External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study

Noa Dagan,[1,2] Chandra Cohen-Stavi,[1] Maya Leventer-Roberts,[1,3] Ran D Balicer[1,4]

[1]Clalit Research Institute, Chief Physician's Office, Clalit Health Services, Tel Aviv, Israel

[2]Computer Science Department, Ben Gurion University of the Negev, Be'er Sheba, Israel

[3]Department of Preventive Medicine and Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[4]Epidemiology Department, Ben Gurion University of the Negev, Be'er Sheba, Israel

Correspondence to: N Dagan noa.dgn@gmail.com

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2017;356:i6755
http://dx.doi.org/10.1136/bmj.i6755

Accepted: 08 December 2016

ABSTRACT

OBJECTIVE

To directly compare the performance and externally validate the three most studied prediction tools for osteoporotic fractures—QFracture, FRAX, and Garvan—using data from electronic health records.

DESIGN

Retrospective cohort study.

SETTING

Payer provider healthcare organisation in Israel.

PARTICIPANTS

1 054 815 members aged 50 to 90 years for comparison between tools and cohorts of different age ranges, corresponding to those in each tools' development study, for tool specific external validation.

MAIN OUTCOME MEASURE

First diagnosis of a major osteoporotic fracture (for QFracture and FRAX tools) and hip fractures (for all three tools) recorded in electronic health records from 2010 to 2014. Observed fracture rates were compared to probabilities predicted retrospectively as of 2010.

RESULTS

The observed five year hip fracture rate was 2.7% and the rate for major osteoporotic fractures was 7.7%. The areas under the receiver operating curve (AUC) for hip fracture prediction were 82.7% for QFracture, 81.5% for FRAX, and 77.8% for Garvan. For major osteoporotic fractures, AUCs were 71.2% for QFracture and 71.4% for FRAX. All the tools underestimated the fracture risk, but the average observed to predicted ratios and the calibration slopes of FRAX were closest to 1. Tool specific validation analyses yielded hip fracture prediction AUCs of 88.0% for QFracture (among those aged 30-100 years), 81.5% for FRAX (50-90 years), and 71.2% for Garvan (60-95 years).

CONCLUSIONS

Both QFracture and FRAX had high discriminatory power for hip fracture prediction, with QFracture performing slightly better. This performance gap was more pronounced in previous studies, likely because of broader age inclusion criteria for QFracture validations. The simpler FRAX performed almost as well as QFracture for hip fracture prediction, and may have advantages if some of the input data required for QFracture are not available. However, both tools require calibration before implementation.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Tools for prediction of osteoporotic fractures are recognised by leading guidelines as an important component of osteoporosis prevention but are underutilised

Of the three most studied fracture prediction tools—QFracture, FRAX, and Garvan—QFracture was the only one developed using data from electronic health records

The adaptation of these tools for automatic implementation in external electronic health record systems is not clear, nor is their relative performance

WHAT THIS STUDY ADDS

Automatic computation of all three tools using data from external electronic health records produced similar results, as has been previously reported, for each of the tools separately (tested in separate cohorts of the same age ranges as the derivation cohort of each tool)

When evaluated using one cohort (for which the age ranges of all tools overlap), QFracture and FRAX yielded high discriminatory performance for hip fracture prediction, with QFracture performing slightly better

This performance gap was much smaller than previously reported by reviews, which compared results from validation studies that tested each of the tools using different age ranges

Introduction

Osteoporotic fractures cause major morbidity and mortality, with many people who experience such fractures rapidly deteriorating in health status and experiencing a lower quality of life.[1 2] This poses a substantial economic burden to health systems, patients, and their families.[3] The burden of osteoporotic fractures is expected to increase as populations age, with the incidence of hip fractures reported to increase 30-fold between the ages of 50 and 90 years.[4] Osteoporotic fractures and re-fractures can be prevented and better managed when people at high risk are identified early.[5 6]

Routine scanning of bone mineral density is recommended in all women and in some guidelines also for men, but despite these recommendations, rates of screening remain low, leaving osteoporosis undiagnosed in many patients.[7-10] Furthermore, the criteria used for bone mineral density to identify those at high risk for osteoporotic fractures are not highly sensitive, as more than half of older women with osteoporotic fractures do not meet the bone mineral density criteria for osteoporosis (T score lower than –2.5).[11]

For these reasons, multiple risk assessment tools based on clinical and personal characteristics have been developed in recent years to identify those at high risk for osteoporotic fractures. The most studied tools are the World Health Organization's FRAX, Garvan, and QFracture, which are all freely available online for public use.[12] Each tool has been developed in different contexts, with FRAX and Garvan based on cohort studies using survey and doctor and patient reported data, and QFracture based on data from electronic health records.[4 13-16] The extent to which each tool has been externally validated varies: FRAX has been validated by 26 studies in nine countries, Garvan by six studies in three countries, and QFracture by three studies within the United Kingdom and the Republic of Ireland.[12] These three tools also differ in their complexity in terms of the number of input variables included, with QFracture using 26 variables, FRAX using 11, and Garvan using five. In addition, FRAX and Garvan offer predictions with or without the input of a pre-existing bone mineral density measurement, whereas QFracture does not include bone mineral density in its algorithm. Supplement 1 summarises the basic features of the three tools.

Although many predictive tools have been developed, few are used to support clinical decision making to identify patients at high risk for osteoporotic fractures.[17] With increasing use of electronic health record systems there is the potential to produce automated personalised fracture risk scores to better direct treatment and reduce the overall burden of osteoporotic fractures. These risk scores can be presented both directly to the patients and made accessible to their doctors through the electronic health record system. Several studies have shown the benefits of improved management of osteoporosis and fracture prevention from electronic health records or electronic software based decision support implementations.[18 19] In determining which of the various prediction tools is adaptable for automatic implementation using electronic health record data, the predictive performance of each tool (both discrimination and calibration), the validation results in various populations, and the availability of types of data required for the tool must be considered.

Although numerous reviews have compared FRAX, Garvan, and QFracture,[12 20-22] to the best of our knowledge their performance has not been directly compared within one population. A few studies have directly compared two of the tools in the same population.[13 23-25] However, the only study to evaluate tool performance in a large population and among both men and women compared old versions of QFracture and FRAX.[13] Several other studies purported to compare two tools but did not validate the predicted risk with observed events over a subsequent follow-up period.[26-29]

Several substantial pitfalls have been highlighted both in comparisons of performance across various tools and in external validations of specific tools. Problems with missing input variables, sample size, and the number of outcome events were noted as limiting the ability of validation studies to provide generalisable results and full validations of original tools.[17] Most studies aiming to compare measures of tool performance were reviews or meta-analyses (not direct comparisons within one population) that relied on results of specific tool validations. The comparability of these validations has been critiqued, because different inclusion criteria and follow-up periods might affect their results.[30] Age, for example, is a major determinant of fracture risk, and thus the choice of age ranges included in specific validation studies were suggested to substantially affect the reported performance of the tool. Furthermore, many of these validations did not present a comparison between the validation and derivation populations used to develop the tools, to shed light on the kind of validation they contribute—ranging between “reproducibility” (evaluating the tool within a population with similar characteristics) and “transportability” (evaluating the tool within a population of different characteristics).[31] The lack of consistency in study designs among previous validation studies presents challenges in the ability to draw meaningful conclusions about which tool offers the best performance.

We compared the performance of the three most commonly studied fracture prediction tools in a single, large population when computed automatically based on electronic health record data. We also conducted a tool specific external validation in an independent population to evaluate the performance of the tools in populations with the same age range as those in which they were developed, thus allowing comparison with previously reported performance.

Methods

Setting

This study used electronic health record data from Clalit Health Services, the largest of four national health funds in Israel. All Israeli residents are covered by one of the health funds and can switch between them at any time; however, switching rates are relatively low—about 1% annually[32]—which allows for consistent longitudinal follow-up. Clalit Health Services is both a healthcare insurer and a provider, thus financing and supplying services to its 4.3 million members, which make up more than half of Israel’s population. Membership of Clalit Health Services comprises the general population, but for historical reasons the organisation has a slightly larger proportion of the older population and those from a lower socioeconomic class.[33]

Study design

In this historical prospective cohort study we compared the probability of hip fracture over five years using FRAX, QFracture, and Garvan, as well as the probability of major osteoporotic fractures over five years using FRAX and QFracture, computed on 1 January 2010 (index date), with fracture events observed up to 31 December 2014 (follow-up period).

In the first part of this study we compared the performance of the three tools, and thus selected a population in which the reported age ranges for all tools overlap. This comparative analysis was conducted for risk of hip fracture. Because the definition used by Garvan for major osteoporotic fractures is much broader than the one used by QFracture and FRAX (vertebral, distal radius, proximal humerus, or hip), we conducted additional analyses only between QFracture and FRAX to compare the performance for predicting major osteoporotic fractures. In the second part of the study we conducted a tool specific external validation for performance in predicting fractures, using cohorts with varying age ranges for each tool.

Study population

The comparative analysis was performed among members of Clalit Health Services aged 50 to 90 years as of the index date, who had at least three years of continuous membership before the index date and through the follow-up period or until death (see fig 1). Therefore, the cohort did not include those who were lost to follow-up. Although FRAX was developed among a population that excluded patients who were treated for osteoporosis,[34] the other two tools were not, and thus for comparative purposes, treatment for osteoporosis was not used as an exclusion criterion (a population of non-treated patients was evaluated in a separate sensitivity analysis).

For the tool specific external validation analyses we used specified age ranges corresponding to those chosen in the original tool development studies: the QFracture analysis included members aged 30-100 years,[16] the FRAX analysis included members aged 50-90 years,[4] and the Garvan analysis included members aged 60-95 years[14] (the official calculator computes risk for 50 or more years, which was the reason why we chose age 50 as the lower limit for the comparative analysis).[35] The rest of the inclusion and exclusion criteria did not differ from those used in the comparative analysis (see fig 1).

To account for real world settings, in the populations of both analyses we included those who died during the follow-up period.

Data sources

The electronic health record data at Clalit Health Services contain comprehensive administrative and clinical data. These include demographic information, diagnoses given in a community or a hospital setting, chronic disease and oncology registries, laboratory results, written prescriptions and prescriptions dispensed, clinical markers (eg, body mass index, smoking status), medical procedures, and imaging data.

Input variables

Input variables included clinical status, prescription drug use, and demographic characteristics, according to the variables used in each of the tools. Supplement 2 lists the codes used to define diagnoses and drug based variables.

To provide as comprehensive data as possible for the prediction tools, we based all input variables of the three prediction tools on information that was last documented as of the index date. Most study variables represent chronic conditions and were consequently taken with no date limitation before the index date. For variables that could potentially change over time (including body mass index, smoking status, alcoholism, nursing home residency, history of falls, and drug use), we took the last relevant documented history with no time limitation, and we also conducted a sensitivity analysis in which the extraction of such variables was limited to two years before the index date. The sensitivity analysis was performed to establish the implications of not limiting the time from which variable data were taken.

Clinical diagnoses—Input variables for diagnosis included history of osteoporotic fractures, secondary osteoporosis, dementia, Parkinson's disease, epilepsy, diabetes and other endocrine conditions, obstructive airways disease, cardiovascular disease, malabsorption, chronic liver disease, chronic kidney disease, rheumatoid arthritis, systemic lupus erythematosus, and documented history of falls. We extracted these diagnoses from community and hospital records, as well as from the Clalit Health Services chronic disease registry, when appropriate. Diagnoses were defined based on the International Classification of Diseases, ninth revision (ICD-9), International Classification of Primary Care (ICPC), and chronic disease registry codes. Diagnoses made in the community setting were further validated based on doctors' accompanying free text diagnosis description, available only in the community records.

Body mass index—This variable was computed from documented weight and height measurements.

Smoking status— In the Clalit Health Services database, smoking status is defined as non-smokers, former smokers, or current smokers. In QFracture, three current smoking categories are provided according to the number of cigarettes smoked daily.[36] To avoid the bias of categorising patients in one of the outlying categories, we assigned Clalit Health Services "current smokers" to the middle category (10-19 cigarettes daily). For FRAX's two category smoking status, we assigned former smokers in our population to the non-smokers category, as was done in the cohorts used to develop FRAX.[37 38]

Alcohol consumption—The Clalit Health Services database does not include information on alcohol intake, so we defined alcohol consumption as a dichotomous (yes or no) variable, based on diagnoses of alcoholism or alcohol induced chronic complications (ICD-10 codes for related psychiatric diagnoses were used for alcoholism in addition to the ICD-9 and ICPC codes). Of the five alcohol consumption categories provided by QFracture, we assigned individuals with alcohol related diagnoses to the fourth level category (7-9 units daily, using the UK's definition of alcohol unit), since the lower categories were unlikely to cause alcohol related complications, and the highest category might overestimate the alcohol consumption for some of the relevant population. Given the inability to distribute individuals without alcohol related diagnoses to the various alcohol consumption levels, we assigned them to the "none" (ie, no alcohol intake) category.

Family history of fractures—A family history of osteoporosis and hip fractures was defined by diagnosis codes indicating such a history and by searching the medical records of the parents of study members, when the family connection was defined within the electronic health record and either parent was a member of Clalit Health Services.

Medication use—We computed variables for medication use, considered only in QFracture and FRAX, based on pharmacy dispensing data. Glucocorticoid use was defined differently by these tools—two prescription records in the last six months by QFracture versus current or past use for more than three months by FRAX. We therefore computed glucocorticoid use as two separate variables. Purchases of antidepressants and hormone replacement therapy medications were included only in the QFracture analyses.

Nursing home care—We considered an individual to be a nursing home resident when the patient's primary clinic or treating doctor were administratively defined as institutional positions.

In cases where there was no documentation of body mass index, weight, or smoking status before the index date (the only variables for which missing data could be identified), we used multiple imputation to complete these values. We also performed a complete case sensitivity analysis without imputed variables.

Outcome variables

Outcome variables included both hip fracture and major osteoporotic fractures, which were defined as fractures of the hip, vertebrae, distal radius, or proximal humerus. These variables were defined based on the records for clinical diagnoses.

Predictive tool risk computation

We computed the five year risk according to QFracture (2012 version) and Garvan based on their full tool equations.[14 16 36 39] To ensure correct automation, we manually validated a few dozen cases against the official calculator sites. Since the current FRAX equations are not published by the authors, we used the FRAX 10 year probability charts calibrated for Israel, stratified by sex, age, body mass index, and number of clinical risk factors, as supplied by the official FRAX site.[37] We multiplied the 10 year probabilities by 0.5 to convert to five year probabilities. The justification for this transformation was established by examining the rate of osteoporotic fracture events over a 10 year period, between 2005 and 2014 (see supplement 3 for further details). All tools were computed without the input of bone mineral density because QFracture does not include this variable and data on bone mineral density were limited in the electronic health record system for the study years.

Statistical analysis

To compare across the three tools, which were developed using different modelling methods, we used the provided risk probabilities for each tool respectively and treated the outcome as if it were a binary variable (fracture or no fracture). This decision was also guided by the clinical application of these risk predictions tools—that doctors and patients perceive the output as risk for the relevant follow-up period, regardless of the methods used to produce it. The closed cohort design facilitated this strategy of treating the outcome as a binary variable, because there was a known outcome for all study members in a fixed follow-up period of five years.[40 41] Since it is clinically important to test the accuracy of the predicted probability of fracture both for people who survive the follow-up period and for those with shorter life spans, we did not account for shortening of the follow-up period due to death.

To evaluate the overall ability of each tool to discriminate between those at low risk and those at high risk we used the area under the receiver operating curve (AUC) in both the comparative and the tool specific external validation analyses. We calculated other discriminatory measures—sensitivity, specificity, positive and negative predictive values, accuracy, and error—for the top 10% and 20% highest risk cut-offs of each tool. In three separate sensitivity analyses we further evaluated the discrimination measures in the comparative analysis: limitation of the time range of variables with less chronic nature, complete case analysis, and a subpopulation that excluded patients who were being treated for osteoporosis in the two years before the index date.

Since the AUC is considered a somewhat crude overall discriminatory measure, that might overlook the contribution of specific risk factors that are not prevalent in the population but are potentially clinically significant for an individual patient's risk prediction,[30] we conducted a reclassification analysis between the two tools with the highest AUCs in the comparative analysis. We report the total numbers of patients classified as low risk and high risk using a top 10% cut-off level for the two tools, as well as measures of net reclassification index analysis.[42] The net reclassification index for events is the rate of events that were correctly reclassified as high risk by the tested tool (usually the tool that incorporates more risk factors) minus the rate of events wrongly reclassified as low risk. The net reclassification index for non-events is the parallel measure, and is the rate of non-events that were correctly reclassified as low risk minus the rate of non-events that were wrongly reclassified as high risk. The overall net reclassification index is the combination of net reclassification index for events and net reclassification index for non-events, whereas the more intuitive weighted net reclassification index is the combination of the same values weighted by the relative size of the groups they represent.[43] We calculated standard errors for all net reclassification index values.[44]

We assessed the calibration of each tool by comparing the average predicted risk with the observed percentage of those who experienced fractures over the follow-up period, stratified by age groups and separately by 10ths of fracture risk. To provide calibration measures that are not based on grouping of individuals into strata, we compiled calibration aparametric curves, calibration slopes, and calibration-in-the-large values[45] using functions by Harrell et al[46] and added these to calibration plots.

Multiple imputation was conducted using 10 iterations and 20 multiple imputations, thus creating 20 full datasets, using functions by Van Buuren et al.[47] We performed all analyses separately on each of these imputed datasets and averaged these to determine the final performance measures. A 95% confidence interval for AUC measures of specific prediction tools as well as for the differences between tools was calculated using Rubin's rules for variance estimation in multiple imputed datasets[45 48] (by taking into account both the AUC variance of 1000 bootstrap samples within each imputed dataset and the variance of the 20 average AUCs between the imputed datasets). Owing to the nature of the net reclassification index analysis, this analysis was only based on one random imputed dataset. Plots were created using a combined dataset that included all of the separate imputed datasets.

All analyses were conducted using R, CRAN version 3.2.2 (mice,[47] ROCR,[49] and rms[46] packages).

Patient involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to study participants or the relevant patient community.

Results

As of 1 January 2010, 1 085 104 members of Clalit Health Services were aged 50 to 90 years. Of those, we excluded 30 289 (2.8%) because they did not meet the criteria for continuous membership (fig 1). The final population for the comparative analysis consisted of 1 054 815 people (54.6% women). This population included 28 091 (2.7%) who experienced a hip fracture and 81 564 (7.7%) who experienced a major osteoporotic fracture during the follow-up period (table 1). Supplement file 4 provides specific fracture rates stratified by age and sex. A total of 113 591 (10.8%) people died during the follow-up period. The average length of follow-up was 4.73 years, with 4 990 557 total person years of follow-up. Overall, 54 849 (5.2%) of the records were imputed for weight and body mass index values, and 34 921 (3.3%) were imputed for smoking status. Table 1 lists the characteristics of the study population by input variables of the three prediction tools, the outcome fracture rates, and which variables were included in each tool.

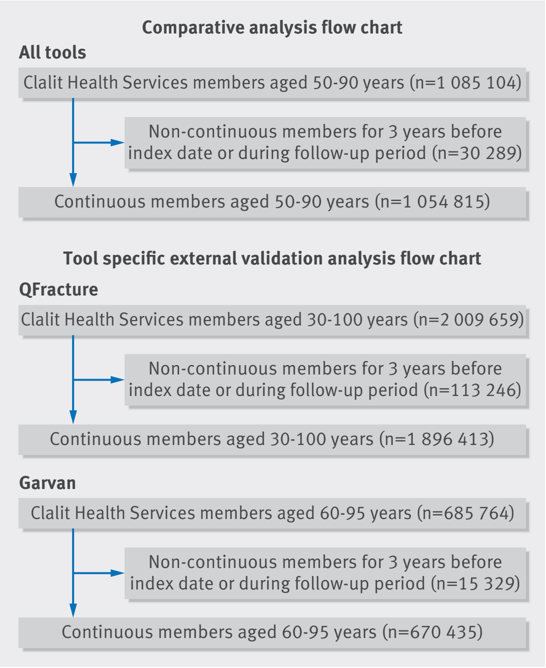
Comparative analysis flow chart
All tools
Clalit Health Services members aged 50-90 years (n=1 085 104)
Non-continuous members for 3 years before index date or during follow-up period (n=30 289)
Continuous members aged 50-90 years (n=1 054 815)
Tool specific external validation analysis flow chart
QFracture
Clalit Health Services members aged 30-100 years (n=2 009 659)
Non-continuous members for 3 years before index date or during follow-up period (n=113 246)
Continuous members aged 30-100 years (n=1 896 413)
Garvan
Clalit Health Services members aged 60-95 years (n=685 764)
Non-continuous members for 3 years before index date or during follow-up period (n=15 329)
Continuous members aged 60-95 years (n=670 435)

Fig 1 | Population flowchart for comparative and tool specific external validation analyses (FRAX external validation population is same as population used for comparative analysis)

In examining the comparative performance across the tools, QFracture had the highest AUC for hip fracture prediction (82.7%, 95% confidence interval 82.4% to 82.9%), followed closely by FRAX (81.5%, 81.3% to 81.7%). Garvan's AUC for hip fracture prediction (77.8%, 77.5% to 78.1%) was lower (table 2). The confidence interval for the difference between the QFracture and FRAX AUCs was 1.0-1.3%, whereas the confidence interval for the difference between the QFracture and Garvan AUCs was 4.7-5.1%. Among the highest 10% risk for hip fracture, as predicted in 2010, QFracture identified 45.1% (sensitivity) of those who went on to experience a hip fracture, FRAX 43.6%, and Garvan 36.9%. By targeting those in the 20% highest risk for hip fracture, QFracture identified 68.9% of hip fractures, FRAX 65.8%, and Garvan 57.1%. The specificity and negative predictive values were high and comparable for all three tools (table 2).

The QFracture and FRAX discriminatory measures for prediction of major osteoporotic fractures were lower than those for hip fracture prediction. AUCs for both tools were close (QFracture: 71.2%, 71.0% to 71.4%; FRAX: 71.4%, 71.2% to 71.6%, table 2). The confidence interval for the difference between the FRAX and QFracture AUCs was 0.1-0.3%. The sensitivity for the top 10% highest risk group was 26.7% for QFracture, compared with 29.0% for FRAX, and the positive predictive value was 20.7% for QFracture and 22.4% for FRAX. Figure 2 presents the comparisons of the receiver operating curves for all three tools in predicting hip fractures and for QFracture and FRAX in predicting major osteoporotic fractures.

The results from the three sensitivity analyses were consistent on the relative performance of the tools in their discriminatory measures to that of the main analysis. Analyses limiting variable data collection to the two years before the index date can be found in supplement 5. Complete case analyses are in supplement 6, and analyses of non-treated patients in supplement 7.

We conducted a reclassification analysis between QFracture and FRAX (the two tools that yielded the highest AUCs) to compare how these tools categorised patients into low risk and high risk groups. QFracture, which incorporates more risk factors than FRAX in its prediction model, was considered the "reclassifying" model in the analysis, so that we could evaluate the prediction increment offered by its added risk factors (table 3). The net proportion of patients who experienced a hip fracture and were correctly reclassified as high risk by QFracture compared with FRAX was 1.50% (net reclassification index for events). The net proportion of patients who experienced a major osteoporotic fracture and were correctly reclassified as high risk by QFracture was −2.31% (net reclassification index for events). For both types of outcomes, the change in the correct reclassification of non-events was less than 0.2%. The net changes in the proportion of patients assigned a more appropriate risk category for prediction of hip

Table 1 | Characteristics of comparative analysis population, by each of the input variables included in QFracture, FRAX, and Garvan

Input variables*	No (%) in study population	Major osteoporotic fracture†	Hip fracture†	QFracture	FRAX	Garvan
Overall	1 054 815 (100)	81 564 (7.7)	28 091 (2.7)			
Age group (years):						
50-59	401 035 (38.0)	15 324 (3.8)	1994 (0.5)	V	V	V
60-69	299 305 (28.4)	17 634 (5.9)	3689 (1.2)			
70-79	222 475 (21.1)	25 871 (11.6)	9465 (4.3)			
80-89	132 000 (12.5)	22 735 (17.2)	12 943 (9.8)			
Sex:						
Men	478 825 (45.4)	23 268 (4.9)	8996 (1.9)	V	V	V
Women	575 990 (54.6)	58 296 (10.1)	19 095 (3.3)			
Ethnicity:						
Black African	12 813 (1.2)	540 (4.2)	135 (1.1)	V	–	–
White	1 042 002 (98.8)	81 024 (7.8)	27 956 (2.7)			
Nursing home residency:						
No	1 041 516 (98.7)	80 155 (7.7)	27 191 (2.6)	V	–	–
Yes	13 299 (1.3)	1409 (10.6)	900 (6.8)			
Body mass index category:						
Obese	309 128 (29.3)	24 417 (7.9)	6833 (2.2)	V	V	V‡
Overweight	405 416 (38.4)	30 296 (7.5)	10 101 (2.5)			
Normal	276 206 (26.2)	23 190 (8.4)	9467 (3.4)			
Underweight	9216 (0.9)	1113 (12.1)	601 (6.5)			
Missing	54 849 (5.2)	2548 (4.6)	1089 (2.0)			
Smoking category:						
Non-smoker	681 698 (64.6)	57 859 (8.5)	20 033 (2.9)	V	V	–
Former smoker	163 185 (15.5)	11 427 (7.0)	3901 (2.4)			
Current smoker	175 011 (16.6)	10 171 (5.8)	3130 (1.8)			
Missing	34 921 (3.3)	2107 (6.0)	1027 (2.9)			
Alcoholism:						
No	1 043 558 (98.9)	80 487 (7.7)	27 623 (2.6)	V	V	–
Yes	11 257 (1.1)	1077 (9.6)	468 (4.2)			
Parental hip fracture:						
No	1 036 081 (98.2)	80 698 (7.8)	27 949 (2.7)	V	V	–
Yes	18 734 (1.8)	866 (4.6)	142 (0.8)			
Parental osteoporotic fracture:						
No	990 039 (93.9)	79 026 (8.0)	27 776 (2.8)	V	–	–
Yes	64 776 (6.1)	2538 (3.9)	315 (0.5)			
Major osteoporotic fracture:						
No	984 128 (93.3)	60 562 (6.2)	19 208 (2.0)	V	V	–
Yes	70 687 (6.7)	21 002 (29.7)	8883 (12.6)			
No of fractures after age 50 years:						
0	898 475 (85.2)	50 897 (5.7)	15 932 (1.8)	–	–	V
1	119 329 (11.3)	19 766 (16.6)	7408 (6.2)			
2	27 171 (2.6)	7307 (26.9)	3111 (11.4)			
≥3	9840 (0.9)	3594 (36.5)	1640 (16.7)			
History of a fall:						
No	990 681 (93.9)	68 371 (6.9)	21 406 (2.2)	V	–	–
Yes	64 134 (6.1)	13 193 (20.6)	6685 (10.4)			
No of falls in past year:						
0	1 031 443 (97.8)	76 082 (7.4)	25 320 (2.5)	–	–	V
1	9746 (0.9)	2292 (23.5)	1214 (12.5)			
2	9986 (0.9)	2272 (22.8)	1106 (11.1)			
≥3	3640 (0.3)	918 (25.2)	451 (12.4)			
Secondary osteoporosis§:						
No	984 123 (93.3)	74 259 (7.5)	25 408 (2.6)	–	V	–
Yes	70 692 (6.7)	7305 (10.3)	2683 (3.8)			
Dementia:						
No	1 030 739 (97.7)	78 019 (7.6)	25 885 (2.5)	V	–	–
Yes	24 076 (2.3)	3545 (14.7)	2206 (9.2)			
Parkinson's disease:						
No	1 032 213 (97.9)	78 322 (7.6)	26 305 (2.5)	V	–	–
Yes	22 602 (2.1)	3242 (14.3)	1786 (7.9)			

(Continued)

Table 1 | Characteristics of comparative analysis population, by each of the input variables included in QFracture, FRAX, and Garvan

Input variables*	No (%) in study population	Major osteoporotic fracture†	Hip fracture†	QFracture	FRAX	Garvan
Epilepsy:						
No	996 622 (94.5)	74 672 (7.5)	25 450 (2.6)	V	–	–
Yes	58 193 (5.5)	6892 (11.8)	2641 (4.5)			
Type 1 diabetes:						
No	1 053 791 (99.9)	81 448 (7.7)	28 046 (2.7)	V	–	–
Yes	1024 (0.1)	116 (11.3)	45 (4.4)			
Type 2 diabetes:						
No	765 591 (72.6)	54 814 (7.2)	17 451 (2.3)	V	–	–
Yes	289 224 (27.4)	26 750 (9.2)	10 640 (3.7)			
Other endocrine disorders:						
No	1 005 799 (95.4)	76 155 (7.6)	26 060 (2.6)	V	–	–
Yes	49 016 (4.6)	5409 (11.0)	2031 (4.1)			
Cancer history:						
No	913 510 (86.6)	66 605 (7.3)	22 106 (2.4)	V	–	–
Yes	141 305 (13.4)	14 959 (10.6)	5985 (4.2)			
Obstructive airways disease:						
No	893 999 (84.8)	65 236 (7.3)	22 149 (2.5)	V	–	–
Yes	160 816 (15.2)	16 328 (10.2)	5942 (3.7)			
Cardiovascular disease:						
No	756 649 (71.7)	50 090 (6.6)	14 494 (1.9)	V	–	–
Yes	298 166 (28.3)	31 474 (10.6)	13 597 (4.6)			
Malabsorption:						
No	1 042 869 (98.9)	80 355 (7.7)	27 671 (2.7)	V	–	–
Yes	11 946 (1.1)	1209 (10.1)	420 (3.5)			
Chronic liver disease:						
No	1 033 492 (98.0)	79 269 (7.7)	27 250 (2.6)	V	–	–
Yes	21 323 (2.0)	2295 (10.8)	841 (3.9)			
Chronic renal disease:						
No	971 965 (92.1)	72 096 (7.4)	23 396 (2.4)	V	–	–
Yes	82 850 (7.9)	9468 (11.4)	4695 (5.7)			
Rheumatoid arthritis:						
No	1 028 482 (97.5)	78 045 (7.6)	26 828 (2.6)	V	V	–
Yes	26 333 (2.5)	3519 (13.4)	1263 (4.8)			
Systemic lupus erythematosus:						
No	1 052 835 (99.8)	81 296 (7.7)	28 003 (2.7)	V	–	–
Yes	1980 (0.2)	268 (13.5)	88 (4.4)			
Drug purchases¶:						
Glucocorticoids:						
No	1 027 475 (97.4)	77 593 (7.6)	26 745 (2.6)	V	V	–
Yes	27 340 (2.6)	3971 (14.5)	1346 (4.9)			
Antidepressants:						
No	951 080 (90.2)	67 760 (7.1)	22 351 (2.4)	V	–	–
Yes	103 735 (9.8)	13 804 (13.3)	5740 (5.5)			
Hormone replacement therapy:						
Yes	8663 (0.8)	416 (4.8)	70 (0.8)	V	–	–
No	1 046 152 (99.2)	81 148 (7.8)	28 021 (2.7)			

V=variables used as input information for specified tool.

*Values within each input variable are sorted by predicted fracture rate—ie, variable's value that has lowest predicted risk (as defended by prediction tools) appears first.

†Fracture rate during follow-up period (2010-14), within population of each subgroup.

‡Garvan uses a weight instead of body mass index.

§Defined by any of following: type 1 diabetes, osteogenesis imperfecta, hyperthyroidism, hypogonadism, premature menopause, malabsorption, and chronic liver disease.

¶Numbers were calculated using QFracture's definition of drug purchase—at least two purchase records in six months before index date. In the case of glucocorticoid use, which is also used by FRAX, the calculation is based on a history of at least 90 days of use (extracted by number of days covered by past purchase records) and resultant numbers, which were similar to those of the QFracture variable, are not presented.

fracture and major osteoporotic fracture by QFracture were 0.08% and –0.36%, respectively.

Table 4 presents the absolute probabilities of hip fracture that were calculated by each of the three tools, and the calibration of these probabilities with the absolute fracture rates that were observed over the five year follow-up period, by sex and age groups. A majority of the observed-to-predicted ratios for hip fractures were greater than 1, indicating underestimation of the risk by all three tools for both men and women and in almost all age groups. The QFracture and Garvan ratios presented a consistent downward trend with the increase in age groups but were steadier across the different age groups for FRAX. The risk underestimation was most prominent for women in Garvan. In addition, Garvan was the only tool to assign lower mean predicted

Table 2 | Comparison of discriminatory measures between QFracture, FRAX, and Garvan of top 10% and 20% high risk score cut-offs by each tool. Values are percentages unless stated otherwise

	Denominator		QFracture		FRAX		Garvan	
Discriminatory measures*	Top 10% risk	Top 20% risk	Measure for top 10% risk	Measure for top 20% risk	Measure for top 10% risk	Measure for top 20% risk	Measure for top 10% risk	Measure for top 20% risk
Hip fractures:								
AUC	NA	NA	82.7		81.5		77.8	
Absolute risk cut-off	NA	NA	4.0	1.8	4.3	2.6	2.7	1.1
Sensitivity	28 091	28 091	45.1 (12 679.6)	68.9 (19 347.8)	43.6 (12 257.4)	65.7 (18 469.8)	36.9 (10 363.9)	57.1 (16 048.6)
Specificity	1 026 724	1 026 724	91.0 (933 922.6)	81.3 (835 108.8)	90.9 (933 500.3)	81.3 (834 230.8)	90.7 (931 606.8)	81.0 (831 809.6)
PPV	105 481	210 963	12.0 (12 679.6)	9.2 (19 347.8)	11.6 (12 257.4)	8.8 (18 469.8)	9.8 (10 363.9)	7.6 (16 048.6)
NPV	949 334	843 852	98.4 (933 922.6)	99.0 (835 108.8)	98.3 (933 500.3)	98.9 (834 230.8)	98.1 (931 606.8)	98.6 (831 809.6)
Accuracy	1 054 815	1 054 815	89.7 (946 602.2)	81.0 (854 456.6)	89.7 (945 757.7)	80.8 (852 700.5)	89.3 (941 970.7)	80.4 (847 858.2)
Error	1 054 815	1 054 815	10.3 (108 212.8)	19.0 (200 358.4)	10.3 (109 057.3)	19.2 (202 114.5)	10.7 (112 844.3)	19.6 (206 956.8)
Major osteoporotic fractures								
AUC	NA	NA	71.2		71.4		–	–
Absolute risk cut-off	NA	NA	6.7	3.9	8.5	5.5	–	–
Sensitivity	81 564	81 564	26.7 (21 777.0)	46.4 (37 879.9)	29.0 (23 628.9)	47.1 (38 446.9)	–	–
Specificity	973 251	973 251	91.4 (889 547.0)	82.2 (800 167.8)	91.6 (891 398.9)	82.3 (800 734.9)	–	–
PPV	105 481	210 963	20.7 (21 777.0)	18.0 (37 879.9)	22.4 (23 628.9)	18.2 (38 446.9)	–	–
NPV	949 334	843 852	93.7 (889 547.0)	94.8 (800 167.8)	93.9 (891 398.9)	94.9 (800 734.9)	–	–
Accuracy	1 054 815	1 054 815	86.4 (911 324.0)	79.4 (838 047.7)	86.7 (915 027.8)	79.6 (839 181.8)	–	–
Error	1 054 815	1 054 815	13.6 (143 491.0)	20.6 (216 767.3)	13.3 (139 787.2)	20.4 (215 633.2)	–	–

NA=not applicable; AUC=area under receiver operating characteristic curve (C statistic); PPV=positive predictive value; NPV=negative predictive value.
Analyses comparing performance for predicting major osteoporotic fractures were conducted only between QFracture and FRAX because Garvan's definition for major osteoporotic fractures is much broader than either tool.
Numbers in parentheses are numerators for measure; numbers contain a decimal component because they are averaged between imputed datasets.
*Assessed with five years of follow-up.

probabilities for women compared with men in the same age groups. The observed-to-predicted ratios by 10ths of risk and sex were also more consistent for FRAX compared with QFracture and Garvan, which presented declining ratios as risk increased (table 5). Figure 3 presents a calibration plot, presenting the observed and predicted rates for each 10th of risk, along with aparametric calibration curves, calibration slopes, and calibration-in-the-large values.

The tool specific external validation analyses consisted of three different cohorts (fig 1): the FRAX validation population was identical to the comparison analysis population (members aged 50-90 years), the QFracture population consisted of 1 896 413 members,

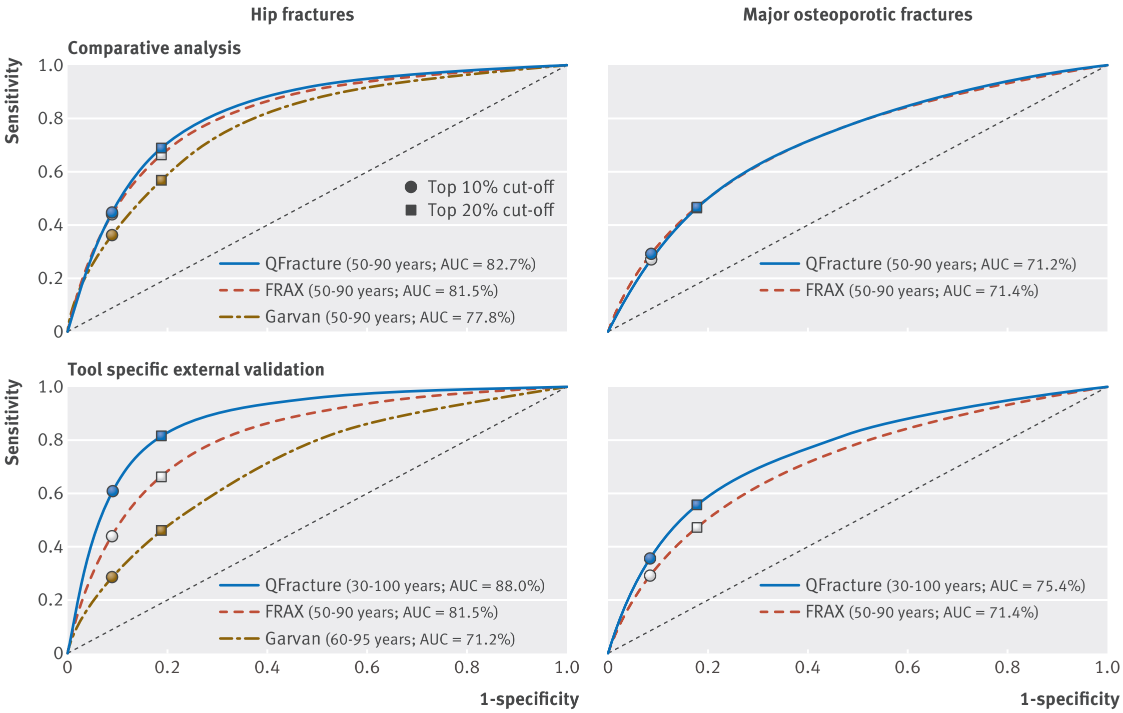

Fig 2 | Receiver operating curves of QFracture, FRAX, and Garvan predictive tools for hip and major osteoporotic fractures during five years of follow-up in comparative and tool specific external validation analyses

Table 3 | Reclassification analysis* for QFracture compared with FRAX

Hip fractures					Major osteoporotic fractures				
QFracture	FRAX		Reclassification measures		QFracture	FRAX		Reclassification measure	
1 054 815 (total population)	Low risk†, 949 332 (90% of population)	High risk†, 105 483 (10% of population)			1 054 815 (total population)	Low risk†, 949 332 (90% of population)	High risk†, 105 483 (10% of population)		
Low risk†, 949 332 (90% of population)	898 268‡ (I) (correctly classified by both models)	35 654‡ (II) (correctly reclassified)	NRI-ne (SE)	0.04% (0.03%)	Low risk†, 949 332 (90% of population)	856 412‡ (I) (correctly classified by both models)	33 123‡ (II) (correctly reclassified)	NRI-ne (SE)	−0.19% (0.03%)
	12 759§ (III) (misclassified by both models)	2651§ (IV) (incorrectly reclassified)	NRI-e (SE)	1.50% (0.27%)		51 863§ (III) (misclassified by both models)	7934§ (IV) (incorrectly reclassified)	NRI-e (SE)	−2.31% (0.14%)
High risk†, 105 483 (10% of population)	3072§ (V). (correctly reclassified)	9609§ (VI) (correctly classified by both models);	NRI (SE)	1.54% (0.27%)	High risk†, 105 483 (10% of population)	6051§ (V) (correctly reclassified)	15 716§ (VI) (correctly classified by both models)	NRI (SE)	−2.50% (0.15%)
	35 233‡ (VII) (incorrectly reclassified)	57 569‡ (VIII) (misclassified by both models)	WNRI (SE)	0.08% (0.05%)		35 006‡ (VII) (incorrectly reclassified)	48710‡ (VIII) (misclassified by both models)	WNRI (SE)	−0.36% (0.05%)

NRI=net reclassification index; NRI-e=net reclassification for events; NRI-ne=net reclassification for non-events; WNRI=weighted net reclassification index; SE=standard error.
NRI calculated as NRI-e+NRI-ne.
NRI-ne calculated as (V−IV)/(III+IV+V+VI.
NRI-e calculated as (II−VI)/(I+II+VII+VIII).
WNRI calculated as NRI-e×((III+IV+V+VI)/IX)+NRI-ne×((I+II+VII+VIII)/IX).
*Calculated with five years of follow-up.
†High risk group defined by each model as study participants who received a risk score in top 10% of risk, and the low risk as the 90% who did not.
‡Non-events.
§Events (people who sustained a fracture).

aged 30-100 years, and the Garvan population included 670 435 members, aged 60-95 years. The population of the QFracture external validation included 31 709 (1.7%) individuals who experienced a hip fracture and 99 058 (5.2%) individuals who experienced a major osteoporotic fracture during the follow-up period. The corresponding rates for the population of the Garvan external validation were 27 897 (4.2%) and 68 859 (10.3%), respectively. Supplement 8 provides a comparison of the prevalence of the risk factors between the populations used to develop the tools (derivation cohorts), and the population of the tool specific external validations in the current study for QFracture[16] and FRAX.[38] The prevalence of risk factors as defined in the final Garvan model were not available for the original Garvan population.[14 15] The current study's QFracture tool specific population was relatively older than QFracture's derivation cohort and was characterised by a greater prevalence (or greater capture rates) of most risk factors. In contrast, the current study's FRAX tool specific population was similar in age to FRAX's derivation cohort, with a smaller share of women and lower prevalence (or lower capture rates) of risk factors.

AUC values for hip fracture in the validation analyses were 88.0% (95% confidence interval 87.8% to 88.2%) for QFracture, 81.5% (81.3% to 81.7%) for FRAX, and 71.2% (70.9% to 71.5%) for Garvan (table 6). The Garvan hip fracture tool was the only one to present sex specific AUC and sensitivity values that were both higher than the overall values. Figure 2 presents the comparisons of the receiver operating characteristic curves for the tool specific external validations. Supplement 9 provides calibration analyses for age and 10ths of risk groups for each of the tool specific external validation cohorts.

Discussion

This study included over one million adults aged 50-90 in a single, general population and directly compared the three most studied fracture prediction tools in an electronic health record system. The discriminatory performance according to the area under the receiver operating curve (AUC) of hip fracture scores for both FRAX and QFracture was high, with the latter performing slightly better, followed by a moderate performance of Garvan. Discriminatory measures for the prediction of major osteoporotic fractures were lower overall than for hip fracture prediction, with very close AUC measures for FRAX and QFracture. Three different sensitivity analyses (see supplements 5-7) examining the impact of input data definitions as well as a different population definition among patients naïve to osteoporosis treatment, have all supported these findings. Given that small differences in the overall AUC (as observed between QFracture and FRAX) may not reflect the entire difference in the discriminative performance for individual patients with a unique set of risk factors, we evaluated the reclassification of individuals between these tools. In examining the value gained from the additional risk factors included in QFracture compared with FRAX, reclassification analysis showed that QFracture had an overall 0.08% net increase and a

Table 4 | Calibration* of observed versus predicted hip fracture rates, by sex and age groups

	Women				Men			
Age range	No of people	Hip fracture rate (%) (No with first hip fracture)	Mean (SD) predicted probability (%)	Observed to predicted ratio	No of people	Hip fracture rate (%) (No with first hip fracture)	Mean (SD) predicted probability (%)	Observed to predicted ratio
QFracture:								
50-54	103 964	0.4 (389)	0.1 (0.002)	3.7	93 755	0.4 (415)	0.1 (0.007)	3.5
55-59	106 372	0.6 (657)	0.2 (0.004)	3.0	96 944	0.5 (533)	0.2 (0.005)	3.2
60-64	93 166	1.0 (954)	0.4 (0.004)	2.4	84 365	0.8 (689)	0.3 (0.007)	2.9
65-69	65 419	2.0 (1284)	0.9 (0.012)	2.2	56 355	1.4 (762)	0.6 (0.015)	2.4
70-74	67 229	3.7 (2520)	1.9 (0.023)	2.0	53 706	2.3 (1250)	1.2 (0.026)	2.0
75-79	58 911	6.8 (4025)	3.9 (0.050)	1.7	42 629	3.9 (1670)	2.6 (0.052)	1.5
80-84	49 633	10.6 (5241)	7.5 (0.087)	1.4	31 830	6.4 (2044)	5.3 (0.088)	1.2
85-89	31 296	12.9 (4025)	11.9 (0.122)	1.1	19 241	8.5 (1633)	9.6 (0.133)	0.9
FRAX:								
50-54	103 964	0.4 (389)	0.2 (0.002)	1.8	93 755	0.4 (415)	0.1 (0.001)	3.0
55-59	106 372	0.6 (657)	0.4 (0.003)	1.7	96 944	0.5 (533)	0.3 (0.002)	2.2
60-64	93 166	1.0 (954)	0.7 (0.005)	1.6	84 365	0.8 (689)	0.4 (0.003)	1.9
65-69	65 419	2.0 (1284)	1.2 (0.010)	1.6	56 355	1.4 (762)	0.8 (0.006)	1.7
70-74	67 229	3.7 (2520)	2.4 (0.019)	1.6	53 706	2.3 (1250)	1.5 (0.011)	1.6
75-79	58 911	6.8 (4025)	4.3 (0.030)	1.6	42 629	3.9 (1670)	2.5 (0.017)	1.6
80-84	49 633	10.6 (5241)	6.0 (0.036)	1.8	31 830	6.4 (2044)	3.4 (0.019)	1.9
85-89	31 296	12.9 (4025)	6.8 (0.037)	1.9	19 241	8.5 (1633)	3.8 (0.020)	2.3
Garvan:								
50-54	103 964	0.4 (389)	0.1 (0.001)	6.9	93 755	0.4 (415)	0.1 (0.001)	5.2
55-59	106 372	0.6 (657)	0.1 (0.002)	5.6	96 944	0.5 (533)	0.2 (0.002)	2.9
60-64	93 166	1.0 (954)	0.2 (0.004)	5.2	84 365	0.8 (689)	0.4 (0.004)	2.1
65-69	65 419	2.0 (1284)	0.4 (0.009)	5.6	56 355	1.4 (762)	0.8 (0.008)	1.6
70-74	67 229	3.7 (2520)	0.6 (0.016)	5.8	53 706	2.3 (1250)	1.6 (0.016)	1.4
75-79	58 911	6.8 (4025)	1.3 (0.033)	5.2	42 629	3.9 (1670)	3.5 (0.033)	1.1
80-84	49 633	10.6 (5241)	2.6 (0.057)	4.1	31 830	6.4 (2044)	7.1 (0.065)	0.9
85-89	31 296	12.9 (4025)	4.7 (0.093)	2.7	19 241	8.5 (1633)	13.9 (0.109)	0.6

*Assessed with five years of follow-up.

0.36% net decrease in the proportion of patients assigned a more appropriate risk category for hip fractures and major osteoporotic fractures, respectively. The combination of these results suggests an overall similar discriminatory performance for QFracture and FRAX, with a small advantage in hip fracture prediction for the former and a small advantage in major osteoporotic fracture prediction for the latter.

The tool specific external validation analyses presented comparable results to those reported in previous individual tool validations of the same age ranges. Despite the identical age ranges that were used for the tool specific external validations, the populations still differed to some extent from the derivation cohorts to which they were compared in terms of overall average age, sex distribution, and prevalence of risk factors (see supplement 8). In addition, the FRAX derivation cohort excluded patients treated for osteoporosis, but the current study found very similar results for FRAX when tested in a cohort with and without these patients (see supplement 7). Owing to these differences, our tool specific external validation analyses provided evidence for the transportability of the tools when considering the spectrum of external validation studies ranging from reproducible to transportable. Furthermore, by comparing performance gaps between tools both in the same population and in populations of different age ranges, our analyses substantiated previous claims of a strong correlation between age spans of the studied population and the observed performance of the tested tool.[12,30]

In an analysis of the calibration measures, FRAX presented the best observed-to-predicted ratios, with the weighted average closest to 1, both across age groups and across predicted risk 10ths. Additionally, the calibration slopes of FRAX were closest to 1, representing better calibration across individuals, on top of the better calibration among groups. The FRAX calibration ratios were also relatively stable, whereas QFracture and Garvan presented a decline in the observed-to-predicted ratios as age increased. A possible contributor to FRAX's more consistent observed-to-predicted ratio across age groups is that it accounts for the competing risk of death, whereas Garvan and QFracture do not.[4] The integration of competing death risk into fracture prediction simulates real world behaviour by assigning lower predicted fracture rates for groups of individuals who have lower life expectancy, such as older people. The issue of whether competing risk of death should be incorporated into fracture prediction tools has been debated in the literature, with some studies accounting for it and others not.[20,25,50-53] Our comparative results observed within a single real world population illustrate that calibration is relatively more consistent when competing risk is incorporated. The observed-to-predicted ratios of QFracture and Garvan also presented a declining trend over 10ths of risk. The trend observed over 10ths is at least in part likely explained by age, Table 5 | Calibration* of observed versus predicted hip fracture rates, by sex and 10ths of risk groups

	Women				Men			
Risk 10th	No of people	Hip fracture rate %) (No with first hip fracture†)	Mean (SD) predicted probability (%)	Observed to predicted ratio	No of people	Hip fracture rate (%) (No with first hip fracture†)	Mean (SD) predicted probability (%)	Observed to predicted ratio
QFracture:								
1	57 599	0.2 (109.8)	0.1 (0.0002)	3.3	47 882	0.2 (108.4)	0.0 (0.0001)	5.6
2	57 599	0.4 (207.2)	0.1 (0.0001)	3.5	47 883	0.3 (130.5)	0.1 (0.0001)	4.0
3	57 599	0.4 (237.6)	0.2 (0.0002)	2.7	47 883	0.4 (176.0)	0.1 (0.0001)	3.7
4	57 599	0.6 (364.8)	0.2 (0.0003)	2.7	47 882	0.4 (195.0)	0.1 (0.0001)	2.9
5	57 599	0.9 (495.1)	0.4 (0.0005)	2.4	47 882	0.5 (258.8)	0.2 (0.0002)	2.8
6	57 599	1.5 (860.0)	0.6 (0.0009)	2.5	47 883	0.9 (421.9)	0.3 (0.0003)	3.1
7	57 599	2.5 (1450.9)	1.0 (0.0017)	2.5	47 883	1.3 (629.3)	0.4 (0.0006)	2.9
8	57 599	4.5 (2618.4)	1.8 (0.0032)	2.5	47 882	2.1 (1022.8)	0.8 (0.0014)	2.7
9	57 599	8.2 (4751.2)	3.6 (0.0077)	2.3	47 882	4.0 (1937.7)	1.6 (0.0040)	2.5
10	57 599	13.9 (8000.2)	13.4 (0.1132)	1.0	47 883	8.6 (4115.8)	9.2 (0.1166)	0.9
FRAX:								
1	57 599	0.3 (150.9)	0.1 (0.0003)	2.7	47 882	0.3 (132.6)	0.1 (0.0002)	4.1
2	57 599	0.4 (239.0)	0.2 (0.0002)	2.3	47 883	0.4 (168.0)	0.1 (0.0002)	2.8
3	57 599	0.5 (306.0)	0.3 (0.0004)	2.0	47 883	0.4 (183.0)	0.2 (0.0002)	2.1
4	57 599	0.6 (358.6)	0.4 (0.0003)	1.6	47 882	0.5 (243.5)	0.2 (0.0002)	2.1
5	57 599	1.1 (657.1)	0.6 (0.0010)	1.8	47 882	0.7 (330.4)	0.3 (0.0005)	2.0
6	57 599	1.7 (956.2)	0.9 (0.0017)	1.8	47 883	0.9 (444.6)	0.5 (0.0006)	1.9
7	57 599	2.6 (1502.0)	1.5 (0.0024)	1.7	47 883	1.5 (698.5)	0.8 (0.0009)	1.8
8	57 599	4.3 (2494.5)	2.6 (0.0032)	1.7	47 882	2.3 (1120.3)	1.3 (0.0024)	1.8
9	57 599	7.9 (4533.5)	4.2 (0.0053)	1.9	47 882	3.8 (1838.0)	2.2 (0.0036)	1.8
10	57 599	13.7 (7897.4)	8.8 (0.0354)	1.6	47 883	8.0 (3837.3)	4.4 (0.0202)	1.8
Garvan:								
1	57 599	0.4 (239.4)	0.0 (0.0001)	21.2	47 882	0.3 (145.8)	0.1 (0.0001)	4.7
2	57 599	0.5 (262.8)	0.0 (0.0001)	11.3	47 883	0.4 (170.7)	0.1 (0.0001)	3.8
3	57 599	0.6 (329.6)	0.1 (0.0001)	9.5	47 883	0.4 (206.2)	0.1 (0.0001)	3.1
4	57 599	0.8 (469.7)	0.1 (0.0001)	9.8	47 882	0.5 (251.0)	0.2 (0.0002)	2.6
5	57 599	1.0 (591.4)	0.1 (0.0001)	8.8	47 882	0.7 (320.6)	0.3 (0.0003)	2.3
6	57 599	1.6 (943.5)	0.2 (0.0002)	9.6	47 883	0.9 (408.4)	0.5 (0.0007)	1.9
7	57 599	2.4 (1404.4)	0.3 (0.0004)	9.3	47 883	1.4 (673.9)	0.8 (0.0014)	1.8
8	57 599	4.2 (2427.9)	0.4 (0.0007)	9.6	47 882	2.2 (1039.3)	1.5 (0.0027)	1.5
9	57 599	7.2 (4173.0)	0.9 (0.0020)	8.4	47 882	3.7 (1762.5)	3.1 (0.0073)	1.2
10	57 599	14.3 (8253.6)	5.9 (0.0868)	2.4	47 883	8.4 (4017.8)	10.8 (0.0921)	0.8

*Assessed with five years of follow-up.
†Case numbers contain a decimal component because they are averaged between imputed datasets.

since higher risk 10ths contained a larger share of older people (data not shown). The overall better calibration of FRAX may also be due to the use of country specific probability charts provided by FRAX.

Strengths and limitations of this study

This study has several strengths in its methods, analyses, and findings and implications for practical real world application. Firstly, we directly compared three well established fracture prediction tools in the same population, thus measuring the differences in performance with minimal effect of confounding. Secondly, the population used for this study was large, had many fracture events (>100 000 major osteoporotic and hip fracture events), included both men and women, and was nationally representative, thereby minimising selection biases. In addition, the population included those who died before the end of follow-up, which simulates real life use of the tools. To our knowledge, no previous study outside of the UK and the Republic of Ireland has validated QFracture in an independent population. Additionally, the tool specific validation analyses allowed the presentation of comparable results to previous reported performance of the specific tools, thus assuring the use of relevant population and input data and further strengthening the results of the comparative analysis.

By testing the applicability and performance of the three predictive tools using data from electronic health records, this study confirmed that the tools are transferrable to an electronic health record system with the potential for automated large scale implementation, even though two of them were not originally developed in this setting. Any organisation that aims to implement these tools into an electronic health record system must make adaptations according to the available data in its database. Despite having to adapt the data according to our electronic health record system for some of the variable categorisations used by the different tools, we observed comparable performance to those found in previous studies across all tools. This shows that the application of these tools in an external electronic health record system can be replicated in other contexts. As further evidence that these tools are applicable

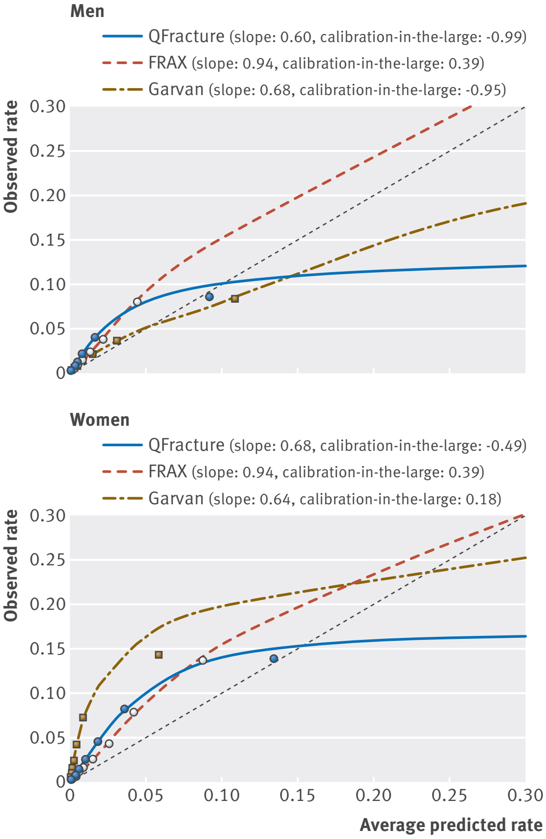

Fig 3 | Calibration plot of hip fracture predictions for QFracture, FRAX, and Garvan in the comparative analysis: five year observed risk and average predicted probabilities by probability 10ths (each mark represents approximately a 10th of the cohort belonging to the same probability 10th)

to an environment of electronic health records, we observed that the fracture rate in strata based on 30 out of 33 total variables considered among the three tools was higher among the strata with known risk factors (see table 1). There were, however, three variables that yielded patterns contrary to the expected direction for fracture risk. Two of these (family history of osteoporosis and hip fractures) were reflections of data limitations, specifically lower rates of well defined family connections for older adults. Smoking status, the third variable that did not follow the anticipated direction, was possibly affected by confounding, since younger men were more likely to be current smokers in the study population (data not shown).

This study has several limitations. While previous studies commonly report probabilities for 10 years of follow-up, we were only able to evaluate the probabilities of fracture risk for five years owing to limited availability of robust baseline data as of 1 January 2005. It has been noted that AUC performance can be potentially affected by the duration of follow-up.[12 30] To address this point, we conducted a preliminary analysis and found that the rate of fracture events is approximately constant, meaning that the cumulative rate of events is linear, as presented in our supplementary material as well as in previous studies.[23] This trend not only substantiates the conversion of 10 year FRAX probabilities into five year

Table 6 | Tool specific external validation of original tools: discriminatory measures* of QFracture, FRAX, and Garvan when evaluated in populations of the same age ranges as the original tools. Values in parentheses are nominators/denominators

	Hip fractures			Major osteoporotic fractures		
Tools	Overall	Men	Women	Overall	Men	Women
QFracture—30-100 years:						
AUC (%)	88.0	85.6	88.6	75.4	68.6	77.4
Sensitivity† (%)	61.4 (19 471.1/31 709)	59.9 (6294.8/10 517)	58.3 (12 362.9/21 192)	35.7 (35 379.2/99 058)	31.4 (10 129.5/32 209)	32.7 (21 849.5/66 849)
Specificity† (%)	90.9 (1 694 534.1/1 864 704)	90.6 (794 409.8/876 852)	91.0 (899310.8/987852)	91.4 (1 643 093.2/1 797 355)	90.8 (776 552.4/855 160)	91.6 (863 140.4/942 195)
FRAX—50-90 years:						
AUC (%)	81.5	79.6	81.5	71.4	68.4	69.8
Sensitivity† (%)	43.6 (12 257.4/28 091)	42.7 (3836.8/8996)	41.3 (7894.3/19 095)	29.0 (23 628.9/81 564)	28.0 (6521.5/23 268)	25.3 (14 770.3/58 296)
Specificity† (%)	90.9 (933 500.3/1 026 724)	90.6 (425 782.8/469 829)	91.1 (507 190.2/556 895)	91.6 (891 398.9/973 251)	90.9 (414 195.5/455 557)	91.7 (474 865.3/517 694)
Garvan—60-95 years‡:						
AUC (%)	71.2	76.5	75.7			
Sensitivity† (%)	28.7 (8013.7/27 897)	35.6 (3050.1/8571)	33.1 (6392.3/19 326)			
Specificity† (%)	90.8 (583 508.7/642 538)	90.8 (259 335.0/285 713)	91.2 (325 602.3/356 825)			

AUC=area under receiver operating characteristic curve (C statistic).
*Assessed with five years of follow-up and calculated for cut-off of top 10% risk for each tool.
†Numerator case numbers contain a decimal component because they are averaged between imputed datasets.
‡Analyses comparing performance for predicting major osteoporotic fractures were only conducted between QFracture and FRAX because Garvan's definition for major osteoporotic fractures is much broader than those for QFracture and FRAX.

probabilities, but also supports an assumption that the performance of the five year probabilities likely reflects the performance of the 10 year probabilities. Furthermore, from a clinical perspective, the five year probabilities might be more useful for prevention and intervention, since the long term safety of bisphosphonates for fracture prevention after five years is unclear.[54 55] Secondly, the evaluation of FRAX relied on probability charts, which provide a cruder risk assessment than complete tool equations. Since FRAX achieved good discrimination, which aligns well with previous studies, it is reasonable to conclude that this limitation did not substantially affect its performance. Finally, as is often the case in electronic health record databases, we did not have extensive data on bone mineral density (in a large enough proportion of the overall population throughout the study timeline) to include it as an input variable for the tools that do offer a bone mineral density option. Given that bone mineral density screening is not performed in a substantial part of the adult population,[7-10] and the fact that QFracture does not include bone mineral density as an input, we have reason to believe that despite this omission, our results are meaningful for practical application.

Comparison to existing research

The AUC results for prediction of hip fracture in the tool specific external validation cohort were comparable to those reported in the development and validation studies for the second version of QFracture, which were 89% for women and 87-88% for men.[16 56] Since the original publications on FRAX and Garvan did not include AUC results for hip fracture prediction without bone mineral density,[4 14] we identified other validation studies of these tools that reported this measure and were performed in populations comprising the same age ranges as the original development studies. AUC results for hip fracture prediction in such FRAX validations ranged from 77% to 79% for both sexes,[57 58] and 83% for women alone,[52] comparable to the 82% that we report for both sexes and for women alone. The AUCs for Garvan from validations that were found relevant for this comparison were 76% in both sexes[25] or 70% and 69% in women and men, respectively,[39] compared with our study's results of 71% in both sexes or 76-77% in women and men separately.

Comparisons of the three tools from reviews and meta-analyses have often concluded that the discrimination performance of QFracture is substantially better than that of FRAX. The conclusion that QFracture has substantially higher performance than other tools has also been cited in practice guidelines, such as the Scottish Intercollegiate Guidelines Network, as justification for using QFracture.[59] However, these conclusions and practice recommendations are based on studies that did not use consistent inclusion criteria, such as sex and age ranges. While we also found QFracture to have better discrimination than either of the other two tools for hip fracture prediction, it was to a lesser extent than previously reported. Additionally, the QFracture discrimination for major osteoporotic fractures was slightly lower than that of FRAX. The possibility that differences in age ranges can substantially affect the observed discriminatory performance of a tool has been previously suggested.[30] Yet this has not been shown within a real world population based study by comparing the performance of one tool in different age ranges. Our tool specific external validation results show that AUC and sensitivity values tended to be higher in populations with wider age ranges than in populations with narrower age ranges (fig 2). For example, the overall AUC of QFracture for hip fracture prediction, which spanned over a 70 year age range, was higher than that of FRAX and Garvan, which spanned over age ranges of 40 and 35 years, respectively. Additionally, by comparing the performance of a specific tool between the comparative and validation analyses, we illustrated how an expansion in age ranges resulted in higher observed discrimination (QFracture, with AUCs of 82.7% *v* 88.0%, respectively), and how a narrowing of the age range results in lower observed discrimination (Garvan, with AUCs of 77.8% *v*71.2%, respectively). In examining the receiver operating characteristic curves for each tool among a population with a narrower age range and among a population with a broader age range, our findings illustrate the direct effect that age has on the performance of these tools (fig 2).

Conclusions and practice implications

Current guidelines for the prevention of osteoporotic fracture incorporate fracture prediction tools; the UK's National Osteoporosis Guideline Group (NOGG)[60] supports the use of age dependent risk cut-offs as an indication for treatment, regardless of bone mineral density, and a lower set of cut-offs as an indication for performing bone mineral density scanning. Another approach, adopted by the National Osteoporosis Foundation (NOF) in the US,[55] recommends using a constant cut-off as indication for treatment in osteopenic patients (with borderline bone mineral density). However, even with clear recommendations from the professional committees, uptake in the adoption of these tools in practice is not widespread.[17 61] This is in part due to the lack of an automated decision support infrastructure that allows clinicians easy access to patients' risk for fractures. The current study showed that two of the three tools assessed offer good discrimination for hip fracture prediction using electronic health record data, which could be incorporated into the electronic health record system and automatically raise an alert for clinicians when a patient is indicated to be at high risk.

Our results were consistent with previous findings that the best discrimination was associated with prediction of hip fractures,[12 20] which are known to be associated with the greatest morbidity and mortality.[62] Because the calibration can be corrected for a local population,[63] the selection of a tool should primarily be based on its discriminative ability. In an electronic health record system where all input variables are available for automatic implementation, our study suggests that QFracture yields the best discriminatory performance for hip fracture prediction. If some of the input variables are not available, FRAX, which performed almost as well, despite not being developed using data from electronic health records, could be a simpler option for implementation of decision support. However, to identify those at risk for all major osteoporotic fractures, FRAX yielded slightly better discrimination, and thus may be preferable. The selected tool should only be used for the age ranges for which it was developed. Since all tools underestimated the risk of hip fractures in our population, the selected tool will require local calibration when implemented in practice, as the guidelines are based on specific risk cut-offs. This calibration will be more straightforward for FRAX, which showed steadier calibration but will require age dependent calibration of the other tools.

To achieve the potential utility of fracture prediction tools in clinical practice and adhere to the fracture prevention guidelines, these tools likely need to be automatically incorporated into electronic health record systems and brought to the attention of primary care doctors only when an action is required, without imposing any additional time burden. This study has shown that automatic implementation of the tools into an external electronic health record system is feasible, and has provided recommendations as to which tool is preferred under different circumstances. Additionally, our findings emphasise the importance of carefully comparing the performance of prediction tools of any kind in similar populations, and, if possible, in the same population. Further research is warranted to evaluate whether automatically generated fracture risk scores made accessible directly to patients or their doctors would increase screening for and treatment of osteoporosis, and ultimately prevent osteoporotic fractures.

We thank our colleagues at the Clalit Research Institute: Sydney Krispin and Carly Davis for their assistance in editing and reviewing the manuscript and Amichay Akriv and Moshe Hoshen for their guidance on the statistical analyses.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Funding: This study was supported internally by the Clalit Research Institute for the design and conduct of the study, data collection, analysis and interpretation of the data, and preparation and review of the manuscript.

Contributors: ND, RDB, and ML-R conceived and designed the study. ND, CC-S, and ML-R analysed and interpreted the data. ND and CC-S drafted the manuscript. All authors critically revised the manuscript for important intellectual content. ND carried out the statistical analysis. CC-S provided administrative, technical, and material support. RDB supervised the study and is the guarantor.

Data sharing: No additional data available.

Ethical approval: This study was approved by the Clalit Health Services research ethics committee.

Transparency: The lead author (ND) confirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 Pisani P, Renna MD, Conversano F, et al. Major osteoporotic fragility fractures: Risk factor updates and societal impact. *World J Orthop* 2016;7:171-81. doi:10.5312/wjo.v7.i3.171.

2 Roux C, Wyman A, Hooven FH, et al. GLOW investigators. Burden of non-hip, non-vertebral fractures on quality of life in postmenopausal women: the Global Longitudinal study of Osteoporosis in Women (GLOW). *Osteoporos Int* 2012;23:2863-71. doi:10.1007/s00198-012-1935-8.

3 Coughlan T, Dockery F. Osteoporosis and fracture risk in older people. *Clin Med (Lond)* 2014;14:187-91. doi:10.7861/clinmedicine.14-2-187.

4 Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 2008;19:385-97. doi:10.1007/s00198-007-0543-5.

5 Lih A, Nandapalan H, Kim M, et al. Targeted intervention reduces refracture rates in patients with incident non-vertebral osteoporotic fractures: a 4-year prospective controlled study. *Osteoporos Int* 2011;22:849-58. doi:10.1007/s00198-010-1477-x.

6 Ruggiero C, Zampi E, Rinonapoli G, et al. Fracture prevention service to bridge the osteoporosis care gap. *Clin Interv Aging* 2015;10:1035-42.

7 Lafata JE, Kolk D, Peterson EL, et al. Improving osteoporosis screening: results from a randomized cluster trial. *J Gen Intern Med* 2007;22:346-51. doi:10.1007/s11606-006-0060-9.

8 Cohen K, Maier D. Osteoporosis: evaluation of screening patterns in a primary-care group practice. *J Clin Densitom* 2008;11:498-502. doi:10.1016/j.jocd.2008.08.104.

9 McNally DN, Kenny AM, Smith JA. Adherence of academic geriatric practitioners to osteoporosis screening guidelines. *Osteoporos Int* 2007;18:177-83. doi:10.1007/s00198-006-0215-x.

10 Powell H, O'Connor K, Greenberg D. Adherence to the U.S. Preventive Services Task Force 2002 osteoporosis screening guidelines in academic primary care settings. *J Womens Health (Larchmt)* 2012;21:50-3. doi:10.1089/jwh.2010.2560.

11 Aspray TJ. Fragility fracture: recent developments in risk assessment. *Ther Adv Musculoskelet Dis* 2015;7:17-25. doi:10.1177/1759720X14564562.

12 Marques A, Ferreira RJ, Santos E, Loza E, Carmona L, da Silva JA. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis[published Online First: Epub Date]. *Ann Rheum Dis* 2015;74:1958-67. doi:10.1136/annrheumdis-2015-207907.

13 Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 2009;339:b4229. doi:10.1136/bmj.b4229.

14 Nguyen ND, Frost SA, Center JR, Eisman JA, Nguyen TV. Development of a nomogram for individualizing hip fracture risk in men and women. *Osteoporos Int* 2007;18:1109-17. doi:10.1007/s00198-007-0362-8.

15 Nguyen ND, Pongchaiyakul C, Center JR, Eisman JA, Nguyen TV. Identification of high-risk individuals for hip fracture: a 14-year prospective study. *J Bone Miner Res* 2005;20:1921-8. doi:10.1359/JBMR.050520.

16 Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* 2012;344:e3427. doi:10.1136/bmj.e3427.

17 Collins GS, Michaëlsson K. Fracture risk assessment: state of the art, methodologically unsound, or poorly reported?*Curr Osteoporos Rep* 2012;10:199-207. doi:10.1007/s11914-012-0108-1.

18 Feldstein A, Elmer PJ, Smith DH, et al. Electronic medical record reminder improves osteoporosis management after a fracture: a randomized, controlled trial. *J Am Geriatr Soc* 2006;54:450-7. doi:10.1111/j.1532-5415.2005.00618.x.

19 DeJesus RS, Angstman KB, Kesman R, et al. Use of a clinical decision support system to increase osteoporosis screening. *J Eval Clin Pract* 2012;18:89-92. doi:10.1111/j.1365-2753.2010.01528.x.

20 Leslie WD, Lix LM. Comparison between various fracture risk assessment tools. *Osteoporos Int* 2014;25:1-21. doi:10.1007/s00198-013-2409-3.

21 Nayak S, Edwards DL, Saleh AA, Greenspan SL. Performance of risk assessment instruments for predicting osteoporotic fracture risk: a systematic review. *Osteoporos Int* 2014;25:23-49. doi:10.1007/s00198-013-2504-5.

22 Rubin KH, Friis-Holmberg T, Hermann AP, Abrahamsen B, Brixen K. Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review. *J Bone Miner Res* 2013;28:1701-17. doi:10.1002/jbmr.1956.

23 Bolland MJ, Siu AT, Mason BH, et al. Evaluation of the FRAX and Garvan fracture risk calculators in older women. *J Bone Miner Res* 2011;26:420-7. doi:10.1002/jbmr.215.

24 Henry MJ, Pasco JA, Merriman EN, et al. Fracture risk score and absolute risk of fracture. *Radiology* 2011;259:495-501. doi:10.1148/radiol.10101406.

25 Sambrook PN, Flahive J, Hooven FH, et al. Predicting fractures in an international cohort using risk factor algorithms without BMD. *J Bone Miner Res* 2011;26:2770-7. doi:10.1002/jbmr.503.

26 Cummins NM, Poku EK, Towler MR, O'Driscoll OM, Ralston SH. clinical risk factors for osteoporosis in Ireland and the UK: a comparison of FRAX and QFractureScores. *Calcif Tissue Int* 2011;89:172-7. doi:10.1007/s00223-011-9504-2.

27 Dobson R, Leddy SG, Gangadharan S, Giovannoni G. Assessing fracture risk in people with MS: a service development study comparing three fracture risk scoring systems. *BMJ Open* 2013;3:e002508. doi:10.1136/bmjopen-2012-002508.

28 Johansen A. QFracture is better than FRAX tool in assessing risk of hip fracture. *BMJ* 2012;345:e4988. doi:10.1136/bmj.e4988.

29 Sandhu SK, Nguyen ND, Center JR, Pocock NA, Eisman JA, Nguyen TV. Prognosis of fracture: evaluation of predictive accuracy of the FRAX algorithm and Garvan nomogram. *Osteoporos Int* 2010;21:863-71. doi:10.1007/s00198-009-1026-7.

30 Kanis JA, Oden A, Johansson H, McCloskey E. Pitfalls in the external validation of FRAX. *Osteoporos Int* 2012;23:423-31. doi:10.1007/s00198-011-1846-0.

31 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24. doi:10.7326/0003-4819-130-6-199903160-00016.

32 Rosen B, Waitzberg R, Merkur S. Israel: Health System Review. *Health Syst Transit* 2015;17:1-212.

33 Gross R, Rosen B, Chinitz D. Evaluating the Israeli health care reform: strategy, challenges and lessons. *Health Policy* 1998;45:99-117. doi:10.1016/S0168-8510(98)00030-X.

34 Kanis JA, Hans D, Cooper C, et al. Task Force of the FRAX Initiative. Interpretation and use of FRAX in clinical practice. *Osteoporos Int* 2011;22:2395-411. doi:10.1007/s00198-011-1713-z.

35 Garvan Institute. Fracture Risk Calculator. 2013. www.garvan.org.au/promotions/bone-fracture-risk/calculator/. Accessed on 8 May 2016.

36 ClinRisk Ltd. QFracture®-2013 risk calculator. 2015. www.qfracture.org/. Accessed on 8 May 2016.

37 World Health Organization Collaborating Center for Metabolic Bone Diseases. FRAX® WHO Fracture Risk Assessment Tool. 2011. www.shef.ac.uk/FRAX/tool.jsp. Accessed on 30 September 2015.

38 Kanis JA, Oden A, Johnell O, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 2007;18:1033-46. doi:10.1007/s00198-007-0343-y.

39 Ahmed LA, Nguyen ND, Bjørnerem Å, et al. External validation of the Garvan nomograms for predicting absolute fracture risk: the Tromsø study. *PLoS One* 2014;9:e107695. doi:10.1371/journal.pone.0107695.

40 Hadorn DC, Draper D, Rogers WH, Keeler EB, Brook RH. Cross-validation performance of mortality prediction models. *Stat Med* 1992;11:475-89. doi:10.1002/sim.4780110409.

41 Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol* 2003;3:21. doi:10.1186/1471-2288-3-21.

42 Pencina MJ, D'Agostino RB Sr, , D'Agostino RB Jr, , Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157-72, discussion 207-12. doi:10.1002/sim.2929.

43 Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology* 2014;25:114-21. doi:10.1097/EDE.0000000000000018.

44 Pencina MJ, D'Agostino RB Sr, , Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;30:11-21. doi:10.1002/sim.4085.

45 Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.*1st ed. Springer, 2009doi:10.1007/978-0-387-77244-8.

46 Harrell FE Jr (2015) rms: Regression Modeling Strategies. R package version 4.4-0.

47 Van Buuren S, Groothuis-Oudshoorn K (2014) mice: Multivariate Imputation by Chained Equations. R package version 2.22.

48 Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 2009;9:57. doi:10.1186/1471-2288-9-57.

49 Sing T SO, Beerenwinkel N, and Lengauer T. (2015) ROCR: Visualizing the Performance of Scoring Classifiers. R package version 1.0-7.

50 Byberg L, Gedeborg R, Cars T, et al. Prediction of fracture risk in men: a cohort study. *J Bone Miner Res* 2012;27:797-807. doi:10.1002/jbmr.1498.

51 Leslie WD, Lix LM, Wu X. Manitoba Bone Density Program. Competing mortality and fracture risk assessment. *Osteoporos Int* 2013;24:681-8. doi:10.1007/s00198-012-2051-5.

52 Pressman AR, Lo JC, Chandra M, Ettinger B. Methods for assessing fracture risk prediction models: experience with FRAX in a large integrated health care delivery system. *J Clin Densitom* 2011;14:407-15. doi:10.1016/j.jocd.2011.06.006.

53 Cooper C, Harvey NC. Osteoporosis risk assessment. *BMJ* 2012;344:e4191. doi:10.1136/bmj.e4191.

54 Black DM, Bauer DC, Schwartz AV, Cummings SR, Rosen CJ. Continuing bisphosphonate treatment for osteoporosis--for whom and for how long?*N Engl J Med* 2012;366:2051-3. doi:10.1056/NEJMp1202623.

55 Cosman F, de Beur SJ, LeBoff MS, et al. National Osteoporosis Foundation. Clinician's Guide to Prevention and Treatment of Osteoporosis. *Osteoporos Int* 2014;25:2359-81. doi:10.1007/s00198-014-2794-2.

56 Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open* 2014;4:e005809. doi:10.1136/bmjopen-2014-005809.

57 Fraser LA, Langsetmo L, Berger C, et al. CaMos Research Group. Fracture prediction and calibration of a Canadian FRAX® tool: a population-based report from CaMos. *Osteoporos Int* 2011;22:829-37. doi:10.1007/s00198-010-1465-1.

58 Leslie WD, Lix LM, Johansson H, Oden A, McCloskey E, Kanis JA. Manitoba Bone Density Program. Independent clinical validation of a Canadian FRAX tool: fracture prediction and model calibration. *J Bone Miner Res* 2010;25:2350-8. doi:10.1002/jbmr.123.

59 Scottish Intercollegiate Guidelines Network (SIGN). Management of osteoporosis and the prevention of fragility fractures: A national clinical guideline. Edinburgh: SIGN; March 2015. SIGN publication No 142.

60 Compston J, Bowring C, Cooper A, et al. National Osteoporosis Guideline Group. Diagnosis and management of osteoporosis in postmenopausal women and older men in the UK: National Osteoporosis Guideline Group (NOGG) update 2013. *Maturitas* 2013;75:392-6. doi:10.1016/j.maturitas.2013.05.013.

61 Bruyère O, Nicolet D, Compère S, et al. Perception, knowledge, and use by general practitioners of Belgium of a new WHO tool (FRAX) to assess the 10-year probability of fracture. *Rheumatol Int* 2013;33:979-83. doi:10.1007/s00296-012-2461-x.

62 Cummings SR, Melton LJ. Epidemiology and outcomes of osteoporotic fractures. *Lancet* 2002;359:1761-7. doi:10.1016/S0140-6736(02)08657-9.

63 Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605. doi:10.1136/bmj.b605.

Supplementary information: Supplementary material includes detailed supporting methodological information, such as diagnoses codes used, and additional analyses conducted (preliminary and sensitivity analyses). It also includes comparative background information between the three tools and comparisons between the current study and previous studies' populations. Lastly, additional calibration analyses are provided