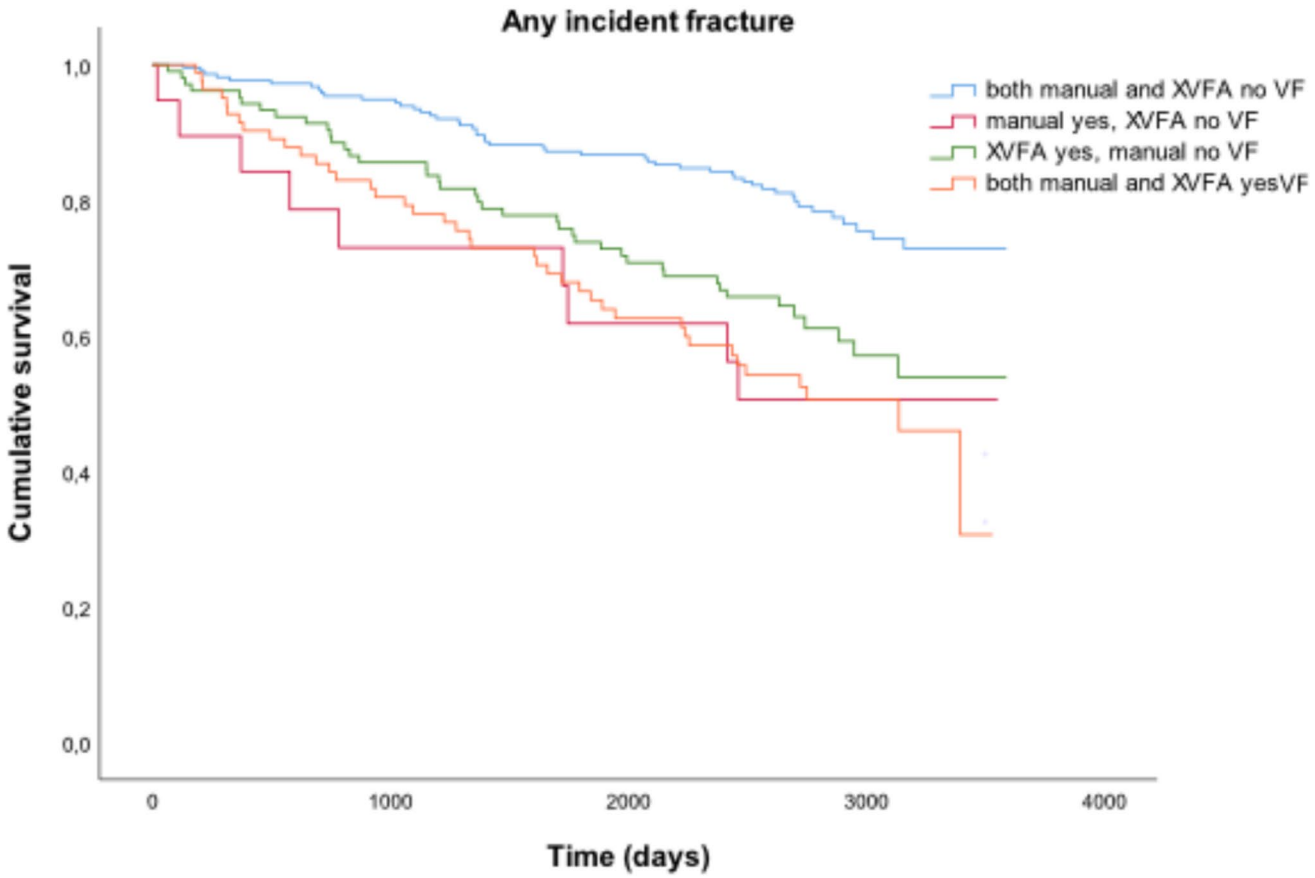

Fig. 3 Kaplan–Meier survival curves for incident fractures in the held-out test set (total $n=423$), stratified by baseline vertebral fracture status assessed by manual VFA and XVFA. Groups were: both negative ($n=217$, blue), manual only ($n=19$, red), XVFA only ($n=104$, green), and both positive ($n=83$, orange). Censoring is accounted for but not shown for clarity. XVFA = explainable vertebral fracture assessment (a model developed using deep learning)

Interestingly, participants with VFs detected by XVFA but not by manual VFA showed higher risk of incident fractures compared with participants with no VFs (Fig. 3). This association remained statistically significant after adjustment for CRFs and FN BMD using Cox regression, suggesting that XVFA may capture VFs with prognostic relevance. Given that XVFA was trained only on manual VF labels and had no access to fracture outcomes, this finding is unlikely to reflect overfitting or data leakage and may represent a clinically meaningful signal rather than a chance occurrence.

Previous studies support the growing role of deep learning in VF detection. In 2019, Derkatch et al. trained a CNN on 8920 VFA images (Lunar Prodigy or iDXA, GE Healthcare), validated against expert readers using mABQ criteria [25], achieving an AUC of 0.94 (95% CI: 0.93–0.95) for VF detection at the patient level, though vertebra-level localization was not specified [26]. Incident fracture data (mean follow-up 3.7 years) for 2813 patients showed adjusted HRs for any non-VF fracture of 1.7 (95% CI: 1.3–2.2) for the CNN and 1.8 (95% CI: 1.3–2.3) for expert readers. These findings are not directly comparable to ours because our definition of incident "any fracture" included VFs, and the mean follow-up was more than twice as long. They used iDXA for the majority of VFAs, which employs improved fan-beam technology, allowing better vertebral visualization than the Hologic Discovery system. Hip fracture incidence was lower in their study (3.0%) than in our test set (7.1%). Among participants with VFs, adjusted HRs for hip fracture were 2.3 (95% CI: 1.5–3.5) for the CNN and 2.4 (95% CI: 1.5–3.7) for experts, similar to XVFA in the test set (HR 2.68 [95% CI: 1.17–6.12]), whereas manual VFA showed no significant associations (Supplement information 1). High hip fracture incidence among women without prevalent VFs likely reflects the advanced age of the cohort (75–79 years), where chronological age is a dominant risk factor [27], and competing mortality risk among women with VFs may mask later hip fractures [28]. Thus, the association between prevalent VFs and hip fracture risk may be partly underestimated. Since 2015, several machine learning (ML) models have been developed for osteoporosis and VF detection [29]. Only two previous studies used Hologic VFAs. Monchka et al. trained vertebra-level CNNs, achieving 91.9%

sensitivity, 99% specificity, and 86.1% F1-score in a test set of 819 labeled VFAs [30]. Hong et al. developed deep learning models to detect VFs (using the mABQ method) and osteoporosis on lateral spine radiographs and VFAs, achieving AUROCs of 0.92 and 0.87, respectively [31]. Individuals with deep learning-detected VFs or osteoporosis had roughly double the fracture risk independent of CRFs and FN BMD. Together with our findings, these studies demonstrate that neural networks can reach diagnostic accuracies comparable to expert readers while enabling reproducible, scalable assessment. XVFA differs in being trained, validated, and tested solely on Hologic VFAs throughout the modeling process. Malgo et al. assessed VFA for VF detection through a retrospective study and meta-analysis including 542 patients and 16 studies (3238 subjects) [9]. The pooled sensitivity and specificity were 0.84 (95% CI: 0.72–0.92) and 0.90 (95% CI: 0.84–0.94), respectively, supporting VFA as a useful tool but suggesting caution in relying solely on it.

Several limitations of XVFA should be noted. Accuracy of keypoint predictions in the first stage directly affects morphometric classification in the second. Noisy images may propagate errors, though the two-stage design mirrors clinical workflow and enhances explainability. XVFA was trained on annotated vertebrae ($n=9862$) without vertebral-level information (T4–L4), achieving an end-to-end AUC of 97% [17]. The end-to-end AUC reflects the entire pipeline, including vertebral detection and anatomical rules, rather than a pure image-to-diagnosis test, which may contribute to its high value. In the present study, manual annotations included only vertebrae T4–L4 ($n=4563$) and excluded vertebrae that were not analyzable, often due to poor image quality, whereas XVFA analyzed all visible vertebrae ($n=5532$) and was not explicitly trained to disregard poorly visible vertebrae. This difference in the number of vertebrae analyzed and the handling of image quality explains why XVFA identified more women with VFs than manual VFA (187 vs 102). Many of these additional XVFA-only detections, which represent asymmetric disagreement and were not counted as fractures in the manual annotation, are classified as false positives when using the manual assessment as reference, but likely reflect increased sensitivity of XVFA, particularly for mild or borderline deformities. The fair kappa value ($\kappa=0.38$) observed in this study does not contradict the high AUC (0.97), as these metrics capture different aspects of performance. Kappa measures agreement at a fixed binary threshold and is sensitive to prevalence and class imbalance, whereas AUC evaluates the model's ability to correctly rank individuals across all possible thresholds. Differences in the number of VFs detected between methods, as highlighted by the Kaplan–Meier analysis, may also contribute to the fair kappa. The high AUC therefore reflects excellent discrimination despite only fair binary agreement.

Derkatch et al. developed a CNN for detecting VFs at the patient level [26], whereas Monchka et al. did so at the vertebra level, providing more detailed information useful for clinical and AI training [30]. Our XVFA also assessed individual vertebrae but did not determine vertebral level, unlike the manual VFA (T4–L4). However, XVFA could evaluate both VF number and severity per image, effectively providing vertebra-level identification. The study population included only ambulatory Swedish women aged 75–80 years, limiting generalizability. Statistical power to detect associations between VFA identified VFs and hip fracture was lower than for other fracture categories. Future studies should include more diverse populations.

Strengths include the large, population-based prospective design with X-ray–verified incident fractures, high data completeness, and standardized morphometric assessment. To our knowledge, this is the first study to evaluate a deep neural network–based method for identifying VFs on DXA-derived VFAs according to both severity and number, and to prospectively examine their associations with incident fractures. The combination of population-based design, objective AI quantification, and clinically validated outcomes represents a robust framework for translating deep learning tools into real-world skeletal health assessment.

Conclusion

This study shows that XVFA can identify VFs and predict incident fractures independently of FRAX risk factors and FN BMD, with more consistent associations than manual assessment. As VFs are often under-recognized, automated XVFA offers a scalable and objective way to improve fracture risk assessment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00198-026-08072-9.

Funding Open access funding provided by University of Gothenburg. This study was funded by the Swedish Research Council (VR), the ALF/LUA grant from the Sahlgrenska University Hospital, and in part through the AIDA project grant DNR 2021—01420.

Data availability All data are available from the corresponding author upon reasonable request.

Declarations

Ethics approval All subjects signed an informed consent prior to participation. The study has been approved by the regional Ethics Review Board in Gothenburg (Dnr 929—12).

Conflict of interest ML has received lecture or consulting fees from Astellas, Amgen, UCB Pharma, Medison Pharma, Sandoz, Gedeon Richter, Jansen-Cilag, Medac, Pharmacosmos, Parexel International, and Crinetics, all outside the submitted work. LJ has received lecture

fees from UCB Pharma, all outside the submitted work. All other authors declare no conflicts of interest.

References

1. Klotzbuecher CM, Ross PD, Landsman PB, Abbott TA 3rd, Berger M (2000) Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. J Bone Miner Res 15:721–739
2. Pongchaiyakul C, Nguyen ND, Jones G, Center JR, Eisman JA, Nguyen TV (2005) Asymptomatic vertebral deformity as a major risk factor for subsequent fractures and mortality: a long-term prospective study. J Bone Miner Res 20:1349–1355
3. Gehlbach SH, Bigelow C, Heimisdottir M, May S, Walker M, Kirkwood JR (2000) Recognition of vertebral fracture in a clinical setting. Osteoporos Int 11:577–582
4. Schousboe JT, Vokes T, Broy SB, Ferrar L, McKiernan F, Roux C, Binkley N (2008) Vertebral fracture assessment: the 2007 ISCD official positions. J Clin Densitom 11:92–108
5. Lorentzon M, Nilsson AG, Johansson H, Kanis JA, Mellstrom D, Sundh D (2019) Extensive undertreatment of osteoporosis in older Swedish women. Osteoporos Int 30:1297–1305
6. Delmas PD, van de Langerijt L, Watts NB, Eastell R, Genant H, Grauer A, Cahall DL, Group IS (2005) Underdiagnosis of vertebral fractures is a worldwide problem: the IMPACT study. J Bone Miner Res 20:557-563
7. Oei L, Koromani F, Breda SJ et al (2018) Osteoporotic vertebral fracture prevalence varies widely between qualitative and quantitative radiological assessment methods: the Rotterdam Study. J Bone Miner Res 33:560–568
8. Szulc P (2018) Vertebral fracture: diagnostic difficulties of a major medical problem. J Bone Miner Res 33:553–559
9. Malgo F, Hamdy NAT, Ticheler C, Smit F, Kroon HM, Rabelink TJ, Dekkers OM, Appelman-Dijkstra NM (2017) Value and potential limitations of vertebral fracture assessment (VFA) compared to conventional spine radiography: experience from a fracture liaison service (FLS) and a meta-analysis. Osteoporos Int. https://doi.org/10.1007/s00198-017-4137-6
10. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
11. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, Skoldenberg O, Gordon M (2017) Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 88:581–586
12. Baum T, Bauer JS, Klinder T, Dobritz M, Rummeny EJ, Noel PB, Lorenz C (2014) Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT. Eur Radiol 24:872–880
13. Murata K, Endo K, Aihara T et al (2020) Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. Sci Rep 10:20031
14. Dong Q, Luo G, Lane NE et al (2022) Deep learning classification of spinal osteoporotic compression fractures on radiographs using an adaptation of the Genant semiquantitative criteria. Acad Radiol 29:1819–1832
15. Nilsson AG, Sundh D, Johansson L, Nilsson M, Mellstrom D, Rudang R, Zoulakis M, Wallander M, Darelid A, Lorentzon M (2017) Type 2 diabetes mellitus is associated with better bone microarchitecture but lower bone material strength and poorer physical function in elderly women: a population-based study. J Bone Miner Res 32:1062–1071
16. Johansson L, Svensson HK, Karlsson J, Olsson LE, Mellstrom D, Lorentzon M, Sundh D (2019) Decreased physical health-related quality of life-a persisting state for older women with clinical vertebral fracture. Osteoporos Int 30:1961–1971
17. Victor Wåhlstrand Skärström LJ, Jennifer Alvén, Mattias Lorentzon, Ida Häggström (2024) Explainable vertebral fracture analysis with uncertainty estimation using differentiable rule-based classification. https://arxiv.org/abs/2407.02926
18. Johansson L, Johansson H, Axelsson KF et al (2022) Improved fracture risk prediction by adding VFA-identified vertebral fracture data to BMD by DXA and clinical risk factors used in FRAX. Osteoporos Int 33:1725–1738
19. Johansson L, Sundh D, Magnusson P, Rukmangatharajan K, Mellstrom D, Nilsson AG, Lorentzon M (2020) Grade 1 vertebral fractures identified by densitometric lateral spine imaging predict incident major osteoporotic fracture independently of clinical risk factors and bone mineral density in older women. J Bone Miner Res 35:1942–1951
20. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E (2008) FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int 19:385–397
21. Blake GM, Rea JA, Fogelman I (1997) Vertebral morphometry studies using dual-energy x-ray absorptiometry. Semin Nucl Med 27:276–290
22. Genant HK, Wu CY, van Kuijk C, Nevitt MC (1993) Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res 8:1137–1148
23. Griffith JF (2015) Identifying osteoporotic vertebral fracture. Quant Imaging Med Surg 5:592–602
24. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, Ni LM, Shum HY (2022) Dino: Detr with improved denoising anchor boxes for end-to-end object detection. https://doi.org/10.48550/arXiv.2203.03605
25. Jiang G, Eastell R, Barrington NA, Ferrar L (2004) Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. Osteoporos Int 15:887–896
26. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD (2019) Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual X-ray absorptiometry. Radiology 293:405–411
27. Johnell O, Kanis JA (2006) An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 17:1726–1733
28. Jalava T, Sarna S, Pylkkanen L et al (2003) Association between vertebral fracture and increased mortality in osteoporotic patients. J Bone Miner Res 18:1254–1260
29. Smets J, Shevroja E, Hugle T, Leslie WD, Hans D (2021) Machine learning solutions for osteoporosis-a review. J Bone Miner Res 36:833–851

30. Monchka BA, Schousboe JT, Davidson MJ, Kimelman D, Hans D, Raina P, Leslie WD (2022) Development of a manufacturer-independent convolutional neural network for the automated identification of vertebral compression fractures in vertebral fracture assessment images using active learning. Bone 161:116427
31. Hong N, Cho SW, Lee YH, Kim CO, Kim HC, Rhee Y, Leslie WD, Cummings SR, Kim KM (2025) Deep learning-based identification of vertebral fracture and osteoporosis in lateral spine radiographs and DXA vertebral fracture assessment to predict incident fracture. J Bone Miner Res 40:628–638