EXTENDED REPORT

The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis

Andréa Marques,[1,2] Ricardo J O Ferreira,[1,2] Eduardo Santos,[1,2] Estíbaliz Loza,[3] Loreto Carmona,[3] José António Pereira da Silva[1,4]

ABSTRACT

Objectives To identify and synthesise the best available evidence on the accuracy of the currently available tools for predicting fracture risk.

Methods We systematically searched PubMed MEDLINE, Embase and Cochrane databases to 2014. Two reviewers independently selected articles, collected data from studies, and carried out a hand search of the references of the included studies. The Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS) checklist was used, and the primary outcome was the area under the curve (AUC) and 95% CIs, obtained from receiver operating characteristic (ROC) analyses. We excluded tools if they had not been externally validated or were designed for specific disease populations. Random effects meta-analyses were performed with the selected tools.

Results Forty-five studies met inclusion criteria, corresponding to 13 different tools. Only three tools had been tested more than once in a population-based setting: FRAX (26 studies in 9 countries), GARVAN (6 studies in 3 countries) and QFracture (3 studies in the UK, 1 also including Irish participants). Twenty studies with these three tools were included in a total of 17 meta-analyses (for hip or major osteoporotic fractures; men or women; with or without bone mineral density).

Conclusions Most of the 13 tools are feasible in clinical practice. FRAX has the largest number of externally validated and independent studies. The overall accuracy of the different tools is satisfactory (>0.70), with QFracture reaching 0.89 (95% CI 0.88 to 0.89). Significant methodological limitations were observed in many studies, suggesting caution when comparing tools based solely on the AUC.

INTRODUCTION

The major clinical consequence of osteoporosis is the occurrence of fragility fractures.[1] Osteoporotic fractures lead to significant suffering, disability and mortality, resulting in enormous costs for individuals and society.[2] Predicting the absolute risk of osteoporotic fractures is, therefore, of the utmost importance to optimise prevention strategies.

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).[3–5] However, BMD has limited sensitivity and specificity in the prediction of fracture.[6–8] In fact, a large number of conditions have been firmly established as risk factors for the occurrence of fragility fractures, independently of BMD, and include age, gender, body mass index, family history of fractures, ethnicity, premature menopause, glucocorticoid use, rheumatoid arthritis, hyperthyroidism, hyperparathryoidism, Cushing's, anorexia nervosa, malabsorption, falls, previous fractures, smoking, high caffeine intake and alcohol abuse.[9–16] These have been combined into prediction algorithms to estimate fracture probability. When applied upon the baseline epidemiology of fragility fractures in a given population, these algorithms or tools provide estimates of absolute risks. The use of these tools, combined with intervention thresholds, is recommended by many international treatment guidelines.[17–19] However, the existing tools differ from each in many relevant aspects: their feasibility, the number and availability of clinical risk factors included, the accessibility of BMD measurements and, finally, their performance in different settings. Such diversity calls for an integrative systematic review (SR) upon which the critical appraisal and selection of tools to be used in clinical practice and research can be based. The existing reviews[20–23] have a number of important limitations, such as exclusion of males, disregard of some relevant prediction algorithms, lack of meta-analysis where applicable and, naturally, omission of important subsequent publications.

The aim of this SR and meta-analysis is to bring together and describe all relevant evidence on the structure and performance of the currently available tools to predict fracture risk in the general population, while overcoming the above limitations.

METHODS

This study was conducted in line with the guidelines of the Cochrane Collaboration and our findings are reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.[24 25]

Protocol and registration

The reviewers (AM, RJOF) and a mentor (JAPS) established the protocol for this SR. Advanced technical advice was obtained from experts (LC, EL, ES). This protocol was not published but is available upon request.

Eligibility criteria

We established the following inclusion criteria for studies:

CrossMark

Linked

► http://dx.doi.org/10.1136/annrheumdis-2015-208245

A. Population—general adult population, both men and women

B. Intervention/test—any fracture risk prediction tool, score, algorithm or other instruments available to predict risk of fracture (with or without BMD measurement)

C. Comparator/control—because we wished to evaluate the performance of prediction tests, we defined the observed occurrence of the event of interest—osteoporotic fracture—as the 'gold standard'

D. Outcome/performance—the primary outcome measure was the area under the curve (AUC) of the fracture risk prediction and its SE, obtained from receiver operating characteristic (ROC) analysis, in the predetermined prediction time-interval. This was chosen as the primary outcome because the AUC represents the accuracy of the predictive model, that is, the probability that a randomly chosen subject with fracture is correctly rated or ranked with greater risk than a randomly chosen individual without fracture[26]

E. Design—cohort studies (either prospective or retrospective) and case–control studies if past data were available for all subjects.

Osteoporotic fracture risk prediction tools were only included in the final analyses if they were developed from an initial population (derivation model) and then externally validated in a different population (validation model), to prevent overestimated accuracy. Studies that included only specific disease populations (eg, chronic renal failure or rheumatoid arthritis patients) were also excluded. We also excluded studies that considered the performance of single variables, such as weight or age. We accepted the definition of major osteoporotic (MOP) fracture adopted by each tool (see below).

Information sources

We only searched published articles. One reviewer (RJOF) performed the electronic search, piloted in PubMed MEDLINE (2003–2014) and then adapted to run also in Cochrane (2003–2014) and Embase (2003–2014). The last search was run on 28 February 2014, with monthly automatic email updates until 6 September 2014. We supplemented electronic searches by checking references cited in published SR and in the articles extracted from the electronic searches. Conference abstracts and unpublished studies were not searched.

Search and study selection

The search strategies included free terms and medical descriptors (eg, MeSH terms) for each PICOD synonym. Some terms used were: Osteoporosis, 'Osteoporotic fractures', 'Risk Assessment', Algorithms, 'Area Under Curve', 'Sensitivity and Specificity', 'Validation Studies' and 'Cohort Studies'. The complete electronic string used for PubMed is provided in online supplementary table S1.

The following limits were applied: (a) articles published after 2003 (as no such studies had been published before then); (b) written in English, Spanish, French, Italian or Portuguese; and (c) performed in humans.

Studies were screened for inclusion over three phases, using Endnote software: (a) we searched and deleted duplicates; (b) two authors (AM and RJOF) independently assessed the electronic search results. They first screened by title and then by abstract. When a title seemed relevant, the abstract was reviewed for eligibility; (3) if any doubt remained, the full text of the article was retrieved and discussed. Arbitration by a third author (JAPS), applied in case of persistent disagreement, took place in

two cases. The reason for exclusion was recorded after the full text screening. The inter-rater agreement between AM and RJOF for selection based on title, abstract and full text, measured with the κ statistic, was 0.99, 0.90 and 0.98, respectively.

The meta-analysis only included articles satisfying, cumulatively, the following four criteria: (a) only validation studies were considered (not the derivation models of the tool); (b) the tool had been validated for the country where the study was performed; (c) the tool had been validated for the outcome of the study (eg, studies employing in the prediction of vertebral fractures, a tool that had only been validated to predict hip fractures, were excluded); and (d) data were reported on at least 100 fracture events (as recommended by Vergouwe et al[27]).

Data collection

All the field researchers (AM, RJOF, ES, EL, LC and JAPS) validated the data extraction form, which was pilot-tested for feasibility and comprehensiveness with five studies and submitted to consensual minor adjustments. The data were extracted by one author (AM) into a Microsoft Excel spreadsheet. Data included the general characteristics of each study and the outcomes measured. A second author (RJOF) confirmed all the data extracted. We contacted some authors in order to obtain additional information, namely regarding required outcome statistical data (CIs and/or SE of AUCs).

Data items

We collected information on the following: (a) study (authors, year, country); (b) methods (study design, inclusion and exclusion criteria, tool(s) evaluated, factors/variables included in the fracture risk estimation, duration of follow-up, adjustment for time of follow-up, number of participants at the start and at the end of follow-up, reasons for loss to follow-up); (c) participants' characteristics (age, sex, race, diseases, medication); (d) fracture characteristics (number per site, ascertainment methods); and (e) outcome results for (i) all fractures, (ii) major fractures and (iii) hip fractures (AUC and SE or 95% CIs).

Risk of bias in individual studies

The quality of each study was independently appraised by two investigators (AM and RJOF) using the Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS-2) checklist,[28] and disagreements were solved by a third researcher (JAPS). We excluded some of the optional items of this checklist and added some new ones, as recommended by QUADAS authors[29] and described by other groups.[20] This resulted in a total of 14 items, all graded as adequate, inadequate or unclear (see online supplementary table S2). This quality assessment was not used to include/exclude data for meta-analysis, except for item 19, which refers to a minimum of 100 events of interest.

Summary measures and synthesis of results

For the synthesis of results, data were pooled and meta-analysis performed using Stata V.12 software (StataCorp, 2011). All results derived from primary studies (AUC and SE) were subjected to double data entry and the pooled AUC with 95% CIs were obtained from random effect meta-analyses by instrument type, fracture site, sex, and whether BMD was included or not.

To test heterogeneity among the studies, the I^2 of Higgins and Thompson was calculated. An I^2 value close to 0% indicates no heterogeneity between studies, close to 25% indicates low heterogeneity, close to 50% indicates moderate heterogeneity, and close to 75% indicates high heterogeneity.[30 31]

RESULTS

We included a total of 45 articles, evaluating 13 different tools. Figure 1 shows the study flow-chart. We identified 3546 articles from PubMed MEDLINE, 571 from Embase and 928 from Cochrane, and selected 60 for detailed review, of which 30 were excluded: 15 did not assess fracture risk prediction tools, 12 did not provide information regarding osteoporotic fracture outcome and 3 were SRs. We identified 15 additional articles through hand searching (n=13) and through saved search email updates (n=2). A total of 45 articles were finally included.

The main characteristics of the 13 tools identified are presented in table 1. The number of factors required for calculation varies from 4 in FRAMO to 31 in the updated QFracture (2012) (see online supplementary table S3). Seven tools include BMD as a risk factor (two as an optional item). Seven tools only predict fracture risk for women. Some tools are available on the internet, the algorithm's formula is published in the article for others, and some are available only on request from the authors. The age range of valid prediction is variable: limited to the interval of 70 to 100 years in FRAMO, to 30 to 99 years in updated QFracture (2012). Most tools were developed for populations above 40–50 years of age. Regarding the time-horizon of prediction, most

tools calculate a 5-year (n=7) or a 10-year risk (n=7). Fracture and Immobilization Score (FRISC) and the updated QFracture (2012) allow the shortest time of prediction (1 year) while some tools provide more than one time-interval, like FRISC with four time-points (1, 3, 5 and 10 years) and the updated QFracture (2012) with 10 time-points (1–10 years). Regarding the types of fracture that is individually predicted, 10 of the 13 tools predict hip fractures and 7 predict major or any osteoporotic fractures. The definition of MOP fracture differs between tools. FRAX considers MOP as the combination of hip, clinical spine, wrist, and humerus.[32] The definition of the updated QFracture is similar, but all vertebral fractures are included, not only the clinical ones.[33] GARVAN's definition of MOP fracture includes all those considered by FRAX plus distal femur, proximal tibia/fibula, distal tibia/fibula, patella, pelvis, rib, sternum, hands and feet (excluding digits).[34]

In addition, FRAMO predicts the mortality risk, and FRISC the immobilisation risk. The 'Computer model for osteoporotic fracture risk' tool provides an estimation of risk reduction after osteoporosis treatment. Finally, regarding the number of published studies assessing each tool, FRAX (with 26 studies in 9 countries), GARVAN (also known as GRX, 6 studies in
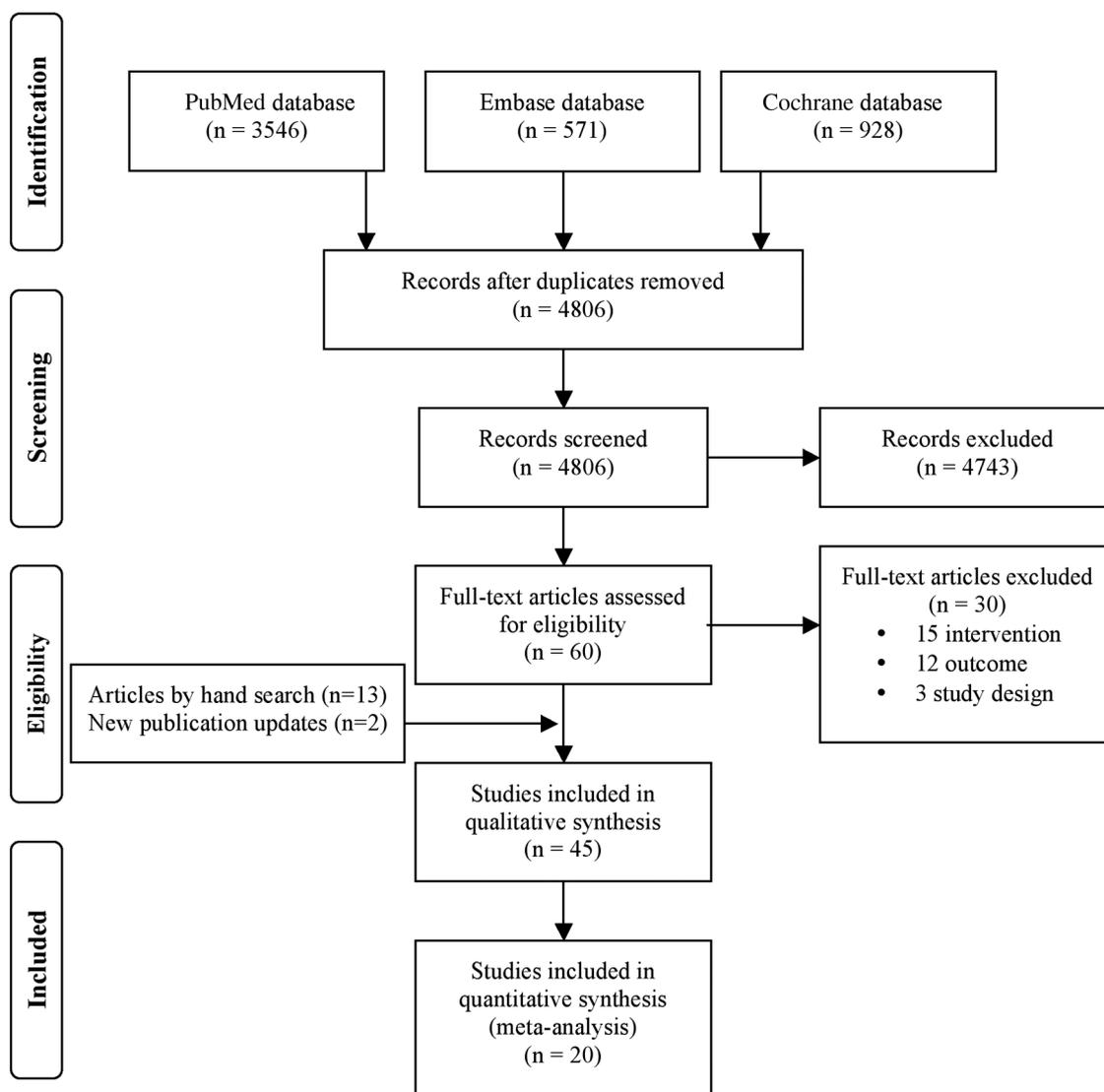

Figure 1 Flow-chart of the article selection.

Table 1 Characteristics of the fracture risk prediction tools

| Tool | Characteristics | | | | |
	Number of clinical risk factors*	BMD	Tool accessibility Gender Age range	Prediction time(s) and outcome(s)	Number of studies
'Computer model for osteoporotic fracture risk'	8	Yes	Request from authors Female only 45–79	5 years Absolute fracture risk Expected absolute risk reduction after treatment	1[35]
FRAMO	4	No	Available in article Male/Female 70–100	2 years Hip fracture risk Mortality	2[36 37]
FRAX	11	Optional	http://www.shef.ac.uk/FRAX Male/Female 40–90	10 years Major osteoporotic fracture risk Hip fracture risk	26[32 38–62]
FRC	12†	Yes	https://riskcalculator.fore.org/ Male/Female ≥45	10 years Hip fracture risk	2[63 64]
FRISC	8	Yes	http://www.biostatistics.jp/prediction/frisc Female only 40–100	1, 3, 5 and 10 years Major osteoporotic fracture risk Immobilisation risk	2[59 65]
FRISK	5	Yes	Available in article Male/Female >60	5 and 10 years Major osteoporotic fracture risk	2[66 67]
GARVAN-GRX	5	Optional	http:/garvan.org.au/promotions/bone-fracture-risk/calculator/ Male/Female 60–96	5 and 10 years Any osteoporotic/fragility fracture risk Hip fracture risk	6[34 39 54 55 68 69]
QFracture	19	No	Removed from website in 2012 Male/Female 30–85	1–10 years Any osteoporotic fracture risk Hip fracture risk	3[33 42 70]
Updated QFracture (2012)	31	No	http://www.qfracture.org/ Male/Female 30–99	1–10 years Any osteoporotic fracture risk Hip fracture risk	1[71]
Score for estimating the long-term risk of fracture in post-menopausal women	8	No	Available in article Females only ≥50	5 years Clinical vertebral fracture risk Clinical osteoporotic fracture risk Hip fracture risk	1[72]
Simplified fracture risk system	5	Yes	Available in article Female only ≥50	10 years Any fracture risk	1[73]
SOF	14	No	Available in article Female only Age unclear	5 years Hip fracture risk	1[74]
WHI	11	No	Request from authors Females only ≥50	5 years Hip fracture risk	1[75]

*Bone mineral density (BMD) was not considered.
Additional descriptions are provide in online supplementary table S3.
†An updated version of the website, dated September 2014, also includes BMD of spine, glucocorticoid exposure, and previous spine fracture, which were not part of the original publication included in this SR. No further publications supporting this change could be found.
FRISC, Fracture and Immobilization Score; FRISK, Fracture Risk prediction; SR, systematic review.

3 countries) and QFracture (3 studies in the UK, 1 including Irish participants) are the most extensively studied. All other tools have been evaluated by only one or two studies.

Methodological quality of the studies
A complete assessment of the quality of the 45 studies, using QUADAS-2, as well as a direct comparison between FRAX, QFracture and GARVAN studies, may be found in online supplementary figures S1 and S2.

Online supplementary table S4 shows the main characteristics of the 45 included articles. Thirty-five of the studies had a longitudinal prospective design, eight a longitudinal retrospective

and one a cross-sectional.[66] We also included a randomised controlled trial (RCT).[43] The mean time of follow-up in the prospective cohorts ranged from 2[36 37 54] to 13.4 years[76] and from 1.7[55] to 11 years (median)[60] in the retrospective cohorts. Five studies evaluated two different tools, and most of them were conducted in North America, Scandinavian, Western Europe, Australia or Japan. Only two studies were multinational. The exclusion criteria were not described in 10 studies and were only briefly mentioned in many others. Only one study stated that no exclusion criteria were applied. The most common exclusion criteria were: unable to walk, use of corticosteroids, bisphosphonates or other bone-active agents, previous history of

hip or MOP fractures, hip replacement and secondary osteoporosis. Participants were mainly recruited from the general population (n=22), but also from osteoporosis screenings (n=12), or were post-menopausal women (n=9). Concerning the total population at baseline, only one study[33] provides this number for both the derivation and validation models, while 14 studies do not provide the baseline numbers, even for the validation model. This number varies from 390[37] to over 1 m.[33] All articles provide the number of participants available for event verification. The majority of studies included only women (n=30), while two studies included only men.[45 64] The participant's age in the validation model ranged from 30[33] to 116 years.[32] The numbers of fractures were usually given for hip and/or MOP fracture, but other sites and other specific outcomes were presented according to the tool (eg, immobilisation)[65] or specific aims of the study (eg, in obese and non-obese subjects).[62] Diagnosis of fractures was based principally on self-report, confirmed by X-rays in 35 studies, or medical records/hospital discharge registers. The highest performances (AUC) were reported for FRAX in China[41] (Hip$_{Women}$ with BMD=0.88; Hip$_{Women}$ without BMD=0.89) and for the updated QFracture[71] (Hip$_{Women}$=0.89; Hip$_{Men}$=0.88). The lowest AUCs (FRAX$_{Men; US; MOP}$=0.54; FRAX$_{Men; UK; MOP}$=0.57) were reported in a retrospective study using a FRAX model which had not been validated for that country and with a very small population.[55]

Meta-analysis

A total of 20 articles were selected for the meta-analysis. The reasons for exclusions are described in online supplementary table S5, the most relevant being number of fractures <100 and AUCs provided only for specific subgroups, for example, as defined by economic status. FRAX provides the greatest specification of outcomes: per site, per gender, with/without BMD. All studies with GARVAN included BMD, while QFracture excludes this measurement. Thus, we performed 10 different meta-analyses for FRAX (15 studies), 3 for GARVAN

(5 studies) and 4 for QFracture (3 studies; we did not include updated QFracture published in 2012 because it only had one external validation study). Regarding the total number of participants included in the meta-analyses, GARVAN had the lowest numbers, QFracture was in between and FRAX had the largest numbers. All meta-analyses showed high heterogeneity, with the exception of one for FRAX and one for GARVAN (moderate heterogeneity). The results of all meta-analyses are presented in table 2. Overall, QFracture obtained the highest AUCs, being above 0.80 in three out of four studies. The three meta-analyses of GARVAN resulted in AUCs of around 0.70. Meta-analyses of studies with FRAX resulted in AUCs of between 0.61 and 0.79.

Pooled AUC data regarding hip fractures are presented in figure 2. This cannot be done for MOP fracture as the definition differs between the three tools.

We compared the risk prediction accuracy of excluded against included studies with meta analysis and found statistically significant higher AUC in the former studies (data not shown).

Summary appraisal of tools

In table 3 we compare aspects of the three different tools deemed relevant for their selection for clinical and research purposes. Most of these features have been mentioned above.

The countries and contexts in which these three major tools have been tested differ considerably. FRAX has been used to evaluate risk of fracture and death in 57 countries and has been the subject of 26 different validation studies in 9 countries. GARVAN was tested only in Australia, New Zealand and Canada. However, it has been proposed that this instrument does not require incorporation of national fracture data.[34 69] QFracture was only validated in the UK (with 88 participants from another country, Ireland, included) but by independent research teams; however, conversely, it has the largest number of participants.

Table 2 Meta-analyses of fracture risk assessment tools according to outcome specifications

Tool	Outcome specifications (BMD/site/sex)	Number of studies	Number of participants	Meta-analysis: random effect model AUC (95% CI)	Heterogeneity, I^2
FRAX (10-year prediction)	Y/MOP/W	n=5[39 41 43 44 60]	14 224	0.67 (0.64 to 0.71)*	80.2%*
	N/MOP/W	n=7[39 41 42 44 47 48 76]	24 726	0.65 (0.63 to 0.68)*	67.6%*
	N/Hip/W	n=9[39 41–44 47 48 53 57]	131 244	0.74 (0.68 to 0.80)*	94.3%*
	Y/Hip/W	n=5[39 41 44 53 57]	115 611	0.79 (0.73 to 0.85)*	93.3%*
	N/MOP/M	n=2[45 47]	11 199	0.63 (0.60 to 0.66)*	0.0%
	N/Hip/M	n=2[45 47]	11 199	0.71 (0.65 to 0.77)*	40.8%
	Y/MOP/B	n=3[46 51]	276 786	0.63 (0.60 to 0.66)*	97.1%*
	Y/Hip/B	n=3[46 51]	276 786	0.77 (0.73 to 0.81)*	69.8%*
	N/MOP/B	n=3[46 51]	276 786	0.61 (0.57 to 0.64)*	96.3%*
	N/Hip/B	n=3[46 51]	276 786	0.67 (0.61 to 0.73)*	94.7%*
GARVAN-GRX (10-year prediction)	Y/Hip/W	n=2[68 39]	5574	0.74 (0.61 to 0.87)*	88.2%*
	Y/MOP/W	n=3[39 68 69]	6932	0.70 (0.64 to 0.75)*	93.8%*
	Y/MOP/M	n=2[68 69]	5010	0.73 (0.68 to 0.78)*	59.0%
QFracture (10-years prediction)	N/MOP/W	n=3[33 70]	1 778 570	0.81 (0.78 to 0.834)*	97.8%*
	N/MOP/M	n=2[33 70]	1 741 983	0.72 (0.67 to 0.76)*	99.2%*
	N/Hip/W	n=3[33 42 70]	1 779 154	0.89 (0.88 to 0.89)*	96.3%*
	N/Hip/M	n=2[33 70]	1 741 983	0.87 (0.86 to 0.88)*	71.0%

Moderate heterogeneity: Higgins I^2 ~50%; high heterogeneity, Higgins I^2 ~75%.
*p<0.05.
AUC, area under the curve; B, both sexes; BMD, bone mass density; Hip, hip fractures; M, men; MOP, major osteoporotic fractures (MOPs are defined differently for the different instruments); N, without BMD; W, women; Y, with BMD.
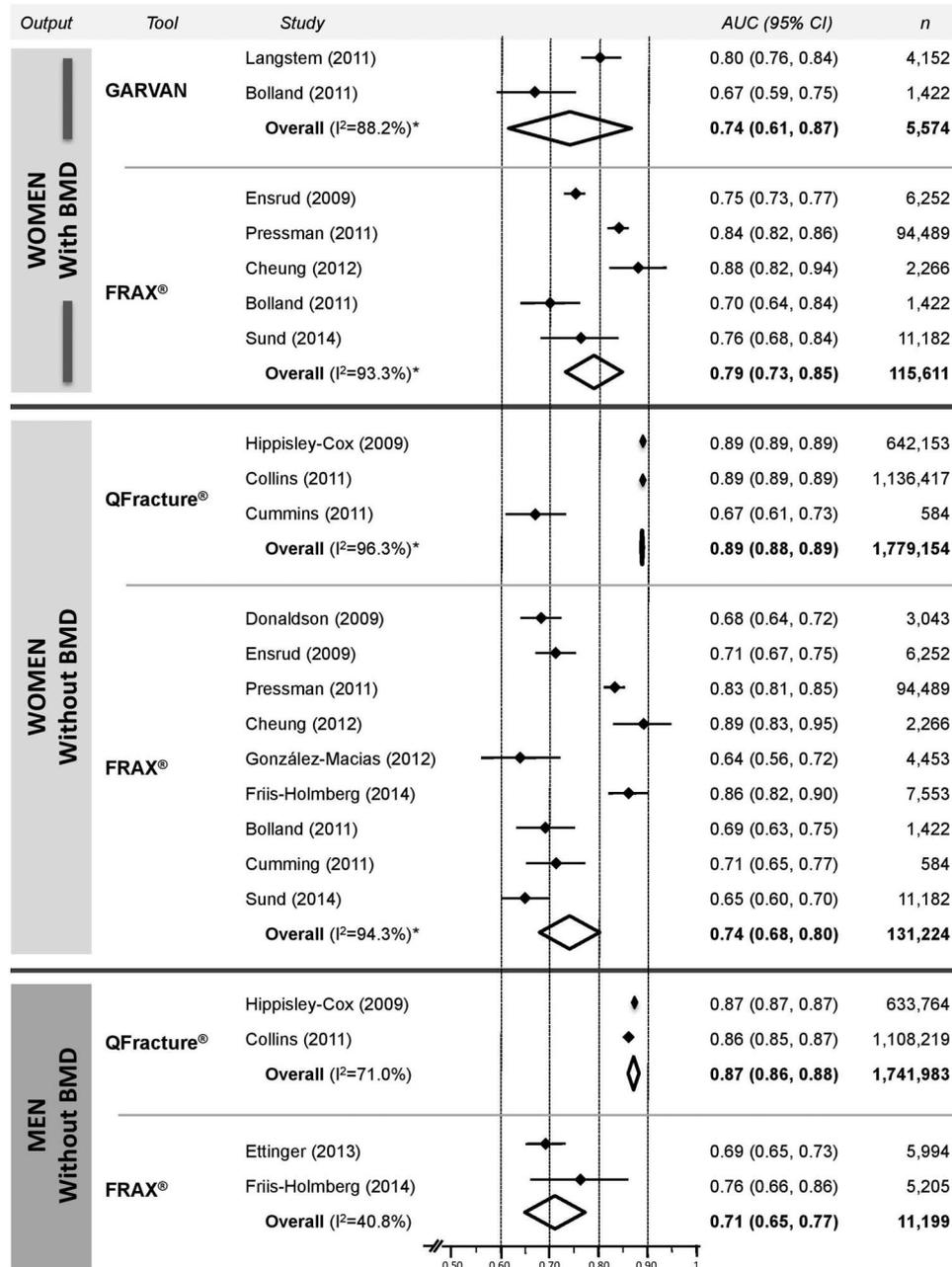
Figure 2 FRAX, GARVAN and QFracture pooled areas under the curve (AUCs) (95% CI) for 10-year hip fracture prediction, according to sex and bone mineral density (BMD) input.

QFracture is associated with the highest AUC, this being achieved at the cost of greater complexity and lower feasibility, given the large number of risk factors considered.

DISCUSSION

This SR identified 13 tools for osteoporotic fracture risk prediction, adding one new instrument (FRISK)[66 67] to the algorithms identified by previous SRs,[20–23] and updating the validation information regarding those already identified. This will help clinicians and researchers select those that best apply to their setting and needs. We have also performed a meta-analysis for 10-year risk prediction of hip and MOP fractures with FRAX, GARVAN and QFracture (for men, women and both genders, with and without BMD). To the best of our knowledge, this is the first meta-analysis on this topic.

The differences between the currently available fracture prediction tools must be emphasised, as caution is required when comparing the results obtained with different instruments. The number of risk factors considered (which varies between 4 and 31), as well as their nature, will have an important impact on feasibility. Differences in output (sex, age, types of fractures and time-intervals of prediction) might affect the applicability of the tool. All instruments predict the risk of osteoporotic fractures but not all provide separate estimations for hip and for major fractures.

On the other hand, our quality assessment of the included studies reveals, as with previous evaluations,[20–23] significant pitfalls in most of the studies, although recent publications appear to be of better quality.[45 47] Among the most important drawbacks is the lack of certainty of unbiased recruitment from the target population.

Table 3 Summary features of the three most studied tools, as deemed relevant for the selection of the instrument in clinical and research settings

	FRAX	QFracture	GARVAN
Feasibility			
Number of clinical risk factors	11	19	5
Requirement for BMD	Optional	No	Optional
Algorithm accessible for individual use	Yes	No*	Yes
Applicability			
Male and female	Yes	Yes	Yes
Age range	40–90	35–100	50–96
Prediction intervals	10	1, 2, …, 10	5, 10
Type of fracture—hip	Yes	Yes	Yes
Type of fracture—MOP	Yes	Yes	Yes
Countries	57	UK only	3
Inclusion in national guidelines	Yes	Yes	Yes
Validity			
Validated in a separate cohort	Yes	Yes	Y (BMD only)
Independent validation†	Yes	Yes	Y (BMD only)
Number of validation studies	26	3	6‡
Population basis for validation, N	4 624 438	3 485 952§	229 162
Population basis for validation, countries	9¶	UK only	3¶
Average quality of studies (QUADAS-2)	Globally similar (see online supplementary figure S2)		
Duration of follow-up equal to tool estimation interval	Yes	Yes (10 year only)	Yes (5 and 10 year)
Consideration of national fracture epidemiology	Yes	No	No
Consideration of background mortality	Yes	No	No
AUC (95% CI)—hip, females, without BMD	0.74 (0.68 to 0.80)	0.89 (0.88 to 0.89)	NA
AUC (95% CI)—hip, females, with BMD	0.79 (0.73 to 0.85)	NA	0.74 (0.61 to 0.87)
AUC (95% CI)—hip, males, without BMD	0.71 (0.65 to 0.77)	0.87 (0.86 to 0.88)	NA
AUC (95% CI)—hip, males, with BMD	0.77 (NA)**	NA	0.85 (NA)**
AUC (95% CI)—MOP	††	††	††

*QFracture was removed from the website in 2012. Only the updated version is now available, but is not suitable for meta-analysis as it has only been the subject of one validation study.
†That is, by independent research groups.
‡Only with BMD.
§Does not include the updated QFracture (2012) study.
¶We did not consider the study that included data from 10 countries.
**One study only.
††Comparison is inadequate because of different definitions of MOP for each tool.
AUC, area under the curve; BMD, bone mineral density; MOP, major osteoporotic fractures; NA, not applicable/not available; QUADAS, Quality Assessment Tool for Diagnostic Accuracy Studies.

There is also a lack of correspondence between the spectrum of participants and the population expected to receive the test in daily practice. This problem was observed in about 50% of included studies and in about 50% of the reports of the three major tools. All the instruments were validated for the general population, but several studies recruited participants from osteoporosis screening settings,[38 40 42 50–52 55 60 63 72 73] while some explicitly excluded people treated for osteoporosis.[41 42 53 55 56 58 63] Reports, unfortunately, do not provide the detailed data necessary for assessing the potential impact of treatment upon fracture prediction. We also verified that two studies excluded individuals previously exposed to glucocorticoids,[42 43] even though this risk factor was included in the risk algorithm under evaluation.

Follow-up time was consistent with the time-horizon of prediction validated for the tool in only a third of the studies. Furthermore, most of those without the required follow-up time[32 33 40 41 43–45 47–50 53–55 61 62 65 68 70 71] did not perform any statistical adjustments for this, which may have influenced the estimated AUCs.

Attrition is a well-known problem in longitudinal epidemiological studies.[77] The attrition rates vary considerably between the included studies, and most of them did not explain these rates. Death is a common cause of attrition in cohort studies of older people,[78] which affects the accuracy of the models. Only some studies in this SR took this into account.[39 41 45 47 48 57 58 61 62 64] One study[38] excluded women who died during follow-up, even though fracture, or its complications, might have been the cause of death.

For practical reasons we will focus our discussion below on FRAX, QFracture and GARVAN, as only these tools have been the subject of more than two validation studies testing exactly the same algorithm. FRISC has three validation studies, but each of them considered a different number of risk factors.

FRAX, GARVAN and QFracture can differentially predict risk in men and women and estimate the risk for hip and MOP fracture. However, the definition of the latter is different in each tool, thus precluding direct comparison.

QFracture and updated QFracture (2012) include a larger number and wider variety of clinical risk factors than FRAX and GARVAN. It is likely that algorithms with the longest lists of risk factors will have feasibility and adherence problems, but also greater accuracy. On the other hand, shorter lists may decrease the accuracy of the prediction. In some studies, the

authors excluded some of required risk factors, which inevitably weakens the robustness of the prediction, even if the impact upon the AUC and c-statistic is typically small.[79 80] In fact, even strong risk factors will have a minimal impact on the AUC if their prevalence in the studied population is low. This may be mistakenly reassuring and, as a rule, prediction tools should be used in strict accordance with the instructions provided by the authors, which in turn reflect the conditions of validation. There are, therefore, several potential caveats in the conclusion that deleting risk factors or opting for simpler ones is a good choice on the basis of the AUC alone.[81]

In FRAX, fracture probability is computed taking both the risk of fracture and the risk of death into account. Neither GARVAN nor QFracture include mortality. Kanis et al[81] have shown that this induces an inadequate continuous increase in the risk predicted by GARVAN in very advanced age. It is possible that the same may happen with QFracture.

Accuracy of estimates
Comparing instruments based on their AUCs, we found important pitfalls related, first and foremost, to differences in the definitions of events and to the participants' characteristics.[81] AUCs also tend to be smaller, the narrower the age range and the longer the duration of follow-up.[81]

To avoid these pitfalls we have: (a) appraised the quality of studies; (b) excluded the original studies, that is, derivation models from meta-analysis; and (c) restricted the comparative analysis to minimally comparable data (hip fractures).

We found that the meta-analysis of studies indicates higher AUCs with QFracture (0.89 and 0.87) than FRAX (0.74 and 0.71) when comparable data are available: hip fractures in women and men, respectively, both without BMD. The 95% CIs in the main two studies and the overall results of QFracture are practically residual and much smaller than those observed for FRAX (0.68 to 0.80 and 0.65 to 0.77), which reflects the larger number of participants in the QFracture studies. QFracture was designed for integration into electronic records systems where all necessary data have already been collected as part of routine care, as in the clinical research databases that served to derive and validate the model. The tool is incorporated into the electronic system allowing automatic calculation. The setting is very convenient but extremely hard to reproduce elsewhere. Derivation and validation were performed in different population samples, but from the same country, which favours a higher AUC. The fact that the tool amenable to meta-analysis (QFracture 2009) is no longer available adds to these difficulties.

Adding BMD to FRAX increases the AUC from 0.74 to 0.79 in women, and from 0.71 to 0.77 in men, but this is still below the values achieved with QFracture (0.89 and 0.87, respectively). Comparing the meta-analysis for GARVAN and FRAX, is only possible for hip fractures in women, using BMD—the results indicate a small numerical advantage for FRAX.

The performance of all these tools was validated for the general population. Thus, their application for specific settings (eg, osteoporosis population, secondary causes of osteoporosis) implies a risk of error. Further studies should also evaluate the threshold for use in clinical practice. Comparison between tools should, ideally, be made in the same population.

Limitations and strengths of this study
Assessing the quality of the studies with QUADAS-2 proved a difficult task, mostly due to poor reporting, and may be controversial as regards some points.

Concerning the meta-analysis, we frequently had to calculate the SE based on other parameters, which may have led to slightly different results (at a centesimal level).

We did not ask authors to provide data on age when this was missing from the publications. This may have slightly influenced the results of meta-analysis, as age may affect the AUC.[81] The only way to adjust our meta-analyses by age was to include studies with similar age bands or to stratify. We did the first but not the second as it was not possible to stratify with the published data.

Using AUC as the outcome for the meta-analysis could also be seen as a limitation, given its weaknesses as discussed above. Furthermore, given that fracture rates differ significantly from country to country, comparison of data obtained in different countries involves some risk of error. However, the vast majority of studies only provide these data.

Among the strengths of this study we would emphasise the comprehensiveness of the literature search and appraisal. Although we did not include so-called 'grey literature' (ie, congress abstracts and unpublished data), the hand search gives us a high degree of confidence that no major studies were missed. No study was excluded for language reasons. We limited our meta-analyses to sets of data that we found to be valid and directly comparable, thus avoiding most of the potential errors in similar exercises. Because we recognised significant heterogeneity, the analyses were performed using the random effects model,[30 31 82] which assumes that the effect of interest is not the same in all studies. This is a more conservative approach, resulting in wider 95% CIs, while, hopefully, reducing the risk of unrealistic assumptions.[30] This was the first meta-analysis performed on data from fracture risk prediction tools.

Conclusions
Thirteen externally validated algorithms designed to predict osteoporotic fracture risk are currently available to clinicians and researchers. Most of these tools are feasible in clinical practice and are simple to access and use. FRAX, QFracture and GARVAN are the most extensively studied tools, with FRAX having the greatest number of independent studies. FRAX was evaluated in a larger number of countries and also allows finer specification of outcomes. Adding BMD to FRAX increases the AUC for hip fractures in both men and women. Studies with QFracture present the highest AUCs; however, this tool has only been studied in the UK and Ireland and requires consideration of 19 clinical factors. This number was actually increased to 31 in the updated version, with a marginal increase in accuracy.

Methodological limitations and risk of bias are present in most studies, but to a lower extent than in the oldest studies. High-quality studies to assess the calibration of fracture prediction tools are still needed. Researchers should use the instruments in accordance with the requirements and indications for which they were validated, in order to allow international unbiased comparisons and better quantitative synthesis.

Author affiliations
[1]Rheumatology Department, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal
[2]Health Sciences Research Unit: Nursing (UICiSA:E), Coimbra, Portugal
[3]Instituto de Salud Musculoesquelética—InMusc, Madrid, Spain
[4]Faculty of Medicine, Clínica Universitária de Reumatologia, University of Coimbra, Coimbra, Portugal

REFERENCES

1. Papaioannou A, Adachi JD, Parkinson W, et al. Lengthy hospitalization associated with vertebral fractures despite control for comorbid conditions. *Osteoporos Int* 2001;12:870–4.
2. Wiktorowicz ME, Goeree R, Papaioannou A, et al. Economic implications of hip fracture: health service use, institutional care and cost in Canada. *Osteoporos Int* 2001;12:271–8.
3. Kanis JA, Melton LJ 3rd, Christiansen C, et al. The diagnosis of osteoporosis. *J Bone Miner Res* 1994;9:1137–41.
4. Looker AC, Wahner HW, Dunn WL, et al. Updated data on proximal femur bone mineral levels of US adults. *Osteoporos Int* 1998;8:468–89.
5. Kanis JA, McCloskey EV, Johansson H, et al. A reference standard for the description of osteoporosis. *Bone* 2008;42:467–75.
6. Kanis JA, Delmas P, Burckhardt P, et al. Guidelines for diagnosis and management of osteoporosis. The European Foundation for Osteoporosis and Bone Disease. *Osteoporos Int* 1997;7:390–406.
7. Abu-Rajab RB, Watson WS, Walker B, et al. Peri-prosthetic bone mineral density after total knee arthroplasty. *J Bone Joint Surg Br* 2006;88:606–13.
8. Watts NB. Is it ethical to use placebos in osteoporosis clinical trials? *Curr Osteoporos Rep* 2004;2:31–6.
9. Kanis JA, Johansson H, Johnell O, et al. Alcohol intake as a risk factor for fracture. *Osteoporos Int* 2005;16:737–42.
10. Kanis JA, Johnell O, Oden A, et al. Smoking and fracture risk: a meta-analysis. *Osteoporos Int* 2005;16:155–62.
11. Kanis JA, Johansson H, Oden A, et al. A family history of fracture and fracture risk: a meta-analysis. *Bone* 2004;35:1029–37.
12. De Laet C, Kanis JA, Oden A, et al. Body mass index as a predictor of fracture risk: a meta-analysis. *Osteoporos Int* 2005;16:1330–8.
13. Johansson H, Kanis JA, Oden A, et al. BMD, clinical risk factors and their combination for hip fracture prevention. *Osteoporos Int* 2009;20:1675–82.
14. Kanis JA, Johansson H, Oden A, et al. A meta-analysis of prior corticosteroid use and fracture risk. *J Bone Miner Res* 2004;19:893–9.
15. Kanis JA, Johnell O, De Laet C, et al. A meta-analysis of previous fracture and subsequent fracture risk. *Bone* 2004;35:375–82.
16. Kanis JA, Borgstrom F, De Laet C, et al. Assessment of fracture risk. *Osteoporos int* 2005;16:581–9.
17. Papaioannou A, Morin S, Cheung AM, et al. 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. *CMAJ* 2010;182:1864–73.
18. Kanis JA, Hans D, Cooper C, et al. Interpretation and use of FRAX in clinical practice. *Osteoporos int* 2011;22:2395–411.
19. Hans DB, Kanis JA, Baim S, et al. Joint Official Positions of the International Society for Clinical Densitometry and International Osteoporosis Foundation on FRAX(®). Executive Summary of the 2010 Position Development Conference on Interpretation and use of FRAX(®) in clinical practice. *J Clin Densitom* 2011;14:171–80.
20. Rubin KH, Friis-Holmberg T, Hermann AP, et al. Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review. *J Bone Miner Res* 2013;28:1701–17.
21. Nayak S, Edwards DL, Saleh AA, et al. Performance of risk assessment instruments for predicting osteoporotic fracture risk: a systematic review. *Osteoporos Int* 2014;25:23–49.
22. Steurer J, Haller C, Hauselmann H, et al. Clinical value of prognostic instruments to identify patients with an increased risk for osteoporotic fractures: systematic review. *PLoS ONE* 2011;6:e19994.
23. Nelson HD, Haney EM, Chou R, et al. Systematic Review to Update the 2002 US Preventive Services Task Force Recommendation. Rockville, MD: Agency for Healthcare Research and Quality (US), 2010.
24. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264–9, w64.
25. 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:e1000100.
26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
27. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172:971–80.
28. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
29. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006;6:9.
30. Moayyedi P. Meta-analysis: can we mix apples and oranges? *Am J Gastroenterol* 2004;99:2297–301.
31. Sousa MR, Ribeiro AL. Systematic review and meta-analysis of diagnostic and prognostic studies: a tutorial. *Arq Bras Cardiol* 2009;92:229–38, 35–45.
32. Kanis JA, Oden A, Johnell O, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 2007;18:1033–46.
33. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ (Online)* 2009;339:1291–5.
34. Nguyen ND, Frost SA, Center JR, et al. Development of a nomogram for individualizing hip fracture risk in men and women. *Osteoporos Int* 2007;18:1109–17.
35. Ettinger B, Hillier TA, Pressman A, et al. Simple computer model for calculating and reporting 5-year osteoporotic fracture risk in postmenopausal women. *J Womens Health* 2005;14:159–71.
36. Albertsson DM, Mellstrom D, Petersson C, et al. Validation of a 4-item score predicting hip fracture and mortality risk among elderly women. *Ann Fam Med* 2007;5:48–56.
37. Albertsson D, Mellstrom D, Petersson C, et al. Hip and fragility fracture prediction by 4-item clinical risk score and mobile heel BMD: a women cohort study. *BMC Musculoskelet Disord* 2010;11:55.
38. Azagra R, Roca G, Martin-Sanchez JC, et al. FRAX® thresholds to identify people with high or low risk of osteoporotic fracture in Spanish female population. *Med Clin* 2015;144:1–8.
39. Bolland MJ, Siu AT, Mason BH, et al. Evaluation of the FRAX and Garvan fracture risk calculators in older women. *J Bone Miner Res* 2011;26:420–7.
40. Brennan SL, Leslie WD, Lix LM, et al. FRAX provides robust fracture prediction regardless of socioeconomic status. *Osteoporos Int* 2014;25:61–9.
41. Cheung EYN, Bow CH, Cheung CL, et al. Discriminative value of FRAX for fracture prediction in a cohort of Chinese postmenopausal women. *Osteoporos Int* 2012;23:871–8.
42. Cummins NM, Poku EK, Towler MR, et al. Clinical risk factors for osteoporosis in Ireland and the UK: a comparison of FRAX and Qfracture Scores. *Calcif Tissue Int* 2011;89:172–7.
43. Donaldson MG, Palermo L, Schousboe JT, et al. FRAX and risk of vertebral fractures: the fracture intervention trial. *J Bone Miner Res* 2009;24:1793–9.
44. Ensrud KE, Lui LY, Taylor BC, et al. A comparison of prediction models for fractures in older women: is one better? *Arch Intern Med* 2009;169:2087–94.
45. Ettinger B, Ensrud KE, Blackwell T, et al. Performance of FRAX in a cohort of community-dwelling, ambulatory older men: the Osteoporotic Fractures in Men (MrOS) study. *Osteoporos Int* 2013;24:1185–93.
46. Fraser LA, Langsetmo L, Berger C, et al. Fracture prediction and calibration of a Canadian FRAX® tool: a population-based report from CaMos. *Osteoporos Int* 2011;22:829–37.
47. Friis-Holmberg T, Rubin KH, Brixen K, et al. Fracture risk prediction using phalangeal bone mineral density or FRAX®?-A Danish cohort study on men and women. *J Clin Densitom* 2014;17:7–15.
48. Gonzalez-Macias J, Marin F, Vila J, et al. Probability of fractures predicted by FRAX® and observed incidence in the Spanish ECOSAP Study cohort. *Bone* 2012;50:373–7.
49. Hillier TA, Cauley JA, Rizzo JH, et al. WHO absolute fracture risk models (FRAX): do clinical risk factors improve fracture prediction in older women without osteoporosis? *J Bone Miner Res* 2011;26:1774–82.
50. Leslie WD, Lix LM, Johansson H, et al. Spine-hip discordance and fracture risk assessment: a physician-friendly FRAX enhancement. *Osteoporos Int* 2011;22:839–47.
51. Leslie WD, Lix LM, Johansson H, et al. Independent clinical validation of a Canadian FRAX tool: fracture prediction and model calibration. *J Bone Miner Res* 2010;25:2350–8.
52. Leslie WD, Lix LM. Absolute fracture risk assessment using lumbar spine and femoral neck bone density measurements: derivation and validation of a hybrid system. *J Bone Miner Res* 2011;26:460–7.
53. Pressman AR, Lo JC, Chandra M, et al. Methods for assessing fracture risk prediction models: experience with FRAX in a large integrated health care delivery system. *J Clin Densitom* 2011;14:407–15.
54. Sambrook PN, Flahive J, Hooven FH, et al. Predicting fractures in an international cohort using risk factor algorithms without BMD. *J Bone Miner Res* 2011;26:2770–7.

55 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;21:863–71.

56 Sornay-Rendu E, Munoz F, Delmas PD, et al. The FRAX tool in French women: how well does it describe the real incidence of fracture in the OFELY cohort. *J Bone Miner Res* 2010;25:2101–7.

57 Sund R, Honkanen R, Johansson H, et al. Evaluation of the FRAX model for hip fracture predictions in the population-based Kuopio Osteoporosis Risk Factor and Prevention Study (OSTPRE). *Calcif Tissue Int* 2014;95:39–45.

58 Tamaki J, Iki M, Kadowaki E, et al. Fracture risk prediction using FRAX®: a 10-year follow-up survey of the Japanese Population-Based Osteoporosis (JPOS) Cohort Study. *Osteoporos Int* 2011;22:3037–45.

59 Tanaka S, Kuroda T, Saito M, et al. Urinary pentosidine improves risk classification using fracture risk assessment tools for postmenopausal women. *J Bone Miner Res* 2011;26:2778–84.

60 Tebe Cordomi C, del Rio LM, Di Gregorio S, et al. Validation of the FRAX predictive model for major osteoporotic fracture in a historical cohort of Spanish women. *J Clin Densitom* 2013;16:231–7.

61 Tremollieres FA, Pouilles JM, Drewniak N, et al. Fracture risk prediction using BMD and clinical risk factors in early postmenopausal women: sensitivity of the WHO FRAX tool. *J Bone Miner Res* 2010;25:1002–9.

62 Premaor M, Parker RA, Cummings S, et al. Predictive value of FRAX for fracture in obese older women. *J Bone Miner Res* 2013;28:188–95.

63 Lo JC, Pressman AR, Chandra M, et al. Fracture risk tool validation in an integrated healthcare delivery system. *Am J Manag Care* 2011;17: 188–94.

64 Ettinger B, Liu H, Blackwell T, et al. Validation of FRC, a fracture risk assessment tool, in a cohort of older men: the Osteoporotic Fractures in Men (MrOS) Study. *J Clin Densitom* 2012;15:334–42.

65 Tanaka S, Yoshimura N, Kuroda T, et al. The Fracture and Immobilization Score (FRISC) for risk assessment of osteoporotic fracture and immobilization in postmenopausal women—a joint analysis of the Nagano, Miyama, and Taiji cohorts. *Bone* 2010;47:1064–70.

66 Henry MJ, Pasco JA, Sanders KM, et al. Fracture Risk (FRISK) Score: Geelong Osteoporosis Study. *Radiology* 2006:190–6.

67 Henry MJ, Pasco JA, Merriman EN, et al. Fracture risk score and absolute risk of fracture. *Radiology* 2011;259:495–501.

68 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:E107–14.

69 Nguyen ND, Frost SA, Center JR, et al. Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks. *Osteoporos Int* 2008;19:1431–44.

70 Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ* 2011;342:d3651.

71 Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* 2012;344:e3427.

72 van Staa TP, Geusens P, Kanis JA, et al. A simple clinical score for estimating the long-term risk of fracture in post-menopausal women. *QJM* 2006;99:673–82.

73 Leslie WD, Tsang JF, Lix LM. Simplified system for absolute fracture risk assessment: clinical validation in Canadian women. *J Bone Miner Res* 2009;24:353–60.

74 Ahmed LA, Schirmer H, Fonnebo V, et al. Validation of the Cummings' risk score; how well does it identify women with high risk of hip fracture: the Tromso Study. *Eur J Epidemiol* 2006;21:815–22.

75 Hundrup YA, Jacobsen RK, Andreasen AH, et al. Validation of a 5-year risk score of hip fracture in postmenopausal women. The Danish Nurse Cohort Study. *Osteoporos Int* 2010;21:2135–42.

76 Tremollieres F, Cochet T, Cohade C, et al. Fracture risk in early postmenopausal women assessed using FRAX. *Joint Bone Spine* 2010;77:345–8.

77 Gustavson K, von Soest T, Karevold E, et al. Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study. *BMC Public Health* 2012;12:918.

78 Brilleman SL, Pachana NA, Dobson AJ. The impact of attrition on the representativeness of cohort studies of older people. *BMC Med Res Methodol* 2010;10:71.

79 Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008;54:17–23.

80 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–35.

81 Kanis JA, Oden A, Johansson H, et al. Pitfalls in the external validation of FRAX. *Osteoporos Int* 2012;23:423–31.

82 Santos E, Cunha M. Interpretação Crítica dos Resultados Estatísticos de uma Meta-Análise: Estratégias Metodológicas. *Millenium* 2013;44:85–98.