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.

Development and Internal Validation of a Diagnostic Model for Fragility Vertebral Fractures in Hospitalized Patients With Thoracolumbar Back Pain

Jinchang Zheng
Affiliated Hospital of Guangdong Medical University
Haoyuan Luo
Affiliated Hospital of Guangdong Medical University
Chuhui Zheng
Affiliated Hospital of Guangdong Medical University
Xiaotong Chen
Affiliated Hospital of Guangdong Medical University
Langfan Xu
Affiliated Hospital of Guangdong Medical University
Xiaoen Zhuang
Affiliated Hospital of Guangdong Medical University
Chanli Zhong
Affiliated Hospital of Guangdong Medical University
Zhen Liang
liangzhen1@gdmu.edu.cn

Affiliated Hospital of Guangdong Medical University

Research Article

Keywords: Fragility vertebral fracture, Thoracolumbar back pain, Diagnostic prediction model, Bone mineral density, Penalized regression

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11577086/v1

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

Additional Declarations: No competing interests reported.

Development and Internal Validation of a Diagnostic Model for Fragility Vertebral Fractures in Hospitalized Patients With Thoracolumbar Back Pain

Jinchang Zheng[1#], Haoyuan Luo[1#] , Chuhui Zheng[1], Xiaotong Chen[1], Langfan Xu[1], Xiaoen Zhuang[1], Chanli Zhong[2], Zhen Liang[1*]

[1]Department of Orthopedic Center, Affiliated Hospital of Guangdong Medical University, Zhanjiang 524001, China.

[2]Department of Clinical Laboratory, Affiliated Hospital of Guangdong Medical University, Zhanjiang 524001, China.

[#]These authors contributed equally to this work.

***Corresponding author**

Zhen Liang, PhD

Department of Orthopedic Center, Affiliated Hospital of Guangdong Medical University, Zhanjiang, Guangdong, China

E-mail: liangzhen1@gdmu.edu.cn;

Tel: +86-759-2387291

ORCID: 0009-0004-0847-5489

Abstract

Background: Thoracolumbar back pain has several causes, and routine measurements alone do not establish whether a symptomatic fracture is present. We developed and internally evaluated a diagnostic model for symptomatic thoracolumbar fragility vertebral fractures in hospitalized patients.

Methods: This retrospective study used routine clinical records of men aged at least

50 years and postmenopausal women admitted with thoracolumbar back pain to the Affiliated Hospital of Guangdong Medical University between January 1, 2023, and September 30, 2026. Fracture status during the index hospitalization was determined from imaging and clinical history. Fifteen candidate predictors entered a group least absolute shrinkage and selection operator logistic model with three-knot restricted cubic splines for continuous predictors. The full cohort was used for development. Internal validation comprised 1000 patient-level bootstrap resamples repeating spline construction, cross-validation, penalty selection, predictor-group selection, and fitting. We assessed discrimination, calibration, prediction stability, and exploratory net benefit. The study was not registered.

Results: The cohort included 1414 patients, of whom 571 (40.4%) had the target fracture outcome. The 12 retained predictors were sex/eligibility stratum, age, lumbar spine and femoral neck bone mineral density, calcium, fasting plasma glucose, triglycerides, low-density lipoprotein cholesterol, 25-hydroxyvitamin D3, osteocalcin, C-terminal telopeptide of type I collagen, and procollagen type I N-terminal propeptide. Apparent area under the receiver operating characteristic curve was 0.822 (95% confidence interval, 0.800 – 0.844). Bootstrap-corrected concordance index, Brier score, and calibration slope were 0.818, 0.171, and 1.218, respectively. Age and lumbar spine bone mineral density were always retained during resampling; biochemical predictors were less consistently selected. The mean absolute difference between bootstrap-model and original-model probabilities was 4.0 percentage points. Corrected net-benefit point estimates exceeded prioritizing all or no patients for

imaging at evaluated thresholds of 0.10 – 0.80.

Conclusions: External validation and prospective assessment are required before use for magnetic resonance imaging prioritization; no clinical cutoff was established. This study adds a complete-pipeline assessment of individual prediction stability to diagnostic modeling in a symptom-defined inpatient cohort.

Keywords: Fragility vertebral fracture; Thoracolumbar back pain; Diagnostic prediction model; Bone mineral density; Penalized regression.

Background

Fragility vertebral fractures can cause pain and functional impairment, and their recognition should prompt assessment and management of underlying osteoporosis [1,2]. In Chinese cities, insurance claims data show that the incidence of clinically diagnosed vertebral fractures and medical expenses among adults aged 50 and above have been on the rise from 2013 to 2017 [3]. For patients hospitalized due to thoracolumbar back pain, the primary concern is whether vertebral fractures are the cause of their symptoms.

Dual-energy X-ray absorptiometry (DXA) provides bone mineral density (BMD) measurements, but these do not identify the cause of back pain. A fragility fracture may be present even when BMD is above the densitometric threshold for osteoporosis [1]. Imaging is therefore needed to characterize a suspected fracture and assess its cause [4]. Bone turnover markers provide information about skeletal remodeling, although sampling conditions and renal function affect their interpretation [5]. Their contribution to diagnosis needs to be assessed in patients presenting with symptoms.

The existing models are designed for different clinical scenarios. Vfrac uses clinical information to assist in identifying elderly women with back pain who should undergo spinal X-ray examination in primary care [6]. Deep learning models based on magnetic resonance imaging (MRI) can assist in identifying fractures after scanning [7]. Other studies have developed nomograms for postmenopausal women with osteoporosis [8], or classified fragility fractures using clinical and laboratory information [9]. The differences in inclusion criteria, outcome definitions and required measurement indicators limited its direct application to symptomatic male and female inpatients.

The evaluation of diagnostic models should also go beyond apparent discrimination. A systematic review identified a high risk of bias in published spinal fracture prediction models and limited evidence supporting clinical implementation [10]. Internal validation should fully consider the selection of predictor variables and parameter tuning [11]. Individual prediction stability is another aspect that needs attention: average performance may mask the variability of probability estimates for specific patients [12].

We developed and internally evaluated a diagnostic model for symptomatic thoracolumbar fragility vertebral fractures in men aged at least 50 years and postmenopausal women hospitalized with thoracolumbar back pain. Using complete-pipeline bootstrap validation, we assessed discrimination, calibration, prediction stability, and exploratory net benefit. The proposed application is diagnostic support before final MRI-based adjudication when BMD and laboratory

results are already available.

Methods

Study Design and Participants

We developed a diagnostic prediction model using clinical records from the Department of Spine Surgery, Affiliated Hospital of Guangdong Medical University. Eligibility was assessed among patients hospitalized with thoracolumbar back pain as their primary complaint between January 1, 2023, and September 30, 2026. This was a single-center retrospective study. Reporting followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD+AI) statement for regression or machine-learning models [13].

Patients who met the inclusion criteria for the study were male patients at least 50 years old or female patients who had reached postmenopausal status. Natural menopause is defined as at least 12 consecutive months without menstruation, and this state cannot be related to pregnancy, medication, surgical intervention or other factors. The exclusion criteria include: men under 50 years old, women who have not experienced natural menopause, individuals with spinal infections or tumors, those who have suffered from high-energy trauma, and patients lacking complete clinical information. Duplicate hospitalization records have been excluded to ensure that each individual is counted only once in the observation group. The sample size was determined based on the count of eligible participants included during the study period.

The Ethics Committee of the Affiliated Hospital of Guangdong Medical University

approved this study on 30 December 2025 (approval no. PJKT2025-330). The committee waived the requirement for individual informed consent. This study was not registered and there is no publicly available research protocol. For more implementation details, please refer to the Supplementary Methods (Additional file 1). The analysis code and the final prediction model can be found in Additional File 2.

This dataset was assembled independently of the dataset used in the previous study published by Chen et al. [8]. As the two studies were conducted in the same institution during partially overlapping recruitment periods, it cannot be ruled out that some participants overlap. The degree has not been quantified. The intended use is to provide diagnostic support before confirmatory MRI assessment when BMD and laboratory results are available. This workflow has not undergone a prospective evaluation.

Outcome definition and evaluation

The primary outcome measure was to record symptomatic thoracolumbar fragility vertebral fractures diagnosed during hospitalization and related to symptoms at the time of visit. Fractures at other anatomical sites were not included in the outcome indicators. Fracture classification integrates imaging manifestations and clinical history consistent with low-energy injuries, including minor falls or sprains. The morphological features and signal abnormalities related to fractures were evaluated and examined by MRI, including the low-signal areas on T1-weighted scans and the high-signal areas on T2-weighted scans.

MRI examinations were independently reviewed by two radiologists. All radiologists

involved in image interpretation, including the senior adjudicating radiologist, were blinded to patients' BMD measurements, laboratory results, and prior clinical diagnoses. The two initial readers were also blinded to each other's assessments until both had submitted their initial interpretations. Disagreements were adjudicated by a third, senior radiologist.

Patients classified as having or not having a fracture underwent imaging assessment sufficient to establish their fracture status. Final outcome classification was based on imaging findings and the relevant clinical history; BMD measurements and laboratory results were not used as diagnostic criteria.

Candidate predictors and clinical measurements

A total of fifteen potential predictors were evaluated, including sex/eligibility stratum, age, body mass index (BMI), bone mineral density (BMD) of the lumbar spine, BMD of the femoral neck, diabetes status, serum calcium levels, serum phosphorus levels, fasting plasma glucose, triglycerides, low-density lipoprotein (LDL) cholesterol, 25-hydroxyvitamin D3 [25(OH)D3], osteocalcin, C-terminal telopeptide of type I collagen (CTX), and procollagen type I N-terminal propeptide (P1NP).

BMD of the lumbar spine and the left femoral neck was assessed during the index hospitalization utilizing a MEDIX90 dual-energy X-ray absorptiometry system (Medilink, France). The values for BMD were obtained directly from the clinical DXA reports. In patients with previous percutaneous vertebroplasty (PVP), treated vertebrae were excluded during routine DXA analysis, and the reported lumbar spine BMD was derived from the remaining vertebrae. Fractured vertebrae without prior

PVP were retained in the lumbar spine BMD analysis. BMD values were expressed in g/cm^2.

Fasting blood samples obtained during the index hospitalization were analyzed by the hospital's Department of Clinical Laboratory. The measurements for osteocalcin, CTX, and P1NP were performed via electrochemiluminescence immunoassays (ECLIA) using Roche cobas e 402/e 801 systems with reagents provided by the manufacturer (Roche Diagnostics). Results for vitamin D were obtained under the test label "25-hydroxyvitamin D3" as indicated in the clinical laboratory reports. The units of measurement are detailed in **Table 1**.

The stratified code for gender/inclusion criteria is as follows: 1 for postmenopausal women and 0 for men aged 50 and above. The code for diabetes is: 1 for existence and 0 for non-existence. The outcome was coded as 1 for symptomatic thoracolumbar fragility vertebral fracture and 0 for absence of this outcome.

Data preparation and descriptive analysis

Data preparation included checks of variable formats, coding, missing values, patient identifiers, and numerical ranges. Unusual values were checked against the source data and retained when confirmed. No automatic outlier deletion or winsorization was applied. All outcome and candidate-predictor values were available for the 1414 included patients; imputation was therefore unnecessary.

Descriptive summaries used the median with the 25th–75th percentiles for continuous measurements and n (%) for categorical characteristics. Between-group descriptive comparisons were not used to select predictors in the revised modeling

procedure.

Model development

The complete analytic cohort is utilized for the development of the model. The updated modeling procedure was established following the exploratory analysis and prior to the official bootstrap validation. This procedure was not established ahead of the examination of the research data.

We fitted a penalized logistic model using group least absolute shrinkage and selection operator regression (group LASSO) [14]. Each clinical predictor formed one penalty group. Continuous predictors were represented by three-knot restricted cubic splines [15], with knots at the 10th, 50th, and 90th percentiles of the relevant training sample. The intercept was not penalized.

Penalty selection used outcome-stratified, patient-level 10-fold cross-validation. The one-standard-error rule was applied to mean held-out binomial deviance. Spline construction and standardization both use only the training data within each fold. No class weighting or outcome rebalancing was applied.

The selected penalty was used to fit the final model to the full cohort. Penalized coefficients were retained without an unpenalized post-selection refit or additional recalibration. Ordinary Wald confidence intervals and P values were not estimated for individual spline coefficients. Numerical settings and implementation details are provided in the Supplementary Methods.

Performance evaluation and internal validation

Discrimination was summarized by the concordance index (C-index), which equals

the area under the receiver operating characteristic curve (AUC) for this binary outcome. Overall probability error was assessed with the Brier score. For calibration-in-the-large, we fitted a logistic regression containing an intercept and the predicted logit as an offset with coefficient fixed at 1. A separate logistic regression estimated the calibration slope, with its intercept unrestricted. The DeLong method provided a descriptive 95% confidence interval (CI) for apparent AUC, conditional on the fitted predictions [16].

Internal validation used 1000 patient-level bootstrap resamples [17]. Within each resample, we rebuilt the spline bases, ran the inner cross-validation, selected the penalty and predictor groups, and fitted the penalized model [11,18]. When a patient appeared more than once in a resample, all copies stayed together in the inner cross-validation.

Each bootstrap-fitted model was evaluated in its resampled data and in the original cohort. For each performance measure, the mean difference between these evaluations was subtracted from the original model’s apparent value. Corrected estimates were reported without confidence intervals. Earlier exploratory analyses and analyst-led decisions were outside this resampling procedure. We did not assess performance by demographic subgroup or formally evaluate fairness.

Calibration curve and prediction stability

Calibration curves were estimated using locally weighted scatterplot smoothing (LOWESS). Point-by-point optimism correction uses the full-process bootstrap model. The curve is presented between the 2.5th percentile and the 97.5th percentile of the

original model's predicted probability and is not extrapolated beyond its supported probability range. The rug chart shows the predicted values for all patients. The confidence band has not been calculated. The smoothing and interpolation settings can be found in the supplementary methods.

Using the same bootstrap models, we counted how often each predictor group was retained. For each patient, we calculated the mean absolute difference between the bootstrap-predicted probabilities and the original model’s probability, then averaged this quantity across patients [12]. We also summarized the within-patient 2.5th–97.5th percentile ranges. These ranges describe variation across fitted models, not confidence intervals. Selection frequencies were not used for further predictor selection.

Exploratory decision curve analysis

Decision curve analysis (DCA) explored the net benefit of model-guided MRI prioritization. Threshold probabilities from 0.05 to 0.80 were evaluated in increments of 0.01. Patients whose predicted probabilities met or exceeded a threshold were classified as eligible for model-triggered prioritization.

Net benefit was calculated as the true-positive proportion minus the false-positive proportion weighted by the threshold odds, $p_t/(1 - p_t)$[19,20]. Model-guided prioritization was compared with prioritizing all patients and prioritizing none. The latter represented no model-triggered prioritization, not withholding clinically indicated imaging or standard care.

Net benefit was optimism-corrected at each threshold using the complete-pipeline

bootstrap models. No smoothing or confidence bands were applied. The threshold range was exploratory, and no clinical operating cutoff was selected. Predictor-measurement costs, MRI capacity, waiting times, and downstream treatment effects were not incorporated.

Model presentation and software

A nomogram was constructed from the unchanged penalized model. Its numerical equivalence to the original model and the standalone prediction function was verified. Exact prediction specifications and verification procedures are described in the Supplementary Methods.

The analysis was completed using version R 4.4.2. Group-penalized regression was implemented using grpreg 3.6.0 [21], spline construction was performed using Hmisc 5.2-5, and receiver operating characteristic (ROC) analysis was conducted using pROC 1.18.5. The numerical settings, random seeds and implementation details are all recorded in the analysis code and supplementary methods.

This model is designed for clinicians who are familiar with spinal diseases and the interpretation of DXA and laboratory reports. Before entering the values, users must verify the identity, unit and completeness of the predictor variables. The standalone prediction function will reject missing or invalid required inputs. Model estimates should not be used when the required measurement values are unavailable or unreliable.

Patient and public participation

Patients and members of the public did not participate in the design, implementation,

interpretation or communication planning of this study.

The use of artificial intelligence tools

ChatGPT (OpenAI) is used as an auxiliary tool for language editing and writing and debugging R analysis code. All text and code generated with assistance have been reviewed and revised by the authors, who are fully responsible for the analysis, interpretation and final manuscript.

Results

Study Population and Clinical Characteristics

In total, 1738 patients admitted due to thoracolumbar back pain were examined. Exclusions comprised 87 men aged under 50 years, 38 non-postmenopausal women, 169 patients with excluded spinal conditions, including infection, tumors, and high-energy trauma, and 30 patients with incomplete data. The final cohort included 1414 patients: 571 (40.4%) with symptomatic thoracolumbar fragility vertebral fractures and 843 (59.6%) without the target fragility vertebral fracture outcome **(Fig. 1)**.

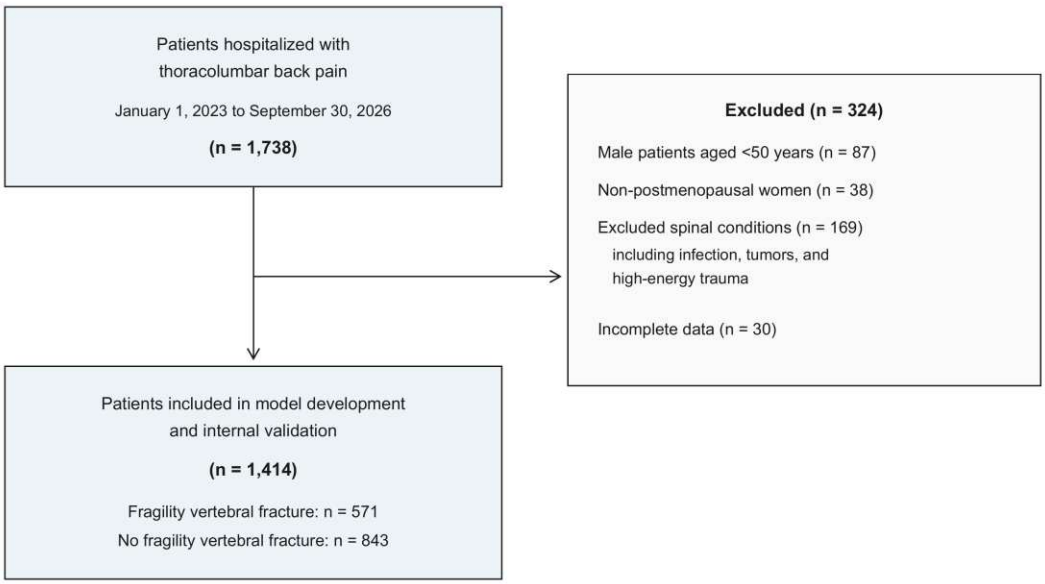


Fig. 1 Participant selection and analytic cohort.

The cohort comprised 514 eligible men (36.4%) and 900 postmenopausal women (63.6%). Median age was 68.0 years (interquartile range [IQR], 60.0 – 76.0), and median BMI was 22.59 kg/m² (IQR, 20.32 – 24.44). Diabetes was recorded in 193 patients (13.6%). The age of patients with fractures was generally higher than that of those without fractures, with median ages of 73.0 years and 65.0 years, respectively. Postmenopausal women accounted for 78.3% in the fracture group, while the proportion was 53.7% in the non-fracture group. The median lumbar bone mineral density (BMD) in the fracture group was 0.602 g/cm², and that in the non-fracture group was 0.768 g/cm². Similarly, the median BMD of the femoral necks in the two groups was 0.690 g/cm² and 0.857 g/cm², respectively. The distribution of all 15 candidate predictor variables is shown in **Table 1**.

Table 1. Characteristics of the analytic cohort according to fragility vertebral fracture status.

Characteristic	Overall (n=1414)	Fragility vertebral fracture (n=571)	No fragility vertebral fracture (n=843)
Sex/eligibility stratum			
Male, age ≥50 years	514 (36.4)	124 (21.7)	390 (46.3)
Postmenopausal female	900 (63.6)	447 (78.3)	453 (53.7)
Age (years)	68.0 (60.0-76.0)	73.0 (65.5-80.0)	65.0 (58.0-72.0)
Body mass index (kg/m^2)	22.59 (20.32-24.44)	21.91 (19.54-23.61)	23.03 (21.09-24.84)
Lumbar spine BMD (g/cm^2)	0.695 (0.585-0.830)	0.602 (0.524-0.683)	0.768 (0.663-0.894)
Femoral neck BMD (g/cm^2)	0.788 (0.671-0.914)	0.690 (0.603-0.794)	0.857 (0.748-0.967)
Diabetes mellitus			
Yes	193 (13.6)	82 (14.4)	111 (13.2)
No	1221 (86.4)	489 (85.6)	732 (86.8)
Calcium (mmol/L)	2.23 (2.16-2.30)	2.22 (2.15-2.29)	2.23 (2.17-2.31)
Phosphorus (mmol/L)	1.15 (1.02-1.27)	1.13 (1.01-1.27)	1.16 (1.03-1.28)
Fasting plasma glucose (mmol/L)	4.77 (4.22-5.50)	4.85 (4.25-5.69)	4.70 (4.19-5.37)
Triglycerides (mmol/L)	1.23 (0.91-1.79)	1.11 (0.84-1.60)	1.32 (0.96-1.95)
LDL cholesterol (mmol/L)	3.16 (2.52-3.83)	3.09 (2.50-3.69)	3.20 (2.52-3.91)
25-Hydroxyvitamin D_3 (ng/mL)	31.11 (24.94-39.04)	29.79 (23.98-36.99)	32.07 (26.12-40.39)
Osteocalcin (ng/mL)	16.80 (11.70-23.40)	16.80 (11.80-23.35)	16.70 (11.70-23.50)
CTX (ng/mL)	0.680 (0.430-0.958)	0.740 (0.420-1.000)	0.650 (0.430-0.920)
P1NP (ng/mL)	52.60 (36.00-73.10)	58.40 (40.20-78.80)	49.30 (33.20-68.25)

Note: Data are shown as median (25th – 75th percentiles) or n (%). Percentages are computed for each column individually. All potential predictors are listed without consideration for the final model selection. BMD: bone mineral density; LDL: low-density lipoprotein; CTX: C-terminal telopeptide of type I collagen; P1NP: N-terminal propeptide of type I procollagen.

Model development and nomogram

The penalty corresponding to minimum cross-validation loss was 0.003765; the one-standard-error rule selected a penalty of 0.017074. Twelve predictor groups were

retained: sex/eligibility stratum, age, lumbar spine BMD, femoral neck BMD, calcium, fasting plasma glucose, triglycerides, LDL cholesterol, 25-hydroxyvitamin D3, osteocalcin, CTX, and P1NP. BMI, diabetes status, and phosphorus were not retained. The final model included one binary predictor and 11 continuous predictors represented by three-knot restricted cubic splines, giving 23 nonzero slope basis coefficients and an intercept. Penalized coefficients and development-cohort spline knots are reported in **Table 2**. The nomogram provides a graphical representation of the unchanged fitted model **(Fig. 2)**.

Table 2. Penalized coefficients and spline knots of the final diagnostic prediction model.

Predictor	Linear or indicator β	Nonlinear β	Knot 1 (P10)	Knot 2 (P50)	Knot 3 (P90)
Intercept	3.490998	-	-	-	-
Postmenopausal female vs male aged ≥50 years	0.319316	-	-	-	-
Age (years)	0.020199	0.023544	55	68	83
Lumbar spine BMD (g/cm^2)	-6.721370	3.747262	0.5013	0.695	0.9727
Femoral neck BMD (g/cm^2)	-1.446552	0.860999	0.5803	0.7885	1.03
Calcium (mmol/L)	-0.023649	0.006828	2.09	2.23	2.377
Fasting plasma glucose (mmol/L)	0.064300	-0.043289	3.8	4.77	6.894
Triglycerides (mmol/L)	-0.135009	0.141249	0.7	1.23	2.575
LDL cholesterol (mmol/L)	-0.013955	-0.031906	1.96	3.16	4.49
25-Hydroxyvitamin D_3 (ng/mL)	-0.004401	0.003956	20.11	31.105	48.014
Osteocalcin (ng/mL)	-0.008239	0.003968	8.139	16.8	31.9
CTX (ng/mL)	-0.000855	-0.005146	0.26	0.68	1.27
P1NP (ng/mL)	0.001061	-0.000969	24.13	52.6	97.58

Note: Coefficients are penalized log-odds coefficients for the specified spline basis on the original predictor scales. Each continuous predictor contributes a linear and a nonlinear basis term;

individual basis coefficients should not be interpreted as constant per-unit odds ratios. Knots correspond to the 10th, 50th, and 90th development-cohort percentiles, calculated using quantile type 7. Postmenopausal women are coded 1 and men aged at least 50 years are coded 0. The model contains 23 nonzero slope basis coefficients and one intercept; this count is not the effective degrees of freedom. Ordinary Wald confidence intervals and P values were not estimated. The basis was constructed using Hmisc::rcspline.eval with inclx = TRUE, norm = 2, and pc = FALSE. Coefficients are displayed to six decimal places. Exact predictions should use the unrounded parameters and accompanying prediction function. P10, P50, and P90 refer to the 10th, 50th, and 90th percentiles, respectively. Additional abbreviations can be found in Table 1.

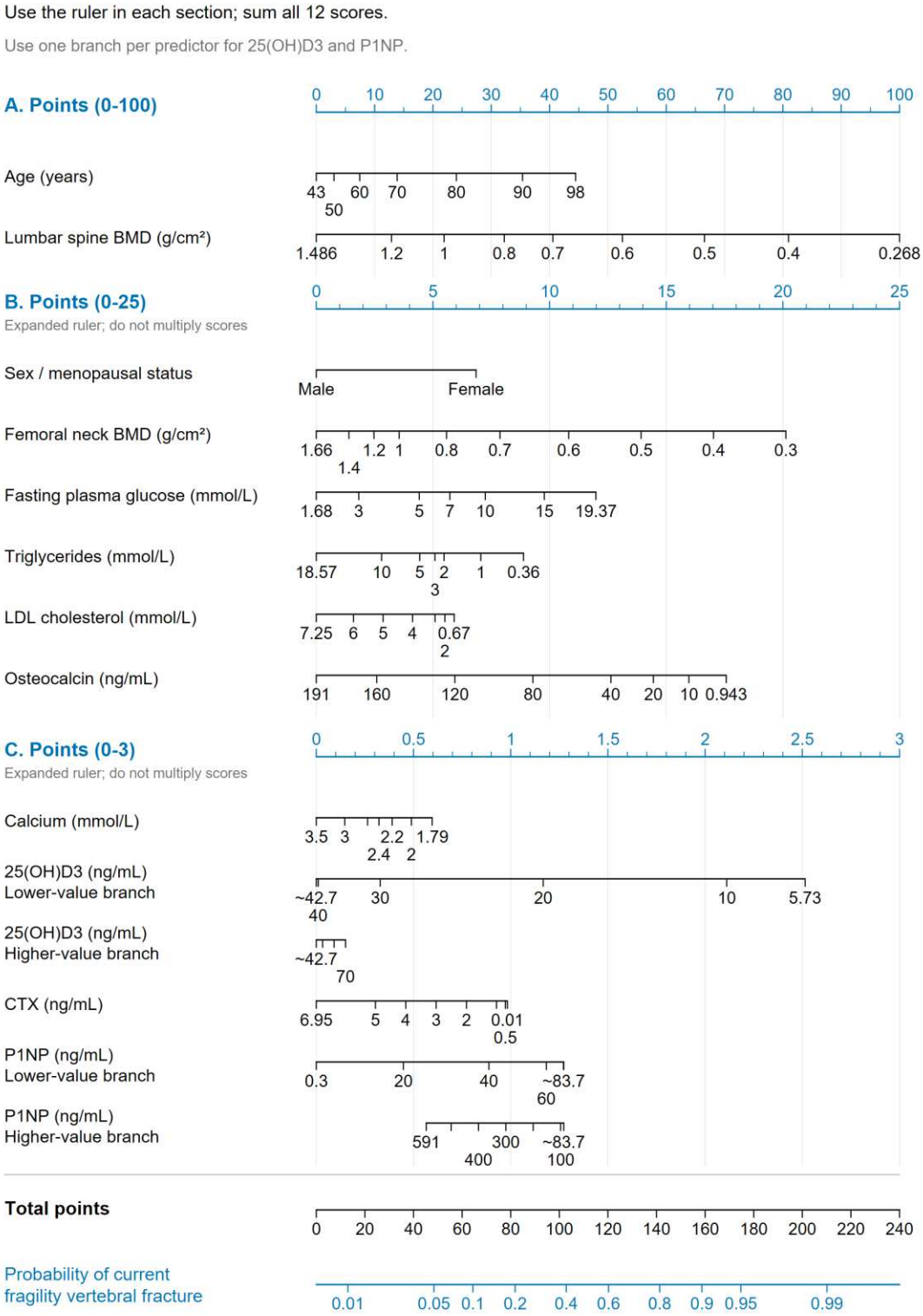


Fig. 2 Nomogram for identifying fragility vertebral fractures during the index hospitalization.

Discrimination and Internal Validation

All 1000 bootstrap repetitions of the complete process were successfully completed.

The calculated apparent AUC was 0.822 (DeLong 95% CI was 0.800 to 0.844), while the C-index after optimism correction was 0.818. The apparent Brier score and the corrected Brier score were 0.168 and 0.171 respectively (refer to **Table 3**). The ROC curve is shown in **Fig. 3**.

Table 3. Apparent and optimism-corrected performance of the diagnostic prediction model.

Performance measure	Apparent	Bootstrap optimism-corrected
C-index (AUC)	0.822	0.818
Brier score	0.168	0.171
Calibration-in-the-large intercept	-0.000001	0.000211
Calibration slope	1.252	1.218

Note: Both columns describe the same development cohort (n = 1414; 571 fracture events). Optimism correction used 1000 patient-level bootstrap resamples repeating the complete revised model-development procedure.

The DeLong 95% confidence interval for apparent AUC was 0.800 – 0.844. This interval is conditional on the fitted predictions and does not account for the complete development process. Confidence intervals for corrected estimates were not calculated.

The estimation of the calibration intercept takes the predicted logit as the offset, and its coefficient is fixed at 1. The calibration slope is evaluated independently, and its intercept is freely estimated. The intercept is displayed to six decimal places, while other indicators are rounded to three decimal places. C-index refers to the concordance index, and AUC represents the area under the receiver operating characteristic curve. The above results reflect the results of internal verification.

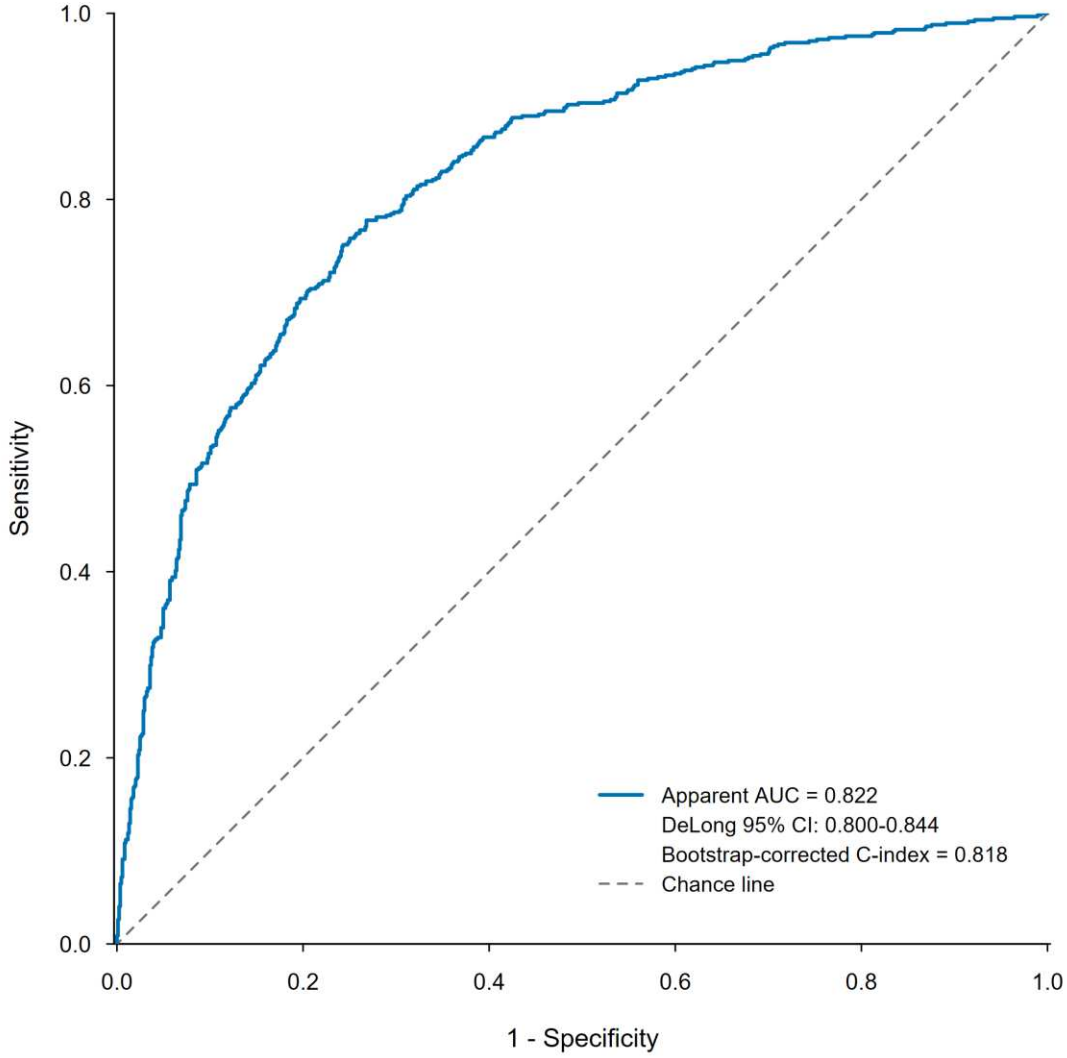


Fig. 3 Discrimination of the diagnostic prediction model in the entire analytic cohort.

Calibration

The calibration-in-the-large is close to zero both before and after optimism correction. The observed calibration slope was 1.252, and the slope after optimism correction was 1.218 (see Table 3). Fig. 4 shows the apparent calibration curve and the calibrated curve after correction.

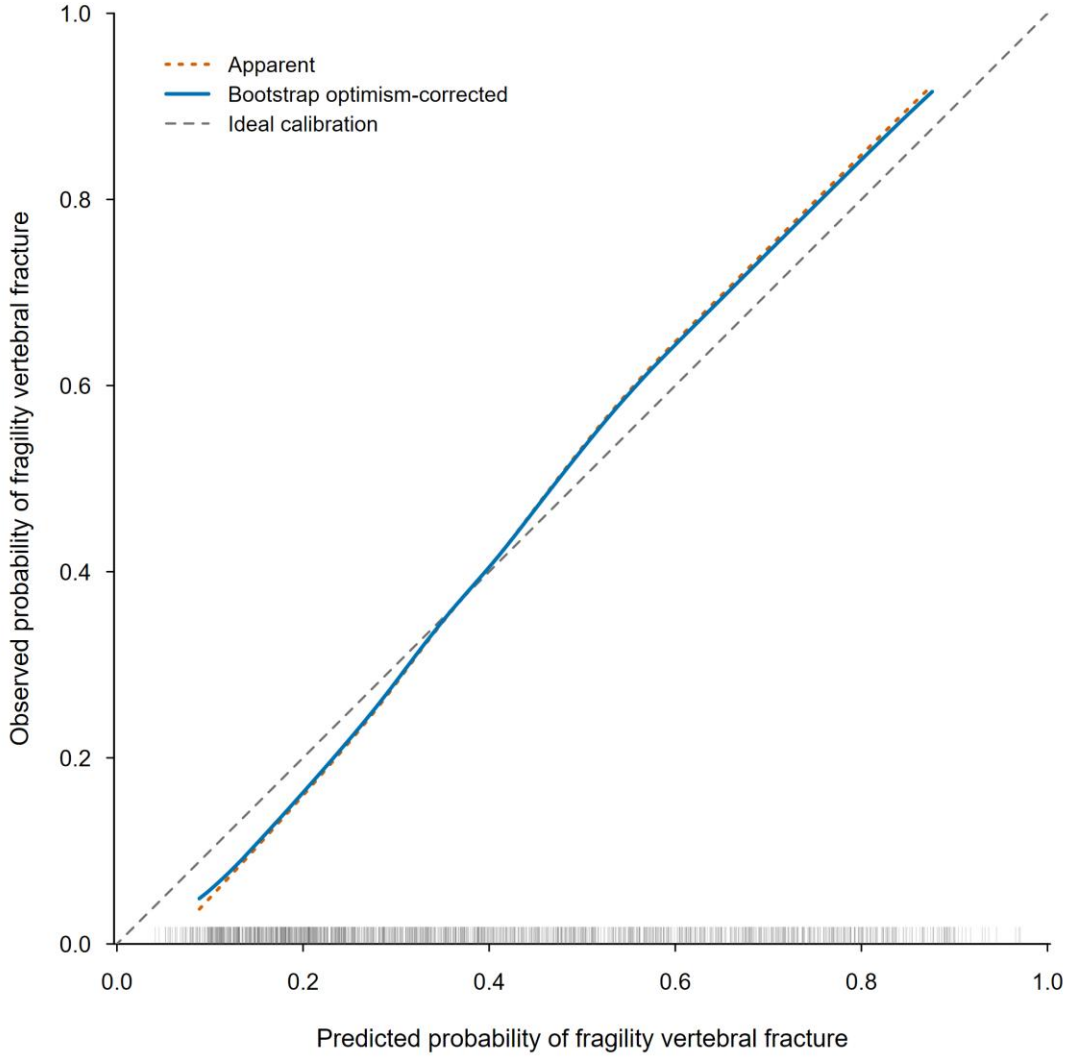


Fig. 4 Apparent and bootstrap optimism-corrected calibration

Predictor-selection and prediction stability

Age and lumbar spine BMD were retained in every bootstrap sample. Femoral neck BMD and sex/eligibility stratum were retained in 95.0% and 89.9% of samples, respectively. Selection frequencies for the eight biochemical predictors retained in the original model ranged from 9.5% for CTX to 32.7% for fasting plasma glucose. The median number of selected predictor groups was five, with a range of 2 to 14.

The average absolute variation in probabilities between the bootstrap model and the original model was 0.0398, which corresponds to 3.98 percentage points. The median

width of the within-patient 2.5th – 97.5th percentile range was 0.1230, or 12.30 percentage points; the 90th percentile of these widths was 0.1750, or 17.50 percentage points.

Exploratory decision curve analysis

Optimism-corrected net-benefit point estimates for model-guided MRI prioritization exceeded both reference strategies at the evaluated thresholds from 0.10 to 0.80 **(Fig. 5)**. When the threshold was set at 0.30, the corrected net benefit amounted to 0.246, in contrast to 0.148 when all patients were prioritized and zero when none were prioritized. At a threshold of 0.80, the corrected net benefit decreased to 0.021. No clinical operating cutoff was determined.

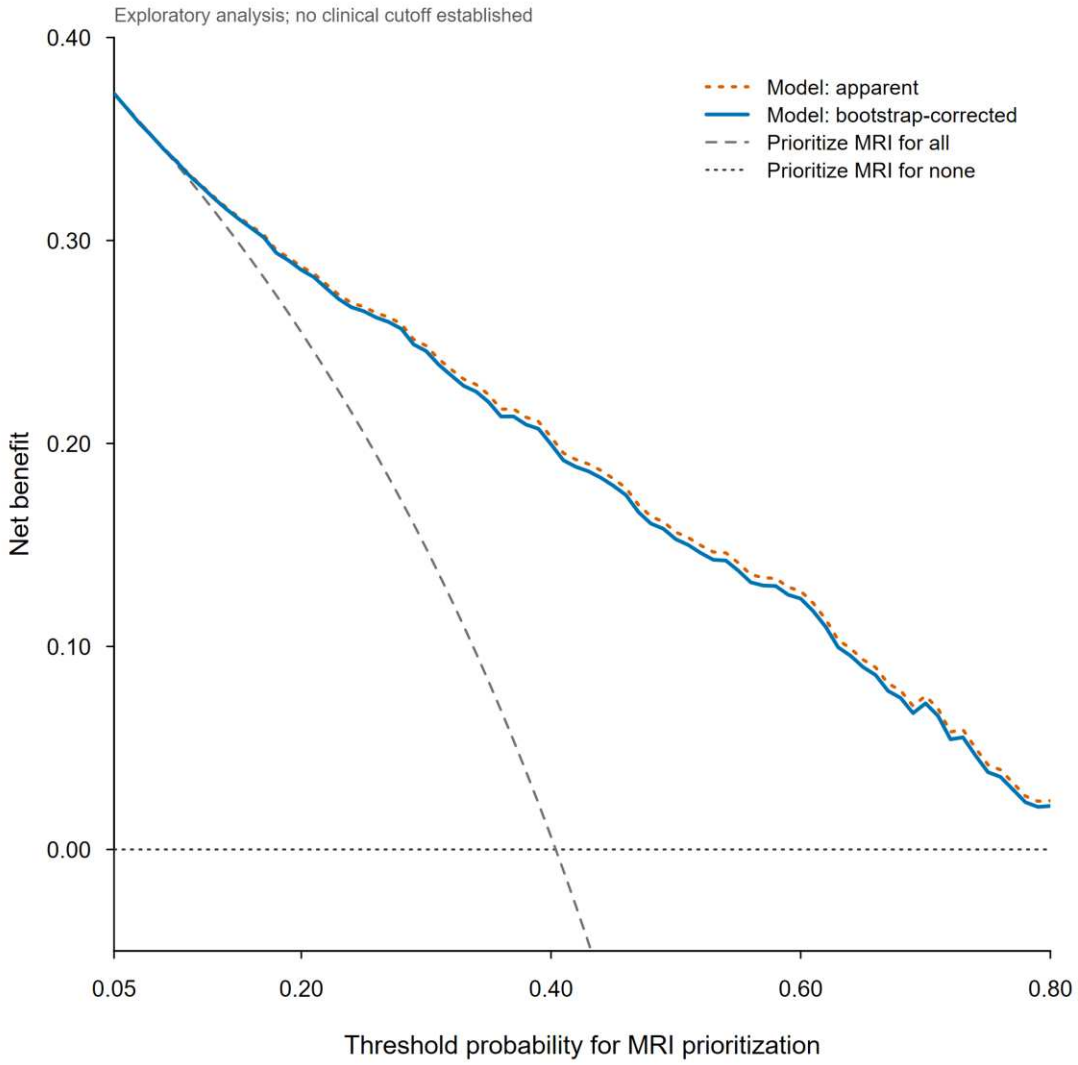


Fig. 5 Exploratory decision curve analysis of MRI-priority triage guided by the model

Discussion

Principal findings and diagnostic relevance

Among 1,414 patients hospitalized due to thoracolumbar back pain, this model was able to distinguish patients with symptomatic thoracolumbar fragility vertebral fractures from those without fractures. Its Bootstrap-corrected C-index was 0.818 and the Brier score was 0.171. Both groups of patients presented with back pain. Therefore, this model addresses a diagnostic discrimination issue encountered in symptomatic hospitalized patients rather than a comparison with asymptomatic

controls. Its intended use is to assist in assessment when DXA and laboratory test results are available but the final diagnosis based on imaging has not yet been established.

This diagnostic purpose is different from predicting the risk of future fractures. Fragility fractures can occur in patients with BMD T-scores higher than the density measurement threshold for osteoporosis [1], while a simple low BMD value cannot determine whether the current pain is attributed to vertebral fractures. This model estimates the current probability of fractures by integrating the information obtained during admission. Whether this estimate will help clinicians prioritize confirmatory evaluations remains to be answered by prospective evaluation.

Positioning within the existing evidence

The Vfrac study focused on an earlier diagnostic decision: among women aged 65 and above who visited primary care due to back pain, who should undergo spinal X-ray examinations. The model incorporated clinical history, physical measurement indicators and pain characteristics, with an AUC of 0.802. After removing the pain description words, the discrimination ability decreased [6]. Our model is applied to the later stage of the assessment pathway, that is, after the completion of DXA and laboratory tests during hospitalization. Vfrac also pointed out information not included in this dataset: The location, nature and activity-related patterns of pain may have complementary value to the measurement indicators included in this article.

Previous studies have combined skeletal and biochemical measurement indicators with other clinical information. Wen et al. incorporated the paravertebral muscle

features derived from MRI into the vertebral fracture model [22]. Our predictor variables do not require MRI-derived measurement indicators. Chen et al. studied 326 postmenopausal women with osteoporosis in our institution and used a randomized development-validation segmentation [8]. This study included eligible men and repeated the model development process in bootstrap resampling. Although the datasets were constructed separately, there was some overlap in the recruitment times, and the degree of overlap among the participants was not evaluated. Therefore, the results of this study should not be regarded as an independent duplicate verification of the research.

Si et al. also classified osteoporotic fragility fractures by integrating clinical, laboratory and BMD information [9]. Therefore, integrating these information sources is not the innovation of this study in itself. The contribution of this analysis lies in focusing on symptomatic thoracolumbar vertebral fractures in hospitalized patients and evaluating the complete model development process, including the selection of predictor variables and the variability of individual predictions.

Various models address distinct inquiries. Wang et al. employed deep learning techniques to detect new vertebral fractures in MRI scans, aiding in the interpretation of images post-examination [7]. Sun et al. conducted an evaluation of fractures that appeared over a five-year follow-up period in postmenopausal women [23]. In contrast, Nie et al. developed a model for new vertebral fractures occurring within three years following vertebroplasty [24]. The discrimination capabilities they reported do not allow for a comparison of the relative effectiveness of our model, as

the outcomes, evaluation periods, and patient demographics vary.

Interpretation of the predictive information

Age and lumbar spine BMD were retained in all 1,000 bootstrap samples. Femoral neck BMD and the sex/eligibility stratum were also frequently retained, in 95.0% and 89.9% of samples, respectively. These selection patterns were more reproducible than those of the biochemical predictors. They do not, however, quantify effect sizes or establish biological importance. In particular, the sex/eligibility variable distinguishes postmenopausal women from men aged at least 50 years; it cannot separate the effects of sex and menopausal status.

The two BMD measurements require different clinical interpretation. Lumbar measurements are susceptible to local structural changes, including degeneration and vertebral compression [25]. Lin et al. demonstrated that spinal implants or bone cement can influence BMD when affected vertebrae remain within the DXA region of interest, even if they are subsequently excluded from the numerical analysis [26]. These findings are relevant to our use of routine clinical DXA reports. Frequent retention of lumbar and femoral neck BMD warrants examining their joint predictive contribution, but it does not show that using both improves performance over either measurement alone.

Indicators of bone turnover can provide insights into bone metabolism that are not revealed by areal bone mineral density. The 2025 international consensus reaffirms that P1NP and the β-isomerized C-terminal telopeptide of type I collagen can be used as reference markers, while emphasizing the limitations of sampling conditions,

448 renal function, and incorporating them into fracture risk assessment algorithms [5].
449 This provides a basis for the study of biochemical detection indicators, but the
450 evidence regarding future fracture risks does not establish its value in diagnosing
451 existing fractures.
452 The timing of measurement is particularly important for this study. Jia et al. reported
453 that there were differences in P1NP and CTX among patients evaluated at different
454 time intervals after vertebral fractures [27]. Ivaska et al. 's research on subsequent
455 fractures in elderly women found that the association between bone turnover markers
456 and fractures depends on the type of marker, age, and predictive time window [28].
457 The admission measurements in this study may simultaneously reflect the existing
458 bone metabolic state and the response to the current fracture, but this study was
459 unable to distinguish between these two types of components. The response to an
460 existing fracture may still provide diagnostic information, but it should not be
461 interpreted as evidence that the measured concentration existed prior to the fracture or
462 caused it.
463 All eight biochemical predictors retained in the original model had substantially lower
464 bootstrap selection frequencies, ranging from 9.5% to 32.7%. Their inclusion should
465 not be interpreted as evidence of reproducible independent risk factors. Conversely,
466 nonselection does not establish clinical irrelevance: diabetes, for example, is
467 associated with bone fragility despite not being retained here [29]. Retention of
468 glucose or lipid measurements does not establish them as causal intervention targets.
469 Comparison with a simpler demographic–BMD model is needed to quantify the

additional predictive value of the biochemical panel.

Reliability beyond discrimination

Although the difference between apparent discrimination and corrected discrimination is small, its explanation depends on the repeated steps of the verification procedure. In addition to coefficient estimation, the selection and tuning of predictor variables may also introduce optimism bias [30]. We conducted research and development using a full cohort [11,17], and repeated spline construction, tuning, predictor group selection, and fitting in each bootstrap sample [18]. The corrected estimates cover all the above-mentioned stages of the revision procedure, but do not cover the early exploratory decision-making.

The small reduction in C-index did not capture all variation in individual predictions. Across the bootstrap models, probabilities differed from the original estimates by an average of about 4.0 percentage points, and the median within-patient percentile-range width was 12.3 percentage points. Such variation matters when an individual probability informs further assessment [12,31].

The corrected calibration slope of 1.218 suggests that predictions were too moderate on the logit scale [32]. Penalization may have contributed, but the analysis did not establish the cause, and the precision of the slope was not quantified. External evaluation should examine absolute probabilities as well as discrimination.

Potential clinical application

The exploratory decision curve supports further research on the priority triage strategy of MRI. Within the evaluation threshold range of 0.10-0.80, the corrected net benefit

point estimates exceeded both reference strategies. This range does not define when clinicians should take action. The setting of the threshold must reflect the consequences of delayed assessment and unnecessary priority triage, rather than being selected from the favorable sections of the decision curve [19].

Implementation also depends on the way and timing of obtaining predictor information. Khalid et al. 's research indicates that self-administered and face-to-face physical measurements may yield different Vfrac assessment results [33]. For this model, relevant considerations include the accessibility and consistency of DXA and laboratory test results. When these results are already available, their proposed uses are most reasonable; if these results are obtained specifically for the computational model, the additional costs and burdens need to be evaluated.

A higher estimated probability may prompt closer assessment or consideration of accelerating imaging examinations. This probability was not evaluated as an indication for bed rest or drug treatment, nor was it used as a rule for postponing MRI. The assessment and management of osteoporosis after the diagnosis of vertebral fractures should follow established guidelines [2,34]. Whether model-guided priority triage can improve waiting time or patient outcomes remains to be verified.

Advantages, limitations, and future directions

This study examined the symptomatic inpatient population, included eligible men and postmenopausal women, and evaluated the complete revision development process. The outcome assessment adopts an independent first-round image interpretation and ruling. Radiologists were blinded to candidate BMD measurements and laboratory

results. The above-mentioned detection indicators have not been included in the reference standards. In addition, by exporting the model specifications and conducting prediction equivalence tests, the fitted model can be reproduced.

The retrospective, single-center design limited the extrapolation of results to external outpatient, emergency and community scenarios. Excluding cases with incomplete records may introduce selection bias. Lumbar BMD was extracted from routine reports: vertebrae that had previously undergone vertebroplasty were excluded when applicable, but untreated fractured vertebrae were retained. Therefore, structural changes may have affected the measurement results and should not be assumed to represent the bone state not affected by fractures. Renal function, osteoporosis treatment, detailed pain characteristics and the time after fracture were not included in the analysis. The performance in each demographic subgroup was not evaluated.

The revised development rules were fixed before formal validation, but bootstrap did not reproduce the previous exploratory analysis or analyst-led decisions. Therefore, there may still be a certain degree of optimism bias. The confidence interval of the corrected performance and the uncertainty band of the decision curve were not estimated. Transportability and the incremental value relative to simpler demographic-BMD models or conventional clinical evaluations have yet to be established.

The next step of assessment should apply the frozen model to an external cohort where the eligibility criteria, predictor measurements, and outcome evaluations are all clearly defined [35]. Comparisons with simpler models should determine whether the

combination of biochemical indicators provides useful additional information. Subsequently, prospective workflow studies need to be conducted to evaluate the accessibility of predictors, the prioritization of imaging examinations, and clinically relevant outcomes.

Conclusions

The diagnostic model identified patients in this inpatient cohort who had symptomatic thoracolumbar fragility vertebral fractures as opposed to those who did not, achieving a bootstrap-corrected C-index of 0.818. Calibration and individual prediction variability require external evaluation. Its additional value over simpler models and its effect on MRI prioritization should be established before clinical implementation.

Figure legends

Fig. 1 Participant selection and analytic cohort.

Between January 1, 2023, and September 30, 2026, a screening was conducted at the Affiliated Hospital of Guangdong Medical University for patients hospitalized due to thoracolumbar back pain. Out of 1,738 individuals, 324 were excluded from the study: 87 were men below 50 years old, 38 were non-postmenopausal women, 169 had spinal conditions that were excluded, such as infections, tumors, and high-energy trauma, and 30 had incomplete data records. The final analytic group included 1,414 patients, comprised of 571 with symptomatic thoracolumbar fragility vertebral fractures and 843 without fragility vertebral fractures. This entire group was utilized for developing models and conducting internal validation.

Fig. 2 Nomogram for identifying fragility vertebral fractures during the index hospitalization.

This nomogram displays the final penalized logistic model with 12 predictors. Each continuous predictor is represented by a three-knot restricted cubic spline. Read each predictor’s score from the Points ruler in its corresponding section (A, B, or C), then sum all 12 scores. Locate the sum on the Total points axis and read the corresponding probability. Area B and Area C use the extended visual scale, but all scores share the same additive score unit and no amplification factor should be applied. For 25-hydroxyvitamin D3 and P1NP, use only the branch containing the patient’s measurement. Male refers to a man who is at least 50 years old. Female refers to postmenopausal women. The score conversion retains the penalty coefficient and the stored spline knots. The manually read value is an approximation. The exact probability should be obtained using the accompanying prediction function. This model estimates the current fracture probability rather than the future fracture risk and cannot replace confirmatory imaging examinations or establish treatment indications. BMD, bone mineral density; LDL, low-density lipoprotein; 25(OH)D3, 25-hydroxyvitamin D3; CTX, C-terminal telopeptide of type I collagen; P1NP, N-terminal propeptide of type I procollagen.

Fig. 3 Discrimination of the diagnostic prediction model in the entire analytic cohort.

The blue receiver operating characteristic (ROC) curve showed the apparent discrimination of 1414 patients (including 571 patients with symptomatic

thoracolumbar fragility vertebral fractures). The area under the apparent ROC curve (AUC) was 0.822 (DeLong 95% confidence interval [CI], 0.800-0.844), and the complete-pipeline bootstrap-corrected concordance index (C-index) was 0.818. This CI is only applicable to the apparent AUC based on the fitted predicted values. No CI estimation was performed for the corrected C-index. The gray dashed diagonal line represents chance discrimination.

Fig. 4 Apparent and bootstrap optimism-corrected calibration.

The orange dotted line represents apparent calibration, the blue solid line represents calibration after optimism correction, and the gray dashed diagonal line represents ideal calibration. Correction was carried out using 1,000 patient-level bootstrap resampling to repeat the revised development process. Curves were estimated using locally weighted scatterplot smoothing (LOWESS; smoothing span, two-thirds; no robust iterations). They are displayed between the 2.5th and 97.5th percentiles of the original model’s predicted probabilities. The rug plot includes all 1,414 patients. The curve is not extrapolated outside its support range. At each display probability value, the number of paired bootstrap evaluations involved in the calculation ranged from 715 to 1,000 times. The confidence band is not shown in the figure.

Fig. 5 Exploratory decision curve analysis of MRI-priority triage guided by the model.

Suppose the action is to prioritize magnetic resonance imaging (MRI) assessment after the predictor results are available. The orange dotted line represents the apparent net benefit of the model, and the blue solid line represents the optimism-corrected net

benefit obtained based on 1,000 full-process bootstrap resampling. The gray dashed curve represents prioritizing all patients for MRI, and the dark gray dotted horizontal line represents no model-triggered prioritization. Non-priority scheduling refers to not conducting model-triggered priority triage, rather than not performing MRI examinations or giving up standard diagnosis and treatment. The threshold probabilities ranging from 0.05 to 0.80 were evaluated with a step size of 0.01. The cut-off value for clinical operation has not been determined. The smoothing processing or confidence bands are not shown in the figure. The graphical display shows that the net benefit range spans from -0.05 to 0.40. The lowest net benefit figures for all prioritized strategies are included in the corresponding coordinate data. This analysis did not evaluate variations in waiting times, treatment choices, or patient outcomes.

List of abbreviations

AUC: area under the curve of the receiver operating characteristic; BMD: bone mineral density; BMI: body mass index; C-index: concordance index; CI: confidence interval; CTX: C-terminal telopeptide of type I collagen; DCA: decision curve analysis; DXA: dual-energy X-ray absorptiometry; ECLIA: electrochemiluminescence immunoassay; IQR: interquartile range; LASSO: least absolute shrinkage and selection operator; LDL: low-density lipoprotein; LOWESS: locally weighted scatterplot smoothing; MRI: magnetic resonance imaging; P1NP: procollagen type I N-terminal propeptide; PVP: percutaneous vertebroplasty; ROC: receiver operating characteristic; TRIPOD+AI: revised Transparent Reporting of a

multivariable prediction model for Individual Prognosis Or Diagnosis guidelines for regression or machine learning models; 25(OH)D3: 25-hydroxyvitamin D3.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. Ethical approval was granted by the Ethics Committee of the Affiliated Hospital of Guangdong Medical University on 30 December 2025 (approval no. PJKT2025-330). The committee waived the requirement for individual informed consent for this retrospective study using de-identified clinical data.

Consent for publication

Not applicable.

Availability of data and materials

Participant-level clinical data are not publicly released because the records may permit identification. Access may be requested from the corresponding author and requires approval from the Ethics Committee of the Affiliated Hospital of Guangdong Medical University. Additional file 2 contains the analysis code, frozen model specification, and standalone prediction function.

The author-written analysis code and accompanying documentation are distributed under the MIT License. This license does not grant access to participant-level clinical data, which remain subject to the stated privacy and ethics requirements.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Big Data Platform of the Affiliated Hospital of Guangdong Medical University; the National Natural Science Foundation of China (Grant No. 82472454 and 82272505); the Scientific Research Project of the Guangdong Provincial Administration of Traditional Chinese Medicine (Grant No. 20251205 and 20261199); the Zhanjiang Science and Technology Development Special Fund (Grant No. 2025A502026 and 2025A502021); the High-Level Talents Scientific Research Start-up Funds of the Affiliated Hospital of Guangdong Medical University (Grant No. GCC2022008 and GCC2024022); the Clinical Research Program of the Affiliated Hospital of Guangdong Medical University (Grant No. LCYJ2019B012); and the Special Project for Clinical and Basic Science and Technology Innovation of Guangdong Medical University (Grant No. GDMULCJC2025049 and GDMULCJC2025063).

The funders had no role in study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Authors' contributions

ZL conceived and designed the study, supervised the research, acquired funding, and critically revised the manuscript for important intellectual content. HL, ChZ, XC, and LX collected and verified the clinical data. XZ contributed to data collection. JZ performed the statistical analyses. JZ and CZh drafted the manuscript. JZ and HL contributed equally to this work. All authors read and approved the final manuscript.

Author initials: JZ, Jinchang Zheng; HL, Haoyuan Luo; ChZ, Chuhui Zheng; XC,

Xiaotong Chen; LX, Langfan Xu; XZ, Xiaoen Zhuang; CZh, Chanli Zhong; ZL, Zhen Liang.

Acknowledgments

The authors thank the Joint Research Center on Anti-Osteoporosis Innovative Drugs of Guangdong Medical University and the International Chinese Musculoskeletal Research Society for their support during the study.

Supplementary information

Additional file 1 (.docx): Supplementary Methods. Details of spline construction, penalty tuning, numerical settings, calibration curve implementation, and prediction specification and verification.

Additional file 2 (.zip): Analysis code and frozen model. R analysis scripts, full-precision model parameters, stored spline knots, the standalone prediction function, and accompanying documentation. No participant-level data are included.

References

[1] Khan AA, Slart RHJA, Ali DS, et al. Osteoporotic Fractures: Diagnosis, Evaluation, and Significance From the International Working Group on DXA Best Practices. Mayo Clin Proc 2024;99:1127–41. https://doi.org/10.1016/j.mayocp.2024.01.011.

[2] LeBoff MS, Greenspan SL, Insogna KL, et al. The clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int 2022;33:2049–102. https://doi.org/10.1007/s00198-021-05900-y.

[3] Zheng X-Q, Xu L, Huang J, et al. Incidence and cost of vertebral fracture in urban China: A five-year population-based cohort study. Int J Surg. 2023;109(7):1910–1918. https://doi.org/10.1097/JS9.0000000000000411.

[4] Khan MA, Jennings JW, Baker JC, et al. ACR Appropriateness Criteria® Management of Vertebral Compression Fractures: 2022 Update. J Am Coll Radiol 2023;20:S102–24. https://doi.org/10.1016/j.jacr.2023.02.015.

[5] 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. https://doi.org/10.1007/s00198-025-07422-3.

[6] Khera TK, Hunt LP, Davis S, et al. A clinical tool to identify older women with back pain at high risk of osteoporotic vertebral fractures (Vfrac): a population-based cohort study with exploratory economic evaluation. Age Ageing 2022;51:afac031. https://doi.org/10.1093/ageing/afac031.

[7] Wang Y-N, Liu G, Wang L, et al. A Deep-Learning Model for Diagnosing Fresh Vertebral Fractures on Magnetic Resonance Images. World Neurosurg 2024;183:e818–24. https://doi.org/10.1016/j.wneu.2024.01.035.

[8] Chen J, Wei B, Cai J, et al. Construction of a nomogram prediction model for individualized prediction of vertebral compression fracture risk in postmenopausal osteoporosis population. Sci Rep 2026;16:9161. https://doi.org/10.1038/s41598-026-40375-z.

[9] Si Z, Hu K, Wang H, et al. An explainable machine learning model for predicting osteoporotic fragility fractures: a retrospective study in South China. Front Med (Lausanne) 2026;13:1869681. https://doi.org/10.3389/fmed.2026.1869681.

[10] Feller D, Wingbermühle R, Oei EHG, et al. Diagnostic prediction models for spinal fractures in individuals with spinal pain or trauma: a systematic review and meta-analysis. eClinicalMedicine 2025;88:103456. https://doi.org/10.1016/j.eclinm.2025.103456.

[11] Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 2024;384:e074819. https://doi.org/10.1136/bmj-2023-074819.

[12] Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J. 2023;65:2200302. https://doi.org/10.1002/bimj.202200302.

[13] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

[14] Meier L, Van De Geer S, Bühlmann P. The Group Lasso for Logistic Regression. Journal of the Royal Statistical Society Series B: Statistical Methodology 2008;70:53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x.

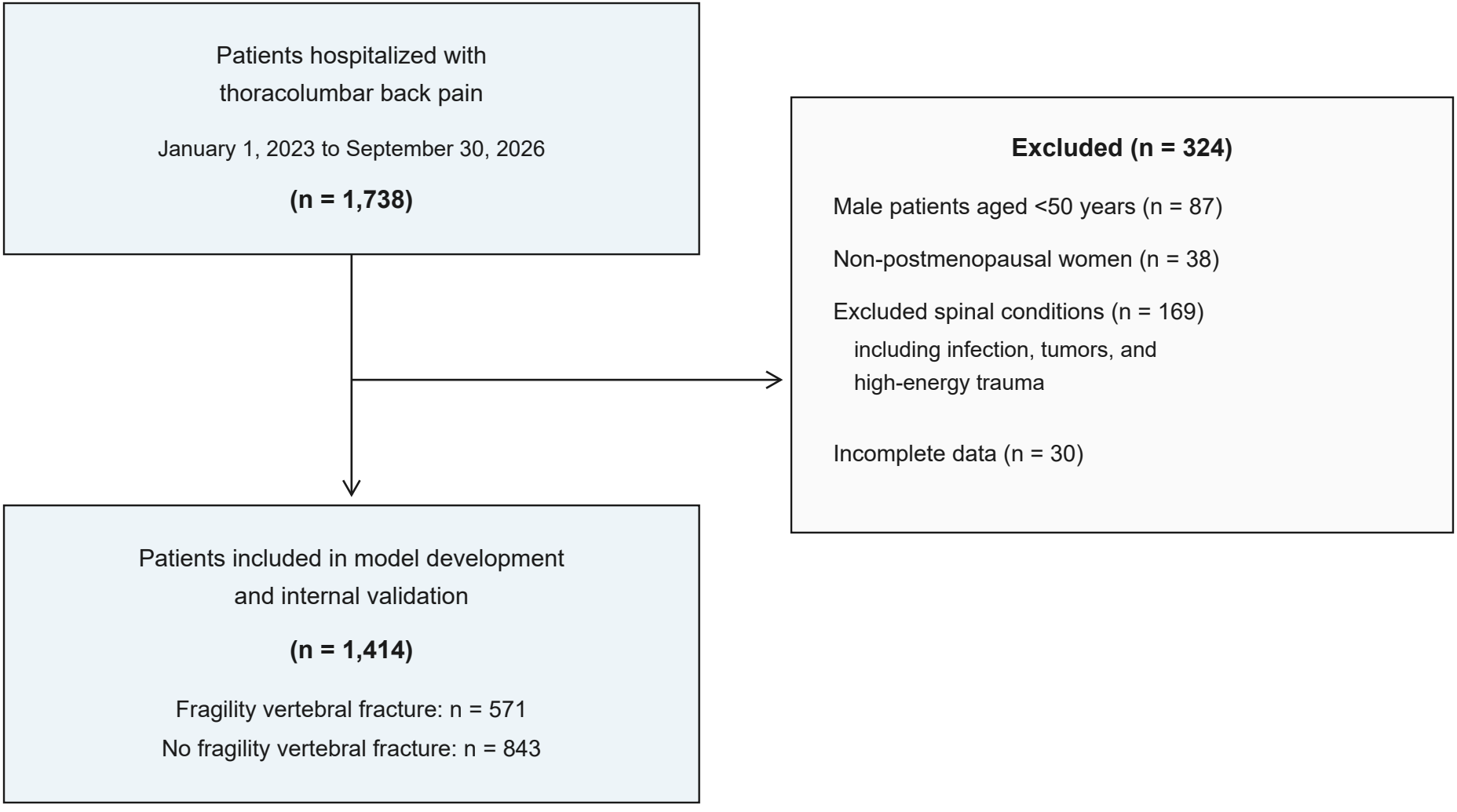
[15] Desquilbet L, Mariotti F. Dose-response analyses using restricted cubic spline functions in public health research. Stat Med 2010;29:1037–57. https://doi.org/10.1002/sim.3841.

[16] 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. https://doi.org/10.2307/2531595.

[17] Steyerberg EW, Harrell FE Jr, Borsboom GJJM, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–781.

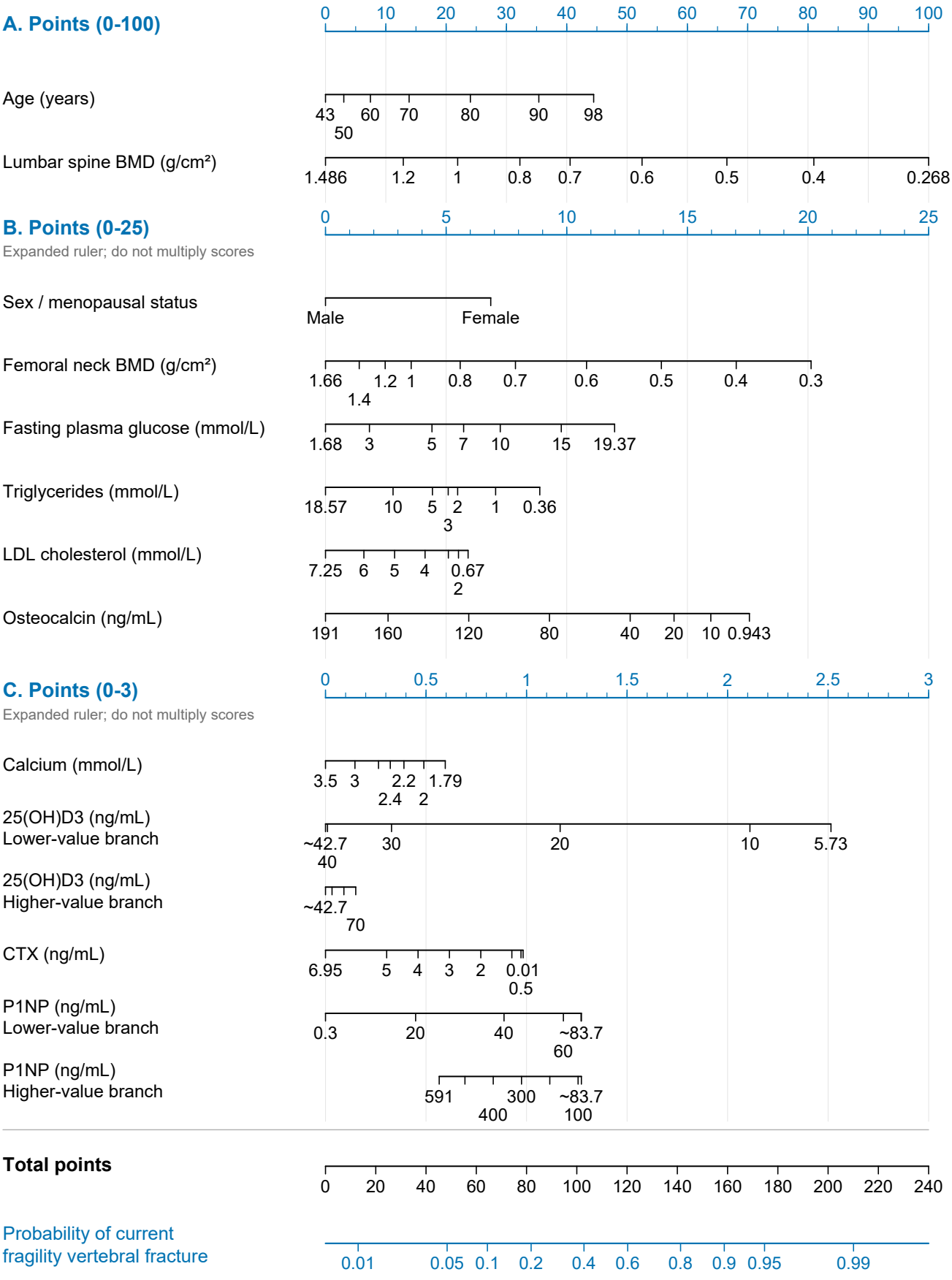
https://doi.org/10.1016/S0895-4356(01)00341-9.
[18] Efthimiou O, Seo M, Chalkou K, et al. Developing clinical prediction models: a step-by-step guide. BMJ 2024;386:e078276. https://doi.org/10.1136/bmj-2023-078276.
[19] Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3:18. https://doi.org/10.1186/s41512-019-0064-7.
[20] Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Making 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.
[21] Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput 2015;25:173–87. https://doi.org/10.1007/s11222-013-9424-2.
[22] Wen Z, Mo X, Zhao S, et al. Study on Risk Factors of Primary Non-traumatic OVCF in Chinese Elderly and a Novel Prediction Model. Orthop Surg 2022;14:2925–38. https://doi.org/10.1111/os.13531.
[23] Sun X, Jing P, Yang Y, et al. Risk prediction of osteoporotic vertebral compression fractures in postmenopausal osteoporotic women by machine learning modelling. Front Med (Lausanne) 2025;12:1664219. https://doi.org/10.3389/fmed.2025.1664219.
[24] Nie M, Chen Z, Shi L, et al. Prediction of new vertebral compression fracture within 3 years after percutaneous vertebroplasty for osteoporotic vertebral compression fracture: Establishment and validation of a nomogram prediction model. PLoS ONE 2024;19:e0303385. https://doi.org/10.1371/journal.pone.0303385.
[25] Slart RHJA, Punda M, Ali DS, et al. Updated practice guideline for dual-energy X-ray absorptiometry (DXA). Eur J Nucl Med Mol Imaging 2025;52:539–63. https://doi.org/10.1007/s00259-024-06912-6.
[26] Lin Y-H, Chou C-S, Teng MMH. The choice of region of interest after spinal procedures alters bone mineral density measurements. PLoS ONE 2023;18:e0285898. https://doi.org/10.1371/journal.pone.0285898.
[27] Jia Z, Tang M, Zhang X, et al. Changes in Bone Turnover Markers after Osteoporotic Vertebral Compression Fractures in Males and Females. BioMed Res Int 2022;2022:5381601. https://doi.org/10.1155/2022/5381601.
[28] Ivaska KK, McGuigan FE, Malmgren L, et al. Bone Turnover Marker Profiling and Fracture Risk in Older Women: Fracture Risk from Age 75 to 90. Calcif Tissue Int 2022;111:288–99. https://doi.org/10.1007/s00223-022-00996-8.
[29] Meier C, Eastell R, Pierroz DD, et al. Biochemical Markers of Bone Fragility in Patients With Diabetes. J Clin Endocrinol Metab 2023;108:e923–36. https://doi.org/10.1210/clinem/dgad255.
[30] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505.

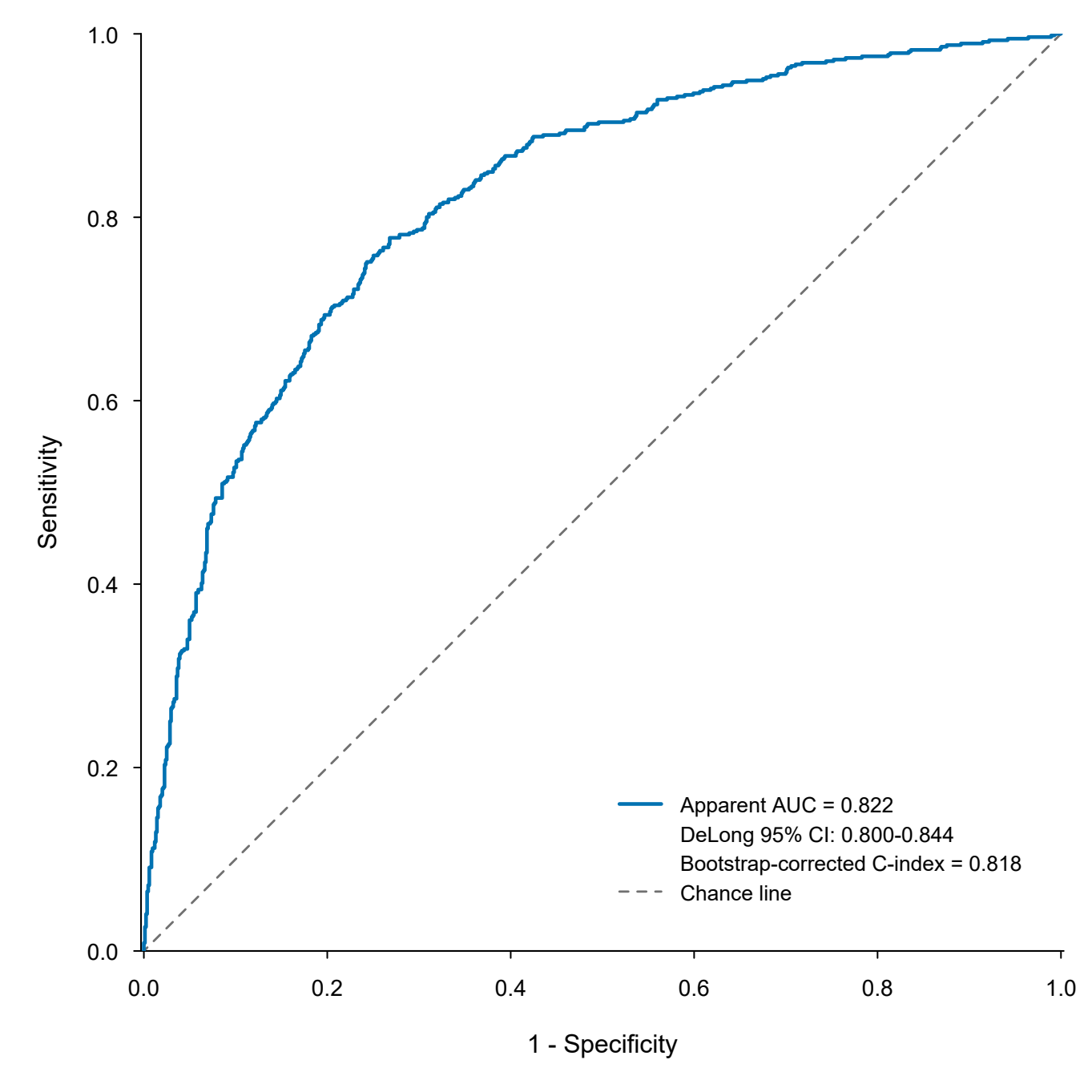
[31] Riley RD, Collins GS, Kirton L, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. BMJ 2025;388:e080749. https://doi.org/10.1136/bmj-2024-080749.
[32] Van Calster B, McLernon DJ, van Smeden M, et al. Topic Group “Evaluating diagnostic tests and prediction models” of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
[33] Khalid TY, Wilmott W, Shere C, et al. Acceptability of self-completion versus face-to-face use of a vertebral fragility fracture clinical decision tool for use in older people with back pain in the UK. Arch Osteoporos 2025;20:102. https://doi.org/10.1007/s11657-025-01586-5.
[34] Gregson CL, Armstrong DJ, Avgerinou C, et al. The 2024 UK clinical guideline for the prevention and treatment of osteoporosis. Arch Osteoporos 2025;20:119. https://doi.org/10.1007/s11657-025-01588-3.
[35] Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024;384:e074820. https://doi.org/10.1136/bmj-2023-074820.

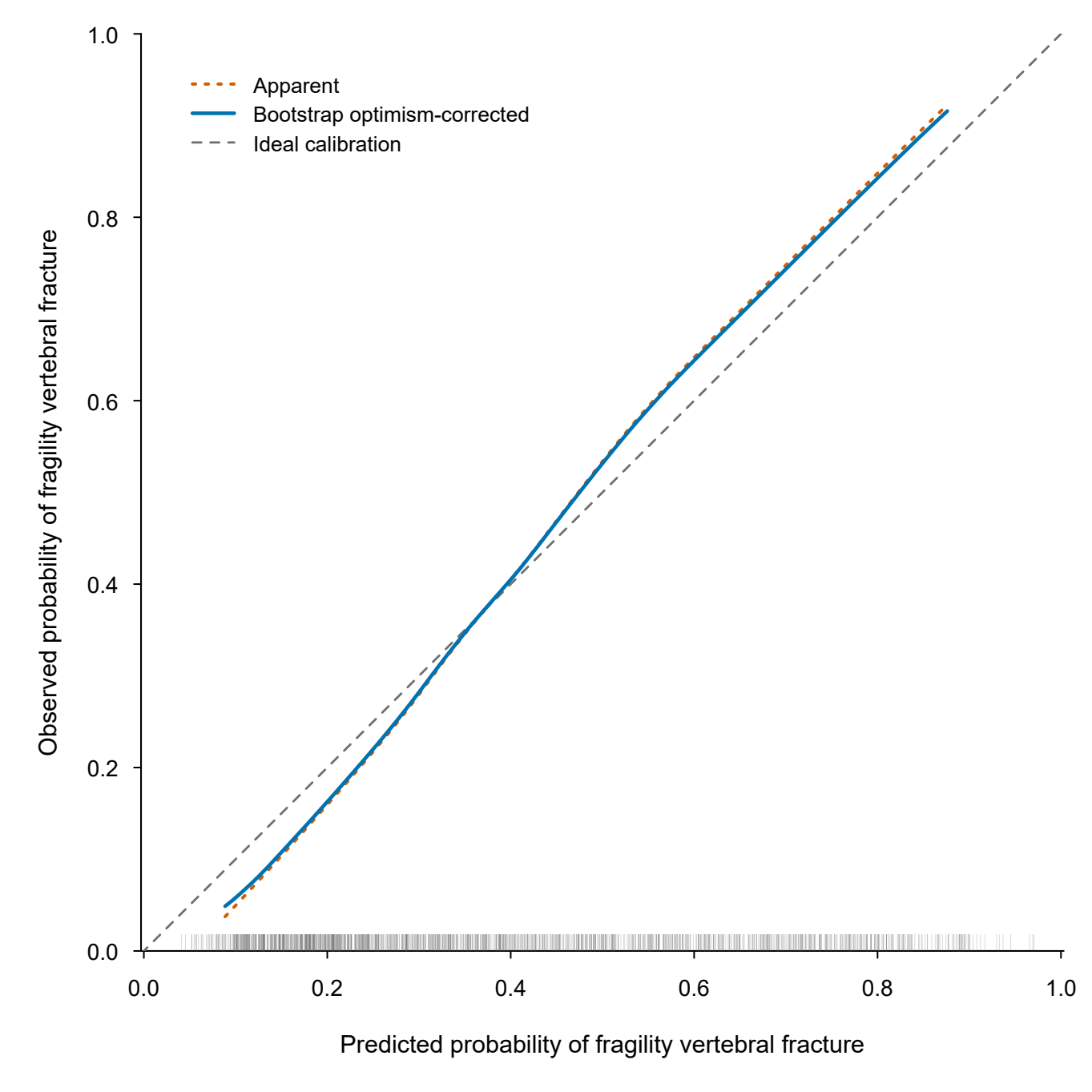

Patients hospitalized with
thoracolumbar back pain
January 1, 2023 to September 30, 2026
(n = 1,738)
Excluded (n = 324)
Male patients aged <50 years (n = 87)
Non-postmenopausal women (n = 38)
Excluded spinal conditions (n = 169)
including infection, tumors, and
high-energy trauma
Incomplete data (n = 30)
Patients included in model development
and internal validation
(n = 1,414)
Fragility vertebral fracture: n = 571
No fragility vertebral fracture: n = 843

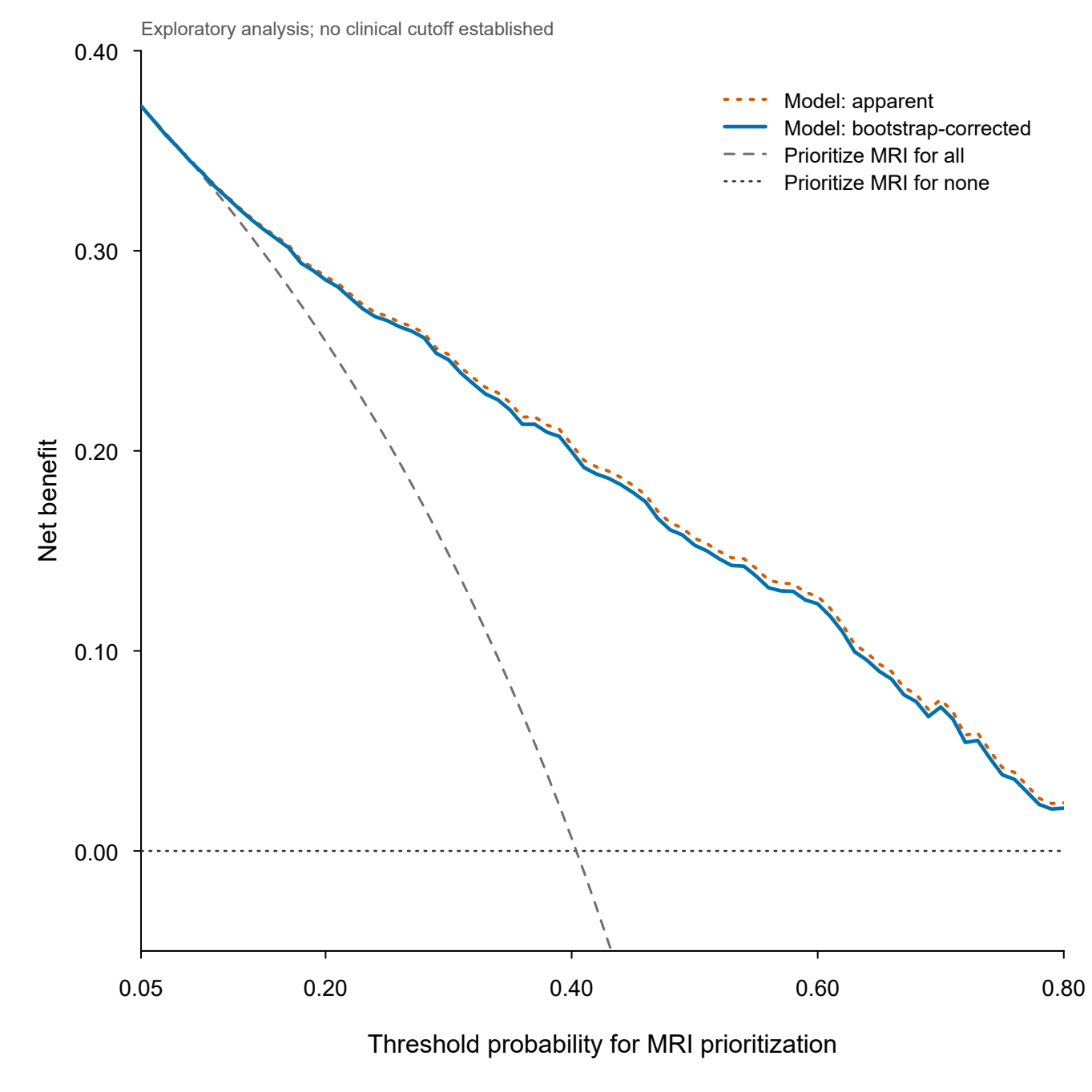
Use the ruler in each section; sum all 12 scores.

Use one branch per predictor for 25(OH)D3 and P1NP.




1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Apparent AUC = 0.822
DeLong 95% CI: 0.800-0.844
Bootstrap-corrected C-index = 0.818
Chance line


Apparent
Bootstrap optimism-corrected
Ideal calibration
Observed probability of fragility vertebral fracture
Predicted probability of fragility vertebral fracture
0.0
0.2
0.4
0.6
0.8
1.0


Exploratory analysis; no clinical cutoff established
Model: apparent
Model: bootstrap-corrected
Prioritize MRI for all
Prioritize MRI for none
Net benefit
0.40
0.30
0.20
0.10
0.00
0.05
0.20
0.40
0.60
0.80
Threshold probability for MRI prioritization

Supplementary Files

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

- Additionalfile1SupplementaryMethods.docx
- Additionalfile2AnalysiscodeandfrozenmodelMIT.zip