BMJ Open

The predictive value of machine learning on fracture risk in osteoporosis:a systematic review and meta-analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-071430
Article Type:	Original research
Date Submitted by the Author:	27-Dec-2022
Complete List of Authors:	Wu, Yanqian; Southeast University Chao, Jianqian; Southeast University Bao, Min; Southeast University Zhang, Na; Southeast University
Keywords:	Public health < INFECTIOUS DISEASES, Epidemiology < ONCOLOGY, Bone diseases < ORTHOPAEDIC & TRAUMA SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

The predictive value of machine learning on fracture risk in osteoporosis:a systematic review and meta-analysis

Yanqian Wu[1], Jianqian Chao[1*] , Min Bao[1],Na Zhang[1]

[1] Key Laboratory of Environmental Medicine Engineering of Ministry of Education, School of Public Health, Southeast University, Nanjing210009, P.R. China

wyqmm0523@163.com(Yanqian Wu,PhD), chaoseu@163.com (Jianqian Chao,PhD), 1689586985@qq.com(Min Bao,PhD),540923028@qq.com(Na Zhang,PhD)

[*]**Corresponding author:** Jianqian Chao, Key Laboratory of Environmental Medicine Engineering of Ministry of Education, School of Public Health, Southeast University, 87 Dingjiaqiao Road, Gulou District, Nanjing 210009, China.

Tel: +86 025 86424437/13813955976.

E-mail: chaoseu@163.com

Word count 3036.

Abstract

Objectives: Early identification of fracture risk in osteoporosis patients is essential. In recent years, machine learning methods have been gradually introduced into this field; however, their predictive value remains controversial. Therefore, we conducted this systematic review and meta-analysis to explore the predictive value of machine learning on fracture risk in patients with osteoporosis.

Setting: Eligible studies were collected from four databases (PubMed, Embase, Cochrane Library and Web of Science) until June. 20th, 2022. A meta-analysis of the C-index was performed using random-effects models, while bivariate mixed-effects models were used for sensitivity and specificity. In addition, a subgroup analysis was performed according to the types of machine learning models and fracture sites.

Participants: Patients were diagnosed with osteoporosis,ML was applied to predict fracture risk and at least one measure of model performance (discrimination or calibration) was reported.

Primary and secondary outcome measures: Outcome variables were measured as C index or AUC , sensitivity(%), specificity(%).

Results: Forty-six studies were included in our meta-analysis, involving 8,869,283 patients, 89 prediction models specifically developed for osteoporosis populations, and 39 validation sets. These models' most commonly used predictors were age, fracture history, body mass index (BMI), bone mineral density (BMD), radiomics data, weight, height, bone mineral density T-score, history of falls, gender, and other chronic diseases. Overall, the C-index of machine learning was 0.76 (95% CI: 0.73,0.79) and 0.74 (95% CI: 0.71,0.77) in the training set and validation set, respectively; and the sensitivity was 0.81 (95% CI: 0.75,0.86)) and 0.82 (95% CI: 0.75,0.87); and the specificity was 0.76 (95% CI: 0.80,0.81) and 0.83 (95% CI: 0.72,0.90), respectively.

Conclusions: Machine learning has an ideal predictive value on fracture risk in patients with osteoporosis and can be used as a potential tool for early identification of fracture risk in osteoporosis patients.

Keywords Osteoporosis Machine learning Fractures Meta-Analysis

Strengths and limitations of this study

- Rigorous literature search and methodology followed to provide reliable estimates.
- We performed a quantitative synthesis to enhance the comparability of ML models.
- ML has an ideal predictive value for fracture risk in patients with osteoporosis.
- Most of the included studies (64%) had a high risk of bias.

Introduction

Osteoporosis is a systemic metabolic bone disease characterized by decreased bone mass and degraded bone tissue microarchitecture, leading to an increased risk of bone fragility and fracture (WHO,1994)[1]. Osteoporosis features low quality of life, high cost, high disability rate and high morbidity rate[2]. It has become a global health problem that threatens human health. According to the World Health Organization, osteoporosis is the second-most serious health issue after cardiovascular diseases[3].

Machine learning (ML) is a subfield of artificial intelligence, which enables computers to "learn" through programs[4].ML models have been applied in the field of osteoporosis and provided new opportunities for fracture prediction. A review by Ferizi et al. (2019) summarized relevant researches on the application of artificial intelligence to the prediction of osteoporosis. It drew a conclusion that new methods for automatic

image segmentation and fracture risk prediction showed promising clinical value[5]. A systematic review by Smets et al. (2021) reviewed the state-of-the-art ML methods and their application in osteoporosis diagnosis and fracture prediction[6]. Another review by Anam et al. (2021) explored the prediction of ML for osteoporosis in trabecular bone. This paper presented a detailed systematic review of ML prediction for trabecular bone diseases using magnetic resonance imaging (MRI) both from a methodology-driven and application perspective[7]. Most studies focused on the role of ML either to predict an indicator of osteoporosis, such as BMD or fractures, or as a tool for automatic segmentation of the images of patients at risk of osteoporosis, rather than predicting osteoporotic fractures.

The present study evaluated the predictive value of ML for fracture risk in osteoporosis patients and provided an evidence-based medical basis for the application of ML in clinical practice.

Materials and methods

This study was conducted in accordance with the Preferred Items of Systematic Review and Meta Analysis (PRISMA) statement (**Table S1**)[8]. The protocol was registered on the international prospective register of systematic reviews (PROSPERO) (Registration No. CRD42022346896).Relevant studies were retrieved from Pubmed, Embase, Cochrane Library, and Web of Science, and the retrieval was as of June 20, 2022. Two researchers independently searched the literature, the search strategy is shown in **Table S2**.

Inclusion criteria were as follows: (1) patients were diagnosed with osteoporosis (OP); (2) ML was applied to predict fracture risk; (3) at least one measure of model performance (discrimination or calibration) was reported; (4) study population included adult patients older than 18 years, mainly including adults, the elderly and postmenopausal women.Exclusion criteria were as follows: (1) studies that only analyzed risk factors without building complete ML models; (2) studies that only included osteoporosis but did not mention fracture risk; (3) studies without available full text (or only abstract available) or data; (4) meta-analyses, reviews, case reports,editorial materials, letters, protocols, errata, and notes.

Two researchers independently extracted data using standardized tables. Any studies excluded after full-text review have been recorded with reasons for their exclusion. The list of extracted items was based on the CHARMS checklist [9], and two data extraction sheets were prepared for developed and validated models. Finally, the extracted data involved the first author, year of publication, country, study design, data source,population group, gender, age, fracture sites, types of predictive models, number of predictors, and outcomes.The risk of bias was assessed using the Prediction Model Risk Of Bias Assessment Tool (PROBAST). The PROBAST contained a large number of questions in four distinct domains: participants, predictors, outcomes, and statistical analysis, reflecting the overall risk of bias and usability[10].

A meta-analysis of the metrics (C-index and accuracy) was performed to evaluate ML models. If the C-index did not report 95% confidence intervals (CI) and standard errors, we estimated the standard errors in reference to the study by Debray TP et al.[11]. A C-index of 0.5 indicates that a model performs no better than chance; 0.6 to 0.7 is considered modest discrimination; 0.71 to 0.8 indicates very good discrimination; and greater than 0.8 is considered strong [12]. When there was a lack of accuracy in the original studies, we calculated it based on the sensitivity and specificity in combination with the number of samples in each subgroup and the number of modeling samples[11]. Considering the differences in variables, ML models and variation parameters included in the studies, the random effects model was preferred for the meta-analysis of C-index, and the bivariate mixed effects model was used for the meta-analysis of sensitivity and specificity.Our

meta-analysis were performed using the software Stata 15.1 (Stata Corporation, College Station, TX, USA) and R4.2.0 (R Development Core Team, Vienna, http://www.R-project.org). A *p* value less than 0.05 was considered statistically significant.

Results

A total of 6,851 studies(1395 from PubMed, 2644 from Embase Database, 157 from Cochrane Library, and 2655 from Web of Science) were searched from the databases. After removing duplicates and screening titles and abstracts, 340 articles remained. According to a full-text review, 46 articles[13-58] were eligible. Forty-four articles presented the development of one or more prediction models for osteoporotic fracture, and twenty-four articles described the validation of one or more models. The search process is shown in **Figure 1**.

Forty-six studies were ultimately included in our meta-analysis, involving 8,869,283 patients and 2,611,525 fracture cases. The majority of studies were conducted in U.S.(n = 10) and Canada (n = 6), and most studies were cohort studies (n = 40) or case-control studies (n = 6). The median age of osteoporosis patients was 69 years (ranging from 48.5 to 84). Most study samples covered postmenopausal women (n = 15). The fracture sites included hip fracture (n = 12), vertebral fracture (n = 11) and multi-site fracture (n = 23). Most studies were based on clinical hospital data (n=14), while some used questionnaire collection data (n=9), osteoporosis registry data (n=9), electronic health records(n=6), and administrative data (n=6). Only 11 articles elucidated the cross-validation method used by their models. The baseline characteristics are shown in **Table 1**.

A total of 89 prediction models were specifically developed for the osteoporosis population in 44 articles, and 39 validation sets were performed in 24 studies. For most of these models, the C-index or the area under the receiver operating characteristic curve (AUC) ranged from 0.58 to 0.98. **Table S3** shows all model-development and validation studies on the ML models for outcome prediction in patients with osteoporosis. The number of participants ranged from 28 to 6,329,986 (median 1026), and the number of fracture events varied from 14 to 2468694 (median 143). Among all the identified prediction models, the logistic regression (35%) was the most commonly studied model, followed by the survival model (17%).

Table 1. Characteristics of included studies in meta-analysis

Author	Year	Country	Data source	Sample population type	Average age,years	Fracture site	ML models
Wu, Q[23]	2020	USA	gene database	older men	74.8	multiple fractures	LR ANN RF BT
Villamor, E[24]	2020	Spain	clinical hospital data	postmenopausal women	81.4	hip fracture	LR SVM ANN RF
Van Geel, Tacm[25]	2011	Netherlands	questionnaire collection data	postmenopausal women	62	vertebral fracture	SM
Ulivieri, F. M[26]	2021	Italy	clinical hospital data	patient	48.5	vertebral fracture	ANN
Yoda, T[27]	2021	Japan	clinical hospital data	patient	77.6	vertebral fracture	CNN
Jiang, X. Z[28]	2013	USA	clinical hospital data	older women	61.4	multiple fractures	LR
Schousboe, J. T[29]	2014	USA	clinical hospital data	older women	75	vertebral fracture	LR
Sandhu, S. K[30]	2010	Australia	electronic health record data	patient	74	multiple fractures	LR
Rubin, K. H[31]	2018	Denmark	administrative data	inhabitant	61.4	multiple fractures	LR
Pluskiewicz, W[32]	2010	Poland	osteoporosis registry data	postmenopausal women	68.5	multiple fractures	LR
Jang, E. J[33]	2016	Korea	questionnaire collection data	inhabitant	61	multiple fractures	LR
Barret A. Monchka[34]	2021	Canada	osteoporosis registry data	inhabitant	75	vertebral fracture	CNN
Mehta, S. D[35]	2020	USA	clinical hospital data	patient	69	vertebral fracture	SVM
Langsetmo, L[36]	2011	Canada	questionnaire collection data	inhabitant	67.6	multiple fractures	SM
Ioannidis, G[37]	2017	Canada	electronic health record data	elderly	61	multiple fractures	DT LR
K. K. Nishiyama[38]	2013	Canada	questionnaire collection data	postmenopausal women	73	multiple fractures	SVM
Kruse, C[39]	2017	Denmark	administrative data	inhabitant	60.8	hip fracture	DT NB
Kolanu, N[40]	2021	Australia	electronic health record data	patient	73.4	multiple fractures	ANN
Kim, H. Y[41]	2016	Korea	administrative data	inhabitant	60	multiple fractures	SM
Hsieh, C. I[42]	2021	China	clinical hospital data	patient	72.2	hip fracture	LR
Hong, N[43]	2021	Korea	clinical hospital data	older women	73	hip fracture	SM BT SVM
Ho-Le, T. P[44]	2017	Australia	osteoporosis registry data	postmenopausal women	69.1	hip fracture	ANN LR KNN SVM
Henry, M. J[45]	2011	Australia	osteoporosis registry data	older women	74	multiple fractures	LR
Galassi, A[46]	2020	Spain	electronic health record data	postmenopausal women	81.4	hip fracture	DT LR RF SVM
FitzGerald, G[47]	2014	California	questionnaire collection data	postmenopausal women	67	multiple fractures	SM
Ferizi, U[48]	2019	USA	osteoporosis registry data	postmenopausal women	62	multiple fractures	LR

Enns-Bray, W. S[49]	2019	USA	clinical hospital data	older women	77.2	hip fracture	LR
Engels, A[50]	2020	Germany	administrative data	patient	75.6	hip fracture	SVM RF LR BT
De Vries, B. C. S[51]	2021	The Netherlands	clinical hospital data	patient	68	multiple fractures	ANN RF SM
Cheung, E. Y[52]	2012	China	electronic health record data	postmenopausal women	62	multiple fractures	SM
Chanplakorn, P[53]	2021	Thailand	osteoporosis registry data	postmenopausal women	68.5	vertebral fracture	SM
Bredbenner, T. L[54]	2014	USA	clinical hospital data	older men	65	hip fracture	LR
Beyaz, S[55]	2020	Turkey	osteoporosis registry data	patient	74.9	multiple fractures	ANN
Berry, S. D[56]	2018	USA	administrative data	inhabitant	84	hip fracture	SM
Beaudoin, C[57]	2021	Canada	administrative data	elderly	75.1	multiple fractures	SM
Baleanu, F[58]	2022	Belgium	clinical hospital data	postmenopausal women	70.1	multiple fractures	LR
Almog, Y. A[59]	2020	USA	electronic health record data	patient	50	vertebral fracture	ANN
Zagorski, P[60]	2021	Poland	questionnaire collection data	postmenopausal women	65.2	hip fracture	LR
Diez-Perez, A[61]	2007	Spain	questionnaire collection data	postmenopausal women	72.3	multiple fractures	SM
Lix, L. M[62]	2018	Canada	osteoporosis registry data	older women	65.6	multiple fractures	LR
Li, Q. J[63]	2021	China	clinical hospital data	patient	70	multiple fractures	LR
Lee, S[64]	2008	Korea	osteoporosis registry data	older women	65	multiple fractures	SVM
Jacobs, J. W. G[65]	2010	Portugal	questionnaire collection data	inhabitant	66	vertebral fracture	LR
Eller-Vainicher, C[66]	2011	Italy	questionnaire collection data	postmenopausal women	68	vertebral fracture	ANN LR
Zhong, B. Y[67]	2017	China	clinical hospital data	patient	72	vertebral fracture	SM
Xiao, X[68]	2021	USA	gene database	postmenopausal women	64.5	hip fracture	SM

*LR: Logistic Regression;ANN:artifificial neural network;SVM = support-vector machine; CNN = convolutional neural network; kNN = k-nearest neighbors; RF = random forests;DT=decision tree;NB=Naive Bayes;BT=Boosted tree;SM=Survival model;AUC=area under the receiver operating characteristic curve;ROC=receiver operating characteristic;PPV=positive predictive value;NPV=negative predictive value.

We roughly classified the fracture risk predictors into seven types: demographics/fracture history, physical examination, lifestyle, comorbidity, drug and nutrient intake, radiomics, and mental state. The classification of fracture risk predictors is presented in **Table 2**. The most commonly used predictors in these models were age (n=59), past fracture history (n=32), body mass index (BMI) (n=31), bone mineral density (BMD) (n=25), radiomics data (n=23), weight (n=23), height (n=21), bone mineral density T-score (n=20), history of falls (n=22), gender (n=17) and other chronic diseases (n=12).

Table 2. Main sorts of predictors included in 89 developed models for osteoporosis patients

Predictors	Number of models
Demographics and fracture history	
Age	59
History of falls	22
Past fracture history	32
Sex	17
Women's menopause age	6
Genetic risk score (GRS)	5
Physical examination	
Body mass index,BMI	31
Bone mineral density，BMD	25
Weight	23
Height	21
Bone mineral density t-score	20
Grip	1
Transfer ability	1
Lifestyle	
Alcohol consumption	7
Smoking	6
Physical activity index	4
Frequent sun exposure	3
Lack of physical exercise	1
Vagrant event	1
Comorbidity	
Other chronic diseases	12
Osteoporosis	8
Rheumatoid arthritis	6
Fracture type	4
Backache	2
Drug and nutrient intake	
Use of hormonal drugs	6

	Calcium intake	3
	Intake of other drugs	2
Radiomics		
	Radiomic data	23
Mental state		
	Cognitive performance	3
	Anxiety/depression	2

The risk of bias assessment of the model-development studies is summarized in **Figure 2**. More than half of these studies had a high risk of bias (n=48; 64%). The risk of bias in most studies was low in terms of participants, predictors and outcome. However, a high or unclear risk of bias in the statistical analysis was observed in all model development and validation studies. More details are shown in **Table S4**.

Thirty-nine studies involving 89 models were included in the meta-analysis of the C-index. Since substantial heterogeneity was present, we performed subgroup analyses based on fracture site and model type. **Table 3** shows the results of the meta-analysis of C-index for ML models in patients with osteoporosis. The forest plot of C-index is presented in **Figure 3**. The pooled C-index was 0.76 (95 % CI: 0.73, 0.79) (I^2 = 99.8%, P < 0.001) in the training set and 0.74 (95 % CI: 0.71, 0.77) (I^2 = 99.8%, P < 0.001) in the validation set. In the training set, Naive Bayes showed the highest predictive performance (pooled C-index = 0.92), followed by artificial neural network (pooled C-index = 0.82), decision trees (pooled C-index = 0.78) and logistic regression (pooled C-index = 0.77). Furthermore, models for vertebral fracture (pooled C-index = 0.79) and hip fracture (pooled C-index = 0.78) performed better than those for multi-site fracture (pooled C-index = 0.72). However, in the validation set, logistic regression (pooled C-index = 0.82) showed the best performance, closely followed by support vector machines (pooled C-index = 0.78), artificial neural network (pooled C-index = 0.73) and boosted tree (pooled c-index = 0.70). Models for vertebral fracture (pooled C-index = 0.84) performed better than those for hip fracture (pooled C-index = 0.73) and multi-site fracture (pooled C-index = 0.71).Across these studies, we extracted 53 estimates of balanced accuracy (the average of the reported sensitivity and specificity), ranging from 0.66 to 1.00. As presented in **Table 4** , the mean sensitivity and specificity of models in the training set were 0.81 (95 % CI: 0.75, 0.86) (I^2 = 99.0%, P < 0.001) and 0.82 (95 % CI: 0.75, 0.87) (I^2 = 99.9%, P < 0.001), respectively. Moreover, the mean sensitivity and specificity of models in the validation set were 0.76 (95 % CI: 0.80, 0.81) (I^2 = 98.9%, P < 0.001)and 0.83 (95 % CI: 0.72, 0.90) (I^2 = 99.9%, P < 0.001), respectively.

Table 3. Results of meta-analysis of C-index statistics for subgroup analysis by fracture site and machine learning type

subgroup	Training dataset		Validation dataset	
	N	C-statistic(95% CI)	N	C-statistic(95% CI)
Fracture site				
Vertebral fracture	13	0.79(0.74,0.85)	5	0.84(0.67,1.00)
Hip fracture	20	0.78(0.73,0.84)	9	0.73(0.65,0.81)
Multi-site fracture	26	0.72(0.70,0.74)	16	0.71(0.65,0.77)
Model type				
LR	27	0.77(0.72,0.82)	8	0.82(0.75,0.88)
ANN	6	0.82(0.74,0.90)	4	0.73(0.56,0.91)
RF	3	0.70(0.68,0.72)	3	0.66(0.59,0.73)
SVM	4	0.76(0.61,0.91)	3	0.78(0.59,0.96)
DT	2	0.78(0.56,0.99)	1	0.69(0.67,0.70)
NB	1	0.92(0.90,0.95)	-	
Survival model	11	0.69(0.67,0.71)	8	0.68(0.67,0.69)
Boosted tree	5	0.73(0.71,0.74)	3	0.70(0.69,0.71)
Overall	59	0.76(0.73,0.79)	30	0.74(0.71,0.77)

Table 4. Results of sensitivity and specificity subgroup analysis by fracture site and machine learning type

subgroup	Training dataset			Validation dataset		
	N	Sensitivity(95% CI)	Specificity(95% CI)	N	Sensitivity(95% CI)	Specificity(95% CI)
Fracture site						
Vertebral fracture	9	0.73(0.59,0.83)	0.92(0.86,0.95)	4	0.87(0.70.0.95)	0.97(0.94,0.98)
Hip fracture	13	0.90（0.82,0.94）	0.82(0.75,0.88)	6	0.84（0.77,0.89）	0.85(0.80,0.89)
Multi-site fracture	15	0.76(0.65,0.84)	0.72(0.57,0.83)	8	0.66（0.61,0.70）	0.69(0.53,0.81)
Model type						
LR	18	0.72(0.64,0.78)	0.77(0.69,0.84)	6	0.74(0.64,0.83)	0.80(0.59,0.92)
ANN	7	0.89(0.77,0.95)	0.92(0.81,0.97)	4	0.82(0.69,0.90)	0.89(0.78,0.95)

RF	1	0.84	0.91	1	0.70	0.47
SVM	5	0.86(0.72,0.94)	0.84(0.61,0.95)	3	0.79(0.72,0.85)	0.89(0.79,0.94)
DT	2	0.97(0.53,1.00)	0.70(0.67,0.73)	-		
NB	1	0.88	0.81	-		
kNN	1	1.00	0.83	1	0.81	0.79
Survival model	1	0.81	0.52	-		
Boosted tree	1	0.59	0.67	1	0.70	0.95
Overall	37	0.81(0.75,0.86)	0.82(0.75,0.87)	16	0.76(0.80,0.81)	0.83(0.72,0.90)

Discussion

The present systematic review provides an overview of all currently available fracture risk prediction models developed or validated in the osteoporosis population to assess the overall ML models for fracture risk. The most commonly used predictors in these models are age, gender, weight, height, BMI, past fracture history, BMD, radiomics data, bone mineral density T-score, history of falls, and other chronic diseases. In general, most predictors included in model-development studies are traditional risk factors. A recent study showed that the most common risk factors for fragility fractures contained decreased bone mineral density, age, gender, low BMI, history of fragility fractures, family history of hip fractures, history of glucocorticoid therapy, smoking, excessive alcohol consumption,lack of vitamin D, early menopause and immobility[59]. This is consistent with some of the common fracture predictors identified in our study. Our study also finds that radiomics data are frequently used as fracture predictors in ML models for osteoporosis. A retrospective, single-center, preliminary investigation by Lim et al. reported that the predictive performance of ML analysis with radiomics features and Abdomen-pelvic CT to diagnose osteoporosis showed high validity with more than 93% accuracy, specificity, and negative predictive value[60].

In terms of the models in the training sets, Naive Bayes performs best (pooled C-index = 0.92), followed by artificial neural network(ANN), decision trees, and logistic regression. As Naive Bayesian algorithm is not affected by missing data and requires less training data set, it is widely used in the medical field as one of the most effective learning algorithms in data mining[61]. The Naive Bayes algorithm can be used to analyze a large number of unknown data. Bayesian models are the statistical method of choice when resource settings are low, especially when sample size and budget are limited[62]. Additionally, ANN is widely used in radiation, urology, inspection, and cardiovascular fields. With its computer processing techniques, ANN can assist in diagnosing various diseases and provide guidance for clinical medication[63-64].

In the validation sets, logistic regression (pooled C-index = 0.82) performed best, closely followed by support vector machines (SVM), ANN and boosted tree. Possibly because logistic regression is very efficient and easy to implement, it's easy to understand, and it outputs calibrated predicted probabilities. An article on prediction models for the outcomes in patients with chronic obstructive pulmonary disease also revealed that logistic regression (n=111; 72%) was the most frequently used modeling method[65]. This is the same as recent findings reported by Kushan et al.[66], their ML models with logistic regression outperformed those with

random forest and decision trees. Moreover, SVM adapts well to small samples and high-dimensional data with a low misclassification rate, and therefore can be used for classification and regression analysis[67].

Most included studies reported multiple metrics, such as sensitivity, specificity, AUC and ROC. The mean sensitivity and specificity of the models in the training set model were 0.81 (95 % CI: 0.75, 0.86) and 0.82 (95 % CI: 0.75, 0.87), respectively, greater than that of the models in the validation set. Therefore, fracture risk prediction using ML models still has promotion space in external validation. However, a single performance metric such as AUC or ROC is insufficient to recommend the application of ML models into clinical practice[6], and it is necessary to combine other multiple metrics.

This systematic review and meta-analysis summarized a large number of studies to comprehensively evaluate the predictive value of ML on fracture risk in patients with osteoporosis. It elucidated the characteristics of the established models and validation studies on existing models. We performed a quantitative synthesis that was never done in previous studies, enhancing the comparability of these models. Furthermore, only the C-index was reported in most predictive models[65,68], but our study used bivariate mixed-effects models for sensitivity and specificity analyses. In the training dataset, sensitivity of hip fracture (pooled C-index = 0.90), performed best, closely followed by multi-site fracture, vertebral fracture. For patients with hip fractures, missed and misdiagnosed radiographs can lead to poor prognosis[69]. ML models have been increasingly used to identify hip fracture risk with high accuracy[29]. ML has a stronger ability to recognize images and can provide a diagnostic scheme with high accuracy for inexperienced physicians to refer to.

Some limitations still need to be considered in the present study. The risk of bias assessment demonstrated that most studies (64%) had a high risk of bias, whether they involved the development or external validation of a prediction model for the osteoporosis population. The main bias came from the analysis, because most studies did not properly handle continuous and categorical variables and reported no method for processing missing values. Only two articles reported the use of median imputation to deal with missing values[13,29], while others did not mention how to deal with missing values. These shortcomings in the methodological quality may be due to a lack of guidelines for the standard reporting of risk prediction studies at that time. In addition, some models were reported with little information that was insufficient for external validation by other researchers, let alone to be implemented in clinical practice. For example, only 11 articles used the K-fold cross-validation method to improve the accuracy of their algorithms[13,14,17,25,28,29,32,34,40,44,45], but most of the eligible articles did not. Models without stringent validation offer limited applicability[66]. Furthermore, we observed large between study heterogeneity in the meta-analyses of C statistics. Potential sources of heterogeneity could be the differences in patients's characteristics, data sources and analysis methods across the validation studies. More than 30% of the research data in included studies came from clinical studies, and clinical data are heterogeneous and usually imbalanced. At last, most ML models did not report balanced accuracy, lack calibration or external validation or decision curves that the generalization of the models need to be further verified, so we suggest that subsequent research can be further refined.

Although our findings lack some evidence due to the limitations mentioned above, the present study can still provide meaningful recommendations for future research and practice. First, as existing ML models for fracture prediction focus on populations in Western countries, more external validation studies on other

populations are needed to widen their application. Second, ML models can identify valuable evidence to support clinicians in making more accurate judgments in highly complex decision-making processes and have certain clinical application value[67]. Future researches need to be more rigorous, robust, and comprehensive when assessing the quality of its clinical application and impact on clinicians and patients. Third, the advances in emerging technologies such as ML have opened a new era of clinical medical research, providing new directions for solving intricate problems with classical statistical methods. However, clinicians currently are not skillful in using such emerging technologies. Therefore, clinicians should be encouraged to improve their ability to use ML so that medical research can become more accurate with the help of ML.

In conclusion, ML has an ideal predictive value for fracture risk in patients with osteoporosis and can be used as a potential tool for early identification of fracture risk in patients with osteoporosis. Therefore, in the future, we can try to construct multi-racial cases, including machine learning of predictors of living habits, eating habits and social background, and then rely on machine learning to develop simple risk assessment tools adapted to multiple races.

Declarations

Funding

This work was supported by National Natural Science Foundation of China (81872711), 2022 Jiangsu University Philosophy and Social Science Research Major Project （2022SJZD141).

Authors Contributions

Yanqian Wu and Jianqian Chao designed the review,developed the inclusion criteria, screened titles and abstracts, appraised the quality of included papers, and drafted the manuscript. Min Bao and Na Zhang collected and cleaned the data,also analysed the data. All authors critically reviewed drafts and approved the final manuscript.

Competing interests

Yanqian Wu, Jianqian Chao , Min Bao and Na Zhang declare that they have no conflict of interest.

Ethical Approval

This study does not involve human participants and ethical approval was not required.

Patient and Public Involvement

No patient involved.

Data sharing statement

The data generated during this study are not publicly available, but a de-identified analytical file is available from the corresponding author on reasonable request.

References

1. Kanis, JA (1994) Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. Osteoporosis international 4:368-381. https://doi.org/10.1007/BF01622200
2. Veronese N, Kolk H, Maggi S(2020) Epidemiology of Fragility Fractures and Social Impact. In: Falaschi P, Marsh D, editors. Orthogeriatrics: The Management of Older Patients with Fragility Fractures [Internet]. 2nd ed. Cham (CH): Springer,2021. Chapter 2. https://doi.org/10.1007/978-3-030-48126-1_2
3. Piscitelli P, Feola M, Rao C, Celi M, Gasbarra E, Neglia C, Quarta G, Liuni FM, Parri S, Iolascon G, Brandi ML, Distante A, Tarantino U (2014)Ten years of hip fractures in Italy: For the first time a decreasing trend in elderly women. World J Orthop 18;5(3):386-91. https://doi: 10.5312/wjo.v5.i3.386.
4. Battina D S (2019) Artificial Intelligence in Software Test Automation: A Systematic Literature Review. International Journal of Emerging Technologies and Innovative Research (www. jetir. org| UGC and issn Approved) ISSN 2349-5162.https://doi.org/ 10.2196/26448
5. Ferizi U, Honig S, Chang G (2019) Artificial intelligence, osteoporosis and fragility fractures. Current opinion in rheumatology 31(4): 368.https://doi.org/ 10.1097/BOR.0000000000000607
6. Smets J, Shevroja E, Hügle T, Leslie WD, Hans D (2021) Machine learning solutions for osteoporosis—a review. Journal of Bone and Mineral Research 36(5): 833-851.https://doi.org/ 10.1002/jbmr.4292
7. Anam M, Hussain M, Nadeem M W et al (2021) Osteoporosis prediction for trabecular bone using machine learning: a review. Computers, Materials & Continua (CMC) 67(1).https://doi.org/ 10.32604/cmc.2021.013159
8. Liberati A, Altman D G, Tetzlaff J et al (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 62(10): e1-34.https://doi.org/ 10.1371/journal.pmed.1000100
9. Palazón‐Bru A, Martín‐Pérez F, Mares‐García E et al (2020) A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models.Statistics in Medicine 39(23): 3207-3225.https://doi.org/ 10.1002/sim.8660
10. Nagendran M, Chen Y, Lovejoy C A et al (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368.https://doi.org/ 10.1136/bmj.m689
11. Debray T P A, Damen J A A G, Riley R D et al (2019)A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Statistical methods in medical research 28(9): 2768-2786.https://doi.org/ 10.1177/0962280218785504
12. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD (2018) Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?Stat Methods Med Res 27:3505-22. https://doi.org/10.1177/0962280217705678
13. Wu Q, Nasoz F, Jung J, Bhattarai B, Han MV (2020) Machine learning approaches for fracture risk assessment: a comparative analysis of genomic and phenotypic data in 5130 older men. Calcified tissue international 107(4): 353-361.https://doi.org/10.1007/s00223-020-00734-y
14. Villamor E, Monserrat C, Del Río L,Romero-Martín JA, Rupérez MJ (2020) Prediction of osteoporotic hip fracture in postmenopausal women through patient-specific FE analyses and machine learning. Computer Methods and Programs in Biomedicine 193: 105484. https://doi.org/10.1016/j.cmpb.2020.105484

15. Van Geel T A C M, Nguyen N D, Geusens P P et al (2011) Development of a simple prognostic nomogram for individualising 5-year and 10-year absolute risks of fracture: a population-based prospective study among postmenopausal women. Annals of the rheumatic diseases 70(1): 92-97. https://doi.org/ 10.1136/ard.2010.131813

16. Ulivieri F M, Rinaudo L, Piodi L P et al (2021) Bone strain index as a predictor of further vertebral fracture in osteoporotic women: An artificial intelligence-based analysis. Plos one 16(2): e0245967.https://doi.org/ 10.1371/journal.pone.0245967

17. Yoda T, Maki S, Furuya T et al (2022) Automated Differentiation Between Osteoporotic Vertebral Fracture and Malignant Vertebral Fracture on MRI Using a Deep Convolutional Neural Network. Spine 47(8): E347-E352.https://doi.org/ 10.1097/BRS.0000000000004307

18. Jiang X, Westermann LB, Galleo GV, Demko J, Marakovits KA, Schnatz PF (2013) Age as a predictor of osteoporotic fracture compared with current risk-prediction models. Obstetrics & Gynecology 122(5): 1040-1046.https://doi.org/ 10.1097/AOG.0b013e3182a7e29b

19. Schousboe JT, Rosen HR, Vokes TJ et al (2014)Prediction models of prevalent radiographic vertebral fractures among older women. Journal of Clinical Densitometry 17(3): 378-385.https://doi.org/ 10.1016/j.jocd.2013.09.021. Epub 2014 Feb 25

20. Sandhu S K, Nguyen N D, Center J R et al (2010) Prognosis of fracture: evaluation of predictive accuracy of the FRAX™ algorithm and Garvan nomogram. Osteoporosis international 21(5): 863-871.https://doi.org/ 10.1007/s00198-009-1026-7. Epub 2009 Jul 25

21. Rubin K H, Möller S, Holmberg T , Bliddal M, Søndergaard J, Abrahamsen B (2018)A new fracture risk assessment tool (FREM) based on public health registries. Journal of bone and mineral research 33(11): 1967-1979.https://doi.org/ 10.1002/jbmr.3528

22. Pluskiewicz W, Adamczyk P, Franek E et al (2010) Ten-year probability of osteoporotic fracture in 2012 Polish women assessed by FRAX and nomogram by Nguyen et al.—Conformity between methods and their clinical utility. Bone 46(6): 1661-1667.https://doi.org/ 10.1016/j.bone.2010.02.012

23. Jang E J, Lee Y K, Choi H J et al (2016) Osteoporotic fracture risk assessment using bone mineral density in Korean: a community-based cohort study[J]. Journal of bone metabolism 23(1): 34-39.https://doi.org/ 10.11005/jbm.2016.23.1.34. Epub 2016 Feb 29

24. Monchka B A, Kimelman D, Lix L M et al (2021) Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: The Manitoba Bone Mineral Density Registry. Bone150: 116017.https://doi.org/ 10.1016/j.bone.2021.116017

25. Mehta S D, Sebro R (2020) Computer-aided detection of incidental lumbar spine fractures from routine dual-energy X-ray absorptiometry (DEXA) studies using a support vector machine (SVM) classifier. Journal of Digital Imaging 33(1): 204-210.https://doi.org/ 10.1007/s10278-019-00224-0

26. Langsetmo L, Nguyen T V, Nguyen N D et al (2011) Independent external validation of nomograms for predicting risk of low-trauma fracture and hip fracture. CMAJ 183(2): E107-E114.https://doi.org/ 10.1503/cmaj.100458. Epub 2010 Dec 20

27. Ioannidis G, Jantzi M, Bucek J et al (2017) Development and validation of the Fracture Risk Scale (FRS) that predicts fracture over a 1-year time period in institutionalised frail older people living in Canada: an electronic record-linked longitudinal cohort study. BMJ open 7(9): e016477.https://doi.org/ 10.1136/bmjopen-2017-016477

28. Nishiyama K K, Macdonald H M, Hanley D A et al (2013)Women with previous fragility fractures can be classified based on bone microarchitecture and finite element analysis measured with HR-pQCT[J]. Osteoporosis International, 2013, 24(5): 1733-1740.https://doi.org/ 10.1007/s00198-012-2160-1
29. Kruse C, Eiken P, Vestergaard P (2017) Machine learning principles can improve hip fracture prediction[J]. Calcified tissue international 100(4): 348-360.https://doi.org/ 10.1007/s00223-017-0238-7
30. Kolanu N, Brown A S, Beech A et al (2021) Natural language processing of radiology reports for the identification of patients with fracture. Archives of Osteoporosis 16(1): 1-8.https://doi.org/ 10.1007/s11657-020-00859-5
31. Kim H Y, Jang E J, Park B J, Kim TY, Shin SA, Ha YC, Jang Sl (2016) Development of a Korean Fracture Risk Score (KFRS) for predicting osteoporotic fracture risk: analysis of data from the Korean National Health Insurance Service. PLoS One 11(7): e0158918.https://doi.org/ 10.1371/journal.pone.0158918
32. Hsieh C I, Zheng K, Lin C, Mei L, Lu L, Li W, Kuo CF (2021) Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. Nature communications 12(1):1-9. https://doi.org/10.1038/s41467-021-25779-x
33. Hong N, Park H, Kim CO et al (2021) Bone Radiomics score derived from DXA Hip images enhances hip fracture prediction in older women. Journal of Bone and Mineral Research 36(9): 1708-1716.https://doi.org/ 10.1002/jbmr.4342
34. Ho-Le TP, Center JR, Eisman JA, Nguyen TV, Nguyen HT (2017) Prediction of hip fracture in post-menopausal women using artificial neural network approach. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4207-4210). https://doi.org/ 10.1109/EMBC.2017.8037784
35. Henry M J, Pasco J A, Merriman E N et al (2011) Fracture risk score and absolute risk of fracture. Radiology 259(2): 495-501.https://doi.org/ 10.1148/radiol.10101406
36. Galassi A, Martín-Guerrero JD, Villamor E , Monserrat C, Rupérez MJ (2020) Risk assessment of hip fracture based on machine learning. Applied Bionics and Biomechanics 2020.https://doi.org/ 10.1155/2020/8880786
37. FitzGerald G, Compston JE, Chapurlat RD, Pfeilschifter J, Cooper C, Hosmer D.W, Gehlbach SH (2014) Empirically based composite fracture prediction model from the Global Longitudinal Study of Osteoporosis in Postmenopausal Women (GLOW). The Journal of Clinical Endocrinology & Metabolism 99(3):817-826..https://doi.org/ 10.1210/jc.2013-3468
38. Ferizi U, Besser H, Hysi P, Jacobs J, Rajapakse CS, Chen C, Chang G (2019) Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from MRI data. Journal of Magnetic Resonance Imaging 49(4): 1029-1038..https://doi.org/ 10.1002/jmri.26280
39. Enns-Bray WS, Bahaloo H, Fleps I, Pauchard Y, Taghizadeh E, Sigurdsson S,Helgason B (2019) Biofidelic finite element models for accurately classifying hip fracture in a retrospective clinical study of elderly women from the AGES Reykjavik cohort. Bone 120:25-37.https://doi.org/ 10.1016/j.bone.2018.09.014
40. Engels A, Reber K C, Lindlbauer I, Rapp K, Büchele G, Klenk J, König HH (2020) Osteoporotic hip fracture prediction from risk factors available in administrative claims data-A machine learning approach. PloS one 15(5):e0232969.https://doi.org/ 10.1371/journal.pone.0232969
41. De Vries B C S, Hegeman J H, Nijmeijer W, Nijmeijer W, Geerdink J, Seifert C, Groothuis-Oudshoorn CG M (2021) Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis. Osteoporosis international 32(3):437-449..https://doi.org/ 10.1007/s00198-020-05735-z

42. Cheung E Y N, Bow C H, Cheung C L, Soong C, Yeung S, Loong C, Kung A (2012) Discriminative value of FRAX for fracture prediction in a cohort of Chinese postmenopausal women. Osteoporosis international:23(3), 871-878.https://doi.org/ 10.1007/s00198-011-1647-5

43. Chanplakorn P, Lertudomphonwanit T, Daraphongsataporn N, Sritara C, Jaovisidha S, Sa-Ngasoongsong P (2021) Development of prediction model for osteoporotic vertebral compression fracture screening without using clinical risk factors, compared with FRAX and other previous models. Archives of Osteoporosis 16(1): 1-9.https://doi.org/ 10.1007/s11657-021-00957-y

44. Bredbenner T L, Mason R L, Havill L M, Orwoll ES,Nicolella DP, Osteoporotic Fractures in Men (MrOS) Study (2014) Fracture risk predictions based on statistical shape and density modeling of the proximal femur. Journal of bone and mineral research 29(9):2090-2100.https://doi.org/ 10.1002/jbmr.2241

45. Beyaz S, Açıcı K, Sümer E (2020) Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Joint diseases and related surgery 31(2): 175.https://doi.org/ 10.5606/ehc.2020.72163

46. Berry S D, Zullo A R, Lee Y, Mor V, McConeghy KW, Banerjee G, Kiel DP (2018) Fracture risk assessment in long-term care (FRAiL): development and validation of a prediction model. The Journals of Gerontology: Series A 73(6):763-769..https://doi.org/ 10.1093/gerona/glx147

47. Beaudoin C, Jean S, Moore L et al (2021) Prediction of Osteoporotic Fractures in Elderly Individuals: A Derivation and Internal Validation Study Using Healthcare Administrative Data. Journal of Bone and Mineral Research 36(12): 2329-2342.https://doi.org/ 10.1002/jbmr.4438

48. Baleanu F, Moreau M, Charles A et al (2022) Fragility fractures in postmenopausal women: development of 5-year prediction models using the FRISBEE study. The Journal of Clinical Endocrinology & Metabolism 107(6): e2438-e2448.https://doi.org/ 10.1210/clinem/dgac092

49. Almog Y A, Rai A, Zhang P et al (2020) Deep learning with electronic health records for short-term fracture risk identification: crystal bone algorithm development and validation. Journal of medical Internet research 22(10): e22550.https://doi.org/ 10.2196/22550

50. Zagórski P, Tabor E, Martela-Tomaszek K, Adamczyk P, Pluskiewicz W (2021) Five-year fracture risk assessment in postmenopausal women, using both the POL-RISK calculator and the Garvan nomogram: the Silesia Osteo Active Study. Archives of Osteoporosis 16(1): 1-8.https://doi.org/10.1007/s11657-021-00881-1

51. Diez-Perez A, Gonzalez-Macias J, Marin F et al (2007) Prediction of absolute risk of non-spinal fractures using clinical risk factors and heel quantitative ultrasound. Osteoporosis International 18(5): 629-639.https://doi.org/ 10.1007/s00198-006-0297-5

52. Lix L M, Leslie W D, Majumdar S R (2018) Measuring improvement in fracture risk prediction for a new risk factor: a simulation. BMC Research Notes, 11(1):1-5.https://doi.org/10.1186/s13104-018-3178-z

53. Li Q, Long X, Wang Y et al (2021) Development and validation of a nomogram for predicting the probability of adjacent segmental fractures after vertebral augmentation of osteoporotic vertebral compression fractures.https://doi.org/ 10.1186/s12891-021-04845-x

54. Lee S, Lee JW, Jeong J W, Yoo DS, Kim S (2008) A preliminary study on discrimination of osteoporotic fractured group from nonfractured group using support vector machine. In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 474-477).https://doi.org/ 10.1109/IEMBS.2008.4649193

55. Jacobs J W G, Da Silva J A P, Armbrecht G et al (2010) Prediction of vertebral fractures is specific for gender and site of bone mineral density measurement. The Journal of Rheumatology 37(1): 149-154.https://doi.org/ 10.3899/jrheum.090731

56. Eller-Vainicher C, Chiodini I, Santi I et al (2011) Recognition of morphometric vertebral fractures by artificial neural networks: analysis from GISMO Lombardia Database. PLoS One 6(11): e27277.https://doi.org/ 10.1371/journal.pone.0027277

57. Zhong B Y, He S C, Zhu H D et al (2017) Risk prediction of new adjacent vertebral fractures after PVP for patients with vertebral compression fractures: development of a prediction model. Cardiovascular and interventional radiology 40(2): 277-284.https://doi.org/ 10.1007/s00270-016-1492-1

58. Xiao X, Wu Q(2021) The Utility of Genetic Risk Score to Improve Performance of FRAX for Fracture Prediction in US Postmenopausal Women. Calcif Tissue Int108(6):746-756.https:// doi:10.1007/s00223-021-00809-4

59. Pisani P, Renna MD, Conversano F, Casciaro E, Di Paola M, Quarta E, Muratore M, Casciaro S (2016) Major osteoporotic fragility fractures: Risk factor updates and societal impact. World J Orthop 18;7(3):171-81. https://doi: 10.5312/wjo.v7.i3.171

60. Lim HK, Ha HI, Park SY, Han J (2021) Prediction of femoral osteoporosis using machine-learning analysis with radiomics features and abdomen-pelvic CT: A retrospective single center preliminary study. PloS one, 16(3), e0247330.https://doi.org/10.1371/journal.pone.0247330

61. Shimizu S (2019) Non-Gaussian methods for causal structure learning. Prevention Science 20(3): 431-441.https://doi.org/10.1007/s11121-018-0901-x

62. Supawattanabodee B , Wiriyasirivaj B (2018). Utility of Bayesian Logistic Regression Model in the Development of a Clinical Risk Score Index for Screening of Osteoporosis in Menopausal Women. Journal of the medical association of Thailand 101(8):193.

63. Ganesan N, Venkatesh K, Rama M A et al (2010) Application of neural networks in diagnosing cancer disease using demographic data. International Journal of Computer Applications 1(26): 76-85.https://doi.org/ 10.1002/med.21658

64. Popescu D, El-Khatib M, El-Khatib H, Ichim L (2022) New Trends in Melanoma Detection Using Neural Networks: A Systematic Review. Sensors (Basel) 22(2):496.https://doi.org/10.3390/s22020496.

65. Bellou V, Belbasis L, Konstantinidis A K et al (2019) Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisa.BMJ 367.https://doi.org/ 10.1136/bmj.l5358

66. De Silva K, Lee W K, Forbes A, Demme RT, Barton C, Enticott J (2020) Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. International journal of medical informatics 143: 104268.https://doi.org/ 10.1016/j.ijmedinf.2020.104268

67. Jain R, Sontisirikit S, Iamsirithaworn S et al (2019) Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC infectious diseases 19(1): 1-16.https://doi.org/ 10.1186/s12879-019-3874-x

68. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, Swart EL, Girbes ARJ, Thoral P, Ercole A, Hoogendoorn M, Elbers PWG (2020) Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46(3):383-400.https://doi.org/10.1007/s00134-019-05872-y

69. Hossain M, Akbar SA, Andrew G (2010) Misdiagnosis of occult hip fracture is more likely in patients with poor

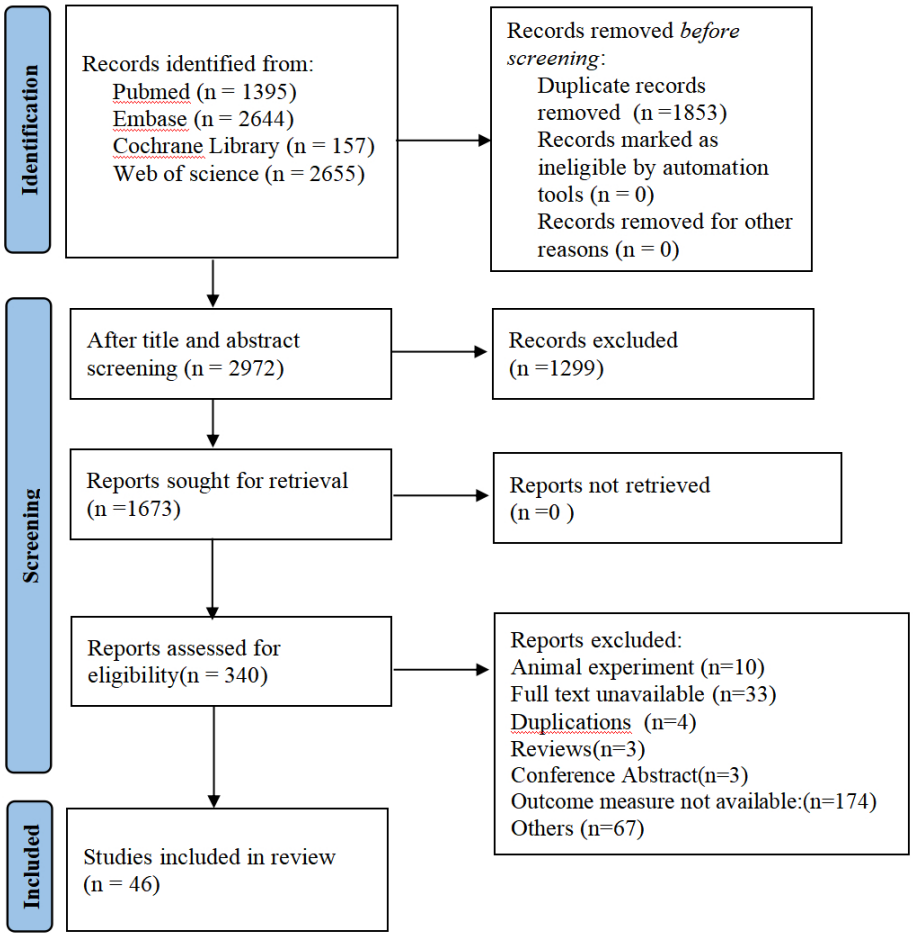
mobility and cognitive impairment. Acta Orthop Belg 76(3):341-6.

Figure legends

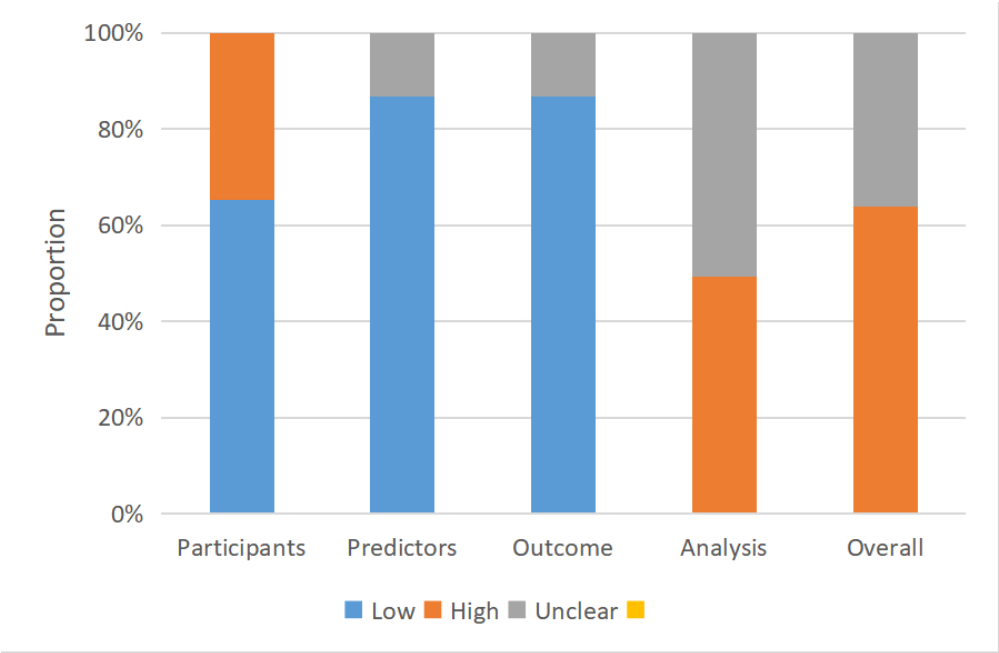
Figure 1 The flow chart of retrieval process

Figure 2 Risk of bias assessment (using PROBAST) based on four domains across 88 machine learning models

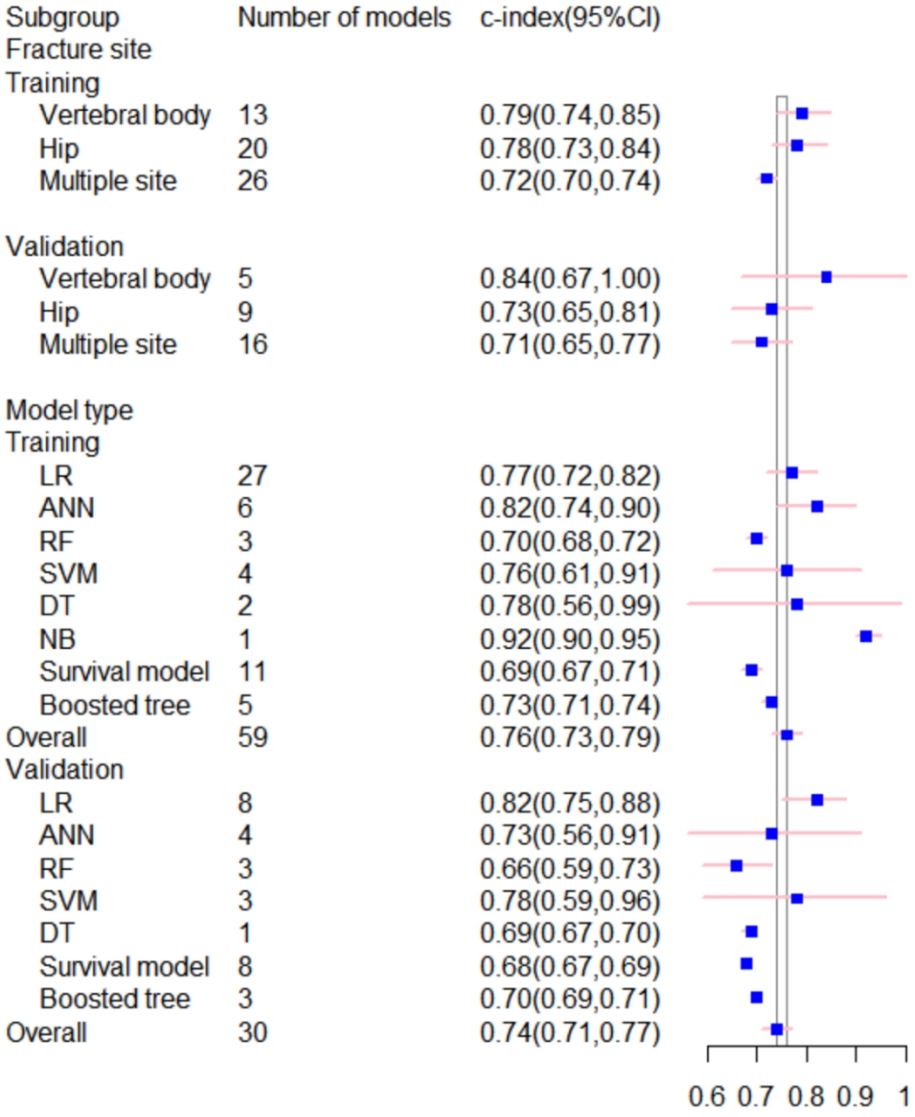
Figure 3 Forest plots for C-index statistics for subgroup analysis by fracture site and machine learning type



392x382mm (72 x 72 DPI)



316x207mm (72 x 72 DPI)



401x502mm (72 x 72 DPI)

Table S1 : PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	2-4
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	4
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	5
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	4-5
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Table S2
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	5-6，Figure 1
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	6
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	6
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Table S3
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	6-7，Figure 2
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	7
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	7
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	7
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	7
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	7-8
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Table3,Table4
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	7
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	6-7
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	7
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	8
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1

Section and Topic	Item #	Checklist item	Location where item is reported
Study characteristics	17	Cite each included study and present its characteristics.	7-8,Table 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	9,Figure 2,Table S4
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	9-10，Table S2
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	10-11
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	10，Figure 3,Table 3
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	10-11,Table 3
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	10-11，Figure 4,Table 4
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	9,Figure 2
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Table 3,Table 4
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	11-15
	23b	Discuss any limitations of the evidence included in the review.	14-15
	23c	Discuss any limitations of the review processes used.	14
	23d	Discuss implications of the results for practice, policy, and future research.	15
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	4,Registration No. CRD42022346896
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	The protocol was registered on PROSPERO
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	15
Competing interests	26	Declare any competing interests of review authors.	16
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Table1,Table S3

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

Table S2 Literature search strategy

1.Pubmed

Search number	Query	Results
#1	"Osteoporosis"[Mesh]	59,962
#2	"Osteoporosis"[Title/Abstract] OR "Osteoporoses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR "age related bone loss"[Title/Abstract] OR "age related bone losses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR ((("bone and bones"[MeSH Terms] OR ("Bone"[All Fields] AND "bones"[All Fields]) OR "bone and bones"[All Fields] OR "Bone"[All Fields]) AND "Losses"[All Fields]) AND "Age-Related"[Title/Abstract])	77,866
#3	"Osteoporosis"[MeSH Terms] OR ("Osteoporosis"[Title/Abstract] OR "Osteoporoses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR "age related bone loss"[Title/Abstract] OR "age related bone losses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR ((("bone and bones"[MeSH Terms] OR ("Bone"[All Fields] AND "bones"[All Fields]) OR "bone and bones"[All Fields] OR "Bone"[All Fields]) AND "Losses"[All Fields]) AND "Age-Related"[Title/Abstract]))	95,574
#4	"Machine Learning"[Mesh]	41,875
#5	"machine learning"[Title/Abstract] OR "transfer learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "prediction model"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "random forest"[Title/Abstract] OR "artificial neural network"[Title/Abstract] OR "ANN"[Title/Abstract] OR "support vector machine"[Title/Abstract] OR "SVM"[Title/Abstract] OR "gradient boosting machine"[Title/Abstract] OR "GBM"[Title/Abstract] OR "Nomogram"[Title/Abstract] OR "XGboost"[Title/Abstract] OR "Logistic"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "external validation"[Title/Abstract]	554,732
#6	"Machine Learning"[MeSH Terms] OR "Machine Learning"[Title/Abstract] OR "transfer learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "prediction model"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "random forest"[Title/Abstract] OR "artificial neural network"[Title/Abstract] OR "ANN"[Title/Abstract] OR "support vector machine"[Title/Abstract] OR "SVM"[Title/Abstract] OR "gradient boosting machine"[Title/Abstract] OR "GBM"[Title/Abstract] OR "Nomogram"[Title/Abstract] OR "XGboost"[Title/Abstract] OR "Logistic"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "external validation"[Title/Abstract]	560,576
#7	"fractures, bone"[MeSH]	198,823
#8	"fractures bone"[Title/Abstract] OR "broken bones"[Title/Abstract] OR "bone broken"[Title/Abstract] OR "bones broken"[Title/Abstract] OR "broken bone"[Title/Abstract] OR "Fractures"[Title/Abstract] OR "Fracture"[Title/Abstract]	282,444
#9	"fractures, bone"[MeSH Terms] OR "fractures bone"[Title/Abstract] OR "broken bones"[Title/Abstract] OR "bone broken"[Title/Abstract] OR "bones broken"[Title/Abstract] OR "broken bone"[Title/Abstract] OR "Fractures"[Title/Abstract] OR "Fracture"[Title/Abstract]	325,361
#10	#3AND #6 AND #9	1,395

2.Cochrane

Search number	Query	Results
#1	MeSH descriptor: [Osteoporosis] explode all trees	4,308
#2	(Osteoporosis):ti,ab,kw OR (Osteoporoses):ti,ab,kw OR (Bone Loss, Age-Related):ti,ab,kw OR (Age-Related Bone Loss):ti,ab,kw OR (Age-Related Bone Losses):ti,ab,kw	11,404
#3	(Bone Loss, Age Related):ti,ab,kw OR (Bone Losses, Age-Related):ti,ab,kw	510
#4	#1 OR #2 OR #3	11,690
#5	MeSH descriptor: [Machine Learning] explode all trees	200
#6	(machine learning):ti,ab,kw OR (Transfer Learning):ti,ab,kw OR (Deep learning):ti,ab,kw OR	8,875

	(Prediction model):ti,ab,kw OR (artificial intelligence):ti,ab,kw	
#7	(random forest):ti,ab,kw OR (artificial neural network):ti,ab,kw OR (ANN):ti,ab,kw OR (Support vector machine):ti,ab,kw OR (SVM):ti,ab,kw	3,025
#8	(Gradient Boosting Machine):ti,ab,kw OR (GBM):ti,ab,kw OR (Nomogram):ti,ab,kw OR (XGboost):ti,ab,kw OR (Logistic):ti,ab,kw	29,213
#9	(Decision tree):ti,ab,kw OR (External validation):ti,ab,kw	1,737
#10	#5 OR #6 OR #7 OR #8 OR #9	39,732
#11	MeSH descriptor: [Fractures, Bone] explode all trees	6,688
#12	(Fractures, Bone):ti,ab,kw OR (Broken Bones):ti,ab,kw OR (Bone, Broken):ti,ab,kw OR (Bones, Broken):ti,ab,kw OR (Broken Bone):ti,ab,kw	8,663
#13	(Fractures):ti,ab,kw OR (Fracture):ti,ab,kw	25,188
#14	#11 OR #12 OR #13	25,283
#15	#4 AND #10 AND #14	157

3.Embase

Search number	Query	Results
#1	'osteoporosis'/exp	144,364
#2	'osteoporosis':ab,ti OR 'osteoporoses':ab,ti OR 'bone loss, age-related':ab,ti OR 'age-related bone loss':ab,ti OR 'age-related bone losses':ab,ti OR 'bone loss, age related':ab,ti OR 'bone losses, age-related':ab,ti	115,173
#3	#1 OR #2	167,079
#4	'machine learning'/exp	300,972
#5	'machine learning':ab,ti OR 'transfer learning':ab,ti OR 'deep learning':ab,ti OR 'prediction model':ab,ti OR 'artificial intelligence':ab,ti OR 'random forest':ab,ti OR 'artificial neural network':ab,ti OR ann:ab,ti OR 'support vector machine':ab,ti OR svm:ab,ti OR 'gradient boosting machine':ab,ti OR gbm:ab,ti OR nomogram:ab,ti OR xgboost:ab,ti OR logistic:ab,ti OR 'decision tree':ab,ti OR 'external validation':ab,ti	814,126
#6	#4 OR #5	1,002,289
#7	'fracture'/exp	362,337
#8	'fractures, bone':ab,ti OR 'broken bones':ab,ti OR 'bone, broken':ab,ti OR 'bones, broken':ab,ti OR 'broken bone':ab,ti OR 'fractures':ab,ti OR 'fracture':ab,ti	343,893
#9	#7 OR #8	445,516
#10	#3 AND #6 AND #9	2,644

4.Web of science

Search number	Query	Results
#1	Osteoporosis (Topic) or Osteoporoses (Topic) or Bone Loss, Age-Related (Topic) or Age-Related Bone Loss (Topic) or Age-Related Bone Losses (Topic) or Bone Loss, Age Related (Topic) or	104,281

	Bone Losses, Age-Related (Topic)	
#2	machine learning (Topic) or Transfer Learning (Topic) or Deep learning (Topic) or Prediction model (Topic) or artificial intelligence (Topic) or random forest (Topic) or artificial neural network (Topic) or ANN (Topic) or Support vector machine (Topic) or SVM (Topic) or Gradient Boosting Machine (Topic) or GBM (Topic) or Nomogram (Topic) or XGboost (Topic) or Logistic (Topic) or Decision tree (Topic) or External validation (Topic)	1,950,484
#3	Fractures, Bone (Topic) or Broken Bones (Topic) or Bone, Broken (Topic) or Bones, Broken (Topic) or Broken Bone (Topic) or Fractures (Topic) or Fracture (Topic)	575,668
#4	#1 AND #2 AND #3	2,655

Table S3 Methodological characteristics of machine learning models developed for outcome prediction in patients with Osteoporosis

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	C-index	Sensitivity
Wu, Q	2020	Train	M	Multiple fractures	361	4104	LR		
Wu, Q	2020	Train	M	Multiple fractures	361	4104	RF		
Wu, Q	2020	Train	M	Multiple fractures	361	4104	BT		
Wu, Q	2020	Train	M	Multiple fractures	361	4104	ANN		
Wu, Q	2020	Test	M	Multiple fractures	90	1026	LR	0.6410	0.7610
Wu, Q	2020	Test	M	Multiple fractures	90	1026	RF	0.7005	0.7000
Wu, Q	2020	Test	M	Multiple fractures	90	1026	BT	0.7100	0.5650
Wu, Q	2020	Test	M	Multiple fractures	90	1026	ANN	0.6910	0.7120
Villamor, E	2020	Train	F	Hip fracture	65	101	LR		
Villamor, E	2020	Train	F	Hip fracture	65	101	SVM		
Villamor, E	2020	Train	F	Hip fracture	65	101	ANN		
Villamor, E	2020	Train	F	Hip fracture	65	101	RF		
Villamor, E	2020	Test	F	Hip fracture	65	101	LR		
Villamor, E	2020	Test	F	Hip fracture	65	101	SVM		
Villamor, E	2020	Test	F	Hip fracture	65	101	ANN		
Villamor, E	2020	Test	F	Hip fracture	65	101	RF		
van Geel, Tacm	2011	Train	F	Vertebral fracture	382	2372	SM		
Ulivieri, F. M	2021	Train	F	Vertebral fracture	56	90	ANN	0.8300	0.7500
Yoda, T	2021	Train	M+F	Vertebral fracture	28	50	CNN	0.9670	0.9250

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	C-index	Sensitivity
Yoda, T	2021	Test	M+F	Vertebral fracture	21	47	CNN	0.9840	0.9810
Jiang, X. Z	2013	Train	F	Multiple fractures	15	615	LR	0.7600	0.8100
Schousboe, J. T	2014	Train	F	Vertebral fracture	2883	7233	LR	0.6790	
Sandhu, S. K	2010	Train	F	Multiple fractures	47	144	LR	0.8400	0.7800
Sandhu, S. K	2010	Train	M	Multiple fractures	18	56	LR	0.7600	0.7400
Rubin, K. H	2018	Train	F	Multiple fractures	11898	647103	LR	0.7500	0.7520
Rubin, K. H	2018	Train	M	Multiple fractures	11851	647103	LR	0.7520	0.6450
Rubin, K. H	2018	Test	F	Multiple fractures	4762	600567	LR	0.8740	0.6000
Rubin, K. H	2018	Test	M	Multiple fractures	4776	600566	LR	0.8510	0.6300
Pluskiewicz, W	2010	Train	F	Hip fracture	1599	2012	LR	0.850	0.7590
Pluskiewicz, W	2010	Train	F	Multiple fractures	1704	2012	LR	0.8790	0.7390
Jang, E. J	2016	Train	M	Multiple fractures	36	363	LR	0.7390	
Jang, E. J	2016	Train	F	Multiple fractures	50	405	LR	0.7180	
Barret A. Monchka	2021	Train	M+F	Vertebral fracture	1470	8920	CNN	0.9500	0.8240
Mehta, S. D	2020	Train	M+F	Vertebral fracture	86	246	SVM	0.9258	0.8950
Mehta, S. D	2020	Test	M+F	Vertebral fracture	22	61	SVM	0.8963	0.8180
Langsetmo, L	2011	Test	M	Multiple fractures	139	1606	SM	0.7000	
Langsetmo, L	2011	Test	F	Multiple fractures	672	4152	SM	0.6900	
Ioannidis, G	2017	Train	M+F	Multiple fractures	3858	22386	DT	0.6690	
Ioannidis, G	2017	Test	M+F	Multiple fractures	1294	7462	DT	0.6870	

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	C-index	Sensitivity
K. K. Nishiyama	2013	Train	F	Multiple fractures	44	88	SVM	0.6800	0.5280
K. K. Nishiyama	2013	Test	F	Multiple fractures	14	28	SVM	0.8000	0.6880
Kruse, C	2017	Train	F	Hip fracture	293	4722	NB	0.9200	0.8800
Kruse, C	2017	Train	M	Hip fracture	47	717	DT	0.8900	1.0000
Kolanu, N	2021	Train	M+F	Multiple fractures	433	5089	ANN		0.9900
Kolanu, N	2021	Test	M+F	Multiple fractures	97	327	ANN		0.6960
Kim, H. Y	2016	Train	M	Multiple fractures	4889	185127	SM	0.6800	
Kim, H. Y	2016	Train	F	Multiple fractures	14951	174126	SM	0.6500	
Kim, H. Y	2016	Test	M+F	Multiple fractures	19915	359255	SM	0.6650	
Hsieh, C. I	2021	Train	M+F	Hip fracture	2254	5164	LR	0.9700	0.8820
Hsieh, C. I	2021	Test	M+F	Hip fracture	922	2060	LR	0.9600	0.8990
Hsieh, C. I	2021	Train	M+F	Vertebral fracture	530	57662	LR	0.9700	0.6960
Hsieh, C. I	2021	Test	M+F	Vertebral fracture	922	3346	LR	0.9400	0.7400
Hong, N	2021	Train	F	Hip fracture	143	433	RF	0.7840	
Hong, N	2021	Train	F	Hip fracture	143	433	BT	0.7680	
Hong, N	2021	Train	F	Hip fracture	143	433	SVM	0.7590	
Hong, N	2021	Train	F	Hip fracture	143	433	BT	0.7580	
Hong, N	2021	Test	F	Hip fracture	34	2029	SM	0.8400	
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	ANN		0.8890
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	LR		0.9070

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	C-index	Sensitivity
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	KNN		1.0000
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	SVM		0.9240
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	ANN		0.8330
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	LR		0.7780
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	KNN		0.8060
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	SVM		0.8060
Henry, M. J	2011	Train	F	Multiple fractures	125	600	LR	0.7000	0.6420
Galassi, A	2020	Train	F	Hip fracture	62	96	LR		0.7033
Galassi, A	2020	Train	F	Hip fracture	62	96	SVM		0.9367
Galassi, A	2020	Train	F	Hip fracture	62	96	DT		0.5967
Galassi, A	2020	Train	F	Hip fracture	62	96	RF		0.8330
FitzGerald, G	2014	Train	F	Multiple fractures	2638	47429	SM	0.6670	
Ferizi, U	2019	Train	F	Multiple fractures	32	92	BT	0.6400	0.5880
Ferizi, U	2019	Train	F	Multiple fractures	32	92	LR	0.6500	0.5490
Ferizi, U	2019	Train	F	Multiple fractures	32	92	LR	0.6700	0.5210
Enns-Bray, W. S	2019	Train	F	Hip fracture	95	254	LR	0.7270	
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	LR	0.7140	
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	RF	0.6860	
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	SVM	0.6600	
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	BT	0.7110	

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	C-index	Sensitivity
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	LR	0.7220	1.0000
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	BT	0.7250	
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	LR	0.6950	1.0000
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	RF	0.6850	
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	SVM	0.6500	
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	BT	0.7020	
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	LR	0.6980	
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	BTt	0.7030	
de Vries, B. C. S	2021	Train	M+F	Multiple fractures	805	7578	SM	0.6970	
de Vries, B. C. S	2021	Train	M+F	Multiple fractures	805	7578	ANN	0.6700	
de Vries, B. C. S	2021	Train	M+F	Multiple fractures	805	7578	RF	0.6870	
de Vries, B. C. S	2021	Test	M+F	Multiple fractures	165	1770	SM	0.6250	
de Vries, B. C. S	2021	Test	M+F	Multiple fractures	165	1770	ANN	0.5880	
de Vries, B. C. S	2021	Test	M+F	Multiple fractures	165	1770	RF	0.5930	
Cheung, E. Y	2012	Train	F	Multiple fractures	106	2266	SM	0.7300	0.8080
Chanplakorn, P	2021	Train	F	Vertebral fracture	179	617	LR	0.6500	0.4300
Bredbenner, T. L	2014	Train	M	Hip fracture	45	472	LR	0.9300	
Beyaz, S	2020	Train	M+F	Multiple fractures	235	2106	CNN		0.8250
Berry, S. D	2018	Train	M	Hip fracture	3541	119874	SM	0.6922	
Berry, S. D	2018	Train	F	Hip fracture	11012	299794	SM	0.7106	

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	C-index	Sensitivity
Berry, S. D	2018	Test	M+F	Hip fracture	28050	858636	SM	0.6800	
Beaudoin, C	2021	Train	M+F	Multiple fractures	57678	307909	SM	0.6810	
Beaudoin, C	2021	Test	M+F	Multiple fractures	21809	273372	SM	0.6790	
Baleanu, F	2022	Train	F	Multiple fractures	410	3560	LR	0.7300	
Almog, Y. A	2020	Train	M+F	Vertebral fracture	2468694	6329986	ANN	0.8120	0.8120
Almog, Y. A	2020	Test	M+F	Vertebral fracture	295479	3476219	ANN	0.6680	0.7070
Zagorski, P	2021	Train	F	Hip fracture	49	389	LR	0.8840	0.9390
Diez-Perez, A	2007	Train	F	Multiple fractures	363	5201	SM	0.6720	
Lix, L. M	2018	Train	F	Multiple fractures	749	31999	LR	0.7060	
Li, Q. J	2021	Train	F	Multiple fractures	49	403	LR	0.8820	
Li, Q. J	2021	Test	F	Multiple fractures	17	159	LR	0.8690	
Lee, S	2008	Train	F	Multiple fractures	47	94	SVM		0.8500
Jacobs, J. W. G	2010	Train	M	Vertebral fracture	58	109	LR	0.5100	
Jacobs, J. W. G	2010	Train	F	Vertebral fracture	98	205	LR	0.7400	0.6700
Eller-Vainicher, C	2011	Train	F	Vertebral fracture	33	372	LR	0.8230	0.3730
Eller-Vainicher, C	2011	Train	F	Vertebral fracture	33	372	ANN	0.6990	0.7480
Zhong, B. Y	2017	Train	M+F	Vertebral fracture	33	256	SM	0.7800	
Zhong, B. Y	2017	Test	M+F	Vertebral fracture	23	165	SM	0.7200	
Xiao, X	2021	Train	F	Hip fracture	25	699	SM	0.8040	

*M:Male;F:Female;LR: Logistic Regression;ANN:artifificial ANN;SVM = support-vector machine; CNN = convolutional ANN; kNN = k-nearest neighbors; RF = random forests;DT=decision tree;BT=Boosted tree;SM=Survival model.

Table S4 Risk of bias assessment grading of the machine learning predictive modelling studies of osteoporosis populations as per the PROBAST criteria

Study	Participants bias	Predictors bias	Outcome bias	Analysis bias	Overall bias rating
Wu, Q	low	low	low	high	high
Wu, Q	low	low	low	high	high
Wu, Q	low	low	low	high	high
Wu, Q	low	low	low	high	high
Villamor, E	high	unclear	unclear	high	high
Villamor, E	high	unclear	unclear	high	high
Villamor, E	high	unclear	unclear	high	high
Villamor, E	high	low	unclear	high	high
Van Geel, Tacm	low	low	low	high	high
Van Geel, Tacm	low	low	low	high	high
Ulivieri, F. M	low	low	low	high	high
Yoda, T	low	low	low	high	high
Jiang, X. Z	low	low	low	high	high
Schousboe, J. T	high	low	low	unclear	high
Sandhu, S. K	high	unclear	unclear	high	high
Rubin, K. H	low	low	low	unclear	unclear
Pluskiewicz, W	high	low	low	unclear	high
Jang, E. J	low	low	low	high	high
Barret A. Monchka	high	low	low	unclear	high
Mehta, S. D	high	unclear	unclear	high	high
Langsetmo, L	low	low	low	unclear	unclear
Ioannidis, G	high	low	low	unclear	high
K. K. Nishiyama	low	low	low	high	high
Kruse, C	low	low	low	unclear	unclear
Kruse, C	low	low	low	unclear	unclear
Kolanu, N	high	low	low	unclear	high
Kim, H. Y	low	low	low	unclear	unclear
Hsieh, C. I	low	low	low	unclear	unclear

Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Ho-Le, T. P	low	low	low	high	high
Ho-Le, T. P	low	low	low	high	high
Ho-Le, T. P	low	low	low	high	high
Ho-Le, T. P	low	low	low	high	high
Henry, M. J	low	low	low	unclear	unclear
Galassi, A	low	low	low	high	high
Galassi, A	low	low	low	high	high
Galassi, A	low	low	low	high	high
Galassi, A	low	low	low	high	high
FitzGerald, G	low	low	low	unclear	unclear
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high
Enns-Bray, W. S	high	low	low	high	high
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
de Vries, B. C. S	high	low	low	unclear	high
de Vries, B. C. S	high	low	low	unclear	high
de Vries, B. C. S	high	low	low	unclear	high
Cheung, E. Y	low	low	low	unclear	unclear
Chanplakorn, P	high	low	low	unclear	high
Bredbenner, T. L	high	unclear	unclear	high	high

Beyaz, S	high	low	low	unclear	high
Berry, S. D	low	low	low	unclear	unclear
Beaudoin, C	high	low	low	unclear	high
Baleanu, F	low	low	low	unclear	unclear
Baleanu, F	low	low	low	unclear	unclear
Almog, Y. A	high	low	low	unclear	unclear
Zagorski, P	low	low	low	high	high
Diez-Perez, A	low	low	low	unclear	unclear
Lix, L. M	low	low	low	unclear	unclear
Li, Q. J	high	low	low	high	high
Lee, S	high	unclear	low	high	high
Jacobs, J. W. G	low	low	low	unclear	unclear
Eller-Vainicher, C	low	low	low	high	high
Eller-Vainicher, C	low	low	low	high	high
Zhong, B. Y	high	low	low	high	high
Xiao, X	low	low	low	high	high

*When a single study included multiple models, risk of bias concerns were assessed for each model.

BMJ Open

The predictive value of machine learning on fracture risk in osteoporosis:a systematic review and meta-analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-071430.R1
Article Type:	Original research
Date Submitted by the Author:	22-Aug-2023
Complete List of Authors:	Wu, Yanqian; Southeast University Chao, Jianqian; Southeast University Bao, Min; Southeast University Zhang, Na; Southeast University
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Epidemiology, Public health
Keywords:	Public health < INFECTIOUS DISEASES, Epidemiology < ONCOLOGY, Bone diseases < ORTHOPAEDIC & TRAUMA SURGERY, Systematic Review

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

The predictive value of machine learning on fracture risk in osteoporosis: a systematic review and meta-analysis

Yanqian Wu[1], Jianqian Chao[1*], Min Bao[1], Na Zhang[1]

[1] Key Laboratory of Environmental Medicine Engineering of Ministry of Education, School of Public Health, Southeast University, Nanjing210009, P.R. China

wyqmm0523@163.com(Yanqian Wu, PhD), chaoseu@163.com (Jianqian Chao, PhD), 1689586985@qq.com(Min Bao, PhD),540923028@qq.com(Na Zhang, PhD)

[*]**Corresponding author:** Jianqian Chao, Key Laboratory of Environmental Medicine Engineering of Ministry of Education/Health Management Research Center, School of Public Health, Southeast University, 87 Dingjiaqiao Road, Gulou District, Nanjing 210009, China.

Tel: +86 025 86424437/13813955976.

E-mail: chaoseu@163.com

Word count 2957.

Abstract

Objectives: Early identification of fracture risk in patients with osteoporosis is essential. Machine learning (ML) has emerged as a promising technique to predict the risk, whereas its predictive performance remains controversial. Therefore, we conducted this systematic review and meta-analysis to explore the predictive efficiency of ML for the risk of fracture in patients with osteoporosis.

Methods: Relevant studies were retrieved from four databases (PubMed, Embase, Cochrane Library, and Web of Science) until May 31, 2023. A meta-analysis of the C-index was performed using a random-effects model, while a bivariate mixed-effects model was used for the meta-analysis of sensitivity and specificity. In addition, subgroup analysis was performed according to the types of ML models and fracture sites.

Results: Fifty-three studies were included in our meta-analysis, involving 15,209,268 patients, 86 prediction models specifically developed for the osteoporosis population, and 41 validation sets. The most commonly used predictors in these models encompassed age, BMI, past fracture history, bone mineral density T-score, history of falls, BMD, radiomics data, weight, height, gender, and other chronic diseases. Overall, the pooled C-index of ML was 0.75 (95% CI: 0.72,0.78) and 0.75 (95% CI: 0.71,0.78) in the training set and validation set, respectively; the pooled sensitivity was 0.79 (95% CI: 0.72,0.84) and 0.76 (95% CI: 0.80,0.81) in the training set and validation set, respectively; and the pooled specificity was 0.81 (95% CI: 0.75,0.86) and 0.83 (95% CI: 0.72,0.90) in the training set and validation set, respectively.

Conclusions: ML has a favorable predictive performance for fracture risk in patients with osteoporosis. However, most current studies lack external validation. Thus, external validation is required to verify the reliability of ML models.

Keywords: Osteoporosis; Machine learning; Fractures; Meta-Analysis

Strengths and limitations of this study

- The latest systematic review and meta-analysis conducted to assess ML models for fracture risk.
- We performed a quantitative synthesis to enhance the comparability of ML models.
- C-index, sensitivity and specificity was performed to evaluate the performance of ML models.
- Several studies were included in the systematic review but excluded from subsequent meta-analyses.
- Most of the included studies lack external validation.

Introduction

Osteoporosis is a systemic metabolic bone disease characterized by decreased bone mass and degraded bone microarchitecture, leading to an increased risk of bone fragility fracture (WHO,1994) [1]. Due to high disability and morbidity rates, high treatment costs, and low quality of life of patients, it has emerged as a global health concern [2]. According to the World Health Organization, osteoporosis is the second-most serious health issue after cardiovascular diseases [3]. This condition may cause fragility fractures that commonly occur in the wrist, spine, and hip. Spine and hip fractures may lead to disability, which not only affects the quality of life and longevity of patients, but also causes enormous medical expenses and a heavy burden of care [4, 5].

Machine learning (ML), a subfield of artificial intelligence, enables computers to "learn" through programs. Compared with traditional statistical methods, ML emphasizes more on the accuracy of prediction and can

detect regularities in multi-dimensional data sets. ML algorithms can be basically divided into supervised learning and unsupervised learning [6]. ML has been applied in the field of osteoporosis, providing a novel method for the prediction of fracture risk. A review by Ferizi et al. (2019) summarized relevant studies on the application of artificial intelligence to the prediction of osteoporosis. It drew a conclusion that ML methods for automatic image segmentation and fracture risk prediction showed a promising clinical value [7]. A systematic review by Smets et al. (2021) reviewed the state-of-the-art ML methods and their application in osteoporosis diagnosis and fracture prediction [8]. Another review by Anam et al. (2021) explored the prediction performance of magnetic resonance imaging for osteoporosis in trabecular bone from a methodology-driven and application perspective [9]. Most studies focused on the role of ML in the prediction of osteoporosis indicators, such as bone mineral density (BMD), or in the automatic segmentation of the images of patients at risk of osteoporosis. However, the efficiency of ML in predicting osteoporotic fractures is understudied.

The present study evaluated the predictive performance of ML for fracture risk in osteoporosis patients, providing an evidence-based medical basis for the application of ML in clinical practice.

Materials and methods

This study was conducted in accordance with the Preferred Items of Systematic Review and Meta-Analysis (PRISMA) statement (**Supplemental Table S1**) [10]. The protocol was registered on the international prospective register of systematic reviews (PROSPERO) (Registration No. CRD42022346896). Relevant studies were retrieved from Pubmed, Embase, Cochrane Library, and Web of Science, and the retrieval was as of May 31, 2023. Two researchers independently searched the literature. The search strategy is shown in **Supplemental Table S2**.

Inclusion criteria were as follows: (1) Patients were diagnosed with osteoporosis; (2) ML was applied to predict fracture risk; (3) At least one measure of model performance (discrimination or calibration) was reported; (4) Study population included adult patients older than 18 years, mainly including adults, older people, and postmenopausal women. Exclusion criteria were as follows: (1) Studies that only analyzed risk factors without building complete ML models; (2) Studies that only included osteoporosis but did not mention fracture risk; (3) Studies without available full text (or only abstract available) or data; (4) Meta-analyses, reviews, case reports, editorial materials, letters, protocols, errata, and notes.

Two researchers independently extracted data using standardized tables. Any studies excluded after full-text review have been recorded with reasons for their exclusion. The list of extracted items was based on the CHARMS checklist [11], and two data extraction sheets were prepared for developed and validated models, respectively. Finally, the extracted data included the first author, year of publication, country, study design, data source, population group, gender, age, fracture sites, types of predictive models, number of predictors, and outcomes. The risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST). The PROBAST contained a large number of questions in four distinct domains: participants, predictors, outcomes, and statistical analysis, reflecting the overall risk of bias and applicability [12].

Meta-analysis of C-index, sensitivity and specificity was performed to evaluate the performance of ML models. If the C-index did not report 95% confidence intervals (CI) and standard errors, we estimated the standard errors in reference to the study by Debray TP et al. [13]. A C-index of 0.5 indicates low discrimination; 0.6 to 0.7 indicates modest discrimination; 0.71 to 0.8 indicates very good discrimination; and greater than 0.8

indicates strong discrimination [14]. When original studies did not report the accuracy, we calculated it based on the sensitivity, specificity, the number of samples in each subgroup, and the number of modeling samples [13]. Given the differences in variables, ML algorithms, and parameters across the studies, the random-effects model was preferred for the meta-analysis of C-index, and the bivariate mixed-effects model was used for the meta-analysis of sensitivity and specificity. Heterogeneity was quantified using I^2 statistics. Sensitivity analysis was performed to further identify the source of heterogeneity by removing each study and re-calculating the pooled effect size of the remaining studies. The meta-analysis was performed using the software Stata 15.1 (Stata Corporation, College Station, TX, USA) and R4.2.0 (R Development Core Team, Vienna, http://www.R-project.org). A p value less than 0.05 was considered statistically significant.

Patient and Public Involvement

No patient involved.

Results

A total of 12,468 studies were searched from the databases, including 2409 from PubMed, 4387 from Embase, 170 from Cochrane Library, and 5502 from Web of Science. After removing duplicates and screening titles and abstracts, 378 articles remained. According to a full-text review, 53 articles [13-67] were included. Fifty-three articles presented the development of one or more prediction models for osteoporotic fracture, while twenty-six articles described the validation of one or more models. The search process is shown in **Figure 1**.

Fifty-three studies were ultimately included in our meta-analysis, involving 15,209,268 patients. Many studies originated from U. S. (n = 11),European(n = 11),and China (n = 8). Most studies were cohort studies (n = 46), and the rest were case-control studies (n = 7). The median age of osteoporosis patients was 68.8 years (ranging from 48.5 to 84). The study population in most studies covered women (n = 24). The fracture sites included multi-site (n = 26), vertebra (n = 14), hip (n = 12), and femur (n = 1). Most studies were based on clinical hospital data (n=19), while some used questionnaire collection data (n=10), osteoporosis registry data (n=9), electronic health records(n=7), and administrative data (n=6). Only 13 articles elucidated the cross-validation method. The baseline characteristics of the included studies are shown in **Supplemental Table S3**.

There were 86 prediction models specifically developed for the osteoporosis population and 41 validation sets. Ninety-eight ML models reported the C-index or the area under the receiver operating characteristic curve (AUC), ranging from 0.50 to 0.98. **Supplemental Table S4** shows all studies on the development and validation of ML models for outcome prediction in patients with osteoporosis. Among all the identified prediction models, the logistic regression (31.4%) was the most commonly used algorithm, followed by the survival model (18%).

The most commonly used predictors in ML models were age (n=72), body mass index (BMI) (n=40), past fracture history (n=35), bone mineral density T-score (n=33), history of falls (n=29), bone mineral density (BMD) (n=28), radiomics data (n=25), weight (n=24), height (n=23), gender (n=20), and other chronic diseases (n=20) (**Table 1**).

Table 1. Main sorts of predictors included in developed models for osteoporosis patients

Predictors	Number of models
Demographics	
Age	72
History of falls	29
Sex	20
Women's menopause age	8
Family genetic history	6
Race	5
Physical examination	
Body mass index,BMI	40
Bone mineral density t-score	33
Bone mineral density，BMD	28
Weight	24
Height	23
Motor ability	10
Lifestyle	
Alcohol consumption	13
Smoking	11
Physical activity	10
Lack of physical exercise	7
Daily activities	5
Limited physical activity	4
Frequent sun exposure	3
Comorbidity	
Past fracture history	35
Other chronic diseases	20
Osteoporosis	8
Rheumatoid arthritis	7
Genetic risk score (GRS)	5
Fracture type	4
Backache	2
Drug and nutrient intake	
Use of hormonal drugs	8
Calcium intake	8
Nutritional status	6
Intake of other drugs	4
Radiomics	

	Radiomic data	25
Mental state		
	Cognitive performance	3
	Anxiety/depression	2

Note: BMD (g/cm^2)

The risk of bias assessment of the included studies is summarized in **Figure 2**. More than half of these studies had a high risk of bias (67%). The risk of bias in most studies was low in terms of participants, predictors, and outcome. However, a high or unclear risk of bias in the statistical analysis was observed in all studies. More details are shown in **Supplemental Table S5**.

Sixty-six training datasets and 32 validation datasets were included in the meta-analysis of the C-index. Since substantial heterogeneity was present, we performed subgroup analyses based on fracture site and model type. **Table 2** shows the results of the meta-analysis of C-index of ML models in predicting osteoporosis. Logistic regression is the most widely used method. The forest plot of C-index is presented in **Supplemental Figure S1 and Figure S2**. The pooled C-index was 0.75 (95 % CI: 0.72, 0.78) (I^2 = 99.7%, P < 0.001) in the training set and 0.75 (95 % CI: 0.71, 0.78) (I^2 = 99.8%, P < 0.001) in the validation set. In the training set, other deep learning method showed the highest predictive performance (pooled C-index = 0.97), followed by convolutional neural network (CNN) (pooled C-index = 0.94), decision trees (pooled C-index = 0.78), and logistic regression (pooled C-index = 0.75). Furthermore, models for vertebral fracture (pooled C-index = 0.80) and hip fracture (pooled C-index = 0.76) outperformed those for multi-site fracture (pooled C-index = 0.70). However, in the validation set, CNN (pooled C-index = 0.98) showed the best performance, closely followed by other deep learning method (pooled C-index = 0.82), logistic regression (pooled C-index = 0.80), and support vector machines (pooled C-index = 0.78). Models for vertebral fracture (pooled C-index = 0.87) outperformed those for hip fracture (pooled C-index = 0.73) and multi-site fracture (pooled C-index = 0.71). Across these studies, we extracted 57 estimates of balanced accuracy (the average of the reported sensitivity and specificity), ranging from 0.59 to 1.00. As presented in **Table 3**, the mean sensitivity and specificity of models were 0.79 (95 % CI: 0.72, 0.84) (I^2 = 99.2%, P < 0.001) and 0.81 (95 % CI: 0.75, 0.86) in the training set (I^2 = 99.9%, P < 0.001), and 0.76 (95 % CI: 0.80, 0.81) (I^2 = 98.9%, P < 0.001) and 0.83 (95 % CI: 0.72, 0.90) in the validation set (I^2 = 99.9%, P < 0.001), respectively. The results of sensitivity analysis show that ML models built for different fracture sites have stable performance in the training and validation sets (**Supplemental Figure S3-S8**).

Table 2. Results of subgroup analysis of C-index by fracture site and machine learning type

subgroup	Training dataset		Validation dataset	
	N	C-statistic(95% CI)	N	C-statistic(95% CI)
Fracture site				
Vertebral fracture	15	0.80(0.74,0.87)	6	0.87(0.71,1.00)
Hip fracture	20	0.76(0.72,0.81)	9	0.73(0.65,0.81)
Multi-site fracture	31	0.70(0.67,0.72)	17	0.71(0.65,0.76)
Model type				
LR	26	0.75(0.72.0.78)	7	0.80(0.73,0.87)
ANN	4	0.73(0.64,0.82)	3	0.66(0.62,0.70)
CNN	2	0.95(0.94,0.96)	1	0.98(0.94,1.00)
RF	3	0.70(0.68,0.72)	3	0.66(0.59,0.73)
SVM	5	0.72(0.60,0.85)	3	0.78(0.59,0.96)
DT	2	0.78(0.56,0.99)	1	0.69(0.67,0.70)
NB	2	0.74(0.39,1.00)	-	
kNN	1	0.51(0.46,0.55)	-	
Survival model	13	0.70(0.69,0.74)	9	0.68(0.67,0.69)
Boosted tree	5	0.71(0.69,0.74)	3	0.70(0.69,0.71)
Enseble learning	1	0.72(0.71,0.73)		
Other DL	2	0.97(0.96,0.97)	1	0.82(0.77,0.87)
Overall	66	0.75（0.72,0.78）	32	0.75(0.71,0.78)

Note: LR:Logistic Regression;ANN:artificial neural network;CNN: convolutional neural network;RF:random forests;SVM:support-vector machine;DT:decision tree;NB:Naive Bayes;kNN:k-nearest neighbors;DL:deep learnimg model.

Table 3. Results of subgroup analysis of sensitivity and specificity by fracture site and machine learning type

subgroup	Training dataset			Validation dataset		
	N	Sensitivity (95% CI)	Specificity (95% CI)	N	Sensitivity (95% CI)	Specificity (95% CI)
Fracture site						
Vertebral fracture	10	0.73(0.61,0.82)	0.91(0.86,0.95)	3	0.87(0.70.0.95)	0.97(0.94,0.98)
Hip fracture	13	0.90（0.82,0.94）	0.82(0.75,0.88)	5	0.84(0.77,0.89)	0.85(0.80,0.89)

Multi-site fracture	18	0.71(0.59,0.81)	0.72(0.60,0.81)	8	0.66(0.61,0.70)	0.69(0.53,0.81)
Model type						
LR	17	0.70(0.63,0.77)	0.73(0.67,0.79)	4	0.66(0.55,0.75)	0.65(0.50,0.77)
ANN	4	0.91(0.70,0.98)	0.93(0.75,0.98)	3	0.78(0.71,0.83)	0.85(0.71,0.93)
CNN	3	0.83(0.81,0.84)	0.91(0.79,0.96)	1	0.98	0.95
RF	1	0.84	0.91	1	0.70	0.46
SVM	6	0.81(0.63,0.92)	0.63(0.13,0.95)	3	0.79(0.72,0.85)	0.89(0.79,0.94)
DT	2	0.97(0.53,1.00)	0.70(0.67,0.73)	-		
NB	2	0.63(0.13,0.95)	0.76（0.70,0.81）	-		
kNN	2	0.95（0.39,1.00)	0.80（0.77,0.83）	1	0.81	0.79
Survival model	1	0.81	0.52	-		
Boosted tree	1	0.59	0.67	1	0.70	0.95
Other DL	2	0.81(0.72,0.87)	0.96(0.93,0.98)	2	0.83(0.74,0.90)	0.95(0.92,0.97)
Overall	41	0.79(0.72,0.84)	0.81(0.75,0.86)	16	0.76(0.80,0.81)	0.83(0.72,0.90)

Discussion

ML is a popular research method that provides new tools for early detection of diseases. This study systematically explored the application of the latest ML methods in predicting fracture risk in osteoporosis. The most commonly used predictors in ML models are age, BMI, past fracture history, bone mineral density T-score, history of falls, BMD, radiomics data, weight, height, gender, and other chronic diseases. In general, most predictors included in model-development studies are traditional risk factors. A recent study showed that the most common risk factors for fragility fractures encompassed decreased bone mineral density, age, gender, low BMI, history of fragility fractures, family history of hip fractures, history of glucocorticoid therapy, smoking, excessive alcohol consumption, lack of vitamin D, early menopause, and immobility [68]. This is consistent with some common fracture predictors identified in our study. Our study also finds that radiomics data are frequently used as a fracture predictor in ML models for osteoporosis. A retrospective, single-center, preliminary investigation by Lim et al. reported ML based on radiomics features and Abdomen-pelvic CT for diagnosing osteoporosis showed high predictive performance, with accuracy, specificity, and negative predictive value exceeding 93% [69].

The present study found that ML methods commonly used in the field of osteoporosis included logistic regression, decision tree, random forest, survival model, support vector machine (SVM), ensemble learning, artificial neural network (ANN), CNN, and the latest deep learning technology. ML has a good performance in the prediction and identification of osteoporosis and fracture. In terms of the models in the training sets, the prediction efficiency of other deep learning method is optimal, followed by CNN, decision trees, and logistic regression. In the validation sets, CNN showed the best performance, closely followed by other deep learning

method, logistic regression, and SVM. Deep learning is more powerful than traditional machine learning algorithms, with a wide range of coverage. Its performance increases with the amount of data [70]. Deep learning has been successfully applied to assist in the diagnosis and prediction of osteoporotic fractures [34, 63]. CNN, a core algorithm of deep learning, is widely used in the field of data analysis and disease prediction with high accuracy [71]. CNN techniques can effectively predict the risk of osteoporotic fractures, enabling clinicians to take timely treatment measures, thereby reducing the occurrence of fractures [19, 23, 47]. Additionally, logistic regression is an efficient, simple, and easy-to-operation ML method that outputs calibrated predicted probabilities. An article on prediction models for the outcomes in patients with chronic obstructive pulmonary disease revealed that logistic regression was the most frequently used modeling method [72]. This is the same as recent findings reported by Kushan et al. [73]. Their ML models based on logistic regression outperformed those based on random forest and decision trees. Moreover, SVM adapts well to small samples and high-dimensional data with a low misclassification rate, and therefore can be used for classification and regression analysis [74].

Most included studies report multiple outcomes, such as sensitivity, specificity, AUC, and ROC. The mean sensitivity of the models in the training set model is 0.79 (95 % CI: 0.72, 0.84), greater than that of the models in the validation set. Most models are internally validated in the same population database and lack external validations in other populations. Only ML models in six articles were externally validated [32, 34, 36, 48, 55, 67]. Therefore, external validations of ML models for predicting fracture risk are needed. However, a single performance measure such as AUC or ROC is insufficient to recommend the application of ML models into clinical practice [8], and multiple measures of performance should be combined.

This systematic review and meta-analysis summarized a large number of studies to comprehensively evaluate the predictive performance of ML for fracture risk in patients with osteoporosis. The characteristics of the established and validated models were described. We performed a quantitative synthesis that was never done in previous studies to compare these models. Furthermore, the meta-analysis of C-index was performed using the random-effects model, since the C-index was reported in most predictive models [72, 75]. Meanwhile, the bivariate mixed-effects model was used for the meta-analysis of sensitivity and specificity. In the training dataset, the sensitivity of hip fracture was the highest, closely followed by multi-site fracture and vertebral fracture. For patients with hip fractures, radiographs may cause missed diagnosis and misdiagnosis, leading to poor prognosis [76]. ML models have been increasingly used to identify hip fracture risk with high accuracy [31]. ML has a stronger power to recognize images and can assist inexperienced clinicians in a highly accurate diagnosis.

Some limitations still need to be considered in the present study. Due to incomplete reporting of indicators, several studies were only included in the systematic review and were excluded from subsequent meta-analyses [61, 63]. Studies conducted in either Western or Asian populations lack external validation, and thus external validations in other populations are needed to widen the application of ML models. The risk of bias assessment demonstrated that most studies (67%) had a high risk of bias, regardless of whether they involved the development or external validation of a prediction model for the osteoporosis population. The main bias came from the statistical analysis, because most studies did not properly handle continuous and categorical variables and reported no method for processing missing values. Only three articles reported the use of median imputation

or multiple interpolation method to deal with missing values [15, 31, 66], while others did not mention how to deal with missing values. These shortcomings in the methodology may be due to a lack of guidelines for the standard reporting of risk prediction studies at that time. In addition, some models were reported with little information, making it unable for other researchers to perform external validation, much less the application in clinical practice. For example, only 12 articles used the K-fold cross-validation method to improve the accuracy of their algorithms [15, 16, 19, 27, 30, 31, 34, 36, 42, 46, 47, 62], but most of the eligible articles did not. Models without stringent validation cannot be widely applied [73]. Many studies have limited applicability in clinical practice because of flawed methodologies or unrepresentative data sets. Future research should give priority to the development of practical algorithms. Furthermore, we observed large heterogeneity in the meta-analysis of C statistics. Potential sources of heterogeneity may be the differences in patients' characteristics, data sources, and analysis methods across the studies. More than 30% of the research data came from clinical studies, and clinical data are heterogeneous and usually imbalanced. At last, most ML models did not report balanced accuracy and lacked calibration or external validation or decision curves. Thus, further research is required to address these issues, improving the generalization of the models.

Despite the limitations mentioned above, the present study can still provide meaningful recommendations for future research and practice. First, the major strength of our study is the rigorous literature search and methodology to provide reliable estimates. This is the latest systematic review and meta-analysis conducted to comprehensively assess ML models for fracture risk. Second, ML models can provide convincing evidence to assist clinicians in making more accurate judgments during highly complex decision-making processes, with certain clinical application values[74]. More rigorous, robust, and comprehensive research is warranted to assess its clinical application and impact on clinicians and patients. Third, the advances in emerging technologies such as ML have opened a new era of clinical medical research, providing new directions for solving intricate problems with classical statistical methods. However, clinicians currently are not skillful in using such emerging technologies. Therefore, it is advisable for clinicians to improve their ability to use ML to make more accurate diagnoses.

In conclusion, ML has a favorable predictive performance for fracture risk in patients with osteoporosis and can be used as a potential tool for early identification of fracture risk in this population. However, most current studies lack external validation. Therefore, future research is needed to validate and improve the existing predictive models for osteoporosis risk rather than developing new models.

Declarations

Funding

This work was supported by National Natural Science Foundation of China (81872711), Postgraduate Research & Practice Innovation Program of Jiangsu Province（KYCX23_0328).

Authors Contributions

Yanqian Wu and Jianqian Chao designed the review,developed the inclusion criteria, screened titles and abstracts, appraised the quality of included papers, and drafted the manuscript. Min Bao and Na Zhang collected and cleaned the data,also analysed the data. All authors critically reviewed drafts and approved the final

manuscript.

Competing interests

Yanqian Wu, Jianqian Chao, Min Bao and Na Zhang declare that they have no conflict of interest.

Ethical Approval

This study does not involve human participants and ethical approval was not required.

Data sharing statement

Data are available upon reasonable request. Data may be obtained from the corresponding author and are not publicly available.

References

1 Kanis JA. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. WHO Study Group. *Osteoporos Int* 1994;4(6):368-81. doi: 10.1007/bf01622200

2 Veronese N, Kolk H, Maggi S. Epidemiology of Fragility Fractures and Social Impact. In: Falaschi P, Marsh D, editors. Orthogeriatrics: The Management of Older Patients with Fragility Fractures. Cham (CH): Springer

Copyright 2021, The Author(s). 2021. p. 19-34.

3 Piscitelli P, Feola M, Rao C*, et al.* Ten years of hip fractures in Italy: For the first time a decreasing trend in elderly women. *World J Orthop* 2014;5(3):386-91. doi: 10.5312/wjo.v5.i3.386

4 Borgström F, Karlsson L, Ortsäter G*, et al.* Fragility fractures in Europe: burden, management and opportunities. *Arch Osteoporos* 2020;15(1):59. doi: 10.1007/s11657-020-0706-y

5 Zimmermann EA, Busse B, Ritchie RO. The fracture mechanics of human bone: influence of disease and treatment. *Bonekey Rep* 2015;4:743. doi: 10.1038/bonekey.2015.112

6 Gupta R, Srivastava D, Sahu M*, et al.* Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers* 2021;25(3):1315-60. doi: 10.1007/s11030-021-10217-3

7 Ferizi U, Honig S, Chang G. Artificial intelligence, osteoporosis and fragility fractures. *Curr Opin Rheumatol* 2019;31(4):368-75. doi: 10.1097/bor.0000000000000607

8 Smets J, Shevroja E, Hügle T*, et al.* Machine Learning Solutions for Osteoporosis-A Review. *J Bone Miner Res* 2021;36(5):833-51. doi: 10.1002/jbmr.4292

9 Anam M, Ponnusamy V-ap, Hussain M*, et al.* Osteoporosis Prediction for Trabecular Bone using Machine Learning: A Review. *Computers, Materials & Continua* 2021;67(1):89--105. doi: 10.32604/cmc.2021.013159

10 Liberati A, Altman DG, Tetzlaff J*, et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;6(7):e1000100. doi: 10.1371/journal.pmed.1000100

11 Palazón-Bru A, Martín-Pérez F, Mares-García E, *et al.* A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models. *Stat Med* 2020;39(23):3207-25. doi: 10.1002/sim.8660

12 Nagendran M, Chen Y, Lovejoy CA, *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Bmj* 2020;368:m689. doi: 10.1136/bmj.m689

13 Debray TP, Damen JA, Riley RD, *et al.* A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019;28(9):2768-86. doi: 10.1177/0962280218785504

14 Snell KI, Ensor J, Debray TP, *et al.* Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2018;27(11):3505-22. doi: 10.1177/0962280217705678

15 Wu Q, Nasoz F, Jung J, *et al.* Machine Learning Approaches for Fracture Risk Assessment: A Comparative Analysis of Genomic and Phenotypic Data in 5130 Older Men. *Calcif Tissue Int* 2020;107(4):353-61. doi: 10.1007/s00223-020-00734-y

16 Villamor E, Monserrat C, Del Río L, *et al.* Prediction of osteoporotic hip fracture in postmenopausal women through patient-specific FE analyses and machine learning. *Comput Methods Programs Biomed* 2020;193:105484. doi: 10.1016/j.cmpb.2020.105484

17 van Geel TA, Nguyen ND, Geusens PP, *et al.* Development of a simple prognostic nomogram for individualising 5-year and 10-year absolute risks of fracture: a population-based prospective study among postmenopausal women. *Ann Rheum Dis* 2011;70(1):92-7. doi: 10.1136/ard.2010.131813

18 Ulivieri FM, Rinaudo L, Piodi LP, *et al.* Bone strain index as a predictor of further vertebral fracture in osteoporotic women: An artificial intelligence-based analysis. *PLoS One* 2021;16(2):e0245967. doi: 10.1371/journal.pone.0245967

19 Yoda T, Maki S, Furuya T, *et al.* Automated Differentiation Between Osteoporotic Vertebral Fracture and Malignant Vertebral Fracture on MRI Using a Deep Convolutional Neural Network. *Spine (Phila Pa 1976)* 2022;47(8):E347-e52. doi: 10.1097/brs.0000000000004307

20 Jiang X, Westermann LB, Galleo GV, *et al.* Age as a predictor of osteoporotic fracture compared with current risk-prediction models. *Obstet Gynecol* 2013;122(5):1040-46. doi: 10.1097/AOG.0b013e3182a7e29b

21 Schousboe JT, Rosen HR, Vokes TJ, *et al.* Prediction models of prevalent radiographic vertebral fractures among older women. *J Clin Densitom* 2014;17(3):378-85. doi: 10.1016/j.jocd.2013.09.021

22 Sandhu SK, Nguyen ND, Center JR, *et al.* Prognosis of fracture: evaluation of predictive accuracy of the FRAX algorithm and Garvan nomogram. *Osteoporos Int* 2010;21(5):863-71. doi: 10.1007/s00198-009-1026-7

23 Rubin KH, Möller S, Holmberg T, *et al.* A New Fracture Risk Assessment Tool (FREM) Based on Public Health Registries. *J Bone Miner Res* 2018;33(11):1967-79. doi: 10.1002/jbmr.3528

24 Pluskiewicz W, Adamczyk P, Franek E, *et al.* Ten-year probability of osteoporotic fracture in 2012 Polish women assessed by FRAX and nomogram by Nguyen et al.-Conformity between methods and their clinical utility. *Bone* 2010;46(6):1661-7. doi: 10.1016/j.bone.2010.02.012

25 Jang EJ, Lee YK, Choi HJ, *et al.* Osteoporotic Fracture Risk Assessment Using Bone Mineral Density in Korean: A Community-based Cohort Study. *J Bone Metab* 2016;23(1):34-9. doi: 10.11005/jbm.2016.23.1.34

26 Monchka BA, Kimelman D, Lix LM, *et al.* Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: The Manitoba Bone Mineral Density Registry. *Bone* 2021;150:116017. doi: 10.1016/j.bone.2021.116017

27 Mehta SD, Sebro R. Computer-Aided Detection of Incidental Lumbar Spine Fractures from Routine Dual-Energy X-Ray Absorptiometry (DEXA) Studies Using a Support Vector Machine (SVM) Classifier. *J Digit Imaging* 2020;33(1):204-10. doi: 10.1007/s10278-019-00224-0

28 Langsetmo L, Nguyen TV, Nguyen ND, *et al.* Independent external validation of nomograms for predicting risk of low-trauma fracture and hip fracture. *Cmaj* 2011;183(2):E107-14. doi: 10.1503/cmaj.100458

29 Ioannidis G, Jantzi M, Bucek J, *et al.* Development and validation of the Fracture Risk Scale (FRS) that predicts fracture over a 1-year time period in institutionalised frail older people living in Canada: an electronic record-linked longitudinal cohort study. *BMJ Open* 2017;7(9):e016477. doi: 10.1136/bmjopen-2017-016477

30 Nishiyama KK, Macdonald HM, Hanley DA, *et al.* Women with previous fragility fractures can be classified based on bone microarchitecture and finite element analysis measured with HR-pQCT. *Osteoporos Int* 2013;24(5):1733-40. doi: 10.1007/s00198-012-2160-1

31 Kruse C, Eiken P, Vestergaard P. Machine Learning Principles Can Improve Hip Fracture Prediction. *Calcif Tissue Int* 2017;100(4):348-60. doi: 10.1007/s00223-017-0238-7

32 Kolanu N, Brown AS, Beech A, *et al.* Natural language processing of radiology reports for the identification of patients with fracture. *Arch Osteoporos* 2021;16(1):6. doi: 10.1007/s11657-020-00859-5

33 Kim HY, Jang EJ, Park B, *et al.* Development of a Korean Fracture Risk Score (KFRS) for Predicting Osteoporotic Fracture Risk: Analysis of Data from the Korean National Health Insurance Service. *PLoS One* 2016;11(7):e0158918. doi: 10.1371/journal.pone.0158918

34 Hsieh CI, Zheng K, Lin C, *et al.* Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nat Commun* 2021;12(1):5472. doi: 10.1038/s41467-021-25779-x

35 Hong N, Park H, Kim CO, *et al.* Bone Radiomics Score Derived From DXA Hip Images Enhances Hip Fracture Prediction in Older Women. *J Bone Miner Res* 2021;36(9):1708-16. doi: 10.1002/jbmr.4342

36 Ho-Le TP, Center JR, Eisman JA, *et al.* Prediction of hip fracture in post-menopausal women using artificial neural network approach. *Annu Int Conf IEEE Eng Med Biol Soc* 2017;2017:4207-10. doi: 10.1109/embc.2017.8037784

37 Henry MJ, Pasco JA, Merriman EN, *et al.* Fracture risk score and absolute risk of fracture. *Radiology* 2011;259(2):495-501. doi: 10.1148/radiol.10101406

38 Galassi A, Martín-Guerrero JD, Villamor E, *et al.* Risk Assessment of Hip Fracture Based on Machine Learning. *Appl Bionics Biomech* 2020;2020:8880786. doi: 10.1155/2020/8880786

39 FitzGerald G, Compston JE, Chapurlat RD, *et al.* Empirically based composite fracture prediction model from the Global Longitudinal Study of Osteoporosis in Postmenopausal Women (GLOW). *J Clin Endocrinol Metab* 2014;99(3):817-26. doi: 10.1210/jc.2013-3468

40 Ferizi U, Besser H, Hysi P, *et al.* Artificial Intelligence Applied to Osteoporosis: A Performance Comparison of Machine Learning Algorithms in Predicting Fragility Fractures From MRI Data. *J Magn Reson Imaging* 2019;49(4):1029-38. doi: 10.1002/jmri.26280

41 Enns-Bray WS, Bahaloo H, Fleps I, *et al.* Biofidelic finite element models for accurately classifying hip fracture in a retrospective clinical study of elderly women from the AGES Reykjavik cohort. *Bone* 2019;120:25-37. doi: 10.1016/j.bone.2018.09.014

42 Engels A, Reber KC, Lindlbauer I, *et al.* Osteoporotic hip fracture prediction from risk factors available in administrative claims data - A machine learning approach. *PLoS One* 2020;15(5):e0232969. doi: 10.1371/journal.pone.0232969

43 de Vries BCS, Hegeman JH, Nijmeijer W, *et al.* Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis. *Osteoporos Int* 2021;32(3):437-49. doi: 10.1007/s00198-020-05735-z

44 Cheung EY, Bow CH, Cheung CL, *et al.* Discriminative value of FRAX for fracture prediction in a cohort of Chinese postmenopausal women. *Osteoporos Int* 2012;23(3):871-8. doi: 10.1007/s00198-011-1647-5

45 Chanplakorn P, Lertudomphonwanit T, Daraphongsataporn N, *et al.* Development of prediction model for osteoporotic vertebral compression fracture screening without using clinical risk factors, compared with FRAX and other previous models. *Arch Osteoporos* 2021;16(1):84. doi: 10.1007/s11657-021-00957-y

46 Bredbenner TL, Mason RL, Havill LM, *et al.* Fracture risk predictions based on statistical shape and density modeling of the proximal femur. *J Bone Miner Res* 2014;29(9):2090-100. doi: 10.1002/jbmr.2241

47 Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. *Jt Dis Relat Surg* 2020;31(2):175-83. doi: 10.5606/ehc.2020.72163

48 Berry SD, Zullo AR, Lee Y, *et al.* Fracture Risk Assessment in Long-term Care (FRAiL): Development and Validation of a Prediction Model. *J Gerontol A Biol Sci Med Sci* 2018;73(6):763-69. doi: 10.1093/gerona/glx147

49 Beaudoin C, Jean S, Moore L, *et al.* Prediction of Osteoporotic Fractures in Elderly Individuals: A Derivation and Internal Validation Study Using Healthcare Administrative Data. *J Bone Miner Res* 2021;36(12):2329-42. doi: 10.1002/jbmr.4438

50 Baleanu F, Moreau M, Charles A, *et al.* Fragility Fractures in Postmenopausal Women: Development of 5-Year Prediction Models Using the FRISBEE Study. *J Clin Endocrinol Metab* 2022;107(6):e2438-e48. doi: 10.1210/clinem/dgac092

51 Almog YA, Rai A, Zhang P, *et al.* Deep Learning With Electronic Health Records for Short-Term Fracture Risk Identification: Crystal Bone Algorithm Development and Validation. *J Med Internet Res* 2020;22(10):e22550. doi: 10.2196/22550

52 Zagórski P, Tabor E, Martela-Tomaszek K, *et al.* Five-year fracture risk assessment in postmenopausal women, using both the POL-RISK calculator and the Garvan nomogram: the Silesia Osteo Active Study. *Arch Osteoporos* 2021;16(1):32. doi: 10.1007/s11657-021-00881-1

53 Díez-Pérez A, González-Macías J, Marín F, *et al.* Prediction of absolute risk of non-spinal fractures using clinical risk factors and heel quantitative ultrasound. *Osteoporos Int* 2007;18(5):629-39. doi: 10.1007/s00198-006-0297-5

54 Lix LM, Leslie WD, Majumdar SR. Measuring improvement in fracture risk prediction for a new risk factor: a simulation. *BMC Res Notes* 2018;11(1):62. doi: 10.1186/s13104-018-3178-z

55 Li Q, Long X, Wang Y, *et al.* Development and validation of a nomogram for predicting the probability of new vertebral compression fractures after vertebral augmentation of osteoporotic vertebral compression fractures. *BMC Musculoskelet Disord* 2021;22(1):957. doi: 10.1186/s12891-021-04845-x

56 Lee S, Lee JW, Jeong JW, *et al.* A preliminary study on discrimination of osteoporotic fractured group from nonfractured group using support vector machine. *Annu Int Conf IEEE Eng Med Biol Soc* 2008;2008:474-7. doi: 10.1109/iembs.2008.4649193

57 Jacobs JW, Da Silva JA, Armbrecht G, *et al.* Prediction of vertebral fractures is specific for gender and site of bone mineral density measurement. *J Rheumatol* 2010;37(1):149-54. doi: 10.3899/jrheum.090731

58 Eller-Vainicher C, Chiodini I, Santi I, *et al.* Recognition of morphometric vertebral fractures by artificial neural networks: analysis from GISMO Lombardia Database. *PLoS One* 2011;6(11):e27277. doi: 10.1371/journal.pone.0027277

59 Zhong BY, He SC, Zhu HD, *et al.* Risk Prediction of New Adjacent Vertebral Fractures After PVP for Patients with Vertebral Compression Fractures: Development of a Prediction Model. *Cardiovasc Intervent Radiol* 2017;40(2):277-84. doi: 10.1007/s00270-016-1492-1

60 Xiao X, Wu Q. The Utility of Genetic Risk Score to Improve Performance of FRAX for Fracture Prediction in US Postmenopausal Women. *Calcif Tissue Int* 2021;108(6):746-56. doi: 10.1007/s00223-021-00809-4

61 Du J, Wang J, Gai X, *et al.* Application of intelligent X-ray image analysis in risk assessment of osteoporotic fracture of femoral neck in the elderly. *Math Biosci Eng* 2023;20(1):879-93. doi: 10.3934/mbe.2023040

62 Wang M, Chen X, Cui W, *et al.* A Computed Tomography-based Radiomics Nomogram for Predicting Osteoporotic Vertebral Fractures: A Longitudinal Study. *J Clin Endocrinol Metab* 2023;108(6):e283-e94. doi: 10.1210/clinem/dgac722

63 Dong Q, Luo G, Lane NE, *et al.* Deep Learning Classification of Spinal Osteoporotic Compression Fractures on Radiographs using an Adaptation of the Genant Semiquantitative Criteria. *Acad Radiol* 2022;29(12):1819-32. doi: 10.1016/j.acra.2022.02.020

64 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(11):2925-38. doi: 10.1111/os.13531

65 Pluskiewicz W, Adamczyk P, Werner A, *et al.* POL-RISK: an algorithm for 10-year fracture risk prediction in postmenopausal women from the RAC-OST-POL study. *Pol Arch Intern Med* 2023;133(3). doi: 10.20452/pamw.16395

66 Kong XK, Zhao ZY, Zhang D, *et al.* Major osteoporosis fracture prediction in type 2 diabetes: a derivation and comparison study. *Osteoporos Int* 2022;33(9):1957-67. doi: 10.1007/s00198-022-06425-8

67 Agarwal A, Baleanu F, Moreau M, *et al.* External validation of FRISBEE 5-year fracture prediction models: a registry-based cohort study. *Arch Osteoporos* 2022;18(1):13. doi: 10.1007/s11657-022-01205-7

68 Pisani P, Renna MD, Conversano F, *et al.* Major osteoporotic fragility fractures: Risk factor updates and societal impact. *World J Orthop* 2016;7(3):171-81. doi: 10.5312/wjo.v7.i3.171

69 Lim HK, Ha HI, Park SY, *et al.* Prediction of femoral osteoporosis using machine-learning analysis with radiomics features and abdomen-pelvic CT: A retrospective single center preliminary study. *PLoS One* 2021;16(3):e0247330. doi: 10.1371/journal.pone.0247330

70 Avci O, Abdeljaber O, Kiranyaz S, *et al.* A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. *Mechanical Systems and Signal Processing* 2021;147:107077. doi: 10.1016/j.ymssp.2020.107077

71 Shen SC, Peña Fernández M, Tozzi G, *et al.* Deep learning approach to assess damage mechanics of bone tissue. *J Mech Behav Biomed Mater* 2021;123:104761. doi: 10.1016/j.jmbbm.2021.104761

72 Bellou V, Belbasis L, Konstantinidis AK, *et al.* Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *Bmj* 2019;367:l5358. doi: 10.1136/bmj.l5358

73 Silva K, Lee WK, Forbes A, *et al.* Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *Int J Med Inform* 2020;143:104268. doi: 10.1016/j.ijmedinf.2020.104268

74 Jain R, Sontisirikit S, Iamsirithaworn S, *et al.* Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC Infect Dis* 2019;19(1):272. doi: 10.1186/s12879-019-3874-x

75 Fleuren LM, Klausch TLT, Zwager CL, *et al.* Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020;46(3):383-400. doi: 10.1007/s00134-019-05872-y

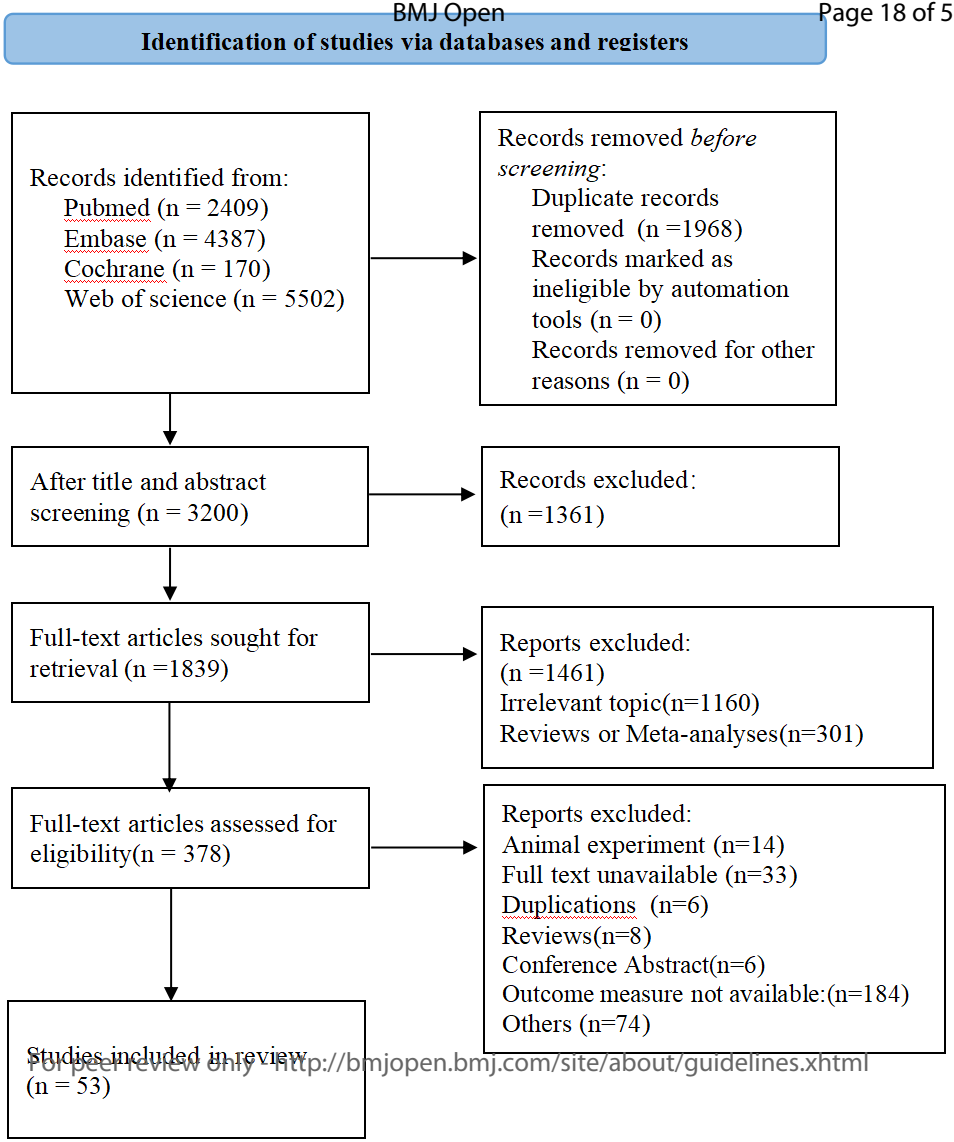
76 Hossain M, Akbar SA, Andrew G. Misdiagnosis of occult hip fracture is more likely in patients with poor mobility and cognitive impairment. *Acta Orthop Belg* 2010;76(3):341-6.

Figure legends

Figure 1 The flow chart of retrieval process

Figure 2 Risk of bias assessment (using PROBAST) based on four domains across

all machine learning models



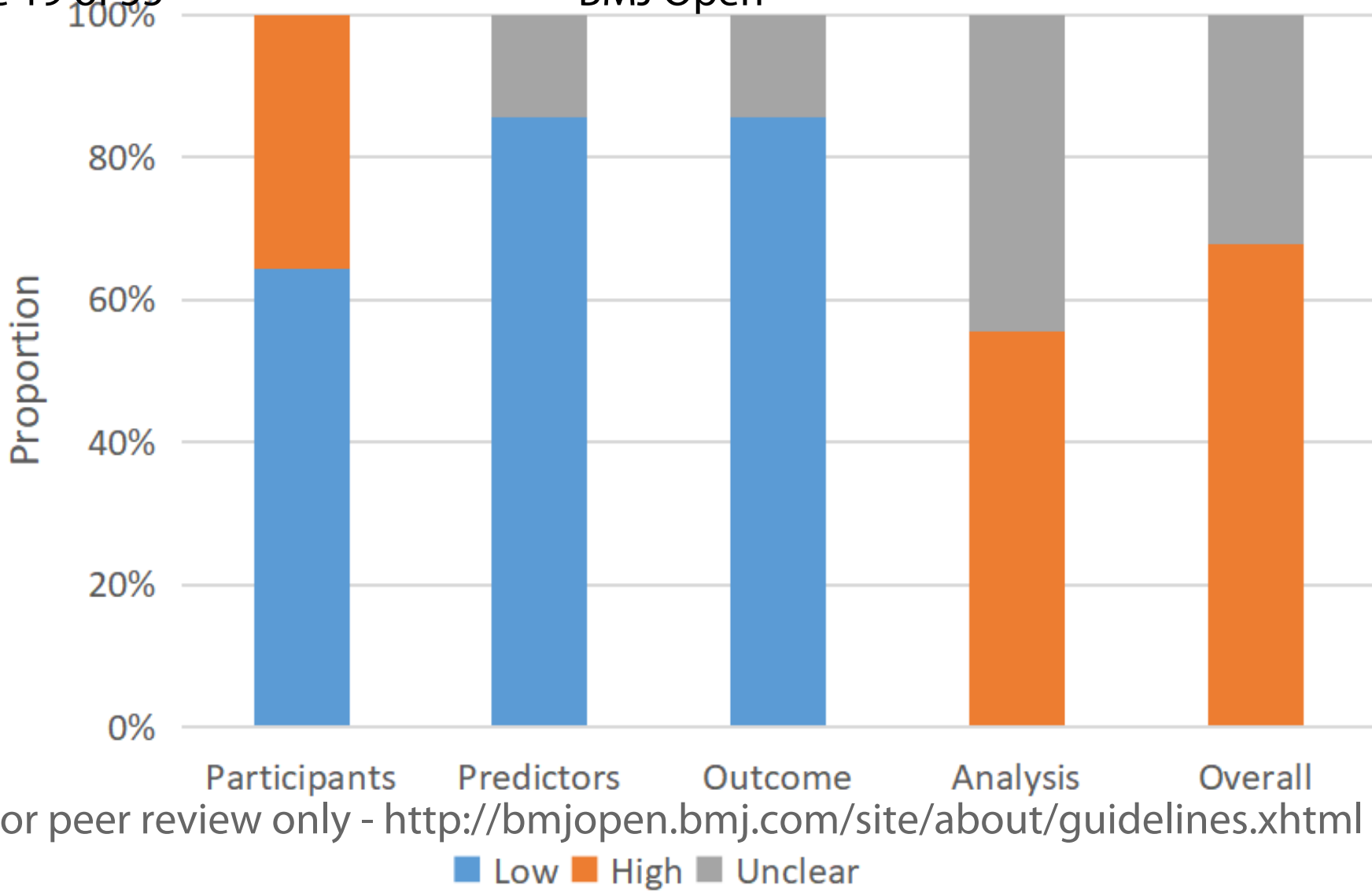


Table S1 : PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	2-3
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	3
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	3
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	3
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Table S2
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	3，Figure 1
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	3
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points,	3

Table S1 : PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
		analyses), and if not, the methods used to decide which results to collect.	
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Table S3,Table S4
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	3-4，Figure 2
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	4
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	4
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	4
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	4
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	3-4
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Table2,Table3
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	3
Reporting bias	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from	3

Table S1 : PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
assessment		reporting biases).	
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	4
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	4
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1
Study characteristics	17	Cite each included study and present its characteristics.	4,Table S3
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	6,Figure 2,Table S5
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	6,Table2,Table3
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	6
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	6,Table 2
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	6-7,Table 2
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	6-8,Table 3
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	6,Figure 2
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Table 2,Table 3

Table S1 : PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	8-10
	23b	Discuss any limitations of the evidence included in the review.	8-10
	23c	Discuss any limitations of the review processes used.	8-10
	23d	Discuss implications of the results for practice, policy, and future research.	10
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	4,Registration No. CRD42022346896
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	The protocol was registered on PROSPERO
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	10
Competing interests	26	Declare any competing interests of review authors.	11
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Table S3,Table S4

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

Table S2 Literature search strategy

1.Pubmed

Search number	Query	Results
#1	"Osteoporosis"[Mesh]	62,328
#2	"Osteoporosis"[Title/Abstract] OR "Osteoporoses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR "age related bone loss"[Title/Abstract] OR "age related bone losses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR ((("bone and bones"[MeSH Terms] OR ("Bone"[All Fields] AND "bones"[All Fields]) OR "bone and bones"[All Fields] OR "Bone"[All Fields]) AND "Losses"[All Fields]) AND "Age-Related"[Title/Abstract])	82,879
#3	"Osteoporosis"[MeSH Terms] OR ("Osteoporosis"[Title/Abstract] OR "Osteoporoses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR "age related bone loss"[Title/Abstract] OR "age related bone losses"[Title/Abstract] OR "bone loss age related"[Title/Abstract] OR ((("bone and bones"[MeSH Terms] OR ("Bone"[All Fields] AND "bones"[All Fields]) OR "bone and bones"[All Fields] OR "Bone"[All Fields]) AND "Losses"[All Fields]) AND "Age-Related"[Title/Abstract]))	100,673
#4	"Machine Learning"[Mesh]	55,536
#5	"machine learning"[Title/Abstract] OR "transfer learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "prediction model"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "random forest"[Title/Abstract] OR "artificial neural network"[Title/Abstract] OR "ANN"[Title/Abstract] OR "support vector machine"[Title/Abstract] OR "SVM"[Title/Abstract] OR "gradient boosting machine"[Title/Abstract] OR "GBM"[Title/Abstract] OR "Nomogram"[Title/Abstract] OR "XGboost"[Title/Abstract] OR "Logistic"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "external validation"[Title/Abstract]OR "cox"[Title/Abstract]	708,017
#6	"Machine Learning"[MeSH Terms] OR "Machine Learning"[Title/Abstract] OR "transfer learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "prediction model"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "random forest"[Title/Abstract] OR "artificial neural network"[Title/Abstract] OR "ANN"[Title/Abstract] OR "support vector machine"[Title/Abstract] OR	869,928

	"SVM"[Title/Abstract] OR "gradient boosting machine"[Title/Abstract] OR "GBM"[Title/Abstract] OR "Nomogram"[Title/Abstract] OR "XGboost"[Title/Abstract] OR "Logistic"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "external validation"[Title/Abstract]OR "cox"[Title/Abstract]	
#7	"fractures, bone"[MeSH]	299,700
#8	"fractures bone"[Title/Abstract] OR "broken bones"[Title/Abstract] OR "bone broken"[Title/Abstract] OR "bones broken"[Title/Abstract] OR "broken bone"[Title/Abstract] OR "Fractures"[Title/Abstract] OR "Fracture"[Title/Abstract]	343,051
#9	"fractures, bone"[MeSH Terms] OR "fractures bone"[Title/Abstract] OR "broken bones"[Title/Abstract] OR "bone broken"[Title/Abstract] OR "bones broken"[Title/Abstract] OR "broken bone"[Title/Abstract] OR "Fractures"[Title/Abstract] OR "Fracture"[Title/Abstract]	325,361
#10	#3AND #6 AND #9	2,409

2.Cochrane

Search number	Query	Results
#1	MeSH descriptor: [Osteoporosis] explode all trees	5,754
#2	(Osteoporosis):ti,ab,kw OR (Osteoporoses):ti,ab,kw OR (Bone Loss, Age-Related):ti,ab,kw OR (Age-Related Bone Loss):ti,ab,kw OR (Age-Related Bone Losses):ti,ab,kw	11,868
#3	(Bone Loss, Age Related):ti,ab,kw OR (Bone Losses, Age-Related):ti,ab,kw	549
#4	#1 OR #2 OR #3	12,188
#5	MeSH descriptor: [Machine Learning] explode all trees	866
#6	(machine learning):ti,ab,kw OR (Transfer Learning):ti,ab,kw OR (Deep learning):ti,ab,kw OR (Prediction model):ti,ab,kw OR (artificial intelligence):ti,ab,kw	10,742
#7	(random forest):ti,ab,kw OR (artificial neural network):ti,ab,kw OR (ANN):ti,ab,kw OR (Support vector machine):ti,ab,kw OR (SVM):ti,ab,kw	3,194
#8	(Gradient Boosting Machine):ti,ab,kw OR (GBM):ti,ab,kw OR (Nomogram):ti,ab,kw OR (XGboost):ti,ab,kw OR (Logistic):ti,ab,kw	32,161

#9	(Decision tree):ti,ab,kw OR (External validation):ti,ab,kw	2,025
#10	#5 OR #6 OR #7 OR #8 OR #9	43,070
#11	MeSH descriptor: [Fractures, Bone] explode all trees	8,166
#12	(Fractures, Bone):ti,ab,kw OR (Broken Bones):ti,ab,kw OR (Bone, Broken):ti,ab,kw OR (Bones, Broken):ti,ab,kw OR (Broken Bone):ti,ab,kw	9,471
#13	(Fractures):ti,ab,kw OR (Fracture):ti,ab,kw	27,252
#14	#11 OR #12 OR #13	27,386
#15	#4 AND #10 AND #14	170

3.Embase

Search number	Query	Results
#1	'osteoporosis'/exp	152,054
#2	'osteoporosis':ab,ti OR 'osteoporoses':ab,ti OR 'bone loss, age-related':ab,ti OR 'age-related bone loss':ab,ti OR 'age-related bone losses':ab,ti OR 'bone loss, age related':ab,ti OR 'bone losses, age-related':ab,ti	121,472
#3	#1 OR #2	176,124
#4	'machine learning'/exp	377,384
#5	'machine learning':ab,ti OR 'transfer learning':ab,ti OR 'deep learning':ab,ti OR 'prediction model':ab,ti OR 'artificial intelligence':ab,ti OR 'random forest':ab,ti OR 'artificial neural network':ab,ti OR ann:ab,ti OR 'support vector machine':ab,ti OR svm:ab,ti OR 'gradient boosting machine':ab,ti OR gbm:ab,ti OR nomogram:ab,ti OR xgboost:ab,ti OR logistic:ab,ti OR 'decision tree':ab,ti OR 'external validation':ab,ti OR 'cox':ab,ti	1,265,200
#6	#4 OR #5	1,485,765
#7	'fracture'/exp	383,399
#8	'fractures, bone':ab,ti OR 'broken bones':ab,ti OR 'bone, broken':ab,ti OR 'bones, broken':ab,ti OR 'broken bone':ab,ti OR 'fractures':ab,ti OR 'fracture':ab,ti	363,644

#9	#7 OR #8	471,174
#10	#3 AND #6 AND #9	4,387

4.Web of science

Search number	Query	Results
#1	Osteoporosis (Topic) or Osteoporoses (Topic) or Bone Loss, Age-Related (Topic) or Age-Related Bone Loss (Topic) or Age-Related Bone Losses (Topic) or Bone Loss, Age Related (Topic) or Bone Losses, Age-Related (Topic)	210,210
#2	machine learning (Topic) or Transfer Learning (Topic) or Deep learning (Topic) or Prediction model (Topic) or artificial intelligence (Topic) or random forest (Topic) or artificial neural network (Topic) or ANN (Topic) or Support vector machine (Topic) or SVM (Topic) or Gradient Boosting Machine (Topic) or GBM (Topic) or Nomogram (Topic) or XGboost (Topic) or Logistic (Topic) or Decision tree (Topic) or External validation (Topic) or Cox (Topic)	3,698,410
#3	Fractures, Bone (Topic) or Broken Bones (Topic) or Bone, Broken (Topic) or Bones, Broken (Topic) or Broken Bone (Topic) or Fractures (Topic) or Fracture (Topic)	1,302,805
#4	#1 AND #2 AND #3	5,502

Table S3 Characteristics of included studies in meta-analysis

Author	Year	Country	Data source	Sample population type	Mean age,years	Fracture site	Total sample, n	Validation Method	ML method	Model evaluation metrics
Wu, Q[15]	2020	USA	gene database	men	74.8	multiple	5130	internal	LR ANN RF BT	AUC Sensitivity Specificity Accuracy
Villamor, E[16]	2020	Spain	clinical hospital	women	81.4	hip	137	internal	LR SVM ANN RF	Accuracy
Van Geel, Tacm[17]	2011	Netherlands	questionnaire collection	women	62	vertebral	2372	-	SM	AUC
Ulivieri, F. M[18]	2021	Italy	clinical hospital	patient	48.5	vertebral	90	-	ANN	Sensitivity Specificity Accuracy ROC
Yoda, T[19]	2021	Japan	clinical hospital	patient	77.6	vertebral	97	internal	CNN	AUC Sensitivity Specificity
Jiang, X. Z[20]	2013	USA	clinical hospital	women	61.4	multiple	615	-	LR	AUC Sensitivity Specificity Accuracy
Schousboe, J. T[21]	2014	USA	clinical hospital	women	75	vertebral	7233	-	LR	AUC ROC
Sandhu, S. K[22]	2010	Australia	electronic health record	patient	74	multiple	200	-	LR	AUC
Rubin, K. H[23]	2018	Denmark	administrative	subjects	61.4	multiple	2495339	internal	LR	AUC ROC Accuracy PPV NPV
Pluskiewicz, W[24]	2010	Poland	osteoporosis registry	women	68.5	multiple	2012	-	LR	ROC AUC
Jang, E. J[25]	2016	Korea	questionnaire collection	subjects	61	multiple	768	-	LR	C-statistics

Barret A. Monchka[26]	2021	Canada	osteoporosis registry	subjects	75	vertebral	12742	internal	CNN	AUC Sensitivity Specificity Accuracy PPV NPV
Mehta, S. D[27]	2020	USA	clinical hospital	patient	69	vertebral	307	internal	SVM	AUC ROC Sensitivity Specificity Accuracy PPV NPV
Langsetmo, L[28]	2011	Canada	questionnaire collection	subjects	67.6	multiple	5758	internal	SM	C-Statistics ROC
Ioannidis, G[29]	2017	Canada	electronic health record	subjects	61	multiple	29848	internal	DT LR	C-statistics
K. K. Nishiyama[30]	2013	Canada	questionnaire collection	women	73	multiple	116	internal	SVM	ROC AUC Sensitivity Specificity Accuracy
Kruse, C[31]	2017	Denmark	administrative	subjects	60.8	hip	7252	internal	DT NB	ROC AUC Sensitivity Specificity
Kolanu, N[32]	2021	Australia	electronic health record	patient	73.4	multiple	5416	external	ANN	ROC AUC Sensitivity Specificity
Kim, H. Y[33]	2016	Korea	administrative	subjects	60	multiple	718508	internal	SM	C-statistics
Hsieh, C. I[34]	2021	China	clinical hospital	patient	72.2	hip	36279	external	Other DL	AUC ROC Sensitivity Specificity Accuracy PPV NPV
Hong, N[35]	2021	Korea	clinical hospital	women	73	hip	2462	internal	SM	C-statistics

Ho-Le, T. P[36]	2017	Australia	osteoporosis registry	women	69.1	hip	1167	external	BT SVM ANN LR kNN SVM	AUC Sensitivity Specificity
Henry, M. J[37]	2011	Australia	osteoporosis registry	women	74	multiple	600	-	LR	AUC ROC Sensitivity Specificity
Galassi, A[38]	2020	Spain	electronic health record	women	81.4	hip	137	internal	DT LR RF SVM	Sensitivity Specificity Accuracy
FitzGerald, G[39]	2014	California	questionnaire collection	women	67	multiple	47429	-	SM	C-statistics
Ferizi, U[40]	2019	USA	osteoporosis registry	women	62	multiple	92	-	LR BT kNN SVM NB	AUC ROC Sensitivity Specificity
Enns-Bray, W. S[41]	2019	USA	clinical hospital	women	77.2	hip	254	-	LR	AUC ROC
Engels, A[42]	2020	Germany	administrative	patient	75.6	hip	78074	internal	SVM RF LR Ensemble learning	AUC ROC
De Vries, B. C. S[43]	2021	The Netherlands	clinical hospital	patient	68	multiple	9348	internal	BT ANN RF SM	C-statistics
Cheung, E. Y[44]	2012	China	electronic health record	women	62	multiple	2266	-	SM	AUC ROC Sensitivity Specificity
Chanplakorn, P[45]	2021	Thailand	osteoporosis registry	women	68.5	vertebral	617	-	SM	AUC ROC
Bredbenner, T.	2014	USA	clinical hospital	men	65	hip	922	internal	LR	AUC

L[46]										ROC
Beyaz, S[47]	2020	Turkey	osteoporosis registry	patient	74.9	multiple	2106	-	ANN	AUC ROC Sensitivity Specificity Accuracy
Berry, S. D[48]	2018	USA	administrative	subjects	84	hip	1278304	external	SM	C-statistics
Beaudoin, C[49]	2021	Canada	administrative	subjects	75.1	multiple	581281	internal	SM	C-statistics
Baleanu, F[50]	2022	Belgium	clinical hospital	women	70.1	multiple	3560	-	LR	AUC ROC
Almog, Y. A[51]	2020	USA	electronic health record	patient	50	vertebral	9806205	internal	ANN	AUC ROC Sensitivity Specificity
Zagorski, P[52]	2021	Poland	questionnaire collection	women	65.2	hip	389	-	LR	AUC ROC Sensitivity Specificity PPV NPV
Diez-Perez, A[53]	2007	Spain	questionnaire collection	women	72.3	multiple	5201	-	SM	AUC ROC
Lix, L. M[54]	2018	Canada	osteoporosis registry	women	65.6	multiple	31999	-	LR	AUC ROC
Li, Q. J[55]	2021	China	clinical hospital	patient	70	multiple	562	internal and external	LR	C-statistics
Lee, S[56]	2008	Korea	osteoporosis registry	women	65	multiple	94	-	SVM	Sensitivity Specificity
Jacobs, J. W. G[57]	2010	Portugal	questionnaire collection	subjects	66	vertebral	314	-	LR	AUC ROC Sensitivity Specificity
Eller-Vainicher, C[58]	2011	Italy	questionnaire collection	women	68	vertebral	372	-	ANN LR	AUC ROC Sensitivity Specificity

Zhong, B. Y[59]	2017	China	clinical hospital	patient	72	vertebral	421	internal	SM	Accuracy C-statistics
Xiao, X[60]	2021	USA	gene database	women	64.5	hip	699	-	SM	AUC
Du,J[61]	2022	China	clinical hospital	subjects	71	femur	120	-	SVM RF DT AdaBoost ANN XGBoost	Accuracy Specificity Recall Precision
Wang,M[62]	2022	China	clinical hospital	subjects	73.4	vertebral	7906	-	SM	AUC
Dong,Q[63]	2022	USA	clinical hospital	men	73.7	vertebral	3792	internal	Other DL	AUC ROC Sensitivity Specificity PPV NPV FDR F1 score Accuracy
Wen,Z[64]	2022	China	clinical hospital	patient	73.5	vertebral	270	internal	LR	AUC ROC Specificity Sensitivity PPV NPV Diagnostic efficiency
Pluskiewicz,W [65]	2023	Poland	questionnaire collection	women	66.4	multiple	640	-	LR	AUC
Kong,X[66]	2022	China	clinical hospital	patient	55.1	multiple	1730	-	SM	AUC NRI IDI
Agarwal,A[67]	2023	Canada	electronic health record	women	70.7	multiple	9716	external	SM	AUC ROC

*LR: Logistic Regression;ANN:artificial neural network;SVM = support-vector machine; CNN:convolutional neural network; kNN: k-nearest neighbors; RF: random forests;DT:decision tree;NB:Naive Bayes;BT:Boosted tree;SM:Survival model;DL:deep learnimg model;AUC:area under the receiver operating

characteristic curve;ROC:receiver operating characteristic;PPV:positive predictive value;NPV:negative predictive value;FDR:R false discovery rate;NRI:net reclassifcation index; IDI:integrated discrimination improvement.

Table S4 Methodological characteristics of machine learning models developed for outcome prediction in patients with Osteoporosis

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Wu, Q	2020	Train	M	Multiple fractures	361	4104	LR	10-fold cross validation	Median interpolation				
Wu, Q	2020	Train	M	Multiple fractures	361	4104	RF	10-fold cross validation	Median interpolation				
Wu, Q	2020	Train	M	Multiple fractures	361	4104	BT	10-fold cross validation	Median interpolation				
Wu, Q	2020	Train	M	Multiple fractures	361	4104	ANN	10-fold cross validation	Median interpolation				
Wu, Q	2020	Test	M	Multiple fractures	90	1026	LR	10-fold cross validation	Median interpolation	0.6410	0.7610	0.4420	0.6980
Wu, Q	2020	Test	M	Multiple fractures	90	1026	RF	10-fold cross validation	Median interpolation	0.7005	0.7000	0.4670	0.7590
Wu, Q	2020	Test	M	Multiple fractures	90	1026	BT	10-fold cross validation	Median interpolation	0.7100	0.5650	0.6930	0.8840
Wu, Q	2020	Test	M	Multiple fractures	90	1026	ANN	10-fold cross validation	Median interpolation	0.6910	0.7120	0.5980	0.8390
Villamor, E	2020	Train	F	Hip fracture	65	101	LR	10-fold cross validation					0.7669
Villamor, E	2020	Train	F	Hip fracture	65	101	SVM	10-fold cross validation					0.7569

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Villamor, E	2020	Train	F	Hip fracture	65	101	ANN	10-fold cross validation					0.7642
Villamor, E	2020	Train	F	Hip fracture	65	101	RF	10-fold cross validation					0.6940
Villamor, E	2020	Test	F	Hip fracture	65	101	LR	10-fold cross validation					0.7309
Villamor, E	2020	Test	F	Hip fracture	65	101	SVM	10-fold cross validation					0.7835
Villamor, E	2020	Test	F	Hip fracture	65	101	ANN	10-fold cross validation					0.6940
Villamor, E	2020	Test	F	Hip fracture	65	101	RF	10-fold cross validation					0.7334
van Geel, Tacm	2011	Train	F	Vertebral fracture	382	2372	SM	Bootstrapping					
Ulivieri, F. M	2021	Train	F	Vertebral fracture	56	90	ANN			0.8300	0.7500	0.8372	
Yoda, T	2021	Train	M+F	Vertebral fracture	28	50	CNN	5-fold cross validation		0.9670	0.9250	0.9490	0.9380
Yoda, T	2021	Test	M+F	Vertebral fracture	21	47	CNN	5-fold cross validation		0.9840	0.9810	0.9490	0.9640
Jiang, X. Z	2013	Train	F	Multiple fractures	15	615	LR			0.7600	0.8100	0.4700	0.5100
Schousboe, J. T	2014	Train	F	Vertebral fracture	2883	7233	LR			0.6790			

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Sandhu, S. K	2010	Train	F	Multiple fractures	47	144	LR			0.8400	0.7800	0.8000	
Sandhu, S. K	2010	Train	M	Multiple fractures	18	56	LR			0.7600	0.7400	0.8000	
Rubin, K. H	2018	Train	F	Multiple fractures	11898	647103	LR			0.7500	0.7520	0.5650	
Rubin, K. H	2018	Train	M	Multiple fractures	11851	647103	LR			0.7520	0.6450	0.6090	
Rubin, K. H	2018	Test	F	Multiple fractures	4762	600567	LR			0.8740	0.6000	0.6990	
Rubin, K. H	2018	Test	M	Multiple fractures	4776	600566	LR			0.8510	0.6300	0.5840	
Pluskiewicz, W	2010	Train	F	Hip fracture	1599	2012	LR			0.850	0.7590	0.7370	
Pluskiewicz, W	2010	Train	F	Multiple fractures	1704	2012	LR			0.8790	0.7390	0.5980	
Jang, E. J	2016	Train	M	Multiple fractures	36	363	LR			0.7390			
Jang, E. J	2016	Train	F	Multiple fractures	50	405	LR			0.7180			
Barret A. Monchka	2021	Train	M+F	Vertebral fracture	1470	8920	CNN			0.9500	0.8240	0.9430	0.9230
Mehta, S. D	2020	Train	M+F	Vertebral fracture	86	246	SVM	10-fold cross validation		0.9258	0.8950	0.9560	0.9350
Mehta, S. D	2020	Test	M+F	Vertebral fracture	22	61	SVM	10-fold cross validation		0.8963	0.8180	0.9740	0.9180
Langsetmo, L	2011	Test	M	Multiple fractures	139	1606	SM			0.7000			

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Langsetmo, L	2011	Test	F	Multiple fractures	672	4152	SM			0.6900			
Ioannidis, G	2017	Train	M+F	Multiple fractures	3858	22386	DT			0.6690			
Ioannidis, G	2017	Test	M+F	Multiple fractures	1294	7462	DT			0.6870			
K. K. Nishiyama	2013	Train	F	Multiple fractures	44	88	SVM	10-fold cross validation		0.6800	0.5280	0.7970	0.6890
K. K. Nishiyama	2013	Test	F	Multiple fractures	14	28	SVM	10-fold cross validation		0.8000	0.6880	0.8850	0.8100
Kruse, C	2017	Train	F	Hip fracture	293	4722	NB	5-fold cross validation	random forest imputation	0.9200	0.8800	0.8100	
Kruse, C	2017	Train	M	Hip fracture	47	717	DT	5-fold cross validation	random forest imputation	0.8900	1.0000	0.6900	
Kolanu, N	2021	Train	M+F	Multiple fractures	433	5089	ANN				0.9900	0.9950	
Kolanu, N	2021	Test	M+F	Multiple fractures	97	327	ANN				0.6960	0.9500	
Kim, H. Y	2016	Train	M	Multiple fractures	4889	185127	SM			0.6800			
Kim, H. Y	2016	Train	F	Multiple fractures	14951	174126	SM			0.6500			
Kim, H. Y	2016	Test	M+F	Multiple fractures	19915	359255	SM			0.6650			
Hsieh, C. I	2021	Train	M+F	Hip fracture	2254	5164	Other DL	4-fold cross validation		0.9700	0.8820	0.9140	0.9000

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Hsieh, C. I	2021	Test	M+F	Hip fracture	922	2060	Other DL	4-fold cross validation		0.9600	0.8990	0.9200	0.9100
Hsieh, C. I	2021	Train	M+F	Vertebral fracture	530	57662	Other DL	4-fold cross validation		0.9700	0.6960	0.9790	0.9500
Hsieh, C. I	2021	Test	M+F	Vertebral fracture	922	3346	Other DL	4-fold cross validation		0.9400	0.7400	0.9730	0.9480
Hong, N	2021	Train	F	Hip fracture	143	433	RF			0.7840			0.7300
Hong, N	2021	Train	F	Hip fracture	143	433	BT			0.7680			0.7200
Hong, N	2021	Train	F	Hip fracture	143	433	SVM			0.7590			0.7400
Hong, N	2021	Train	F	Hip fracture	143	433	BT			0.7580			0.7300
Hong, N	2021	Test	F	Hip fracture	34	2029	SM			0.8400			
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	ANN	5-fold cross validation			0.8890	0.8610	0.8630
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	LR	5-fold cross validation			0.9070	0.8640	0.8670
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	KNN	5-fold cross validation			1.0000	0.8330	0.8460
Ho-Le, T. P	2017	Train	F	Hip fracture	54	700	SVM	5-fold cross validation			0.9240	0.9690	0.9660
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	ANN	5-fold cross			0.8330	0.8770	0.8730

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
								validation					
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	LR	5-fold cross validation			0.7780	0.8180	0.8150
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	KNN	5-fold cross validation			0.8060	0.7930	0.7940
Ho-Le, T. P	2017	Test	F	Hip fracture	36	467	SVM	5-fold cross validation			0.8060	0.8160	0.8150
Henry, M. J	2011	Train	F	Multiple fractures	125	600	LR			0.7000	0.6420	0.6620	
Galassi, A	2020	Train	F	Hip fracture	62	96	LR				0.7033	0.7146	0.7081
Galassi, A	2020	Train	F	Hip fracture	62	96	SVM				0.9367	0.6292	0.8077
Galassi, A	2020	Train	F	Hip fracture	62	96	DT				0.5967	0.7446	0.6587
Galassi, A	2020	Train	F	Hip fracture	62	96	RF				0.8330	0.9231	0.8710
FitzGerald, G	2014	Train	F	Multiple fractures	2638	47429	SM			0.6670			
Ferizi, U	2019	Train	F	Multiple fractures	32	92	BT	23-fold cross validation		0.6200	0.5880	0.6670	0.6390
Ferizi, U	2019	Train	F	Multiple fractures	32	92	LR	23-fold cross validation		0.6200	0.5600	0.7010	0.6510
Ferizi, U	2019	Train	F	Multiple fractures	32	92	LR	23-fold cross validation		0.6190	0.5400	0.7010	0.6420

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Ferizi, U	2019	Train	F	Multiple fractures	32	92	SVM	23-fold cross validation		0.5910	0.4490	0.7440	0.6410
Ferizi, U	2019	Train	F	Multiple fractures	32	92	kNN	23-fold cross validation		0.5060	0.2690	0.7420	0.5760
Ferizi, U	2019	Train	F	Multiple fractures	32	92	NB	23-fold cross validation		0.5650	0.4520	0.6790	0.6020
Enns-Bray, W. S	2019	Train	F	Hip fracture	95	254	LR			0.7270			
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	LR	10-fold cross validation		0.7140			
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	RF	10-fold cross validation		0.6860			
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	SVM	10-fold cross validation		0.6600			
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	BT	10-fold cross validation		0.7110			
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	Ensemble learning	10-fold cross validation		0.7220	1.0000		
Engels, A	2020	Train	M+F	Hip fracture	6115	20456	BT	10-fold cross validation		0.7250			
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	LR	10-fold cross validation		0.6950	1.0000		

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	RF	10-fold cross validation		0.6850			
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	SVM	10-fold cross validation		0.6500			
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	BT	10-fold cross validation		0.7020			
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	Ensemble learning	10-fold cross validation		0.6980			
Engels, A	2020	Test	M+F	Hip fracture	1529	57618	BT	10-fold cross validation		0.7030			
de Vries, B. C. S	2021	Train	M+F	Multiple fractures	805	7578	SM			0.6970			
de Vries, B. C. S	2021	Train	M+F	Multiple fractures	805	7578	ANN			0.6700			
de Vries, B. C. S	2021	Train	M+F	Multiple fractures	805	7578	RF			0.6870			
de Vries, B. C. S	2021	Test	M+F	Multiple fractures	165	1770	SM			0.6250			
de Vries, B. C. S	2021	Test	M+F	Multiple fractures	165	1770	ANN			0.5880			
de Vries, B. C. S	2021	Test	M+F	Multiple fractures	165	1770	RF			0.5930			
Cheung, E. Y	2012	Train	F	Multiple fractures	106	2266	SM			0.7300	0.8080	0.5170	
Chanplakorn, P	2021	Train	F	Vertebral fracture	179	617	LR			0.6500	0.4300	0.8600	

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Bredbenner, T. L	2014	Train	M	Hip fracture	45	472	LR	10-fold cross validation		0.9300			
Beyaz, S	2020	Train	M+F	Multiple fractures	235	2106	CNN	5-fold cross validation			0.8250	0.6930	0.7770
Berry, S. D	2018	Train	M	Hip fracture	3541	119874	SM			0.6922			
Berry, S. D	2018	Train	F	Hip fracture	11012	299794	SM			0.7106			
Berry, S. D	2018	Test	M+F	Hip fracture	28050	858636	SM			0.6800			
Beaudoin, C	2021	Train	M+F	Multiple fractures	57678	307909	SM			0.6810			
Beaudoin, C	2021	Test	M+F	Multiple fractures	21809	273372	SM			0.6790			
Baleanu, F	2022	Train	F	Multiple fractures	410	3560	LR			0.7300			
Almog, Y. A	2020	Train	M+F	Vertebral fracture	2468694	6329986	ANN			0.8120	0.8120		0.1920
Almog, Y. A	2020	Test	M+F	Vertebral fracture	295479	3476219	ANN			0.6680	0.7070		0.1140
Zagorski, P	2021	Train	F	Hip fracture	49	389	LR			0.8840	0.9390	0.7120	
Diez-Perez, A	2007	Train	F	Multiple fractures	363	5201	SM			0.6720			
Lix, L. M	2018	Train	F	Multiple fractures	749	31999	LR			0.7060			
Li, Q. J	2021	Train	F	Multiple fractures	49	403	LR			0.8820			
Li, Q. J	2021	Test	F	Multiple fractures	17	159	LR			0.8690			

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Lee, S	2008	Train	F	Multiple fractures	47	94	SVM				0.8500	0.4900	
Jacobs, J. W. G	2010	Train	M	Vertebral fracture	58	109	LR			0.5100			
Jacobs, J. W. G	2010	Train	F	Vertebral fracture	98	205	LR			0.7400	0.6700	0.7100	
Eller-Vainicher, C	2011	Train	F	Vertebral fracture	33	372	LR			0.8230	0.3730	0.9030	0.6380
Eller-Vainicher, C	2011	Train	F	Vertebral fracture	33	372	ANN			0.6990	0.7480	0.8780	0.8130
Zhong, B. Y	2017	Train	M+F	Vertebral fracture	33	256	SM			0.7800			
Zhong, B. Y	2017	Test	M+F	Vertebral fracture	23	165	SM			0.7200			
Xiao, X	2021	Train	F	Hip fracture	25	699	SM			0.8040			
Du,J	2022	Train	M+F	Femur fracture		96	SVM					0.6250	
Du,J	2022	Train	M+F	Femur fracture		96	RF					0.5000	
Du,J	2022	Train	M+F	Femur fracture		96	DT					0.5833	
Du,J	2022	Train	M+F	Femur fracture		96	Boosted tree					0.5000	
Du,J	2022	Train	M+F	Femur fracture		96	ANN					0.5833	
Du,J	2022	Train	M+F	Femur fracture		96	Boosted tree					0.5417	
Du,J	2022	Test	M+F	Femur fracture		24	SVM						0.9167

Author	Year	Data set	Gender	Fracture site	Events	Sample size	Model type	Verification method	Missing value processing	C-index	Sensitivity	Specificity	Accuracy
Du,J	2022	Test	M+F	Femur fracture		24	RF						0.8333
Du,J	2022	Test	M+F	Femur fracture		24	DT						0.9167
Du,J	2022	Test	M+F	Femur fracture		24	Boosted tree						0.8750
Du,J	2022	Test	M+F	Femur fracture		24	ANN						0.9583
Du,J	2022	Test	M+F	Femur fracture		24	Boosted tree						0.9167
Wang,M	2022	Train	M+F	Vertebral fractur	72	7906	SM	10-fold cross validation		0.820			
Dong,Q	2022	Train	M+F	Vertebral fractur		3413	Other DL			0.990	0.5980	0.9990	0.9950
Dong,Q	2022	Test	M+F	Vertebral fractur		379	Other DL			0.820	0.9770	0.9510	0.9510
Wen,Z	2022	Train	M+F	Vertebral fractur	208	220	LR			0.854	0.7310	0.8460	
Wen,Z	2022	Test	M+F	Vertebral fractur	50	50	LR			0.979	0.8942	0.9545	
Pluskiewicz,W	2023	Train	F	Multiple fractures	129	640	LR			0.660			
Kong,X	2022	Train	M+F	Multiple fractures	109	1730	SM	Bootstrapping	Mean interpolation	0.803			
Agarwal,A	2023	Test	F	Multiple fractures	264	9716	SM			0.710			

*M:Male;F:Female;LR:Logistic Regression;ANN:artifificial neural network;SVM:support-vector machine; CNN:convolutional ANN; kNN:k-nearest neighbors; RF:random forests;DT:decision tree;BT:Boosted tree;SM:Survival model;NB:Naive Bayes;DL:deep learnimg model.

.

Table S5 Risk of bias assessment grading of the machine learning predictive modelling studies of osteoporosis populations as per the PROBAST criteria

Study	Participants bias	Predictors bias	Outcome bias	Analysis bias	Overall bias rating
Wu, Q	low	low	low	high	high
Wu, Q	low	low	low	high	high
Wu, Q	low	low	low	high	high
Wu, Q	low	low	low	high	high
Villamor, E	high	unclear	unclear	high	high
Villamor, E	high	unclear	unclear	high	high
Villamor, E	high	unclear	unclear	high	high
Villamor, E	high	low	unclear	high	high
Van Geel, Tacm	low	low	low	high	high
Ulivieri, F. M	low	low	low	high	high
Yoda, T	low	low	low	high	high
Jiang, X. Z	low	low	low	high	high
Schousboe, J. T	high	low	low	unclear	high
Sandhu, S. K	high	unclear	unclear	high	high
Rubin, K. H	low	low	low	unclear	unclear
Pluskiewicz, W	high	low	low	unclear	high
Jang, E. J	low	low	low	high	high
Barret A. Monchka	high	low	low	unclear	high
Mehta, S. D	high	unclear	unclear	high	high
Langsetmo, L	low	low	low	unclear	unclear
Ioannidis, G	high	low	low	unclear	high
K. K. Nishiyama	low	low	low	high	high

Kruse, C	low	low	low	unclear	unclear
Kruse, C	low	low	low	unclear	unclear
Kolanu, N	high	low	low	unclear	high
Kim, H. Y	low	low	low	unclear	unclear
Hsieh, C. I	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Hong, N	low	low	low	unclear	unclear
Ho-Le, T. P	low	low	low	high	high
Ho-Le, T. P	low	low	low	high	high
Ho-Le, T. P	low	low	low	high	high
Ho-Le, T. P	low	low	low	high	high
Henry, M. J	low	low	low	unclear	unclear
Galassi, A	low	low	low	high	high
Galassi, A	low	low	low	high	high
Galassi, A	low	low	low	high	high
Galassi, A	low	low	low	high	high
FitzGerald, G	low	low	low	unclear	unclear
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high
Ferizi, U	high	unclear	unclear	high	high

Enns-Bray, W. S	high	low	low	high	high
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
Engels, A	low	low	low	unclear	unclear
de Vries, B. C. S	high	low	low	unclear	high
de Vries, B. C. S	high	low	low	unclear	high
de Vries, B. C. S	high	low	low	unclear	high
Cheung, E. Y	low	low	low	unclear	unclear
Chanplakorn, P	high	low	low	unclear	high
Bredbenner, T. L	high	unclear	unclear	high	high
Beyaz, S	high	low	low	unclear	high
Berry, S. D	low	low	low	unclear	unclear
Beaudoin, C	high	low	low	unclear	high
Baleanu, F	low	low	low	unclear	unclear
Baleanu, F	low	low	low	unclear	unclear
Almog, Y. A	high	low	low	unclear	unclear
Zagorski, P	low	low	low	high	high
Diez-Perez, A	low	low	low	unclear	unclear
Lix, L. M	low	low	low	unclear	unclear
Li, Q. J	high	low	low	high	high
Lee, S	high	unclear	low	high	high
Jacobs, J. W. G	low	low	low	unclear	unclear
Eller-Vainicher, C	low	low	low	high	high

Eller-Vainicher, C	low	low	low	high	high
Zhong, B. Y	high	low	low	high	high
Xiao, X	low	low	low	high	high
Du,J	low	low	low	high	high
Du,J	low	low	low	high	high
Du,J	low	low	low	high	high
Du,J	low	low	low	high	high
Du,J	low	low	low	high	high
Du,J	low	low	low	high	high
Wang,M	high	low	low	high	high
Dong,Q	low	low	low	unclear	unclear
Wen,Z	high	low	low	high	high
Pluskiewicz,W	low	low	low	high	high
Kong,X	high	low	low	high	high
Agarwal,A	low	low	low	unclear	unclear

*When a single study included multiple models, risk of bias concerns were assessed for each model.

Fig.S1 Forest plots for subgroup analysis of C-index statistics by fracture site and machine learning type in training set

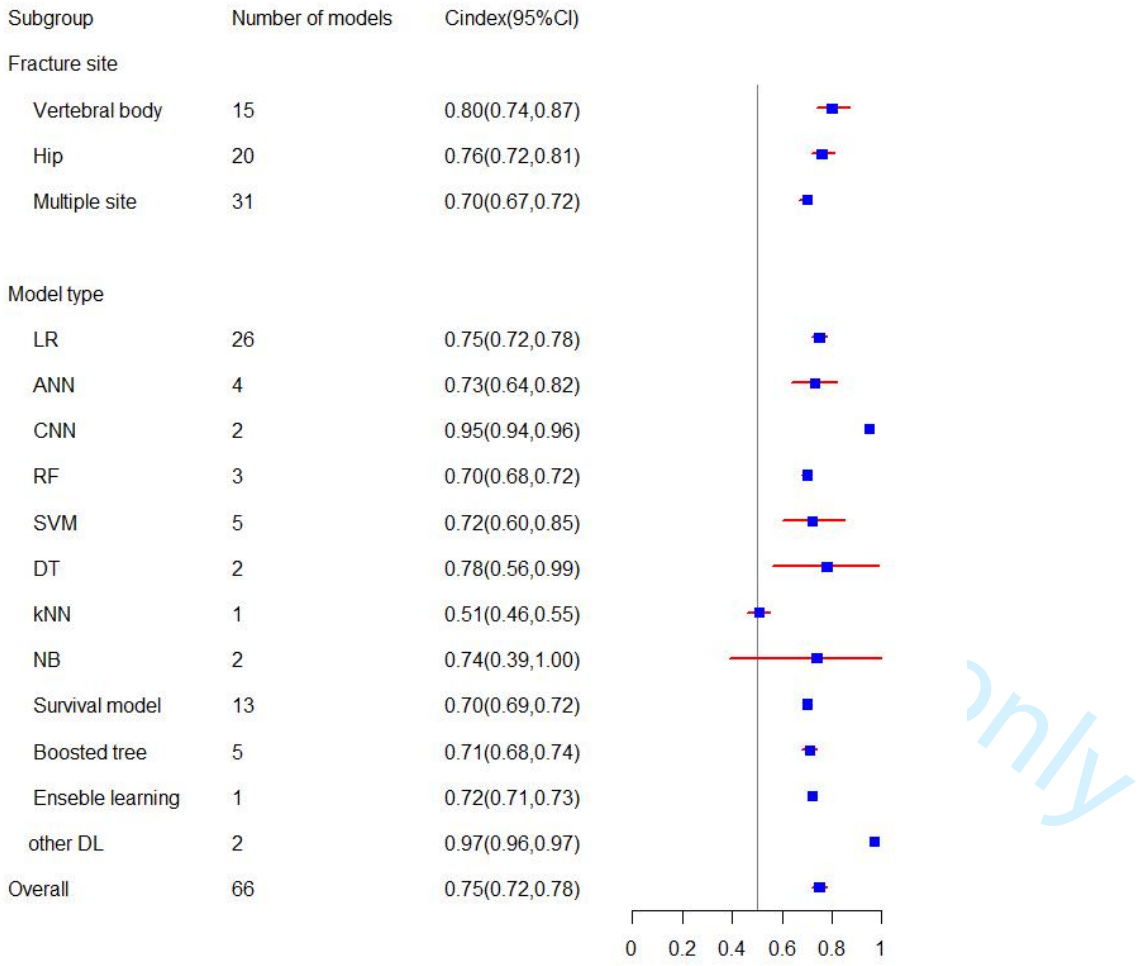


Fig.S2 Forest plots for subgroup analysis of C-index statistics by fracture site and machine learning type in validation set

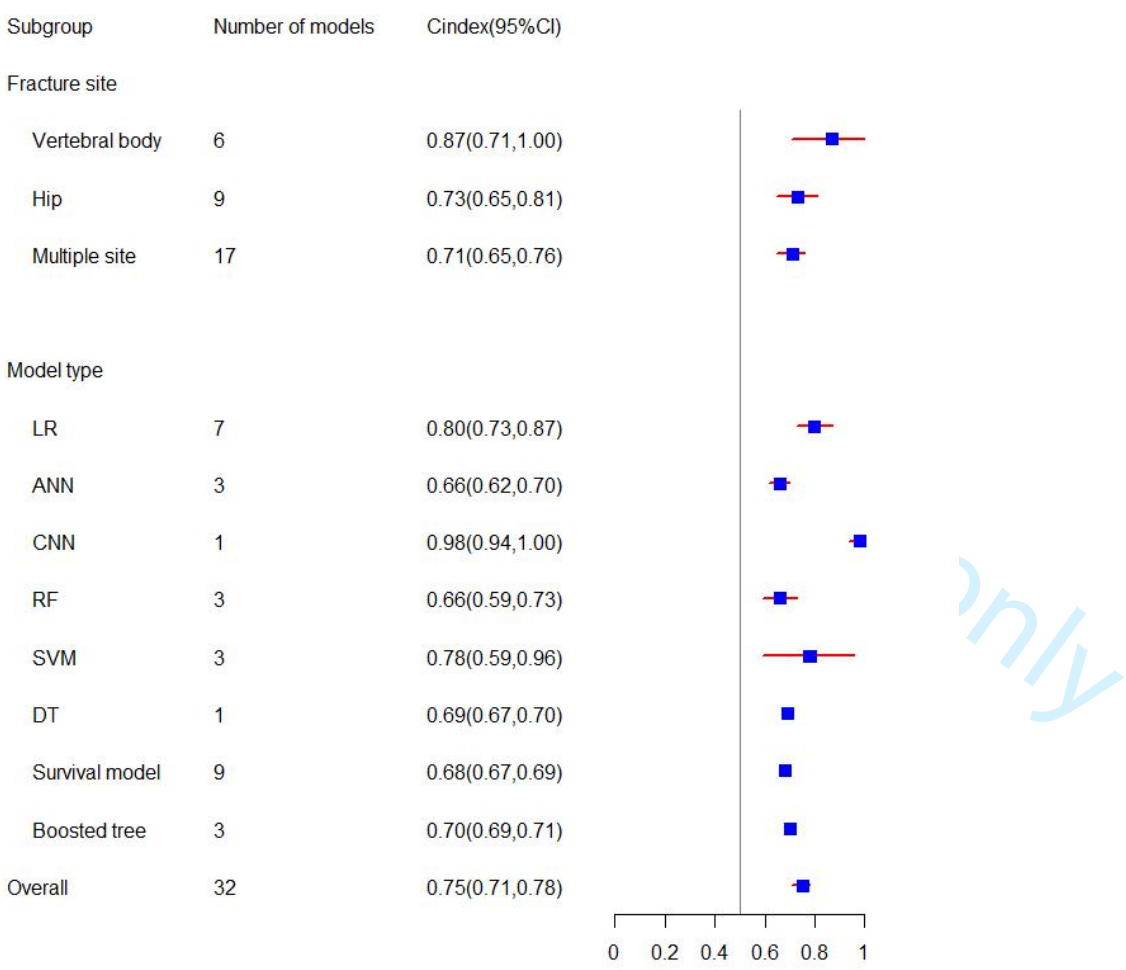


Fig.S3 Sensitivity analysis of multiple fracture model in training set

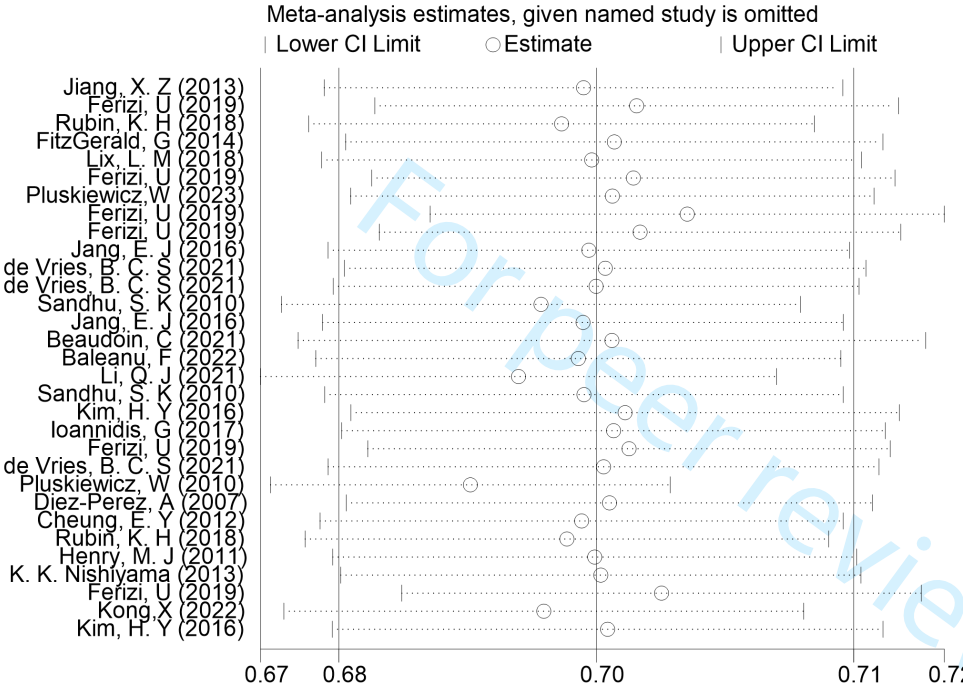


Fig.S4 Sensitivity analysis of vertebral fracture model in training set

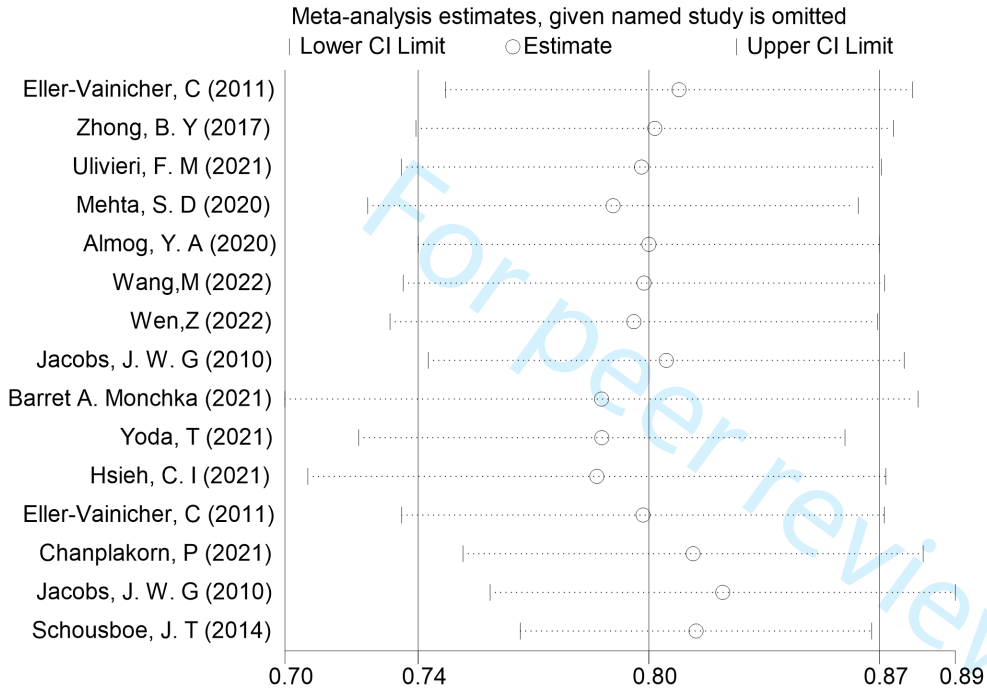


Fig.S5 Sensitivity analysis of hip fracture model in training set

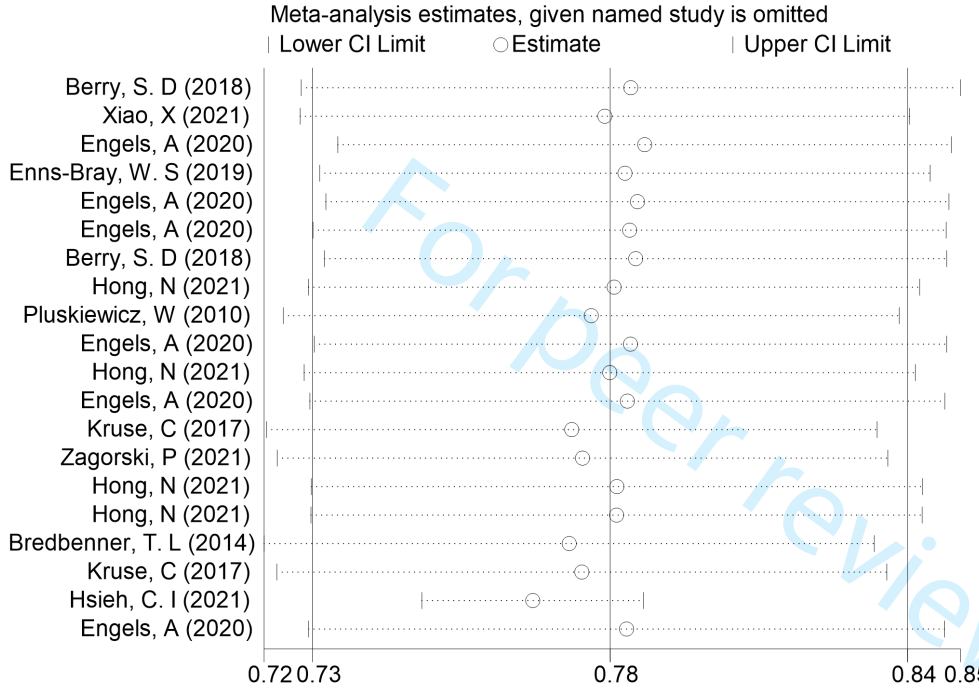


Fig.S6 Sensitivity analysis of multiple fracture model in validation set

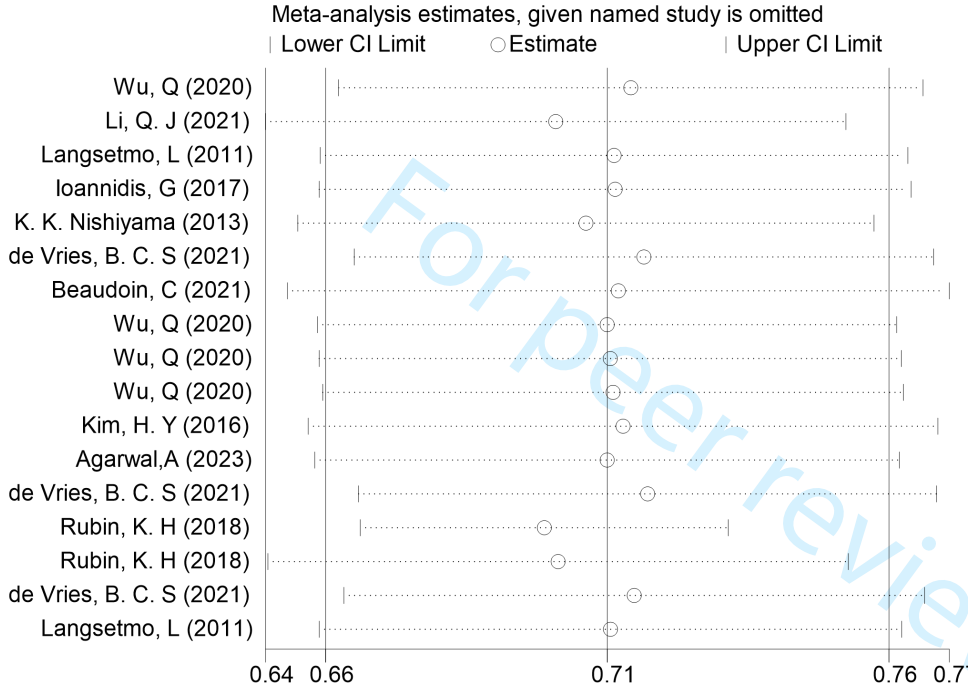


Fig.S7 Sensitivity analysis of vertebral fracture model in validation set

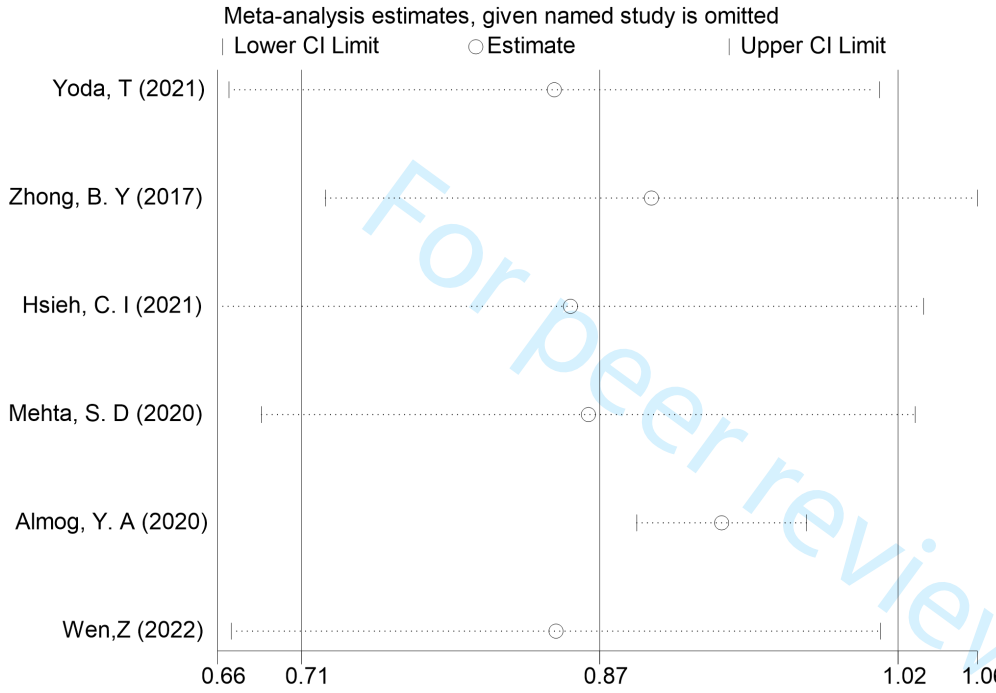


Fig.S8 Sensitivity analysis of hip fracture model in validation set