Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bone Turnover Markers Improve Prediction of Adjacent Vertebral Refracture After Percutaneous Vertebroplasty: An Explainable Machine Learning Study

Enci Sun
Jining Medical University
Zepeng Xiang
Jining Medical University
Guangdong Gao
Jining Medical University
Qingwen Wu
Affiliated Hospital of Jining Medical University
Hong Wang
China Medical University
Qingwei Li
plasurg0618@163.com

Affiliated Hospital of Jining Medical University

Research Article

Keywords: Osteoporotic vertebral compression fracture, Percutaneous vertebroplasty, Adjacent vertebral refracture, Bone turnover markers, β-C-terminal telopeptide of type I collagen, Machine learning, Explainable artificial intelligence

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11534354/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bone Turnover Markers Improve Prediction of Adjacent Vertebral Refracture After Percutaneous Vertebroplasty: An Explainable Machine Learning Study

Enci Sun[1,3], Zepeng Xiang[1,3], Guangdong Gao[1,3], Qingwen Wu[3], Hong Wang[2,4], Qingwei Li[3,4]*

Affiliations

1 Jining Medical University, Jining, Shandong 272000, China

2 China Medical University, Shenyang, Liaoning, China

3 Department of Spine Surgery, Affiliated Hospital of Jining Medical University, Jining, Shandong, China

4 Department of Spine Surgery, Dalian Central Hospital, Dalian, Liaoning, China

*Corresponding author

Corresponding Authors:

Hong Wang

Department of Spine Surgery, Dalian Central Hospital, Dalian, Liaoning, China

China Medical University, Shenyang, Liaoning, China

Tel: +86 17661345209

E-mail: plasurg0618@163.com

Qingwei Li

Department of Spine Surgery, Affiliated Hospital of Jining Medical University

No. 89 Guhuai Road, Jining, Shandong 272000, China

China Medical University, Shenyang, Liaoning, China

Tel: +86 17661345209

E-mail: plasurg0618@163.com

Abstract

Objective: To investigate whether bone turnover markers provide predictive information beyond conventional clinical risk factors for adjacent vertebral refracture after percutaneous vertebroplasty (PVP) in older patients with osteoporotic vertebral compression fractures (OVCFs), and to to develop an explainable risk stratification model.

Methods: We retrospectively included patients with OVCFs who underwent PVP between June 2019 and August 2024. The final analytic cohort comprised 1,730 patients, including 316 who developed adjacent vertebral refracture and 1,414 who did not. Patients were randomly divided at an approximately 8:2 ratio into a training cohort (n = 1,383) and an independent test cohort (n = 347). Eight prespecified predictors were evaluated: age, sex, body mass index (BMI), bone mineral density (BMD), β-C-terminal telopeptide of type I collagen (β-CTX), N-mid osteocalcin (N-OC), total procollagen type I N-terminal propeptide (tPINP), and 25-hydroxyvitamin D (25-OH-D). Six machine learning models were developed, including logistic regression (LR), decision tree (DT), random forest (RF), extra trees (ET), gradient boosting (GB), and extreme gradient boosting (XGB), and were evaluated in the independent test cohort. Model discrimination was compared using DeLong tests. Multivariable logistic regression, restricted cubic spline analyses, and SHapley Additive exPlanations (SHAP) were used to characterize associations and model interpretability. Robustness was further assessed using feature-set restriction, 1st-to-99th percentile Winsorization, and threshold sensitivity analyses based on out-of-fold predictions from the training cohort.

Results: Among 1,730 patients, 316 (18.3%) developed adjacent vertebral refracture. In the independent test cohort, ET achieved the numerically highest area under the receiver operating characteristic curve (AUC) of 0.872 (95% CI, 0.829–0.915), whereas RF achieved an AUC of 0.869 (95% CI, 0.826–0.912); the difference was not statistically significant after Benjamini-Hochberg adjustment (P = 0.709). RF had the highest accuracy (0.818) and F1 score (0.583), with a Brier score of 0.106, and was therefore selected as the primary model for

subsequent interpretation. SHAP analysis identified β-CTX as the most influential predictor, followed by BMD, N-OC, and 25-OH-D. In feature-set sensitivity analysis, the full RF model yielded an AUC of 0.862 (95% CI, 0.817–0.906), whereas a reduced model containing only the five bone-related variables yielded an AUC of 0.848 (95% CI, 0.801–0.896), with no significant difference between the two models (P = 0.339). In contrast, a model based only on age, sex, and BMI had substantially lower discrimination (AUC, 0.570; 95% CI, 0.494–0.646). After Winsorization, the AUC was 0.867 (95% CI, 0.824–0.910), which did not differ significantly from the original analysis (P = 0.078). Threshold sensitivity analysis further showed that lowering the classification threshold increased sensitivity and negative predictive value at the expense of specificity.

Conclusions: Bone turnover-related variables, particularly β-CTX, provide clinically relevant predictive information beyond conventional clinical characteristics for identifying patients at increased risk of adjacent vertebral refracture after PVP. An RF model based on eight routinely available variables demonstrated good discrimination and robust performance for internal risk stratification. However, given the retrospective design and imperfect calibration (calibration slope, 1.640), the model is currently better suited for risk stratification than for precise estimation of individual absolute risk.

Keywords: Osteoporotic vertebral compression fracture; Percutaneous vertebroplasty; Adjacent vertebral refracture; Bone turnover markers; β-C-terminal telopeptide of type I collagen; Machine learning; Explainable artificial intelligence

introduction

Osteoporotic vertebral fracture (OVF) is one of the most common fragility fractures and can lead to pain, spinal deformity, impaired mobility, and reduced quality of life, while substantially increasing the risk of subsequent fractures(1).Previous studies have suggested that the risk of refracture is particularly high during the early period after a vertebral fracture. Therefore, early identification of patients with OVF who are at high risk of

refracture is important for developing individualized follow-up strategies and secondary fracture prevention measures.

Refracture in patients with OVF is influenced by multiple factors, including age, sex, body mass index, bone mineral density (BMD), and bone metabolic status. Although BMD is an important indicator for fracture risk assessment, it cannot fully reflect the dynamic process of bone remodeling when considered alone. Total procollagen type I N-terminal propeptide (tPINP) and β-isomerized C-terminal telopeptide of type I collagen (β-CTX-I) reflect bone formation and bone resorption, respectively, and are currently recommended by the International Osteoporosis Foundation and the International Federation of Clinical Chemistry and Laboratory Medicine as reference bone turnover markers for osteoporosis research and management (2). Therefore, integrating demographic characteristics, BMD, and bone metabolism-related biomarkers may provide a more comprehensive assessment of refracture risk. In recent years, machine learning has been applied to predict the risk of refracture after vertebral augmentation in patients with osteoporotic vertebral compression fractures (OVCFs), demonstrating promising predictive potential(3).However, most existing studies have been retrospective, and some models rely on radiomic features or relatively large numbers of predictors. In addition, insufficient external validation and limited transparency in model reporting have restricted model interpretability and clinical translation. SHapley Additive exPlanations (SHAP) can quantify the contribution of individual variables to model predictions and facilitate interpretation of both global and patient-level prediction behavior in complex machine learning models(4).Therefore, in the present study, we aimed to develop and compare multiple machine learning models based on a limited number of readily available clinical variables to predict the risk of adjacent vertebral refracture after percutaneous vertebroplasty (PVP) in patients with OVCFs. We further used SHAP analysis to evaluate key predictors and their individual contributions to model predictions, with the goal of developing a risk assessment approach that balances predictive performance, clinical accessibility, and interpretability. The overall study workflow is shown in Figure 1.

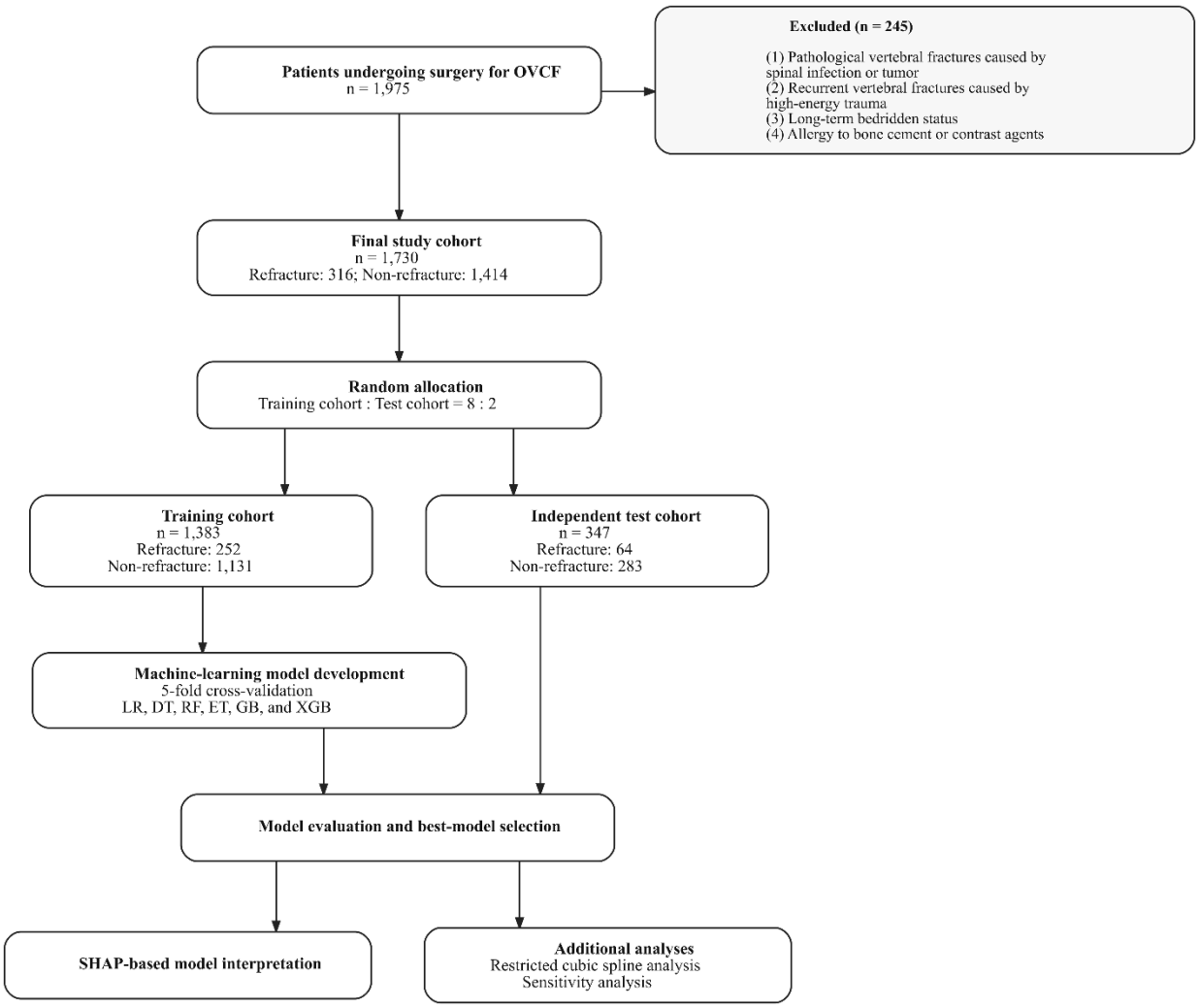


Figure 1. Study flow diagram. A total of 1,975 patients with osteoporotic vertebral compression fractures (OVCFs) who underwent percutaneous vertebroplasty (PVP) were initially screened. After exclusion of 245 patients according to the prespecified exclusion criteria, 1,730 patients were included in the final analytic cohort, including 316 with adjacent vertebral refracture and 1,414 without refracture. Patients were randomly allocated at an approximately 8:2 ratio to a training cohort (n = 1,383) and an independent hold-out test cohort (n = 347). Machine-learning models were developed using 5-fold cross-validation within the training cohort and evaluated in the hold-out test cohort.

2. Materials and Methods

2.1 Study Design and Participants

This was a single-center retrospective cohort study. We consecutively collected data from patients with osteoporotic vertebral compression fractures (OVCFs) who underwent percutaneous vertebroplasty (PVP) at

the Affiliated Hospital of Jining Medical University between June 2019 and August 2024. A total of 1,975 patients were initially screened, and 1,730 patients were ultimately included in the analysis according to the prespecified inclusion and exclusion criteria. The exclusion criteria were as follows: (1) pathological vertebral fractures caused by spinal infection or tumors; (2) recurrent vertebral fractures resulting from high-energy trauma; (3) long-term bedridden status; and (4) allergy to bone cement or contrast agents. In total, 245 patients were excluded.

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Affiliated Hospital of Jining Medical University (approval No. 2024-08-C024).

2.2 Outcome Definition and Follow-up

The primary outcome was adjacent vertebral refracture. Inpatient and outpatient records from discharge after the initial PVP procedure through the end of the study observation period in August 2024 were retrieved from the hospital electronic medical record system. Adjacent vertebral refracture was defined as a newly developed vertebral compression fracture involving a vertebra adjacent to the initially treated vertebra that occurred after the index PVP procedure and before August 2024, was confirmed by radiography or computed tomography (CT), and resulted in rehospitalization for treatment.

Patients who developed an adjacent vertebral refracture during the observation period were classified into the refracture group. For patients without refracture, follow-up was censored in August 2024. Among the 1,730 patients included in the final analysis, 316 experienced adjacent vertebral refracture and 1,414 did not.

2.3 Clinical Data and Predictor Variables

Demographic characteristics, bone mineral density, and bone metabolism-related data were extracted from the electronic medical records and laboratory information system during the index hospitalization for PVP. Eight predictor variables were ultimately included in the models: age, sex, body mass index (BMI), bone mineral density (BMD), β-C-terminal telopeptide of type I collagen (β-CTX), N-mid osteocalcin (N-OC), total

procollagen type I N-terminal propeptide (tPINP), and 25-hydroxyvitamin D (25-OH-D). Lumbar spine BMD was measured using quantitative computed tomography (QCT). β-CTX, N-OC, tPINP, and 25-OH-D were obtained from blood tests performed after hospital admission.

The final analytic cohort had complete data for all eight prespecified predictor variables and the outcome; therefore, no data imputation was required. To assess potential selection bias associated with the exclusion process, we compared the distributions of age, sex, and BMD between the initially screened cohort and the final analytic cohort and found no significant differences. Nevertheless, these comparisons cannot completely exclude selection bias resulting from the prespecified exclusion criteria.

2.4 Univariable and Multivariable Logistic Regression Analyses

To investigate the associations between the candidate predictors and refracture risk, we first performed univariable logistic regression analyses to estimate the unadjusted associations of individual variables with refracture, expressed as crude odds ratios (ORs). Subsequently, based on prior clinical knowledge and the existing literature, all eight prespecified predictors—age, sex (male vs female), body mass index (BMI), bone mineral density (BMD), β-C-terminal telopeptide of type I collagen (β-CTX), N-mid osteocalcin (N-OC), total procollagen type I N-terminal propeptide (tPINP), and 25-hydroxyvitamin D (25-OH-D)—were simultaneously entered into a multivariable logistic regression model to estimate their independent associations with refracture after adjustment for the other potential confounders. Results are reported as ORs with corresponding 95% confidence intervals (CIs). To minimize bias arising from the exclusion of potentially important predictors, we did not use stepwise variable selection based on univariable *P*-value thresholds (e.g., $P < 0.05$); instead, all prespecified variables were retained in the multivariable model (5).

2.5 Dataset Partitioning and Model Development

The entire cohort of 1,730 patients was randomly divided at an approximately 8:2 ratio into a training cohort and an independent test cohort. The training cohort comprised 1,383 patients, including 252 with refracture and 1,131 without refracture. The independent test cohort comprised 347 patients, including 64 with refracture and 283 without refracture.

Six machine learning models were developed using the training cohort: logistic regression (LR), decision tree (DT), random forest (RF), extra trees (ET), gradient boosting (GB), and extreme gradient boosting (XGB). Model training and hyperparameter tuning were performed in R (version 4.5.2) using 5-fold cross-validation, with the area under the receiver operating characteristic curve (ROC-AUC) used as a primary criterion for model optimization and hyperparameter selection. The final RF model was implemented using the ranger algorithm, an efficient C++-based implementation of random forests that supports classification, regression, and survival analysis(6).The optimal hyperparameters for the final RF model were mtry = 2, splitrule = "gini", and min.node.size = 10. All model training and hyperparameter tuning were performed exclusively within the training cohort; the independent test cohort was not used at any stage of model development.

2.6 Model Performance Evaluation

The predictive performance of all models was evaluated in the independent test cohort. Discrimination was assessed using the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC), with 95% CIs reported for ROC-AUC. Using fixed classification thresholds derived from out-of-fold (OOF) predictions in the training cohort, we further calculated accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score.

Overall probabilistic prediction error was assessed using the Brier score. In accordance with current methodological recommendations for clinical prediction models, model performance was evaluated by considering discrimination, calibration, and overall prediction error rather than relying solely on ROC-AUC

(7).Pairwise comparisons of ROC-AUCs between models were performed using the nonparametric DeLong method (8).Because multiple model comparisons were conducted, the Benjamini-Hochberg (BH) procedure was applied to control the false discovery rate (9).For the RF model ultimately selected for subsequent interpretation, 1,000 bootstrap resamples were used to generate a bias-corrected calibration curve, and the Brier score, calibration intercept, and calibration slope were also reported. Decision curve analysis (DCA) was further performed to evaluate the net benefit of the model across a range of threshold probabilities relative to the **treat-all** and **treat-none** strategies (10).

2.7 Model Interpretability Analysis

SHapley Additive exPlanations (SHAP) were used to interpret the RF model selected as the primary model for subsequent interpretation.SHAP values were used to assess the relative importance and direction of effect of each predictor on model output at the global level. Patient-level explanations were also generated to quantify the contribution of individual features to the predicted refracture risk for a given patient, thereby improving model transparency and clinical interpretability (4).

2.8 Restricted Cubic Spline Analysis

To further explore potential nonlinear associations between bone-related continuous variables and the risk of adjacent vertebral refracture, restricted cubic spline (RCS) analyses were performed separately for β-CTX, BMD, N-OC, tPINP, and 25-OH-D. Each RCS model included four knots located at the 5th, 35th, 65th, and 95th percentiles of the distribution of the corresponding variable. When evaluating a given target variable, the model was adjusted for the remaining seven predictors to assess both the overall association and potential nonlinearity between that variable and refracture risk.

195 **2.9 Subgroup and Interaction Analyses**

196 To evaluate the stability of RF model discrimination across different patient characteristics, prespecified
197 subgroup analyses were conducted in the independent test cohort. Patients were stratified by sex (female vs
198 male), age (<75 vs ≥75 years), BMI (<24 vs ≥24 kg/m²), β-CTX (<468.4 vs ≥468.4), and 25-OH-D (<18.28 vs
199 ≥18.28). The ROC-AUC and corresponding 95% confidence interval (CI) were calculated for each subgroup.
200 Bootstrap methods were used to estimate the between-subgroup differences in AUC and their 95% CIs within
201 each stratification factor, and *P* values from multiple subgroup comparisons were adjusted using the
202 Benjamini-Hochberg (BH) procedure.

203 In addition, exploratory interaction analyses were performed in the full study cohort (n = 1,730). With β-CTX
204 specified as the primary variable of interest, its interactions with sex, age, BMI, BMD, and 25-OH-D were
205 evaluated separately within a multivariable logistic regression framework. Interactions were assessed by
206 comparing nested models with and without the corresponding interaction term using likelihood ratio tests. The
207 *P* values from the five interaction tests were further adjusted for multiple comparisons using the BH procedure.
208 For interactions that remained statistically supported after multiple-comparison adjustment, simple effects of
209 β-CTX were further estimated at representative levels of the interacting variable to facilitate interpretation of
210 heterogeneity in the association. Subgroup and interaction analyses were considered exploratory and were not
211 used to develop additional independent prediction models.

212

213 **2.10 Sensitivity Analyses**

214 Three types of sensitivity analyses were performed to assess the robustness of model performance and the
215 main study findings. First, feature-set restriction analyses were conducted to evaluate the stability of RF model
216 performance across different sets of predictors. Second, continuous predictors were Winsorized at the 1st and
217 99th percentiles, after which the RF model was refitted and its performance was compared with that of the

corresponding non-Winsorized model. Third, threshold sensitivity analyses were performed using alternative classification thresholds derived exclusively from out-of-fold (OOF) predicted probabilities in the training cohort. In addition to the OOF Youden threshold, sensitivity-oriented thresholds corresponding to an OOF sensitivity of at least 80% and at least 90% were determined and then fixed for evaluation in the independent test cohort.

2.11 Development of the Clinical Risk Calculator

To enhance model accessibility and facilitate its potential clinical translation, a web-based risk calculator based on the final random forest (RF) model was developed using R Shiny. The calculator directly loads the frozen ranger RF model object and does not retrain, update, or recalibrate the model during use.

After users enter the eight predictor variables—age, sex, body mass index (BMI), lumbar spine bone mineral density (BMD) measured by quantitative computed tomography (QCT), β-C-terminal telopeptide of type I collagen (β-CTX), N-mid osteocalcin (N-OC), total procollagen type I N-terminal propeptide (tPINP), and 25-hydroxyvitamin D (25-OH-D)—the calculator provides an individualized predicted probability of adjacent vertebral refracture and a reference classification threshold derived from out-of-fold (OOF) predictions in the training cohort. The web-based calculator has been deployed on a public server and is publicly accessible online(11). The tool is currently intended for research and risk-stratification purposes only and has not yet undergone independent external validation.

2.12 Statistical Analysis

All statistical analyses and machine learning modeling were performed using R software (version 4.5.2). Continuous variables are presented as medians with interquartile ranges [median (IQR)] and were compared between groups using the Wilcoxon rank-sum test. Categorical variables are presented as frequencies and

241 percentages [n (%)] and were compared using the Pearson chi-square test. All statistical tests were two-sided,
242 and $P < 0.05$ was considered statistically significant.
243
244 **3.1 Study Population and Baseline Characteristics**
245 A total of 1,975 patients with OVCFs who underwent PVP were screened. After exclusion of 245 patients according to
246 the prespecified criteria, 1,730 patients were included in the final analysis **(Figure 1)**. Among them, 316 patients (18.3%)
247 developed adjacent vertebral refracture during the observation period, whereas 1,414 patients (81.7%) did not.
248 Compared with patients without refracture, those who developed refracture were older [74.00 (69.00–80.00) vs 71.00
249 (65.00–77.00) years, $P < 0.001$], had a higher proportion of men (23.4% vs 16.3%, $P = 0.003$), and had lower BMD
250 [38.20 (22.03–58.00) mg/cm³ vs 56.65 (38.20–78.00) mg/cm³, $P < 0.001$]. In addition, patients in the refracture group
251 had higher levels of β-CTX, N-OC, and tPINP and lower levels of 25-OH-D (all $P < 0.001$). No significant difference in
252 BMI was observed between the two groups ($P = 0.403$) **(Table 1).**
253 The entire cohort was randomly divided at an approximately 8:2 ratio into a training cohort (n = 1,383) and an independent
254 test cohort (n = 347). Refracture occurred in 252 patients (18.2%) in the training cohort and in 64 patients (18.4%) in the
255 test cohort. No statistically significant differences were observed between the training and test cohorts in outcome
256 distribution, age, sex, BMI, BMD, β-CTX, N-OC, tPINP, or 25-OH-D (all $P > 0.05$), indicating that the baseline
257 characteristics of the two cohorts were generally comparable **(Supplementary Table S 1).**
258 **Table 1.Baseline characteristics of the study population according to adjacent vertebral refracture status.**

Characteristic	Overall	Non_refracture	Refracture	P_value
Age, years	72.00 (66.00–78.00)	71.00 (65.00–77.00)	74.00 (69.00–80.00)	<0.001
Sex, n (%)				0.003

Characteristic	Overall	Non_refracture	Refracture	P_value
Female	1425 (82.4%)	1183 (83.7%)	242 (76.6%)	
Male	305 (17.6%)	231 (16.3%)	74 (23.4%)	
BMI, kg/m²	23.19 (20.76–25.46)	23.19 (20.76–25.50)	23.13 (20.45–25.39)	0.403
BMD, mg/cm³	53.35 (34.62–74.70)	56.65 (38.20–78.00)	38.20 (22.03–58.00)	<0.001
β-CTX	447.25 (249.43–731.80)	360.70 (243.03–650.30)	743.50 (537.70–901.48)	<0.001
N-OC	20.75 (13.70–31.50)	19.99 (13.28–29.70)	24.70 (16.40–36.70)	<0.001
tPINP	55.23 (39.05–70.38)	53.77 (38.09–68.93)	61.62 (44.80–80.67)	<0.001
25-OH-D	21.78 (16.47–26.68)	22.06 (17.05–27.17)	19.39 (14.41–24.93)	<0.001

Data are presented as median (IQR) for continuous variables and n (%) for categorical variables. Continuous variables were compared using the Wilcoxon rank-sum test, whereas categorical variables were compared using the Pearson χ^2 test or Fisher's exact test, as appropriate.

3.2 Logistic Regression and Multicollinearity Analyses

Univariable logistic regression analyses showed that age, sex, BMD, β-CTX, N-OC, tPINP, and 25-OH-D were all significantly associated with adjacent vertebral refracture, whereas BMI was not. After all eight predictors were simultaneously entered into the multivariable logistic regression model, older age (OR = 1.02, 95% CI, 1.01–1.04; P = 0.009) and male sex (OR = 2.54, 95% CI, 1.80–3.58; $P < 0.001$) remained associated with higher odds of refracture. Each 10 mg/cm³ increase in BMD was associated with lower odds of refracture (OR = 0.89, 95% CI, 0.84–0.95; $P < 0.001$), whereas each 100-unit increase in β-CTX (OR = 1.28, 95% CI, 1.21–1.34; $P < 0.001$) and each 5-unit increase in N-OC

(OR = 1.08, 95% CI, 1.05–1.12; $P < 0.001$) were associated with higher odds of refracture. Each 10-unit increase in tPINP was associated with lower odds of refracture (OR = 0.93, 95% CI, 0.88–0.98; $P = 0.011$), and each 5-unit increase in 25-OH-D was associated with an OR of 0.86 (95% CI, 0.79–0.93; $P < 0.001$). BMI remained nonsignificant in the multivariable model (OR = 1.03, 95% CI, 0.99–1.07; $P = 0.146$) **(Table 2;Figure 2).**

Assessment of multicollinearity showed that variance inflation factors (VIFs) for all predictors ranged from 1.037 to 1.470, with the maximum value well below the commonly used threshold of 5. Spearman correlation analysis identified β-CTX and tPINP as the most strongly correlated pair of predictors ($|\rho| = 0.458$), and all pairwise absolute correlation coefficients were <0.70, indicating no substantial multicollinearity overall (**Figure 3**). Nevertheless, the moderate correlation between β-CTX and tPINP may have contributed to the reversal in the direction of the tPINP coefficient in the multivariable model.

Notably, tPINP showed an apparently "protective" association after multivariable adjustment, which was opposite to the direction observed in the univariable analysis (OR ≈ 1.10) and to that generally reported in the existing literature. Therefore, the adjusted OR for tPINP should be interpreted as a conditional association after simultaneous adjustment for markers of bone resorption, rather than as evidence of an independent biological protective effect. Similarly, the higher odds of refracture observed in men than in women in the present study (OR = 2.54) should not be considered inconsistent with the generally higher fracture risk reported in women with osteoporosis. This finding may reflect the characteristics of this specific surgical cohort, in which male patients undergoing PVP may have had more severe disease, such as lower baseline BMD or more active bone turnover, and consequently a higher observed rate of adjacent vertebral refracture. Because potentially important confounders, including fracture severity and prior anti-osteoporosis treatment, were not included in the present analysis, the observed sex effect should be regarded as exploratory and requires validation in external datasets.

Table 2. Univariable and multivariable logistic regression analyses of factors associated with adjacent vertebral refracture.

Variable	Crude OR (95% CI)	Crude P value	Adjusted OR (95% CI)	Adjusted P value
Age, per 1 year	1.04 (1.03–1.06)	<0.001	1.02 (1.01–1.04)	0.009
Male vs Female	1.57 (1.16–2.11)	0.003	2.54 (1.80–3.58)	<0.001
BMI, per 1 unit	0.98 (0.95–1.01)	0.271	1.03 (0.99–1.07)	0.146
BMD, per 10 units	0.79 (0.75–0.82)	<0.001	0.89 (0.84–0.95)	<0.001
β-CTX, per 100 units	1.29 (1.24–1.34)	<0.001	1.28 (1.21–1.34)	<0.001
N-OC, per 5 units	1.14 (1.11–1.18)	<0.001	1.08 (1.05–1.12)	<0.001
tPINP, per 10 units	1.10 (1.06–1.15)	<0.001	0.93 (0.88–0.98)	0.011
25-OH-D, per 5 units	0.83 (0.76–0.90)	<0.001	0.86 (0.79–0.93)	<0.001

OR, odds ratio; CI, confidence interval; BMI, body mass index; BMD, bone mineral density; β-CTX, beta-C-terminal telopeptide of type I collagen; N-OC, N-mid osteocalcin; tPINP, total procollagen type I N-terminal propeptide; 25-OH-D, 25-hydroxyvitamin D. Crude ORs were obtained from univariable logistic regression models, whereas adjusted ORs were derived from the multivariable logistic regression model. $P < 0.05$ was considered statistically significant.

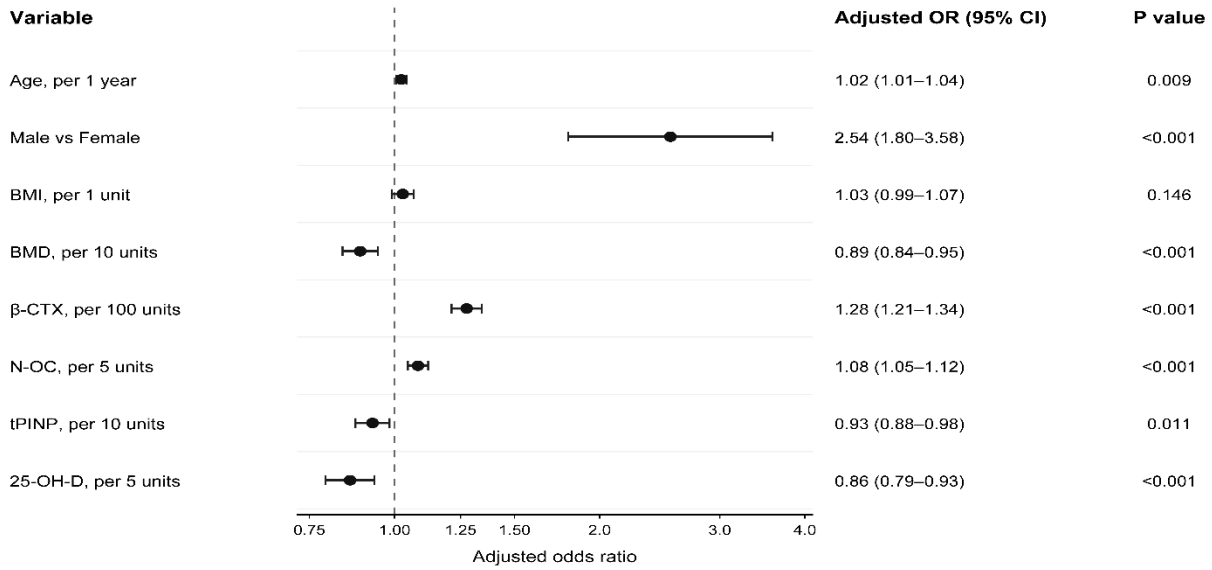


Figure 2**.** Multivariable logistic regression analysis of factors associated with adjacent vertebral refracture. Adjusted odds ratios (ORs) and corresponding 95% confidence intervals (CIs) are shown for the eight prespecified predictors.

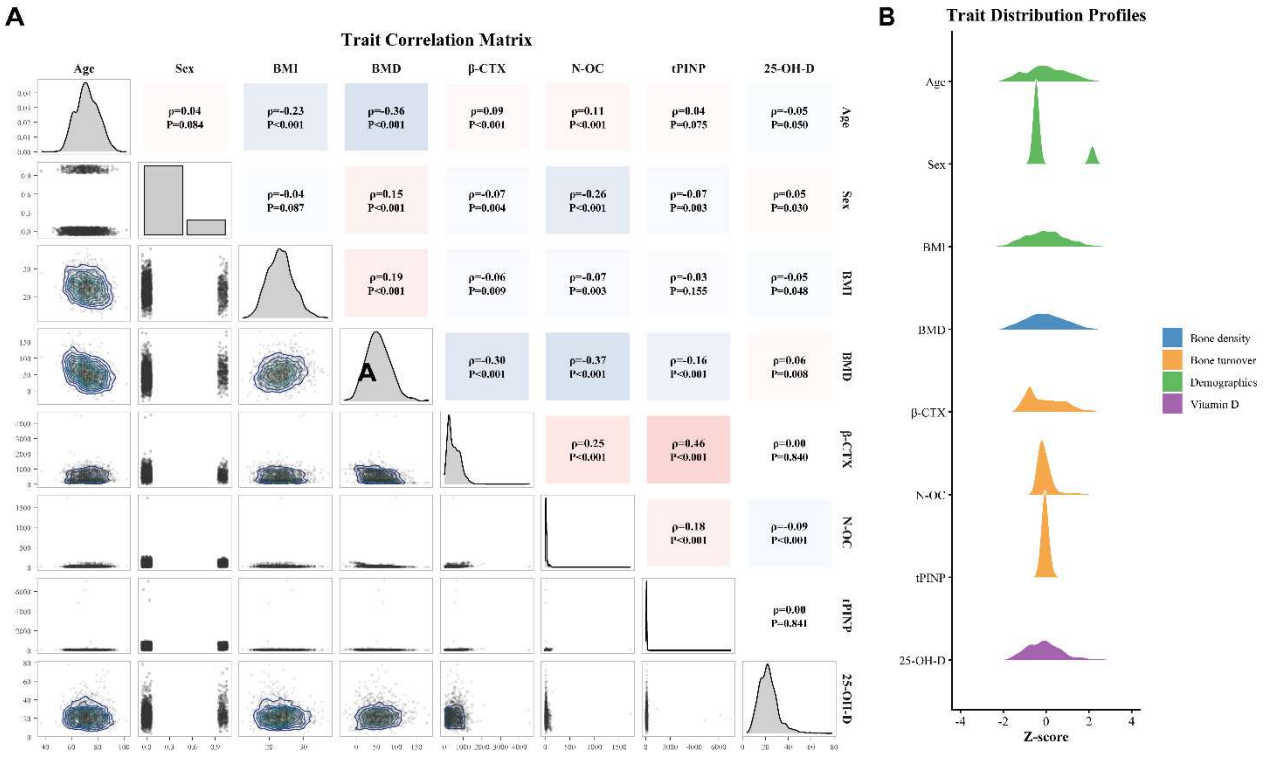


Figure 3. Correlation and distribution characteristics of the candidate predictors.(A) Spearman correlation matrix for the eight candidate predictors. Correlation coefficients (ρ) and corresponding *P* values are shown in the upper triangle; variable distributions and pairwise scatter/density patterns are displayed on the diagonal and in the lower triangle, respectively. (B) Standardized distribution profiles of the eight predictors based on Z-scores, grouped according to demographic, bone density, bone turnover, and vitamin D-related domains. No strong pairwise correlation was observed among the predictors.

3.3 Predictive Performance of the Machine Learning Models

The predictive performance of the six models was compared in the independent test cohort. ET achieved the numerically highest ROC-AUC of 0.872 (95% CI, 0.829–0.915), followed by RF [0.869 (95% CI, 0.826–

0.912)] and GB [0.866 (95% CI, 0.822–0.910)]. The AUCs for XGB, LR, and DT were 0.848, 0.806, and 0.760, respectively (**Figure 4**A; Table 3). ET also achieved the highest PR-AUC (0.644), followed by RF (0.638) (Table 3).

At the fixed classification thresholds derived from OOF predictions in the training cohort, RF achieved the highest accuracy (0.818) and F1 score (0.583), with a sensitivity of 0.688, specificity of 0.848, PPV of 0.506, and NPV of 0.923. GB achieved the highest sensitivity (0.844) and NPV (0.951). RF and ET both had a Brier score of 0.106, the lowest among all models (Table 3).

Pairwise DeLong tests showed that, after Benjamini-Hochberg (BH) adjustment, the AUCs of RF, ET, and GB were significantly higher than that of LR (BH-adjusted $P = 0.005$, 0.001, and 0.006, respectively) and were also significantly higher than that of DT (all $P < 0.001$). No statistically significant difference in AUC was observed between RF and ET (BH-adjusted $P = 0.709$) or between RF and GB (BH-adjusted $P = 0.771$). The difference between RF and XGB did not reach statistical significance after adjustment for multiple comparisons (BH-adjusted $P = 0.065$)（Supplementary Table S 2）.

Decision curve analysis (DCA) showed that, across a range of threshold probabilities, several models provided a positive net benefit relative to the treat-all and treat-none strategies, with RF demonstrating a consistently favorable net benefit (**Figure 4**B). Calibration of the RF model was evaluated using 1,000 bootstrap resamples for bias correction. The RF model had a Brier score of 0.106, a calibration intercept of 0.076, and a calibration slope of 1.640. The apparent and bootstrap bias-corrected calibration curves are shown in **Figure 4**C.

Although ET achieved the numerically highest AUC in the independent test cohort, its AUC did not differ significantly from that of RF. In comparison, RF achieved the highest accuracy and F1 score while sharing the lowest Brier score with ET (Table 3). Considering discrimination, threshold-dependent classification performance, overall probabilistic prediction error, and the requirements for subsequent model interpretation, RF was selected as the primary model for subsequent SHAP interpretation, subgroup analyses, and sensitivity

analyses.

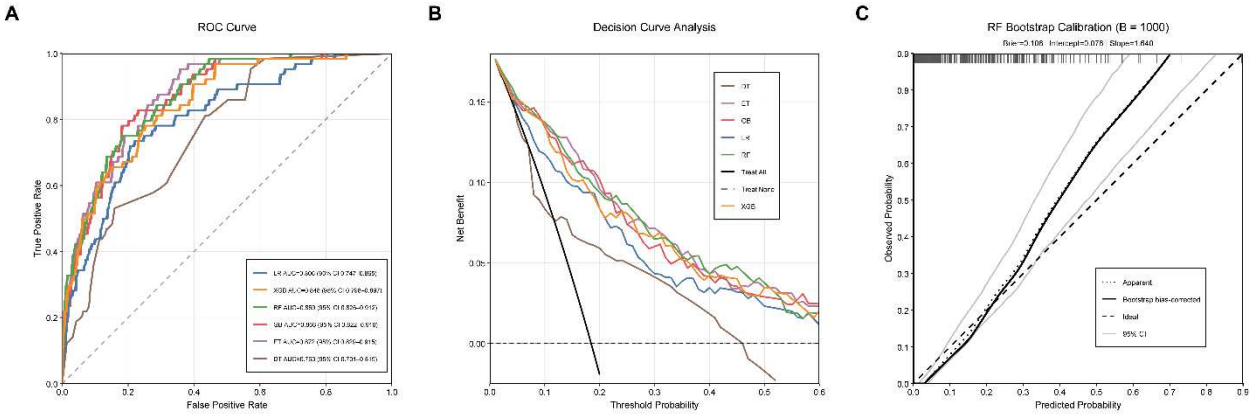


Figure 4. Discrimination, clinical utility, and calibration of the machine-learning models.**(A)** Receiver operating characteristic (ROC) curves of the six prediction models in the independent test cohort. **(B)** Decision curve analysis (DCA) comparing the net benefit of the six models across a range of threshold probabilities. **(C)** Calibration of the random forest model using 1,000 bootstrap resamples, showing the apparent, bootstrap bias-corrected, and ideal calibration curves. RF, random forest; LR, logistic regression; DT, decision tree; ET, extra trees; GB, gradient boosting; XGB, extreme gradient boosting.

Table 3**. Predictive performance of the six machine learning models in the test cohort.**

Model	AUC (95% CI)	PR-AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 Score	Brier Score
LR	0.806 (0.747–0.865)	0.532	0.755	0.734	0.760	0.409	0.927	0.525	0.118
DT	0.760 (0.701–0.819)	0.429	0.666	0.625	0.675	0.303	0.888	0.408	0.145
RF	0.869 (0.826–0.912)	0.638	0.818	0.688	0.848	0.506	0.923	0.583	0.106
ET	0.872 (0.829–0.915)	0.644	0.795	0.719	0.813	0.465	0.927	0.564	0.106
GB	0.866 (0.822–0.910)	0.614	0.720	0.844	0.693	0.383	0.951	0.527	0.108
XGB	0.848 (0.798–0.897)	0.608	0.700	0.828	0.671	0.363	0.945	0.505	0.110

Abbreviations: AUC, area under the receiver operating characteristic curve; PR-AUC, area under the precision-recall curve; CI, confidence interval; LR, logistic regression; DT, decision tree; RF, random forest; ET, extra trees; GB, gradient boosting; XGB, extreme gradient boosting; PPV, positive predictive value; NPV, negative predictive value. Classification metrics were calculated using model-specific cutoffs determined by the Youden index from out-of-fold predictions in the training cohort. Brier scores were calculated from predicted probabilities.

3.4 Interpretability Analysis of the RF Model

SHAP analysis demonstrated substantial differences in the contributions of individual predictors to the RF model output. Based on global SHAP importance, β-CTX had the greatest predictive contribution, followed by BMD, N-OC, and 25-OH-D, whereas sex, age, tPINP, and BMI made relatively smaller overall contributions (Figure 5A). The distribution of SHAP values further showed that higher β-CTX levels generally shifted model predictions toward a higher probability of refracture, whereas higher BMD generally shifted predictions toward a lower probability of refracture. Male sex also predominantly contributed in the direction of higher predicted risk (**Figure 5**A).

SHAP interaction analysis showed that β-CTX had the largest mean absolute main-effect SHAP value (0.091), while the interaction between β-CTX and tPINP showed a relatively large contribution (mean |SHAP interaction| = 0.021) (**Figure 5**B). These findings indicate that predictions generated by the RF model reflect both individual predictor effects and interactions between predictors.

Patient-level SHAP analysis further illustrated how individual predictors contributed to the model output for specific patients. For example, the representative low-risk patient (ID 1259) had a predicted refracture probability of 0.1%, with lower β-CTX and higher BMD being the major factors shifting the predicted probability downward (**Figure 5**C). Patient-level model output trajectories and distributions of SHAP contributions further demonstrated that predicted probabilities were jointly determined by multiple predictors

and that the direction and magnitude of contribution of the same predictor varied across individual patients (**Figure 5**D–E).

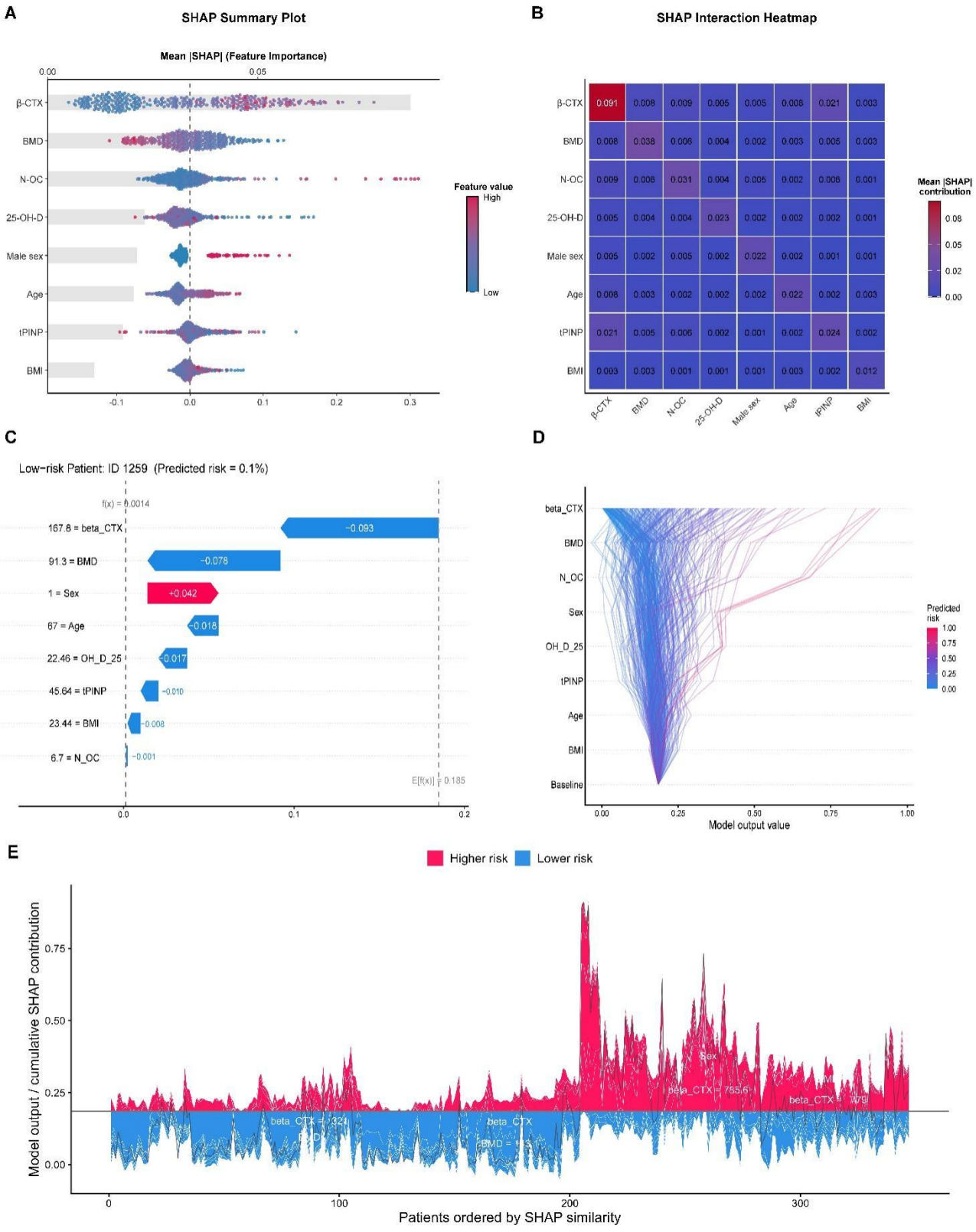


Figure 5. SHAP-based interpretation of the random forest model.(A) SHAP summary plot showing global feature

importance and the direction of individual feature contributions to model predictions. (B) SHAP interaction heatmap showing the magnitude of main and pairwise interaction effects. (C) Individual SHAP explanation for a representative low-risk patient (ID 1259). (D) Patient-level model-output trajectories across predictors. (E) Distribution of cumulative SHAP contributions across patients. Positive SHAP values indicate contributions toward a higher predicted probability of adjacent vertebral refracture, whereas negative values indicate contributions toward a lower predicted probability.

3.5 Nonlinear Associations of Continuous Bone-Related Variables

To further characterize the shapes of the associations between the prespecified continuous bone-related variables and adjacent vertebral refracture, restricted cubic spline (RCS) analyses were performed for β-CTX, BMD, N-OC, tPINP, and 25-OH-D.

β-CTX showed a significant overall association with adjacent vertebral refracture as well as a significant nonlinear relationship (P for overall < 0.001; P for nonlinearity < 0.001). BMD likewise demonstrated both a significant overall association and significant nonlinearity (P for overall < 0.001; P for nonlinearity = 0.003). For tPINP, both the overall association and the test for nonlinearity were statistically significant (P for overall = 0.002; P for nonlinearity = 0.016).

N-OC was significantly associated with refracture overall (P for overall < 0.001), but no significant evidence of nonlinearity was observed (P for nonlinearity = 0.688). Similarly, 25-OH-D showed a significant overall association with refracture (P for overall = 0.001), whereas the test for nonlinearity was not statistically significant (P for nonlinearity = 0.372) **(Figure 6).**

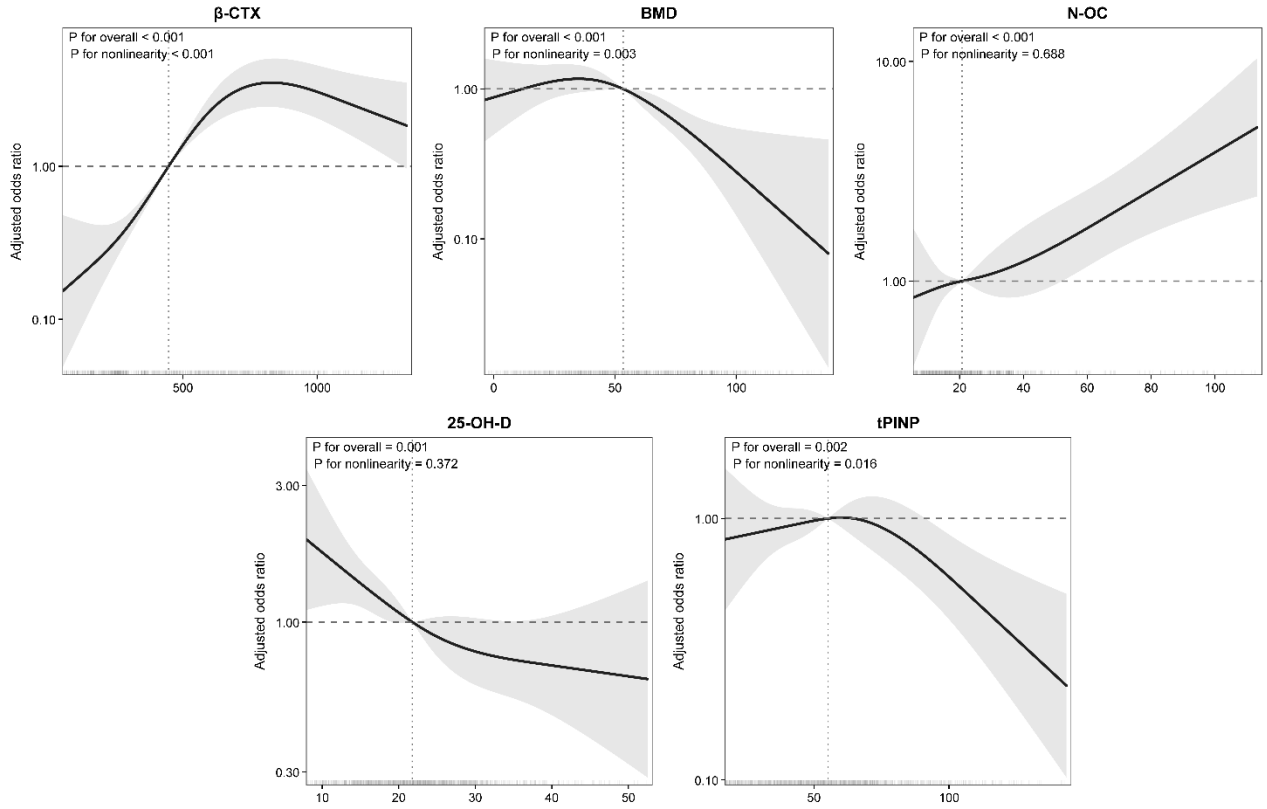


Figure 6. Restricted cubic spline analyses of bone-related continuous variables and adjacent vertebral refracture. Restricted cubic splines were fitted for β-CTX, BMD, N-OC, 25-OH-D, and tPINP using four knots at the 5th, 35th, 65th, and 95th percentiles. Each model was adjusted for the remaining seven predictors. Solid lines represent adjusted ORs and shaded areas indicate 95% CIs. The horizontal dashed line denotes an OR of 1.0. *P* values for overall association and nonlinearity are shown in each panel.

3.6 Subgroup and Interaction Analyses

The discriminative performance of the RF model was further evaluated across prespecified subgroups in the independent test cohort. The female subgroup included 275 patients with 50 refracture events and had an AUC of 0.848 (95% CI, 0.796–0.900), whereas the male subgroup included 72 patients with 14 events and had an AUC of 0.964 (95% CI, 0.927–1.000). The between-group difference in AUC was 0.116 and remained statistically significant after Benjamini-Hochberg (BH) adjustment ($P < 0.001$). In the subgroup with β-CTX <468.4, the RF model achieved an AUC of 0.941 (95% CI, 0.889–0.992), compared with 0.755 (95% CI, 0.674–0.836) in the subgroup with β-CTX ≥468.4. The corresponding AUC

difference was 0.186 (BH-adjusted $P < 0.001$) (**Figure 7;Supplementary Table S3)**.

In contrast, no statistically significant differences in AUC were observed between the age subgroups (<75 vs ≥75 years: AUC, 0.884 vs 0.834), BMI subgroups (<24 vs ≥24 kg/m²: AUC, 0.851 vs 0.896), or 25-OH-D subgroups (<18.28 vs ≥18.28: AUC, 0.908 vs 0.854), with BH-adjusted $P = 0.308$ for all comparisons (Figure 7; **Supplementary Table S3**).

Exploratory interaction analyses showed a statistically significant interaction between β-CTX and BMD (interaction OR = 1.03, 95% CI, 1.01–1.04; raw $P = 0.002$; BH-adjusted $P = 0.011$). A significant interaction was also observed between β-CTX and 25-OH-D (interaction OR = 0.97, 95% CI, 0.94–0.99; raw $P = 0.005$; BH-adjusted $P = 0.012$). The interaction between β-CTX and sex was statistically significant before multiple-comparison adjustment (raw $P = 0.034$) but did not remain significant after BH correction (BH-adjusted $P = 0.056$). No significant interactions were observed between β-CTX and age or BMI (BH-adjusted $P = 0.474$ and 0.783, respectively) **(**Supplementary Table S4**).**

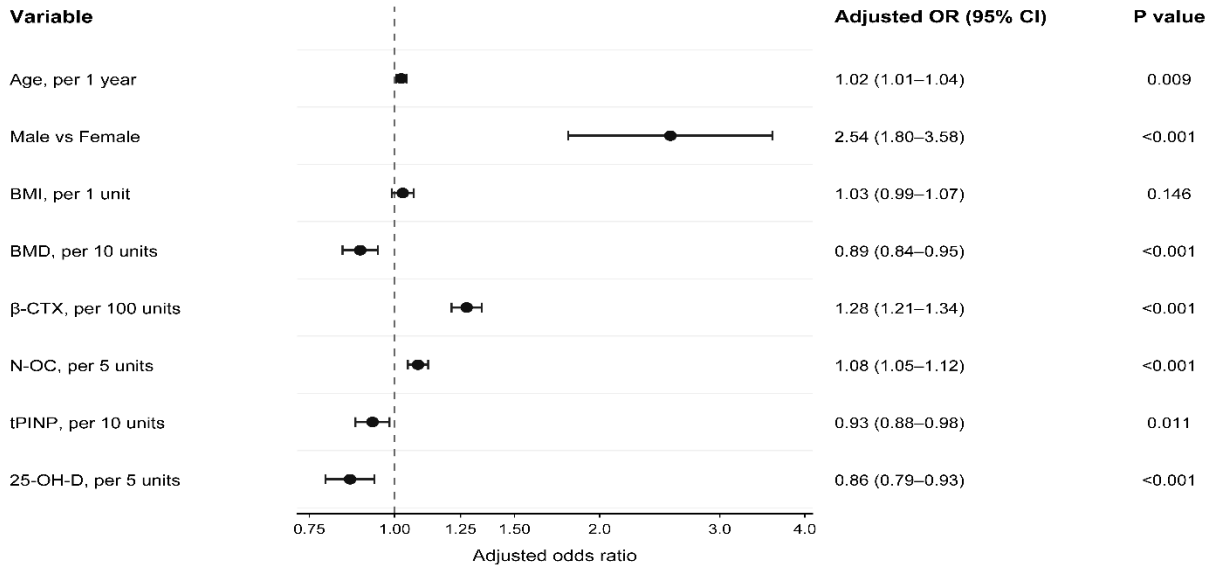


Figure 7.Subgroup performance of the random forest model in the independent test cohort**.**AUCs and corresponding 95% confidence intervals are shown for prespecified subgroups according to sex, age, BMI, β-CTX, and 25-OH-D. The numbers of patients and refracture events are presented for each subgroup. P values represent between-subgroup differences in AUC and were adjusted for multiple comparisons using the Benjamini–Hochberg procedure. AUC, area under the receiver operating characteristic curve; CI, confidence interval; BMI, body mass index; β-CTX, beta-C-terminal

telopeptide of type I collagen; 25-OH-D, 25-hydroxyvitamin D.

3.7 Sensitivity Analyses

Multiple sensitivity analyses generally supported the robustness of the main findings for the RF model. First, in the feature-set restriction analysis, the RF model incorporating all eight predictors achieved an AUC of 0.862 (95% CI, 0.817–0.906) with a Brier score of 0.107. When the model was restricted to the five bone-related variables (BMD, β-CTX, N-OC, tPINP, and 25-OH-D), the AUC was 0.848 (95% CI, 0.801–0.896) and the Brier score remained 0.107. Compared with the full model, the AUC decreased by 0.013, and the difference was not statistically significant (DeLong $P = 0.339$). In contrast, the basic clinical model including only age, sex, and BMI yielded an AUC of 0.570 (95% CI, 0.494–0.646) with a Brier score of 0.163. Relative to the full model, the AUC decreased by 0.292, and the difference was statistically significant (BH-adjusted $P < 0.001$) **(**Supplementary Figure S 1**).**

Second, continuous predictors were Winsorized at the 1st and 99th percentiles, and the RF model was refitted within the same sensitivity-analysis framework. After Winsorization, the model achieved an AUC of 0.867 (95% CI, 0.824–0.910) and a Brier score of 0.106. By comparison, the corresponding non-Winsorized RF model had an AUC of 0.862 (95% CI, 0.817–0.906) and a Brier score of 0.107. The AUC difference was 0.005 and did not reach statistical significance by DeLong testing ($P = 0.078$) **(**Supplementary Table S5**).**

Finally, different operating thresholds were determined from OOF predicted probabilities in the training cohort and then fixed for application to the independent test cohort. The OOF Youden threshold was 0.243, yielding a sensitivity of 0.688, specificity of 0.848, PPV of 0.506, NPV of 0.923, accuracy of 0.818, F1 score of 0.583, and balanced accuracy of 0.768 in the independent test cohort. When a threshold of 0.167 was selected to achieve an OOF sensitivity of at least 80% in the training cohort, sensitivity in the test cohort increased to 0.859 and NPV to 0.955, whereas specificity decreased to 0.671; accuracy and F1 score were 0.706 and 0.519, respectively. When a threshold of 0.120 was selected to achieve an OOF sensitivity of at least 90%, test-cohort sensitivity further increased to 0.969 and NPV to 0.988, while specificity

decreased to 0.562; accuracy and F1 score were 0.637 and 0.496, respectively. The balanced accuracies for the three threshold strategies were 0.768, 0.765, and 0.765, respectively **(**Supplementary Table S6**).**

4. Discussion

4.1 Main Findings and Clinical Context

In this study of 1,730 patients with OVCFs who underwent PVP, we developed and compared six machine learning models using eight readily available clinical variables and further evaluated their predictive performance, interpretability, and robustness through SHAP analysis, restricted cubic spline analysis, subgroup analyses, and multiple sensitivity analyses. The principal findings were as follows. First, the incidence of adjacent vertebral refracture was 18.3%, and age, sex, BMD, β-CTX, N-OC, tPINP, and 25-OH-D were independently associated with refracture; significant nonlinear associations were observed for β-CTX, BMD, and tPINP. Second, ET achieved the numerically highest AUC, whereas RF showed a more balanced overall performance in terms of AUC, classification performance, Brier score, and clinical net benefit. Because the difference in AUC between RF and ET was not statistically significant, RF was selected for subsequent model interpretation. Third, SHAP analysis identified β-CTX, BMD, and N-OC as the major drivers of model predictions. Fourth, restricting the model to the five bone-related variables resulted in only a modest reduction in discrimination, whereas a model including only age, sex, and BMI showed substantially poorer performance. Finally, sensitivity analyses using Winsorization and alternative classification thresholds generally supported the robustness of the model findings.

The 18.3% incidence of adjacent vertebral refracture observed in the present study was broadly consistent with the burden of secondary vertebral fractures reported in previous studies. A 2025 systematic review and meta-analysis including 3,821 patients treated with PVP reported an overall incidence of secondary vertebral fracture of 16.6%, whereas Tang et al. reported an incidence of adjacent vertebral fracture of 9.1% among 1,077 patients

undergoing PVP; previous large cohorts have reported rates ranging from approximately 7.3% to 29.0%(12)(13).Differences in incidence across studies may be attributable to variation in outcome definitions, patient selection, and duration of follow-up. In the present study, older age was associated with a higher risk of refracture, consistent with the findings of Tang et al., who reported that advanced age was associated with a greater likelihood of adjacent vertebral fracture(12).Lower BMD was also a consistent risk indicator in our cohort, in agreement with recent meta-analyses and findings from other PVP and percutaneous kyphoplasty cohorts(13)(14).

Notably, in the multivariable analysis, men had higher odds of refracture than women (OR = 2.54), a finding that is inconsistent with a recent meta-analysis identifying female sex as a risk factor for secondary fractures(13).Therefore, the sex effect observed in the present study is more appropriately interpreted as a conditional association within this specific surgical cohort and should not be taken to imply that men have a higher fracture risk than women in the general population with osteoporosis. Patient selection, the restriction of the outcome to adjacent vertebral fractures, and residual confounding from unmeasured factors may all have contributed to this association.

4.2 Unique Predictive Value of Bone Metabolism-Related Markers

Bone-related variables demonstrated particularly important predictive value in the present study. Lumbar spine BMD measured by QCT was inversely associated with refracture risk, and RCS analysis indicated a nonlinear relationship. Previous systematic reviews have shown that volumetric BMD derived from QCT can effectively distinguish patients with osteoporosis with and without vertebral fractures, and its association with vertebral fracture risk may be stronger than that of areal BMD measured by dual-energy X-ray absorptiometry (DXA)(15).More notably, β-CTX was not only significantly and positively associated with refracture risk in the multivariable logistic regression analysis but also showed the highest overall importance in the SHAP analysis and a pronounced nonlinear association with the outcome. β-CTX reflects bone resorption activity. Previous

meta-analyses and the most recent ESCEO/IOF/IFCC consensus have shown that higher β-CTX levels are independently, although moderately, associated with future fracture risk, and β-CTX and PINP continue to be recommended as reference bone turnover markers in osteoporosis research(2)(16).Our findings further suggest that, in patients undergoing PVP, combining information on bone turnover with BMD may be more informative for identifying the risk of adjacent vertebral refracture than relying on demographic characteristics alone.

The findings for tPINP require more cautious interpretation. Although unadjusted tPINP levels were higher in patients who developed refracture, after simultaneous adjustment for the other seven variables, higher tPINP was associated with lower odds of refracture; RCS analysis also demonstrated a significant nonlinear association with the outcome. Previous studies have generally suggested that higher PINP levels are associated with increased fracture risk(2)(16)， Therefore, the inverse association observed after multivariable adjustment in the present study should not be interpreted as evidence that increasing tPINP has a protective effect. Bone formation and bone resorption are intrinsically coupled components of the bone remodeling process, and β-CTX and tPINP were moderately correlated in our cohort ($|\rho| = 0.458$). When multiple bone turnover markers are entered simultaneously into a model in the presence of nonlinear relationships, regression coefficients represent conditional associations rather than isolated biological effects of individual markers. The latest international consensus likewise notes that interactions between bone turnover markers and other fracture risk factors remain incompletely understood(2).Accordingly, the directional finding for tPINP should be regarded as a signal requiring further prospective validation. Similarly, N-OC was independently associated with refracture and ranked among the leading predictors in the SHAP analysis; however, the standardized evidence supporting N-OC as an individual fracture prediction marker remains less established than that for β-CTX and PINP and therefore warrants validation in external cohorts.

4.3 Selection and Performance of the Machine Learning Models

With respect to model performance, ET achieved the highest AUC (0.872), whereas RF achieved an AUC of

0.869, with no statistically significant difference between the two models. RF also demonstrated the highest accuracy and F1 score and achieved a low Brier score, supporting its selection for subsequent model interpretation. Jiang et al. previously developed a random survival forest model incorporating MRI radiomic features and clinical factors to predict new vertebral fractures after vertebral augmentation, with a 2-year time-dependent AUC of 0.839 in the external validation cohort, suggesting that machine learning can capture complex risk patterns that may be difficult to characterize using conventional approaches (17).However, direct comparisons based solely on AUC should be avoided because outcome definitions, follow-up strategies, and validation designs differ across studies. A 2026 systematic review further showed that, although existing machine learning studies of refracture after treatment for OVCF generally reported favorable predictive performance, most were limited by retrospective designs, relatively small sample sizes, lack of external validation, and incomplete reporting(17).The relatively large sample size, complete locked dataset, and independent hold-out test cohort in the present study improved the stability of model evaluation to some extent, but they do not substitute for true multicenter external validation.

4.4 Interpretability and the Clinical Value of a Concise Predictor Set

The interpretability analyses further highlighted the potential value of using a concise set of routinely available predictors. SHAP analysis not only identified β-CTX, BMD, and N-OC as the major drivers of the RF model but also demonstrated that the direction and magnitude of contribution of the same predictor varied across individual patients. By decomposing complex model predictions into feature-specific contributions, SHAP has become an important approach for interpreting tree-based models. More broadly, the value of explainable artificial intelligence in health care lies in supporting both predictive performance and clinical interpretability(18).More importantly, in the feature-set sensitivity analysis, restricting the model to the five bone-related variables reduced the AUC only modestly, from 0.862 to 0.848, and the difference was not statistically significant (P = 0.339), whereas a model including only age, sex, and BMI yielded an AUC of

0.570. These findings suggest that the major discriminative information in the model was derived from BMD and bone metabolic status, and they support the feasibility of risk stratification using a limited number of relatively accessible bone-related variables. However, these results should not be interpreted as indicating that demographic characteristics have no predictive value; rather, they suggest that demographic variables alone are insufficient to achieve satisfactory individualized prediction.

The subgroup and threshold sensitivity analyses provided additional information regarding potential model application. No significant differences in AUC were observed across subgroups defined by age, BMI, or 25-OH-D, suggesting that the discriminative performance of the RF model was generally stable across these patient characteristics. In contrast, the AUC was higher in men than in women and higher in patients with lower β-CTX than in those with higher β-CTX. These findings should be interpreted cautiously. The male subgroup in the test cohort included only 72 patients and 14 refracture events, and the relatively high AUC may therefore have been influenced by the limited sample size and number of events. Likewise, the difference in AUC between β-CTX subgroups reflects heterogeneity in model discrimination rather than evidence of differential causal effects of β-CTX on refracture risk. Accordingly, these subgroup findings should be regarded as exploratory.

In addition, the OOF threshold analysis showed that shifting the operating threshold from the Youden-based strategy toward sensitivity-oriented thresholds increased sensitivity in the test cohort from 0.688 to 0.859 and 0.969, with corresponding NPVs of 0.955 and 0.988, while specificity decreased to 0.671 and 0.562, respectively. This trade-off suggests that, if the model is used in the future for postoperative screening of high-risk patients, a more sensitivity-oriented decision threshold may be selected according to the intended clinical objective to reduce missed high-risk cases. However, the resulting increase in false-positive classifications and health care resource utilization will require evaluation in prospective clinical utility studies.

559 **4.5 Strengths and Limitations**

560 This study has several strengths. First, the cohort included 1,730 patients, representing a relatively large sample
561 compared with many existing machine learning studies of refracture in patients with OVCF and providing a
562 stable data foundation for model development and performance evaluation. Second, only eight routinely
563 available variables encompassing demographic characteristics, QCT-derived BMD, and circulating bone
564 metabolism markers were used, thereby avoiding reliance on high-dimensional radiomic features or extensive
565 laboratory testing. This approach may reduce data acquisition burden and computational complexity while
566 enhancing the potential generalizability of the model. Third, model performance was evaluated
567 comprehensively. In addition to the conventional ROC-AUC metric, we reported PR-AUC, Brier score,
568 calibration curves, and decision curve analysis. Bootstrap bias correction, restricted cubic spline analysis,
569 SHAP-based interpretation, subgroup analyses, and multiple sensitivity analyses were also performed to assess
570 predictive performance, nonlinear associations, model interpretability, and robustness from multiple
571 perspectives, in accordance with current TRIPOD+AI recommendations for transparent reporting of machine
572 learning prediction models
573 (7).

574 Several important limitations should also be acknowledged. First, adjacent vertebral refracture was defined as
575 an event that was confirmed by imaging and led to rehospitalization for treatment. This definition may have
576 failed to capture asymptomatic or mildly symptomatic refractures in patients who did not seek medical
577 attention, potentially leading to underestimation of the true event rate and introducing outcome ascertainment
578 bias. Second, follow-up time was not incorporated into the modeling framework. Outcomes observed over
579 varying follow-up durations were simplified into a binary endpoint; consequently, the current model estimates
580 cumulative risk over the observation period rather than risk at a prespecified time point, such as 1 year after
581 PVP. Future studies should therefore consider time-to-event approaches, including Cox proportional hazards

models or random survival forests, to provide more precise and dynamic risk estimates.

With respect to calibration, although the RF model showed a low Brier score (0.106) and good discrimination (AUC, 0.869), the calibration slope was 1.640, which deviated substantially from the ideal value of 1 and indicated imperfect calibration of the predicted probabilities. Accordingly, the current tool may be more appropriate for relative risk stratification than for precise estimation of individual absolute risk. External validation and, if necessary, recalibration are required before the model is used for individual-level probability estimation in clinical practice.

In addition, this was a single-center retrospective study, and the randomly partitioned test cohort still represents internal rather than external validation. Several potentially important covariates were not available, including bone cement distribution, vertebral morphology, adherence to anti-osteoporosis therapy, renal function, and history of falls. Renal function is particularly relevant because it may influence circulating β-CTX and tPINP levels and thereby confound the associations between bone turnover markers and refracture

(2).Furthermore, the retrospective nature of the data precluded complete standardization of blood sampling time, fasting status, and other preanalytical conditions for bone turnover marker measurements, which may have introduced additional measurement variability. Taken together, these limitations indicate that the present model should currently be regarded as a candidate tool for risk stratification, and multicenter prospective external validation and recalibration are required before broader clinical implementation.

4.6 Conclusion

In this study, we developed and internally validated an explainable random forest-based risk stratification model using eight routinely available clinical and bone metabolism-related variables to identify patients at increased risk of adjacent vertebral refracture after PVP for OVCFs. The model demonstrated good discrimination in the independent test cohort (AUC, 0.869) and a favorable clinical net benefit, with β-CTX, BMD, and N-OC identified as the major contributors to model predictions.

However, given the single-center retrospective design and imperfect calibration (calibration slope, 1.640), the model should currently be considered a tool for postoperative risk stratification rather than a tool for precise estimation of individual absolute risk. Future multicenter prospective studies are required to externally validate and recalibrate the model. The model may assist clinicians in identifying high-risk patients and optimizing secondary fracture prevention strategies.

Declarations

Ethics Approval and Informed Consent

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of our institution on November 26, 2025 (Approval No.: 2024-08-C024). Due to the retrospective nature of the study, the requirement for informed consent was waived by the ethics committee.

Trial number: Not applicable.

Consent for Publication

Not applicable.

Availability of Data and Materials

The datasets generated and/or analyzed during the current study are not publicly available because they contain patient-level clinical information. De-identified data may be available from the corresponding author upon reasonable request and approval from the institutional ethics committee.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

E.S. Z.X. , and G.G. contributed to the study conception, design, and data collection. H.W. and Q.L. supervised the study, provided critical revisions, and approved the final manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

The authors would like to thank the Department of Spine Surgery of our institution for their support in data collection and clinical follow-up.

References

[1] Jang HD, Kim EH, Lee JC, et al. Management of osteoporotic vertebral fracture: Review update 2022. Asian Spine J. (2022);16:934-946. doi:10.31616/asj.2022.0441

[2] Bhattoa HP, Vasikaran S, Trifonidi I, et al. Update on the role of bone turnover markers in the diagnosis and management of osteoporosis: A consensus paper from the european society for clinical and economic aspects of osteoporosis, osteoarthritis and musculoskeletal diseases (ESCEO), international osteoporosis foundation (IOF), and international federation of clinical chemistry and laboratory medicine (IFCC). Osteoporos Int. (2025);36:579-608. doi:10.1007/s00198-025-07422-3

[3] Shi J, Xiao J, Ai Y, et al. Machine learning-based risk predictive model for postoperative refractures in patients with osteoporotic vertebral compression fractures: A systematic review and critical appraisal. BMC musculoskelet disord. (2026);27:394. doi:10.1186/s12891-026-09759-0

[4] Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clin Transl Sci.

(2024);17:e70056. doi:10.1111/cts.70056

[5] Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: A step-by-step guide. BMJ. (2024);386:e078276. doi:10.1136/bmj-2023-078276

[6] Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. (2017);77:1-17. doi:10.18637/jss.v077.i01

[7] TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. (2024);385:q902. doi:10.1136/bmj.q902

[8] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. (1988);44:837-845. doi:10.2307/2531595

[9] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc: B (Methodol). (1995);57:289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

[10] Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Mak: Int J Soc Med Decis Mak. (2006);26:565-574. doi:10.1177/0272989X06295361

[11] Web-based calculator for adjacent vertebral refracture risk after percutaneous vertebroplasty. Posit Connect. Accessed August 31, 2026. https://ovcf-risk-calculator.share.connect.posit.cloud/

[12] Tang B, Liu L, Cui L, et al. Analysis of adjacent vertebral fracture after percutaneous vertebroplasty: Do radiological or surgical features matter? Eur Spine J. (2024);33:1524-1532. doi:10.1007/s00586-023-08092-7

669 [13] Marselou E, Kelekis A, Dimitriadis Z, Koumantakis GA. Risk factors for refracture or new vertebral
670 compression fractures after percutaneous vertebroplasty: A systematic review and meta-analysis.
671 Osteoporos Int. (2025);36:1297-1311. doi:10.1007/s00198-025-07523-z

672 [14] Yang W, Zou K, Lin X, et al. Risk factors for new vertebral fractures after percutaneous vertebroplasty
673 or percutaneous kyphoplasty in the treatment of osteoporotic vertebral compression fractures. Front
674 Med. (2025);12. doi:10.3389/fmed.2025.1514894

675 [15] Chen L, Wu X yi, Jin Q, Chen G yao, Ma X. The correlation between osteoporotic vertebrae fracture
676 risk and bone mineral density measured by quantitative computed tomography and dual energy X-ray
677 absorptiometry: A systematic review and meta-analysis. Eur Spine J. (2023);32:3875-3884.
678 doi:10.1007/s00586-023-07917-9

679 [16] Tian A, Ma J, Feng K, et al. Reference markers of bone turnover for prediction of fracture: a meta-
680 analysis. J Orthop Surg Res. (2019);14:68. doi:10.1186/s13018-019-1100-6

681 [17] Jiang Y, Cai J, Zeng Y, et al. Development and validation of a machine learning model to predict
682 imminent new vertebral fractures after vertebral augmentation. BMC Musculoskelet Disord.
683 (2023);24:472. doi:10.1186/s12891-023-06557-w

684 [18] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial
685 intelligence for healthcare: A systematic review of the last decade (2011-2022). Comput Methods
686 Programs Biomed. (2022);226:107161. doi:10.1016/j.cmpb.2022.107161

687

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- Table2.docx
- Table3.docx
- SupplementaryTableS1.docx
- SupplementaryTableS2.docx
- SupplementaryTableS3.docx
- SupplementaryTableS4.docx
- SupplementaryTableS5.docx
- SupplementaryTableS6.docx