Effect of competing mortality risks on predictive performance of the QFracture risk prediction tool for major osteoporotic fracture and hip fracture

Impact of competing mortality risks on predictive performance of the QFracture risk prediction tool for major osteoporotic fracture and hip fracture: external validation cohort study in a UK primary care population

Authors

Shona Livingstone MSc

Daniel R Morales PhD

Megan McMinn PhD

Chima Eke PhD

Prof Peter T Donnan PhD

Prof Bruce Guthrie PhD

Author affiliations

S Livingstone, Dr Morales and PT Donnan: Division of Population Health and Genomics, University of Dundee

Megan McMinn: Centre for Population Health Sciences, Usher Institute, University of Edinburgh

B Guthrie and C Eke: Advanced Care Research Centre,

Corresponding author:
Bruce Guthrie, Doorway 3, Old Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG
bruce.guthrie@ed.ac.uk

Keywords: Fracture risk; primary prevention; risk-prediction; external validation; QFracture; competing mortality risk

Word count: 4284

Abstract

Objectives: To externally evaluate QFracture for predicting major osteoporotic fracture (MOF) and hip fracture.

Design, setting and participants: Linked primary care, hospital admission and mortality data from the Clinical Research Practice Datalink GOLD. People aged 30-99 years with up-to-standard linked data for at least one year were eligible. MOF was defined as any hip, distal forearm, proximal humerus or vertebral crush fracture, ascertained from GP, hospital discharge and mortality data. QFracture 10-year predicted MOF and hip fracture risk was calculated, and performance evaluated versus observed 10-year fracture risk in the whole population, and in subgroups of age and comorbidity. QFracture calibration was examined with and without accounting for competing non-fracture mortality risk.

Results: There were 2,747,409 women with 95,598 MOF and 36,400 hip fractures, and 2,684,730 men with 34,321 MOF and 13,379 hip fractures. Incidence of all fractures was higher than in QFracture internal derivation. Competing mortality risk was more common than fracture from middle-age onwards. QFracture discrimination in the whole population was excellent or good for MOF and hip fracture (Harrell's-C in women 0.813 and 0.918 respectively; in men 0.738 and 0.888), but was poor to moderate in age subgroups (e.g. Harrell's-C in women and men aged 85-99 respectively 0.576 and 0.624 for MOF, and 0.624 and 0.637 for hip). Without accounting for competing risks, QFracture systematically under-predicted fracture risk in all models, more so for MOF than hip fracture, and more so in older people. Accounting for competing risks, QFracture still under-predicted in the whole population but showed considerable over-prediction in older age-groups and people with high comorbidity at high fracture risk.

Conclusion: QFracture systematically under-predicts fracture risk (because of fracture under-ascertainment) and over-predicts in older and comorbid people (because of competing mortality). The use of QFracture in its current form needs reviewing, particularly in people at high risk of death from other causes.

What is already known on this topic

- QFracture is recommended by the National Institute of Health and Care Effectiveness to predict risk of fracture and to guide decisions to start bisphosphonates, on the basis of previous validation studies showing good predictive performance.
- Previous validation studies have followed the QFracture derivation study in not including fractures recorded in hospital discharge data, and in not accounting for competing mortality risk.

What this study adds

- The observed incidence of fracture was higher in this study (which included hospital recorded fractures) than in QFracture derivation and validation studies (which did not).
- Despite excellent discrimination in the whole population, there was systematic under-prediction of fracture risk by QFracture, and systematic over-prediction in older and more comorbid people once competing mortality risk was accounted for.
- Calibration is sufficiently poor that the use of QFracture for clinical prediction needs reviewing, particularly in people at high risk of death from other causes.

Introduction

Fragility or low-impact fractures are a common consequence of osteoporosis and osteopenia, and a major cause of morbidity, disability and in some cases death. Bisphosphonates reduce hip and vertebral fracture risk in people with osteoporosis,[1] and guidelines internationally recommend pharmacological treatment for people at high risk of fracture.[1-4] In the UK, guidelines recommend using a fracture risk prediction tool in middle-age and older who have risk factors for fracture, with bone mineral density (BMD) measurement reserved for further risk stratification in those at intermediate risk.[2,4] In the US, guidelines from the US Bone Health and Osteoporosis Foundation (previously known as National Osteoporosis Foundation) recommend similar use of prediction tools for middle-aged people but additionally recommend routine use of BMD measurement in older people.[5] Risk-stratified guideline recommendations like this are increasingly used by guideline-developers to target treatment to those with the greatest capacity to benefit, but the effectiveness of this strategy critically depends on the performance of the risk-prediction tools used.

A number of fracture risk prediction tools have been created, although only two have been subject to repeated external validation (QFracture and Garvan).[6,7] The first version of QFracture[8] was externally validated in a different UK primary care dataset and found to have excellent discrimination and calibration (discrimination is the ability of the prediction tool to correctly differentiate between people who experience a fracture and those who do not, whereas calibration refers to how closely the predicted and observed probabilities agree).[9] Subsequently, Dagan et al externally validated the updated QFracture algorithm and the Garvan prediction tool in an Israeli dataset. QFracture had very good discrimination but discrimination was only moderate for Garvan, and both tools systematically under-predicted fracture risk.[7] FRAX has been internally validated in several datasets, with FRAX discrimination reported as good but calibration rarely assessed.[6,10] However, FRAX cannot be externally validated because the underlying FRAX algorithm has never been made public which prevents full independent evaluation.[7] Dagan et al also presented an external validation of FRAX in their analysis, but FRAX predictions were not based on full FRAX estimates of risk because the prediction equation is not published.[7] Based on the approximate FRAX risk used, they also found significant under-prediction of fractures for this tool.

In the UK, NICE recommends the use of either QFracture or FRAX to inform decisions to initiate bisphosphonate treatment, but recognise that estimated fracture risk for individuals can vary considerably between tools.[1,2] FRAX has been shown to over-predict fracture risk when the same method of fracture ascertainment as QFracture derivation was used.[2,8,11] Two possible reasons for these differences are: (1) how fractures are identified in the derivation of each tool, with QFracture using codes in primary care records and mortality data[12] and FRAX using self-report and hospital

records[13] (these may be incomplete in different ways); (2) FRAX accounts for competing mortality risks but QFracture does not. Competing risk of (non-fracture) mortality is a known problem in risk prediction, which arises because standard modelling methods assume that patients who are censored before the intended end of follow-up have the same risk of fracture as those who are not censored. While this assumption may be reasonable for loss to follow-up due to changing address, it is clearly false when someone dies. Not accounting for competing mortality risk causes over-prediction of risk of fractures, which is likely to be more of a problem in older people and those with multimorbidity.[14-16]

The aim of this study was therefore to externally validate QFracture, and specifically to compare prediction in relation to better ascertained fracture rates, and to examine the effect of competing risk on predictive performance.

Methods

Data source and population. Linked GP (Clinical Practice Research Datalink [CPRD] Gold), mortality registration (Office of National Statistics [ONS]), and hospital inpatient (Hospital Episode Statistics [HES]) data were used. The data are similar to the QFracture derivation dataset in terms of the inclusion of linked primary care and mortality data, but we also included linked hospital admission data for fracture ascertainment. To be included, patients had to be: permanently registered with a general practice contributing up-to-standard data for at least one year; have linkage to HES discharge and ONS mortality data; and be aged ≥30 years and <100 years. Cohort entry was the latest of the dates on or after 01/01/04. Cohort exit was the date of the earliest of: first relevant fracture event; death; deregistration from the general practice; date of the last data collection from the practice; or the end of the study on 31/3/16. All outcomes and predictors are recorded blind to the study hypothesis as recorded as part of routine clinical care. No formal power calculation was done, since the study size is determined by the data available in CPRD which was considered sufficient.[17]

Outcomes. Two outcomes were modelled as per QFracture – major osteoporotic fracture (MOF) and hip fracture.[12] Major osteoporotic fracture was defined as hip, vertebral, wrist or proximal humeral fractures ascertained from codes in the GP electronic health record (using Read codes, which has been shown to have high positive predictive value for hip fracture[18]), HES discharge diagnoses (ICD-10 codes recorded in the primary position ie the reason for admission), or ONS death registration (ICD-10 codes). QFracture does not publish codes used to define these outcomes, so we derived our own which are described in supplementary tables S1 and S2. MOF recorded before study entry was

used as a predictor variable. MOF or hip fracture recorded after the index date were used as the outcome variable, with the date of the event taken as the first record of fracture.

Prediction model. We implemented the published QFracture®-2016 risk model (under GNU Lesser General Public Licence v3) and calculated QFracture predicted 10-year risk of a major osteoporotic fracture and the risk of a hip fracture for all patients in our cohort. As with fracture outcomes, we derived codesets for each predictor which are described in supplementary tables S3 to S5. The key difference from QFracture derivation was that QFracture allowed body mass index (BMI), alcohol and smoking status recorded *after* the date of study entry but *before* any fracture outcome to be used in prediction, whereas in this analysis we restricted predictor values to those recorded before study entry only to avoid using future information in prediction.

Comorbidity. For each patient at baseline, we calculated the Charlson Comorbidity Index (CCI) based on primary care Read codes.[19] CCI was not used in prediction, but was used to stratify the analysis of discrimination and calibration by level of comorbidity (CCI score grouped into 0, 1, 2, and 3+).

Missing data. The extent and management of missing data is detailed in supplementary table S6. As with QFracture derivation, those with missing ethnicity were assumed to be white. For missing BMI, smoking status, and alcohol status, Multivariate Imputation by Chained Equations[20] was used to generate five imputed datasets which were combined using Rubin's rules[21]. Morbidities and prescribing used for prediction were assumed to absent if not recorded, the same as QFracture derivation, reflecting that morbidity and prescribing recording in CPRD is generally good.[22,23]

Statistical methods. As recommended by reporting guidance,[24] initial analysis compared the study population and fracture rates in this study with previously published QFracture derivation and validation cohorts (although variable reporting across previously published papers means that the comparison population varies depending on the data available).[8,9,12] The performance of the QFracture®-2016 risk score was assessed by examining discrimination and calibration. We used Harrell's C-statistic, truncated to only include pairs where the earliest survival time is no later than 10 years after entry (a C-statistic of 0.5 indicates discrimination that is no better than chance, whereas a C-statistic of 1 indicates perfect discrimination). Two additional measures of discrimination were calculated, the D statistic of Royston and Sauerbrei (which is based on the separation in event-free survival between patients with predicted risk scores above and below the median; higher values indicate greater discrimination),[25] and a related R-squared statistic estimating explained variation for censored survival data.[26]

Calibration was assessed for ten equally-sized groups (deciles) of participants ranked by predicted risk, by plotting observed proportions versus predicted probabilities. We estimated observed risk for censored data in two ways: (1) using the standard Kaplan-Meier estimator (which is consistent with the assumptions made in QFracture derivation in that it does not account for competing risks); and (2) the Aalen Johansen estimator (an extension to allow for competing events, non-fracture death in this case).[27] All models were fitted in R-4.0.0 and STATA 11.2. Plots were generated separately by sex, for all patients and for subgroups of age and CCI based on summary statistics pooled across the imputed datasets.

The study funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all of the data and the final responsibility to submit for publication.

Results

There were 2,747,409 women and 2,684,730 men included in the analysis, with a mean age of 50.7 and 48.5 years respectively (table 1). The study population was similar to the previously published QFracture internal validation population in term of mean age, sex, BMI and ethnicity but there was a higher recorded prevalence of previous major osteoporotic fracture, nursing or care home residence, and many long-term conditions including type 2 diabetes, history of falls, dementia, cancer, asthma or chronic obstructive pulmonary disease, chronic renal disease, malabsorption, and epilepsy or anticonvulsant prescription. For the population evaluated in relation to major osteoporotic fracture, median follow-up was 5.7 (IQR 2.2-10.5) years in women and 5.6 (2.2-10.4) years in men. For hip fracture, median follow-up was 5.9 (IQR 2.2-10.6) years in women and 5.7 (2.2-10.4) years in men.

The crude incidence of both major osteoporotic fracture and hip fracture was higher in women than men (MOF 6.12 per 1000 person-years in women vs 2.26 in men; hip fracture 2.30 vs 0.88 respectively) (supplementary tables S7 and S8). There was a marked age gradient for both outcomes with sex differences being larger in older age (eg in women aged 30-34, MOF 0.95/1000 person-years rising to 33.53 aged 80-99; in men aged 30-34 1.02/1000 person-years rising to 15.42 aged 80-99) (supplementary tables S9 and S10). Across the whole population, MOF incidence in this study was 4.22/1000 person-years of follow-up, compared to 2.45/1000 in the previously published updated QFracture internal validation cohort,[12] and 2.89/1000 in a previously published CPRD validation cohort.[12] For hip fracture, overall incidence was 1.60/1000 person-years, compared to 1.32/1000 in the same previously published CPRD validation cohort.[28] Two-thirds (64,163; 67.1%) of MOF in women and half (17,276; 50.3%) of MOF in men were in people aged 65 years and over. For

Table 1: Baseline data in external validation cohort and in previously published QFracture internal validation cohort[4]

	This study external validation cohort		Previous study QFracture internal validation cohort* [12]
	Women N= 2747409 (50.6)	Men N=2684730 (49.4)	All patients N=1583373
Mean (SD) age	50.7 (17.4)	48.5 (15.6)	50 (1.6)
Mean (SD) body mass index	26.6 (6.0)	27.1 (4.8)	26.1 (4.6)
Women (%)	2747409 (50.6)		804 563 (50.8)
Ethnicity			
White or not recorded	2614423 (95.2)	2556923 (95.2)	1 493 455 (94.3)
Indian	25420 (0.9)	27087 (1.0)	17 670 (1.1)
Pakistani	11121 (0.4)	12316 (0.5)	6489 (0.4)
Bangladeshi	3473 (0.1)	4972 (0.2)	4191 (0.3)
Other Asian	18896 (0.7)	17758 (0.7)	10 779 (0.7)
Black Caribbean	4780 (0.2)	4030 (0.2)	10 144 (0.6)
Black African	22736 (0.8)	20776 (0.8)	17 367 (1.1)
Chinese	7358 (0.3)	5517 (0.2)	5206 (0.3)
Other ethnic group	39202 (1.4)	35351 (1.3)	18 072 (1.1)
Smoking status			
Non-smoker	1146025 (41.7)	807294 (30.1)	773 198 (48.8)
Ex-smoker	390520 (14.2)	439503 (16.4)	257 087 (16.2)
Light (<10 cigarettes/day)	135272 (4.9)	125229 (4.7)	94 400 (6.0)
Moderate (10-19 cigarettes/day)	188078 (6.8)	190990 (7.1)	113 757 (7.2)
Heavy (10+ cigarettes/day)	107288 (3.9)	158134 (5.9)	86 787 (5.5)
Current smoking amount not recorded	43957 (1.6)	78372 (2.9)	65 106 (4.1)
Not recorded	780226 (26.8)	963580 (33.0)	193 038 (12.2)
Alcohol status			
None	570900 (20.8)	317208 (11.8)	330 695 (20.9)
<1 unit/day	854476 (31.1)	548761 (20.4)	402 847 (25.4)
1-2 units/day	561603 (20.4)	669776 (24.9)	287 441 (18.2)
3-6 units/day	52785 (1.9)	224507 (8.4)	84 478 (5.3)
7-9 units/day	5750 (0.2)	38273 (1.4)	8743 (0.6)
>9 units/day	2993 (0.1)	9583 (0.7)	7429 (0.5)
Not recorded	698902 (25.4)	866,622 (32.3)	461 740 (29.2)
Previous major osteoporotic fracture	152417 (5.5)	113520 (4.2)	27 907 (1.8)
Parental history of osteoporosis or hip fracture	10561 (0.4)	1077 (0.0004)	4227 (0.3)
Nursing or care home resident	16819 (0.6)	7455 (0.3)	1535 (0.1)
Condition or prescription			
Type 1 diabetes	8747 (0.3)	12008 (0.4)	4322 (0.3)
Type 2 diabetes	81715 (3.0)	100009 (3.7)	43 437 (2.7)
History of falls	153841 (5.6)	74368 (2.8)	17 382 (1.1)
Dementia	34892 (1.3)	15036 (0.6)	7791 (0.5)
Cancer	94090 (3.4)	67380 (2.5)	28 203 (1.8)
Asthma or COPD	355014 (12.9)	303541 (11.3)	113 175 (7.1)
Cardiovascular disease	156577 (5.7)	195378 (7.3)	77 824 (4.9)
Chronic liver disease	6093 (0.2)	6753 (0.3)	3216 (0.2)
Chronic renal disease	33274 (1.2)	24395 (0.9)	3413 (0.2)
Parkinson's Disease	7585 (0.3)	8348 (0.3)	3650 (0.2)
Rheumatoid arthritis or SLE	11970 (0.4)	32950 (1.2)	10 091 (0.6)
Malabsorption	34884 (1.3)	27122 (1.0)	8026 (0.5)
Endocrine disorders	25089 (0.9)	5866 (0.2)	7882 (0.5)
Epilepsy or prescribed anticonvulsants	66145 (2.4)	59214 (2.2)	26 271 (1.7)
Prescribed antidepressants	66145 (2.4)	59214 (2.2)	111 229 (7.0)
Prescribed corticosteroid	37169 (1.4)	22632 (0.8)	30 998 (2.0)

Prescribed oestrogen only HRT	33679 (1.2)	127 (0.0)	14 988 (0.9)

SLE: systemic lupus erythematosus, COPD: chronic obstructive pulmonary disease, TIA: transient ischaemic attack

* Only reports whole population so cannot stratify by sex

hip fracture, 32,339 (88.8%) in women and 10.167 (76.0%) were in people aged 65 years and over (supplementary tables S7 and S8).

Although MOF and hip fracture incidence both increased with age in both men and women, the incidence of non-fracture mortality increased more steeply with age (particularly in men). Non-fracture death had similar incidence to MOF in young people, and increasingly greater incidence with age, being over four times as common as MOF in women aged 90-99 and almost ten times as common as MOF in men aged 90-99 years (figure 1, supplementary tables S15 and S16). Non-fracture death had higher incidence than hip fracture at all ages.

In the whole population, QFracture discrimination for MOF was excellent in women (C=0.813) and good in men (C=0.738), and for hip fracture was excellent in both sexes (women C=0.918, men C=0.888) (table 2). However, stratified by age, for both outcomes discrimination was poor to moderate in older adults where fracture prediction is recommended[1] (e.g. for MOF, aged 65-74, C=0.616 for women and 0.660 for men; aged 85-99, C=0.576 for women and C=0.624 for men) (table 2). Stratified by CCI, in all strata discrimination was good for MOF and good-to-excellent for hip fracture.

Table 2: Discrimination and model fit for Major Osteoporotic Fracture and Hip Fracture*

	Women Major Osteoporotic Fracture			Men Major Osteoporotic Fracture		
	Harrell's C	**D**	**R-squared**	**Harrell's C**	**D**	**R-squared**
All patients	0.813 (0.811,0.815)	2.25 (2.24,2.27)	54.8 (54.5,55.1)	0.738 (0.735,0.741)	1.76 (1.74,1.78)	42.4 (41.9,43.0)
Age-group						
30-64	0.709 (0.706,0.712)	1.30 (1.28,1.32)	28.8 (28.2,29.4)	0.625 (0.621,0.630)	0.84 (0.81,0.86)	14.4 (13.6,15.1)
65-74	0.616 (0.612,0.620)	0.71 (0.69,0.73)	10.7 (10.1,11.4)	0.660 (0.653,0.668)	1.00 (0.95,1.04)	19.2 (17.9,20.6)
75-84	0.615 (0.612,0.619)	0.67 (0.65,0.69)	9.6 (9.1,10.2)	0.652 (0.645,0.659)	0.91 (0.87,0.95)	16.4 (15.2,17.6)
85-99	0.576 (0.570,0.581)	0.38 (0.35,0.42)	3.4 (2.9,4.0)	0.624 (0.613,0.636)	0.67 (0.60,0.73)	9.6 (8.0,11.3)
CCI						
0	0.795 (0.793,0.798)	2.08 (2.06,2.10)	50.8 (50.4,51.2)	0.668 (0.664,0.673)	1.22 (1.20,1.25)	26.3 (25.4,27.1)
1	0.801 (0.797,0.805)	2.08 (2.05,2.10)	50.7 (50.1,51.4)	0.730 (0.723,0.737)	1.64 (1.59,1.68)	39.0 (37.7,40.2)
2	0.747 (0.742,0.753)	1.60 (1.56,1.63)	37.8 (36.9,38.8)	0.727 (0.719,0.736)	1.54 (1.49,1.60)	36.3 (34.6,37.9)
3+	0.712 (0.706,0.718)	1.30 (1.26,1.33)	28.7 (27.5,29.8)	0.724 (0.715,0.733)	1.46 (1.40,1.51)	33.7 (32.0,35.4)
	Women hip fracture			**Men hip fracture**		
	Harrell's C	**D**	**R-squared**	**Harrell's C**	**D**	**R-squared**
All patients	0.918 (0.915,0.921)	3.26 (3.24,3.28)	71.7 (71.4,71.9)	0.888 (0.882,0.893)	3.19 (3.16,3.23)	70.9 (70.4,71.3)
Age-group						
30-64	0.832 (0.823,0.841)	2.24 (2.19,2.30)	54.6 (53.4,55.8)	0.765 (0.755,0.776)	1.88 (1.82,1.94)	45.8 (44.1,47.4)
65-74	0.694 (0.687,0.701)	1.20 (1.16,1.24)	25.7 (24.4,27.0)	0.705 (0.694,0.716)	1.29 (1.23,1.36)	28.5 (26.5,30.5)
75-84	0.664 (0.659,0.669)	0.95 (0.92,0.98)	17.7 (16.8,18.5)	0.679 (0.670,0.687)	1.08 (1.03,1.13)	21.7 (20.1,23.3)
85-99	0.601 (0.595,0.608)	0.51 (0.47,0.55)	5.8 (5.0,6.7)	0.637 (0.623,0.651)	0.75 (0.67,0.82)	11.8 (9.8,13.9)
CCI						
0	0.924 (0.919,0.929)	3.36 (3.33,3.39)	72.9 (72.6,73.3)	0.852 (0.844,0.860)	2.84 (2.79,2.89)	65.8 (64.9,66.6)
1	0.899 (0.893,0.905)	2.92 (2.88,2.96)	67.1 (66.4,67.7)	0.872 (0.861,0.882)	2.89 (2.82,2.96)	66.7 (65.6,67.7)
2	0.839 (0.831,0.846)	2.24 (2.19,2.29)	54.5 (53.4,55.5)	0.808 (0.796,0.821)	2.17 (2.09,2.25)	53.0 (51.1,54.7)
3+	0.783 (0.775,0.792)	1.75 (1.70,1.80)	42.2 (40.8,43.5)	0.782 (0.770,0.794)	1.90 (1.83,1.97)	46.4 (44.5,48.2)

CCI: Charlson Comorbidity Index.

* Harrell's C takes values from 0.5 (no better than chance) to 1 (perfect discrimination). A difference of >0.1 has been proposed as indicating a meaningful difference in discrimination.[25] R-squared takes values from 0 (no variation in the outcome is explained by the risk model) to 100% (the risk model explains all variation in the outcome).

Calibration plots are shown in figures 2 to 4, and supplementary figures S2-S9. Where observed MOF rates were estimated without accounting for competing risk (left-hand panels in figures 2-3 and supplementary figures S2-S5), in the whole population for both men and women there was under-prediction of fracture risk at all levels of predicted risk. Stratified by age, there was under-prediction in all age-groups and at all levels of predicted risk except in the very highest predicted risk decile in people aged 80-99 where there was over-prediction. Similar patterns were seen when stratified by CCI with under-prediction in all groups except the most multimorbid at the highest levels of predicted risk.

When observed MOF rates were estimated accounting for competing risk (right-hand panels in figures 2-3 and supplementary figures S2-S5), in the whole population, there was somewhat less under-prediction with some over-prediction in women at highest predicted risk. Stratified by age, under-prediction was present in younger age-groups but to a lesser degree than without accounting for competing risk, but there was considerable over-prediction in women aged 85-99 at higher risk and the majority of men aged 85-99, and over-prediction in men and women aged 75-84 at the highest levels of predicted risk. Notably in these older age-groups, observed MOF risk was either flat or decreased as the decile of predicted risk increased. Similar patterns were seen when stratified by CCI with over-prediction of fracture risk in the most multimorbid (CCI=3+) and in people with CCI=2 at the highest level of predicted risk.

For hip fracture, where observed hip fracture rates were estimated without accounting for competing risk (left-hand panels in figures 4-5 and supplementary figures S6-S9), in the whole population there was larger under-prediction of fracture risk than for MOF at all levels of predicted risk for both women and men. Stratified by age, there was under-prediction in all age-groups and at all levels of predicted risk except the highest two predicted risk deciles in women aged 80-99 where there was large over-prediction of risk, and similar over-prediction in the highest risk decile for men aged 80-99. Similar patterns were seen when stratified by CCI with under-prediction in all groups except the most multimorbid at the highest levels of predicted risk.

When observed hip fracture rates were estimated accounting for competing risk (right-hand panels in figures 4-5 and supplementary figures S6-S9), in the whole population, there was somewhat less under-prediction with some over-prediction in women at highest predicted risk. Stratified by age, under-prediction was less in younger age-groups, but there was considerable over-prediction in both sexes aged 85-99 at higher predicted risk, and over-prediction in both sexes aged 75-84 at the highest levels of predicted risk. As with MOF, in these two older age-groups, observed hip fracture rates were flat or declined across all ten deciles of increasing predicted risk. Similar patterns were

seen when stratified by CCI with over-prediction of fracture risk in the most multimorbid (CCI=3+) and in people with CCI=2 at the highest level of predicted risk.

Discussion

Summary of findings

This external validation of the QFracture risk prediction tool found that it has very good to excellent discrimination in the whole population aged 30-99 years, but has poor to good discrimination in important sub-groups including older patients and those with higher levels of multimorbidity. In contrast, calibration was very poor. When evaluated in its own terms (without accounting for competing risk), QFracture showed consistent under-prediction for both MOF and hip fracture. The most likely explanation for this finding is that fracture ascertainment in this study is more complete since it includes fractures recorded during hospital admission in addition to those recorded in GP EHRs and mortality registration. In this study in women, 14802 (13.5%) of MOF and 6911 (19.0%) of hip fracture were only recorded in hospital admission data, compared to 6,305 (18.4%) MOF and 2,515 (19.1%) hip fractures in men. Restricting fracture ascertainment to GP and mortality data (to match the previously published internal[12] and external validation studies[9,28]), the higher observed incidence of hip fracture in this study was largely explained, but only partially explained for MOF (supplementary tables S11-S14, supplementary figure S1). Additionally the earliest study entry year in this study is 2004 compared to QFracture derivation where it is 1998, and recording of fractures in GP data is likely to have improved over time.

When evaluated against observed fractures estimated accounting for competing mortality risk, then under-prediction in general reduced (because failing to account for competing risk causes over-prediction) but there was large over-prediction at higher levels of predicted risk in older people and in people with more complex multimorbidity. Notably, in people aged 85-99 and people with CCI of 3 or more, calibration was extremely poor with observed risk flat or even declining across deciles of increasing predicted risk. QFracture therefore has two causes of poor calibration which operate in different directions. It under-predicts in all patients because derivation is based on incomplete ascertainment of fracture, and it over-predicts in people with high competing risk of death (primarily the old and the more multimorbid).

Strengths and limitations

The strengths of the study includes the use of linked population data, study conduct consistent with methodology recommendations,[24,29] publication of codesets to facilitate replication, and explicit consideration of both performance in important subgroups and competing mortality risks. The high prevalence of missing data for some predictors is an important limitation that is a problem common

to all studies using routine data. Reflecting that QFracture used post-baseline information for some variables whilst we did not, there was more missing data for body mass index and smoking in this study compared to QFracture internal derivation, although there was similar missingness for alcohol status and ethnicity (supplementary table S6). We used multiple imputation under the assumption that data is missing at random, which is likely reasonable for the imputed variables in this context. Additionally, censoring is common with median follow-up of five to six years in this study, similar to others using this kind of data,[9,15] including QFracture derivation and validation studies.[8,9,12] Although we explicitly accounted for censoring due to death in this study, our analysis like others using this kind of data still assumes that people who deregister from a CPRD practice have the same fracture risk as those who do not. This assumption is very likely strong in older people where deregistration due to moving into extra-care housing or a care home might be associated with higher fracture risk. Studies which can continue to follow-up participants even if they move practice would allow this to be examined, which is increasingly possible with the expansion of data linkage driven by the COVID-19 pandemic. A further limitation is that humeral fractures are most commonly recorded in GP data without specifying whether proximal or more distal, and we therefore defined non-site specific humeral fractures as proximal humerus which may lead to some misclassification (some false positives). However, the majority of humeral fractures are proximal[30] and only including humeral fractures specified as proximal would lead to larger misclassification (a larger number of false negatives). We were also not able to validate identified fractures against a gold-standard manual ascertainment of medical records, but our observed hip fracture rates are similar to registry data.[30] Finally, the QFracture prediction tool being evaluated does not include data on bone mineral density (BMD) because this is not routinely available, and because one of the guideline recommended uses of the tool is to identify those who would benefit from BMD measurement. Including BMD in prediction would be expected to improve predictive performance, but exploration of this was outside the scope of this analysis.

Comparison with other literature

The first version of QFracture[8] was independently externally validated in a similar dataset to this one (THIN) and found to have excellent discrimination and calibration in the whole population,[9] and the updated version (as evaluated in this study)[12] was externally validated in CPRD by the QFracture derivation team again finding excellent discrimination and calibration in the whole population.[28] In this study, discrimination in the whole population for both MOF and hip fracture was similarly excellent. However, given the very large differences in fracture incidence across the age-range studied, any prediction tool where the whole population includes everyone aged 30 to 99 will have excellent discrimination.[31,32] Stratifying by age, discrimination varied from poor to moderate (as

expected when the most powerful predictor of fracture is partially removed by examining age subgroups).[31,32] Unlike these previously published validations in UK data,[8,9,12] calibration was poor.

This study differs from these previously published validations in two ways. First, fracture outcome ascertainment additionally included fractures recorded during hospital admission (as well as those recorded in primary care EHRs and mortality data), and the primary care data used is more recent so recording of fractures may well have improved. Better ascertainment of fractures would be expected to lead to under-prediction by QFracture as observed in this study. Consistent with this, an Israeli external validation using both community and hospital data for ascertainment also observed considerable under-prediction by QFracture.[7] However, since the codesets used by QFracture and in previous validations are unpublished, we cannot examine the extent to which differences relate to different choices of fracture codes to include. Second, this study examined calibration against observed outcomes estimated in the same way as previous external validations (using the Kaplan-Meier estimator which does not account for competing mortality risk) and additionally accounting for competing risk (using the Aalen-Johansen estimator). As expected,[14,16,31] accounting for competing risks led to large changes in observed risk in older people and those with more multimorbidity where non-fracture death is more common, consistent with over-prediction by QFracture in people with high competing mortality risk (despite under-prediction in all patients due to incomplete fracture ascertainment in QFracture derivation).

Implications for policy, practice and research

QFracture and similar clinical prediction tools[28] which include a very wide age range typically have excellent discrimination, but that likely reflects that age is a very powerful predictor of most outcomes.[31,32] As found in this study, excellent discrimination in the whole population is compatible with poor discrimination and very poor calibration in the subgroups most at risk of the outcome (older people and those with high morbidity). Examination of discrimination and calibration stratified by age (and other important predictors where applicable) provides a better indication of predictive performance from a clinical perspective. Future research could examine whether fracture prediction models that are more tailored to different age-groups (including pre- and post-menopausal in women) provide better prediction (since for example, osteoporosis may dominate fracture risk in younger people, whereas falls risk may be important in older people).

QFracture in its current form has two major problems. First, this study and a previous external validation[7] in Israel found that it under-predicts risk in general, mostly likely because its derivation is based on incomplete fracture ascertainment. This could be resolved either by recalibration of the existing QFracture tool, or derivation of a new version with better fracture ascertainment. Second, it

does not account for competing mortality risks which leads to considerable over-prediction in people at high risk of death from other causes, notably older people and those with high multimorbidity. Similar over-prediction has been observed for cardiovascular risk prediction models[15,33,34] but the impact is greater for fracture risk prediction because fracture-related death is a smaller proportion of total mortality than cardiovascular disease. This could be resolved by derivation of new models which explicitly account for competing risk. Both of these problems are resolvable, and we emphasise that the problem is not with fracture risk prediction per se, but with the particular implementation of the current version of QFracture.

The FRAX fracture risk prediction tool is also recommended by NICE and does account for competing mortality risk, but systematic external validation is not possible because the prediction algorithm is not publicly available.[6,10] Dagan et al report an external validation of FRAX in Israeli data from primary and secondary care, finding similar levels of under-prediction to QFracture (although their analysis did not account for competing mortality risk).[7] However, FRAX risk prediction was only approximate based on the count of clinical risk factors, rather than based on the actual FRAX risk equation, because the FRAX prediction algorithm has never been made publicly available and therefore replicable. Although FRAX does account for competing mortality risk, exactly how it does this and its performance in external validation remains uncertain. Publication of the full algorithm would allow direct and fair comparison with other tools to identify the optimal tool for different contexts.[7]

There are implications for clinical risk-stratification and for decision-making by patients and clinicians. Bisphosphonates are cost-effective at relatively low thresholds of predicted risk,[1] but misclassification will occur given poor calibration. It is also recommended that decision-making with individuals considers expected benefit for the individual, but patient decision aids mostly rely on being able to reasonably accurately predict individual risk.[35] From this perspective, risk-stratification with the current version of QFracture will under-predict fracture risk in younger and less multimorbid people (and therefore underestimate expected benefit of treatment) and will over-predict fracture risk in older people and those with high multimorbidity (and will therefore overestimate expected benefit of treatment).

There is therefore a need to derive, internally validate, and externally validate new fracture risk prediction models which are based on data with better ascertainment of outcomes, and which account for competing mortality risk. Equally, prediction in the very old requires specific attention, building on small existing studies of prediction in this population.[36] There are plans to update the FRAX model which does account for competing mortality,[37] but publication of the prediction

algorithm will be critical for establishing its external validity will be critical for establishing its external validity.[24]

Conclusion

This study found that QFracture under-predicts in general because its derivation is based on incomplete fracture ascertainment, and considerably over-predicts in groups with high risk of death from other causes because it does not account for competing mortality risk. Its use in clinical practice therefore needs review, particularly in people at high risk of death from other causes.

Author contribution

The study was conceived of and designed by BG, DRM, and PTD who obtained the funding. All authors contributed to study design and interpretation. SL, CE and MM led data management and SL led analysis supported in both by BG, DRM and PTD. SL and BG drafted the paper, which all authors reviewed and edited. SL, BG, MM and DRM verified the underlying data.

Competing interests

No competing interests to declare. BG reports funding from NIHR, Legal and General PLC, Medical Research Council, and Chief Scientist Office unrelated to this study. DM reports funding from NIHR, Chief Scientist Office and Tenovus unrelated to this study. PTD reports funding from EU Health FP7 and Chief Scientist Office unrelated to this study.

Ethics approval

The study was approved by the Clinical Practice Research Datalink Independent Scientific Advisory Committee protocol 16_248.

Funding

This study/project is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research Programme (project reference 15/12/22). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The authors had full and sole access the data, and the funder had no role in the conduct of the research or the decision to publish. Dr. Eke is supported by a research grant from Legal and General PLC (Advanced Care Research Centre), and Dr. Morales by a Wellcome Trust Clinical Research Development Fellowship (Grant 214588/Z/18/Z).

Public involvement

Public contributors were involved in the design and conduct of the study as members of the study steering group, and we would like to acknowledge the contribution of Graham Bell and Alison Allen.

Data sharing

The data controller is the Clinical Practice Research Datalink (CPRD), and under the data licence granted, the authors are not allowed to share data. Researchers can apply to CPRD directly for access to the raw data.

Figure legends

Figure 1: Major osteoporotic fracture (MOF), hip fracture and non-fracture death incidence in women and men

Figure 2: Calibration for major osteoporotic fracture in women without accounting for competing risks (left hand) and accounting for competing risks (right hand). Coloured line (observed risk) above matching black line (predicted risk) indicates under-prediction; below black line indicates over-prediction.

Figure 3: Calibration for major osteoporotic fracture in men without accounting for competing risks (left hand) and accounting for competing risks (right hand). Coloured line (observed risk) above matching black line (predicted risk) indicates under-prediction; below black line indicates over-prediction.

Figure 4: Calibration for hip fracture in women without accounting for competing risks (left hand) and accounting for competing risks (right hand). Coloured line (observed risk) above matching black line (predicted risk) indicates under-prediction; below black line indicates over-prediction.

Figure 5: Calibration for hip fracture in men without accounting for competing risks (left hand) and accounting for competing risks (right hand). Coloured line (observed risk) above matching black line (predicted risk) indicates under-prediction; below black line indicates over-prediction.

References

1. National Institute for Health and Care Excellence. Bisphosphonates for treating osteoporosis. London, UK: National Institute for Health and Care Excellence, 2017.
2. National Institute for Health and Care Excellence. Short clinical guideline CG146 - Osteoporosis: fragility fracture risk. London: National Institute for Health and Care Excellence, 2012.
3. National Osteoporosis Guideline Group. Osteoporosis: clinical guideline for prevention and treatment. Sheffield: National Osteoporosis Guideline Group, 2014.
4. National Osteoporosis Guideline Group. Clinical guideline for the prevention and treatment of osteoporosis. National Osteoporosis Guidelines Group on behalf of Bone Research Society, British Geriatrics Society, British Orthopaedic Association, British Orthopaedic Research Society, International Osteoporosis Foundation, Osteoporosis 2000, Osteoporosis Dorset, Primary Care Rheumatology Society, Royal College of General Practitioners, Royal Osteoporosis Society, Royal Pharmaceutical Society, and Society for Endocrinology, 2017.
5. Cosman F, de Beur SJ, LeBoff MS, et al. Clinician's Guide to Prevention and Treatment of Osteoporosis. Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA 2014;25(10):2359-2381. (In eng). DOI: 10.1007/s00198-014-2794-2.
6. 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. Ann Rheum Dis 2015;74(11):1958-67. (In eng). DOI: 10.1136/annrheumdis-2015-207907.
7. Dagan N, Cohen-Stavi C, Leventer-Roberts M, Balicer RD. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. BMJ 2017;356:i6755. DOI: 10.1136/bmj.i6755.
8. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores2009.
9. Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores2011.
10. 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(8):1033-1046. DOI: 10.1007/s00198-007-0343-y.
11. Hippisley-Cox J, Coupland C. Validation of QFracture compared with FRAX: analysis prepared for NICE 2011. Nottingham, UK: University of Nottingham, 2011.
12. 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. (Journal Article). DOI: 10.1136/bmj.e3427.
13. 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(4):385-397. (In English). DOI: 10.1007/s00198-007-0543-5.
14. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. International Journal of Epidemiology 2012;41(3):861-870. DOI: 10.1093/ije/dyr213.
15. Livingstone S, Morales DR, Donnan PT, et al. Effect of competing mortality risks on predictive performance of the QRISK3 cardiovascular risk prediction tool in older people and those with comorbidity: external validation population cohort study. The Lancet Healthy Longevity 2021;2(6):e352-e361. DOI: https://doi.org/10.1016/S2666-7568(21)00088-X.
16. Leslie WD, Lix LM, Wu X. Competing mortality and fracture risk assessment. Osteoporos Int 2013;24(2):681-688. (In English). DOI: 10.1007/s00198-012-2051-5.
17. Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer, 2009.

18. Van Staa TP, Abenhaim L, Cooper C, Zhang B, Leufkens HGM. The use of a large pharmacoepidemiological database to study exposure to oral corticosteroids and risk of fractures: validation of study population and results. Pharmacoepidemiology and Drug Safety 2000;9(5):359-366. DOI: 10.1002/1099-1557(200009/10)9:5<359::AID-PDS507>3.0.CO;2-E.
19. Metcalfe D, Masters J, Delmestri A, et al. Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases. BMC Medical Research Methodology 2019;19(1):115. DOI: 10.1186/s12874-019-0753-5.
20. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011;45(3):67. DOI: 10.18637/jss.v045.i03.
21. Rubin D. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons, 1987.
22. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology 2015;44(3):827-836. DOI: 10.1093/ije/dyv098.
23. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. British Journal of Clinical Pharmacology 2010;69(1):4-14. DOI: 10.1111/j.1365-2125.2009.03537.x.
24. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology 2014;14:40. DOI: 10.1186/1471-2288-14-40.
25. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Statistics in Medicine 2004;23:723-748. DOI: DOI: 10.1002/sim.1621.
26. Altman D, Vergouwe Y, Royston P, Moons K. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
27. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 2007;26(11):2389-2430. DOI: 10.1002/sim.2712.
28. 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(8). DOI: 10.1136/bmjopen-2014-005809.
29. Collins G, Reitsma J, Altman D, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ 2015;350:g7594.
30. Bergh C, Wennergren D, Möller M, Brisby H. Fracture incidence in adults in relation to age and gender: A study of 27,169 fractures in the Swedish Fracture Register in a well-defined catchment area. PLoS One 2020;15(12):e0244291. (In eng). DOI: 10.1371/journal.pone.0244291.
31. Kanis JA, Oden A, Johansson H, McCloskey E. Pitfalls in the external validation of FRAX. Osteoporos Int 2012;23(2):423-431. DOI: 10.1007/s00198-011-1846-0.
32. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928-35. (In eng). DOI: 10.1161/circulationaha.106.672402.
33. Koller MT, Leening MJG, Wolbers M, et al. Development and Validation of a Coronary Risk Prediction Model for Older U.S. and European Persons in the Cardiovascular Health Study and the Rotterdam Study. Annals of Internal Medicine 2012;157(6):389-397. DOI: 10.7326/0003-4819-157-6-201209180-00002.
34. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. Epidemiology 2009;20(4):555-561. DOI: 10.1097/EDE.0b013e3181a39056.
35. National Institute for Health and Care Excellence. Bisphosphonates for treating osteoporosis: Patient decision aid https://www.nice.org.uk/guidance/ta464/resources London, UK: National Institute for Health and Care Excellence, 2019.

36. Lam M-T, Sing C-W, Li GHY, Kung AWC, Tan KCB, Cheung C-L. Development and Validation of a Risk Score to Predict the First Hip Fracture in the Oldest Old: A Retrospective Cohort Study. The Journals of Gerontology: Series A 2020;75(5):980-986. DOI: 10.1093/gerona/glz178.
37. Vandenput L, Johansson H, McCloskey EV, et al. Update of the fracture risk prediction tool FRAX: a systematic review of potential cohorts and analysis plan. Osteoporos Int 2022. DOI: 10.1007/s00198-022-06435-6.