

# Comparison between various fracture risk assessment tools

W. D. Leslie · L. M. Lix

Received: 24 April 2013 / Accepted: 24 May 2013 / Published online: 25 June 2013  
© International Osteoporosis Foundation and National Osteoporosis Foundation 2013

**Abstract** The suboptimal performance of bone mineral density as the sole predictor of fracture risk and treatment decision making has led to the development of risk prediction algorithms that estimate fracture probability using multiple risk factors for fracture, such as demographic and physical characteristics, personal and family history, other health conditions, and medication use. We review theoretical aspects for developing and validating risk assessment tools, and illustrate how these principles apply to the best studied fracture probability tools: the World Health Organization FRAX®, the Garvan Fracture Risk Calculator, and the QResearch Database's QFractureScores. Model development should follow a systematic and rigorous methodology around variable selection, model fit evaluation, performance evaluation, and internal and external validation. Consideration must always be given to how risk prediction tools are integrated into clinical practice guidelines to support better clinical decision making and improved patient outcomes. Accurate fracture risk assessment can guide clinicians and individuals in understanding the risk of having an osteoporosis-related fracture and inform their decision making to mitigate these risks.

**Keywords** Fractures · Osteoporosis · Prediction tools · Statistical models

## Introduction

The presence of osteoporosis is a major risk factor for the development of fractures of the hip, proximal humerus, vertebra, and forearm (often termed the “major osteoporotic fracture” sites) though many other skeletal sites are also at increased risk of fracture [1]. The consequences of fracture include increased mortality, morbidity, institutionalization, and economic costs [2–6]. Moreover, all osteoporosis related fractures can lead to significant long-term disability and decreased quality of life [7, 8]. Worldwide, the number of people who have suffered a prior osteoporotic fracture was estimated to be 56 million in 2000 with approximately 9 million new osteoporotic fractures each year [9]. As the prevalence of osteoporosis increases with age, the global burden of osteoporosis is projected to rise markedly over the coming decades due to an increasing number of elderly individuals in the population [10].

In the absence of a defining fracture, the diagnosis of osteoporosis is based on the measurement of bone mineral density (BMD) by dual X-ray absorptiometry (DXA). The World Health Organization provided an operational definition of osteoporosis given as a BMD that lies 2.5 standard deviations or more below the average mean value for young healthy women [T-score  $\leq -2.5$  standard deviation (SD)] based upon a standardized reference site (the femoral neck) and a standard reference range for both men and women (the NHANES III data for women aged 20–29 years) [11–13]. BMD measurement from DXA provides a relative estimate of fracture risk along a continuum, increasing 1.4- to 2.6-fold for every SD decrease in BMD [14, 15]. The accuracy of BMD measurements using central DXA to predict osteoporotic fractures is comparable to the use of blood pressure measurement for prediction of stroke and is superior to serum cholesterol as a predictor of myocardial infarction [15]. Although reduced bone mass is an important and easily quantifiable measurement, studies have shown that most

---

W. D. Leslie · L. M. Lix  
University of Manitoba, Winnipeg, Canada  
L. M. Lix  
e-mail: lisa.lix@med.umanitoba.ca

W. D. Leslie (✉)  
Department of Medicine (C5121), 409 Tache Avenue,  
Winnipeg, Canada R2H 2A6  
e-mail: bleslie@sbgh.mb.ca

fractures occur in individuals with a BMD T-score above the defining cutoff for osteoporosis [16]. The T-score categorization of BMD can be credited for contributing to increased awareness of osteoporosis as a significant health problem, but with the benefit of hindsight and almost two decades of additional research, the limitations of the T-score are also apparent and include: **site dependence in age-related bone loss**, uncertainty about the appropriate reference population for use in men and non-Caucasians, inability to separate cortical from trabecular compartments, and the difficulty in characterizing a complex three-dimensional volume with a two-dimensional areal projection as measured by central DXA.

The suboptimal sensitivity and specificity of using BMD alone for prediction of fracture risk has led to the development of new risk prediction algorithms that estimate fracture probability using additional risk factors for fracture, such as demographic and physical characteristics, personal and family history, other health conditions, and medication use. Risk assessment tools are primarily intended for use by non-experts and are, by necessity, (over)simplifications of highly complex relationships. No risk assessment tool can include all known risk factors for osteoporosis and fracture, or nuances associated with the severity and temporal relationships. In primary care, a tool that works “most of the time” is preferable to no tool at all. Since no tool can capture the scientific knowledge and clinical experience of the content expert, clinical judgment must always be exercised in applying results to the individual patient. That said, even the expert can benefit from prediction tools which provide a benchmark and contribute to greater consistency in the clinical approach.

This review examines theoretical aspects of risk assessment tools and illustrates how these principles apply to the best studied fracture probability tools: the World Health Organization FRAX<sup>®</sup>, the Garvan Fracture Risk Calculator, and the QResearch Database’s QFractureScores.

## Theoretical considerations

### Model construction

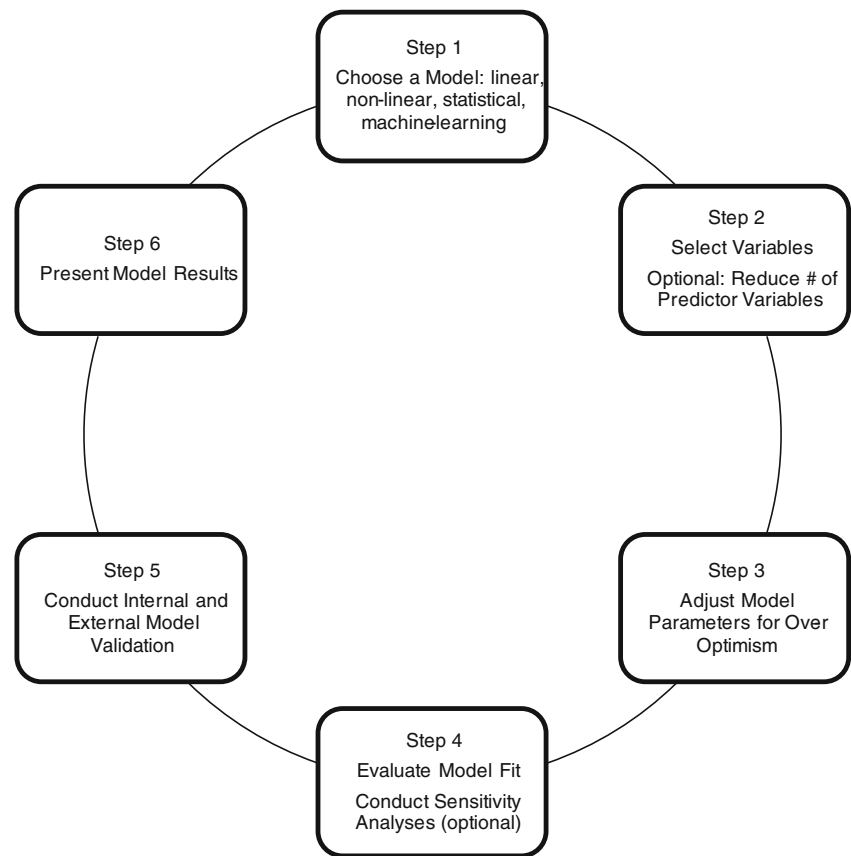
Multivariable parametric, semi-parametric, non-parametric, and machine-learning statistical models have all been used to produce risk estimates for a variety of health outcomes, including fracture. Key considerations in developing these estimates include the rationale for selection of a statistical or machine-learning model, selection of predictor variables or variable reduction techniques to develop a parsimonious model (i.e., simplest model with the smallest number of variables to achieve maximum predictive performance), adjustment of parameter estimates to ensure accurate prediction, evaluation of model performance, with internal and external validation to ensure replicability, and generalizability of risk estimates

[17–19]. Figure 1 identifies the major steps in the development of risk estimates. The process is iterative, so that researchers may move either backward or forward through these steps as they refine their modeling approach and its application to a dataset.

At the outset, it should be noted that risk prediction algorithms are often criticized because they provide marginal (i.e., average) risk estimates, which are relevant for populations and not for individuals. However, the vast majority of recommendations about treatment and practice in the medical literature are also based on marginal algorithms. While the utility of adopting a “personalized medicine” approach, which uses genetic characteristics of the individual in addition to traditional risk factors to produce risk estimates, is beginning to be explored for some health outcomes, there have been few, if any, studies that have investigated this topic for fracture risk estimation. Moreover, recent evidence suggests that improvements in predictive performance that arise from the addition of genetic information to traditional risk prediction models are small [20, 21].

Parametric or semi-parametric models for producing risk estimates include multiple linear regression for continuous outcomes (e.g., for prediction of BMD scores) and logistic or Cox proportional hazards regression models for dichotomous outcomes (e.g., fracture presence versus absence) [22]. The logistic model is adopted when observation time for risk estimation of the event of interest (e.g., fracture) is fixed for each individual while the Cox model is frequently adopted when the duration of the observation period for the event of interest varies and there is the potential for censoring (i.e., loss of follow-up) for individuals. Non-parametric or machine learning models such as classification trees, recursive partitioning techniques, artificial neural networks, and support vector machines may result in improved prediction accuracy when there is a non-linear relationship between the predictor variables and the outcome or some mathematical function of the outcome. Trees and recursive partitioning techniques may be appealing to clinicians because they use a hierarchical classification process involving a series of yes-no decisions; this is similar to the way clinicians make diagnostic decisions. The goal is to place each individual into a class in which the incidence of the outcome is either high or low. Machine-learning models are advantageous for very large datasets in which there may be hundreds or even thousands of potential variables to include in the prediction model [23]. Accordingly, exploratory studies to develop new prediction models may benefit from the use of machine-learning models. Hybrid models have been proposed to improve prediction accuracy; for example, Abu-Hanna and de Keizer [24] developed a prediction model that combines the classification and regression tree with the logistic regression model; the latter is applied within the specific patient sub-populations created by the sequence of classifications that produce the tree nodes.

**Fig. 1** Major steps in risk assessment model development



The authors found that this approach resulted in improved precision for predicting the risk of intensive care unit outcomes such as death.

**Selection of a parsimonious set of predictor variables is crucial when either total sample size or the number of events per predictor is small. Rules of thumb suggest that at least 10 events per predictor variable (and preferably many more) must be available in the data to produce stable parameter estimates in the logistic regression model; simulation studies have shown that with less than 10 events per variable, accuracy and precision of the model parameter estimates will be less than optimal [25]. As well, variable selection techniques such as forward, backward, and stepwise selection have limited power when sample size is small [26] and may therefore result in misleading conclusions about variable retention in the final model.**

Different statistical or machine-learning models may result in discordant conclusions about the variables to be retained in the final model because they rest on different assumptions about the nature of the relationships between the outcome and the predictors and/or are based on different estimation methods that may not result in equivalent conclusions about relative importance. The selection of predictor variables should therefore be based, at least in part, on theoretical and clinical considerations to provide face validity for the model. In theory, all predictor variables

suspected of being associated with the outcomes of interest should be considered as model candidates. However, it need not be the case that all predictor variables are assumed to have a causal relationship with the outcome; some measures are selected because of their known association with other variables that may not be measured in the available data.

As well, the selection of model predictors should be based on consideration of the potential for accurate measurements to be collected for the variables. For example, self-reported height will be less accurate than measured height. Measurement error may result in biased estimates of association between the outcome and predictor variables, thereby reducing prediction accuracy.

While a priori variable reduction can often be achieved with clinical input, empirical variable reduction techniques, such as principal components analysis, can aid in the development of a parsimonious model; such techniques are particularly important when collinearity (i.e., high correlations) exists among the predictor variables under investigation, which can result in poor model performance. At the same time, principal components analysis, or other variable-reduction techniques, must be used cautiously and this analysis may not an appropriate choice if the predictor variables are not continuous and normally distributed. As well, a single variable that represents a linear combination of multiple collinear variables may be difficult for clinicians

and patients to interpret. Li et al. used principal components analysis to combine different measures of bone fragility, based on imaging data, to construct a hip fracture risk model [27]. Exploration of collinearity among a set of variables by using the variance inflation factor, a measure of the amount of inflation in the variance of the estimated model parameters when there is correlation among the predictor variables, may facilitate decision making about the use of principal components analysis as a variable reduction technique. Collinearity may be particularly problematic between main effect and interaction effect terms; centering of the variables by subtracting the mean score will help to reduce collinearity.

Accuracy of the prediction model is often better in the original (i.e., developmental) population than in new patient populations. Overfitting of the model to the data may result in over-optimistic estimates of the model parameters. Overfitting commonly occurs when a model has too many parameters relative to the number of observations. As well, use of the model parameters developed from the original population in a new patient population does not take account of random variation in the estimates, which again can result in overly optimistic estimates of model parameters in the original data. Correction (i.e., shrinkage) of parameter estimates is recommended prior to their application in new populations. A number of methods to shrink the parameter estimates have been proposed. Moons et al. [28] compared different shrinkage techniques in a logistic regression model and recommended a penalized version of maximum likelihood estimation, although the researchers noted that this method is not available in existing statistical software packages such as SAS. Steyerberg et al. [29] noted that a simple linear shrinkage factor worked well in a small dataset for predicting 30-day mortality.

Likelihood ratio tests and penalized measures of the likelihood function, such as the Akaike Information Criterion (AIC) and Bayesian–Schwarz Information Criterion (BIC), are commonly used to evaluate model fit for parametric and semi-parametric models based on maximum likelihood estimation. The AIC and BIC measures add a penalty for including more variables in the model. These statistics can be used to assess the impact on model fit of inclusion of additional risk factors into the model or their interactions. Measures of variation, such as  $R^2$  and pseudo- $R^2$  statistics [30], provide an indication of explanatory power of the model.

### Assessing model performance

Discrimination (the model's ability to distinguish between individuals who do or do not experience the event of interest) and calibration (agreement between observed and predicted event rates for groups of individuals) are key aspects of predictive performance of risk algorithms. These

two concepts are quite different and it is possible for a model to be well-calibrated but provide poor discrimination, and vice versa. For example, a model that reports the same average fracture risk for everyone in the population could be perfectly calibrated (i.e., observed and predicted fractures across the whole population are the same) but would clearly provide no useful discrimination (i.e., individuals developing fractures or remaining fracture-free are scored the same). Alternatively, multiplying or dividing all risk scores by 100 would obviously strongly alter the number of predicted events (i.e., calibration), but high-risk and low-risk individuals would still show the same relative ordering and discrimination would be the same.

The Brier score, which provides an indication of agreement between an observed binary outcome and the predicted probability of that outcome, is a measure of both calibration and discrimination. Lower Brier scores indicate improved model accuracy [31, 32]. Scores can range from 0 to 0.25 for a non-informative model, assuming 50 % incidence of the outcome. If the outcome incidence is lower, the maximum value of the Brier score for a non-informative model will be lower. Spiegelhalter's  $z$ -test is used to evaluate the statistical significance of the calibration component of the Brier score [33]; a significant test result indicates poor calibration.

The ability of the model to discriminate between individuals with and without the outcome is commonly assessed using the  $c$ -statistic [34, 35], which corresponds to the area under the receiver operation characteristic (ROC) curve for binary outcomes. The  $c$ -statistic ranges from zero to one, with a value of one representing perfect prediction and a value of 0.5 representing chance prediction. A value between 0.7 and 0.8 is typically considered to demonstrate acceptable predictive performance, while a value greater than 0.8 is indicative of excellent predictive performance. The difference in  $c$ -statistics for two nested models (i.e., two models in which the predictor variables in one model are a subset of the predictor variables in a second model) serves as an indication of the improvement in discriminative performance when new risk factors are added to a baseline model. DeLong et al. [36] proposed a test of the statistical significance of the change in the  $c$ -statistic for nested models.

Some newer methods to assess discrimination and classification include variants of the  $c$ -statistic for survival analyses [37, 38] that allow for censoring of observations, and methods that involve analysis of risk reclassification, that is, comparisons of the classification of individuals under different risk algorithms [39]. Reclassification tables and net reclassification improvement statistics provide clinicians with tools to understand the potential uncertainty in risk estimates, or to explore the effect of a novel risk factor on discriminative performance. The latter describes how much more frequently appropriate reclassification occurs than

inappropriate reclassification when a new risk algorithm is adopted. Statistically significant changes in the *c*-statistic may not always correspond with changes in risk categorization. Alternatively, ROC analyses show a lack of sensitivity to additional risk factors even when the presence (or absence) of such risk factors can make a difference in determining whether an individual patient lies below or above an intervention threshold. For example, the addition of a strong but uncommon new risk factor (hazard ratio 3.0 with prevalence 1 %) would only increase the *c*-statistic from 0.7 to 0.703, but could easily alter the decision to treat an individual with that risk factor [40]. This underscores the importance of examining multiple measures of discrimination and classification in the development of a risk prediction algorithm [41, 42].

### Model validation

Validation is one of the most important steps in the development of a risk algorithm, but is also one of the most overlooked steps [17]. Both internal and external validation analyses are recommended. The former refers to the process of assessing the reproducibility of the risk algorithm in the same population, while the latter refers to the generalizability of the algorithm to other patient populations. Internal validation can be achieved by developing the risk algorithm in a random subset of the sample and then evaluating its performance in a separate random subset of the sample. For an external validation, predictions are calculated from the developed risk algorithm and then tested in new data that are sampled from a different population than the one in which the model was developed. This could include a population from a different country, facility, or health care provider.

Primary methods for internal validation are split-sample, cross-validation, and bootstrap validation [43, 44]. In split-sample validation, a portion of the sample (e.g., 70 %) is used to construct the risk algorithm and the remaining portion (e.g., 30 %) is used to validate the risk algorithm. Cross-validation uses the same methodology as split-sample validation, but repeats the process of constructing the development and validation samples. For example, a 10-fold cross-validation methodology proceeds as follows: the sample of data is divided into 10 subsets of (approximately) equal size. Then, 90 % (i.e., nine tenths) of the cohort is used to estimate the risk algorithm and estimate risk for the nine tenths of the sample, as well as for the entire sample. Optimism bias is the difference between the prediction error for the entire sample and the prediction error for the nine-tenths sample. This process is repeated 10 times, each time leaving out one tenth of the sample, to ensure that each sample participant has a predicted risk from a regression model for which he/she is excluded. The average of the 10 optimism bias estimates is computed to provide a measure of internal validity (e.g., prediction accuracy). The most

extreme form of cross-validation is the jackknife method, in which  $N-1$  sample participants ( $N$  is the total sample size) are used for model development, with validation on the individual who was excluded.

Bootstrap validation is a popular approach to assess the reproducibility of a risk algorithm. A bootstrap sample is created by randomly sampling with replacement from the sample. The bootstrap sample contains the same number of individuals (i.e.,  $N$ ) as in the original sample. However, because sampling is done with replacement, the data for an individual could appear more than once in a single bootstrap sample. The model is developed in the bootstrap sample and validated in the original sample. The difference in the prediction error for the two models indicates the optimism in model performance. The sampling process is usually repeated a large number of times (e.g.,  $B = 500$  or 1,000 bootstrap samples) in order to obtain stable results. Some researchers adjust measures of calibration and discrimination based on the average values produced from the bootstrap methodology because the adjusted measures will better reflect the expected performance in a new population.

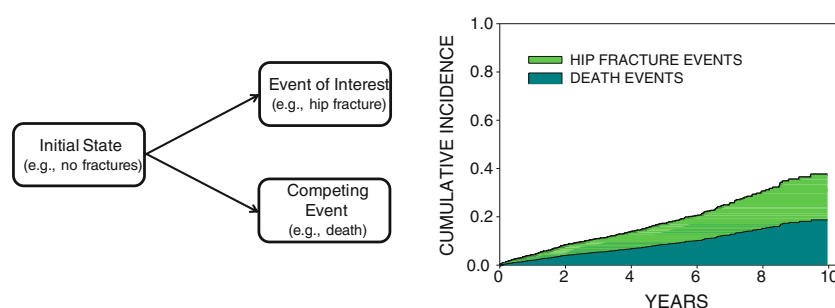
External validation techniques can be distinguished by their focus on new populations defined by time, space, or investigators [44]. For example, a risk algorithm might be externally validated by applying it to a population from the same geographic region (e.g., same country or region) but at a different point in time. Or the algorithm might be validated on a population from a different country or region. The most convincing external validation occurs when the risk algorithm is used by an independent investigator in a different dataset on a sample that is both temporally and geographically distinct from the developmental population. While external validation of multiple linear and logistic regression models has been explored in a number of studies, validation of the Cox regression model has only recently been explored [45].

### Competing risk and other considerations

While the preceding discussion has focused on the primary considerations in the development of a risk algorithm, there are other elements of model development that should be considered. For fracture, competing risks are particularly important to consider in order to produce unbiased estimates of fracture risk (Fig. 2) [46]. Treating the competing event as “censored” at the time of occurrence is inappropriate because after a competing event has occurred, fracture is no longer possible. For example, in order to estimate the risk of hip fracture among osteoporotic patients, subjects are followed from a baseline date (such as date of diagnosis) until the date of hip fracture, death, or study closing date. A patient who dies without hip fracture during the study period is considered to have a “competing event” at their



**Fig. 2** Competing risks multistate model with two outcome states: hip fracture and death. Analytical methods can separate the cumulative incidence functions related to each outcome



date of death. A patient who is alive and fracture-free at the end of the study is considered to be censored. The conventional Cox proportional hazard model considers only two endpoints: the outcome of interest (event) and loss to follow-up (censoring). Since many of the risk factors for fracture are also risk factors for death (e.g., older age, prior fracture, and medical comorbidities), failure to consider competing mortality may result in an overestimate of fracture probability. One analysis found that in subgroups with higher mortality (men, age >80 years, high fracture probability, or presence of diabetes) failure to account for competing mortality overestimated major osteoporotic fracture probability by 16–56 % with the standard nonparametric (Kaplan–Meier) method and 15–29 % with the standard parametric (Cox) model [47]. However, to date only a few statistical software packages have specific functions that consider competing outcomes [48].

A further consideration should be given to missing data, which can threaten both the internal and external validity of a risk prediction algorithm [49–51]. The influence of missing data increases as the proportion of observations increases. Ignoring missing observations can reduce statistical power to detect associations between the outcome and predictor variable. As well, if the mechanism of missing data is non-ignorable, that is, if the pattern of missingness is associated with one or more variables that are not contained in the study dataset, then the missingness can lead to erroneous inferences about the strength of association and also affect the predictive ability of the final model because the individuals for whom data are available are not a random subset of the study population.

Since Little and Rubin [52] first proposed the multiple imputation method to address missing data, it has become the primary approach in the statistical analysis of incomplete data. The multiple imputation method can be applied in combination with risk prediction models. However, this method does assume that the pattern of missingness is ignorable, that is, that it is associated with one or more measured variables in the dataset that can be used to develop a predictive model for the missingness. The primary steps in the imputation process are as follows: First,  $m > 1$  complete datasets are obtained by replacing the missing values with  $m$  imputed plausible values. Then the  $m$

complete datasets are analyzed using standard statistical analyses. The estimates of the parameters of interest from the  $m$  complete datasets are combined, typically by averaging [53]. In the averaging process, both between- and within-imputation variance is computed, both of which are needed in order to produce correct inferences about the statistical significance of risk factors in the algorithm.

### Clinical fracture prediction tools

We now review fracture probability algorithms that have been validated in at least one cohort independent from the original derivation population: the World Health Organization FRAX®, the Garvan Fracture Risk Calculator, and the QResearch Database's QFractureScores. Each of these fracture probability tools was developed in accordance with the general principles outlined above and followed the major developmental steps illustrated in Fig. 1. The fundamental design elements of these tools are summarized in Table 1, with a summary of the independent validation studies in Table 2. Tools developed to identify individuals with low BMD (e.g., SCORE, OST, and ORAI) do not provide a direct estimate of fracture probability and were not included in this review, although it is worth noting that some of these have also been shown to stratify fracture risk [54–56].

#### FRAX ([www.shef.ac.uk/FRAX](http://www.shef.ac.uk/FRAX))

FRAX® was developed by the WHO Collaborating Centre for Metabolic Bone Diseases for estimation of individual 10-year osteoporotic (clinical spine, hip, forearm, and proximal humerus) and hip fracture probability [57]. In addition to a prior fragility fracture, age, sex, body mass index, and additional risk factors for fractures were identified including the prolonged use of glucocorticoids, secondary osteoporosis, rheumatoid arthritis, a parental history of hip fracture, current cigarette smoking, and alcohol intake of three or more units per day. The clinical risk factors (CRFs) were determined in a series of meta-analyses using data from nine prospective population-based cohorts from around the world [58].

**Table 1** Fundamental elements of fracture risk prediction tools with at least one independent validation cohort

Prediction tool, URL	Risk factors	Outputs	Derivation population	Distinctive characteristics
FRAX (Fracture Risk Assessment Tool) [58] <a href="http://www.shef.ac.uk/FRAX">www.shef.ac.uk/FRAX</a>	<ul style="list-style-type: none"> <li>• Age, sex, BMI</li> <li>• Prior fragility fracture, glucocorticoid use <math>\geq 3</math> months, secondary osteoporosis, rheumatoid arthritis, parental hip fracture, current cigarette smoking, alcohol intake of <math>\geq 3</math> units/day (yes/no)</li> <li>• Femoral neck BMD or T-score (optional)</li> </ul>	<ul style="list-style-type: none"> <li>• 10-year major osteoporotic fracture (clinical vertebrae, hip, forearm, proximal humerus)</li> <li>• 10-year hip fracture</li> </ul>	<ul style="list-style-type: none"> <li>• Nine international prospective cohorts (46,340 subjects)</li> </ul>	<ul style="list-style-type: none"> <li>• Meta-analysis for clinical risk factors</li> <li>• Population-specific calibration</li> <li>• Includes fracture and mortality interaction terms</li> <li>• Competing mortality risk</li> </ul>
Garvan Fracture Risk Calculator (Dubbo nomogram) [92, 93] <a href="http://www.garvan.org.au/bone-fracture-risk">www.garvan.org.au/bone-fracture-risk</a>	<ul style="list-style-type: none"> <li>• Age, sex</li> <li>• Fractures after age 50 (none, 0, 1, 2, <math>\geq 3</math>)</li> <li>• History of falls in the previous 12 months (none, 0, 1, 2, <math>\geq 3</math>)</li> <li>• Femoral neck BMD (or T-score) or weight</li> </ul>	<ul style="list-style-type: none"> <li>• 5- or 10-year any osteoporotic fracture (hip, clinical vertebrae, wrist, metacarpal, humerus, scapula, clavicle, distal femur, proximal tibia, patella, pelvis, and sternum)</li> <li>• 5- or 10-year hip fracture</li> </ul>	<ul style="list-style-type: none"> <li>• DOES (Dubbo Osteoporosis Epidemiology Study) (1,358 women and 858 men aged <math>\geq 60</math> years)</li> </ul>	<ul style="list-style-type: none"> <li>• Includes dose-response for number of prior fractures and falls</li> </ul>
QFractureScores [95, 96] <a href="http://www.qfracture.org">www.qfracture.org</a>	<ul style="list-style-type: none"> <li>• Age, sex, 10 ethnic origins</li> <li>• Height, weight</li> <li>• Smoking (4 levels), alcohol intake (5 levels), diabetes (type 1, type 2), previous fracture, parental osteoporosis or hip fracture, living in a nursing or care home, history of falls, dementia, cancer, asthma/COPD, cardiovascular disease, chronic liver disease, chronic kidney disease, Parkinson's disease, rheumatoid arthritis/SLE, malabsorption, endocrine problems, epilepsy or anticonvulsant use, antidepressant use, steroid use, HRT use</li> </ul>	<ul style="list-style-type: none"> <li>• 1- to 10-year osteoporotic fracture (clinical spine, hip, distal forearm; humerus fracture was included with the 2012 version)</li> <li>• 1 to 10 year hip fracture</li> </ul>	<ul style="list-style-type: none"> <li>• 357 England and Wales general practices for derivation and 178 practices for internal validation (QResearch Database) (more than 1 million women and more than 1 million men in the derivation cohort)</li> </ul>	<ul style="list-style-type: none"> <li>• Includes dose-response for smoking, alcohol intake, type of diabetes</li> <li>• BMD is not an input variable</li> </ul>

BMI body mass index, BMD bone mineral density, HRT hormone replacement therapy, COPD chronic obstructive pulmonary disease

**Table 2** Independent validation studies for fracture risk prediction tools described in Table 1

Author	Population	Performance	Comments
<b>FRAX</b>			
Kanis et al. 2007 [63]	11 international cohorts (230,486 persons, 1,208,528 person years of follow-up), country-specific FRAX	<ul style="list-style-type: none"> <li>• HF prediction: averaged gradient of risk 1.83 without BMD, 2.52 with BMD; averaged AUC 0.66 without BMD, 0.74 with BMD</li> <li>• Other osteoporotic fractures: averaged gradient of risk 1.53 without BMD, 1.57 with BMD; averaged AUC 0.60 without BMD, 0.62 with BMD</li> </ul>	Did not account for the effect of competing mortality risk
Somay-Rendu et al. 2010 [64]	867 French women (OFELY cohort), French FRAX	<ul style="list-style-type: none"> <li>• MOF AUC 0.75 without BMD, 0.78 with BMD</li> <li>• Observed incidence of MOF higher than predicted probability</li> </ul>	
Trémollières et al. 2010 [65]	2,651 French women (MENOS cohort), French FRAX	<ul style="list-style-type: none"> <li>• MOF AUC 0.63 without BMD, 0.78 with BMD</li> </ul>	
Leslie et al. 2010 [67]	36,730 Canadian women and 2,873 Canadian men (Manitoba BMD cohort), Canadian FRAX	<ul style="list-style-type: none"> <li>• HF AUC 0.830 with BMD, 0.793 without BMD</li> <li>• Major osteoporosis fracture AUC 0.694 with BMD, 0.663 without BMD</li> <li>• 10-year HF for FRAX with BMD observed 2.7 % vs. predicted 2.8 % in women, observed 3.5 % vs. predicted 2.9 % in men</li> <li>• 10-year MOF for FRAX with BMD observed 12.0 % vs. predicted 11.1 % in women, observed 10.7 % vs. predicted 8.4 % in men</li> </ul>	
Fraser et al. 2011 [68]	4,778 Canadian women and 1,919 Canadian men (CaMos cohort), Canadian FRAX	<ul style="list-style-type: none"> <li>• HF AUC 0.80 with BMD, 0.77 without BMD</li> <li>• Major osteoporosis fracture AUC 0.69 with BMD, 0.66 without BMD</li> <li>• 10-year MOF probability for FRAX with BMD predicted 5.4 % vs. observed 6.4 % in men, predicted 10.8 % vs. observed 12.0 % in women</li> <li>• 10-year HF probability for FRAX with BMD predicted 2.7 % vs. observed 2.7 % in women, predicted 1.3 % vs. observed 2.4 % in men</li> <li>• The observed/expected ratio for fracture was 1.79 (95 % CI 1.44–2.21) without BMD and 1.94 (95 % CI 1.45–2.54) with BMD using UK FRAX model in Polish women</li> </ul>	Fractures could only be assessed in long-term survivors
Czerwinski et al. 2011 [69]	501 Polish women (convenience sample), UK FRAX	<ul style="list-style-type: none"> <li>• MOF AUC 0.67 without BMD, 0.69 with BMD</li> </ul>	Only 43 self-reported MOF and 4 HF
Tamaki et al. 2011 [70]	815 Japanese women (population-based JPOS cohort), Japanese FRAX	<ul style="list-style-type: none"> <li>• 10-year probability of HF agreed with observed HF overall (7.6 % vs. 7.6 %) and by age decade (<i>p</i> value was non-significant)</li> </ul>	BMD was not available
Rubin et al. 2011 [71]	3,636 Danish women (registry-based linkage study), Swedish FRAX	<ul style="list-style-type: none"> <li>• HF AUC 0.78 without BMD</li> <li>• Major fractures AUC 0.61 without BMD</li> </ul>	BMD was not available, fractures were self-reported
Sambrook et al. 2011 [104]	19,586 women from 723 primary-care practices in 10 countries (GLOW Study), country-specific FRAX	<ul style="list-style-type: none"> <li>• HF AUC 0.70, 0.69 without BMD</li> </ul>	
Bolland et al. 2011 [105]			



**Table 2** (continued)

Author	Population	Performance	Comments
	1,422 women (placebo-controlled trial of calcium supplements), New Zealand FRAX	<ul style="list-style-type: none"> <li>• MOF AUC 0.64, 0.62 without BMD</li> <li>• FRAX with BMD underestimated MOF and HF (ratios 0.5 and 0.8, respectively)</li> <li>• FRAX without BMD MOF (ratio 0.7) but overestimated HF (ratio 1.4)</li> <li>• MOF AUC 0.66 without BMD, AUC 0.67–0.68 with BMD</li> <li>• Osteoporotic fracture AUC 0.665 HF AUC 0.710</li> </ul>	Fractures could only be assessed in long-term survivors, fractures were self-reported
Henry et al. 2011 [102]	600 Australian women, UK and US FRAX		Retrospective study
Cummins et al. 2011 [98]	246 postmenopausal women from Ireland and the UK with recent low-trauma fracture and 338 non-fracture controls (6 centers), UK FRAX		
Ertinger et al. 2012 [73]	5,891 White US men (MrOS cohort), US White FRAX	<ul style="list-style-type: none"> <li>• HF discrimination AUC 0.77 with BMD, 0.69 without BMD</li> <li>• MOF AUC 0.67 with BMD vs. 0.63 without BMD</li> <li>• Observed to predicted HF without BMD ratios 0.9–1.1, but risk with BMD was underestimated in the highest risk quintile</li> <li>• MOF risk slightly underestimated without BMD (ratio 0.7–0.9) and with BMD (ratio 0.7–1.1)</li> <li>• FRAX explained 7–17 % of all fractures and 41–60 % of HF</li> <li>• AUC without BMD 0.62 MOF, 0.64 HF</li> <li>• Estimated/observed ratios were 0.66 for MOF and 1.10 for HF</li> </ul>	Femoral neck BMD only available in a small subset
Byberg et al. 2012 [74]	5,921 Swedish men (ULSAM cohort), Swedish FRAX		Limited by the 3-year follow-up, lack of data on BMD or clinical vertebral fractures, no competing mortality risk adjustment
Gonzalez-Macias et al. 2012 [75]	5,201 Spanish women (ECOSAP Study cohort), Spanish FRAX		Fractures could only be assessed in long-term survivors, fractures were self-reported
Tebe Cordomi et al. [76]	1,231 Spanish women (CETIR database), Spanish FRAX	<ul style="list-style-type: none"> <li>• MOF AUC 0.61 with BMD</li> <li>• Observed/expected MOF ratio 3.9</li> </ul>	Small numbers in some subgroups (4 predicted vs. 9 observed obese women in the lowest category for FRAX probability with BMD)
Premaor et al. 2013 [72]	6,049 White US women (SOF cohort), US White FRAX	<ul style="list-style-type: none"> <li>• AUCs similar in obese and non-obese women (18.5 % were obese)</li> <li>• Calibration good in both groups for MOF using FRAX with BMD</li> <li>• HF risk underestimated (most marked among obese women in the lowest category for FRAX probability with BMD)</li> </ul>	
Garvan FRC			
Langsetmo et al. 2010 [94]	4,152 Canadian women and 1,606 Canadian men (CaMos cohort)	<ul style="list-style-type: none"> <li>• Low-trauma fracture Harrell's C 0.69 for women and 0.70 for men</li> <li>• HF Harrell's C 0.80 for women and 0.85 for men</li> </ul>	BMD was not available, fractures were self-reported
Sanbrook et al. 2011 [104]	19,586 women from 723 primary-care practices in 10 countries (GLOW Study)	<ul style="list-style-type: none"> <li>• HF AUC 0.76 without BMD</li> <li>• Osteoporotic fractures AUC 0.64 without BMD</li> <li>• HF AUC 0.67</li> </ul>	
Bolland et al. 2011 [105]			

**Table 2** (continued)

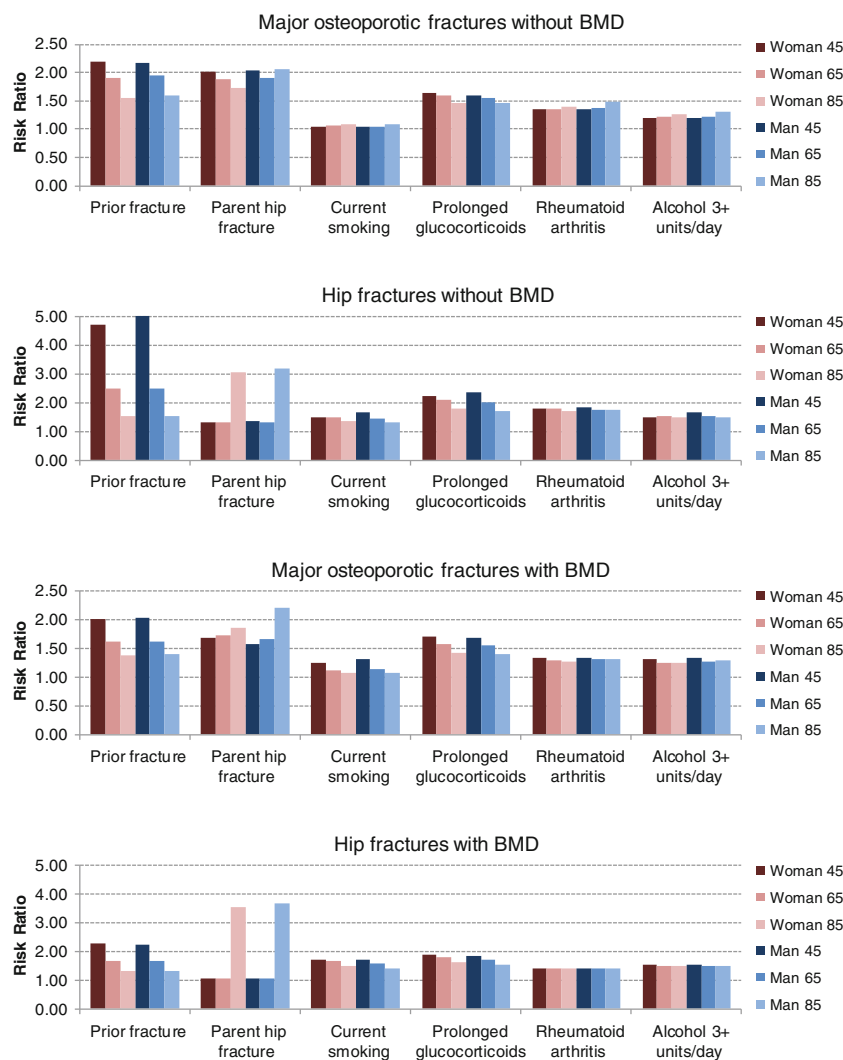
Author	Population	Performance	Comments
Henry et al. 2011 [102]	1,422 women (placebo-controlled trial of calcium supplements)	<ul style="list-style-type: none"> <li>• Osteoporotic fracture AUC 0.64</li> <li>• Predicted/observed osteoporotic fracture ratio 1.0, HF ratio 1.5</li> <li>• MOF AUC 0.66 without BMD, AUC 0.70 with BMD</li> </ul>	Fractures could only be assessed in long-term survivors, fractures were self-reported
QFractureScores	600 Australian women		
Collins et al. 2011 [97]	2.2 million adults from 364 general practices in UK, Wales, Scotland, and Northern Ireland (THIN database)	<ul style="list-style-type: none"> <li>• Osteoporotic fracture AUC 0.82 in women and 0.74 in men</li> <li>• HF AUC 0.89 in women and 0.86 in men</li> <li>• Calibration plots adhered closely to the line of identity</li> </ul>	Used earlier version of QFractureScores (Web version 1)
Cummins et al. 2011 [98]	246 postmenopausal women from Ireland and the UK with recent low-trauma fracture and 338 non-fracture controls (6 centers)	<ul style="list-style-type: none"> <li>• Osteoporotic fracture AUC 0.668, HF AUC 0.637</li> </ul>	Retrospective study, used earlier version of QFractureScores (Web version 1)

MOF major osteoporotic fracture, HF hip fracture, AUC area under the curve

Interactions between clinical risk factors are incorporated into the FRAX algorithm. As a result, there is no single weighting for a clinical risk factor in FRAX. However, Fig. 3 illustrates the relative importance of various risk factors and how they are affected by age and sex. In estimating hip fracture probability the contribution of prior fracture attenuates with advancing age whereas for parental hip fracture it increases with advancing age. Body mass index (BMI) is strongly associated with fracture risk when BMD is unknown but is no longer important when the effect of BMD has been considered. FRAX also adjusts for competing mortality, and the competing mortality approach used by FRAX is unique among the risk prediction models. Individuals may have equivalent hazards for fracture, but if they differ in terms of hazard for death then this will affect the 10-year fracture probability. For example, smoking is a risk factor for fracture but also increases the risk for death. Thus, the increased mortality associated with smoking reduces the importance of smoking as a risk factor for fracture. Figure 4 shows how older age and lower BMD, which are independent risk factors for death, affect 10-year fracture hip fracture probability in men and women. Ten-year major fracture probability tends to increase with age to peak around 80–85 years and then declines as the rising death hazard exceeds the rising fracture hazard. Men with extremely low BMD (femoral T-score  $-4.0$ ) are predicted to have such high mortality that this actually blunts the expected age-related increase.

Population-specific FRAX tools are customized to the fracture and mortality epidemiology in that specific region [57]. The initial release of FRAX in 2008 covered nine countries (including four ethnic calculators for the USA), while the most recent version includes 47 countries. In recognition of the large international variability in fracture and mortality rates [59], the FRAX tool is customized (calibrated) based upon the fracture and death epidemiology within that country. Some country tools have been updated, but the fundamental FRAX algorithm has not changed [60]. Minimal data requirements for constructing a FRAX tool are sex- and age-specific mortality and hip fracture rates (5-year subgroups). In many countries, such data are relatively easy to obtain. In contrast, non-hip fracture data considered by FRAX (clinical spine, distal forearm, and proximal humerus) are quite difficult to accurately collect at the population level. Where high-quality data are not available, the assumption is commonly made that the ratio of these non-hip to hip fracture rates is similar to that observed from population-based data obtained in Malmö, Sweden [61, 62]. As seen in Fig. 5, FRAX is sensitive to the large differences in osteoporotic fracture rates between different populations which show more than a 10-fold variation [59]. The US ethnic calculators provide sex- and age-specific predictions that are roughly parallel but differ quantitatively (White > Asian = Hispanic > Black). In contrast,

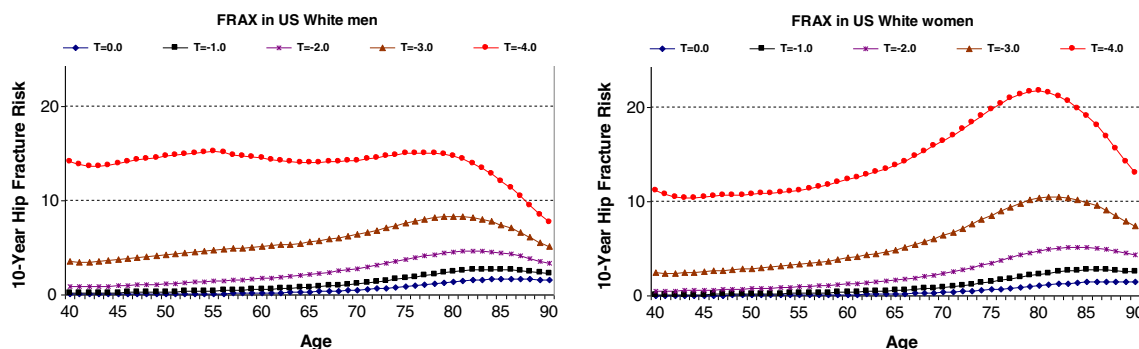
**Fig. 3** Relative importance of individual clinical risk factors in FRAX for selected combinations of age and sex (comparator—no additional risk factors, US White FRAX tool, BMI 25 kg/m<sup>2</sup>, and T-score −2.5)



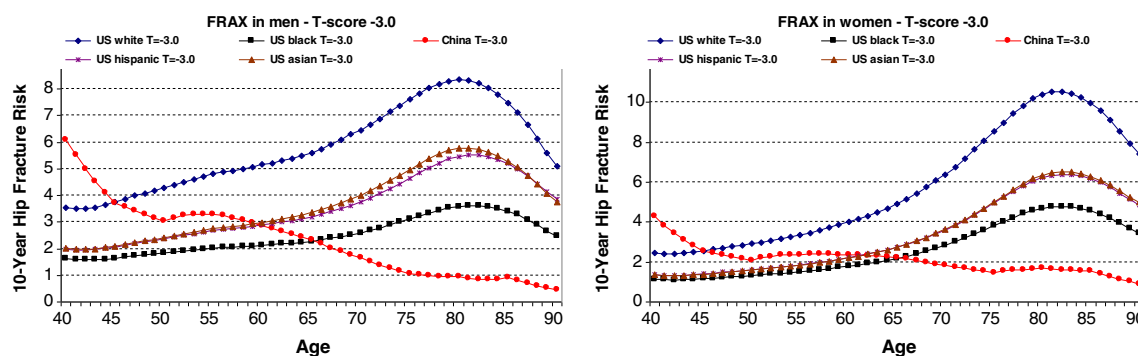
the Chinese calculator has a very different configuration and also diverges from the US Asian calculator with 10-year fracture probabilities that decline with age because of a larger competing mortality effect in China.

Fracture discrimination was assessed in nine primary derivation cohorts (46,340 subjects with 189,852 person years of follow-up) and in 11 additional validation cohorts (230,486 persons with 1,208,528 person years of follow-up)

[63]. Risk stratification with FRAX including BMD was superior to FRAX without BMD or BMD alone. In the primary derivation cohorts, the gradient of risk for hip fracture increased from 1.84 to 2.91 [area under the curve (AUC) from 0.67 to 0.78] with the inclusion of BMD, and for other osteoporotic fractures it increased from 1.55 to 1.61 (AUC increased from 0.62 to 0.63) with the inclusion of BMD. For hip fracture prediction in the validation



**Fig. 4** Effect of age and femoral neck T-score on FRAX major osteoporotic fracture probability (assumes BMI 25 kg/m<sup>2</sup> and no other risk factors)



**Fig. 5** Effect of ethnicity on FRAX hip fracture probability (assumes femoral neck T-score  $-3.0$ , BMI  $25 \text{ kg/m}^2$ , and no other risk factors)

cohorts, averaged gradient of risk (1.83 without BMD and 2.52 with BMD) and AUC (0.66 without BMD and 0.74 with BMD) were similar to the derivation cohorts. For other osteoporotic fractures, the validation cohorts again gave only marginally lower gradients of risk (1.53 without BMD and 1.57 with BMD) and AUC (0.60 without BMD and 0.62 with BMD). A strong but opposite age interaction was seen for hip fracture (lower gradient of risk for older age) compared with other osteoporotic fractures (increasing gradient of risk with advancing age) whether this was based upon BMD only, clinical risk factors alone, or clinical risk factors with BMD.

A limited number of studies have performed independent assessments of FRAX to predict subsequent fracture, but differ widely in sample size, methodology (particularly incorporation of competing mortality risk), and assessments (discrimination versus calibration). These and other methodologic factors may affect interpretation of validation studies [40].

In 2010, Sornay-Rendu et al. [64] examined 867 French women age 40 years and over from the OFELY (Os des Femmes de Lyon) cohort which included 95 incident major osteoporotic and 17 incident hip fractures. Predicted probabilities of fracture were considerably greater in women with than women without incident fractures. The observed incidence of major osteoporotic fractures was found to be higher than the predicted probability, but analysis did not account for the effect of competing mortality. AUC for fracture discrimination for major osteoporotic fractures was 0.75 (95 % confidence interval, CI, 0.71–0.79) without BMD and 0.78 (95 % CI 0.72–0.82) with BMD]. FRAX was not significantly better than femoral neck BMD alone [AUC 0.74 (95 % CI 0.71–0.77)] or femoral neck BMD and age [AUC 0.79 (95 % CI 0.75–0.81)]. Trémollières et al. [65] examined a separate group of 2,651 French women from the MENOS (Menopause et Os) cohort who sustained 145 major osteoporotic fractures (13 hip fractures) during the follow-up period. Once again, fracture discrimination was good for major osteoporotic fractures (AUC 0.63, 95 %

CI 0.56–0.69) which was not better than BMD alone (AUC 0.66, 95 % CI 0.60–0.73).

A FRAX tool for Canada was developed, based upon 2005 national hip fracture and 2004 mortality data [66]. The accuracy of the fracture predictions was assessed in two large, independent cohorts: the Canadian Multicentre Osteoporosis Study (one of the population-based FRAX<sup>®</sup> derivation cohorts) and the Manitoba Bone Density Program (a long-term observational clinical cohort that is independent of the FRAX<sup>®</sup> derivation cohorts) [67, 68]. Analyses for the Manitoba BMD cohort (36,730 women and 2,873 men) and CaMos cohort (4,778 women and 1,919 men) were similar and showed that the Canadian FRAX tool generated fracture risk predictions that were consistent with observed fracture rates across a wide range of risk categories in both clinical and average populations. In the Manitoba cohort [67], hip fracture discrimination from receiver operating curve analysis was 0.830 (95 % CI 0.815–0.846) and for major osteoporosis-related fractures was 0.694 (95 % CI 0.684–0.705), results comparable to values reported in the derivation and validation of the cohorts studied by the WHO Collaborating Centre [63]. Fracture discrimination using FRAX with BMD was better than FRAX without BMD (hip fracture AUC 0.793, major osteoporosis fracture AUC 0.663) or femoral neck BMD alone (hip fracture AUC 0.801, major osteoporosis fracture AUC 0.679). The 10-year estimate for hip fractures in all women was 2.7 % (95 % CI 2.1–3.4 %) with a predicted value of 2.8 % for FRAX with BMD, and in men the observed risk was 3.5 % (95 % CI 0.8–6.2 %) with a predicted value of 2.9 %. The 10-year estimate of osteoporotic fracture risk for all women was 12.0 % (95 % CI 10.8–13.4 %) with a predicted value of 11.1 % for FRAX with BMD, and in men, the observed risk was 10.7 % (95 % CI 6.6–14.9 %) with a predicted value of 8.4 %. Discrepancies were observed within some subgroups but generally were small. For CaMos [68], results were similar with the FRAX estimates using BMD and CRFs superior to BMD alone or CRFs alone for both major osteoporotic fractures and hip fractures. For major osteoporotic fractures, FRAX with BMD gave AUC 0.69 (95 % CI 0.67–0.71) vs.

FRAX without BMD 0.66 (95 % CI 0.63–0.68) and femoral neck T-score alone 0.66 (95 % CI 0.64–0.69). For hip fractures, FRAX with BMD gave AUC 0.80 (95 % CI 0.77–0.83) vs. FRAX without BMD 0.77 (95 % CI 0.73–0.80). Mean overall 10-year FRAX probability with BMD for major osteoporotic fractures was not significantly different from the observed value in men [predicted 5.4 % vs. observed 6.4 % (95 % CI 5.2–7.5 %)] and only slightly lower in women [predicted 10.8 % vs. observed 12.0 % (95 % CI 11.0–12.9 %)]. FRAX was well calibrated for hip fracture assessment in women [predicted 2.7 % vs. observed 2.7 % (95 % CI 2.2–3.2 %)] but underestimated risk in men [predicted 1.3 % vs. observed 2.4 % (95 % CI 1.7–3.1 %)].

The importance of correct calibration was noted when the UK FRAX tool was used to assess fracture risk in 501 Polish women referred for BMD testing (convenience sample) [69]. Self-reported incident fractures 9–12 years later were assessed by telephone interview. The observed/expected ratio for fracture was 1.79 (95 % CI 1.44–2.21) without BMD and 1.94 (95 % CI 1.45–2.54) with BMD indicating that the UK model significantly overestimated fracture risk in Polish women. Results could be biased as fractures could only be assessed in long-term survivors and death competes with fractures. Average life expectancy in Poland is 3.7 years less than in the UK, and use of the UK FRAX model would be expected to overestimate fracture probability in a Polish setting.

A small Japanese study (43 self-reported major osteoporotic fractures and four hip fractures) was reported by Tamaki et al. [70] using the Japanese Population-Based Osteoporosis Study (JPOS) cohort. The number of observed major osteoporotic or hip fracture events were found to be consistent with FRAX predictions, and again there was significant stratification in fracture risk (AUC major osteoporotic fracture 0.67, 95 % CI 0.59–0.75 without BMD; 0.69, 95 % CI 0.61–0.76 with BMD).

Rubin et al. [71] performed a registry linkage study using baseline questionnaire data from 3,636 Danish women with FRAX hip fracture probabilities calculated from the Swedish tool. Predicted and observed risks estimates incorporated adjustment for 10-year survival rates. The predicted 10-year hip fracture risk was 7.6 % overall with observed risk also 7.6 %, ranging from 0.3 % at the age of 41–50 years (observed risk 0.4 %) to 25.0 % at the age of 81–90 years (observed risk 24.0 %) (*p* value was non-significant, overall and by age decade). For the closely related Scandinavian countries of Sweden and Denmark, a single FRAX tool may be sufficient.

Premaor et al. [72] recently examined the question of whether FRAX was applicable to obese older women using 6,049 white women from the US Study of Osteoporotic Fractures (SOF) cohort. Fracture discrimination from AUC was similar in obese and non-obese women. Calibration was good in both groups for prediction of major osteoporotic

fractures using FRAX with BMD, but hip fracture risk was found to be underestimated (most marked among obese women in the lowest category for FRAX probability with BMD, four predicted vs. nine observed).

Ettinger et al. [73] examined 5,891 men age 65 years and older (374 with incident to major osteoporotic fractures, 161 incident hip fractures). Hip fracture discrimination (AUC 0.77 with BMD vs. 0.69 without BMD) was better than for major osteoporotic fractures (AUC 0.67 with BMD vs. 0.63 without BMD). Inclusion of BMD significantly improved the overall net reclassification index for major osteoporotic fractures and hip fractures. Observed to predicted fracture ratios according to probability quintiles showed good calibration for hip fracture prediction without BMD (ratios 0.9–1.1), but hip fracture risk was significantly underestimated in the highest risk quintile when BMD was used in the calculation. Conversely, major osteoporotic fracture risk was underestimated without BMD (predicted ratio 0.7–0.9) and also when BMD was used (predicted ratio 0.7–1.1).

Byberg et al. [74] examined 5,921 men age 50 years and older from Sweden in the Uppsala Longitudinal Study of Adult Men (ULSAM) in which 585 individuals sustained fracture (189 with hip fractures). FRAX explained 7–17 % of all fractures and 41–60 % of hip fractures. Including additional comorbidity, medication and behavioral factors improved overall fracture prediction. Femoral neck BMD was only available in a small subset of those age 82 years and older.

Gonzalez-Macias et al. [75] examined 5,201 women age 65 years and older in a 3-year prospective follow-up study in Spain (201 with major osteoporotic fractures, 50 with incident hip fractures) using data from the ECOSAP (Ecografía Osea en Atención Primaria) Study. AUC for FRAX without BMD was 0.62 for major osteoporotic fractures and 0.64 for hip fractures. Estimated to observed fracture ratios of 0.66 and 1.10, respectively, were likely influenced by the limited duration of the follow-up (i.e., 3 years), lack of data on clinical vertebral fractures, and lack of competing mortality risk adjustment. Another Spanish study from Tebe Cordomi et al. [76] conducted a retrospective cohort study of 1,231 women aged 40–90 years (222 with at least one self-reported fracture after baseline assessment). AUC for major osteoporotic fracture was 0.61 (95 % CI 0.57–0.65) estimated with BMD. The number of observed fractures was 3.9 times higher than the expected number (95 % CI 3.4–4.5). Fractures were self-reported at a follow-up survey at least 10 years after baseline assessment, but there was a large rate of non-response/non-participation (855 of 2,086). Fractures could only be assessed in long-term survivors, and excluding individuals who died prior to 10 years could bias results leading to overestimation in calibration.

The FRAX tool been endorsed and integrated into clinical practice guidelines by several national bodies [77–85]. Given the paradigm shift introduced with FRAX, many questions have arisen regarding its use in specific and



challenging circumstances. A set of joint positions were developed by the International Society of Clinical Densitometry (ISCD) and International Osteoporosis Foundation (IOF) in 2010, and 28 recommendations were eventually jointly endorsed [86, 87].

#### FRAX variants

Simplified models may have a role in settings where it is not feasible to use the full FRAX tool or where the clinical risk factor data cannot be collected.

The Foundation for Research and Education (FORE) fracture risk calculator (FRC) provides a simplified construct of FRAX for the US using the same clinical risk factors with fixed relative risk, no interactions, and no competing mortality adjustment ([riskcalculator.fore.org](http://riskcalculator.fore.org)) [88]. Performance was assessed in 94,489 women age 50 years and older with BMD measured at baseline from the Kaiser Permanente Northern California database (1,579 with incident hip fractures during mean 6.6 years of follow-up). The AUC was 0.83 (95 % CI 0.82–0.84) without BMD and 0.85 (95 % CI 0.84–0.86) with BMD. The FRC underestimated observed fracture risk by 30–40 %, possibly due to the high rate of hormone therapy (40 % of the cohort). The FRC was updated in 2012 (version 2.0) to include vertebral fracture as a separate risk factor, spine BMD (together with the usual femoral neck BMD), and consideration of oral glucocorticoid dose.

The original version of the FRC has also been evaluated in 5,893 men from the MrOS cohort [73]. Ratios of observed predicted probabilities were close to unity indicating good calibration. Hip fracture discrimination was higher for hip fracture (AUC 0.71 without BMD and 0.79 with BMD) than for major osteoporotic fracture (AUC 0.6 without BMD and 0.70 with BMD). Estimated 10-year fracture probabilities were performed with the Kaplan–Meier method which does not adjust for competing mortality. BMD improved overall performance as assessed with net reclassification indices (8.5 % for hip and 4.0 % for major osteoporotic fracture). A large divergence in predicted hip fracture risk can be seen for FRC and FRAX with advancing age in women and men, with a 3-fold differential by age 85 years.

An alternative approach of FRAX simplification was developed by the Canadian Association of Radiologists and Osteoporosis Canada (called the CAROC tool) based upon a semi-quantitative approach that estimates 10-year major osteoporotic fracture risk as low (less than 10 %), moderate (10–20 %), and high (greater than 20 %) [89, 90]. Using the Canadian FRAX tool [66], sex-specific cutoffs based upon age and femoral neck T-score were derived that would assign individuals to the low, moderate, and high risk categories assuming no additional clinical risk factors [91]. A further refinement was made to ensure that an osteoporotic T-score

would not be classified as low risk ([www.osteoporosis.ca/multimedia/pdf/CAROC.pdf](http://www.osteoporosis.ca/multimedia/pdf/CAROC.pdf)). Under this paradigm, fragility fracture after age 40 (excluding craniofacial, hands, feet, and ankles) and recent prolonged systemic glucocorticoid use (prednisone equivalent 7.5 mg daily or greater for at least 3 months in the prior year) affect the basal fracture risk from sex, age, and femoral neck T-score. Each of these two clinical risk factors increases the risk category (i.e., from low to moderate, or moderate to high). Using the same large Canadian cohorts from the FRAX validation work [67, 68], the 10-year fracture outcomes were shown to agree with the assigned risk category indicating good calibration [91]. There was also a high level of agreement between the simplified and standard FRAX categories (88–89 % agreement) with low rates of reclassification under FRAX.

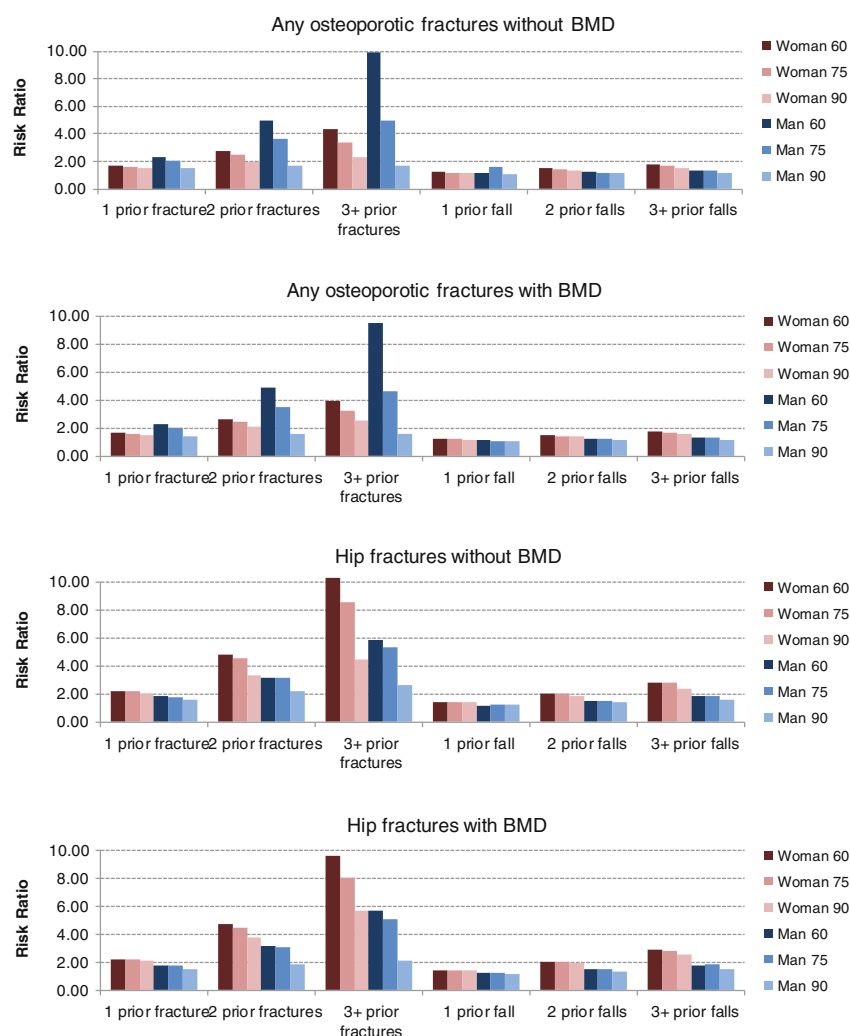
#### Garvan fracture risk calculator

([www.garvan.org.au/bone-fracture-risk](http://www.garvan.org.au/bone-fracture-risk))

The Dubbo Osteoporosis Epidemiology Study (DOES) was initiated in 1989 and involves follow-up of over 3,500 participants. Based upon 426 clinical fractures in women (96 hip) and 149 clinical fractures in men (31 hip) excluding digits, 5- and 10-year fracture probability nomograms were constructed [92, 93]. Inputs include age, sex, femoral neck BMD (optional), body weight, history of prior fractures after age 50 years (none, 0, 1, 2, and 3 or more) and history of falls in the previous 12 months (none, 0, 1, 2, and 3 or more). Risk factors that are relatively uncommon in the general population (e.g., glucocorticoid use and specific medical conditions) are not included. There is an assumption of additivity regarding number of fractures and number of falls. If femoral neck BMD is not available, then weight is used as a surrogate. The model has only been calibrated for the Australian population. It does not include an explicit competing mortality risk adjustment. Some sense of the relative importance of various risk factors and how they are affected by age and sex is illustrated in Fig. 6. The apparent declining risk ratios in older individuals, most dramatically seen for multiple prior fractures, reflects risk “saturation” whereby probability can approach but never exceed 100 %: the same increase in fractures *odds* ( $=\text{probability}/[1 - \text{probability}]$ ) gives an attenuated increase in fracture *probability* ( $=\text{odds}/[1 + \text{odds}]$ ) in those with higher basal risk scores. This effect is particularly important for strong risk factors (risk ratios  $>2$ ) and when basal risk exceeds 10–20 %. For example, a 90-year-old man with T-score  $-2.5$  has a 64 % risk for any fracture over the next 10 years, but when combined with two or more prior fractures his risk is  $>99$  %.

The Garvan algorithm has been independently evaluated in the Canadian population (4,152 women and 1,606 men age 55–95 years at baseline) with 8.6 years of follow-up (699 low-trauma fractures including 97 hip fractures) [94]. Fracture discrimination and calibration were found to be generally good

**Fig. 6** Relative importance of individual clinical risk factors in the Garvan Fracture Risk Calculator for selected combinations of age and sex (comparator—no additional risk factors, weight 60 kg for women and 70 kg for men, and T-score  $-2.5$ )



in both women and men. For low-trauma fractures, the concordance between predicted risk and fracture events (Harrell's  $C$  which is similar to AUC) was 0.69 among women and 0.70 among men. For hip fractures, the concordance was 0.80 among women and 0.85 among men. Observed 10-year low-trauma fracture risk agreed with the predicted risk for all risk subgroups except in the highest risk quintile in men and women (observed risk lower than predicted). Observed 10-year hip fracture risk agreed with the predicted risk for all risk subgroups except in the highest quintile for women (observed risk lower than predicted).

QFractureScores ([www.qfracture.org](http://www.qfracture.org))

The largest prospective database for osteoporotic fracture prediction has been from England and Wales using 357 general practices for derivation and 178 practices for validation in the initial analysis (QResearch Database) [95]. This provided more than 1 million women and more than 1 million men age 30–85 years in the derivation cohort with

24,350 incident osteoporotic fractures in women (9,302 hip fractures) and 7,934 osteoporotic fractures in men (5,424 hip fractures). The risk calculator includes numerous clinical risk factors but does not include BMD. It provides outputs of any osteoporotic fracture (hip, wrist, or spine) and hip fracture over a user selected follow-up period from 1 year to 10 years. The QFractureScores algorithm was updated in 2012, with inclusion of a number of new risk factors, removal of several others, and inclusion of humerus fractures as one of the osteoporotic fractures [96]. It provides calibration for 10 different ethnic origins. In addition to age, sex, and ethnicity, the algorithm includes smoking status (four levels), alcohol consumption (five levels), diabetes (type 1 and type 2), previous fracture, parental osteoporosis or hip fracture, living in a nursing or care home, history of falls, dementia, cancer, asthma/COPD, cardiovascular disease, chronic liver disease, chronic kidney disease, Parkinson's disease, rheumatoid arthritis/SLE, malabsorption, endocrine problems, epilepsy or anticonvulsant use, antidepressant use, steroid use, HRT use, height, and weight. The 2012 version provided a further

improvement in AUC for osteoporotic fracture prediction (AUC 0.79 in women and 0.71 in men) and hip fracture prediction (AUC 0.89 in women and 0.88 in men).

An independent validation study was performed using 364 general practices from the THIN database (2.2 million adults aged 30–85 years with 25,208 osteoporotic and 12,188 hip fractures) [97]. The validation cohort gave AUC discrimination for osteoporotic fracture of 0.82 in women and 0.74 in men, and for hip fracture of 0.89 in women and 0.86 in men. Calibration plots adhered closely to the line of identity. QFractureScores explained 63 % of the variation in hip fracture risk in women and 60 % of the variation in men (49 % and 38 % for osteoporotic fracture risk).

A small retrospective comparison of FRAX and QFractureScores was conducted in 246 postmenopausal women aged 50–85 years from six centers in Ireland and the UK with recent low-trauma fracture and 338 non-fracture control women [98]. AUCs for fracture discrimination were similar in QFractureScores and FRAX (0.668 vs. 0.665) and also for hip fractures (0.637 vs. 0.710). The difference in AUC measures between these two studies is striking. The broad age range used in the initial derivation/validation work may have resulted in inflated performance measures since osteoporotic fractures are unlikely before age 50. Additional assessments of QFractureScores in older women and men are needed.

#### Other models

There are a number of other fracture risk calculators and risk assessment tools that have been developed, but are awaiting independent validation. These include but are not limited to the Fracture INDEX (Study of Osteoporotic Fractures), an 11-factor model for 5-year hip fracture risk assessment (Women's Health Initiative), and FRISK (Australia Geelong Osteoporosis Study), a simple four-item risk score (Rotterdam Study and Longitudinal Aging Study Amsterdam) [99–103].

#### Comparisons

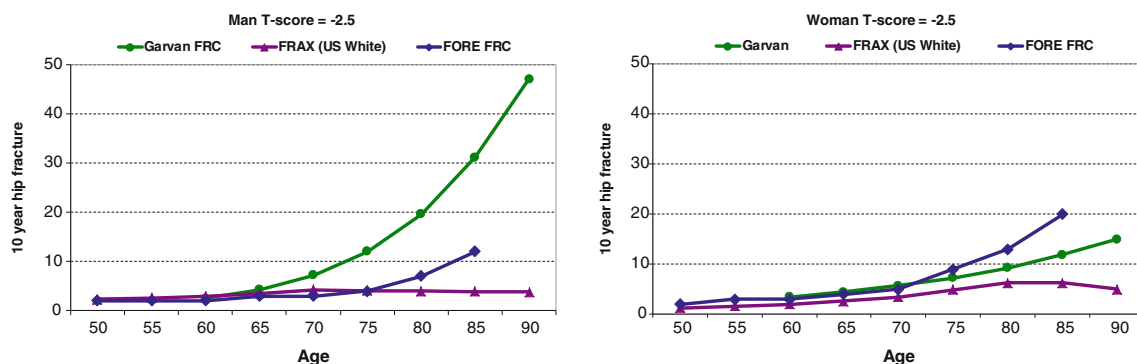
A direct comparison of 10-year hip fracture predictions for the various risk assessment tools discussed above is presented in Fig. 7 (Garvan FRC, FRAX US White, and FORE FRC with BMD set to a T-score −2.5) and Fig. 8 (Garvan FRC, FRAX US White, and QFractureScores without BMD). FRAX (without BMD) and QFractureScores track each other closely across the age spectrum, whereas FRAX gives lower estimates than Garvan FRC or FORE FRAC in older individuals, especially older men. The age- and sex-dependent divergence reflect the effect of computing mortality which is only explicitly represented in the FRAX formulation. For example, the FORE FRC (which is a FRAX variant without interactions

or adjustment for competing mortality risk) is similar to FRAX at age 65 years, but by age 85 years gives values that are 3-fold greater. Divergence becomes particularly extreme at age 90 years with T-score −2.5 for Garvan FRC versus FRAX with a 3-fold difference in women and over a 10-fold difference in men. One corollary of these differences is that these tools are not interchangeable in their clinical application: different tools and guidelines could identify very different numbers of individuals for treatment based upon the same intervention cutoff (e.g., 3 % 10-year hip fracture risk).

A limited number of studies have performed “head to head” comparisons of these fracture prediction tools. Using 2-year self-report fracture data, Sambrook et al. [104] reported on 19,586 women from the GLOW cohort age 60 years or older who are not receiving osteoporosis treatment (880 women reported incident fractures including 69 hip fractures and 468 “major fractures” and 583 “osteoporotic fractures”). For hip fracture prediction, the AUC was 0.78 for FRAX with BMD, 0.76 for Garvan FRC, and 0.78 for age and prior fracture alone. For major fractures, the AUC was 0.61 for FRAX and for osteoporotic fractures was 0.64 for Garvan FRC, neither of which was better than age and prior fracture alone.

Bolland et al. [105] compared FRAX (New Zealand) and the Garvan FRC in 1,422 healthy New Zealand women, mean age 74 years, participating in a 5-year randomized, placebo-controlled trial of calcium supplements. Women were contacted average 8.8 years post-enrollment about fracture events (self-reported). No follow-up information was available for 248 women who died or for a further 53 women who could not be contacted. Hip fracture discrimination was similar for FRAX with BMD (0.70, 95 % CI 0.64–0.77), FRAX without BMD (0.69, 95 % CI 0.63–0.76), and Garvan FRC (0.67, 95 % CI 0.60–0.75). For FRAX-defined osteoporotic fractures, Garvan-defined osteoporotic fractures, and all fractures, the AUCs were slightly lower (range 0.60–0.64). The Garvan FRC was well calibrated for Garvan-defined osteoporotic fractures (predicted/observed ratio 1.0) but overestimated hip fractures (ratio 1.5). FRAX with BMD underestimated FRAX-defined major osteoporotic fracture risk and hip fracture risk (ratios 0.5 and 0.8, respectively) and gave divergent results when used without BMD (ratios 0.7 for major osteoporotic fractures and 1.4 for hip fractures). Goodness-of-fit testing showed significant deviation from a perfectly calibrated model ( $p < 0.01$ ) in all cases except for hip fracture prediction from FRAX without BMD. Neither FRAX nor Garvan FRC provided better discrimination than age and BMD alone. Results could be biased as fractures could only be assessed in long-term survivors.

Fracture discrimination with FRAX and Garvan was also compared in a report from Henry et al. [102] in 600 Australian women age 60 years and older using FRAX tools for the UK and US. AUCs were similar for major osteoporotic fracture prediction without BMD (AUC 0.66) and with BMD (AUC 0.67–0.70).



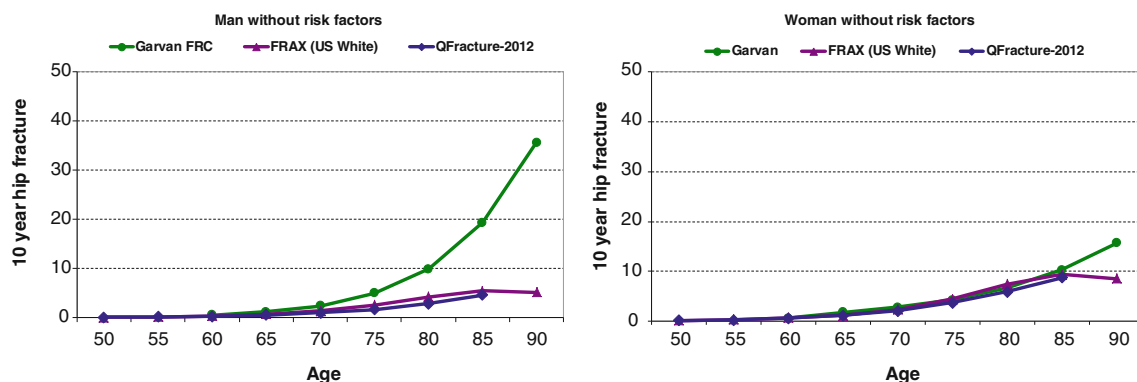
**Fig. 7** Comparison of 10-year hip fracture predictions for risk prediction tools with BMD (assumes no additional risk factors, BMI 25 kg/m<sup>2</sup>, and T-score -2.5)

The only independent evaluation of QFractureScores (Web version 1) to date has been a small retrospective comparison with FRAX in 246 postmenopausal women aged 50–85 years from six centers in Ireland and the UK with recent low-trauma fracture compared with 338 non-fracture control women [98]. AUCs for osteoporotic fracture discrimination were similar in QFractureScores and FRAX (0.668 vs. 0.665) and also for hip fractures (0.637 vs. 0.710). The difference in AUC measures compared with the initial reports from the QFractureScores authors are striking. This may reflect the broad age range used in the initial derivation/validation work. Whether similar results would be seen with the 2012 version of QFractureScores is uncertain.

### Clinical implications

A risk prediction tool on its own does not necessarily alter patient management. A recent review of updated guidelines around the world found a diversity of approaches [106]. Some guidelines have embraced fracture risk as the preferred decision making approach, others are still largely dictated by BMD, whereas others are a hybrid. The National Osteoporosis Foundation (NOF) Guidelines are an example of the latter. Treatment is

recommended for individuals with an osteoporotic T-score, clinical osteoporosis based upon low-trauma spine or hip fracture, with a secondary role for fracture risk prediction in individuals with low bone mass (osteopenia) where major osteoporotic fracture risk exceeds 20 % or hip fracture risk exceeds 3 % under the US FRAX tools [77]. The Osteoporosis Canada guidelines recommend treatment initiation based upon clinical osteoporosis (hip fracture, spine fracture, or multiple fragility fractures) or where major osteoporotic fracture probability exceeds 20 % under FRAX or the simplified CAROC tool [85]. The National Osteoporosis Guidelines Group (NOGG), working in collaboration with many other societies from the UK, recommends that treatment be considered in women with a prior fragility fracture (BMD measurement is optional) or when major osteoporotic fracture probability with FRAX exceeds an age-adapted treatment threshold. The intervention threshold at each age is set at a risk equivalent to that associated with a prior fracture, resulting in a lower threshold in younger individuals and a higher threshold in older individuals [107]. The NOGG approach has fully embraced fracture risk in guiding therapy and restricts BMD to individuals whose fracture risk is close to the intervention threshold: individuals with fracture risk well below the intervention or well above the intervention threshold are not recommended for BMD testing. A similar approach has been



**Fig. 8** Comparison of 10-year hip fracture predictions for risk prediction tools without BMD (assumes no additional risk factors, BMI 25 kg/m<sup>2</sup>, and weight 60 kg for women and 70 kg for men)



advocated for the European setting by the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO) and International Osteoporosis Foundation (IOF) [108]. Judging from the diversity of approaches, “once size does not fit all” when it comes to how risk prediction is currently integrated into clinical practice guidelines.

## Future directions

It is self-evident that no risk prediction model can include all possible risk factors for fracture: even if such a tool could be created, it would be impossibly detailed and unwieldy. More than 80 secondary causes of osteoporosis were specified in the US Surgeon General’s report on osteoporosis [109]. Moreover, not all risk factors can be easily or reliably measured. Finding the “right” balance between complexity and simplicity is analogous to the tradeoff between sensitivity and specificity. There is no perfect cutoff, and needs will vary depending upon clinical context and data availability. In primary care where there is a low prevalence of high-risk medication use, it may be quite reasonable to omit glucocorticoids and antineoplastic medications. However, in a rheumatology, organ transplant, or breast cancer clinic, this would clearly be unacceptable.

Several investigators have noted that very simple prediction models (e.g., age, BMD, and prior fracture) can discriminate fractures as well as more complex models such as FRAX [70, 105, 110–112]. This is not altogether surprising given the insensitivity of general measures of test performance (e.g., ROC) to detect incremental improvement in risk classification from additional risk factors [39, 40]. For example, the prevalence of high-dose glucocorticoid use in the general population is very low (~1 %) and excluding this risk factor from the assessment is barely noticeable at the population level despite the importance it has for the individual risk prediction.

Even where a clinical risk factor is omitted from the model, clinical judgment must be brought to bear with a qualitative consideration of the importance of the missing information. BMD is a requirement for some but not all risk prediction models, and when included is typically based upon the femoral neck. Attempts to incorporate additional sites of measurement (e.g., lumbar spine) have been proposed [113]. Inclusion of additional skeletal measures (e.g., bone turnover markers, trabecular bone score) adds to the complexity with uncertain benefit in terms of the improvement in overall patient outcomes [86, 87]. How to best reconcile these competing and divergent forces remains unclear. Similarly, fundamental design consideration such as the importance of competing mortality warrant further discussion. As noted in this review, competing mortality can dramatically impact on the risk prediction measurement, particularly among population sub-groups with high mortality risk such as the elderly, men, those with serious comorbidities, and those at highest risk for fracture [47].

Finally, the ultimate question is whether patients selected for treatment based upon a risk prediction model actually benefit in terms of fracture prevention. To date, some retrospective analyses have suggested that this may be the case though there is ongoing controversy regarding the utility of pharmacologic intervention in individuals without significant reductions in BMD [114]. Whether formal clinical trials will be undertaken to address this question is uncertain.

## Conclusions

Prognostic models for fracture risk assessment can guide clinicians and individuals in understanding the risk of having an osteoporosis-related fracture and inform their decision making to mitigate these risks. However, to be useful for these purposes, risk estimates must be derived from valid and accurate models. Validity must be established in different populations, as well as over time. Competing health events must be considered to produce unbiased estimates. Model development should follow a systematic and rigorous methodology around variable selection, model fit evaluation, performance evaluation, and internal and external validation. As well, authors who report the results of fracture risk prediction models must ensure complete and accurate reporting of all stages of model development to ensure replicability of results, so that future research can build on current developments in a consistent manner [115]. Finally, accurate risk prediction is only an academic exercise unless practitioners know how to use this information. Careful consideration must always be given to the integration of risk prediction tools into clinical practice guidelines to support better clinical decision making and improved patient outcomes. The statistician George E.P. Box noted “All models are wrong but some are useful”. Every risk prediction tool has limitations but it can still be valuable tools to complement clinical judgment if these limitations are understood.

**Sources of support** L.M.L. is supported by a Manitoba Research Chair.

**Conflicts of interest** William Leslie: Speaker bureau: Amgen, Novartis. Research grants: Novartis, Amgen, Genzyme. Advisory boards: Novartis, Amgen. Lisa Lix: Research grant: Amgen.

## References

1. Stone KL, Seeley DG, Lui LY et al (2003) BMD at multiple sites and risk of fracture of multiple types: long-term results from the Study of Osteoporotic Fractures. *J Bone Miner Res* 18:1947–1954
2. Wiktorowicz ME, Goeree R, Papaioannou A et al (2001) Economic implications of hip fracture: health service use, institutional care and cost in Canada. *Osteoporos Int* 12:271–278
3. Papaioannou A, Adachi JD, Parkinson W et al (2001) Lengthy hospitalization associated with vertebral fractures despite control for comorbid conditions. *Osteoporos Int* 12:870–874



4. Center JR, Nguyen TV, Schneider D et al (1999) Mortality after all major types of osteoporotic fracture in men and women: an observational study. *Lancet* 353:878–882
5. Johnell O, Kanis JA, Oden A et al (2004) Mortality after osteoporotic fractures. *Osteoporos Int* 15:38–42
6. Ioannidis G, Papaioannou A, Hopman WM et al (2009) Relation between fractures and mortality: results from the Canadian Multicentre Osteoporosis Study. *CMAJ* 181:265–271
7. Adachi JD, Ioannidis G, Berger C et al (2001) The influence of osteoporotic fractures on health-related quality of life in community-dwelling men and women across Canada. *Osteoporos Int* 12:903–908
8. Hallberg I, Rosenqvist AM, Kartous L et al (2004) Health-related quality of life after osteoporotic fractures. *Osteoporos Int* 15:834–841
9. Johnell O, Kanis JA (2006) An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos Int* 17:1726–1733
10. Melton LJ III (2003) Epidemiology worldwide. *Endocrinol Metab Clin North Am* 32:1–13
11. Kanis JA, Melton LJ III, Christiansen C et al (1994) The diagnosis of osteoporosis. *J Bone Miner Res* 9:1137–1141
12. Looker AC, Wahner HW, Dunn WL et al (1998) Updated data on proximal femur bone mineral levels of US adults. *Osteoporos Int* 8:468–489
13. Kanis JA, McCloskey EV, Johansson H et al (2008) A reference standard for the description of osteoporosis. *Bone* 42:467–475
14. Johnell O, Kanis JA, Oden A et al (2005) Predictive value of BMD for hip and other fractures. *J Bone Miner Res* 20:1185–1194
15. Marshall D, Johnell O, Wedel H (1996) Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *BMJ* 312:1254–1259
16. Cranney A, Jamal SA, Tsang JF et al (2007) Low bone mineral density and fracture burden in postmenopausal women. *CMAJ* 177:575–580
17. Moons KG, Kengne AP, Woodward M et al (2012) Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 98:683–690
18. Moons KG, Kengne AP, Grobbee DE et al (2012) Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 98:691–698
19. Lloyd-Jones DM (2010) Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation* 121:1768–1777
20. Nguyen TV, Eisman JA (2013) Genetic profiling and individualized assessment of fracture risk. *Nat Rev Endocrinol* 9:153–161
21. Janssens AC, van Duijn CM (2009) Genome-based prediction of common diseases: methodological considerations for future research. *Genome Med* 1:20
22. Callas PW, Pastides H, Hosmer DW (1998) Empirical comparisons of proportional hazards, Poisson, and logistic regression modeling of occupational cohort data. *Am J Ind Med* 33:33–47
23. Datta S, DePadilla LM (2006) Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. *Stat Meth* 3:79–92
24. Abu-Hanna A, de Keizer N (2003) Integrating classification trees with local logistic regression in Intensive Care prognosis. *Artif Intell Med* 29:5–23
25. Peduzzi P, Concato J, Kemper E et al (1996) A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 49:1373–1379
26. Steyerberg EW, Eijkemans MJ, Habbema JD (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 52:935–942
27. Li W, Kornak J, Harris TB et al. (2009) Hip fracture risk estimation based on principal component analysis of QCT atlas: a preliminary study. *Proc. SPIE* 7262, Medical Imaging 2009: Biomedical Applications in Molecular, Structural, and Functional Imaging 7262:doi:10.1117/12.811743
28. Moons KG, Donders AR, Steyerberg EW et al (2004) Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* 57:1262–1270
29. Steyerberg EW, Eijkemans MJ, Harrell FE Jr et al (2000) Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 19:1059–1079
30. Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. *Biomet* 78:691–692
31. Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*:1–3
32. Blattenberger G (1985) Separating the Brier score into calibration and refinement components: a graphical exposition. *Am Stat* 26–32
33. Spiegelhalter DJ (1986) Probabilistic prediction in patient management and clinical trials. *Stat Med* 5:421–433
34. Ikeda M, Ishigaki T, Yamauchi K (2002) Relationship between Brier score and area under the binomial ROC curve. *Comput Methods Programs Biomed* 67:187–194
35. Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15:361–387
36. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837–845
37. Harrell FE Jr, Lee KL, Califf RM et al (1984) Regression modelling strategies for improved prognostic prediction. *Stat Med* 3:143–152
38. Chambless LE, Diao G (2006) Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med* 25:3474–3486
39. Steyerberg EW, Vickers AJ, Cook NR et al (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21:128–138
40. Kanis JA, Oden A, Johansson H et al (2012) Pitfalls in the external validation of FRAX. *Osteoporos Int* 23:423–431
41. Pressman AR, Lo JC, Chandra M et al (2011) Methods for assessing fracture risk prediction models: experience with FRAX in a large integrated health care delivery system. *J Clin Densitom* 14:407–415
42. Leslie WD, Lix LM (2011) Absolute fracture risk assessment using lumbar spine and femoral neck bone density measurements: derivation and validation of a hybrid system. *J Bone Miner Res* 26:460–467
43. Steyerberg EW, Harrell FE Jr, Borsboom GJ et al (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 54:774–781
44. Steyerberg EW, Bleeker SE, Moll HA et al (2003) Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* 56:441–447
45. Royston P, Altman DG (2013) External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 13:33
46. Satagopan JM, Ben-Porat L, Berwick M et al (2004) A note on competing risks in survival data analysis. *Br J Cancer* 91:1229–1235
47. Leslie WD, Lix LM, Wu X (2013) Competing mortality and fracture risk assessment. *Osteoporos Int* 24:681–688
48. Scrucca L, Santucci A, Aversa F (2007) Competing risk analysis using R: an easy guide for clinicians. *Bone Marrow Transplant* 40:381–387
49. Allison PD (ed) (2002) Missing data. Sage, Thousand Oaks
50. Enders CK (ed) (2010) Applied missing data analysis. Guilford, New York
51. Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. *Psychol Methods* 7:147–177

52. Little RJA, Rubin DB (eds) (2002) Statistical analysis with missing data. Wiley, New York
53. Mayer B, Muche R, Hohl K (2012) Software for the handling and imputation of missing data—an overview. *J Clin Trials* 2:103–111
54. Gourlay ML, Powers JM, Lui LY et al (2008) Clinical performance of osteoporosis risk assessment tools in women aged 67 years and older. *Osteoporos Int* 19:1175–1183
55. Rud B, Hilden J, Hyldstrup L et al (2009) The Osteoporosis Self-Assessment Tool versus alternative tests for selecting postmenopausal women for bone mineral density assessment: a comparative systematic review of accuracy. *Osteoporos Int* 20:599–607
56. Schwartz EN, Steinberg DM (2006) Prescreening tools to determine who needs DXA. *Curr Osteoporos Rep* 4:148–152
57. Kanis JA, Oden A, Johansson H et al (2009) FRAX and its applications to clinical practice. *Bone* 44:734–743
58. Kanis JA, on behalf of the World Health Organization Scientific Group. Assessment of osteoporosis at the primary health-care level. Technical Report. Accessible at [http://www.shef.ac.uk/FRAX/pdfs/WHO\\_Technical\\_Report.pdf](http://www.shef.ac.uk/FRAX/pdfs/WHO_Technical_Report.pdf). 2007. Published by the University of Sheffield
59. Kanis JA, Oden A, McCloskey EV et al (2012) A systematic review of hip fracture incidence and probability of fracture worldwide. *Osteoporos Int* 23:2239–2256
60. McCloskey E, Kanis JA (2012) FRAX updates 2012. *Curr Opin Rheumatol* 24:554–560
61. Kanis JA, Johnell O, Oden A et al (2000) Long-term risk of osteoporotic fracture in Malmö. *Osteoporos Int* 11:669–674
62. Kanis JA, Oden A, Johnell O et al (2001) The burden of osteoporotic fractures: a method for setting intervention thresholds. *Osteoporos Int* 12:417–427
63. Kanis JA, Oden A, Johnell O et al (2007) 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* 18:1033–1046
64. Sornay-Rendu E, Munoz F, Delmas PD et al (2010) The FRAX tool in French women: how well does it describe the real incidence of fracture in the OFELY cohort? *J Bone Miner Res* 25:2101–2107
65. Tremollieres FA, Pouilles JM, Drewniak N et al (2010) Fracture risk prediction using BMD and clinical risk factors in early postmenopausal women: sensitivity of the WHO FRAX tool. *J Bone Miner Res* 25:1002–1009
66. Leslie WD, Lix LM, Langsetmo L et al (2011) Construction of a FRAX((R)) model for the assessment of fracture probability in Canada and implications for treatment. *Osteoporos Int* 22:817–827
67. Leslie WD, Lix LM, Johansson H et al (2010) Independent clinical validation of a Canadian FRAX tool: fracture prediction and model calibration. *J Bone Miner Res* 25:2350–2358
68. Fraser LA, Langsetmo L, Berger C et al (2011) Fracture prediction and calibration of a Canadian FRAX(R) tool: a population-based report from CaMos. *Osteoporos Int* 22:829–837
69. Czerwinski E, Kanis JA, Osieleń J et al (2011) Evaluation of FRAX to characterise fracture risk in Poland. *Osteoporos Int* 22:2507–2512
70. Tamaki J, Iki M, Kadowaki E et al (2011) Fracture risk prediction using FRAX(R): a 10-year follow-up survey of the Japanese Population-Based Osteoporosis (JPOS) Cohort Study. *Osteoporos Int* 22:3037–3045
71. Rubin KH, Abrahamsen B, Hermann AP et al (2011) Fracture risk assessed by Fracture Risk Assessment Tool (FRAX) compared with fracture risk derived from population fracture rates. *Scand J Public Health* 39:312–318
72. Premaor M, Parker RA, Cummings S et al (2013) Predictive value of FRAX for fracture in obese older women. *J Bone Miner Res* 28:188–195
73. Ettinger B, Ensrud KE, Blackwell T et al. (2013) Performance of FRAX in a cohort of community-dwelling, ambulatory older men: the Osteoporotic Fractures in Men (MrOS) study. *Osteoporos Int* 24:1185–1193
74. Byberg L, Gedeberg R, Cars T et al (2012) Prediction of fracture risk in men: a cohort study. *J Bone Miner Res* 27:797–807
75. Gonzalez-Macias J, Marin F, Vila J et al (2012) Probability of fractures predicted by FRAX(R) and observed incidence in the Spanish ECOSAP Study cohort. *Bone* 50:373–377
76. Tebe Cordomi C, Del Rio LM, Di GS et al. (2013) Validation of the FRAX predictive model for major osteoporotic fracture in a historical cohort of Spanish women. *J Clin Densitom* 16:231–237
77. Dawson-Hughes B (2008) A revised clinician's guide to the prevention and treatment of osteoporosis. *J Clin Endocrinol Metab* 93:2463–2465
78. Dawson-Hughes B, Tosteson AN, Melton LJ III et al (2008) Implications of absolute fracture risk assessment for osteoporosis practice guidelines in the USA. *Osteoporos Int* 19:449–458
79. Kanis JA, Johnell O, Oden A et al (2008) FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 19:385–397
80. Kanis JA, McCloskey EV, Johansson H et al (2008) Case finding for the management of osteoporosis with FRAX((R))—assessment and intervention thresholds for the UK. *Osteoporos Int* 19:1395–1408
81. Lippuner K, Johansson H, Kanis JA et al (2010) FRAX assessment of osteoporotic fracture probability in Switzerland. *Osteoporos Int* 21:381–389
82. Kanis JA, Burlet N, Cooper C et al (2008) European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporos Int* 19:399–428
83. Fujiwara S, Kasagi F, Masunari N et al (2003) Fracture prediction from bone mineral density in Japanese men and women. *J Bone Miner Res* 18:1547–1553
84. Neuprez A, Johansson H, Kanis JA et al (2009) [A FRAX model for the assessment of fracture probability in Belgium]. *Rev Med Liege* 64:612–619
85. Papaioannou A, Morin S, Cheung AM et al (2010) 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. *CMAJ* 182:1864–1873
86. Kanis JA, Hans D, Cooper C et al (2011) Interpretation and use of FRAX in clinical practice. *Osteoporos Int* 22:2395–2411
87. Hans DB, Kanis JA, Baim S et al (2011) Joint Official Positions of the International Society for Clinical Densitometry and International Osteoporosis Foundation on FRAX((R)) Executive Summary of the 2010 Position Development Conference on Interpretation and Use of FRAX((R)) in Clinical Practice. *J Clin Densitom* 14:171–180
88. Lo JC, Pressman AR, Chandra M et al (2011) Fracture risk tool validation in an integrated healthcare delivery system. *Am J Manag Care* 17:188–194
89. Siminoski K, Leslie WD, Frame H et al (2005) Recommendations for bone mineral density reporting in Canada. *Can Assoc Radiol J* 56:178–188
90. Siminoski K, Leslie WD, Frame H et al (2007) Recommendations for bone mineral density reporting in Canada: a shift to absolute fracture risk assessment. *J Clin Densitom* 10:120–123
91. Leslie WD, Berger C, Langsetmo L et al. (2011) Construction and validation of a simplified fracture risk assessment tool for Canadian women and men: results from the CaMos and Manitoba cohorts. *Osteoporos Int* 22:1873–1883
92. Nguyen ND, Frost SA, Center JR et al (2007) Development of a nomogram for individualizing hip fracture risk in men and women. *Osteoporos Int* 18:1109–1117
93. Nguyen ND, Frost SA, Center JR et al (2008) Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks. *Osteoporos Int* 19:1431–1444

94. Langsetmo L, Nguyen TV, Nguyen ND et al. (2011) Independent external validation of nomograms for predicting risk of low-trauma fracture and hip fracture. *CMAJ* 183:E107
95. Hippisley-Cox J, Coupland C (2009) Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 339:b4229
96. Hippisley-Cox J, Coupland C (2012) 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* 344:e3427
97. Collins GS, Mallett S, Altman DG (2011) Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ* 342:d3651
98. Cummins NM, Poku EK, Towler MR et al (2011) Clinical risk factors for osteoporosis in Ireland and the UK: a comparison of FRAX and QFractureScores. *Calcif Tissue Int* 89:172–177
99. Black DM, Steinbuch M, Palermo L et al (2001) An assessment tool for predicting fracture risk in postmenopausal women. *Osteoporos Int* 12:519–528
100. Robbins J, Aragaki AK, Kooperberg C et al (2007) Factors associated with 5-year risk of hip fracture in postmenopausal women. *JAMA* 298:2389–2398
101. Henry MJ, Pasco JA, Sanders KM et al (2006) Fracture Risk (FRISK) Score: Geelong Osteoporosis Study. *Radiology* 241:190–196
102. Henry MJ, Pasco JA, Merriman EN et al (2011) Fracture risk score and absolute risk of fracture. *Radiology* 259:495–501
103. Pluijm SM, Koes B, de Laet C (2009) A simple risk score for the assessment of absolute fracture risk in general practice based on two longitudinal studies. *J Bone Miner Res* 24:768–774
104. Sambrook PN, Flahive J, Hooven FH et al (2011) Predicting fractures in an international cohort using risk factor algorithms without BMD. *J Bone Miner Res* 26:2770–2777
105. Bolland MJ, Siu AT, Mason BH et al (2011) Evaluation of the FRAX and Garvan fracture risk calculators in older women. *J Bone Miner Res* 26:420–427
106. Leslie WD, Schousboe JT (2011) A review of osteoporosis diagnosis and treatment options in new and recently updated guidelines on case finding around the world. *Curr Osteoporos Rep* 9:129–140
107. Compston J, Cooper A, Cooper C et al (2009) Guidelines for the diagnosis and management of osteoporosis in postmenopausal women and men from the age of 50 years in the UK. *Maturitas* 62:105–108
108. Kanis JA, McCloskey EV, Johansson H et al (2013) European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporos Int* 24:23–57
109. Office of the Surgeon General (US). (2004) Bone Health and Osteoporosis: A Report of the Surgeon General. Rockville (MD): Office of the Surgeon General (US) Available from: <http://www.ncbi.nlm.nih.gov/books/NBK45513/>
110. Chen P, Klege JH, Adachi JD et al (2009) Vertebral fracture status and the World Health Organization risk factors for predicting osteoporotic fracture risk. *J Bone Miner Res* 24:495–502
111. Donaldson M, Palermo L, Schousboe JT, Ensrud K, Hochberg MC, Cummings SR (2009) FRAX and risk of vertebral fractures: The Fracture Intervention Trial (FIT). *J Bone Miner Res* 24:1793–1799
112. Ensrud KE, Lui LY, Taylor BC et al (2009) A comparison of prediction models for fractures in older women: is more better? *Arch Intern Med* 169:2087–2094
113. Leslie WD, Lix LM, Johansson H et al (2011) Spine–hip discordance and fracture risk assessment: a physician-friendly FRAX enhancement. *Osteoporos Int* 22:839–847
114. Kanis JA, McCloskey E, Johansson H et al (2012) FRAX((R)) with and without bone mineral density. *Calcif Tissue Int* 90:1–13
115. van den BT, Heymans MW, Leone SS et al (2013) Overview of data-synthesis in systematic reviews of studies on outcome prediction models. *BMC Med Res Methodol* 13:42