01 Sep 2012

Developing Novel Prognostic Biomarkers for Multivariate Fracture Risk Prediction Algorithms

Ernest K. Poku

Mark R. Towler
Missouri University of Science and Technology, mtowler@mst.edu

Niamh M. Cummins

Jeff D. Newman

Recommended Citation

E. K. Poku et al., "Developing Novel Prognostic Biomarkers for Multivariate Fracture Risk Prediction Algorithms," *Calcified Tissue International*, vol. 91, no. 3, pp. 204 - 214, Springer, Sep 2012.
The definitive version is available at https://doi.org/10.1007/s00223-012-9627-0

Calcif Tissue Int (2012) 91:204–214
DOI 10.1007/s00223-012-9627-0

ORIGINAL RESEARCH

Developing Novel Prognostic Biomarkers for Multivariate Fracture Risk Prediction Algorithms

Ernest K. Poku · Mark R. Towler · Niamh M. Cummins · Jeff D. Newman

Received: 25 April 2012 / Accepted: 15 June 2012 / Published online: 11 July 2012

Abstract Multivariate prediction algorithms such as FRAX® and QFractureScores provide an opportunity for new prognostic biomarkers to be developed and incorporated, potentially leading to better fracture prediction. As more research is conducted into these novel biomarkers, a number of factors need to be considered for their successful development for inclusion in these algorithms. In this review, we describe two well-known multivariate prediction algorithms for osteoporosis fracture risk applicable to the UK population, FRAX and QFractureScores, and comment on the current prognostic tools available for fracture risk; dual X-ray assessment, quantitative ultrasonography, and genomic/biochemical markers. We also highlight the factors that need to be considered in the development of new biomarkers. These factors include the requirement for prospective data, collected in new cohort studies or using archived samples; the need for adequate stability data to be provided; and the need for appropriate storage methods to be used when retrospective data are required. Area under the receiver operating characteristic curve measures have been found to have limited utility in assessing the impact of the addition of new risk factors on the predictive performance of multivariate algorithms. New performance evaluation measures, such as net reclassification index and integrated discrimination improvement, are increasingly important in the evaluation of the impact of the addition of new markers to multivariate algorithms, and these are also discussed.

Keywords Algorithms · DXA · Fracture · FRAX · Prognostic

The authors have stated that they have no conflict of interest.

E. K. Poku (✉) · J. D. Newman
Cranfield Health, Cranfield University, Cranfield, UK
e-mail: e.poku@cranfield.ac.uk

M. R. Towler
Inamori School of Engineering, Alfred University, Alfred, NY, USA

M. R. Towler
Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

N. M. Cummins
Centre for Interventions in Infection, Inflammation and Immunity, Graduate Entry Medical School, University of Limerick, Limerick, Ireland

The introduction of multivariate algorithm-based fracture risk assessment tools such as FRAX® has broadened the risk factors considered important for osteoporotic fracture risk [1]. These risk calculators are modifiable and therefore can incorporate appropriately validated new prognostic markers for fracture risk in the future. In particular, additional markers for bone quality factors that are linked to fracture risk would be beneficial [2]. Bone quality refers in part to the organic matrix of bone, but it also describes a set of characteristics that influence strength, such as microarchitecture, remodeling, and damage accumulation.

Traditionally, osteoporosis has been defined using bone mineral density (BMD) as measured by T-scores. In recent years, there has been a move away from T-scores as the operating definition of osteoporosis, in favor of the use of absolute risk of fracture and risk calculators that are based on algorithms that estimate those risks [3]. This movement has led to an evolved definition of the disease that incorporates more clinical risk factors (CRF) and that bases treatment decisions on absolute risk of fracture thresholds over a 10 year period rather than T-scores [3]. This paradigm shift brings osteoporosis into line with other conditions, such as heart disease, where patient risks are assessed on an absolute risk-of-event basis over a 10 year period [4]. This major change in the definition of osteoporosis creates an opportunity for new prognostic biomarkers to be identified and incorporated into risk assessment quickly and efficiently via the absolute risk approach once they meet the appropriate clinical evidence requirements.

There is a need for improved prognostic factors in osteoporosis as a result of the increasing burden of fracture on the population and the resultant high mortality rates. Burge et al. [5] estimated that there were more than 2 million fractures in the United States in 2005, resulting in direct health care costs of $17 billion. The authors projected that this number would grow by 50 % by 2025 as a result of the aging of the population (the “gray tsunami”). The number of fractures will reach over 3 million a year, at an annual cost of $25 billion, over the same time period [5]. It has been established that while incidence rates of hip fracture may be relatively low, excess mortality is significant, at between 8 and 36 % compared with community-based controls during the first year [6]. It has also been noted that it would be beneficial to treat women earlier than is current practice, ideally in the perimenopausal stage, when bone mass is near its lifetime peak, in order for the benefits of early preventative treatment to be realized [7]. The downside is that longer follow-up times may be needed than the currently standard 10-year period to conduct clinical trials that demonstrate the long-term benefits of early treatment.

Here we review the current state of the art and consider the steps required to develop a new prognostic marker with sufficient clinical evidence to justify inclusion in current fracture risk calculators. We include current risk factors and their evidence bases, methodologies for introducing new risk factors, and new techniques available to evaluate their performance in terms of health and cost-effectiveness.

Current Prognostic Fracture Risk Calculators

The increased interest in regression model-based risk calculators developed from established cohort studies has been driven by the need to develop more accurate models for who is likely to experience a fracture and when the fracture will occur. Another important issue is the lack of availability of dual X-ray assessment (DXA) machines in many countries. Also, DXA performance is not optimal for detecting osteoporotic fracture risk as a result of poor predictive sensitivity, and therefore the use of additional CRFs in combination with DXA could help increase the sensitivity of diagnosis without impairing specificity [8]. Health economic evaluations have indicated that it is most effective to implement mass screening programs using an initial assessment with CRFs followed by DXA evaluation in high-risk subjects [9]. Mass screening can therefore be justified with the support of non-BMD prognostic markers to enhance overall prognostic performance in combination with DXA. Early work to combine CRFs into prediction models for fracture risk to supplement DXA was conducted by Black et al. [10]. Subsequent work has resulted in three validated fracture risk prediction models that are currently available online: FRAX, QFractureScores, and the Garvan model. The Garvan and Black models were developed in Australian and U.S. populations, respectively. The FRAX model is currently the most widely used. In order to provide an illustrative comparison, two of these models, FRAX and QFractureScores, both of which are available for UK populations, are described in more detail.

FRAX

The World Health Organization Collaborating Centre for Metabolic Bone Diseases (University of Sheffield, Sheffield, UK), led by Kanis [7], developed the FRAX risk calculator to improve osteoporosis risk assessment. The algorithm, which uses a Poisson regression model to estimate risk, was developed with data from nine population cohorts and validated in another 11 cohorts comprising over 1 million patient-years. FRAX can calculate 10 year risk probabilities with or without the inclusion of femoral neck BMD. Table 1 lists the CRFs currently considered to have sufficient clinical evidence to justify their inclusion in FRAX.

There are a number of general and methodology-specific limitations in the FRAX initiative [8]. The calculator does not consider medications that influence fracture risk, and other factors such as the risk of falls and the presence of biochemical markers of bone turnover have been excluded because of the lack of large prospective studies validating their use. Additionally, risk factors are quantified in a binary fashion, rather than by using multiple state options. A wide number of risk factors were considered for inclusion, but only nine were thought to have sufficient evidence to justify their inclusion in the model [7]. The developers consider FRAX to be a platform technology into which new risk factors can be incorporated as they become available [3]. CRFs used in isolation do not predict fracture risk as strongly as a BMD measurement. However, CRFs in combination with BMD provide an enhanced predictive ability over BMD alone.

Health screening modeling has demonstrated that the combined use of CRF and BMD in FRAX leads to a higher positive predictive value, a lower number of subjects required to treat to prevent one fracture, and enhanced

Table 1 Clinical risk factors evaluated by the FRAX and QFractureScores algorithms

Clinical risk factor	FRAX	QFractureScores
Age	X	X
Sex	X	X
Weight	X	X
Height	X	X
Previous fracture	X	
Parental hip fracture/osteoporosis	X	X
Smoking	X	X
Glucocorticoids[a]	X	X
Rheumatoid arthritis	X	X
Secondary osteoporosis[b]	X	
Alcohol intake	X	X
Femoral neck bone mineral density	X	
Asthma		X
Heart attack/stroke		X
Falls		X
Chronic liver disease		X
Tricyclic antidepressants		X
Type 2 diabetes		X
Hormone replacement therapy		X
Endocrine problem		X
Malabsorption		X
Menopausal symptoms		X

[a] In QFractureScores, the use of "steroids" is recorded rather than glucocorticoids

[b] In QFractureScores secondary causes of osteoporosis are not recorded as a single entity but are recorded separately (as shown above)

sensitivity in 55-, 60-, and 65-year-olds over BMD alone [11]. This indicates that additional non-BMD prognostic factors could enhance the overall performance of predictive tools for fracture risk. The FRAX developers selected a 10 year horizon partly on the basis of the likely treatment duration, and also on the basis of the limitations of the available clinical evidence, as few relevant studies had more than 10 years of follow-up data [12]. However, it may also be clinically useful to predict the 20 year or lifetime risks for younger women in order to earlier identify those who are significantly at risk of a fragility fracture in the future, which may be used to justify more regular screening that may result in non-pharmaceutical interventions and lifestyle advice at an earlier stage for higher-risk individuals. Early intervention at perimenopause could result in greater maintenance of bone mass and a reduction in the rate of loss in later life [13]. Barr et al. have shown that screening for osteoporosis between the ages of 45 and 54 and following up with hormone replacement therapy leads to reduced fracture incidence [14]. The incidence of hip fracture rises significantly in women aged between 70 and 90, and clinical studies indicate that between these ages, the prognostic performance of BMD as determined by DXA falls by more than the performance of CRFs [7]. There may therefore be an argument to focus on CRFs and exclude BMD as a risk factor when identifying elderly women who would benefit from treatment.

The development and rapid acceptance of FRAX is an acknowledgement by the medical community of the importance of non-BMD risk factors in predicting osteoporotic fracture. The use of BMD within FRAX does improve prediction [11], but the identification of additional risk factors with the potential to replace BMD would be beneficial to widen the use of osteoporosis screening, particularly in lower-income countries where DXA is often unavailable. The advantage of including non-BMD-based CRFs that can be collected in a questionnaire format by a risk algorithm is that these can be obtained at low cost and can add significantly to the prognostic power of BMD or, in the absence of BMD can provide an acceptable decision-making tool for clinicians.

QFractureScores

The developers of the QFractureScores algorithm (http://www.qfracture.org) implemented a very different approach to that of the FRAX developers. Their aim was to develop an algorithm that was prognostic without the requirement for diagnostic testing that introduces an external cost to the prevention program. The QResearch database, a validated database of risk factors and outcome data collected from primary care practices in the UK, was used to develop the algorithm [15]. This database contains the health records of over 11 million people in England and Wales. The QResearch database contains information on 1,174,232 men and 1,183,633 women, aged between 30 and 85, and 7,898,208 (female) and 8,049,306 (male) observation years were used in developing the algorithm. In the female group, 24,350 incident fractures and 9,302 hip fractures were recorded. The risk factors assessed in the database are outlined in Table 1.

The hazard ratios (HR) and coefficients in the model were derived by the Cox proportional hazard regression model. To validate the QFractureScores model, hip fracture prognostic performance in a separate defined QResearch group was compared with the actual events over a 10 year period and with the predictions generated by FRAX in the same cohort. The validation group contained 653,789 women, and the average hip fracture incidence rate was 1.15 % (range 1.13–1.17 %) [15]. QFractureScores has also been externally validated in a UK-based population using records in The Health Improvement Network (THIN) database (http://www.thin-uk.com), which added an additional 13 million observation years. The observed results closely matched those observed in the internal validation study, adding further evidence for the integrity of the QFractureScores approach [16]. The developers of QFracture have recently released a new algorithm incorporating additional risk factors such as ethnicity and previous fracture on the basis of their analysis of the prospective cohort study, QResearch, which has improved predictive performance over the original QFracture algorithm [17].

A Comparison of FRAX and QFractureScores

FRAX and QFractureScores were compared using the validation cohort in the original QFractureScores study [15]. QFractureScores resulted in better discrimination compared with FRAX using the D statistic. The values were 0.11 higher in women; any difference exceeding 0.1 is considered important. The authors attribute the performance of QFractureScores to the fact that FRAX uses data from multiple international databases rather than from a single national data source, as is the case with the QResearch database. The FRAX algorithm generated an area under the receiver operating characteristic curve (AUC) value of 0.845 for female hip fracture, and QFractureScores had a value of 0.89 for the same event. However, the use of these data for a direct comparison of FRAX and QFractureScores may not be appropriate because of the difficulties encountered in comparing AUCs between studies, particularly when adjustments have not been made for differences in major predictive factors, such as age, between studies [18]. Recent work in an independent UK- and Irish-based population using only CRFs indicated that FRAX and QFractureScores were reasonably well correlated ($R = 0.857$) for hip fracture, suggesting that both tools could be of value in primary care settings [19].

In addition to the differences in outcomes predicted, there are methodological differences between the two algorithms. DXA measures are not considered in QFractureScores, whereas they are an important variable in FRAX. Additionally, mortality is considered in FRAX but not in QFractureScores; death as a risk factor becomes increasingly important with age, particularly in people older than 80, and this should be considered in any comparison of the two models in older subjects. In terms of input factors to the algorithm, as shown in Table 1, QFractureScores does not consider prior fracture as it was developed in subjects without a prior fracture, which gives the algorithm a different weighting to FRAX. The clinically relevant outcomes predicted by the two algorithms also differ, as shown in Table 2.

Table 2 Comparison of outcomes

Fracture	FRAX	QFractureScores
Hip	X	X
Clinical vertebrae	X	X
Humerus	X	
Wrist	X	
Distal radius		X

Current Prognostic Biomarkers

DXA

DXA has been shown to be predictive for hip fracture at the femoral neck with different odds ratios depending on the age of the subject; a 50-year-old has been shown to have a risk of 3.68 (95 % CI 2.61–5.19) and an 80-year-old to have a risk of 2.28 (95 % CI 2.09–2.50) [20]. Incidence rates increase with age, but the predictive power of DXA for 10 year hip fracture reduces with age. Additionally, DXA has the adoption challenges of cost, availability, and effectiveness in women younger than 65 [21]. This age group has been identified as important for making long-term treatment decisions that will greatly affect future fracture rates; a group DXA is currently unable to support using mass screening [21, 22].

Quantitative Ultrasonography

Quantitative ultrasonography (QUS) is an alternative technique to DXA for assessing BMD and has been available since the early 1990s. Hans et al. demonstrated the prognostic power of QUS in women with a mean age of 80.4 years over a 2 year follow-up [23]. The relative risk for hip fracture was 2.0 (1.6–2.4) for broadband ultrasound attenuation and 1.9 (1.6–2.4) for speed of sound compared with 1.9 (1.6–2.4) for BMD as measured by DXA in the same study. There has always been a view that QUS measures more aspects of bone structure (e.g., microarchitecture) than just BMD and as a result provides some measure of bone quality [24]. Langton and Langton [25] reported linear regression fit (R^2) values between broadband ultrasound attenuation and elasticity (Young's modulus) in calcaneus bone of between 65 and 77 %, indicating a relationship between the two values. The potential to incorporate some bone quality measures into an overall assessment of fracture risk has clear clinical utility [25], and there is now some clinical evidence that QUS is prognostic of hip fracture over a 10 year period. A 1 standard deviation (SD) decrease in broadband ultrasound attenuation gave a HR for non-vertebral fracture of 1.414 (1.236–1.616), and a 1 SD change in speed of sound resulted in a HR of 1.359 (1.193–1.548) [26].

Since the move to measurement of absolute risks for risk assessment, there has been a reappraisal of the diagnostic potential of QUS. In a recent study of 1,455 participants aged between 64 and 76, followed up over 10.3 years and including 79 fracture cases, an algorithm incorporating both QUS and known CRFs, including smoking, prior fracture, and alcohol intake, achieved comparable results to DXA. The combination of QUS and CRFs achieved a HR of 2.04 (1.55–2.69) per SD compared with a HR of 2.26 (1.74–2.95) for BMD. The authors concluded that in terms of absolute risk, the use of QUS is comparable with DXA [27]. The move to absolute risk for assessing future fracture risk appears to offer some additional opportunities for QUS to gain wider acceptance, but the number of long-term prospective studies required to confirm the results of Moayyeri et al. [27] will continue to be a barrier to its wider acceptance. The adoption of QUS in clinical practice has also been limited as a result of issues with the maintenance of instrument precision and accuracy, as well as reproducibility in practice.

Biochemical Markers

A number of studies have shown that biochemical markers of bone remodeling are capable of predicting fracture risk [28, 29]. These biomarkers have the advantage of reflecting global skeletal activity whereas BMD measurements assess only a small portion of the skeleton at a specific site. Garnero et al. [30] demonstrated that cross-linked C telopeptides of type I collagen (CTX) is prognostic of hip fracture in older women, with an odds ratio for hip fracture of 2.2 (1.3–3.6), which was independent of bone mass. The use of BMD and CTX in combination generates a higher hip fracture odds ratio of 4.8. Although these studies demonstrated the utility of CTX to predict fracture, the patient population has limited clinical utility. The EPIDOS study was conducted in an older population (over 74 years of age), and the study had a short (3 years) follow-up period. The evidence for CTX's clinical utility for the prevention of future fracture over a longer period and in younger women is still to be developed [29]. Other bone turnover markers shown to be predictive include serum osteocalcin, serum procollagen type I C propeptide, and urinary deoxypyridinoline, but they all currently lack the required level of clinical evidence to justify inclusion in the FRAX algorithm [29]. Biochemical markers have the advantage of being easily measured in a serum or urine sample; however, this also means that issues of biological variability can arise.

Genomic Markers

Osteoporosis is a polygenic disease, involving a large variety of gene products implicated in both bone modeling and remodeling. A number of candidate genes have already been identified including those that code for the following: vitamin D receptor, estrogen receptor, insulin growth factor, parathyroid hormone, and type I collagen. Twin studies have been widely used to assess the importance of genotype in the osteoporotic condition, finding that between 60 and 85 % of BMD variance is genetically determined [31, 32]. Research has also been conducted on non-BMD risk factors; Mann and Ralston investigated the genetic influence on non-BMD CRFs including body mass index, age at menopause, and smoking history. A statistically significant relationship was found between a gene that encodes for collagen type 1 alpha 1 (*COL1A1*) and body mass index and fracture risk [33]. However the other CRFs were found to be nonsignificant. An association between the polymorphism for transcription factor Sp1 in the gene *COL1A1* and bone health has also recently been reported. The presence of at least one copy of the T allele was associated with osteoporotic fractures, but not with low BMD, in postmenopausal white women aged 50–70 years [34]. The increasing use of whole genome studies to investigate disease brings new hope for improved clinical utility with genetic tests, but prospective studies will be required to establish a compelling link to future fracture [35]. The limitations of the genetic research are that most studies to date have focused on the link between genotype and BMD rather than future fracture risk [36]. It is also probable that a prognostic test based on whole genome analysis is likely to be prohibitively expensive for mass screening in the foreseeable future.

Development of New Prognostic Biomarkers

Because of the limitations of BMD and the existing non-BMD-based markers, there is a need to identify new prognostic markers that could enhance the overall performance of tools like FRAX. Demonstrating that these new biomarkers are predictive of fracture risk rather than correlated with DXA T-scores requires the use of prospective study data with substantial follow-up times. Kanis et al. [1] have discussed the cohorts considered suitable for deriving data for a risk calculator and have shown that hundreds of thousands of person-years are required.

Importance of Cohort Studies

In order to develop a completely novel prognostic marker for osteoporotic fracture risk, there is a requirement to collect patient samples at baseline and then to follow the patient for a number of years. Because of the low incidence rate of hip fractures in postmenopausal women (less than 5 %), cohorts in excess of 10,000 subjects could be required to ensure sufficient events have occurred over a 10 year study, making the costs and time commitment for new studies substantial. This is especially the case when the women of interest are perimenopausal, and therefore the incidence rate of fracture is particularly low over the next decade [7]. An alternative, more cost-effective option is to apply a retrospective cohort approach using an existing well-established cohort in which samples were collected in the past and then followed up for hip fracture in subsequent years. Osteoporosis cohorts of this type that are long established and well known include the Aberdeen Prospective Osteoporosis Screening Study [37] and the European Prospective Osteoporosis Study [38]. However, a challenge with retrospective cohort approaches is the restriction to existing samples and data already collected, which may be suboptimal for the new marker of interest. This limitation can mean that data required in the predictive algorithm may not have been collected at baseline, either for an individual patient or for the entire group. Studies can manage this problem by using multiple imputation for individuals; however, for the entire cohort, it may not be possible to replicate data for the risk factor. The missing risk factor may result in different results that should be considered in any overall interpretation of the study. If the number of risk factors missing render the retrospective cohort study approach impractical, an alternative approach would be to include the new risk factor into a prospective clinical study that incorporates treatment. The advantage of commencing a completely new study is the ability to examine any biomarkers of interest and any end points of interest. Incorporating the new risk factor as an arm in a study such as the Screening of Older Women for Prevention of Fracture (SCOOP) study [39] may provide an intermediate approach between the lower cost and speed of a solely retrospective study and the high costs and long duration of a long-term prospective fracture study. The SCOOP study evaluates a FRAX- and DXA-based screening method compared with standard screening methods followed by treatment, and the primary outcome is the number of fractures in each arm. This 5 year study will provide evidence of the performance of the predictive algorithm on the most important clinical outcome: fractures.

In order to enable a retrospective study to be carried out in a timely and cost-effective manner without having to test tens of thousands of archived samples, nested case-control designs are attractive. Sample types previously collected in published osteoporosis studies have included DXA scores as well as bone, blood, urine, and skin samples [40–42]. Because the incidence rate of hip fracture is less than 3 % in the age range with most clinical utility, 50–70 years of age, the use of case-to-control ratios of 1:3 or more is recommended [43]. As an example of this nested case-control approach, envisage a scenario where a new technique has been developed that can extract bone quality information from X-ray images. If we assume that a 1:3 case-to-control ratio is sufficient, rather than conducting a completely new study, we instead could retrospectively examine archived X-ray images from 100 fracture events and 300 controls. These 400 data points could then be evaluated more cost- and time-effectively than the traditional prospective approach.

The developers of the FRAX algorithm developed substantial evidence requirements for the inclusion of risk factors, including their use in a number of studies, accumulated person-years in trials, and follow-up durations [7]. For new biomarkers to be accepted into risk calculators without prohibitive barriers to entry, it is proposed that the following acceptance criteria be used: a follow-up time of at least 5 years and independent verification in two cohorts using a training set developed in a separate cohort. This approach would allow additional risk factors to be incorporated for applications where DXA is not available.

In order to take advantage of the retrospective cohort study approach, new prognostic markers must make use of stored or archived samples. This means that new prognostic methods that cannot use previously archived samples will require prospective studies to be fully validated. This will be a significant evidence barrier for the development of some novel techniques.

Sample Stability Considerations

Several years of follow-up are required to collect sufficient clinical data for prospective studies, and when archived samples are used in a study, the effect of the ageing process on the archived samples need to be taken into consideration. It is essential to demonstrate that archived samples will yield similar or identical results to previous work upon reanalysis, or that any changes observed are consistent and can be accounted for in subsequent calculations. This question has previously been explored in the literature in a limited way; a major challenge is the requirement to evaluate long-term storage for each sample type and analyte. UK Biobank has developed a protocol for the collection of blood and urine with a view to long-term storage that is based on a review of the literature; they established the need to freeze samples at particular temperatures for particular applications for long-term storage [40]. It is likely that any new biomarker would need to explore the use of accelerated aging on fresh biological samples to mimic archived samples stored for several years in order to establish the viability of testing the samples for a new analyte. Possible approaches include the use of calculations such as the Arrhenius equation, but it is challenging to mimic aging processes that can be measured in decades using this process [44].

Performance Evaluation Measures

There has been increased recognition in the recent academic literature that there is a need for additional measures to assess the performance of different prognostic risk factors and multivariate risk models beyond what is offered by the receiver operating characteristic (ROC) curve [4]. The ROC curve has been observed to perform poorly as a measure of prognostic performance in population-based cohorts in which the disease has a low prevalence; a graphical example of the increased ROC performance observed by the addition of CRFs to BMD is shown in Fig. 1. This is the situation in osteoporotic hip fractures where the incidence rate is low, but the consequences for health are very serious in terms of increased mortality [6]. McClish [45] has suggested solutions to improve the clinical utility by analyzing just a portion of the ROC curve. The full area under the ROC curve approach was criticized for equally weighting false-positive rates that may not reflect the clinical outcomes in a number of conditions. Calibration remains an important evaluation measure for predictive models; it assesses the ability of the model to accurately predict the incidence rate for the event compared with the rate observed in reality. Graphical examples of calibration comparing the predictive performance of FRAX and QFracture in a UK cohort are provided in Figs. 2 and 3 [15]. The Hosmer–Lemeshow test is commonly used to report the goodness of fit of the predicted and observed incidence rates [46].

New approaches to evaluate the performance of additional clinical markers in a multivariate model have been proposed that move beyond whether a new prognostic test offers good discrimination between the cases and controls (as evaluated by the ROC curve) to whether it significantly changes the classification of the subject who may be at risk [4]. The addition of risk factors into prognostic models can result in small changes in the AUC, which do not reflect the changes in risk category that result from the new information [46]. Cook notes that many new biomarkers may have clinically relevant odds ratios (between 1.5 and 2.0), but these will have only a modest impact on the ROC curves [46]. New reclassification metrics are able to address the weaknesses of these measures in terms of perfect discrimination using ROC curves. It is now being argued that these new measures are more important to prognostic models than the traditional ROC curve and AUC measure.

More novel techniques that provide additional information on the relative performance of predictive models are net benefit analysis [47], decision curve analysis [48], the Pepe method [49], net reclassification index (NRI) [46], and integrated discrimination analysis (IDI) [50]. Net benefit approaches allow a broader evaluation of the clinical usefulness of a predictive model by incorporating information on clinical management strategies. These models can be complex to develop, and decision curve analysis has the advantage of providing an evaluation of net benefit using a simpler model that requires no additional data on costs or treatment effectiveness. The Pepe method provides additional information on the performance of a model by classifying the subjects on the basis of the proportion above and below selected thresholds and their case and non-case status.

Two new measures of prognostic performance in particular are gaining popularity: the NRI and the IDI. The NRI is a measure that quantifies the number of subjects correctly reclassified as diseased and correctly reclassified as healthy on the basis of the addition of a new biomarker. IDI is similar to NRI but uses probabilities rather than risk categories [50]. There is some debate on the most appropriate way to use these new measures. Pencina et al. [50] argue that for the evaluation of a new marker, an additional measure, IDI, is required, rather than using NRI in isolation. This measure describes the difference between the

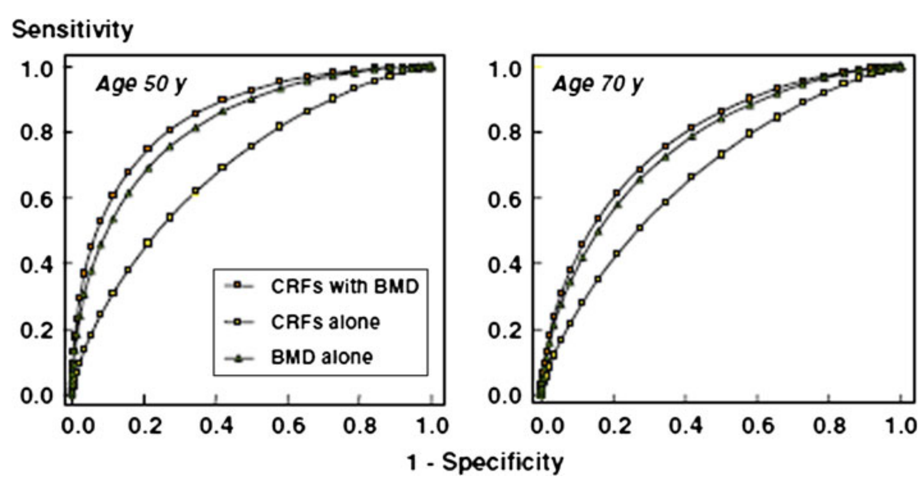

Fig. 1 Receiver operating characteristic curves for the risk score for hip fracture prediction at the ages of 50 and 70 years [20]. Image used with permission of the World Health Organization Collaborating Centre for Metabolic Bone Diseases, University of Sheffield. FRAX® is registered to J. A. Kanis, University of Sheffield

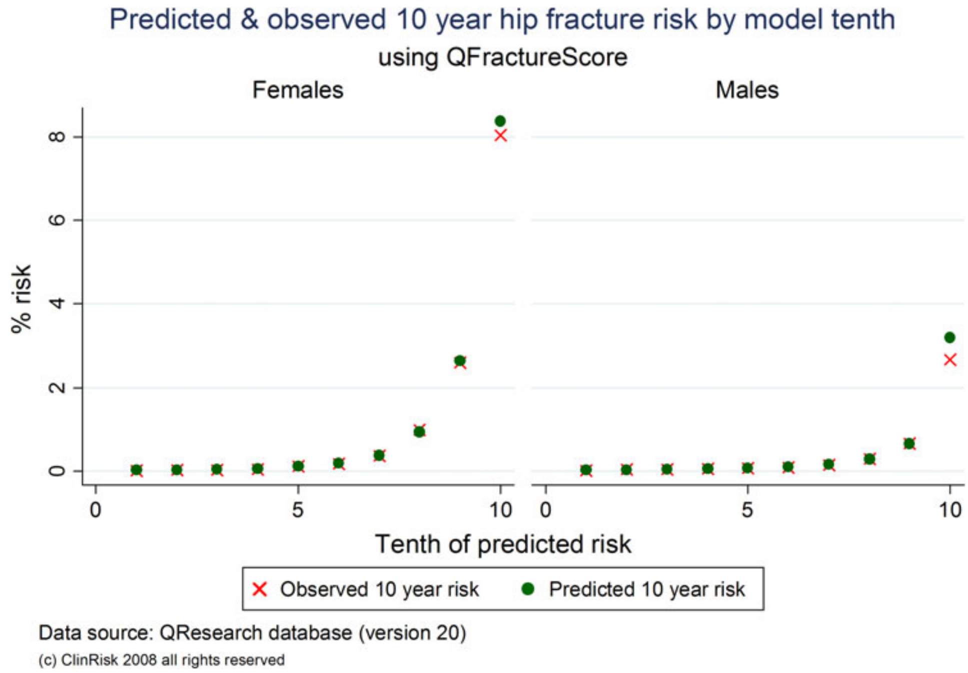

Fig. 2 Predicted to observed risk of hip fracture using QFractureScores [15]

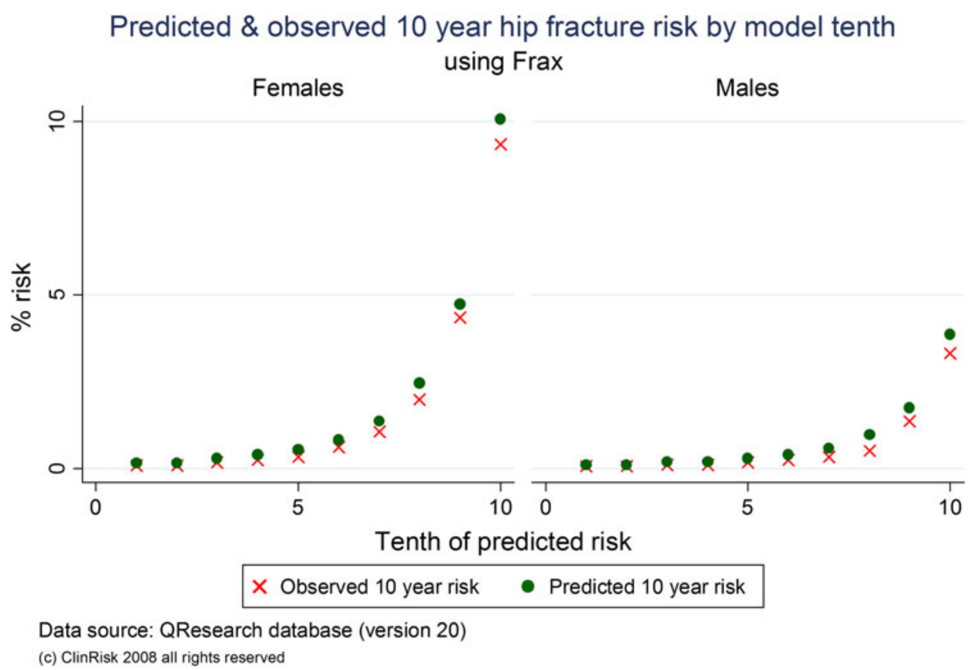

Fig. 3 Predicted to observed risk of hip fracture using the FRAX® [15]

improvement in average sensitivity and any change in average "one minus specificity" and can be seen as an alternative to AUC that is appropriate for use when adding a new marker to a multivariate prediction algorithm.

It is notable that the authors of these articles state that these new measures can be used to evaluate whether an expensive new biomarker should be introduced from an economic standpoint. However, the criteria to evaluate this have not yet been published in detail. Published health economic studies typically use odds ratio, relative risk, or AUC to evaluate the economic performance of tests [51]. It is clear from recent work that the limited impact of some new additive tests on AUC measures restrains the ability of clinical practitioners to evaluate cost-effectiveness because the AUC is not taking the reclassification of subjects into account. A recent study has explored the possibility of evaluating cost-effectiveness using NRI as an alternative to traditional relative risk based approaches and have investigated how measures of discrimination, classification, and costs can be linked [52]. Pencina et al. [50] offer a process that weighs the NRI on the basis of the cost saving when a person moves up in classification compared to incurred costs when they move down in classification, caused by misclassification. An example would be when a person no longer receives unnecessary treatment as a result of reclassification.

These measures have previously been used to explore the performance of cardiac markers and are now also being used in osteoporosis studies [53]. Donaldson et al. compared a simple BMD and age model with the FRAX model using the Cook and Pepe methods in the Study of Osteoporotic Fractures. AUC in both models was similar for hip fracture (0.75 vs. 0.76), but the novel methods were able to differentiate the predictive models by identifying differences in who is correctly and incorrectly classified. A total of 8 % of cases were not treated in error, but 18 % of non-cases were correctly not treated, according to an analysis using the Pepe method when the FRAX model was compared with the simple model.

Conclusions

Online fracture risk assessment tools offer significant opportunities for novel biomarkers to be used for fracture risk prediction. The challenges created by the requirement to demonstrate predictive power over time frames in excess of a decade in a disease with a relative low incidence rate is challenging, particularly when there are no archived samples to draw on. The recent work with the QResearch database indicates that better predictive performance can be achieved by the addition of more risk factors, if appropriately validated. Although the prevalence of osteoporosis is high, the incidence rate of the most damaging event, hip fracture, is relatively low—less than 5 % per year in the population of interest. The development of new prognostic markers has a significant barrier that is based on the long follow-up time during which events occur. Using retrospective studies with archived samples, intervention studies, and nested case-control and case-cohort approaches may substantially improve the development times for the adoption of new biomarkers. The limited number of archived samples available and their stability over long durations of time will be key considerations in the development of these approaches.

Several alternative prognostic biomarkers have been evaluated to date, but as yet none has provided the evidence base to supersede DXA. It may be that the way forward now is to use these tools in combination with DXA and, where cost-effective, as a prescreening tool to select subjects for DXA testing. This review has set out some of the considerations required by researchers seeking to incorporate new risk factors into the existing prediction algorithms. There is significant scope in the field of osteoporosis for the following: increased use of real patient data, increased use of archived samples, increased use of end points with real clinical utility, i.e., hip fracture, and for prognostic-based end points like NRI and IDI to be applied. These new techniques could ultimately lead to the development of a new generation of prognostic tools to improve patient care for people with osteoporosis.

References

1. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E (2008) FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int 19:385–397
2. Bouxsein ML (2003) Bone quality: Where do we go from here? Osteoporos Int 14(suppl 5):S118–S127
3. McCloskey EV, Johansson H, Oden A, Kanis JA (2009) From relative risk to absolute fracture risk calculation: the FRAX algorithm. Curr Osteoporos Rep 7:77–83
4. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115:928–935
5. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A (2007) Incidence and economic burden of osteoporosis-related fractures in the United States, 2005–2025. J Bone Miner Res 22:465–475
6. Abrahamsen B, van Staa T, Ariely R, Olson M, Cooper C (2009) Excess mortality following hip fracture: a systematic epidemiological review. Osteoporos Int 20:1633–1650
7. Kanis J (2007) Assessment of osteoporosis at the primary health care level. World Health Organization, Lyon
8. Kanis JA, Oden A, Johansson H, Borgström F, Ström O, McCloskey E (2009) FRAX® and its applications to clinical practice. Bone 44:734–743
9. Kanis JA, Johnell O (2005) Requirements for DXA for the management of osteoporosis in Europe. Osteoporos Int 16: 229–238
10. Black DM, Steinbuch M, Palermo L, Dargent-Molina P, Lindsay R, Hoseyni MS, Johnell O (2001) An assessment tool for predicting fracture risk in postmenopausal women. Osteoporos Int 12:519–528
11. Johansson H, Kanis JA, Oden A, Johnell O, McCloskey E (2009) BMD, clinical risk factors and their combination for hip fracture prevention. Osteoporos Int 20:1675–1682
12. Johnell O, Kanis JA, Oden A, Johansson H, De Laet C, Delmas P, Eisman JA, Fujiwara S, Kroger H, Mellstrom D, Meunier PJ, Melton LJ 3rd, O'Neill T, Pols H, Reeve J, Silman A, Tenenhouse A (2005) Predictive value of BMD for hip and other fractures. J Bone Miner Res 20:1185–1194
13. Garton MJ, Cooper C, Reid D (1997) Perimenopausal bone density screening—Will it help prevent osteoporosis? Maturitas 26:35–43
14. Barr RJ, Stewart A, Torgerson DJ, Reid DM (2010) Population screening for osteoporosis risk: a randomised control trial of medication use and fracture risk. Osteoporos Int 21:561–568
15. 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 339:1291–1295
16. 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 (in press)
17. 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 (in press)
18. Kanis JA, Oden A, Johansson H, McCloskey E (2012) Pitfalls in the external validation of FRAX. Osteoporos Int 23:423–431
19. Cummins NM, Poku EK, Towler MR, O'Driscoll OM, Ralston SH (2011) Clinical risk factors for osteoporosis in Ireland and the UK: a comparison of FRAX and QFractureScores. Calcif Tissue Int 89:172–177
20. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, Burckhardt P, Cooper C, Christiansen C, Cummings S, Eisman JA, Fujiwara S, Gluer C, Goltzman D, Hans D, Krieg MA, La Croix A, McCloskey E, Mellstrom D, Melton LJ 3rd, Pols H, Reeve J, Sanders K, Schott AM, Silman A, Torgerson D, van Staa T, Watts NB, Yoshimura N (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
21. Schousboe JT (2008) Cost effectiveness of screen-and-treat strategies for low bone mineral density: How do we screen, who do we screen and who do we treat? Appl Health Econ Health Policy 6:1–18
22. Stewart A, Kumar V, Reid DM (2006) Long-term fracture prediction by DXA and QUS: a 10-year prospective study. J Bone Miner Res 21:413–418
23. Hans D, Krieg MA (2009) Quantitative ultrasound for the detection and management of osteoporosis. Salud Publica Mex 51(suppl 1):S25–S37
24. Bauer DC, Gluer CC, Cauley JA, Vogt TM, Ensrud KE, Genant HK, Black DM (1997) Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women: a prospective study. Arch Intern Med 157:629–634
25. Langton CM, Langton DK (2000) Comparison of bone mineral density and quantitative ultrasound of the calcaneus: site-matched correlation and discrimination of axial BMD status. Br J Radiol 73:31–35
26. Zhu K, Devine A, Prince RL (2009) Quantitative ultrasound measurements predict fracture in older women: a 10-year longitudinal study. Bone 44:S118
27. Moayyeri A, Kaptoge S, Dalzell N, Bingham S, Luben RN, Wareham NJ, Reeve J, Khaw KT (2009) Is QUS or DXA better for predicting the 10-year absolute risk of fracture? J Bone Miner Res 24:1319–1325
28. Garnero P, Hausherr E, Chapuy M-, Marcelli C, Grandjean H, Muller C, Cormier C, Bréart G, Meunier PJ, Delmas PD (1996) Markers of bone resorption predict hip fracture in elderly women: the EPIDOS prospective study. J Bone Miner Res 11:1531–1538
29. Vasikaran S, Eastell R, Bruyère O, Foldes AJ, Garnero P, Griesmacher A, McClung M, Morris HA, Silverman S, Trenti T, Wahl DA, Cooper C, Kanis JA (2011) Markers of bone turnover for the prediction of fracture risk and monitoring of osteoporosis treatment: a need for international reference standards. Osteoporos Int 22:391–420
30. Garnero P, Sornay-Rendu E, Chapuy MC, Delmas PD (1996) Increased bone turnover in late postmenopausal women is a major determinant of osteoporosis. J Bone Miner Res 11:337–349
31. Ralston SH (2005) Genetic determinants of osteoporosis. Curr Opin Rheumatol 17:475–479
32. Arden NK, Baker J, Hogg C, Baan K, Spector TD (1996) The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. J Bone Miner Res 11:530–534
33. Mann V, Ralston SH (2003) Meta-analysis of *COL1A1* Sp1 polymorphism in relation to bone mineral density and osteoporotic fracture. Bone 32:711–717
34. Navarro MC, Sosa M, Del Pino-Montes J, Torres A, Salido E, Saavedra P, Corral-Gudino L, Montilla CA (2007) Collagen type 1 (*COL1A1*) Sp1 binding site polymorphisms is associated with osteoporotic fractures but not with bone density in post-menopausal women from the Canary Islands: a preliminary study. Aging Clin Exp Res 19:4–9
35. Klee EW, Hoppman-Chaney NL, Ferber MJ (2011) Expanding DNA diagnostic panel testing: Is more better? Exp Rev Mol Diagn 11:703–709
36. Brown MA (2005) Genetic studies of osteoporosis—a rethink required. Calcif Tissue Int 76:319–325
37. Barr R, Macdonald H, Stewart A, McGuigan F, Rogers A, Eastell R, Felsenberg D, Glüer C, Roux C, Reid DM (2010) Association between vitamin D receptor gene polymorphisms, falls, balance and muscle power: results from two independent studies (APOSS and OPUS). Osteoporos Int 21:457–466
38. Robbins JA, Schott AM, Garnero P, Delmas PD, Hans D, Meunier PJ (2005) Risk factors for hip fracture in women with high BMD: EPIDOS study. Osteoporos Int 16:149–154
39. Shepstone L, Fordham R, Lenaghan E, Harvey I, Cooper C, Gittoes N, Heawood A, Peters TJ, O'Neill T, Torgerson D, Holland R, Howe A, Marshall T, Kanis JA, McCloskey E (2012) A pragmatic randomised controlled trial of the effectiveness and cost-effectiveness of screening older women for the prevention of fractures: rationale, design and methods for the SCOOP study. Osteoporos Int (in press)
40. Elliott P, Peakman TC (2008) The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol 37:234–244
41. Patel R, Blake GM, Fogelman I (2007) Evaluation of osteoporosis using skin thickness measurements. Calcif Tissue Int 81:442–449
42. Saito M, Marumo K (2010) Collagen cross-links as a determinant of bone quality: a possible explanation for bone fragility in aging, osteoporosis, and diabetes mellitus. Osteoporos Int 21:195–214
43. Langholz B, Thomas DC (1990) Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. Am J Epidemiol 131:169–176
44. Kenis G, Teunissen C, De Jongh R, Bosmans E, Steinbusch H, Maes M (2002) Stability of interleukin 6, soluble interleukin 6 receptor, interleukin 10 and CC16 in human serum. Cytokine 19:228–235
45. McClish DK (1989) Analyzing a portion of the ROC curve. Med Decis Making 9:190–195
46. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 54:17–23
47. Rapsomaniki E, White IR, Wood AM, Thompson SG, Emerging Risk Factors Collaboration (2012) A framework for quantifying net benefits of alternative prognostic models. Stat Med 31:114–130
48. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26:565–574
49. Pepe MS, Feng Z, Gu JW (2008) Comments on "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond," by M. J. Pencina et al. Stat Med. 27:173–181. doi:10.1002/sim.2929
50. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172
51. Schwenkglenks M, Lippuner K (2007) Simulation-based cost-utility analysis of population screening-based alendronate use in Switzerland. Osteoporos Int 18:1481–1491

52. Pencina MJ, D'Agostino RB, Steyerberg EW (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 30:11–21
53. Donaldson MG, Cawthon PM, Schousboe JT, Ensrud KE, Lui LY, Cauley JA, Hillier TA, Taylor BC, Hochberg MC, Bauer DC, Cummings SR, Study of Osteoporotic Fractures (SOF) (2011) Novel methods to evaluate fracture risk models. J Bone Miner Res 26:1767–1773