Osteoporosis International (2025) 36:1535–1544
https://doi.org/10.1007/s00198-025-07642-7

REVIEW

Predictive capacity of fracture risk assessment tools: overview of systematic reviews

Griselda-Adriana Cruz-Priego[1,2,3] · Berenice Araiza-Nava[1,2] · Lucía Méndez-Sánchez[1,2,3] · Delfino Vargas-Chanes[4] · Patricia Clark[1,2,3]

Received: 4 July 2024 / Accepted: 17 July 2025 / Published online: 8 September 2025

Abstract

Purpose Conduct an overview of systematic reviews of the current fracture risk prediction tools in use.

Material and Methods We included systematic reviews (SRs) that assessed the predictive ability of any tool, score, algorithm, or other instrument for fracture risk. The primary outcome measure was the area under the curve (AUC) representing predicted fracture risk within a specified timeframe obtained from receiver operating characteristic (ROC) analysis. We included SRs that studied both men and women with fractures in the general adult population.

Results The review identified 26 different tools currently in use to predict fracture risk. Within these tools a total, 21,717 different prediction variables were found. Among the different tools, a different number of factors were used ranging from the BWC model that used a single predictor variable to the GSOS tool that incorporated 21,717 predictor variables in its model (including many individual SNPs).

Regarding the performance of the tools, AUC ranging from 0.58 to 0.90. None of the models had a prediction capacity greater than 90%. Most of the models are within the range of 0.7 and 0.75, but it cannot be said that any specific one stands out over the others. Rather, a fluctuating behavior is observed in all models within the different studies.

The discrimination of the two most frequently validated models, including FRAX with and without BMD, varied among the studies with AUC/C index ranging from 0.58 to 0.90, respectively. Other commonly validated model, including the Garvan Model showed AUC between 0.57 to 0.84.

Conclusions The vast majority of the models performance is within the range of 0.7 and 0.75. To compare the performance of different tools when predicting fracture, it is very important to consider the differences between prediction tools, the number of risk factors considered, as well as the nature of the variables as they will have an important impact on the feasibility of its use in clinical practice. Likewise, differences in the prediction results may depend on sex, age, types of fractures, as well as the temporal intervals of the prediction and could affect the use of the tools in the daily clinical routine.

Keywords Bone mineral density · Fracture risk assessment tools · Osteoporosis

✉ Patricia Clark
clark@unam.mx

Griselda-Adriana Cruz-Priego
adriana.crpg@ciencias.unam.mx

Berenice Araiza-Nava
berenian2507@gmail.com

Lucía Méndez-Sánchez
luciamendezs@gmail.com

Delfino Vargas-Chanes
dvchanes@gmail.com

[1] Clinical Epidemiology Research Unit, Hospital Infantil de Mexico "Federico Gomez", Mexico City, Mexico

[2] Faculty of Medicine of, National Autonomous University of Mexico (Universidad Nacional Autónoma de México), Mexico City, Mexico

[3] Cochrane, UNAM-Mexico Group, Mexico City, Mexico

[4] University Development Studies Program, National Autonomous University of Mexico, Mexico City, Mexico

Introduction

Osteoporosis is a skeletal disease characterized by low bone mineral density (BMD) and deterioration of bone architecture which conduced to reduced bone strength and, consequently, an increased susceptibility to occurrence of fractures [1, 2]. The operational definition of osteoporosis provided by the WHO is a bone mineral density (BMD) 2.5 or more SDs below the average value for young healthy individuals of the same gender and ethnic background (T-score ≤ -2.5) [2]. Osteoporosis is three times more common in women than in men, partly because women have a lower peak bone mass and partly because of the hormonal changes that occur at the menopause [2].

Some reports say that at least one in 5 men and one in 3 women will suffer a fragility fracture [3]. Fractures not only cause temporary or permanent physical disability [4, 5], but can also affect the overall quality of life of those affected [6, 7] and place a significant burden on healthcare systems [7]. Mainly due to the fact that BMD decreases with age, the incidence of fractures increases exponentially until it becomes an alarming public health problem [8].

In patients who have not yet suffered a fracture, BMD measurement by dual energy X-ray absorptiometry (DXA) is commonly used to identify people with osteoporosis or low BMD. However, many studies have shown that BMD measurement alone does not reliably predict whether an individual will experience a fracture [6, 9–11]. Since the pathogenesis of fractures depends on many factors other than low bone mineral density, scientific evidence has explored different risk factors that could be involved in the occurrence of fragility fractures in addition to BMD [12–14].

Joint efforts by different research groups have been given the task of compiling in systematic literature reviews the different fracture risk prediction tools that have been created or adapted from the different fracture risk factors identified in the different primary studies. In these systematic reviews, a significant number of different tools have been reported that differ in the different clinical risk factors considered, the number of variables included in each prediction model, the populations evaluated, the accessibility of the tools to measure BMD, and therefore the predictive capacity that each tool currently has.

Such a panorama and due to the complexity of the problem, a need arises to carry out an overview of systematic reviews where all the existing evidence is collected on the different fracture risk prediction tools and the predictive capacity of all of them that are currently used in clinical practice, since this information can be useful for health professionals that allows them to choose the tool that adapts to their conditions in their clinical environment, population and objective of prediction, but in the same way this review will help to visualize possible points on improvement of these different tools. Therefore, the objective of this review was to conduct a comprehensive analysis of systematic reviews, specifically focusing on fracture prediction tools with external validation and summarize the evidence on these tools in their predictive capacity.

Methods

Design and registration

The study design was established as a systematic review of reviews and followed the methodology proposed by the Cochrane Collaboration [15] and the Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA) [16]. The detailed protocol of the present study has been previously published (Priego, G. A. C. 2022, August 31). Predictive capacity of fracture risk assessment tools: Overview of Systematic Reviews. https://doi.org/https://doi.org/10.17605/OSF.IO/7SK2M).

Selection criteria

The criteria for inclusion were as follows: Systematic reviews of observational studies either prospective or retrospective cohorts, as well as case–control studies. All articles were identified as meta-analysis or systematic review in the title or abstract. Each of these reviews aimed to assess the predictive ability of validated tools for fracture risk prediction.

We included systematic reviews that assessed the predictive ability of any fracture risk prediction tool, score, algorithm, or any other instrument to predict fracture risk (with or without BMD measurement). SRs that included primary studies that have studied the observed occurrence of the event of interest (fractures) were taken into account. The fractures have to be confirmed by reports, or medical records.

The primary outcome measure was the area under the curve [17] of the predicted fracture risk and its SE, at the specified time interval that were obtained from the receiver operating characteristic (ROC) analysis. We included systematic reviews that had studied both men and women with fractures in the general adult population.

In the same way, we excluded reviews that include primary studies based on animal models, also the SRs which have included tools lacking external validation, as well as investigations evaluating intermediate or surrogate outcomes only. Reviews were not excluded by language or date of publication.

Identification and selection of studies

An independent, paired systematic search for SRs was performed in the following electronic libraries: MEDLINE, Cochrane, Epistemonikos, TripDatabase, PROSPERO, and grey literature (Worldcat, Manchester library search, Health knowledge) without restrictions by date of publication or language. The search terms used were as follows: "Models risk fracture, Prediction risk fracture, Validity tool for risk fracture, Validity tools for risk of bone fracture, Fracture risk assessment tool, Validity of models for risk of bone fractures, Precision of models for risk of bone fractures, Precision of models for risk fracture, Validity tools for prediction of risk fracture, Validity of risk fracture index, Models of risk fracture, Validity risk fracture instrument, Validity risk fracture assessment tool, Validation studies fracture risk assessment tool, Risk fracture assessment tool, Osteoporosis risk assessment instrument, Accuracy and precision risk fracture models, Accuracy and precision risk fracture tool, Risk fracture scale, Osteoporosis risk fracture models, Prediction models risk fracture, Mathematical models risk fracture prediction, Risk fracture index" (Supplementary Table 1).

SRs as a filter was used when available in electronic libraries, and after the general search. The complete search strategy and its adaptations in the different libraries are found in supplementary material I. Additionally, a manual search was carried out in the list of references of the reviews found in the primary search.

Two reviewers assessed the secondary studies found (G-A C-P and B A-N); first by title and abstract to exclude articles that were not systematic reviews or meta-analyses related to the objective of the review. Later they were reviewed in full text.

The selection criteria were applied independently by both reviewers while the full texts were being analyzed. Conflicts were resolved by a third reviewer (L MS).

Overlap assessment

A citation matrix was constructed with the primary studies of the included systematic reviews and the overlap was calculated with the "Corrected Covered Area" theorem described by Pieper et al. [18].

Data collection

Data was extracted from each review included in a Microsoft Excel® spreadsheet in a previously defined form. One review author (G-A C-P) extracted the data and the second carried out the accuracy and completeness check of the abstracted data (B A-N).

Information regarding the objective, population (type of primary studies), intervention, comparator, outcome, and participants (type of patients in the primary studies) was extracted from each of the included reviews. We analyzed the methodology used in each of the SRs (Electronic Search and evaluation of the risk of bias of the primary studies). However, for the present overview, we did not review any of the primary studies as stated and detailed in the Cochrane handbook for systematic reviews to not repeat a review of the original trials.

Methodological quality assessment

Two review authors independently assessed the methodological quality of the included reviews using the "assessment of risk of bias in systematic reviews (ROBIS)" tool. This tool has three phases; to assess relevance, identify concerns in the review process, and judge risk of bias. The second of these phases is divided into four domains: eligibility criteria, identification and selection, data collection and evaluation of studies, as well as synthesis and findings. In each domain, the information used to support the judgment, the signaling questions, and the judgment about concern about the risk of bias are studied.

Synthesis strategy

Due to the heterogeneity of the data, a meta-analysis of these SRs was not acceptable; therefore, it was decided to carry out only a qualitative synthesis of the information.

Results

The search carried out in the digital libraries as mentioned in methods, independently by two authors of this review; up to October 2023, it yielded a total of 20,583 systematic reviews, of which 3093 were in PUBMED, 193 in the Cochrane Library, 1429 in Epistemonikos, 5940 in TripDatabase, 13 records in PROSPERO, and the rest of them were gray literature. Seventy-five SRs related by title were found. When analyzing by abstract, 10 were excluded. In this way, a total of 6 systematic reviews were of potential interest according to the pre-established selection criteria, to be analyzed in full text [19–24]. After this process, a total of six systematic reviews were included in this review of reviews (PRISMA chart) (Fig. 1).

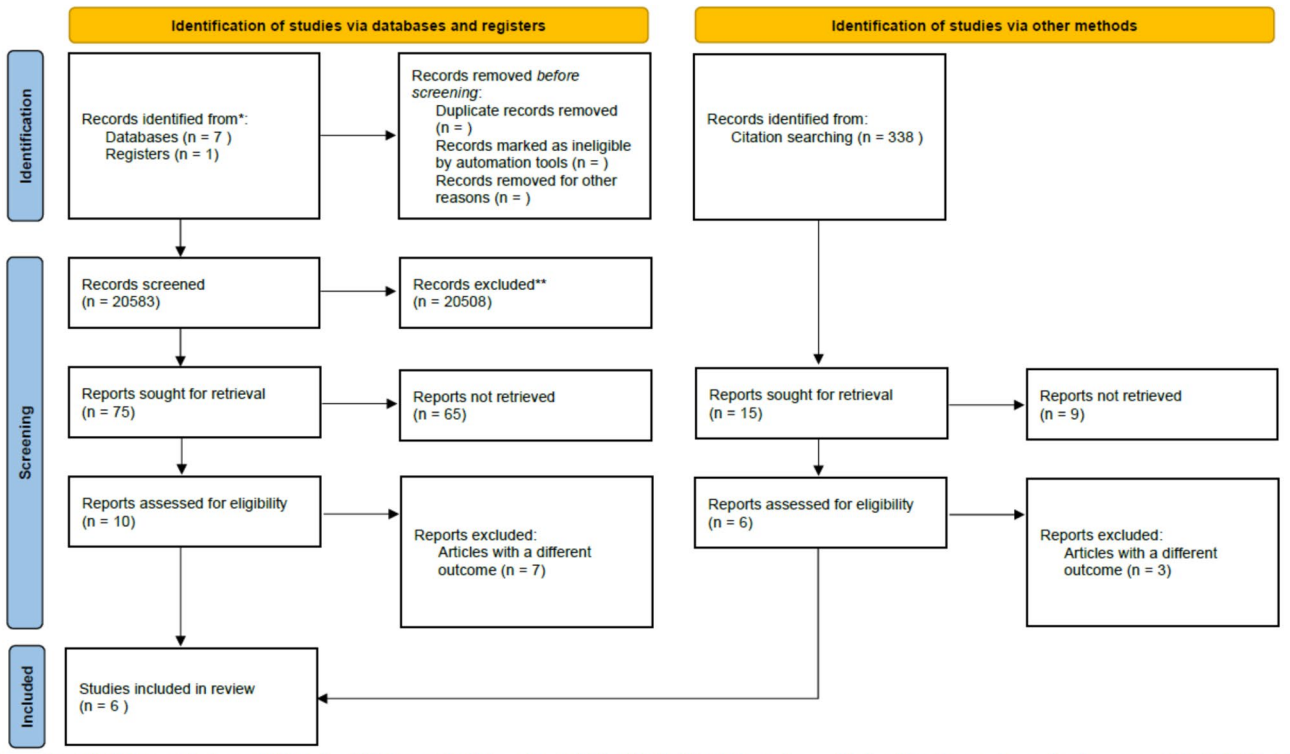

Fig. 1 Flowchart of studies included in the overview

Description of included reviews

The reviews were published between 2013 and 2023. The number of primary studies in the included reviews ranged from 6 to 68, and after the overlap between the SRs was reviewed, 107 primary studies were identified. The characteristics of each primary article are described in Supplementary Table 2. The population size ranged from 390 to 4,726,046. All of the SRs comprised prospective or retrospective observational studies, RCTs, and cross-sectional studies. The age of the participants in most primary studies was over 40 years old; with the exception of the FRAMO tools that only included people between 70 and 100 years old and the QFracture that consider people between 30 and 99 years old. The characteristics of the systematic reviews included are detailed in Table 1.

To facilitate the interpretation of the predictive performance of fracture risk assessment tools, and given the heterogeneity in study methods and populations, we focused our main analysis on tools that had been externally validated in at least two independent studies. We extracted the number of publications reporting each fracture risk prediction tool as identified in the included systematic reviews. This count includes studies with reported use or performance metrics of the tools, regardless of whether they were explicitly labeled as external validations.

Table 2 summarizes the reported publications per tool, along with whether internal and/or external validation was conducted. A total of 26 tools were identified; however, only a subset of them ($n = 11$) had been evaluated in at least two external validation studies, suggesting stronger evidence of reproducibility across different populations. These tools were prioritized in the synthesis of our main results. Tools with limited external validation ($n < 2$) were included in Supplementary Material to ensure completeness without overloading the main findings.

This structured overview highlights the disparity in the degree of validation across available tools. The limited number of models with robust external validation underscores the challenges in generalizing predictive performance across diverse populations and clinical settings.

The objective of these six reviews was to evaluate the predictive capacity of the tools for fractures. From the primary studies that have reported the follow-up time, it was observed that this varied from 2 to 13.4 years. However, only in 25% of all studies the follow-up time was comparable to the prediction time of the tools (the most common of these being the prediction of 10-year fracture risk).

Table 1 Characteristics of the systematic reviews included

Author and year	Title	Population	Test index	Reference method
Beaudoin et al. 2018	Performance of predictive tools to identify individuals at risk of non-traumatic fracture: a systematic review, meta-analysis, and meta-regression	Adults of different ethnicities	FRAX, Garvan, QFracture 2009, QFracture 2012, FRC, WHI, OSIRIS, ORAI, OST, SCORE, FRISC, mSOF, FRA-HS, FRAMO	Fracture observation in follow-up
Marques, A. et al. 2015	The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis	Adults of different ethnicities	"Computer model for osteoporotic fracture risk," FRAX, GARVAN, Qfracture, FRC, FRISC, GARVAN-GRX, Qfracture, QFracture (2012), Simplified fracture risk system, SOF, WHI	Fracture observation in follow-up
Hass Rubin et al. 2013	Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review	Women over 40 years old	ABONE, BWC, EPESE, Ettinger and Colleagues, SOF, FRAMO, FRAX, FRISC, Garvan, OC, ORAI, OST, Qfracture, Qfracture-updated, SCORE, SOF, WHI, Van Staa and Colleagues	Fracture observation in follow-up
Sun X et al. 2022	Prediction models for osteoporotic fractures risk: a systematic review and critical appraisal	Adults of different ethnicities	WHI, Garvan, FRAX, FRAX + TBS, QFracture, Updated QFracture, FRISC, FRISK, FRA-HS, GSOS	Fracture observation in follow-up
Adami, G. et al. 2023	A systematic review on the performance of fracture risk assessment tools: FRAX, DeFRA, FRA-HS	Adults of different ethnicities	FRAX, DeFRA, FRA-HS	Fracture observation in follow-up
Adami, H et al. 2013	Performance of fracture risk assessment tools by race and ethnicity: a systematic review for the ASBMR task force on clinical algorithms for fracture risk	American population over 40 years of age of different races and ethnicities	FRAX and Garvan with and without BMD	Fracture observation in follow-up

Table 2 Summary of reported publications for fracture risk prediction tools

Tool	Internal validation	External validation	Reported publications for each tool (*n*)
Computer model for osteoporotic fracture risk	Yes	Yes	1
FRAMO	Yes	Yes	2
FRAX	Yes	Yes	50
FRC	Yes	Yes	2
FRISC	Yes	Yes	2
FRISK	Yes	Yes	2
GARVAN	Yes	Yes	16
Q-Fracture	Yes	Yes	4
Updated Q-Fracture (2012)	Yes	Yes	2
Score for estimating the long-term risk of fracture in post menopausal women	Yes	Yes	1
Simplified fracture risk system	Yes	Yes	1
SOF	Yes	Yes	3
WHI	Yes	Yes	3
OSIRIS	Yes	Yes	1
ORAI	Yes	Yes	1
OST	Yes	Yes	1
SCORE	Yes	Yes	1
FRAX + TBS	Yes	Yes	5
FRA-HS	Yes	Yes	1
ABONE	Yes	Yes	1
BWC	Yes	Yes	1
EPESE	Yes	Yes	1
Pentosidine + FRISC	Yes	Yes	1
GSOS	Yes	Yes	1
DeFRA	Yes	Yes	1

This table lists the fracture risk assessment tools identified in the included systematic reviews. It shows whether each model underwent internal and/or external validation, along with the number of reported publications per tool. Tools with at least two publications suggesting external validation were prioritized in the main analysis; others are detailed in supplementary materials

Note: The number of publications refers to the studies identified in the included systematic reviews that reported using or validating each tool. In some cases, although the study design suggests external validation (e.g., testing in an independent cohort with performance metrics), it was not explicitly labeled as such by the original systematic review authors

In about three-quarters of the studies, hip fracture was confirmed by medical records or by confirmatory radiography. Of the remaining percentage, most of them were self-reported or no data for this variable was available.

A total of 26 different tools were found and are currently used to predict fracture risk. These are broken down in Supplementary Table 3, as well as the variables considered in each of them. These tools have been developed and validated in different countries around the world: Canada, USA, Denmark, Sweden, Australia, New Zealand, Poland, UK, Ireland, Israel, France, Japan, Spain, China, Italy, Germany, Netherlands, Finland, Thailand, Norway and Portugal; while there were no models developed using data from Africa, Central and South America, and the Middle East.

Regarding the number of predictive risk factors used in these tools, a total of 21,717 different prediction variables were found (Supplementary Table 3). Among the different tools, a different number of risk factors are used, ranging from the model that has a single predictor variable (BWC) to the tool that has the greatest number of predictor risk factors (GSOS) that has 21,717 variables (many of these risk factors are single nucleotide polymorphisms (SNPs) from GWAS).

A majority of models contained similar predictors, such as age (84.61%) and weight (80.0%). Other common variables were previous fractures (73.07%), femoral neck BMD (50%), smoking (42.30%), height (34.61%), sex (34.61%), use of glucocorticoids (30.76%), and rheumatoid arthritis (30.76%). These data are described in Supplementary Table 3.

Most of the tools were developed using logistic regression and Cox proportional hazards models, although for a minority of the primary studies information on the mathematical model was not available. Validation of these tools were done using mainly three methods: cross validation, geographical validation and in some of them, the bootstrapping technique.

Model performance was assessed by discrimination and calibration. Discrimination is often quantified by the area under the receiver operating characteristic curve [17]. AUC less than 0.5 suggests no discrimination, 0.5 to 0.7 is poor, 0.7 to 0.8 is acceptable, 0.8 to 0.9 is excellent, and higher than 0.9 is outstanding [20].

The AUC values ranged from 0.58 to 0.90. Around a quarter of the studies these values were not available. None of the models had a prediction capacity greater than 90%. The vast majority of the models are within the range of 0.7 and 0.75, meaning they had acceptable to excellent performance.

The discrimination of the four most frequently validated models (Table 3), including FRAX with BMD (for MOF and for hip fracture), and FRAX without BMD (for MOF and for hip fracture) varied among the studies, with AUC/C index that ranged from 0.58 to 0.90, respectively. There were some FRAX extension models based on FRAX predictors and other predictors, such as FRAX plus trabecular bone score (TBS). Other commonly validated models, including the Garvan Model 1 and Garvan Model 2 in females, showed AUC between 0.70 and 0.85.

Comparisons of multiple tools within individual cohorts

A limited number of cohort studies included in the systematic reviews evaluated the predictive performance of two or more fracture risk assessment tools within the same population. This approach enables a more robust comparison under uniform conditions. For example, Dagan et al. (2017) [25] assessed FRAX, FRAX + TBS, and QFracture in a large Israeli cohort, reporting notable differences in discrimination (AUCs) between tools despite a shared population framework. Similarly, Holloway-Kew et al. (2019) [26] and Bolland et al. (2011) [27] compared FRAX and Garvan models, highlighting variation in performance based on sex and fracture type. While these comparative analyses offer valuable insights, differences in subpopulation characteristics, model inputs, and statistical handling still pose limitations to drawing definitive conclusions.

Table 3 Predictive capacity of the most used tools

Tool	Range of predictive capacity
FRAX with BMD	0.59 to 0.88
FRAX without BMD	0.58 to 0.90
FRAX plus trabecular bone score (TBS)	0.85
Garvan with BMD (10-year prediction)	0.70 to 0.85
Garvan with BMD (5-year prediction)	0.78 to o.79
QFracture 2009 (10-year prediction)	0.86 to 0.89
QFracture 2012 (5-year prediction)	0.83

Methodological quality of included systematic reviews

For the methodological quality, we assessed the risk of bias in the included SRs using the ROBIS tool. Globally, a 50% risk of unclear bias was observed in the reviews, mostly given by the assessment of synthesis and findings in the reviews as well as the studies eligibility criteria. Two systematic reviews were rated at high risk of bias, which is equivalent to 33.3% of the total of SR's, and only one review was rated at low risk of bias. Only one of the included systematic reviews provided information about any priori register of their protocol. The summary and evaluation of the risk of bias are presented in Fig. 2.

Regarding the domain of the study eligibility criteria, 3 of the SRs had unclear concern and 3 had low concern, mainly for not providing information about having an a priori protocol, and the appropriateness of their restrictions on eligibility criteria based on information electronic databases sources. The identification and selection of the domain of studies presented two studies with high concern and the remaining four with low concern due to poor information on the justification of the restrictions of the search periods, in one case without additional search methods other than databases and the appropriateness of publication, language, and date restrictions.

Two studies were rated high concerns and low concern of six SR's, given some concerns about lack of information about efforts made to minimize errors in data extraction. Finally, for the synthesis and findings domain, three studies were rated as having unclear concerns, two as having high concern, and only one at low risk of bias. The main reasons for these ratings were due to concerns about the appropriateness of the synthesis given the nature and similarity of the research questions, the study designs and results, how heterogeneity was addressed, and the methods (or the absence of them) used to demonstrate the solidity of the results.

Discussion

This report summarizes the evidence related to the performance of fracture risk prediction tools based on six systematic reviews. However, due to the heterogeneity of the

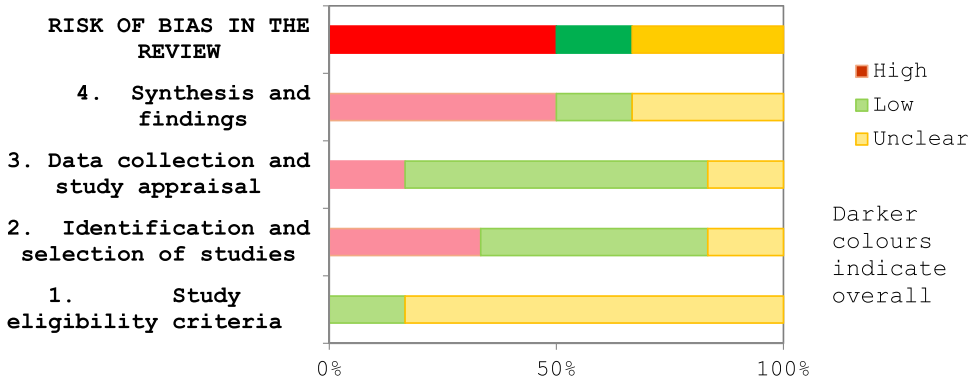

Fig. 2 Summary and graph of risk of bias assessment, assessed with ROBIS tool

populations evaluated across the studies, it was not possible to pool the data for a more in-depth quantitative analysis.

As shown, there are currently at least 26 fracture risk prediction tools that have validation studies: computer model for osteoporotic fracture risk, FRAX, FRAX + TBS, FRAMO, FRC, FRISC, FRISK, GARVAN-GRX, QFracture, Updated QFractured, Score for estimating the long-term risk of fracture in postmenopausal women, Simplified fracture risk system, SOF, WHI, OSIRIS, ORAI, OST, SCORE, FRA-HS, ABONE, BWC, EPESE, Pentosidine + FRISC, GSOS, and DeFRA.

In clinical practice, FRAX, QFracture, and Garvan are among the most frequently used tools. Additionally, some models—such as FRA-HS and WHI—have demonstrated good external performance and hold potential clinical value.

The number of risk factors included in these tools' ranges widely, from just 1 to over 27,000, and their nature significantly affects the feasibility of implementation in routine care. Variations in predictive outcomes depending on sex, age, fracture type, and prediction time horizon may also impact clinical applicability. While all tools estimate the risk of osteoporotic fractures, not all distinguish between hip fractures and major osteoporotic fractures.

The complexity of summarizing predictive capacity in a single range arises from several factors, particularly the characteristics of the populations used in validation studies, while some studies used the general population, some others included specific patients with osteoporosis or at risk of osteoporosis. Unfortunately, in some of the reviews, the eligibility criteria were not completely well defined and, as in the vast majority of reviews, there was no registration of a review protocol.

Another factor complicating the assessment of tool performance is the length of follow-up. As noted in the results, only a small proportion of studies had follow-up periods that matched the tools' prediction horizons. In cases where follow-up and prediction periods did not align, no statistical adjustments were made to correct for this discrepancy—likely affecting reported performance metrics.

No single tool consistently outperforms the others across individual studies. Nevertheless, all models demonstrate variability in performance depending on the specific study context.

Using the area under the receiver operating characteristic curve (AUC) to compare predictive performance presents significant limitations. As previously discussed, the AUC is highly sensitive to underlying population characteristics—such as age distribution, baseline fracture risk, and follow-up duration—making direct cross-study comparisons problematic [28]. Kanis et al. (2012) emphasized that AUC may underestimate clinical utility, especially when differences in calibration and decision thresholds are not considered [29]. Similarly, Halligan et al. (2015) [30] pointed out that reliance on AUC alone can obscure clinically relevant differences in predictive accuracy when applied to heterogeneous populations. These limitations are particularly relevant in this overview, where the included systematic reviews applied varied methodologies and often lacked harmonization of comparator groups. Consequently, while AUC remains a commonly reported metric, it should be interpreted with caution and ideally supplemented with calibration measures, decision curve analysis, or net benefit metrics to provide a more complete assessment of model performance, as Steyerberg et al. (2010) [31] highlight in their proposed framework combining discrimination, calibration, and clinical utility.

The variability observed in the predictive capacity of some tools, such as the FRAX calculator, could be hypothesized to be due to the large spectrum of populations in which they have been evaluated but not validated. Validation of the instrument has only been carried out in Japan, England, Canada, and New Zealand, while in other countries, only calibrations have been performed. This suggests that predictions may need revision to account for differences across racial or ethnic groups. This is particularly important, since it would be necessary to study whether the weight of the different risk factors within the calculator is the same for all types of populations. On the other hand, speaking especially of calibrations, the performance of the tool will depend on the

quality of the epidemiological data that the countries must carry out this process. In this way, if the fracture data in the countries are underdiagnosed, it could lead to the prediction not being as accurate as reality. Therefore, this represents a clear avenue for future investigation.

The performance of prediction algorithms also depends heavily on the validation methodology. When tools are tested in populations similar to those used in their development, predictive performance is often more favorable. However, such results may not generalize to other settings or patient groups.

The broad range of variables included in these tools reflects the wide array of factors that may influence fracture risk prediction and what directly tells us about the complexity of the phenomenon to be modeled. This complexity is due to physical, physiological, and biological factors; this in turn can lead to there not being a total consensus on the most suitable factors for predicting fragility fractures and that are a limitation to being able to completely predict the phenomenon and only remain an approximation.

Of course, clinical researchers are invited to form an international consensus between the different tools that are currently used to evaluate the contribution of the different variables used in prediction, collecting those that are most significant for prediction. Additionally, researchers should aim to provide greater transparency regarding the procedures and mathematical methods used in model development, the score weights applied in the final calculators, and whether predictions from different tools can be effectively used to guide treatment decisions. Only knowledge of all these areas involved in prediction will allow first contact doctors to know which tool to use in their clinical environment.

The relevance of these considerations lies in the clinical implications of the predictive capacity of the tools. For instance, if a model predicts a fracture risk of 52% in a given population, how confident should clinicians be in prescribing first-line anti-osteoporotic treatments versus second-line options? The choice of therapy hinges not only on the magnitude of predicted risk but also on the reliability of the prediction tool in that specific context.

A key strength of this overview lies in the rigorous adherence to the methodology proposed by the Cochrane Collaboration for conducting overviews of systematic reviews, which is widely recognized as the gold standard for evidence synthesis. This ensures a transparent, structured, and reproducible approach. Furthermore, we prioritized tools with external validation in at least two independent studies, enhancing the robustness of our findings. Comparative evaluations of models within shared cohorts also provided more reliable insights into their relative predictive performance. Nonetheless, several limitations must be acknowledged. The considerable heterogeneity across study designs, populations, and prediction horizons precluded a quantitative meta-analysis. Many tools lack validation in regions such as Africa, Latin America, and the Middle East, limiting their generalizability. Additionally, the methodological quality of the included reviews was variable, with a substantial proportion rated as having high or unclear risk of bias. These constraints should be considered when interpreting the results and highlight the need for future multicenter comparative studies with standardized methodologies.

Conclusion

This overview summarizes the current landscape of fracture risk prediction tools, highlighting the variability in their development, validation, and predictive performance. While over two dozen tools are available, only a limited subset have undergone multiple external validations, and even fewer have been directly compared within the same cohort. Comparisons based on AUC alone are inherently limited due to cohort differences and methodological variability, underscoring the need for cautious interpretation. Future research should prioritize head-to-head comparisons of validated tools in diverse populations, using a broader set of performance metrics beyond AUC. Such efforts are essential to guide clinicians in selecting the most appropriate tool for their specific patient populations and clinical settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00198-025-07642-7.

Acknowledgements The author gratefully acknowledges the support of the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), as well as the Posgrado en Ciencias Médicas, Odontológicas y de la Salud de la Universidad Nacional Autónoma de México (UNAM), during the postgraduate studies of Ms. Griselda-Adriana Cruz-Priego (CVU 1004560).

Data availability All the necessary data is presented in the article.

Code availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Conflicts of interest None.

References

1. NIH Consensus Development Panel on Osteoporosis Prevention Diagnosis, and Therapy (2001) Osteoporosis prevention, diagnosis, and therapy. JAMA 285(6):785–95
2. World Health Organization (2003) Prevention and management of osteoporosis: report of a WHO scientific group. (WHO Technical Report Series; no. 921). World Health Organization, Geneva

3. Scientific Advisory Council of Osteoporosis Canada (2011) Osteoporosis Canada 2010 guidelines for the assessment of fracture risk. Can Assoc Radiol J 62(4):243–50
4. Johnell O, Kanis J (2005) Epidemiology of osteoporotic fractures. Osteoporos Int 16(Suppl 2):S3-7
5. Auais MA, Eilayyan O, Mayo NE (2012) Extended exercise rehabilitation after hip fracture improves patients' physical function: a systematic review and meta-analysis. Phys Ther 92(11):1437–1451
6. Woolf AD, Pfleger B (2003) Burden of major musculoskeletal conditions. Bull World Health Organ 81(9):646–56
7. Liang W, Chikritzhs T (2016) The effect of age on fracture risk: a population-based cohort study. J Aging Res 2016:5071438
8. Camal Ruggieri IN, Cícero AM, Issa JPM, Feldman S (2021) Bone fracture healing: perspectives according to molecular basis. J Bone Miner Metab 39:311–31
9. Armas LA, Recker RR (2012) Pathophysiology of osteoporosis: new mechanistic insights. Endocrinol Metab Clin North Am 41(3):475–86
10. Kanis JA (2002) Diagnosis of osteoporosis and assessment of fracture risk. Lancet 359(9321):1929–36
11. Cranney A, Jamal SA, Tsang JF, Josse RG, Leslie WD (2007) Low bone mineral density and fracture burden in postmenopausal women. CMAJ 177(6):575–80
12. Compston J, Cooper A, Cooper C, Gittoes N, Gregson C, Harvey N et al (2017) UK clinical guideline for the prevention and treatment of osteoporosis. Arch Osteoporos 12(1):43
13. Rabar S, Lau R, O'Flynn N, Li L, Barry P (2012) Risk assessment of fragility fractures: summary of NICE guidance. BMJ (Clinical research ed) 345:e3698
14. Papaioannou A, Morin S, Cheung AM, Atkinson S, Brown JP, Feldman S et al (2010) 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. CMAJ 182(17):1864–73
15. Pollock M, Fernandes RM, Becker LA, Pieper D, Hartling L (2020) Chapter V: overviews of reviews. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (eds) Cochrane handbook for systematic reviews of interventions, 2nd ed. John Wiley & Sons, Chichester, pp 149–174
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Moher D et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372(71):1–9
17. Neu CM, Rauch F, Manz F, Schœnau EJ (2001) Modeling of cross-sectional bone size, mass and geometry at the proximal radius: a study of normal bone development using peripheral quantitative computed tomography. Osteoporos Int 12:538–47
18. Pieper D, Antoine SL, Mathes T, Neugebauer EA, Eikermann M (2014) Systematic review finds overlapping reviews were not mentioned in every other overview. J Clin Epidemiol 67(4):368–375
19. Beaudoin C, Moore L, Gagné M, Bessette L, Ste-Marie LG, Brown JP et al (2019) Performance of predictive tools to identify individuals at risk of non-traumatic fracture: a systematic review, meta-analysis, and meta-regression. Osteoporos 30(4):721–740
20. Marques A, Ferreira RJ, Santos E, Loza E, Carmona L, da Silva JA (2015) The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. Ann Rheum Dis 74(11):1958–1967
21. Rubin KH, Friis-Holmberg T, Hermann AP, Abrahamsen B, Brixen K (2013) Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review. J Bone Miner Res 28(8):1701–1717
22. Sun X, Chen Y, Gao Y, Zhang Z, Qin L, Song J et al (2022) Prediction models for osteoporotic fractures risk: a systematic review and critical appraisal. Aging Dis 13(4):1215–1238
23. Adami G, Biffi A, Porcu G, Ronco R, Alvaro R, Bogini R et al (2023) A systematic review on the performance of fracture risk assessment tools: FRAX, DeFRA FRA-HS. J Endocrinol Investig 46(11):2287–2297
24. Fink HA, Butler ME, Claussen AM, Collins ES, Krohn KM, Taylor BC et al (2023) Performance of fracture risk assessment tools by race and ethnicity: a systematic review for the ASBMR task force on clinical algorithms for fracture risk. J Bone Miner Res 38(12):1731–1741
25. Dagan N, Cohen-Stavi C, Leventer-Roberts M, Balicer RD (2017) External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. BMJ (Clinical research ed) 356:i6755
26. Holloway-Kew KL, Zhang Y, Betson AG, Anderson KB, Hans D, Hyde NK et al (2019) How well do the FRAX (Australia) and Garvan calculators predict incident fractures? Data from the Geelong osteoporosis study. Osteoporos Int 30(10):2129–2139
27. Bolland MJ, Siu AT, Mason BH, Horne AM, Ames RW, Grey AB et al (2011) Evaluation of the FRAX and Garvan fracture risk calculators in older women. J Bone Miner Res 26(2):420–427
28. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115(7):928–935
29. Kanis JA, McCloskey EV, Johansson H, Cooper C, Rizzoli R, Reginster JY (2013) European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 24(1):23–57
30. Halligan S, Altman DG, Mallett S (2015) Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol 25(4):932–939
31. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N et al (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21(1):128–138