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 Silva1,4

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, hyperparathyroidism, 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 reviews20–23 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:

unpublished studies were not searched. We supplemented electronic searches by checking references cited in published SR and in the articles published before then; (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 al25).

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 authors29 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 I² of Higgins and Thompson was calculated. An I² 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. The definition of the updated QFracture is similar, but all vertebral fractures are included, not only the clinical ones. 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). 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 1960 Marques A, et al. Ann Rheum Dis 2015;74:1958–1967. doi:10.1136/annrheumdis-2015-207907

Figure 1 Flow-chart of the article selection.
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). 41 The mean time of follow-up in the prospective cohorts ranged from 2 to 13.4 years and from 1.7 to 11 years 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 fracture, bone mineral density (BMD) was not considered.


---

Table 1 Characteristics of the fracture risk prediction tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Number of clinical risk factors*</th>
<th>BMD</th>
<th>Gender accessibility</th>
<th>Tool accessibility</th>
<th>Age range</th>
<th>Prediction time(s) and outcome(s)</th>
<th>Number of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer model for osteoporotic fracture risk'</td>
<td>8</td>
<td>Yes</td>
<td>Request from authors</td>
<td>Request from authors</td>
<td>Female only 45–79</td>
<td>5 years Absolute fracture risk Expected absolute risk reduction after treatment</td>
<td>138</td>
</tr>
<tr>
<td>FRAMO</td>
<td>4</td>
<td>No</td>
<td>Available in article</td>
<td>Available in article</td>
<td>Male/Female 70–100</td>
<td>2 years Hip fracture risk Mortality</td>
<td>236 37</td>
</tr>
<tr>
<td>FRAX</td>
<td>11</td>
<td>Optional</td>
<td><a href="http://www.shef.ac.uk/FRAX">http://www.shef.ac.uk/FRAX</a></td>
<td>Male/Female 40–90</td>
<td>10 years Major osteoporotic fracture risk Hip fracture risk</td>
<td>26238 38-42</td>
<td></td>
</tr>
<tr>
<td>FRC</td>
<td>12†</td>
<td>Yes</td>
<td><a href="https://riskcalculator.fore.org/">https://riskcalculator.fore.org/</a></td>
<td>Male/Female ≥45</td>
<td>10 years Hip fracture risk</td>
<td>53 64</td>
<td></td>
</tr>
<tr>
<td>FRISC</td>
<td>8</td>
<td>Yes</td>
<td><a href="http://www.biostatistics.jp/prediction/frisc">http://www.biostatistics.jp/prediction/frisc</a></td>
<td>Female only 40–100</td>
<td>1, 3, 5 and 10 years Major osteoporotic fracture risk Immobilisation risk</td>
<td>59 65</td>
<td></td>
</tr>
<tr>
<td>FRISK</td>
<td>5</td>
<td>Yes</td>
<td>Available in article</td>
<td>Available in article</td>
<td>Male/Female &gt;60</td>
<td>5 and 10 years Major osteoporotic fracture risk</td>
<td>66 67</td>
</tr>
<tr>
<td>GARVAN-GRX</td>
<td>5</td>
<td>Optional</td>
<td><a href="http://garvan.org.au/promotions/bone-fracture-risk/calculator/">http://garvan.org.au/promotions/bone-fracture-risk/calculator/</a></td>
<td>Male/Female 60–96</td>
<td>5 and 10 years Any osteoporotic/fragility fracture risk Hip fracture risk</td>
<td>64 39 54 55 68 69</td>
<td></td>
</tr>
<tr>
<td>QFracture</td>
<td>19</td>
<td>No</td>
<td>Removed from website in 2012</td>
<td>Male/Female 30–85</td>
<td>1–10 years Any osteoporotic fracture risk Hip fracture risk</td>
<td>33 42 70</td>
<td></td>
</tr>
<tr>
<td>Updated QFracture (2012)</td>
<td>31</td>
<td>No</td>
<td><a href="http://www.qfracture.org/">http://www.qfracture.org/</a></td>
<td>Male/Female 30–99</td>
<td>1–10 years Any osteoporotic fracture risk Hip fracture risk</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>Score for estimating the long-term risk of fracture in post-menopausal women</td>
<td>8</td>
<td>No</td>
<td>Available in article</td>
<td>Available in article</td>
<td>Females only ≥50</td>
<td>5 years Clinical vertebral fracture risk Clinical osteoporotic fracture risk Hip fracture risk</td>
<td>172</td>
</tr>
<tr>
<td>Simplified fracture risk system</td>
<td>5</td>
<td>Yes</td>
<td>Available in article</td>
<td>Available in article</td>
<td>Female only ≥50</td>
<td>10 years Any fracture risk</td>
<td>173</td>
</tr>
<tr>
<td>SOF</td>
<td>14</td>
<td>No</td>
<td>Available in article</td>
<td>Available in article</td>
<td>Female only Age unclear</td>
<td>5 years Hip fracture risk</td>
<td>174</td>
</tr>
<tr>
<td>WHI</td>
<td>11</td>
<td>No</td>
<td>Request from authors</td>
<td>Request from authors</td>
<td>Females only ≥50</td>
<td>5 years Hip fracture risk</td>
<td>175</td>
</tr>
</tbody>
</table>

*Bone mineral density (BMD) was not considered.

Additional descriptions are provide in online supplementary table S3.

1 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.
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. 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.
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) to the algorithms identified by previous SRs 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 emphasized, 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, significant pitfalls in most of the studies, although recent publications appear to be of better quality. Among the most important drawbacks is the lack of certainty of unbiased recruitment from the target population.

---

**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.

<table>
<thead>
<tr>
<th>Output</th>
<th>Tool</th>
<th>Study</th>
<th>AUC (95% CI)</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOMEN With BMD</td>
<td>GARVAN</td>
<td>Langstern (2011)</td>
<td>0.80 (0.76, 0.84)</td>
<td>4,152</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bolland (2011)</td>
<td>0.67 (0.59, 0.75)</td>
<td>1,422</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Overall (P=88.2%)</strong> *</td>
<td>0.74 (0.61, 0.87)</td>
<td>5,574</td>
</tr>
<tr>
<td></td>
<td>FRAX®</td>
<td>Ensrud (2009)</td>
<td>0.75 (0.73, 0.77)</td>
<td>6,252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pressman (2011)</td>
<td>0.84 (0.82, 0.88)</td>
<td>94,469</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cheung (2012)</td>
<td>0.88 (0.82, 0.94)</td>
<td>2,266</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bolland (2011)</td>
<td>0.70 (0.64, 0.84)</td>
<td>1,422</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sund (2014)</td>
<td>0.76 (0.68, 0.84)</td>
<td>11,182</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Overall (P=93.3%)</strong> *</td>
<td>0.79 (0.73, 0.85)</td>
<td>115,611</td>
</tr>
<tr>
<td>WOMEN Without BMD</td>
<td>FRAX®</td>
<td>Donaldson (2009)</td>
<td>0.68 (0.64, 0.72)</td>
<td>3,043</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ensrud (2009)</td>
<td>0.71 (0.67, 0.75)</td>
<td>6,252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pressman (2011)</td>
<td>0.83 (0.81, 0.85)</td>
<td>94,469</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cheung (2012)</td>
<td>0.89 (0.83, 0.95)</td>
<td>2,266</td>
</tr>
<tr>
<td></td>
<td></td>
<td>González-Macias (2012)</td>
<td>0.64 (0.56, 0.72)</td>
<td>4,453</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fris-Holmberg (2014)</td>
<td>0.86 (0.82, 0.90)</td>
<td>7,553</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bolland (2011)</td>
<td>0.69 (0.63, 0.75)</td>
<td>1,422</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cumming (2011)</td>
<td>0.71 (0.65, 0.77)</td>
<td>584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sund (2014)</td>
<td>0.65 (0.60, 0.70)</td>
<td>11,182</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Overall (P=94.3%)</strong> *</td>
<td>0.74 (0.68, 0.80)</td>
<td>131,224</td>
</tr>
<tr>
<td>MEN Without BMD</td>
<td>QFracture®</td>
<td>Hippisley-Cox (2009)</td>
<td>0.87 (0.87, 0.87)</td>
<td>633,764</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Collins (2011)</td>
<td>0.86 (0.85, 0.87)</td>
<td>1,108,219</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Overall (P=71.0%)</strong></td>
<td>0.87 (0.86, 0.88)</td>
<td>1,741,983</td>
</tr>
<tr>
<td></td>
<td>FRAX®</td>
<td>Ettlinger (2013)</td>
<td>0.69 (0.65, 0.73)</td>
<td>5,994</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fris-Holmberg (2014)</td>
<td>0.76 (0.86, 0.88)</td>
<td>5,205</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Overall (P=40.8%)</strong></td>
<td>0.71 (0.65, 0.77)</td>
<td>11,199</td>
</tr>
</tbody>
</table>
Clinical and epidemiological research

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>FRAX</th>
<th>QFracture</th>
<th>GARVAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clinical risk factors</td>
<td>11</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>Requirement for BMD</td>
<td>Optional</td>
<td>No</td>
<td>Optional</td>
</tr>
<tr>
<td>Algorithm accessible for individual use</td>
<td>Yes</td>
<td>No*</td>
<td>Yes</td>
</tr>
<tr>
<td>Male and female</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Age range</td>
<td>40–90</td>
<td>35–100</td>
<td>50–96</td>
</tr>
<tr>
<td>Prediction intervals</td>
<td>10</td>
<td>1, 2, ... 10</td>
<td>5, 10</td>
</tr>
<tr>
<td>Type of fracture—hip</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Type of fracture—MOP</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Countries</td>
<td>57</td>
<td>UK only</td>
<td>3</td>
</tr>
<tr>
<td>Inclusion in national guidelines</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**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.

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 while some explicitly excluded people treated for osteoporosis.39 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,41 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 time32 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 study38 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.\textsuperscript{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.\textsuperscript{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 \textit{et al}\textsuperscript{82} 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.\textsuperscript{83} AUCs also tend to be smaller, the narrower the age range and the longer the duration of follow-up.\textsuperscript{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, it 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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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:En), Coimbra, Portugal
3. Instituto de Salud Musculoesquelética—IMUSC, Madrid, Spain
4. Faculty of Medicine, Clínica Universitaria de Reumatología, University of Coimbra, Coimbra, Portugal

**Acknowledgements** The authors thank John Kanis, Julia Hippisley-Cox and Tuan V. Nguyen, the main authors of the FRAX, QFracture and GARVAN studies, respectively, for critical review of this article and constructive comments. They also would like to thank the researchers who kindly provided additional, unpublished data from their studies: Bruce Ettinger, Gary S. Collins, John Kanis, Lisa Langsetmo and Niamh Cummins.

Contributors All authors of this research paper have directly participated in the planning, execution, or analysis of this study; have read and approved the final version submitted and gave the necessary attention to ensure the integrity of the work.

Funding AM was supported for this study by an educational grant from the Portuguese Health Directorate.

Competing interests AM and JPS were involved in validation of the FRAX algorithm for the Portuguese population.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES


26 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis

Andréea Marques, Ricardo J O Ferreira, Eduardo Santos, Estibaliz Loza, Loreto Carmona and José António Pereira da Silva


Updated information and services can be found at:
http://ard.bmj.com/content/74/11/1958

These include:

Supplementary Material
Supplementary material can be found at:
http://ard.bmj.com/content/suppl/2015/08/06/annrheumdis-2015-207907.DC1
http://ard.bmj.com/content/suppl/2015/08/06/annrheumdis-2015-207907.DC2

References
This article cites 80 articles, 9 of which you can access for free at:
http://ard.bmj.com/content/74/11/1958#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

ARD Lay summaries (83)
Editor’s choice (158)
Calcium and bone (725)
Musculoskeletal syndromes (4951)
Osteoporosis (137)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/