

CRANFIELD UNIVERSITY

SCHOOL OF HEALTH

PhD THESIS

Academic Year 2011-2012

Ernest Kofi Poku

The clinical utility and cost-effectiveness of non-bone mineral density
based prognostic tests for osteoporotic fracture

Supervisor: Dr Jeff Newman

FRIDAY 29TH JUNE 2012

© Cranfield University 2012. All rights reserved. No part of this publication
may be reproduced without permission of the copyright owner.

Abstract

This thesis investigated a number of aspects of the clinical utility and cost-effectiveness of clinical risk factors, online multivariate algorithms, DXA and a nail-based Raman spectroscopy test, BQT for fracture risk using archived nail samples, and questionnaire data from the Nurses' Health Study which followed women for up to 23 years.

The results showed that the BQT in combination with CRFs improved the results over CRFs alone, using logistic regression and Cox's proportional hazards analysis. The improvement seen was larger using the Cox model, indicating that time is an important factor.

The multivariate algorithms, FRAX and QFractureScores were compared in retrospective and cohort models and found to be predictive, but the relative performance of the two algorithms was highly dependent on the input data.

Reclassification is an exciting new approach to evaluating the addition of new biomarkers in multivariate algorithms and was found in the Nurses' Health Study to provide better discrimination than AUC.

Cost-effectiveness analysis using Markov and decision tree approaches showed that the BQT with a low cut-off in combination with DXA was consistently on the cost-effective frontier, indicating that this new biomarker would be an integral part of any mass screening strategy.

In conclusion, it is clear that the use of the BQT can enhance the performance of clinical risk factors and, with further improvements, the combination may offer a cost-effective alternative to the use of DXA for mass screening in multivariate algorithms.

Acknowledgements

My sincere gratitude goes to Dr Jeff Newman, my PhD supervisor, and to Joan Houghton, Nichola White, Kal Karim and the staff of Cranfield University who welcomed me into the fold and provided me with the support I needed to make this experience a positive one.

I would like to thank Dr Mark Towler for approaching me with the idea for a new test for osteoporosis all those years ago; without him, I would never have started this journey. I would also like to acknowledge the support of Dr Niamh Cummins, whose ideas and input were invaluable to my learning experience.

I would like to acknowledge the staff and investors in Crescent Diagnostics for their support in the clinical trials we conducted; the Crescent experience has been one that will always stay with me.

My thanks go to Dr Diane Feskanich of the Nurses' Health Study who kindly allowed me to use data from this study in this thesis, and all the staff and patients involved in trials. Without their enthusiasm and interest, in what was a very unusual test, this project would never have made the progress it had.

I would like to acknowledge Danna Millett, Michelle Poku, Alice Poku, Kwasi Poku, Sandra Poku, Jenny Kendrick, Pippa Anderson and Christine Larson for their support.

Lastly, and most importantly, I would like to thank my family, my wife Susan and my son Fernán, without their understanding and supporting comments this project would not have been possible.

Publications

1. N. M. Cummins, E. K. Poku, M. R. Towler, O. M. O'Driscoll, and S. H. Ralston.
"Clinical risk factors for osteoporosis in Ireland and the UK: A comparison of FRAX and QFractureScores." *Calcified Tissue International*, 89 [2] 172-7 (2011) (Appendix A)
2. J Renwick Beattie, Mark R Towler, Diane Feskanich, John Armhein, Niamh M Cummins, Tom C Bakker Schut and Ernest Poku. "Raman spectroscopy of toenail as prognostic indicator of future hip fracture risk." *23rd International Conference on Raman Spectroscopy*. Conference Abstract. Bangalore, India, August 2012. (Appendix B).
3. E.K. Poku, M.R. Towler, N.M Cummins, J.D. Newman "Developing novel prognostic biomarkers for multivariate fracture risk prediction algorithms". *Calcified Tissue International*, 91 [3]: 204-214 (2012). (Appendix C).

CONTENTS

CHAPTER 1 : LITERATURE REVIEW	1
1.1 Medical diagnostics today	1
1.2 Objectives of the literature review	1
1.3 Scope of the literature review	2
1.4 Osteoporosis care today	3
1.5 The causes of osteoporosis.....	8
1.6. Available Prognostic Tools.....	14
1.7 Treatment options	32
1.8 Why the current prognostic methods need to be improved	40
1.9 Development work to date on a nail structure test for fracture risk.....	48
1.10 Clinical trial design to assess prognostic test performance	63
1.11 The economic evaluation of osteoporosis screening	67
1.12. Literature review summary	79
1.13 Research questions: aim and objectives.....	81
 CHAPTER 2 : THE POTENTIAL OF RAMAN SPECTROSCOPY OF THE TOENAIL AS A PROGNOSTIC INDICATOR OF FUTURE FRACTURE RISK: A 20- YEAR STUDY.....	 84
2.1 ABSTRACT.....	84
2.2 INTRODUCTION.....	85
2.3 METHODS	89
2.4 RESULTS	95
2.5 DISCUSSION	106
 CHAPTER 3 : A COMPARISON OF FRAX AND QFRACTURESCORES IN A CASE-CONTROL STUDY IN IRELAND AND THE UK.....	 110
3.1 ABSTRACT.....	111
3.2 INTRODUCTION.....	112
3.3 METHODS	115
3.4 RESULTS	119
3.5 NOGG RECOMMENDATIONS.....	131
3.6 DISCUSSION	133
3.7 CONCLUSIONS.....	138

CHAPTER 4 : A COMPARISON OF THE QFRACTURESCORES AND FRAX RISK CALCULATORS AND A REGRESSION MODEL WITH THE ACTUAL INCIDENCE OF FRACTURE IN A NESTED CASE-CONTROL WITHIN THE NURSES' HEALTH STUDY COHORT139

4.1 ABSTRACT	139
4.2 INTRODUCTION.....	140
4.3 METHODS	141
4.4 RESULTS	147
4.5 DISCUSSION	156

CHAPTER 5 : ADVANCED AND NOVEL METHODS FOR EVALUATING BIOMARKER ADDED PROGNOSTIC PERFORMANCE IN THE NURSES' HEALTH STUDY161

5.1 ABSTRACT	161
5.2 INTRODUCTION.....	162
5.3 METHODS	165
5.4 RESULTS	169
5.5 DISCUSSION	178
5.6 CONCLUSIONS.....	181

CHAPTER 6 : TRADITIONAL AND ALTERNATIVE HEALTH ECONOMIC MEASURES TO EVALUATE THE COST-EFFECTIVENESS OF OSTEOPOROTIC PROGNOSTIC BIOMARKERS183

6.1. ABSTRACT	183
6.2. INTRODUCTION.....	184
6.3 METHODS	185
6.4 RESULTS	202
6.5 DISCUSSION	210

CHAPTER 7 OVERALL CONCLUSIONS.....215

7.1 THE NEED	215
7.2 BQT DEVELOPMENT	215
7.3 MULTIVARIATE ALGORITHMS.....	216
7.4 PROGNOSTIC MARKER PERFORMANCE EVALUATION.....	216
7.5 COST-EFFECTIVENESS ANALYSIS	217
7.6 FURTHER WORK	217
7.7 RECOMMENDATIONS	218

REFERENCES.....	219
APPENDIX A.....	241
APPENDIX B.....	248
APPENDIX C.....	251
APPENDIX D.....	263

List of Figures

Figure 1.1: The DALYs for osteoporosis compared with other major chronic diseases. Adapted from ‘The burden of osteoporosis’ (Kanis, 2006).....	7
Figure 1.2: Trabecular bone and cortical bone shown to demonstrate their different structures(Photo by Paul Crompton © University of Wales College of Medicine)	8
Figure 1.3: Screenshot of the input webpage of the UK version of FRAX [®] (Kanis <i>et al.</i> , 2009)	24
Figure 1.4: NOGG charts to aid the selection of patients for treatment or further assessment based on fracture risk with and without BMD. The x-axis represents the age of the subjects and the y-axis represents the 10-year probability of a fracture (Kanis <i>et al.</i> , 2009).....	33
Figure 1.5: Raman spectrogram of a human nail and an explanation of the how the peak height ratio for each subject was calculated	49
Figure 1.6: Raman spectrum from bone collagen (Matousek <i>et al.</i> , 2006)	51
Figure 1.7: Raman Spectra from nail keratin in the human fingernail (Moran <i>et al.</i> , 2007) ...	51
Figure 1.8: Figures 6 and 7 overlaid (Cummins, 2009).....	51
Figure 1.9: Image showing the ventral, intermediate and dorsal layers in a nail sample, from Farren <i>et al.</i> , 2004	55
Figure 1.10: The descriptions for the changes in energy levels due to the Raman effect	58
Figure 1.11: A schematic of a confocal micro-Raman spectroscopy instrument. (reproduced from Petry <i>et al.</i> 2003).....	59
Figure 1.12: Steps in the statistical classification of Raman data.....	61
Figure 1.13: The separation of the data using techniques such as PCA	62
Figure 1.14: Relapse-free Kaplan-Meier survival curves for a lung cancer trial from (Clark <i>et al.</i> , 2003).....	65
Figure 2.1: Raman spectra from a female toenail stored in desiccator and room temperature for six days. The abbreviations of the file names: RTD1BS2 (Room Temperature, Day 1, Bottom, Spot2), DESD1TS2 (DESiccator, Day 1, Top, Spot2)	95
Figure 2.2: Raman spectra of toenail soaked in de-ionized water for three time periods, 24hrs, 48hrs and 72 hrs.	96
Figure 2.3: Raman spectra of nail following UV light exposure at durations of up to 120 minutes. The chart on the left shows the normalised spectra and the chart on the right the unmodified spectra.....	97

Figure 2.4: Raman spectra of nail following temperature-based changes at temperatures between 25 deg C and 190 deg C.....	98
Figure 2.5: Mean case and control spectra and the difference between cases and controls magnified one hundred times. The case and control spectra are visible as a single line because they overlap to the naked eye.....	100
Figure 2.6: Areas under receiver operator characteristic curve for prediction models.....	101
Figure 2.7: Area under receiver operator characteristic curve for prediction models in 51 hip fracture cases and 48 controls up to 15 years post nail collection	103
Figure 2.8: Areas under receiver operator characteristic curve for prediction	104
Figure 2.9: Area under receiver operator characteristic curve for prediction models in hip fracture cases up to 15 years post nail collection and their matched controls.	105
Figure 3.1: NOGG assessment and treatment thresholds in relation to fracture probability. The first chart is used when BMD is included in FRAX and the second chart is used when BMD information is available.	113
Figure 3.2. Correlation between FRAX and QFractureScores absolute risks (- BMD) for major fracture (A) and hip fracture (B).....	123
Figure 3.3: Trendline View: Major Fracture for FRAX and QFracture with and without BMD.	124
Figure 3.4: Trendline View: Hip Fracture risk for the three treatment scenarios: FRAX with BMD, FRAX without BMD and QFracture	125
Figure 3.5: ROC curves for the FRAX with BMD (middle line), FRAX without BMD (outer line) and QFracture (inner line) algorithms and major fracture.....	127
Figure 3.6: A comparison of selected CRFs and algorithm outputs	129
Figure 4.1: AUC graph for the FRAX algorithm.....	149
Figure 4.2: AUC graph for the QFractureScores Model.....	150
Figure 4.3: AUC graph for BMI risk factor	151
Figure 4.4: Comparison of ROC Curves for QFracture, FRAX, logistic regression and BMI in the NHS Cohort.....	152
Figure 4.5: Comparison of ROC Curves in NHS subgroup of fracture up to 16 years from baseline	155
Figure 5.1: Kaplan-Meier Curve for the 82 events and 82 censored events.....	169
Figure 5.2: Event timing for those above and below a BMI of 25	170
Figure 5.3: Time to Fracture based on dividing the study group based on HRT use	172

Figure 5.4: Time to fracture based on dividing the study group by Raman score	173
Figure 5.5: Hazard Function: Cox's model clinical risk factors in the NHS group.....	174
Figure 5.6: Hazard Function: Cox's model clinical risk factors	174
Figure 6.1: Decision tree diagram for the four diagnostic strategies	195
Figure 6.2: Markov decision tree with two states well and fracture	196
Figure 6.3: Markov Model Tree.....	200
Figure 6.4: Cost-effectiveness chart: Decision Tree: Four diagnostic scenarios are shown.	202

List of Tables

Table 1.1: Diagnostic options for osteoporosis based on T-scores derived from BMD assessment.....	4
Table 1.2: Comparison of odds ratios for fracture risk at different BMD measurement sites (Marshall, D. 1996).....	14
Table 1.3: Potentially important genetic markers for osteoporosis	15
Table 1.4: Changes in $\alpha\alpha/\beta\beta$ CTX ratio with age, adapted from (Leeming <i>et al.</i> , 2009)	18
Table 1.5: Key Characteristics of primary and validation cohorts used in the development of FRAX, adapted from (Kanis <i>et al.</i> , 2007).....	23
Table 1.6: Risk factors selected for inclusion in the FRAX [®] algorithm.....	23
Table 1.7: Odds ratios for BMD + clinical risk factors in isolation and in combination, adapted from (Kanis <i>et al.</i> , 2007)	25
Table 1.8: Fracture prediction performance characteristics according to age and risk assessment strategy in women, adapted from Johansson <i>et al.</i> (Johansson <i>et al.</i> , 2009)	26
Table 1.9: QFractureScore identified risk factors.....	29
Table 1.10: Significant associations in QFracture algorithm.....	30
Table 1.11: Intervention thresholds for osteoporosis in the UK, adapted from Kanis <i>et al.</i> (Kanis <i>et al.</i> , 2005).....	34
Table 1.12: Effectiveness of various preventative treatments on different types of fracture (adapted by Compston from Compston (Compston, 2000)	36
Table 1.13: Changes in spinal BMD and fracture risk at various sites based on the treatment regime, adapted from (Cefalu, 2004).....	41
Table 1.14: Categories of test results and the potential errors that can arise.....	44
Table 1.15: Clinical factors which influence the growth rates of nail and bone	52
Table 1.16: The major Raman vibrational modes in the human fingernail	55
Table 1.17: The annual treatment costs and fracture prevention effectiveness of osteoporosis treatments in the UK	72
Table 2.1: Clinical risk factors at time of toenail collection in hip fracture	99
Table 2.2: Prediction of hip fracture by Raman test score and clinical risk factors ^a (CRF) among postmenopausal women in the Nurses' Health Study	102
Table 2.3: Diagnostic performance for the different diagnostic scenarios based on alternative cut-off values	105
Table 3.1: CRFs evaluated by the FRAX and QFractureScores Algorithms	114

Table 3.2: Participating centres in the FRAN study	115
Table 3.3: A comparison of the CRFs used by FRAX and QFracture against the data available in the study group.....	117
Table 3.4: Ineligible Subjects in the study group	119
Table 3.5: CRF characteristics of the total study sample split into controls and the fracture group (n=584).	120
Table 3.6: Characteristics of the CRFs in the study sample by geographical region.	121
Table 3.7: Absolute fracture risks (%) for the total study sample and	126
Table 3.8: Diagnostic performance of the different algorithms.....	128
Table 3.9: Pairwise comparison of the FRAX and QFracture algorithms for major fracture	128
Table 3.10: Comparison of the AUCs of CRFs with the FRAX and QFracture algorithms .	130
Table 3.11: NOGG Recommendations for the total study sample for the QFracture and FRAX algorithms for hip and major fracture (n=584).	132
Table 4.1: Clinical Risk Factor data available from the NHS Cohort	143
Table 4.2: CRF data available for the study population	147
Table 4.3: Smoking characteristics in the NHS population.....	147
Table 4.4: Tests of Normality used to evaluate the key CRFs.....	148
Table 4.5: Summary of results for QFracture, FRAX, logistic regression and BMI in the NHS Cohort	153
Table 4.6: Pairwise comparison of four ROC curves developed in the NHS Cohort.....	153
Table 4.7: A comparison of predicted fracture rates in QFracture and FRAX models	154
Table 4.8: AUC values for the NHS subgroup of fracture up to 16 years from baseline.	155
Table 5.1: BMI Case Summary for Kaplan-Meier Analysis	170
Table 5.2: Case summary for alcohol consumption (>24 grams).....	171
Table 5.3: Case summary for Kaplan-Meier analysis of subject divided by Alcohol (any use) and no alcohol consumption	171
Table 5.4: Chi-Square and p-values for each function of the Cox model.....	175
Table 5.5: Cook Method	176
Table 5.6: Pepe Method	176
Table 5.7: Results summary of the improvement gain from adding BQT to the CRF model.	178
Table 6.1: Table of parameters used in base case and their ranges used in the decision tree model.....	189

Table 6.2: Base case predictive value data for the four diagnostic strategies.....	191
Table 6.3: Relative Risks per SD for the four diagnostic strategies for high risk and low risk patients	193
Table 6.4: Relative Risk factors used in the different diagnostic scenarios	194
Table 6.5: Cost data used at the different states of the Markov model.....	197
Table 6.6: Diagnostic test hazard ratios for the high risk and low risk cases in the four diagnostic scenarios	198
Table 6.7: Cost-effectiveness results table for the four diagnostic options	203
Table 6.8: RRR, ARR and NNT results for the four diagnostic scenarios	203
Table 6.9: Cost-effectiveness values for the four diagnostic scenarios	207
Table 6.10: ARR, RRR and NNT results for the four diagnostic scenarios	207
Table 6.11: Reclassification Table for the CRF only model and with BQT added to the model.....	208

Notation

ANOVA:	Analysis of variance between groups
ARR:	Absolute Risk Reduction
AUC:	Area Under the ROC curve
BMD:	Bone Mineral Density
BMI:	Body Mass Index
BQT:	Nail Raman Test/Bone Quality Test
BUA:	Broadband Ultrasound Attenuation
CEA:	Cost-effectiveness Analysis
CERES:	Cranfield Collection of E-Research
CI:	Confidence Interval
CRF:	Clinical Risk Factors
CTX:	Collagen type 1 cross-linked C-telopeptide
CUA:	Cost Utility Analysis
CV:	Coefficient of Variation
DALY:	Disability Adjusted Life Year
DNA:	Deoxyribonucleic acid
DPD:	Deoxypyridinoline
DXA:	Dual energy X-ray Absorptiometry
EMSC:	Extended Multiplicative Signal Correction
FN:	False Negative
FP:	False Positive
FRAN:	Fracture Risk Assessment Nail Study
FRAX:	WHO Fracture Risk Assessment Tool
FTIR:	Fourier Transform Infrared spectroscopy
HR:	Hazard Ratio
HRT:	Hormone Replacement Therapy

HRTUSE:	Use of Hormone Replacement Therapy
ICER:	Incremental Cost-effectiveness Analysis
IDI:	Integrated Discrimination Improvement
ISO:	International Standards Organisation
LDA:	Linear Discriminant Analysis
LOSO:	Leave one subject out
MHRA:	Medicines and Healthcare Products Regulatory Agency
MRI:	Magnetic Resonance Imaging
MWU:	Mann-Whitney Test
NA:	Not applicable
NHS:	Nurses' Health Study
NICE:	National Institute for Clinical Excellence
NOF:	National Osteoporosis Foundation
NOGG:	National Osteoporosis Guideline Group
NNT:	Number needed to treat
NPV:	Negative Predictive Value
NRI:	Net reclassification index
NTX:	collagen type 1 cross-linked N-telopeptide
OR:	Odds Ratio
PC:	Principal Component
PCA:	Principal Component Analysis
PH:	Proportional Hazards
PPV:	Positive Predictive Value
PTH:	Intermittent parathyroid hormone
QUALEFFO:	Quality of life questionnaire of the European Foundation for Osteoporosis
QALY:	Quality Adjusted Life Year

QCT:	Quantified Computed Tomography
QUS:	Quantitative Ultrasonography
RANKL:	Receptor Activation of Nuclear factor kB Ligand
RCT:	Randomised Controlled Trial
RNA:	Ribonucleic Acid
ROC:	Receiver Operating Characteristic
RR:	Relative Risk
RRR:	Relative Risk Reduction
SD:	Standard Deviation
SERM:	Selective Estrogen Receptor Modulator
SOF:	Study of Osteoporotic Fracture
SORS:	Spatially Offset Raman Spectroscopy
SOS:	Speed of Sound
SPSS:	Statistical Analysis Software
SQRT:	Square Root
SR:	Strontium Ranelate
SS:	Disulphide bond
TGF-beta:	Transforming Growth Factor Beta
TN:	True Negative
TP:	True Positive
UK:	United Kingdom
US:	United States of America
UV:	Ultraviolet
WHO:	World Health Organisation

Chapter 1 : Literature Review

1.1 Medical diagnostics today

In 2011, the global *in vitro* diagnostics market was worth \$44bn (Roche, 2011), approximately 5% of the overall healthcare expenditure according to the Lewin Group (The Lewin Group, 2005). The Lewin Group estimates, however, that diagnostics influences between 60% and 70% of healthcare decisions, which illustrates the importance of diagnostics to the overall healthcare system.

Hospitals and reference lab testing organisations dominate the testing market, but there has been a steady increase in the proportion of the overall market carried out in near-patient environments as more sophisticated care is increasingly conducted outside hospitals. Traditionally, the diagnostics industry has focused on diagnosing the presence of diseases, but the trend in recent years has been towards developing tools that are more predictive and prognostic in nature, which can enable the selection of the most appropriate drugs for a particular patient (Urdea *et al.*, 2009). This allows earlier preventive treatment and a delay in healthcare costs until later in life.

Currently, the large diagnostic companies mostly develop new predictive and prognostic biomarkers in-house but an increasing number are developed by hospitals, academic institutions and early-stage companies using venture capital funding for their research activities. One important area of human health research is osteoporotic fracture prevention and diagnosis, where there is a compelling need for better prognosis and early treatment. This thesis evaluates some new potential prognostic tools for future fracture prediction and evaluates their impact on the economics of healthcare.

1.2 Objectives of the literature review

The objective of this literature review is to establish and comment on:

- The healthcare burden of osteoporosis
- The level of medical knowledge about risk factors other than bone mineral density for hip fracture
- The progress made in the use of Raman spectroscopy of the nail as a biomarker for fracture risk

- The knowledge gaps preventing the implementation of a mass screening protocol for osteoporosis fracture risk in the UK and US
- The health economic techniques used to evaluate the cost-effectiveness of screening for osteoporosis.

1.3 Scope of the literature review

During the development of this review, English language peer-reviewed journals and major publications were examined. The review has focused on academic publications over the last decade, as there have been substantial changes in thinking about the management and prevention of the condition, thus reducing the relevance of older publications.

The following sources were consulted for the literature review:

- NICE guidance on osteoporosis
- National Osteoporosis Foundation Clinician Guidelines
- Standard setting and regulatory bodies (ISO and MHRA)
- Peer-reviewed journals and conference proceedings from journals, including Osteoporosis International, Bone and the Journal of Bone Mineral Research
- Government and other regulatory body publications (such as NICE)
- World Health Organization reference texts
- World Intellectual Property Organisation website for patents and patent applications
- British Library EthOS
- Cranfield CERES.

The search strategy used the following keywords: bone quality, bone fracture, FRAX, osteoporosis, bone fragility, bone mineral density, bone markers, hip fracture, ultrasound, cohort studies, collagen cross-links, human nail, fracture risk, principal component analysis, cost-effectiveness, and health economics. SCOPUS and PubMed were used to carry out searches.

The review has focused on papers concerned with the following issues:

- The definition of osteoporosis
- The clinical need for better osteoporosis prediction
- The clinical validity of risk factors for osteoporosis
- The clinical utility of bone mineral density (BMD) testing to predict future fracture
- The clinical utility of non-BMD based risk predictors for future fracture
- The cost-effectiveness requirements for patient selection for treatment
- The diagnostic potential of the human nail
- Confocal Raman spectroscopy
- Methods to conduct cohort studies
- The cost-effectiveness of screen and treat strategies to prevent fracture.

1.4 Osteoporosis care today

1.4.1 Definition of osteoporosis

Osteoporosis is the clinical term used to describe a bone condition that causes low trauma bone fracture in postmenopausal women.

The World Health Organization defined it in 1994:

Osteoporosis is a systemic, degenerative skeletal condition characterized by low bone mass and micro-architectural deterioration of bone, leading to reduced bone strength and an increased risk of fracture. (WHO, 1994)

At the same time, a panel of experts at the WHO quantified the definition to aid physicians in selecting patients for treatment. The revised criteria stated:

Osteoporosis is defined as a BMD that lies 2.5 standard deviations or more below the average value for young healthy women (a T-score of < -2.5 SD). (Kanis, 2007)

Kanis *et al.* report that bone mass and risk of fracture are correlated such that lower bone mass leads to increased relative risk of fracture in postmenopausal women. In particular, they report that a woman with a T score of -3.5 is twice as likely to have a fracture as a woman with a T score of -2.5 .

This bone mass threshold definition has been used worldwide for diagnosis and intervention decisions for the last decade. Table 1.1 displays the different diagnoses that can result from the calculation of T-score using DXA, it further shows the resultant estimated fracture risk. This table was used as the reference standard for identifying women who would benefit from treatment. Women with a diagnosis of osteopenia or osteoporosis would be recommended some form of preventive intervention.

Table 1.1: Diagnostic options for osteoporosis based on T-scores derived from BMD assessment.

Diagnosis	T-Score	Fracture risk
Normal	More than -1	Low
Osteopenic	From -1 to -2.5	Medium
Osteoporosis	Less than -2.5	High
Severe osteoporosis	Under -2.5 + fragility fracture	Very high

The diagnostic reference standard and most commonly used technique for measuring BMD and providing T-scores is dual X-ray absorptiometry (DXA). DXA is considered the most validated and prognostic technique. Kanis *et al.* demonstrate that femoral neck BMD measurements, in particular, are highly prognostic of hip fracture and conclude that the femoral neck should be the reference measurement site by which all other sites are benchmarked (Kanis *et al.*, 2007). The authors conclude by acknowledging the move away from DXA as the sole reference for osteoporosis. The authors should have considered concluding that DXA should be the reference standard for establishing BMD and so distinguish this diagnosis from osteoporosis and future fracture risk in general.

Since 2008, there has been a move away from T-scores as the operating definition of osteoporosis to the use of percentage risk of fracture derived from absolute risks of fracture, and risk calculators based on logistic regression algorithms to establish those risks (McCloskey *et al.*, 2009). This has led to a new definition of the disease, which incorporates broader clinical risk factors, and bases treatment decisions on absolute risk of fracture thresholds over a 10-year period rather than T-scores. This philosophical change brings

osteoporosis into line with other conditions such as heart disease, where patient risks are assessed on an absolute risk of event basis over a 10-year period (Conroy *et al.*, 2003).

The new US National Osteoporosis Foundation (NOF) (NOF, 2008) definition of osteoporosis introduced in 2009 is a departure from the original 1994 definition:

Postmenopausal women and men age 50 and older presenting with the following should be considered for treatment:

- *A hip or vertebral (clinical or morphometric) fracture*
- *T-score ≤ -2.5 at the femoral neck or spine after appropriate evaluation to exclude secondary causes*
- *Low bone mass (T-score between -1.0 and -2.5 at the femoral neck or spine) and a 10 year probability of a hip fracture $\geq 3\%$ or a 10 year probability of a major osteoporosis-related fracture $\geq 20\%$ based on the US-adapted WHO algorithm.*

This major change in the definition of osteoporosis creates an opportunity for new clinical risk factors to be identified and incorporated into a risk assessment quickly and efficiently.

1.4.2 The global healthcare burden of hip fracture

Today, in the seven major countries in the developed world, the US, Japan, and the EU5 (Germany, France, Spain, UK and Italy) there are 143 million women over the age of 50 (United Nations Population Information Network, 2009).

In the developed world, the lifetime risk of a fracture is believed to be between 30% and 40% (Kanis, 2007). It has also been reported that hip fracture incidences show exponential increase with age: the annual incidence rate rises from 0.4 per 1,000 women at 50 years of age to 68 per 1,000 in women aged 85 (Kanis *et al.*, 2002), so older people are particularly at risk from this condition.

Low impact (Wainwright *et al.*, 2005) hip, wrist and vertebral fractures are the most common type of osteoporotic fracture (Wainwright *et al.*, 2005; Burge *et al.*, 2007). Burge *et al.* also demonstrated that hip fractures account for 72% of the costs of fracture in the US healthcare system, but for a minority of the number of fractures. Hip fractures have the highest mortality rates of any type of fracture; they comprise 0.1% of the global burden of all

diseases worldwide and 1.4% of the burden in the developed world, thus hip fractures are a major contributor to the worldwide burden of morbidity and mortality (Johnell and Kanis, 2004). Between 15% and 24% of women will die within 12 months of a hip fracture (Boonen *et al.*, 2004).

It has been estimated that 37% of non-Hispanic white women in the US would be eligible for treatment under the new US NOF guidelines (Dawson-Hughes *et al.*, 2009) and that 37-50% of US women are osteopenic or osteoporotic as defined by the WHO (Looker *et al.*, 1997). These studies conclude that a significant proportion of the over-50s are at risk of osteoporotic fracture, which places a significant burden on healthcare systems around the world.

Burge *et al.* estimated that there were more than 2 million fractures in the US in 2005, resulting in direct healthcare costs of \$17bn in that year alone (Burge *et al.*, 2007). They also projected that this number would grow by 50% by 2025 due to the ageing of the general population, while the number of fractures will reach over 3 million fractures a year at an annual cost of \$25bn by 2025. In the UK, Burge *et al.* estimated, using a Markov model, a cost of £1.8 billion for 190,000 fractures in 2000 (Burge *et al.*, 2001); this was projected to rise to £2.1 billion per annum and 230,000 fractures by 2020. It is interesting to note that the cost per fracture in 2005 in the US was \$8,500 (£5,666 in 2010 US/UK exchange rates) and £9,473 for the equivalent UK figure in 2000. The large difference in values is assumed to be due to significantly higher costs of long-term care in the UK.

In recent years, there has been increased recognition that disability adjusted life year (DALY) outcomes for osteoporosis are worse than for many cancers (Johnell and Kanis, 2006) due to the debilitating consequences of a fracture. This has resulted in an increased awareness of the wider societal risks of osteoporosis within the medical community. The DALY measure takes the impact of disability on the remaining years of life into consideration when assessing the impact of a disease. Figure 1.1 shows the impact of osteoporosis in terms of DALY compared with other major diseases categories. It can be seen that osteoporosis is the second most debilitating condition as defined by DALYs of the conditions listed.

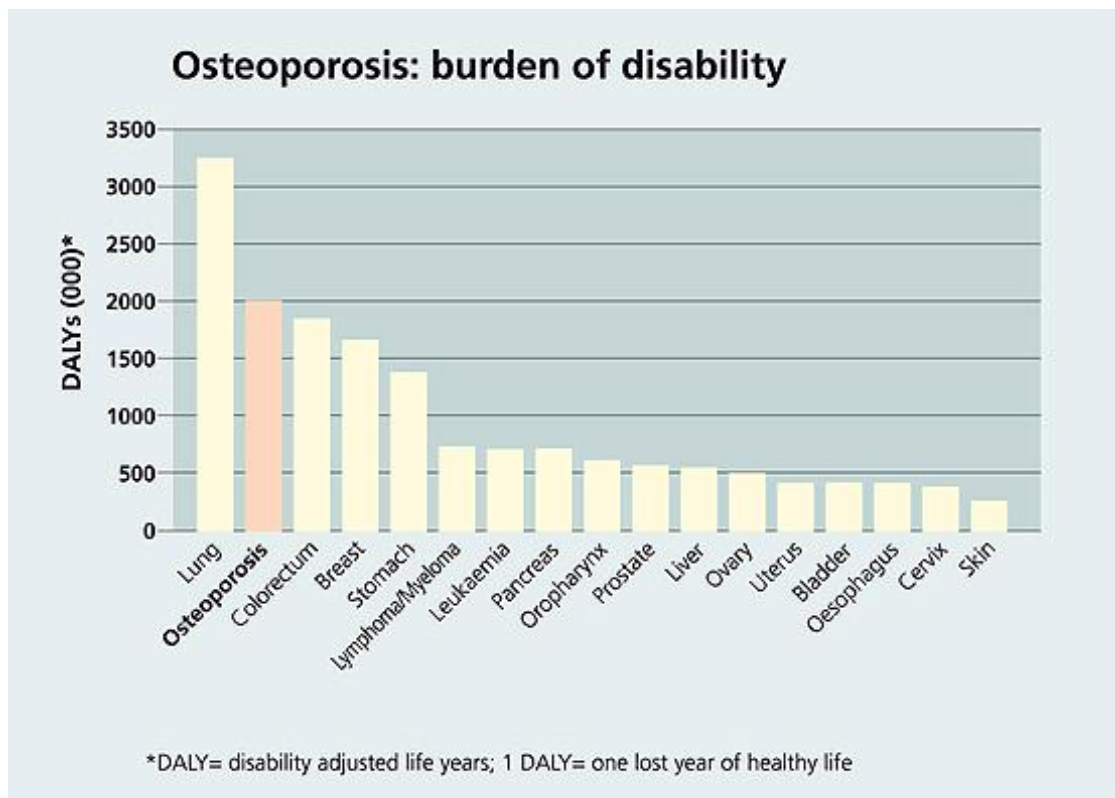


Figure 1.1: The DALYs for osteoporosis compared with other major chronic diseases.
Adapted from 'The burden of osteoporosis' (Kanis, 2006)

There is also now an increased awareness of the impact of osteoporosis and bone fracture on quality of life. Studies using the QUALEFFO, a validated questionnaire specifically focused on quality of life issues in osteoporosis, have raised awareness of the issue (Lips, 1997). A number of studies have shown that older women with hip fractures have significantly lower quality of life compared with age-matched controls (Randell *et al.*, 2000; Bianchi *et al.*, 2005; Lips *et al.*, 1999). Lips reported that an expert panel of the US National Osteoporosis Foundation quantitatively estimated the loss in quality of life utility in the 12 months following hip fracture at 0.4681 caused by hospital and nursing home stays; this compared with a loss of 0.0464 for the 12 months after a distal (non-hip) fracture, a difference in outcome of an order of magnitude (Lips and van Schoor, 2005). These figures are based on a quality-adjusted life year (QALY) scale of between zero (death) and one (perfect health). The authors note that these questionnaires may not reflect the true picture due to the cognitive impairment of older people answering the questions; however, overall, it is apparent that hip fractures have a significant impact on quality of life.

It is clear that the healthcare burden of osteoporosis is significant and that the identification and treatment of women at risk of hip fracture should be a priority for healthcare systems worldwide.

1.5 The causes of osteoporosis

1.5.1. Bone development

Human bone consists of two types of material, trabecular bone and cortical bone. Trabecular bone forms the core of the bone, is more porous than cortical bone and exhibits different elastic properties (Rho *et al.*, 1997; Bayraktar *et al.*, 2004). As early as 1983, Parfitt *et al.* demonstrated that trabecular structure was significantly thinner in hip fracture patients compared with age-matched healthy controls, and they concluded that these structural changes are key to understanding fracture risk (Parfitt *et al.*, 1983). Figure 1.2 shows a cross-section of bone and the trabecular and cortical bone sections can be seen. The cortical bone is much denser and lines the outer surface of the bone, the trabecular bone lies deeper in the bone and has a visible matrix.

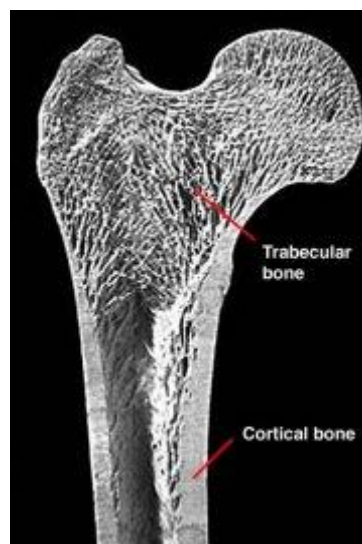


Figure 1.2: Trabecular bone and cortical bone shown to demonstrate their different structures(Photo by Paul Crompton © University of Wales College of Medicine)

Bone is continually being remodelled throughout the human lifetime; the remodelling process within bone is controlled by the level of osteoblast and osteoclast activity (Leeming *et al.*, 2009). Loss of bone strength is associated with an increase in the activity rate and in the

number of osteoclasts in trabecular bone compared with the number and activity rate of the osteoblasts. Osteoclasts are responsible for removing bone (bone resorption), while osteoblasts are cells responsible for rebuilding bone (bone reformation) (Watts, 1999). The ratio between these two cell types changes with age and this ratio affects the type and quantity of bone created (Leeming *et al.*, 2009). To maintain the best quality of bone tissue, bone remodelling on a continuous basis is key (Viguet-Carrin *et al.*, 2006). Leeming *et al.* argue that the rate of bone remodelling is a key component in determining the quality of bone created (Leeming *et al.*, 2009).

A requirement for good bone formation is that the action of the osteoblasts and osteoclasts is coordinated and takes place in a specific order. It is estimated that, following full skeletal growth in adulthood, cortical and trabecular bone remodelling occurs at a rate of 10% per year (Russell, 2001).

1.5.2 Bone strength determinants

At a fundamental level, it can be considered that bones break when the load to which they are subject exceeds the ability of the bone to bear that load (Felsenberg and Boonen, 2005). Felsenberg and Boone introduced the concept of the 'bone quality framework' as a methodology for understanding the determinants of bone strength.

In the bone quality framework, the following factors are important in bone strength:

Structural properties

- Geometry: size and shape of bone
- Micro-architecture: trabecular architecture and cortical thickness/porosity

Material properties

- Mineral: mineral to matrix ratio and crystal size
- Collagen: type and cross-links
- Microdamage and microfracture.

In their view, trabecular bone architecture is critical to bone strength. They refer to a paper by Borah and Babul in which microCT was used to evaluate trabecular bone structure and identify differences between osteoporotic and healthy controls (Borah *et al.*, 2001).

Other authors have taken a similar view of the determinants of bone strength and have settled on four physical determinants of bone strength (Genant *et al.*, 1996; Frost, 2001; Ferretti *et al.*, 1992; Augat *et al.*, 1996):

1. The bone's mechanical properties, or *materials property factor*.
2. The level of fatigue damage to the bone through use, the *microdamage factor*.
3. The quantity of bone, the bone *mass factor*.
4. The size and shape of the bone, the *architectural factor*.

Together, changes in these factors determine bone strength and recent research efforts have tried to identify invasive and non-invasive assessment techniques to evaluate all of these factors. Assessing the bone mass factor, which is commonly quantified using DXA, has achieved the best results to date. The material's property, architectural and microdamage factors are collectively called bone quality factors, a term that is used in this thesis.

Osteoporosis is a broad term used to describe changes in bone structure rather than the more clinically relevant outcome of hip fracture. In other words, a person can be diagnosed as osteoporotic without having a clinically relevant outcome for a number of years, or even at all in the future. The relationship between osteoporosis and hip fracture is similar to the relationship between high cholesterol and myocardial infarction; the link exists, but is not absolute, as other risk factors could be important (Nicholls and Hazen, 2005). Originally, the term osteoporosis described all the relevant bone factors but, in recent years, has become a proxy for BMD alone.

A relationship between BMD and hip fracture has long been established in the literature. In 1993, Cummings *et al.* demonstrated that each standard deviation change in femoral neck density increased the age-adjusted risk of hip fracture 2.6 times in post-menopausal women aged over 65 (Cummings *et al.*, 1993).

Current research efforts are attempting to quantify the other three bone strength determinants. The state of the art is described in the following sections.

1.5.3 Bone mechanical properties factor

The bone mechanical properties factor describes the properties of the bone that provide its elasticity and fracture toughness (Frost, 2001).

As stated in section 1.5.1, the osteoblasts are responsible for the formation of bone proteins, which provide structural strength. Of these proteins, type I collagen is the most abundant. It undergoes structural changes as it ages, which reduces bone strength and, if the bone is not remodelled with better quality bone, the risk of fracture increases (Leeming *et al.*, 2009). The material properties of crystal size, collagen matrix and the mineral-to-matrix ratio are also influential in maintaining good bone strength (Ottani *et al.*, 2001).

Bone mineral density tests provide a degree of information on the level of mineralisation in the bone (vBMD) (Henry *et al.*, 2004). The information is, however, limited because the collagen matrix in bone is not mineralised in a uniform manner and the mineral particles are oriented in the direction of the collagen fibres. This may influence fracture toughness and elasticity in the bone (Boivin and Meunier, 2002; Žižak *et al.*, 2000). BMD testing cannot quantify these types of bone changes.

In the collagen protein, cross-linking structure is key to its overall integrity. The concentration of collagen cross-links in bone in subjects with osteoporosis has been found to be lower than in age-matched controls (Oxlund *et al.*, 1996). Felsenberg *et al.* considered this lower level of cross-linking to be an explanation of why fracture rates are different in individuals with similar levels of trabecular bone material (Felsenberg and Boonen, 2005).

Two studies have demonstrated the link between collagen cross-link formation and bone toughness, stiffness and elasticity (Banse *et al.*, 2002; Vashishth, 2007). Wang *et al.* also demonstrated that collagen cross-linking reduces with age significantly (Wang *et al.*, 2002). Fourier transform infrared (FTIR) analysis of iliac crest bone samples from women with osteoporosis and healthy matched controls has identified significant differences ($P < 0.01$) between collagen cross-link ratios, as defined by the ratio of reducible (pyridinium) to non-reducible (dehydro-dihydroxylysinoxynorleucine) cross-links (Paschalis *et al.*, 2004). This body of work supports the theory that material property changes are occurring in the bones of women at higher risk of fracture and suggests FTIR provides a consistent and quantifiable method to assess those changes, albeit a method that can only be used to analyse bone samples in-vitro, which limits its clinical utility.

There are also published studies which suggest a relationship between fractures and low levels of certain forms of collagen cross-links (Saito *et al.*, 2006a; Saito *et al.*, 2006b).

Taken together these papers represent a growing body of evidence linking collagen cross-link quantity and quality with ageing, osteoporosis and, ultimately, fracture risk. The evidence published to date is not a complete explanation of the mechanism linking bone collagen structure and osteoporosis, because it is difficult to obtain information on bone collagen in well-controlled longitudinal or randomised controlled studies. It is ethically difficult to obtain physical bone samples from large numbers of healthy individuals due to the invasive nature of the procedure; this is a requirement to meet applicable longitudinal study criteria. Animal models, or small-scale cadaver studies by researchers such as Wang *et al.*, have been widely used to infer mechanisms of action; but while helpful in understanding the underlying mechanisms, this type of work does not produce results with clinical utility in humans (Wang *et al.*, 2002). To achieve clinically relevant outcomes, non-invasive tests of bone collagen with the potential for human patient application are required (Saito and Marumo, 2009).

1.5.4 Bone microdamage factor

Bone develops microcracks as a result of daily use, which causes cyclic loading to be exerted on the bone. These cracks can have lengths of up to 20 μm . Non-invasive techniques for examining microdamage do not exist today; techniques such as micro-indentation and confocal laser microscopy (Yin *et al.*, 2009) require direct access to the bone. Yin showed that microdamage in cortical bone is fracture resistant, but this may not apply for trabecular bone. A review by Augat and Schorlemmer considered the evidence for a link between microdamage and increased risk of fracture to be inconclusive (Augat and Schorlemmer, 2006). There is clearly more research work required in this area, but the requirement for bone samples to conduct this type of research and the difficulty of linking the results to long-term prediction of fracture risk due to the requirement for longitudinal studies will continue to limit progress in this area.

1.5.5. Bone architectural factor

Quantified computed tomography (QCT) can be used to quantify the architectural factor and assess bone size and geometry (Griffith and Genant, 2008). In a review, Griffith and Genant describe the progress made using a number of techniques, including QCT, magnetic resonance imaging (MRI) and radiography, to assess bone size and geometry to assist in fracture prediction (Griffith and Genant, 2008). These techniques are, however, only of academic interest at this stage due to their high cost and/or lack of availability in wider clinical care settings.

Majumdar *et al.* noted the potential of MRI to provide information on the biomechanical properties of bone. However, again the costs and availability of MRI mean it would have limited application in primary care (Majumdar, 1998).

One novel technique in the literature is DXL Calscan (Demetech, Sweden) which uses a combination of dual x-ray assessment and laser measurement of the calcaneus to provide a T-score based on a combination of BMD and bone architecture measurement. The laser allows measurement inaccuracies caused by adipose tissues to be reduced. The authors showed, in a prospective study of 4,398 Swedish women, an ability to predict 78% of hip fractures (Brismar *et al.*, 2010). The women were on average 70 years old and the follow-up period was 4 years. A limitation in the study is that, while DXA data was collected and described, a direct comparison of the two measures was not carried out.

1.5.6 Bone quality

This review has established that there are currently few non-invasive technologies that are able to assess bone quality factors – neither the bone's mechanical properties nor the level of fatigue damage to the bone through use, being the microdamage factor. The DXL Calscan results do demonstrate some progress in incorporating bone architectural factors into clinical practice. The other bone quality factors will require similar studies but, before such studies are undertaken, a better understanding of the mechanisms of action linking these factors to fracture outcomes is required.

1.6. Available Prognostic Tools

1.6.1 Dual X-ray absorptiometry for fracture prognosis

Dual X-ray absorptiometry (DXA) is the current reference standard for osteoporosis. The diagnostic procedure requires a subject to lie on a table for around ten minutes while an X-ray scan of their entire body is conducted. The technology has been widely available since the late 1980s and has been shown to be prognostic of hip fracture in a meta-analysis of 11 prospective studies (Marshall *et al.*, 1996).

Table 1.2: Comparison of odds ratios for fracture risk at different BMD measurement sites (Marshall, D. 1996)

Site of measurement	Forearm	Hip fracture	Vertebral fracture	All fractures
Distal radius	1.7(1.4-2.0)	1.8(1.4-2.2)	1.7(1.4-2.1)	1.4(1.3-1.6)
Femoral neck	1.4(1.4-1.6)	2.6(2.0-3.5)	1.8(1.1-2.7)	1.6(1.4-1.8)
Lumbar spine	1.5(1.3-1.8)	1.6(1.2-2.2)	2.3(1.9-2.8)	1.5(1.4-1.7)

Age-adjusted increase in risk of fracture (with 95% confidence interval) in women for every 1 SD decrease in bone mineral density (by absorptiometry) below the mean value for age.

Table 1.2 shows the various potential measurement sites and provide information on the risk of fracture at those sites, it can be seen that, DXA taken at the femoral neck is the best predictor of hip fracture. It is interesting to note that the authors reported figures for the prognostic power of DXA that were superior to those for serum cholesterol as a predictor of cardiovascular disease. A significant area of concern with the widespread use of DXA is that, while BMD measurements are prognostic of fracture risk in the general population, they are still unable to identify the majority of individuals who are at risk of fracture (Marshall *et al.*, 1996). Additionally, there are limitations with the study; the majority of DXA performance studies used in the meta-analysis were carried out in populations with a mean age of over 65. Another limitation is the length of the studies; the investigators in the longest study conducted it over a 24-year period, but it had only 194 subjects. Additionally, the majority of

studies had follow-up periods of less than 10 years. This is an issue when long-term diagnostic decisions are made using DXA in younger age groups or over longer timeframes.

Global DXA usage is not sufficiently widespread to effect a major reduction in the incidence of fracture. Even in the US, Harrington *et al.* reported testing rates of less than 24% in 2002 in at-risk women in a study in Ohio (Harrington *et al.*, 2002); this is due to the high costs and level of support required to manage a DXA screening service. In most countries in the world, there is a gap in the provision of diagnosis that DXA is not able to fill for cost and performance reasons (Kanis, 2007).

1.6.2. Genomic markers for fracture prognosis

There has been significant research interest in genomic markers as potential osteoporosis biomarkers, as they could influence fracture risk by altering bone mass, quality and rates of bone loss in different individuals. The seven genes in table 1.3 were originally considered important candidate genes for osteoporosis (Ralston, 2003).

Table 1.3: Potentially important genetic markers for osteoporosis

Vitamin D receptor	TGF-beta
Oestrogen receptor – alpha	Parathyroid hormone receptor
Type I collagen, COL1A1 and COL1A2	Interleukin 1 receptor antagonist

A meta-analysis of 16 studies investigated COL1A1 and its gene variants and their influence on BMD (Mann *et al.*, 2001), in which different gene variants resulted in significantly different BMD values in postmenopausal women. Another meta-analysis of 12 studies identified an association between the COL1A1 genotype and fracture at an odds ratio of 1.52 (1.27-1.81). Mann *et al.* also investigated the genetic influence within known non-BMD risk factors including bone mass index (BMI), age at menopause and smoking history. They only found a statistically significant relationship between COL1A1 and BMI. This highly cited paper also investigated the relationship between COL1A1 and type I collagen proteins, identifying a statistically significant relationship ($P < 0.05$) between COL1A1 and type I

collagen protein ratios. The authors concluded that COL1A1 genotype influences collagen gene regulation and can lead to abnormal collagen chain production, which affects bone quality.

Twin studies have been widely used to assess the importance of genotype in osteoporosis, finding that between 60% and 85% of BMD variance is genetically influenced (Arden *et al.*, 1996). While genetic relationships in osteoporosis were considered potentially key to understanding fracture risk a decade ago, the link between genetic influenced BMD and clinically relevant fracture has been considered less important in recent years. The heritability of hip fracture is high in fractures that occur in the under 65s (68%), but reduces significantly by the time a woman reaches her 80s (Michaëlsson *et al.*, 2005). Michaëlsson also showed that gene markers were highly influential in extreme forms of bone loss caused by rare bone diseases such as osteogenesis imperfecta.

While there was a lot of initial excitement about the prospect of using genetics to predict future fracture, it has become clear that osteoporosis is not dependent on a few highly influential genes, but rather on hundreds of genes with a small influence on the overall risks (Ralston and De Crombrughe, 2006). The increasing use of whole-genome studies to investigate disease brings new hope for improved clinical utility with genetic tests, but longitudinal studies will be required to establish a compelling link to future fracture. For these multi-gene tests to have clinical utility, significantly cheaper whole genome testing will be required.

Recent research has identified a broader range of candidates that could be influential in bone mineral density. Reppe *et al.* report the following genes as candidates based on an analysis of transcriptional changes in bone samples obtained from within iliac bone in postmenopausal women:

- ACSL3 (acyl-CoA synthetase long-chain family member 3)
- NIPSNAP3B (nipsnap homolog 3B)
- DLEU2 (deleted in lymphocytic leukaemia 2)
- C1ORF61 (chromosome 1 open reading frame 61)
- DKK1 (dickkopf homolog 1)
- SOST (sclerostin)
- ABCA8, (ATP-binding cassette, sub-family A, member 8)
- AFFX-M27830-M-at (an uncharacterised marker).

Dickkopf 1 is a genomic marker of increasing interest in the research community. It has been associated with increased bone resorption and it has been shown that suppression of its activity can lead to a reversal of the damaging effects of excessive resorption (Diarra *et al.*, 2007).

These genes explained 62% of the BMD variation as described by T-score. This was reduced when Z-scores and BMI were considered (Reppe *et al.*, 2010). This type of transcriptome microarray analysis of DNA and RNA represents the future of genome-based research in the osteoporosis field. There is, however, a need to identify additional markers to increase prognostic power and to link the markers to fracture risk using longitudinal studies rather than T-scores.

1.6.3 Biochemical markers for fracture prognosis

There are an increasing number of urine- and serum-based biochemical markers used clinically to manage osteoporosis. There are two major categories, as follows:

- Markers of bone resorption – CTX, N-telopeptide (NTX), cathepsin K and deoxypyridinoline (DPD) are urine or serum-based markers for bone resorption
- Markers of bone formation – osteocalcin, bone specific alkaline phosphatase and procollagen peptides (N and C terminal).

CTX is a very popular marker and is considered to be indicative of post-translational changes in the bone, including isomerisation ($\alpha\alpha/\beta\beta$ CTX ratio) and the creation of advanced glycation end products (AGEs) (Leeming *et al.*, 2009). CTX changes may signify bone quality changes but, as this term has yet to be defined, this statement is purely the opinion of the authors.

Table 1.4 adapted from Leeming *et al.* demonstrates how the CTX ratio changes in bone as it ages. As can be seen, the α protein dominates when bone is young and the β protein dominates when bone is old.

Table 1.4: Changes in $\alpha\alpha/\beta\beta$ CTX ratio with age, adapted from (Leeming *et al.*, 2009)

	Young bone	Aged bone	Old bone
	$\alpha\alpha\alpha\beta$	$\alpha\alpha\beta\beta$	$\alpha\beta\beta\beta$
$\alpha\alpha\beta\beta$ ratio in urine	High	Medium	Low
Clinical Situation	Osteoporotic woman with high bone turnover	Healthy individual	Treatment restores turnover to premenopausal level

A number of studies have shown that bone-remodelling markers are capable of predicting fracture risk (Garnero *et al.*, 1996a). Garnero demonstrated that CTX is prognostic of hip fractures in older women, with an odds ratio for hip fracture of 2.2 (1.3–3.6). The marker was independent of bone mass. BMD and CTX in combination generate a higher hip fracture odds ratio of 4.8 (Garnero *et al.*, 1996b). Garnero proposed in a later study (the Epidémiologie de l'Ostéoporose or EPIDOS cohort study) that a combination of clinical risk factors, such as BMI and the results of a urinary CTX test, could be used in place of DXA as CTX performed as well as DXA in older women (Garnero *et al.*, 1998).

While the Garnero studies demonstrated the utility of CTX to predict fractures, the patient population has limited clinical utility. Garnero *et al.* conducted the EPIDOS study in an older population (over 74 years of age) and the study had a short, 3-year follow-up. The evidence for CTX's clinical utility for the prevention of future fracture over a longer period and in younger women is still lacking.

Another EPIDOS study, which compared older women with premenopausal women over a period of up to 3.3 years of follow-up, confirmed the initial findings that markers of bone formation are prognostic of hip fracture (Chapurlat *et al.*, 2000). Urinary CTX and urinary free DPD were independent predictors of fracture 1.67 (1.19–2.32), and 2.07 (1.49–2.9) respectively. These strong prognostic value figures used a T-score cut-off of –2.0 for high risk, rather than the more conventional value of –2.5. It would have been helpful to see how sensitive the results were to the cut-off value and to understand why –2.0 rather than –2.5 was used.

A study of a number of bone turnover markers and their ability to diagnose hip fracture was carried out in the Study of Osteoporotic Fracture (SOF), involving a US study on women over the age of 65, which included 9,704 non-black women and had 3.8 years of follow-up (Bauer, 2001). The study evaluated a number of markers for their relationship to hip fracture; CTX, NTX, DPD were all examined to assess the risk of fracture in women in the lowest quartile of results in comparison to the rest. The authors found no significant associations and, therefore, the clinical utility of bone turnover markers to predict fracture risk remains uncertain.

There are also concerns with these types of markers about their intra-subject and inter-subject variability. Warnick reported coefficients of variation (CV) within the results of

approximately 20% in individuals over a period of one week (G Russell Warnick, 2001). This level of variability also limits the clinical utility of these markers (Rosen *et al.*, 2000).

1.6.4 Quantitative ultrasonography

Quantitative ultrasonography (QUS) is an alternative technique to DXA for assessing bone mineral density and has been available since the early 1990s. Sound waves passed through bone can measure BMD; higher density material slows the speed of the waves as it passes through. Most testing is carried out at the calcaneus (the ankle bone) because it comprises mostly trabecular bone (95%) and has flat surfaces that aid reliable diagnosis (Guglielmi and de Terlizzi, 2009).

Two measures are made, namely speed of sound (SOS) through the bone and broadband ultrasound attenuation (BUA). SOS and BUA correlate to BMD as measured by DXA and are used to predict osteoporotic fractures. Like DXA, they work best in the over-65s (Hans *et al.*, 1996; Guglielmi *et al.*, 2003; Khaw *et al.*, 2004). Hans demonstrated the prognostic power of QUS in women with a mean age of 80.4 years over a 2-year follow-up period. The relative risk for hip fracture was 2.0 (1.6–2.4) for BUA and 1.9 (1.6–2.4) for SOS compared with 1.9 (1.6–2.4) for BMD, as measured by DXA in the same study.

There has always been a view that QUS measures more aspects of bone structure than just BMD and, as a result, provides some measure of bone quality (Bauer *et al.*, 1997). Langton *et al.* reported linear regression fit (R^2) values between BUA and elasticity (Young's modulus) in calcaneus bone of between 65% and 77%, indicating a relationship between the two values. The potential to incorporate some bone quality measures into an overall assessment of fracture risk has clear clinical utility (Langton and Langton, 2000) and there is now some clinical evidence that QUS is prognostic of hip fracture over a 10-year period; however, more evidence is required (Zhu *et al.*, 2009).

Some authors are concerned about the precision and repeatability of measures using QUS. There are difficulties comparing readings from different instruments owing to measurement repeatability that limits its use in clinical practice (Glüer *et al.*, 2004). Precision values of 1.5% for BUA and 1% for SOS have been achieved, but only in carefully controlled environments and, even then, the prognostic results only reached equivalence with DXA (Stewart and Reid, 2000).

Since 2008, there has been a move to the use of absolute risks for assessing which patients to treat preventively for osteoporosis and, in this context, there has been a reappraisal of the diagnostic potential of ultrasound. In a study of 1,455 people aged between 64 and 76, which included 79 fractures and had 10.3 years of follow-up, an algorithm incorporating both QUS and known clinical risk factors including smoking, prior fracture and alcohol intake achieved good results. The combination of QUS and clinical risk factors achieved a hazard ratio of 2.04 (1.55–2.69) per SD compared with a HR of 2.26 (1.74–2.95) for BMD. The authors concluded that, in terms of absolute risk, the use of QUS is comparable with DXA (Moayyeri *et al.*, 2009).

QUS has achieved limited clinical acceptance due to its measurement variability and inconsistency between devices, in addition to the perception that it offers lower prognostic performance when compared with DXA. The move to absolute risks for assessing fracture risk appears to offer some additional opportunities for the technique to gain wider acceptance, but the number of long-term prospective studies required to confirm the results of Moayyeri *et al.* and the precision issue will continue to be barriers to its wider acceptance.

1.6.5 Risk calculators for fracture prognosis

The drive for new diagnostic methods

The increased interest in regression model risk-based calculators developed from cohort studies has been driven by the need to develop more accurate models for who will have a fracture, when they will have a fracture and the lack of availability of DXA machines in many countries (Kanis *et al.*, 2005). DXA performance is not optimal for detecting osteoporosis due to poor predictive sensitivity, and the use of additional clinical risk factors in combination with DXA could help increase the sensitivity of diagnosis without impairing specificity (Kanis *et al.*, 2009).

The unfavourable health economics of the widespread deployment of DXA make it unlikely that it will become a mass screening modality without the support of other prognostic markers to enhance overall prognostic performance (Kanis and Johnell, 2005). The authors also took the view that neither BMD alone nor a case-finding strategy alone were sufficient to predict hip fracture risk adequately. As a result, primary care physicians require additional

decision-making support to identify which women to treat. Researchers have now undertaken a research programme to validate alternative clinical risk factors for osteoporosis to enhance DXA. Two algorithm-based risk calculators, FRAX and QFractureScores, have been developed to exploit the prognostic power of clinical risk factors for future fracture.

FRAX – a new questionnaire-based tool

The WHO Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK, led by John Kanis, developed the FRAX risk calculator to improve osteoporosis risk assessment. In a paper published in 2005, the team proposed a regression model approach that incorporated additional clinical risk factors. They found that such an approach enriched the diagnosed population by identifying women at higher risk, more so than DXA, when used in isolation (De Laet *et al.*, 2005).

This work was further developed in subsequent years and culminated in a validated algorithm for the calculation of a 10-year fracture risk in men and post-menopausal women, called FRAX. The algorithm, which uses a Poisson regression model to make the risk estimate, was developed using data from nine population cohorts and validated in another 11 cohorts comprising over 1 million patient years (Kanis, 2007). Table 1.5 provides a summary of the key data from the primary and validation cohorts used in FRAX. The FRAX team customised the algorithm for different countries to improve accuracy in different populations. For example, for the UK population, the algorithm incorporates the appropriate weightings for the UK and takes local mortality rates into account (Johansson *et al.*, 2009). Many risk factors were considered for inclusion, but only nine were felt to have sufficient evidence to justify their inclusion in the model (Kanis, 2007). The nine selected are shown in table 1.6. Figure 1.3 shows a screenshot of the web page used to enter the risk data into the FRAX algorithm.

Table 1.5: Key Characteristics of primary and validation cohorts used in the development of FRAX, adapted from (Kanis *et al.*, 2007)

Cohort	Number	% female	Person-years	Hip fracture	Other osteoporotic fracture	Age (mean)
(a) Primary cohorts	46,340	68	189,852	850	3,318	65
(b) Validation cohorts	230,486	100	1,208,528	3,360	15,183	63

Cohorts used included Rotterdam, Rochester, Gothenburg I and II, and the Study of Osteoporotic Fracture.

Table 1.6 below shows the risk factors currently considered to have sufficient evidence to justify their inclusion in FRAX[®].

Table 1.6: Risk factors selected for inclusion in the FRAX[®] algorithm

Current age	Rheumatoid arthritis
Gender	Parental history of hip fracture
Prior osteoporotic fracture	Current smoking
Femoral neck BMD	Alcohol intake (3 or more drinks/d)
Low body mass index (BMI)	
Oral glucocorticoids use > 5 mg/d for 3 months (ever)	

Age, BMD and BMI are entered as continuous variables into FRAX[®].

The following factors are entered into FRAX[®] as dichotomous variables:

- Sex (male/female)
- Prior history of fracture (Y/N)
- Familial history of fracture (Y/N)
- Current smoking (Y/N)
- Alcohol consumption (greater than 3 units a day) (Y/N)
- Long-term corticosteroid use (Y/N)
- Rheumatoid arthritis (Y/N)

FRAX WHO Fracture Risk Assessment Tool

HOME | CALCULATION TOOL | FAQ | REFERENCE

Your Country : **UK** Name / ID : Patient [About the risk factors](#)

Questionnaire:

1. Age (between 40-90 years) or Date of birth
 Age: Date of birth: Y: M: D:

2. Sex ☐ Male ☒ Female

3. Weight (kg)

4. Height (cm)

5. Previous fracture ☐ No ☒ Yes

6. Parent fractured hip ☐ No ☒ Yes

7. Current smoking ☐ No ☒ Yes

8. Glucocorticoids ☐ No ☒ Yes

9. Rheumatoid arthritis ☐ No ☒ Yes

10. Secondary osteoporosis ☐ No ☒ Yes

11. Alcohol 3 more units per day ☐ No ☒ Yes

12. Femoral neck BMD
 T-score

BMI:

The ten year probability of fracture (%) with BMD

Major osteoporotic fracture	23.9
Hip fracture:	8.0

Figure 1.3: Screenshot of the input webpage of the UK version of FRAX[®] (Kanis *et al.*, 2009)

There are a number of potential general and methodology-specific limitations in the FRAX[®] initiative (Kanis *et al.*, 2009). The authors acknowledge that the algorithm will be of limited value to expert clinicians, but believe primary care physicians will find the new algorithm of most use. The burden of the disease resides in primary care, and therefore this is a reasonable justification for the development of the tool. The methodological limitations are that the calculator does not take drug response into consideration and other relevant risk factors, such as falls risk, and biochemical markers of bone turnover have been excluded due to the lack of large international prospective studies validating their use. The developers consider FRAX[®] to be a ‘platform technology’ into which new risk factors can be incorporated as they become available.

The tool is also limited by the relatively few countries that have sufficient data to enable country-specific algorithms to be developed, and low to middle income countries are most in need of alternatives to BMD testing (Kanis *et al.*, 2009). The model is also limited by the inability to consider more than two dosage states for risk factors, such as smoking and alcohol consumption, where levels of consumption have an effect. This, in part, is a weakness

of the studies used to develop the calculator – many did not include dose response questions. Additionally, the stringent criteria used to select studies eliminated some work that incorporated this kind of data. This suggests that, while FRAX may be the gold standard, there may be scope for other risk calculators to offer alternative solutions that incorporate additional risk factors. The developers used a Poisson regression model to build the FRAX model, which means there is scope for other risk calculators to build models using alternative regression techniques that might provide superior prognostic ability.

Table 1.7 shows the odds ratios for 10-year fracture risk using BMD alone, including the FRAX clinical risk factors (CRFs) for hip fracture (Kanis, 2007). It can be seen that the odds ratios for a fracture resulting from low BMD or clinical risk factors is much higher at a younger age than at an older age despite the fact that the incidence rate of fracture is much lower at a younger age.

Table 1.7: Odds ratios for BMD + clinical risk factors in isolation and in combination, adapted from (Kanis *et al.*, 2007)

Age	BMD only	CRFs	BMD+CRFs
50	3.68(2.61-5.19)	2.05(1.58-2.65)	4.23(3.12-5.73)
60	3.07(2.42-3.89)	1.95(1.63-2.33)	3.51(2.85-4.33)
70	2.78(2.39-3.23)	1.84(1.65-2.05)	2.91(2.56-3.31)
80	2.28(2.09-2.50)	1.75(1.62-1.90)	2.42(2.18-2.69)
90	1.70(1.50-1.93)	1.66(1.47-1.87)	2.02(1.71-2.38)

It can be seen in table 1.7, which was developed from the FRAX calculator, clinical risk factors used by themselves as predictors of fracture are weaker than BMD alone. In combination, however, the two offer an enhancement over BMD alone. Table 1.7 also indicates that BMD is highly prognostic of fracture in the 50-60 year old age range, so anyone with low BMD at such a young age is therefore at high risk of fracture. Given the low incidence rate of hip fracture between the ages of 50 and 60 and for the subsequent 10-year period, the clinical utility of using DXA on its own for mass screening for fracture risk is

limited in this age group. The very low prevalence of hip fracture in women between 60 and 70 years of age (less than 1% hip fracture risk per annum) results in the treatment of a very small number of women between 50 and 60 years of age. Health screening modelling has demonstrated that the combined use of FRAX and BMD leads to a higher positive predictive value (PPV), a lower number needed to treat to prevent a fracture and enhanced sensitivity in 55, 60 and 65-year-olds (Johansson *et al.*, 2009). This is shown in table 1.8 where it can be seen that sensitivity and PPV of the test increases with age and when CRFs and BMD are used in combination.

Table 1.8: Fracture prediction performance characteristics according to age and risk assessment strategy in women, adapted from Johansson *et al.* (Johansson *et al.*, 2009)

Age (years)	Assessment strategy		
	CRFs alone	BMD alone	CRFs and BMD
Sensitivity (%)			
50	10.4	35.3	35.9
55	11.9	26.9	33.2
60	12.6	34.7	38.6
65	28.7	55.9	55.8
70	46.4	63.5	62.8
PPV (%)			
50	5.3	1.3	2.5
55	6.6	2.3	4.2
60	5.6	3.7	5.7
65	5.3	6.1	8.6
70	7.5	9.3	11.9

The developers selected a 10-year horizon based on likely treatment duration and on the limitations of the available clinical evidence. Few studies had more than 10 years of follow-up information. It may be more clinically useful to identify the 20-year or lifetime risks for patients and make an early identification of women who are more at risk of a fragility fracture. By identifying at-risk women earlier, it becomes possible to treat them while they still have their peak bone mass near the menopause and intervene to maintain this bone mass

at this point. Barr *et al.* have shown that screening for osteoporosis between the ages of 45–54 and following up with treatment leads to reduced fracture incidence (Barr *et al.*, 2009). Their study used hormone replacement therapy (HRT). Doctors no longer prescribe HRT for osteoporosis, but the study demonstrated the potential of the approach.

When the incidence of hip fracture rises significantly in the 70 and 80 age group, clinical studies indicate that the prognostic performance of BMD, as determined by DXA, falls by more than the performance of the clinical risk factors (Kanis, 2007). There may be a case for using a case-finding strategy and omitting BMD to identify older women who would benefit from treatment.

The development and rapid acceptance of FRAX is an acknowledgement by the medical community of the importance of non-BMD risk factors in predicting osteoporosis. The use of BMD with FRAX[®] does improve prediction (Johansson *et al.*, 2009), but the continued reliance on BMD to obtain a strong performance from the algorithm means that the identification of additional risk factors that are strong enough to replace BMD would be beneficial. Identifying easily identifiable risk factors is a global priority. A cost-effective alternative test could replace DXA in many countries and result in expanded screening programmes, particularly in lower income countries.

FRAX[®] represents a significant step forward in the thinking about osteoporosis and as a ‘platform technology’, it creates an opportunity for the development of new risk factors. These new risk factors could significantly strengthen the calculator and reduce its reliance on DXA tests for BMD, which is particularly needed for the 50-65 age group where DXA testing cannot be economically justified (Kanis, 2007).

The use of 20 cohort studies in the development of FRAX[®] has resulted in an extremely robust and well-validated tool. The nature of the source data limits its value in low and middle income countries where access to BMD testing is limited. One approach to strengthen the calculator for use in countries where DXA availability is limited would be to create another version of the product that incorporates risk factors validated in fewer cohorts (say three) rather than 20; this would open up the algorithm to other risk factors and potentially strengthen it in a clinically useful way for lower income countries. The great advantage of clinical risk factors in the algorithm is that they can be obtained without cost; also, while they may not have the prognostic value of BMD alone, they add significantly to the prognostic

power of BMD and, in the absence of BMD, they can provide an acceptable decision-making tool for clinicians.

A potential methodological weakness of the risk calculator approach to improving the identification of at-risk women is that it avoids the issue of knowing exactly what the mechanisms of action in osteoporosis are. Introducing these risk factors without a better understanding of the mechanisms of action will not aid the development of new treatments to combat the condition.

Another area where the FRAX risk calculator can be improved is in the quantification of the relative risk-reduction benefits of alternative lifestyle changes, or supplement/pharmaceutical interventions. FRAX does not provide information on the relative merits or potential patient benefits of different treatments; rather, the decisions on recommending lifestyle changes or pharmaceutical intervention are currently left entirely to the doctor's discretion. There is currently no widely approved advice to guide them.

QFractureScores

The developers of the QFractureScores algorithm (www.qfracture.org) used a very different approach to the FRAX developers. Their aim was to develop an algorithm that was prognostic without the requirement for laboratory tests, which introduce an external cost to the prevention programme. Data from the QResearch database, a validated database of risk factors and outcome data collected from primary care practices in the UK (Hippisley-Cox and Coupland, 2009), was used to develop the algorithm. This database contains the health records of over 11 million people in England and Wales. The QResearch database contains information on 1,174,232 men and 1,183,633 women aged between 30 and 85, and 7,898,208 (women) and 8,049,306 (men) observation years were used in developing the algorithm. In the female group, there were 24,350 incident fractures and 9,302 hip fractures. A limitation of the work was that no BMD information was available in the database.

Table 1.9 shows the risk factors assessed in the database:

Table 1.9: QFractureScore identified risk factors

Current age	Rheumatoid arthritis
Gender	Parental history of hip fracture
Townsend deprivation score	Current smoking (three states)
Cardiovascular disease	Alcohol intake (six states)
Low body mass index (BMI)	History of falls
Asthma	Type 2 diabetes
Oral glucocorticoids use > 5 mg/d for 3 months (ever)	Gastrointestinal disease
Chronic liver disease	HRT prescriptions
Endocrine conditions	Tricyclic anti-depressant use
Menopausal symptoms	

The hazard ratios and coefficients in the model were derived using Cox's proportional hazards regression model. To validate the QFractureScores model, a hip fracture prognostic performance in a separate defined QResearch group was compared with the actual events over a 10-year period and with the predictions generated by FRAX[®] in the same cohort. The validation group contained 653,789 women and there was an average hip fracture incidence rate of 1.15% (1.13–1.17) (Hippisley-Cox and Coupland, 2009).

The following variables shown in table 1.10 had statistically significant associations with fracture risk:

Table 1.10: Significant associations in QFracture algorithm

Use of HRT	Smoking risk
Use of alcohol	Parental history of osteoporosis
Rheumatoid arthritis	Cardiovascular disease
Type 2 diabetes	Asthma
Tricyclic anti-depressant use	Use of corticosteroids
History of falls	Menopausal symptoms
Chronic liver disease	Gastrointestinal malabsorption
Endocrine disorders	Age
BMI	

The strongest risk factors (adjusted hazard ratios) for hip fracture were:

- Alcohol use > 9 units/day: 2.93
- Heavy smoker: 1.87
- Rheumatoid arthritis: 1.82
- Type 2 diabetes 1.79
- Liver disease: 1.75
- History of falls: 1.69

A comparison of FRAX[®] and QFractureScores was made using the validation cohort. The FRAX[®] algorithm explained 53.83% (54.43% - 55.12%) of female variation, while QFractureScores explained 57.29% (57.18% - 58.09%) of the variation. QFractureScores was therefore slightly superior in this cohort to FRAX[®]. The authors attribute the slightly superior performance of QFractureScores to the fact that FRAX[®] uses data from multiple international databases rather than data derived from the QResearch database. The authors note the low sensitivity of current prognostic tools to predict fracture risk. They did not, however, provide information on the sensitivity and specificity of the algorithm in different age groups in the

published paper. Additionally the requirement to separate women on the basis of alcohol use in excess of 9 units a day which is a rare occurrence suggests that this marker has limited clinical utility beyond this analysis.

A comparison of FRAX[®] and QFractureScores

Both algorithms represent a significant step forward in the identification of women at risk for osteoporosis. The QFractureScore algorithm expands the number of risk factors under consideration for the condition; however, the use of multiple imputation-derived data in the QFractureScores calculator to replace absent BMI, smoking status and alcohol use data is a weakness, as is the lack of integration with BMD measures. The FRAX[®] developers considered BMD information important in the ultimate treatment decision, while data from FRAX[®] papers indicate that performance in combination with DXA is superior to risk factors alone. QFractureScores does offer assessment to younger individuals as its range is between 30 and 85, and it also offers many more dose-response levels compared with FRAX[®], which offers the opportunity for more targeted assessment.

The two organisations behind the algorithms have taken very different approaches to the wider use of their algorithms. The developers of the QFractureScores algorithm have made its underlying coefficients and weightings publicly available, whereas the FRAX[®] team has chosen to keep that information secret to maintain the integrity of the algorithm. This policy prevents changes outside of central control and stops multiple versions of the calculator entering circulation. The two organisations have also used different regression techniques to develop their algorithms. QFractureScores uses Cox's proportional hazards, whereas FRAX[®] uses Poisson's regression.

To establish whether FRAX[®] or QFractureScores is consistently superior, a comparison of both algorithms in a completely independent database is required. This study could also provide sensitivity and specificity information for QFractureScores, which would aid in assessing its potential for mass screening.

1.7 Treatment options

1.7.1 Treatment thresholds in the UK and US

With the introduction of FRAX[®] and the move to absolute risk for determining whom to treat, a requirement has arisen for new treatment threshold guidelines. The UK and US have been among the first to issue new guidelines on thresholds for treatment levels for their own countries (National Osteoporosis Guideline Group, 2009). Figure 1.4 illustrates the UK NOGG guidelines, two charts have been developed, the second, where BMD is not taken shows based on age and percentage risk of fracture whether treatment, no treatment or BMD measurement is required, the other considers only whether treatment is required. These guidelines consider factors such as prevalence of the disease in the population, the performance of the treatment options and health economic factors (Kanis *et al.*, 2008c). In the development of the NOGG guidelines, for example, the authors assumed generic alendronate would cost £24 per month (2008 values) and used a treatment effectiveness outcome of a 35% reduction in fracture risk.

The recommended fracture risk threshold level for treatment has a significant impact on the potential clinical utility of new prognostic markers. Highly effective drugs with few side effects enable the use of less accurate prognostic tools and lower intervention thresholds since the drugs can be safely administered more widely. Health economic issues and screening strategy are therefore important considerations when developing new prognostic markers with adequate clinical utility in the field of osteoporotic fracture prevention.

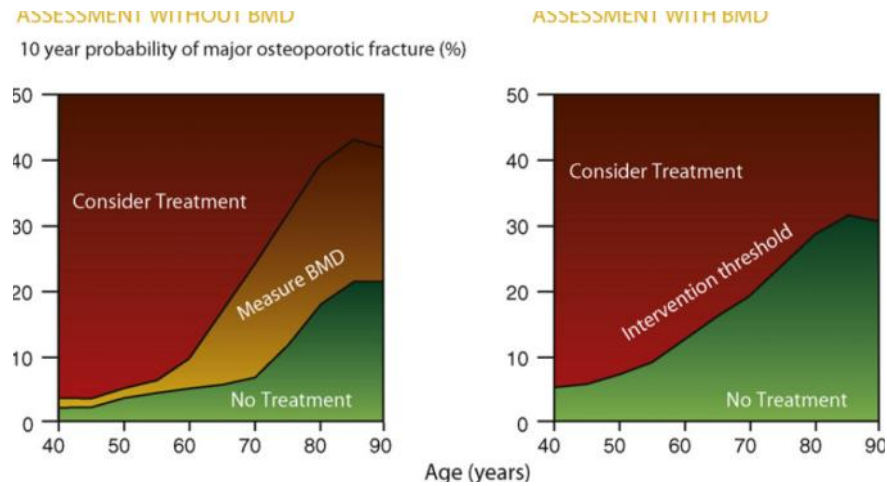


Figure 1.4: NOGG charts to aid the selection of patients for treatment or further assessment based on fracture risk with and without BMD. The x-axis represents the age of the subjects and the y-axis represents the 10-year probability of a fracture (Kanis *et al.*, 2009)

National Osteoporosis Foundation (NOF) Guidelines

The new Clinician's Guide from the NOF (National Osteoporosis Foundation, 2008) states that *postmenopausal women or men over 50 with a T-score of -2.5 or lower at the hip or spine or with a prior hip or spine fracture should be treated.*

In addition, based on absolute fracture risk calculation, patients with low bone mass (T-score between -1.0 and -2.5 at the femoral neck, total hip or spine) should be treated when there is a 10-year probability of hip fracture, that is $\geq 3\%$ or a 10-year probability of a major osteoporosis related fracture that is $\geq 20\%$ based on the US-adapted WHO algorithm. It is important to note that the WHO algorithm is for untreated patients to help decide when to treat, and does not apply to patients already taking an osteoporosis medication.

National Osteoporosis Guideline Group (NOGG) guidelines for intervention

- *Postmenopausal women with a prior fragility fracture should be considered for treatment without the need for further risk assessment, although BMD measurement may sometimes be appropriate, particularly in younger postmenopausal women.*
- *Assessment by the FRAX[®] tool should be undertaken in:*
 - *Men aged 50 years or more (with or without fracture) but with a WHO risk factor or a BMI < 19 kg/m²*
 - *All postmenopausal women without fracture but with a WHO risk factor or a BMI < 19 kg/m².*

Source: (National Osteoporosis Guideline Group, 2009) (Compston *et al.*, 2009)

The treatment thresholds in the UK are higher than in the US, which is driven by health economic considerations. Table 1.11 shows the intervention thresholds in the UK and how they increase with age. The lower threshold represents the point at which further assessment is required and the intervention threshold when treatment is required.

Table 1.11: Intervention thresholds for osteoporosis in the UK, adapted from Kanis *et al.* (Kanis *et al.*, 2005)

Fracture probability (%)			
Age (years)	Lower assessment	Upper assessment	Intervention
50	6	9	7.5
55	7	12	10
60	8.2	15	12.5
65	9.5	19	15
70	11	24	20
75	14	30	25
80	18	36	30

Using a 10-year fragility fracture risk treatment threshold of 3% for hip fractures and 20% for wrist fractures as per the NOF guidelines, the at-risk population in the US eligible for

treatment declines from 53% of non-Hispanic white women to 41%. Overall, 37% of women of all races would be eligible for treatment under the new guidelines (Dawson-Hughes *et al.*, 2009).

A study in Pennsylvania showed that only around 15% of postmenopausal women are typically treated with preventive medicine for osteoporosis (Lee *et al.*, 2006). Another study conducted in a population showed that, in Boston, only 17.5% of postmenopausal women were on medication (Dawson-Hughes *et al.*, 2002). While these studies are not national in scope, they suggest substantial under-treatment in the population.

1.7.2 Therapeutic options

With earlier identification of post-menopausal women at risk of hip fracture, low-cost treatments such as risedronate are effective at reducing fracture risk when compared with no treatment (Borgström *et al.*, 2009).

There are currently three main categories of pharmaceutical treatment approved for use to prevent fractures (Compston, 2009). The first are anti-resorptives, which include bisphosphonates (alendronate, etidronate, ibandronate, zoledronate and risedronate), hormone replacement therapy and raloxifene, a selective estrogen receptor modulator (SERM). The second are anabolic treatments, including Preotact and teriparatide (human parathyroid hormone). Finally, there is a broad category that includes treatments such as strontium ranelate, and upcoming novel therapies at the trial stage. This category includes Prolia (Denosumab), a newly-approved monoclonal antibody for osteoporotic fracture prevention.

Not all the available drugs are equally effective. The data shown in table 1.12, adapted from research by Compston, shows that alendronate, risedronate, HRT and strontium ranelate are the most effective treatments for hip fracture risk reduction (Compston, 2000).

Table 1.12: Effectiveness of various preventative treatments on different types of fracture (adapted by Compston from Compston (Compston, 2000))

Intervention	Vertebral	Non-vertebral	Hip
Alendronate	+	+	+
Risedronate	+	+	+
Etidronate	+	-	-
HRT	+	+	+
Ibandronate	+	-	-
Preotact	+	-	-
Raloxifene	+	-	-
Teriparatide	+	+	-
Strontium ranelate	+	+	

+: Positive effect -: Anti-fracture efficacy yet to be established

Bisphosphonates

Bisphosphonates are synthetic analogues of the naturally occurring compound pyrophosphate, which slow down bone resorption rates (Compston, 2000).

This class of treatment is the most widely used today and is considered highly effective compared with the hormone-based treatments previously available (Ben Greener, 2009).

Many studies have demonstrated the anti-fracture efficacy of the class. A meta-analysis of bisphosphonate treatment by Stevenson *et al.* indicated the class can achieve hip fracture risk reductions of up to 35% (Stevenson *et al.*, 2005; Ben Greener, 2009).

Bisphosphonates are the first-line therapy in most developed countries and their efficacy has been reported in many randomised controlled trials in postmenopausal women. Data pooled from four clinical trials has indicated that risedronate can reduce the risk of hip fracture by up to 60% over 3 years ($p = 0.003$) (McClung *et al.*, 2001). This result, while impressive, is based on a study carried out on women between 70 and 79 years of age who had very severe osteoporosis, and whose BMD scores were four standard deviations below the young mean.

Therefore, this study shows the efficacy of these drugs in older women with severe osteoporosis but does not prove efficacy in younger women with normal BMD over longer timeframes.

There are concerns with compliance with treatment regimens, 75% of women have a compliance rate of less than 80% and that this low level of compliance has resulted in an increased risk of all-causes hospitalisation (Huybrechts *et al.*, 2006). In another study, 42.5% of patients stopped treatment within a year, increasing to 52.1% by the 5-year mark (Solomon *et al.*, 2005). One limitation of these studies is the lack of quantified psychological information on the reasons for the reduction in compliance. Although they do quantify an important issue, more work on the underlying causes of low compliance is required.

Other researchers, e.g. Simon *et al.*, have explored these issues in a different way. In a study comparing once-daily treatments with once-weekly treatments, they established that patients prefer once-weekly treatments due to the inconvenience of remaining upright for 30 minutes every day following dosing with once-daily bisphosphonates (Simon *et al.*, 2002). This issue is now being addressed by the introduction of treatments administered once every 12 months by intravenous infusion, such as zoledronic acid (Black *et al.*, 2007). In this paper, Black *et al.* showed that a single dose of zoledronic acid reduced the risk of vertebral fracture by 77% over a 3-year period in women with an average age of 73.

Intermittent parathyroid hormone (PTH)

Intermittent parathyroid hormone is a treatment regime that involves the intermittent (typically daily) injection of parathyroid hormone into osteoporotic patients. The drug itself is used in two forms, “a genetically engineered 34 amino acid protein with the designation teriparatide (recombinant DNA origin) or recombinant human PTH 1-34” (Deal, 2004).

Intermittent PTH is the most widely used anabolic therapy. Anabolic agents promote bone formation that increases BMD (Rosen and Bilezikian, 2001). Researchers believe PTH controls the calcium homeostatic mechanism, which is key to regulating calcium levels in the body. Reeve *et al.* described the first study to demonstrate the anabolic effects of PTH in 1980. They took biopsy samples of the iliac crest and were able to demonstrate a doubling in bone formation using PTH (Reeve *et al.*, 1980). Subsequent clinical trials have confirmed the efficacy of PTH in reducing fracture risk. Neer *et al.* reported a reduction of 50% in non-spine fractures when compared with a placebo group (Neer *et al.*, 2001).

Lindsay *et al.* also demonstrated that fracture risk reductions endure after the cessation of treatment. This study is relevant because PTH use is limited to a 24-month timeframe (Lindsay *et al.*, 2004), the effects of which last longer than the initial treatment period, thus helping to justify the wider use of PTH.

The potential of bone formation therapies and the effectiveness of PTH has led to a great deal of interest in its mechanism of action. It is interesting to note, however, that no anti-fracture efficacy at the hip has been demonstrated, and this is a key limitation of the therapy (Neuprez and Reginster, 2008).

Strontium ranelate (SR)

Strontium ranelate is an orally active drug that consists of two atoms of stable strontium and one molecule of ranelic acid. Strontium has similar properties to calcium and is treated in the same way by the body's metabolic processes (Marie *et al.*, 2001).

Neuprez *et al.* believe SR influences both bone resorption and bone formation. There is more clinical evidence supporting its bone resorption efficacy than bone formation efficacy in the view of Neuprez *et al.* (Neuprez and Reginster, 2008). They report that SR inhibits bone resorption by affecting osteoclast activity. Ruiz *et al.* conclude from a review of over 800 preclinical and clinical studies that SR improves osteoblast activity, but they take the view that the link to a claim of bone formation for SR needs further evidence (Ruiz *et al.*, 2009). Patients on SR do have increased levels of bone formation compared with those on a placebo, based on serum concentrations of bone-specific alkaline phosphatase (Meunier *et al.*, 2004); therefore, limited evidence for a link to bone formation exists, but more studies would be helpful.

Reginster *et al.* report a 19% reduction in non-vertebral fractures in the Treatment of Peripheral Osteoporosis (TROPOS) trial after one year in women over 74 years of age using SR, and a 36% reduction in hip fracture in a high risk T-score group compared with a low risk group (Reginster *et al.*, 2005). Meunier *et al.* also report a 41% reduction in vertebral fracture after 3 years of treatment using SR. These studies are important because they used large patient samples, 3,640 and 1,649 respectively.

Selective estrogen receptor modulators (SERM)

SERMs influence bone health by interfacing with the estrogen receptors and inducing them to operate as estrogen receptor agonists and antagonists on a tissue dependent basis (Cranney and Adachi, 2005). In bone tissue, SERMs act as agonists and the bone responds as if in the presence of estrogen, which acts to reduce bone resorption. The term SERM relates to any compound that interacts with an estrogen receptor and has similar effects on the tissue. In a review of randomised controlled trials of raloxifene, Cranney *et al.* reported that, in the only study that was adequately powered, namely the Multiple Outcomes of Raloxifene Evaluation trial (MORE), the reduction in risk of vertebral risk fracture was 30% after 3 years of treatment (Cranney and Adachi, 2005). The authors of the MORE trial study also reported a non-significant reduction in non-vertebral fractures (Ettinger *et al.*, 1999).

Prolia (Denosumab)

The current research focus has been on developing therapies that support bone formation. One emerging area of interest is in antibodies that bind to the nuclear factor kb ligand (RANKL). RANKL has been shown to be a major governor of osteoclast activity (Hsu *et al.*, 1999). Denosumab, a monoclonal antibody intravenously administered on a 3 to 6 monthly basis, was designed to effectively bind with RANKL and inhibit its activity. In a study of 412 subjects of an average age of 62, the use of Denosumab resulted in an increase in BMD of between 3% and 6.7% compared with a decrease of 0.8 per cent after 12 months for controls ($P < 0.001$) (McClung *et al.*, 2006).

In the FREEDOM trial at the 2008 American Society for Bone and Mineral Research meeting, Cummings *et al.* reported a 68% reduction in vertebral fracture, a 20% reduction in non-vertebral fracture and a 40% reduction in hip fracture after 3 years of follow-up in patients taking Denosumab on a 6-monthly basis. If similar results were shown in a repeat study, this would represent the best anti-fracture efficacy of any osteoporosis drug available today (Cummings SR *et al.*, 2008).

1.8 Why the current prognostic methods need to be improved

1.8.1 BMD does not predict the majority of fractures

A lack of confidence in BMD as a satisfactory predictor of clinically relevant fracture has led to the introduction of FRAX and an associated increase in the relative importance of clinical risk factors in osteoporosis care. Cefalu, in a review paper, noted two issues that affect the accuracy of BMD measurement: the natural inter-subject variability in BMD scores (and associated measurement inaccuracies), and the fact that BMD is only one of many factors that influence fracture risk (Cefalu, 2004).

Another area of concern with BMD is its ability to predict fracture in the patient population where the burden of fracture occurs. Early papers in the field such as Miller *et al.* note that over 85% of women who fracture have BMD T-scores below -2.5 . However, the paper did not consider the requirement to intervene earlier to prevent the fracture occurring in the first place (Miller *et al.*, 1996). In a subsequent analysis of the SOF, Wainwright *et al.* made a number of important discoveries (Wainwright *et al.*, 2005). They established that 54% of the 243 women in their study who had a hip fracture did not have a diagnosis of osteoporosis as defined by a T-score less than -2.5 . They also note the lack of clinical evidence supporting the use of current anti-fracture therapies in women known to have high BMD. This clearly creates a dilemma for clinicians, who are aware of the burden of fracture in the general population but are unsure if the current treatments are effective in the high BMD group.

An Australian study established that, while the highest rate of fractures was in the osteoporotic group, this represented only 26.9% of the total number of fractures as 73.1% of fractures occurred in women without osteoporosis as defined by BMD (Pasco *et al.*, 2006). This study, conducted over a 5.1 year follow-up period in women between 60 and 94 years of age, adds to the view that, while BMD is useful in a limited numbers of cases, alternative risk factors to identify the vast majority of women at risk of hip fractures are required.

1.8.2 BMD and treatment outcomes are not correlated

Table 1.13 below demonstrates a significant discrepancy between the fracture reduction percentages seen using a number of well-established drugs and the resultant percentage change in associated BMD levels. The Fracture Intervention Trial (FIT) study, as an example, had a 44% reduction in fracture risk for a 6.8% increase in BMD. Additionally the relationship between the two is not linear; there is a large reduction in fracture risk for relatively smaller increases in BMD and the dosage regime is not correlated to reductions in fracture risk (Cefalu, 2004). Table 1.13 presents the data, and shows that vertebral fracture risk reduction exceeds spine BMD increase in all cases. Cefalu concludes by stating that the use of BMD as a surrogate or proxy for anti-fracture efficacy has limited clinical utility by itself. It is clear from the work that BMD is a poor surrogate for assessing the effectiveness of drug treatments to reduce fracture risk.

Table 1.13: Changes in spinal BMD and fracture risk at various sites based on the treatment regime, adapted from (Cefalu, 2004)

Study (duration)	Drug	Spine BMD increase	Vertebral fracture risk reduction	Non-vertebral fracture risk reduction
VERT-NA (3 years)	Risedronate	4.3%	41%	39%
VERT-MN (3 years)	Risedronate	5.9%	49%	33%
FIT-prevalent (3 years)	Alendronate 5 mg/day	6.2%	47%	12%
FIT-absent (4 years)	Alendronate	6.8%	44%	20%
MORE (3 years)	Raloxifene	~2.6%	50%	10%
PROOF (5 years)	Salmon calcitonin	0.6%	33%	12%
Teriparatide (5 years)	Teriparatide	8.6%	65%	35%

FIT: Fracture Intervention Trial, MORE: Multiple Outcomes of Raloxifene Evaluation, PROOF: Prevent Recurrence of Osteoporotic Fractures, VERT-MN and VERT-NA: Vertebral Efficacy With Risedronate Therapy – Multinational and North American trials respectively.

A meta-analysis by Cummings *et al.* of 12 trials which involved different anti-resorptive therapies, and which monitored spine BMD fracture risk reduction compared with placebo,

reported a 54% reduction in fracture risk correlated with an 8% increase in BMD. They conclude that the change in BMD reflects only a small proportion of the change in fracture risk (Cummings *et al.*, 2002).

There is also a significant time difference between treatment efficacy and fracture risk reduction that is becoming apparent using BMD-based measurement tools. Raloxifene has been shown to significantly reduce the risk of fracture almost immediately, with only 6-12 months required for a response to be seen, compared with up to 3 years for spinal bone mineral density to show a change of 2-3% (Qu *et al.*, 2005).

A study of teriparatide showed reductions in hip fracture of 53% and vertebral fracture of 65% for BMD increases of between 9% and 15% (Marcus *et al.*, 2003). This performance is better in comparison with BMD than bisphosphonates, but teriparatide is predominately a second-line therapy. The study also showed that therapeutic response to the treatment was independent of the initial BMD T-score for the patients across a range of between -2.1 and -3.3. This suggests that women with clinical risk factors but without a diagnosis of osteoporosis could benefit from this kind of treatment. The literature overall indicates a weak association between patients' responses to drug treatment and the consequential changes in BMD.

Cefalu described this measurement problem very well (Cefalu, 2004):

“Although BMD may increase with therapy, the increase is not measurable until later, and the overall increase is too small to account for the timing and magnitude of fracture risk reduction.”

A key question that arises from this research is whether the drugs currently used to reduce fracture risk are effective in the osteopenic and normal patient groups, with investigators conducting most research on women with a diagnosis of osteoporosis by T-score in the treatment group. The conclusion Pasco *et al.* reached in their Australian study is that women with normal and osteopenic BMD are at equivalent fracture risk to osteoporotic women and would, therefore, benefit from treatments to increase their BMD. This assertion requires further evidence than is presented in their paper, but seems reasonable. It is also interesting to

note the number of drug response papers that use vertebral fracture rather than hip fracture as an endpoint, as the majority of bisphosphonate drugs have shown good efficacy at the vertebral fracture site. This is despite the fact that the healthcare burden arises from hip fracture. There is a need for more studies to prove the efficacy of bisphosphonates and other drugs at the hip.

1.8.3 The factors influencing mass screening for fracture risk

Screening has been defined by Eddy as:

“The application of a test to detect a potential disease or condition in a person who has no known signs or symptoms of that disease or condition.” (Eddy, 2004)

Osteoporosis is a disease which is asymptomatic before fracture occurs, which fits the criteria for screening very well. Additionally, as discussed in section 1.4, a high number of women will have a fracture in their lifetimes; this will significantly adversely affect their disability adjusted life years. As a result, a screening strategy to identify women who would benefit from preventive treatment may be justified.

At this time in the UK or US there are no mass screening programmes for osteoporosis, but programmes exist for other conditions. A prominent example is breast cancer, as screening programmes for this condition have been running since 1988 in the UK. The rationale for introducing any screening programme requires two competing needs to be quantified and balanced: the expected reduction in mortality and/or disability from screening for the condition, and an evaluation of this benefit compared to the inconvenience and risk associated with the screening protocol (Karsten Juhl Jørgensen and Peter C Gøtzsche, 2010).

Factors involved in formulating a screening strategy

The sensitivity and specificity of a diagnostic tool drives its mass screening potential:

- Sensitivity is a measure of a test's ability to correctly identify people who have the disease as diseased
- Specificity is a measure of a test's ability to correctly identify people as healthy if they do not have the disease.

An ideal test would have a sensitivity and specificity of 100% but, in the absence of such a test, the sensitivity and specificity in the context of the effectiveness of the drugs available and their costs must be considered in any decision to implement a screening programme.

For any test, there are four possible outcomes for a particular patient:

1. A true positive (TP) result where the test has correctly identified an ill person as such.
2. A false positive (FP) result where the test has incorrectly identified a healthy person as ill.
3. A true negative (TN) result where the test has correctly identified a healthy person as such.
4. A false negative (FN) result where the test has incorrectly identified an ill person as healthy.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN})$$

Table 1.14: Categories of test results and the potential errors that can arise

	Ill subjects	Healthy subjects
Positive test	TP	FP (Type I Error)
Negative test	FN (Type II Error)	TN

Table 1.14 shows the categories of diagnostic results and the errors which can arise. A type I error occurs when a healthy subject is identified as unhealthy and a type II error occurs when a diseased subject receives a negative result. The ideal test maximises the true positive and

true negative predictive values. The number of false positives and false negatives will have a significant impact on the screening strategy depending on the magnitude of the harms experienced by people incorrectly diagnosed.

Other measures used to evaluate the performance of diagnostic tests include the positive predictive value (PPV) and negative predictive value (NPV). These measures evaluate the effectiveness of the test for identifying those at risk and those not at risk:

- $PPV = TP / (TP + FP)$
- $NPV = TN / (TN + FN)$

This is particularly important in developing a screening strategy for a future risk. The predictive values of the test are the most relevant values when clinicians assess future risk. The PPV and NPV differ from the sensitivity and specificity of the test because they take the prevalence of the disease into consideration.

Harms

With any screening programme, subsequent diagnosis of osteoporosis and treatment can lead to a number of harms that clinicians should consider before implementing a treatment programme.

Screening using DXA exposes women to radiation, the inconvenience of hospital attendance and anxiety about whether they have the condition (Rimes *et al.*, 1999). Women diagnosed with osteoporosis have more fears than women considered to have normal bone mineral density (Lyles *et al.*, 1993).

A false positive result can lead in the case of osteoporosis to unnecessary treatment that can expose women to a number of risks, including osteonecrosis of the jaw, a condition which affects 1% of women on bisphosphonate treatment (Woo *et al.*, 2006). These potential harms need to be evaluated in the context of drug and diagnostic test performance.

The main factors considered in the analysis of whether a screening programme is required include:

- Gender
- Age
- Testing frequency
- The cost and accuracy of the test
- The cost and performance of the intervention required.

A key measure in defining the viability of a screening strategy is the number needed to treat (NNT) to prevent an occurrence of the condition. By way of comparison, Humphrey *et al.* report the number needed to treat to prevent one death from breast cancer over a 14-year timeframe as 1,224 women (Eddy, 2004; Humphrey *et al.*, 2004). Eddy *et al.* estimate 121 women aged 60 to 64 will need to be treated to prevent one hip fracture (Eddy, 2004). While these figures do not consider the cost of treatment, they show the relative impact of treatment and allow the relative seriousness of the clinical outcome to be considered in any decision to implement a screening strategy.

Clinicians have not accepted DXA, the current gold standard, as a mass screening tool for osteoporotic fracture risk. A report by the UK's Royal College of Physicians recommended that mass screening is not carried out using BMD due to its low sensitivity (Royal College of Physicians, 1999). Using a DXA-derived specificity of 90% results in a sensitivity of only 18%, which is too low for an effective screening technology (Kanis, 2007).

1.8.4 Benefits of screening at the menopause

The age at which a woman is first screened can influence the overall effectiveness of a programme, so screening at different ages needs to be considered in any screening strategy evaluation. A 2007 WHO report on osteoporosis considered the relative merits of screening at the menopause compared with screening later in life (Kanis, 2007). Screening at the menopause could potentially be beneficial because bone loss is rapid at this stage and slows down later in life. However, an issue with screening at the menopause is the relatively low incidence rate of the fractures in women at this stage compared with later life. The WHO

report estimates the 10-year risk of a woman in her 50s at around 1% while, for a woman in her late 70s, it is over 6% (Kanis, 2007).

A longer risk timeframe than the current 10-year period for time to fracture is required to enable the future high fracture risk of women in their 50s to be considered in any treatment decision. From the menopause, a period of 20 years for fracture risk assessment would be more applicable to assess the risks in younger women. On balance, based on the current technology available and issues with compliance with treatment in younger women, the WHO report considers treating women at 65 years of age rather than 50 to be more appropriate.

The introduction of additional clinical risk factors and ultimately FRAX[®] has created the possibility of screening for the condition. Kanis *et al.* have shown that clinical risk factors can have a positive influence on screening strategies by increasing the sensitivity of BMD measures when used in combination with those other factors (Kanis, 2007).

The current peer-reviewed literature does not support the introduction of screening for osteoporosis using BMD due to current performance characteristics. It is interesting to note the lack of peer-reviewed evidence on the ability of screening to prevent fractures, until recently. Humphrey *et al.* reported in 2004 the absence of studies describing the benefits of screening for preventing fracture (Humphrey *et al.*, 2004).

More data on the benefits of earlier screening is becoming available. For example, Barr *et al.* report on a randomised control trial to reduce fracture risk (Barr *et al.*, 2009). In the study, they were able to demonstrate, in 4,800 women between 45 and 54 years of age, the benefits of HRT on a population screened using BMD. Over a 9-year follow-up period, there was a 25.9% reduction in fracture risk compared with the control group. While this study used a now obsolete treatment, HRT, the study demonstrated that screening does result in increased treatment and in a reduction in hip fracture rates. The finding that relatively young women showed such an effect over a 9-year timeframe suggests that screening women under 65 may be valuable, and the results may be even better in the 55-60 age group due to their higher fracture risk.

1.9 Development work to date on a nail structure test for fracture risk

1.9.1 Clinical validation to date

A number of clinical studies conducted at the University of Limerick have identified a potential link between nail structure, low trauma fracture risk and osteoporosis. The researchers identified Raman spectroscopy as the ideal technique to examine the human nail for this condition, and all the early studies were conducted on the fingernail. This test is being developed for commercial applications under the name BQT[®], and it is referred to in this thesis as BQT.

The first study published on the concept by Pillay *et al.* identified differences between the Raman spectral signatures in fingernails collected from healthy women ($n = 4$) and women with osteoporosis as defined by BMD at the hip or lumbar spine ($n = 4$) (Pillay *et al.*, 2005). The Raman spectra were collected using a Raman spectrometer over a range between 400 cm^{-1} and 1800 cm^{-1} . The S-S peak of interest lies at 510 cm^{-1} wavenumbers. The difference in the average half-width maxima for the disulphide peaks between the two groups was nearly statistically significant ($p = 0.06$). This study had a very small number of subjects and so the results were preliminary.

A validation study was then carried out in a larger group by Towler *et al.* in 169 women (Towler *et al.*, 2007); 85 of these women were premenopausal, 84 were postmenopausal and 39 had a history of fracture. In the postmenopausal group, which contained 21 women with a history of fracture and 63 women without a history of fracture, a statistically significant difference between their average S-S peak height was observed ($p = 0.0397$). To ensure consistent measurement between subjects, the measurements were taken relative to the CH_2 peak, at the far end of the range collected at 1450 cm^{-1} . This measure was used in preference to the original method of calculating the half-width maxima at the S-S peak height following discussions with the Raman instrument manufacturer, the half-width maxima represents the height of the peak at the point at which the peak is half its maximum width, it was felt an alternative measuring approach would produce superior results. Figure 1.5 shows a typical Raman spectrum for a human nail and illustrates how the S-S and CH_2 peak height measurements to assess osteoporotic status can be taken. The advantage of this approach was

that using another peak to compare with the S-S peak provided a consistent reference point and eliminated variability in results from the nails caused by non-clinically relevant measurement factors. The measurements were taken for each peak by measuring from peak to trough and the peak height ratio between the two peaks calculated. This ratio represented the value of interest.

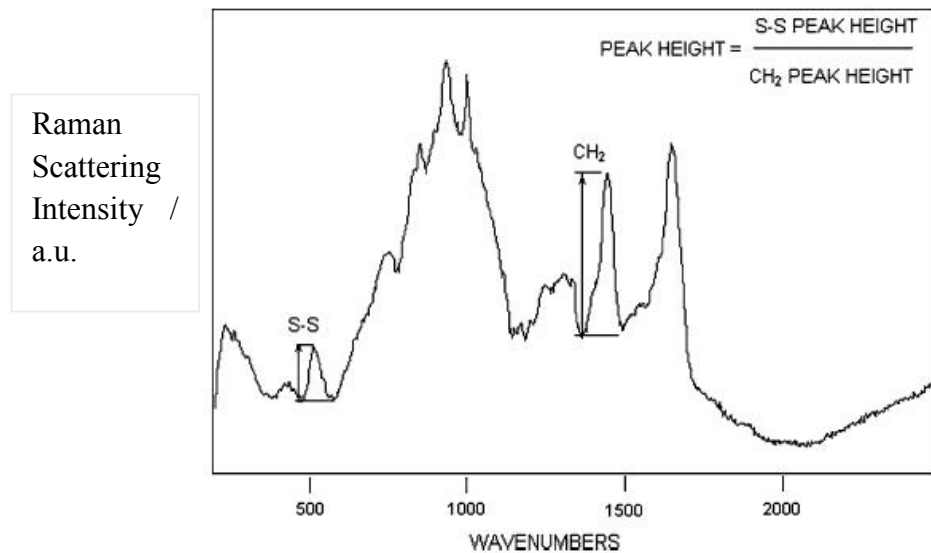


Figure 1.5: Raman spectrogram of a human nail and an explanation of how the peak height ratio for each subject was calculated

Following this work, the investigators conducted two larger retrospective case-control studies. These studies, PREFRAN (n = 134) and FRAN (n = 624), are both described in detail in a PhD thesis by Cummins (Cummins, 2009). In the PREFRAN study, no statistically significant differences were observed between the control and fracture groups, but it is possible that this result was caused by insufficient discrimination in terms of fracture outcomes between the patient groups due to the recruitment process. There were some potential recall bias issues depending on whether women had experienced low impact fractures.

In the FRAN study, the investigators observed statistically significant differences in post-hoc analysis ($P < 0.0001$) between the control and fracture groups. Superior test performance was observed in the osteopenic group compared with the osteoporotic group, suggesting the test may work in women with high BMD scores.

These studies, conducted before the introduction of FRAX[®], were retrospective studies using nail samples collected after the fracture events. A limitation is that the identified fracture subjects were not verified as having experienced low trauma fracture by medical records. There is, therefore, a need for additional studies with larger patient numbers that meet higher evidence criteria; these should be longitudinal cohort studies to assess the ability of the technique to identify absolute fracture risk over periods of 10 years or more.

1.9.2 Relationship between nail growth and bone growth

The novel claims of the Raman spectroscopy-based nail test depend on a measureable relationship between nail growth mechanisms and bone growth mechanisms. There are three approaches that can be used to establish this relationship: first, a mechanistic approach exploring the common processes in the development of the material content of bone and nail; second, an experimental approach monitoring changes in human or animal models; and, third, an observational approach identifying the common diseases between the two groups.

Experimental studies linking nail and bone

Good evidence already exists to show that changes in bone collagen increases bone fragility (Banse *et al.*, 2002; Vashishth, 2007), but there is limited published evidence linking nail and collagen structures. Cummins showed in her thesis that there are similar Raman peaks in nail and collagen structures (Cummins, 2009).

Figures 1.6 – 1.8 show bone collagen and nail keratin Raman signatures followed by an overlaid comparison between the two. Figure 1.8 demonstrates the similarity in height of the CH₂ peak at 1450 cm⁻¹ wavenumbers.

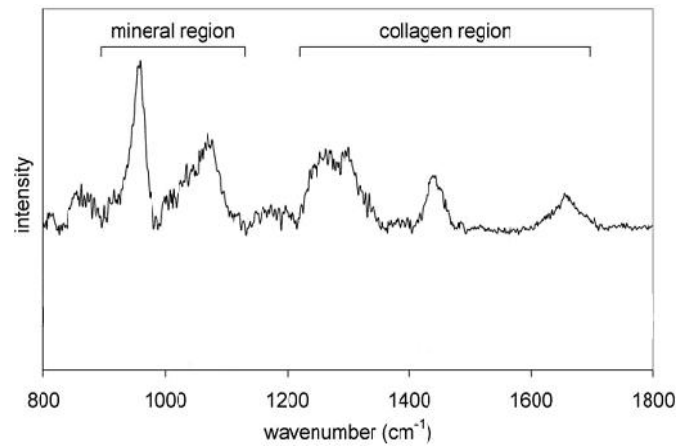


Figure 1.6: Raman spectrum from bone collagen (Matousek *et al.*, 2006)

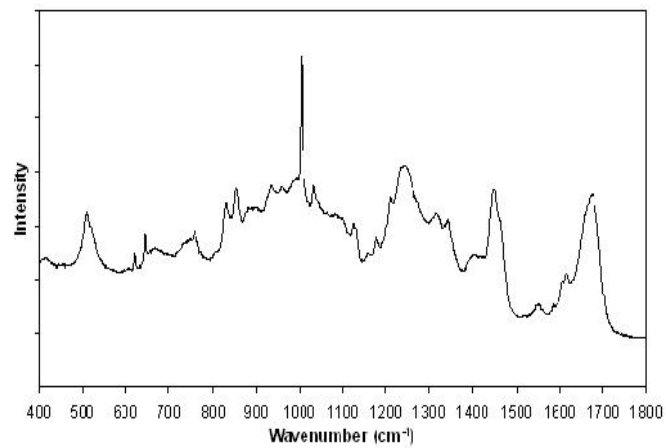


Figure 1.7: Raman Spectra from nail keratin in the human fingernail (Moran *et al.*, 2007)

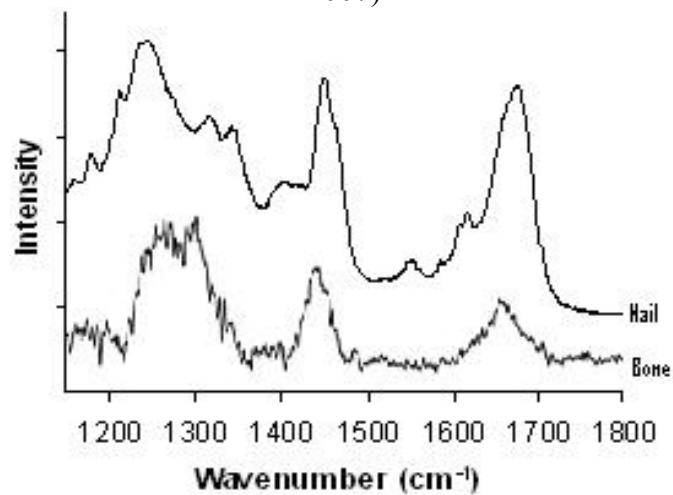


Figure 1.8: Figures 6 and 7 overlaid (Cummins, 2009)

Figure 1 1.8 shows that bone collagen and nail keratin share some common peaks, which may explain the correlation between nail and bone. Towler *et al.* have concluded from this relationship that the link between the two materials is driven by disulphide formation (Towler *et al.*, 2007).

Ohgitani *et al.* reported in 2005 a relationship between nail calcium, magnesium content and lumbar bone mineral density (Ohgitani *et al.*, 2005). Fingernail and toenail clippings were collected from 169 women and 115 men between 20 and 80 years of age. Statistically significant differences were seen in a group of women in their 60s; in these women, calcium concentrations reduced with BMD scores, for fingernails ($p = 0.0016$) and for toenails ($p=0.0215$). Some contradictory results were seen in pre- and postmenopausal women between toenails and fingernails. Calcium levels were significantly lower in the fingernails ($p= 0.031$) of postmenopausal women, but not in their toenails ($p = 0.6574$).

Observational studies

Geyer *et al.* identified a number of clinical factors that have an effect on the growth rates of both nail and bone, these are listed in table 1.15 (Geyer *et al.*, 2004).

Table 1.15: Clinical factors which influence the growth rates of nail and bone

Faster nail growth	Slower nail growth
Male sex	Female sex
Youth	Ageing
Calcium medication	Smoking
Hyperthyroidism	Hypothyroidism
Vitamin D medication	
Oral contraceptives	

Based on the assumption that faster nail growth represents a protective effect and slower nail growth a negative effect, the link between growth and osteoporosis looks interesting. Slower-growing nail is also known to become more brittle (Galimont-Collen *et al.*, 2007). Smoking, ageing and being female are known risk factors for fracture risk (Kanis *et al.*, 2007). Youth, calcium consumption and oral contraceptives are all known protective factors.

Mechanism of action

Research to date has not yet confirmed the underlying mechanism of action linking bone and nail. Underlying any potential link are relationships between three factors:

Nail protein structure – bone collagen structure – bone fragility

The link between bone collagen and fragility is well established in the literature, while the link between nail proteins and fracture is still under investigation. A link between nail and collagen is currently unknown.

Towler *et al.* have hypothesised that the link is due to a relationship between cysteine or sulphur and bone and between bone strength and nail strength. They hypothesised that a relationship exists between bone and nail because both create disulphide bonds and utilise cysteine absorption processes. Disulphide bonding plays a role in the integrity of structural proteins in both keratin (nail) and collagen (bone). The study did not prove this relationship because it explored the relationship between keratin structure and BMD. The relationship between nail structure, bone fracture and BMD can therefore be considered indirect and the use of measures of nail structure to identify women at risk of osteoporosis is in effect using the nail as a surrogate marker (Towler *et al.*, 2007).

Independent corroboration of the link between nail and bone has been reported by Pandey *et al.* (Pandey *et al.*, 2008). They suggest that the nail samples of osteoporotic patients are subject to greater mechanical creep than the nail samples of healthy subjects.

Ogitani *et al.* proposed that the periosteum of the phalangeal bone is the conduit for the biological processes linking nail and bone. The periosteum is a layer of connective tissue which overlays all bones and lies between bone and the skin. They estimate that nail shares the blood flow of bone for up to five months while the nail is under the skin in the nail bed. They also note some contradictory results in the relationship between mineral concentration in nail and BMD, but attribute these to the use of different measurement techniques. They

also attribute the difference between toenail and fingernail results to differences in contact between the nail bed and blood circulation in the toenail and fingernail. They state that the level of contact between the toenail and the periosteum is lower, resulting in lower correlations between the two factors (Ohgitani *et al.*, 2005).

Clinical work conducted by Cummins as part of her PhD thesis has established that diagnostic measures using the peak heights of SS, CH₂, Amide I were able to differentiate between cases and controls in the FRAN study. (Cummins, 2009); This indicates that many features in the nail could be used to assess fracture risk.

The different hypotheses proposed to date need further investigation. The leading candidates are:

1. The periosteum channel linking bone formation to nail formation.
2. Cysteine processes linking the two matrices.

Pre-clinical studies that can evaluate the bone matrix and nail matrix directly in a number of different scenarios may be the best way to conduct studies and explore potential mechanisms linking keratin and bone proteins. Information from the key features in the Raman spectra of nail linked to fracture may also generate additional pathways for a mechanism of action.

1.9.3 Aspects of the human nail for diagnostic applications

Nail structure

Nail consists of fibrous composite keratin in three layers (Farren *et al.*, 2004). The overall thickness of the human nail has been estimated as 202 microns (Ohmi *et al.*, 2000). Keratin's fracture toughness is derived predominately from the disulphide bonds present in cysteine amino acids (Akhtar and Edwards, 1997). Using scanning electron microscopy, Farren *et al.* identified dorsal lower and ventral upper layers consisting of flat overlapping slate-like sheets, lined up in the same plane as the nail, and an intermediate layer which contains more keratin fibres than the upper and lower layers and which are oriented transversely, parallel to the free end of the nail. Figure 1.9 shows the different layers in nail. The intermediate layer is substantially stronger than the dorsal or ventral layers in the longitudinal direction and seems

to provide the structural integrity to the nail, owing to ‘long narrow cells’ in this layer (Farren *et al.*, 2004).

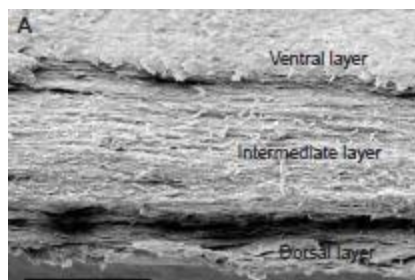


Figure 1.9: Image showing the ventral, intermediate and dorsal layers in a nail sample, from Farren *et al.*, 2004

Williams *et al.* were the first to evaluate human nail using Raman techniques, linking specific chemical bonds in the protein structures to Raman wavenumber shifts (Williams *et al.*, 1994). Gniadecka *et al.* identified the important vibrational modes in nail, these are shown in table 1.16 (Gniadecka *et al.*, 1998).

Table 1.16: The major Raman vibrational modes in the human fingernail

Vibrational mode	Peak position
V(C=O) Amide I	1649 cm ⁻¹
Amide III	1273 cm ⁻¹
CH ₃	2993 cm ⁻¹
CH ₂ (lipids) and CH ₃ (proteins):	1447 cm ⁻¹
Acyl backbone chain	1030 cm ⁻¹
V(S-S)	511 cm ⁻¹

Analytic Procedure

The nail is analysed by clipping a 3mm wide sample from the fingernail or toenail. The nail is then placed on a silicon slide which has a groove to accept the nail sample and a second slide which has a 2mm hole in it to allow the laser to reach the nail without passing through the

slide is placed on top. The two slides are clamped together to flatten the nail. The slide combination is then placed in the instrument for analysis.

Nail stability

To conduct longitudinal trials using nail samples that meet the evidence requirements for clinical adoption, there is a requirement to either collect nail samples at baseline and follow the subjects for 10 years or more, or to use archived nail samples. Archived nail samples collected in studies with adequate follow-up procedures provide a more timely and cost-effective research route. There is, therefore, a requirement to demonstrate from the literature that archived nails retain their structural integrity after 10 years to allow them to be used in cohort trials.

An important aspect of assessing the suitability of nail samples as diagnostic samples is an understanding of the factors that affect the structural integrity of nail over extended storage periods. Edwards *et al.* demonstrated, in a comparison of 500-year-old mummified baby nails and the nails of a modern baby, little difference in the Raman signatures (Edwards *et al.*, 2002). He showed, however, that freezing, bacterial attack and desiccation processes were harmful to keratin-bearing substances. It has been shown that the majority of the water in keratin is in the bound form and therefore processes such as freezing and drying can be harmful to the overall integrity of the nail (Gniadecka *et al.*, 1998). These studies indicate that keeping the nail dry and at a consistent temperature will support its preservation over potentially hundreds of years.

Nail stability can vary in two ways:

1. On a month by month basis, *in vivo*. making long-term extrapolations from the data invalid.
2. Based on the age on the nail, when tested.

There is currently no evidence in the literature describing the changes of nail structure over time in healthy individuals or long-term stability studies on nails collected at a particular point in time using Raman spectroscopy. The work by Edwards and Gniadecka indicates that samples kept dry will remain intact over long periods, but this has not been definitively demonstrated by any research work found.

The differences between fingernail and toenail

Some work using Raman spectroscopy has established that fingernail and toenails are structurally different (Widjaja and Seah, 2006). The mean spectra between the two types of nails were found to be subtly different at wavenumbers 511, 1003, 1062, 1270, 1340, 1449 and 1460 cm^{-1} respectively. Interestingly, some of these correspond to the structurally important S-S bond. Human fingernails are known to grow three times more slowly than human toenails (BEAN, 1953), which supports Widjaja's findings of differences between the groups.

Biomedical applications of the human nail

In 1997, Lawson *et al.* considered the biomedical applications of the human nail from Raman examination and felt that fingernails could potentially provide diagnostic information on metabolic disorders, drug consumption, and infection (Lawson *et al.*, 1997). Nail samples in the Nurses' Health Study, a Harvard University based cohort study, have been successfully used to evaluate nicotine use in smokers as a predictor of coronary heart disease (Al-Delaimy *et al.*, 2008). There are currently however, no commercial diagnostic applications exploiting the relationship between Raman spectroscopy and the human nail.

Slotnick and Nriagu have considered the validity of using human nails as a biomarker of arsenic and selenium exposure and concluded that, since nail can reflect long-term exposure, it is a more suitable biomarker for long-term exposure than blood or urine (Slotnick and Nriagu, 2006). Jablecka *et al.* have recently demonstrated differences in the dielectric properties between the nails in healthy and diabetic subjects (Jablecka *et al.*, 2009).

1.9.4 Raman spectroscopy

The Raman effect

The Raman effect is the common name for a physical phenomenon first identified by C.V. Raman in 1928, and Raman scattering as it is now known involves the inelastic scattering of light which is incident on a material (Raman and Krishnan, 1928). Raman scattering is only a very small component of the reflected light – typically, only one photon of light for every 10^7 photons delivered will be scattered inelastically (Petry *et al.*, 2003). The rest are scattered

elastically, also known as Rayleigh scattering, which results in the light being reflected at the same wavelength at which it encountered the molecule. Figure 1.10 shows the various descriptions for the different scattering effects that happen in the Raman effect.

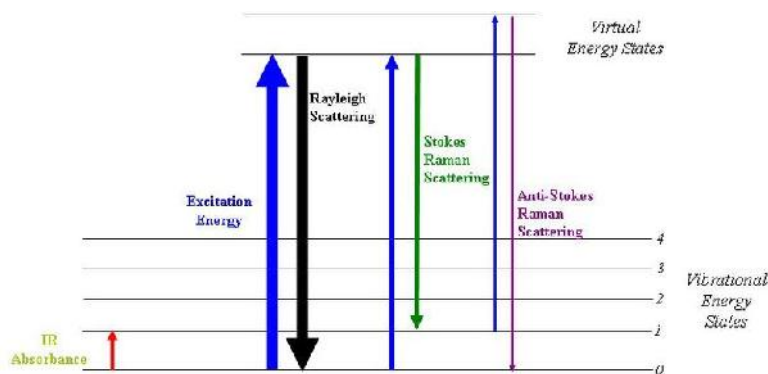


Figure 1.10: The descriptions for the changes in energy levels due to the Raman effect

The change in state in the photon is caused by the molecule with which the photon is undergoing a change in its vibrational or rotational mode, which results in a decrease or increase in energy in the material and the photon. This change in energy results in a change in wavelength in the photon and is known as Raman shift.

Raman spectroscopy

Raman spectroscopy is an analytical technique used in a wide range of research fields to analyse the structure of materials. The technique exploits the Raman effect in a measurable manner. The analytical use of the Raman effect required the invention of the laser, which acts as a strong source of monochromatic light. When used in conjunction with sensitive filtering and detection equipment to separate the inelastically and elastically scattered signals, the Raman signal was reliably isolated. In recent years, Raman spectroscopy has become an increasingly applied research technique and relatively low cost instruments are available from a number of commercial suppliers. The technique is attractive in a number of fields because of the minimal sample preparation required before analysis.

Confocal Raman spectroscopy

As discussed earlier, the human nail contains a number of layers and so there is a requirement to control the depth at which the readings of vibrational modes are taken to ensure information from the correct layer is collected. Confocal microscopy offers a well-validated technique to carry out this type of investigation. Conventional microscopy receives reflections from throughout the material of interest. By using a pinhole, the light observed is constrained to the region of interest, established by the operator manually focusing the laser on the particular region using the microscope objective (Petry *et al.*, 2003). Figure 1.11 shows a schematic for a confocal micro-Raman spectroscopy instrument.

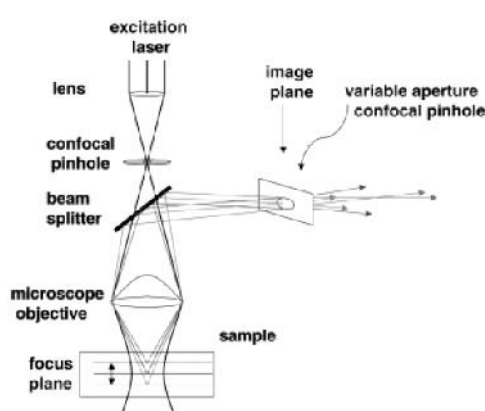


Figure 1.11: A schematic of a confocal micro-Raman spectroscopy instrument.
(reproduced from Petry *et al.* 2003)

Of particular interest has been the work of Caspers *et al.* in establishing the use of Raman spectroscopy to examine the different layers of skin, a similar material to nail (Caspers *et al.*, 2003; Caspers *et al.*, 2001). Skin has two important layers: the top layer, which is 40 μm thick, and the lower layer, which has a thickness of between 1 mm and 4 mm. Caspers *et al.* reported that axial resolutions of around 5 μm had been achieved using confocal Raman spectroscopy. This level of resolution enables the different layers of the human nail to be isolated.

Raman spectroscopy and biomedical applications

Raman spectroscopy is considered a tool with potential in a range of biomedical diagnostic applications (Lawson *et al.*, 1997; Petry *et al.*, 2003).

Its adoption has been limited by high fluorescence from biological samples, which tends to overwhelm the small Raman signal and reduce its diagnostic potential (Petry *et al.*, 2003). This has been less of an issue in recent years as the sensitivity of instruments has improved, allowing the Raman signal to be more reliably separated from fluorescent signals.

Raman spectroscopy to evaluate osteoporosis

Raman spectroscopy is a well-established technique for *in vitro* analysis of bone samples. Work done by Boskey and Paschalis confirms the validity of the technique to evaluate collagen cross-linking in bone (Paschalis *et al.*, 2004; Boskey *et al.*, 2009).

For *in vivo* applications, Raman spectroscopy has historically suffered from the restriction of limited penetration into the skin, which limits its direct applications in osteoporosis due to an inability to penetrate the skin layers to reach the bone. The strong fluorescence and Raman signals at the surface mask the deeper located weaker signals with the diagnostic information (Stone *et al.*, 2007). In recent years a new technique, spatially offset Raman spectroscopy (SORS), has been developed by Matousek *et al.* (Matousek *et al.*, 2006; Draper *et al.*, 2005). Its potential as a tool for analysing bone *in vivo* for fracture risk prediction is clear, as the investigators have obtained sample data with adequate resolution at depths of between 2 mm and 4 mm. However, the trials to date have been retrospective and pre-clinical so the data should be seen as preliminary. The depth at which adequate resolution has been achieved is continually increasing, with the most recent results from a depth of 8.7 mm for the assessment of breast cancer tissue (Matousek and Stone, 2009). These studies, while exciting, provide limited insight into the clinical potential of the SORS technique to predict future fracture. The type of longitudinal study required to incorporate SORS-based prognostic tests into tools such as FRAX[®] and QFractureScores would require a significant investment of time and money; potentially, a 10-year study would be required in the absence of sufficient archived bone samples. The potential, therefore, will be restricted to taking readings at the ankle and wrist for the foreseeable future.

1.9.5 Classification techniques for Raman spectra

Introduction

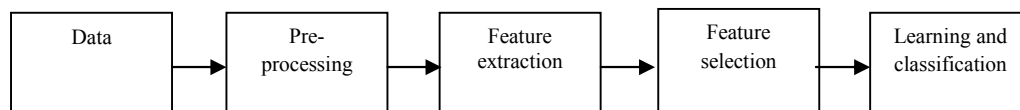


Figure 1.12: Steps in the statistical classification of Raman data

To develop an algorithm that can effectively classify Raman spectra data, a number of analytical steps are required, figure 1.12 outlines the steps required to classify the data.

Statistical classification techniques use a training set of data to learn the key relationships in the data and then create a set of rules that can be used to classify the data in an independent test set. The training set of data is used for the algorithm to learn the key features in the data and so create a classification system to separate the groups. This process is also known as machine learning.

Simple direct measurement of peak height and width measurement in the spectra are sometimes not able to deliver the required discrimination between the groups; for higher levels of discrimination, it is sometimes necessary to use a multivariate technique that considers the whole spectrum. Multivariate statistical classification techniques are able to extract appropriate information, even when it is widely distributed throughout a spectrum.

Feature extraction

Principal component analysis (PCA) is a well-established technique for the reduction of biological data. The technique examines an entire spectrum and resolves the important features into a series of principal components. This step reduces the amount of data and makes further analysis by other techniques more manageable. The principal components contain information about the relationship between the spectra and the event being modelled in the form of a score and a loading for each component. The first components identified are more important in the prediction model than later components; this is because they contain more information linking the spectral variation of interest to the prognosis than lower order

components. Typically, most applications will only use the first few principal components to avoid over-fitting of the model to the data.

The technique is often implemented using mathematical software programs such as Matlab (MathWorks, Massachusetts, USA). While the technique has proved useful, it has the limitation that the relationships between the individual chemical species and the algorithm cannot be determined (Shinzawa *et al.*, 2009). Another limitation of PCA is that, although it is a linear analytical tool that is simple to use, it may underperform with biological data, which is often non-linear. PCA has had success in the field of bioinformatics and microarray analysis (Wang and Zheng, 2008) and is a well-established platform on which to carry out feature extraction. Figure 1.13 illustrates the change that takes place in the transition from data space to feature space.

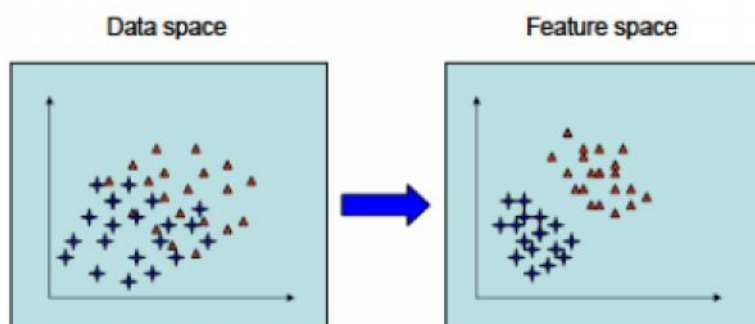


Figure 1.13: The separation of the data using techniques such as PCA

Learning and classification

Once data reduction has been completed using PCA, linear discriminant analysis (LDA) is used to separate the features identified by PCA. This technique draws a straight line through different regions of the data at the best angle it can to differentiate between the two classes of data, such as high risk and low risk scores. While this technique is well established, the use of linear methods limits its performance with complex data sets.

Validation

All the techniques are validated in the first instance within their own dataset using leave-one-out procedures. This technique trains the model by leaving out one of the samples and evaluating the performance of the model on the remaining samples. The process is repeated with a different sample left out each time until all the samples have been tested. The resulting cumulative error provides an assessment of the robustness of the algorithm.

The algorithm is then validated in a set of data independent from the dataset used to develop the model, which is the most robust test for a new algorithm.

1.10 Clinical trial design to assess prognostic test performance

1.10.1 Randomised controlled trials

The best type of trial to establish causality is the randomised controlled trial (RCT) (Moon and Gould, 2000). This trial design follows both a control group and an intervention group to establish whether the intervention (in this case, diagnostic testing and selection for treatment) results in superior outcomes compared with those in the control group. The benefit of RCTs is that the randomisation process eliminates any potential confounders. The main issues are costs and scale to ensure sufficient randomisation. As a result, RCTs are rarely undertaken and it is unlikely the costs could be justified to validate a single new risk factor for osteoporosis.

1.10.2 Observational studies

Another type of trial, which investigators consider less robust but which is far more common, is the observational or cohort study. In this design, all the subjects at the start of the study are free from the condition and are observed over a period to establish who develops the condition and why. This kind of study can be either prospective, where the subjects are followed from baseline forward, or retrospective, where data collected historically is used to compare with outcomes at the present time. These studies can require hundreds of thousands of people to achieve sufficient sample size if rare diseases are being investigated. There are long-running cohort studies aiming to establish causality for a number of chronic diseases,

such as the Nurses' Health Study in the US, the Framingham Heart Study, also in the US, and the Rotterdam Study (Karasik *et al.*, 2002; Feskanich *et al.*, 2002b; Schuit *et al.*, 2004).

Any cohort study must be externally valid – for example, are the results of a study carried out on nurses valid for policy decisions in the wider population? For the study to be internally valid, three factors must be controlled:

1. Bias (systematic error): A common form is selection bias, where people with a particular condition respond to a survey about that condition at a higher rate than the average person in the population.
2. Confounders: This is a problem where a hypothesis assuming cause and effect between two variables is interfered with by another variable. The confounder masks the relationship between the two variables of interest or strengthens it, making the relationship seem stronger than it is in reality.
3. Random error: This error arises by pure coincidence. Analytical errors arising from randomness can be controlled for by maximising the sample size and the duration of the study.

1.10.3 Survival analysis

When conducting a trial to assess the risk of an event over a predetermined period, survival analysis is a well-established technique to compare the effectiveness of different interventions. Osteoporotic fracture trials are well suited to survival analysis because the aim is to evaluate the time taken for a particular event (in this case, fracture) to occur. The technique then enables the differences in time to event occurrence between the two groups over the set period to be evaluated. For osteoporotic fracture prevention, treatment with an effective drug should result in a longer time to fracture in the treatment group compared with the control group. Alternatively, women identified as being at higher risk using a prognostic test should have shorter times to the event when compared with women identified as having a lower risk.

Standard statistical techniques such as ANOVA (analysis of variation between groups) are not effective for this analysis due to the non-normal distribution of these kinds of events (Clark *et al.*, 2003). Censoring (where the value of an observation is only partially known) is also an issue, supporting the use of survival analysis in situations where not all subjects will

have an event by the end of the study. For example, in the case of a 10-year period for hip fracture analysis, the outcome of the patient (death or fracture) is not known in all cases. In this situation, survival analysis is an appropriate technique as it assumes that subjects no longer in the study will have the same survival rates as those who remain in the study (Katz, 2001a).

The two key outputs of any survival analysis are survival probability $S(t)$, i.e. how likely is a person to avoid the event in the specified time frame, and the hazard rate $h(t)$, i.e. what is the likelihood of that event happening to a person at any given time instant.

A number of techniques are available to estimate the survival rate from observed data. The most commonly used are Kaplan-Meier (Kaplan and Meier, 1958) and Cox's proportional hazards (Cox, 1972).

Using the Kaplan-Meier method, a non-parametric estimate can be generated, which can be shown graphically as a Kaplan-Meier survival curve, figure 1.14.

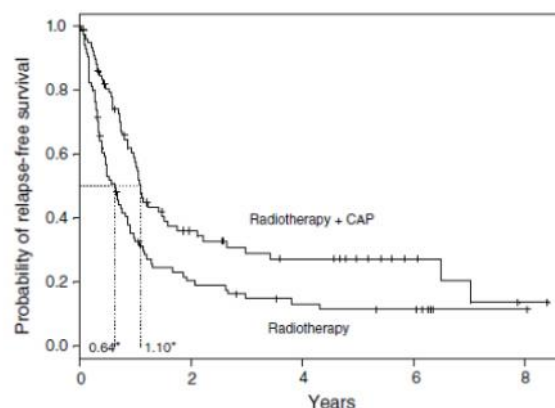


Figure 1.14: Relapse-free Kaplan-Meier survival curves for a lung cancer trial from (Clark *et al.*, 2003)

An issue with Kaplan-Meier and log rank analysis is that they are univariate analyses and only suitable for use with two factors. Where more than two variables need to be included, a multivariate analysis technique needs to be used, commonly known as a regression analysis (Bradburn *et al.*, 2003). These techniques create an equation that includes weightings for multiple variables and thereby incorporates all the relevant factors in the final analysis. Cox's proportional hazards, a semi-parametric method, is the most widely used multivariate technique. A key assumption of this model is that there is a linear change between the factors

influencing the outcome and the logarithm of the relative hazard. For events where the prevalence of the incident is lower than 5%, proportional hazard analysis is not appropriate, therefore Poisson regression models should be used (Katz, 2001b).

1.10.4 Absolute risks, relative risks and hazard ratios

Cohort studies allow absolute risks to be calculated, the risk of an individual having that condition in the general population (Scott, 2008). In osteoporosis, under the new WHO guidelines developed by Kanis, these absolute risks are expressed as a percentage probability over a 10-year period (Kanis, 2007). This timeframe has been selected for clinical and health economic reasons and, in other medical conditions, different periods might be used. A risk could be 1% or 10%; its relative importance will depend on the seriousness of the ultimate health outcomes and the prevalence of the condition in the population.

The hazard ratio is considered similar to relative risk but is specifically used in survival analyses. Like relative risk ratios, the measure reflects the difference between individuals with a positive result and those with a negative result, but with the introduction of a time-based factor. It is, however, only used when a particular period is being considered, as in the case of fracture risk over a 10-year period.

Absolute risks cannot be calculated using case-control studies due to the lack of information of the prevalence of the disease in the population in a case-control study. This statistical issue has a significant impact on the clinical evidence requirements for new prognostic markers seeking to be integrated with FRAX[®] and similar tools in the future. Absolute risks can, however, be estimated from cohort studies using approaches developed by Langholz *et al.* (Langholz and Borgan, 1997).

1.11 The economic evaluation of osteoporosis screening

1.11.1 Introduction

In the modern healthcare environment, there is an increasing emphasis on ensuring that interventions carried out can be justified on both clinical utility and health economic grounds.

The most commonly used evaluation tools are cost-utility analysis (CUA) and cost-effectiveness analysis (CEA). CEA enables the relative merits of different interventions for the same condition to be evaluated. The comparative measure selected is entirely dependent on the condition and intervention of interest and is measured in financial terms as a cost per event. An example would be cost per case of breast cancer detected or prevented.

Health economists use CUA to evaluate an intervention's impact on the subject's quality of life. A number of validated tools are available to evaluate quality of life measures (Lips and van Schoor, 2005). The advantage of this type of evaluation is that it allows health economists to compare the intervention with alternatives for a range of different health conditions at a societal level. The most common measure used in these studies is the quality adjusted life year (QALY) (Drummond, 2005). Typically, an intervention would be justified when the intervention cost per QALY is lower than the QALY benefit. The disability adjusted life year (DALY) has been introduced as an alternative to QALY; this is because QALY has been applied to specific populations, while DALY uses a standard set of measures that can be applied globally across disease categories to support WHO decision making processes. The DALY measures loss of functionality caused by disease, whereas the QALY measures quality of life at various health states (Sassi, 2006). Since most osteoporosis studies focus on comparing between choices in osteoporosis rather than the relative merits of treating osteoporosis rather than other conditions, QALY is used as the most common cost-utility measure in the literature. A key-related outcome consideration is the incremental cost-effectiveness ratio (ICER). This measure reflects the elasticity of change in cost to the outcome measure, either QALY or fractures prevented, depending on the analysis. The strategy with the lowest ICER is the most effective overall.

1.11.2 Health economics definitions in osteoporosis

In osteoporosis-focused health economic studies, the cost per fracture prevented has been the focus of cost-effectiveness studies, with DALY or QALY the focus for cost-utility studies.

In some earlier studies, the measure of cost per case of osteoporosis identified has been used as the comparison measure (Schousboe, 2008). However, this measure is defined by T-score and therefore does not have a direct relationship to a health outcome. It is, therefore, a poorer measure to use than cost per fracture prevented which can be linked to CUA.

The QALY impact of an intervention to prevent osteoporotic fracture can be evaluated by assigning a quality of life impact to different types of fractures, then assigning established figures for a life year and finally incorporating the impact of the intervention in terms of number of fractures prevented. A successful intervention would typically be justified when the intervention cost per QALY is lower than the QALY benefit from the reduced number of fractures resulting from the intervention. It has been reported that ICERs per QALY of less than \$30,000 justify intervention (Schousboe, 2008).

1.11.3 The economics of osteoporosis screening

Health economic studies in the field of osteoporosis have focused on the following areas:

- The economic burden of the disease to society
- The impact of the disease on quality of life
- Cost-effective treatment thresholds for the diseases
- The relative merits of different interventions.

Papers referring to the societal economic burden of the disease and its impact on quality of life have been cited in section 1.4.2. Tosteson has written on the impact of treatment thresholds on the health economics of the disease (Tosteson *et al.*, 2008). In these papers, she indicates that treatment for osteoporosis becomes effective at a threshold 10-year risk of fracture of 3% assuming a cost effectiveness threshold of \$60,000 per QALY gained. This US-focused model assumed a \$600 per year drug cost applied over a 5-year time period and a 35% reduction in fracture risk from taking treatment. In the light of the move to a 10-year

risk of fracture, new health economic models are required to assess the potential impact of the alternative screening methodologies.

A large number of the early published health economic studies used HRT as the preventative treatment. While this was an effective solution for osteoporosis, its side effects now prevent its use and, therefore, the results presented have limited use today. Meta-analysis has also shown that treatment costs and not screening costs are generally evaluated in these studies (Schousboe, 2008). The researchers found that only ten studies were conducted solely using DXA for case-finding patients for treatment, most studies relied only on the judgement of the physician. For a mass screening protocol to be justified, the case for more reliance on the screening tools available and a strong economic case to justify using those tools needs to be made. Another limitation of these studies is that most use hypothetical cohorts rather than specifically designed RCTs or actual cohorts.

1.11.4 Developing a health economic model

In order to develop an acceptable health economic model in the osteoporosis field, a number of assumptions need to be made. Each aspect and the types of assumptions made are described in this section.

Preventative strategy

In order to justify developing a model, a preventative strategy needs to be devised to drive the comparison of alternatives; for example, “comparing non-intervention against screen & treat using DXA in women between 65 and 80 over a twenty year time period” (Schwenkglens and Lippuner, 2007). These two options govern the scenarios being compared in the study.

The modelling framework

Health economists typically use decision trees or Markov models to evaluate health economic choices. Decision trees assign a probability to each outcome, whereas Markov models are able to consider the timeframe in which an event takes place (Drummond, 2005). This is useful when long periods can be involved before events occur. The Markov state transition model has been the preferred methodology to evaluate the relative merits of interventions in osteoporosis (Schwenkglens and Lippuner, 2007). One of its advantages is that transitions from one state to another depend on the current state and are not affected by previous states.

A complex state transition model in osteoporosis might include the following states (Tosteson and Weinstein, 1991):

- Well and living in the community
- Hip fracture
- Nursing home
- Disabled and in nursing home
- Breast cancer
- Ischaemic heart disease
- Dead.

A simple model might only include:

- Well and living in the community
- Hip fracture
- Dead.

In a Markov states transition model, a subject will be assumed to start in one state and move through other health states, ultimately ending in death. The model is run as a computer simulation allowing a number of different scenarios to be evaluated in software packages, such as TreeAge (Treeage Software, Massachusetts, USA).

In one paper evaluating the cost-effectiveness of DXA screening in a Japanese population (Nagata-Kobayashi *et al.*, 2002), five Markov states were selected:

- 1) In complete health.
- 2) In an acute state of hip fracture.
- 3) Able to walk outside after hip fracture.
- 4) Unable to walk outside after hip fracture.
- 5) Dead.

It is interesting to note the confusion in the literature over the use of terms. In this publication, a CUA is carried out but it is described as a CEA. This is a common issue in the literature around health economics.

Incidence levels

The next step in the model development will be to establish an incidence rate in the population of interest. A number of peer-reviewed studies and official publications like the WHO Osteoporosis Guidelines can be used to derive this information (Kanis, 2007). Burge *et al.* have carried out the most recent work on this in the US population (Burge *et al.*, 2007). They estimate that the incidence of hip fracture in the 50-64 age group of the female population in 2005 was 13,420, in the 65-74 age group around 25,288 and, in the 75-84 age group, 84,274 fractures per year.

Relative risks

The relative risk identified for an individual using a particular prognostic technology is an important input factor into the model. This information can come from meta-analyses on a particular prognostic tool or from original research into the new test. It has been well established by meta-analysis that the risk of hip-fracture for women with a diagnosis using BMD is 2.6 per SD change in BMD (Marshall *et al.*, 1996). Most studies used to generate this figure were in women over 65 years of age and the follow-up periods were under 5 years. This compared with a figure of 1.6 per SD change in BMD observed in the Aberdeen Prospective Osteoporosis Screening Study (APOSS) study, conducted in perimenopausal women over a 9-year time frame, a much lower value indicating that BMD predictive performance is lower in younger women (Stewart *et al.*, 2006).

Therapeutic effects

Once a clinician identifies a woman as being at high risk of osteoporotic fracture, a number of interventions can be offered to reduce that risk. The early studies in the field evaluated the use of HRT (Nagata-Kobayashi *et al.*, 2002). However, with the move to bisphosphonates as the dominant prescribing class, the drug commonly used in economic evaluations is alendronate. New drugs coming onto the market like Prolia (denosumab) will result in a new round of health economic evaluations due to its superior anti-fracture performance (Cummings SR *et al.*, 2008).

Key factors considered in assessing a drug are:

- Cost
- Patient Compliance
- Effectiveness in reducing fracture
- Required duration of intervention.

For example, alendronate now costs less than £300 a year, is considered to have a compliance rate of 50%, is considered to be effective over a 5-year timeframe and to reduce the incidence of hip fracture by 30% (Sculpher *et al.*, 1999).

Table 1.17 summarises a number of treatment options for fracture prevention and their associated fracture preventative effect. (Adapted from the study by Sculpher *et al*):

Table 1.17: The annual treatment costs and fracture prevention effectiveness of osteoporosis treatments in the UK

Treatment	Hip Fracture effect	Annual Treatment cost
HRT	30-50%	£23 - £168 pa
Bisphosphonates	34%	£160 pa
Calcium + Vitamin D	30%	£120 pa
Calcitonin	24%	Not available in UK

Direct and indirect costs of fracture

A number of studies have been carried out which act as benchmarks for quantifying the direct and indirect costs resulting from a hip fracture (Burge *et al.*, 2007; Gabriel *et al.*, 2002; Gabriel *et al.*, 1999). The Gabriel study in 2002 was of particular interest in terms of the short-term impacts of fracture. They used a prospective study approach to determine incremental costs of fractures over the normal annual costs of managing an age-matched individual using data from a Minnesota database. Using a case-control approach, they were able to demonstrate that, for 1,263 case/control pairs, with a median age of 73.8 in the year prior to fracture, the costs allocated in the local medical system were \$761 and \$625 respectively for cases and controls. In the year immediately following a fracture, the costs

were \$3,884 and \$712 respectively for all fractures. For hip fractures, the incremental costs were \$11,756 and had a range of between \$15,579 and \$5,150 in the 25th and 75th percentiles. These figures are extremely accurate because the local medical system allowed accurate costs of procedures to be identified using the Mayo Cost Data Warehouse. The level of transparency of costs in the US is much higher than in many parts of the world due to the economic model used to manage healthcare. The study was advanced because it demonstrated that incremental fracture costs were lower than originally believed, due to the elimination of background costs associated with being older in general. The authors noted that previous studies had overestimated first-year costs for a hip fracture at between \$26,400 and \$36,800, but had underestimated the cost of lower impact fractures.

Researchers have also investigated the longer term, on-going costs of fracture. The Burge *et al.* paper considered the future economic burden of osteoporosis-related fractures in the US (Burge *et al.*, 2007), by using figures reported for long-term care by Braithwaite *et al.* (Braithwaite *et al.*, 2003). Braithwaite estimated the lifetime cost of hip fractures at \$81,300. They also estimated that 1.8 life years are lost as a result of a hip fracture. This figure included costs for the fracture, later hospital visits, rehabilitation support, home care and nursing home care.

There is some confusion over whether non-osteoporotic fracture-derived costs should be included in the cost models. Most studies limit their analysis to osteoporosis fracture-based costs, such as hospital stays, because these events are more readily quantified. Most studies have not considered the costs associated with the process of screening, but rather focus on the benefits of a particular treatment vs. placebo according to a review study conducted at York University (Sculpher *et al.*, 1999). This approach has generally been taken because the focus has been on the relative merits of various treatments and their effectiveness rather than a complete disease management approach for the condition.

1.11.5 Papers on screen and treat strategies

A number of peer-reviewed studies have been conducted to evaluate the cost-effectiveness of screening women for osteoporosis to prevent fractures in combination with particular treatments. It has been observed that, between 2000 and 2008, seven papers were published

which assessed so-called “screen and treat” strategies (Schousboe, 2008). An overview of activity in the field is provided in this section.

Brown *et al.* evaluated a screen and treat strategy using the Royal College of Physician guidelines (Brown *et al.*, 2001). DXA sensitivity and specificity were set at 50% and 80% respectively. The adherence rate to HRT treatment was set at 70% and the rate was set at 94% for alendronate in year one. The study showed that the costs of treating all women to prevent fracture were higher than selecting women based on BMD score, at £32,594 and £23,867 respectively when clinicians used HRT, and £171,067 and £14,067 when alendronate was used. They concluded that, the more expensive the preventative treatment, the more important a screening strategy becomes for overall cost-effectiveness. the impact on the costs of a screening strategy using alendronate, the more expensive drug can clearly be seen, the costs drop dramatically when a screen and treat strategy is introduced. The study provides compelling evidence to support the use of screening for high cost drugs, but it could be argued today that the rates of compliance and DXA performance data used in the paper are too high for the study to be realistic. Other limitations of the paper are that it is completely hypothetical and draws its input data from a wide range of sources rather than being experimentally based.

In 2002, the National Screening Committee issued a report entitled ‘Osteoporosis and criteria for screening’. In this report, they concluded that there was a need for mass screening, but that DXA with its use of radiation and high cost would not be a suitable testing modality. They hoped that ultrasound would improve sufficiently to fulfil the need (National Screening Committee, 2002).

A paper by Sim *et al.* investigating QUS as a pre-screen for DXA in a UK population established that QUS was too expensive to act as an effective pre-screen at current costs. They used cost per case of osteoporosis identified as their comparative effectiveness measure (Sim *et al.*, 2005). This paper highlights the practical clinical limitation of using highly correlated technologies of similar cost as a pre-screen to DXA.

In 2005, Schousboe *et al.* wrote the first mass screen and treat strategy evaluation to include diagnostic costs for osteoporosis. This study was designed to evaluate the cost-effectiveness

of universal bone densitometry in women aged 65 and over diagnosed using DXA. A Markov model with eight health states was used and the intervention drug was alendronate. This hypothetical model used incidence rates from the US-based Rochester cohort. The cost per QALY gained was \$43,000 for women aged 65, but this figure dropped to \$5,600 for women aged 75 (Schousboe *et al.*, 2005) over the patient's lifetime. Schousboe has also evaluated the potential of bone turnover markers to cost-effectively reduce fractures, noting that half of all fractures occur in post-menopausal women with no history of fracture. Women over the age of 70 with high bone turnover and with osteopenia were selected for treatment. High bone turnover was quantified as top quartile for the whole population by the investigators. The costs per QALY gained were \$34,000 and \$50,000 for women with T-scores below -2.0 and -1.5 respectively, assuming effective treatment (Schousboe *et al.*, 2007). A flaw with the model developed was that it is highly sensitive to the relative risk chosen for the bone turnover markers. An RR of 3.0 was selected; if the RR were reduced to 1.25 the cost per QALY gained would increase to over \$89,000 making the approach unattractive. An RR of 3.0 is on the high side based on the literature; Garnero *et al.* reported a RR of 2.2 (1.3-3.6) and 1.9 (1.1-3.2) for CTX and free D-Pyr respectively (Garnero *et al.*, 1996a). It was interesting to note that the sensitivity analysis showed the economic viability to be relatively insensitive to the cost of the diagnostic test in a range of between \$28 and \$112.

A paper written on a hypothetical Thai patient population between the ages of 45 and 55 to evaluate a number of competing screening techniques, including universal testing, non-intervention, clinical risk factors, DXA and QUS, offers useful insights (Panichkul *et al.*, 2006). The paper reaches the conclusion that non-intervention is the best approach due to the low reduction in fracture rate in women screened between these ages over a 5-year follow-up time frame. The best performing screening strategy was screening by risk factors in combination with DXA at \$60.33 (US dollars) per fracture prevented, while the worst was universal treatment with the various combinations of QUS, DXA and risk factors lying in between. This conclusion is a missed opportunity, as they may have been able to demonstrate the superiority of a screening strategy over non-intervention over a longer follow-up period due to the higher number of fractures that would occur. The paper has a number of other limitations including the use of a hypothetical cohort, the use of HRT as the preventative treatment, and the relatively high cost of the intervention compared to the lower direct medical costs in Thailand.

Schwenkglenks and Lippuner evaluated a DXA-based screen and treat strategy against no screening in a Swiss group. They developed a hypothetical-based cohort using Markov analysis with four states: alive without fracture, alive with a hip fracture, alive with a peripheral fracture and dead. Swiss data sources were used to obtain the other information required. RRs of 2.7 were used for women identified as osteoporotic as per data found in the Rotterdam study. The investigators estimated the cost of each screening test to be 300 Swiss francs. They found that screening was cost-effective in women over the age of 70 based on an ICER threshold of 50,000 Swiss francs per QALY (Schwenkglenks and Lippuner, 2007).

A more recent, but similar, analysis to Schwenglenks was carried out in a paper by Mueller and Gandjour. This study evaluated the cost-effectiveness of using DXA and CRFs for osteoporosis screening in postmenopausal women compared to no screening in Germany (Mueller *et al.*, 2009). A hypothetical Markov model containing 10,000 women was developed. They considered DXA to have a sensitivity of 34% and a specificity of 90%. They state a sensitivity of 80% and specificity of 50% for CRFs. They note that the combined use of DXA and CRFs produces a sensitivity of 60% and a specificity of 74%. They separated women onto three decades, namely 60, 70 and 80 years of age. A single drug, alendronate, was used in the cost utility study based on its known cost-effectiveness. A 9-state Markov model was used, its complexity caused by the requirements for states for hip fracture, vertebral fracture and forearm fracture. In women between 60 and 70, no screening had the lowest ICER at €4,607. In women in their 70s, a strategy combining DXA and CRFs dominates a CRF-only strategy because it is more effective at lower overall cost; the ICER for CRFs alone was €6,641. In women in their 80s, due to the high risk of fracture, the treatment for all women scenario was cost-effective. Using a cost-effectiveness ratio threshold of €30,000, the authors conclude that using CRFs alone is not cost-effective as a screening platform for women over the age of 70. While their data indicates that using CRFs over DXA is superior in the under-70s, the authors do not commit to this conclusion because their data in this age range is based only on prior vertebral fractures and they had a limited number in the study. A limitation in the study was the attempt to assess the viability of screening women in their 60s over a 10-year fracture horizon which, due to the low incidence level of fractures, has small study numbers. A 20-year horizon could be better for women in their 60s due to the much higher fracture incidence over that time frame. This is a limitation

in what is otherwise a high quality study with comprehensive consideration of both screening and treatment costs. The authors also raise an interesting point in noting that, in certain age groups, women selected based on CRFs benefit from treatment. This is a weakness in a DXA scanning-focused strategy because pharmaceutical intervention has been shown to be effective in women not selected based on BMD. Ultimately, they do not recommend the use of CRFs alone for screening women for treatment, citing uncertainty in which are the best CRFs to use.

1.11.6 The next generation of studies

New studies are required

It has been observed by Schousboe that a new generation of clinical studies will be required due to the new WHO guidelines, which recommend a focus on absolute risk of fracture rather than BMD testing for case finding (Schousboe, 2008).

A study which converts a standardised QALY threshold for mass screening into a “cost per fracture prevented” value, which can be used to justify mass screening, is also required. Studies to date have focused on relative comparisons of screen and treat strategies, but have not established a viable model in the context of the new 10-year fracture risk regime that can be applied across studies.

Potential screen and treat strategies

The ultimate aim of any screen and treat strategy is to develop an approach that cost-effectively evaluates women for fracture risk at the peri-menopausal stage for their lifetime risk.

This review of the literature has shown that the following strategies are available:

1. Treat all women.
2. Treat all women with a positive DXA score.
3. Treat all women with a positive DXA score AND CRF risks.
4. Treat all women with a positive DXA score OR CRF risks.
5. Pre-screen women with CRFs and then use DXA to select.
6. Use CRFs to treat on their own.

Investigators have identified a number of screen and treat options depending on the performance of the non-BMD risk factors. The use of non-BMD risk factors is attractive because data to date indicates that they are not correlated with DXA and, therefore, can add value in combination using an AND or OR logic strategy.

Pre-screen approaches using CRFs or QUS (Panichkul *et al.*, 2006) have had limited success due to the high correlation between CRFs and QUS and DXA in terms of low sensitivity. The nail test, which appears to be uncorrelated with DXA, could add substantial information depending on the screening approach used (Towler *et al.*, 2007).

New outcome measures

One approach that could be taken to justify mass screening would be to investigate whether the cost of mass screening the population to prevent a hip fracture would be less than the lifetime cost of a fracture based on estimates created from literature searches. A starting value could be the \$81,300 calculated by Braithwaite *et al.* (Braithwaite *et al.*, 2003).

1.11.7 Conclusions

It has been established that mass screening using DXA is not justified (National Screening Committee, 2002). It has, however, been shown that using DXA or clinical risk factors, or both, to select women for expensive treatment is justified (Mueller *et al.*, 2009).

It is interesting to note that the studies reviewed show that the cost of screening has no bearing on the viability of a screening programme (Kanis, 2007). The real driver of whether

screening should be carried out is the performance of the test in combination with the performance of the drug. This is because the costs of treatment over a 10-year period are in order of magnitude larger than the cost of screening, and the financial impact of a single hip fracture is significant.

There is a lack of publications exploring the following from a health economic viewpoint:

- The use of questionnaire-based risk algorithms
- The new absolute risk factor guidelines being implemented by national bodies
- The impact on fracture rates of testing women around the menopause with long-term follow-up
- Non-BMD fracture predictors
- The costs in countries outside the US and UK.

Only two papers were found, Mueller *et al.* and the paper based on Thai subjects, which included clinical risk factors. All the other papers solely consider DXA or non-intervention. There is a clear need for a new generation of studies which evaluate DXA in comparison and in combination with the new risk calculator-based methods for assessing future risk of fracture and other non-BMD based methods for assessing fracture risk.

1.12. Literature review summary

It is clear from the literature review that the thinking regarding prevention of fractures has undergone significant changes in recent years. These changes are due to the move from the use of relative risks to absolute risks for identifying women at risk. This has resulted in a move from tools with the ability to diagnose osteoporosis, as defined by T-score, to an emphasis on tools with the prognostic ability to predict fracture. However, there is a lack of clinical data on the long-term prognostic power of a number of osteoporotic clinical risk factors in a number of patient groups. The majority of studies to date have focused on following older women over relatively short time periods, i.e. under 5 years.

The burden of the disease is significant and growing and its mechanisms of action need to be better understood to enable better prognostic tools to be introduced. It is clear from this review that there are significant limitations in using BMD as measured by DXA to manage the condition. A number of effective treatment options for fracture risk mitigation already exist and clinicians can use these treatments with the right prognostic tools. While the clinical evidence for most treatments indicate superior vertebral anti-fracture efficacy compared with efficacy at peripheral sites, such as the hip, a number of treatment options still have clinical utility at the hip, which is an important fracture site. There is, therefore, a clear need for mass screening for this condition to minimise the burden of the disease.

The clinical community is open to new prognostic markers for the disease to help address the gap between the good treatment options and the poor prognostic tools to identify women at risk of the condition. The FRAX[®] risk calculator is a popular tool and a viable mechanism for the introduction of new risk factors into clinical practice as they are developed. More research to identify new non-BMD factors is required, but the need for long-term longitudinal data to validate new risk factors represents a significant barrier to their adoption. There is also a need to more widely validate the new online risk calculators in patient groups. Surrogate and direct measurement of bone will provide the most likely breakthroughs in the prediction of who will fracture.

Work at the University of Limerick has identified the nail as a biomarker that could be a prognostic indicator of fracture risk. A longitudinal study evaluating the marker in the target population is required to validate it. The mechanism of action has yet to be definitively established, but the leading hypotheses suggest a relationship through physiological and/or pathological processes, potentially through the blood.

The most suitable technique to extract the clinically relevant data is Raman spectroscopy, but this technique requires sophisticated analytical techniques to extract the data of interest.

It is clear that the potential techniques for the analysis of complex data such as Raman spectra are numerous and growing rapidly. Non-linear analysis techniques are gaining ground over older techniques, such as PCA, as more complex biological data analyses are carried out and better performance is required.

The cohort or observation study is the most appropriate vehicle to validate new risk factors for osteoporosis, and survival analysis is the most suitable tool for comparison of the

different risk factors. The use of archived samples provides an elegant solution to the challenge of producing long-term follow-up data without significant investments of time and money. There is limited data on the long-term stability of nail samples and this is required to provide reassurance that the use of archived nail samples is a valid approach.

The field of health economics to assess the cost-effectiveness of osteoporosis prevention is a young one. Therefore, more work on the impact of non-BMD predictors and the new drug classes like Prolia on the potential for implementing mass screening protocols is required.

In summary, the field of fracture prevention research is a dynamic one. The recent policy changes have created opportunities for new non-BMD markers. There are a number of research questions still to be answered and this thesis will address some of them.

1.13 Research questions: aim and objectives

1.13.1 Aim

The aim of this thesis is to investigate the clinical utility and cost-effectiveness of a number of new non-BMD based prognostic tools for osteoporotic fracture risk. It is clear from the literature review that there is a need for more and easier to use prognostic tools for the identification of women at risk of fracture.

1.13.2 Objectives

To meet the research aims, five inter-related studies were carried out.

Study 1

A retrospective cohort study to establish the prognostic performance of Raman spectra collected from the human toenail using an algorithm developed in the study and validated using leave-one-out techniques.

This work is novel because a prognostic investigation of nail as a predictor of osteoporosis fracture was conducted for the first time. This study uses Raman spectra collected from the Nurses' Health Study as part of the work conducted by Crescent Diagnostics Ltd.

Study 2

An evaluation of the results from the QFractureScores and FRAX[®] questionnaires in a UK multicentre cross-sectional study group and their potential impact on clinical practice.

This study has two aims: to confirm the findings of the first published QFractureScores paper, which underestimates fracture risk compared with FRAX[®], and to confirm that the two questionnaires produce similar prognostic results for women in different parts of the UK. This work is original because no comparison of the two questionnaires in a UK population has been conducted. This risk calculator data entry work was conducted in collaboration with Dr Niamh Cummins (University of Limerick, Ireland). The estimated risk scores were derived from the risk calculators using clinical risk data available in a previously conducted postmenopausal osteoporosis study. This study has been published in a peer-reviewed journal, Cummins *et al. Calcified Tissue International*, 89 [2] 172-7 (2011) .

Study 3

A comparison of non-BMD risk predictors to identify healthy and at-risk women in a longitudinal study group.

The ability of the following predictors to assess fracture risk was assessed:

- FRAX[®]
- QFractureScores
- An internally developed regression model from the Nurses' Health Study in the US.

The null hypothesis is that the two validated questionnaires and the newly-developed model produce similar prognostic results for women.

This study is original for two reasons. First, an analysis of the clinical risk factors in combination with the Nurses' Health Study database has not yet been conducted. Second, a comparison between these risk factors in the FRAX[®] and QFractureScores calculators in a US-based longitudinal study has not been conducted.

Study 4

An evaluation of the NHS study results using novel evaluation techniques.

The opportunity to evaluate the Nurses' Health Study (NHS) results using Cox's proportional hazards was used to evaluate whether they can produce additional information compared with logistic regression measured using area under the receiver operating characteristic curve (AUC) and net reclassification index performance measures.

Study 5

A comparative evaluation of the cost-effectiveness for mass screening of nail test, nail test + clinical risk factors, and clinical risk factors alone based on the previous study results.

This study is original because no screening cost-effectiveness comparison of these risk predictors has been carried out. The AUCs, positive predictive value, negative predictive value, specificity and sensitivity data for the different prognostic measures in conjunction with the costs of preventive drug treatment and the tests was used to compare their efficacy. The clinical utility of the tests was evaluated using measures including number needed to treat to prevent a fracture, and mean costs per fracture prevented including and excluding drug costs.

Chapter 2 : The potential of Raman spectroscopy of the toenail as a prognostic indicator of future fracture risk: a 20-year study

2.1 ABSTRACT

Purpose: The use of clinical risk factors in bone fracture prediction tools such as FRAX[®] has created an opportunity for new prognostic biomarkers to be introduced into clinical practice. The objective of this study was to investigate the potential of a test based on Raman spectroscopy of nail samples, the BQT, as a predictive tool for hip fracture.

Methods: A literature review and stability study were undertaken to ensure that the use of archived nails was appropriate. Archived toenail samples from 82 postmenopausal women aged 50 to 63 in the Nurses' Health Study were analysed using Raman spectroscopy, with a hip fracture of up to 20 years after nail collection, and 81 age-matched controls. A model to calculate a test score was developed based on the spectroscopy results and cross-validated using leave-one-subject-out analysis. The ability of the test score to predict hip fracture was tested with these same women in models, with and without clinical risk factors (CRFs), by comparing the odds ratios (OR) per 1 SD increase in standardised predictive values.

Results: The test score successfully distinguished between the hip fracture cases and controls ($p=0.004$). With only the test score as a predictor, a statistically significant OR of 1.59 for hip fracture was found (95% CI 1.15-2.21). Additional adjustment for the CRFs increased the OR to 2.50 (95% CI 1.69 – 3.70), which was significantly better than a model with CRFs alone ($p=0.0017$).

Conclusions: The BQT is a potential new tool for predicting hip fracture and may be superior to the use of CRFs alone. Further studies are required to evaluate the test in an independent sample and to compare it with measures of bone mineral density.

Keywords: FRAX, fracture risk, Nurses' Health Study, osteoporosis, Raman spectroscopy, toenail.

2.2 INTRODUCTION

The current reference standard for the diagnosis of osteoporosis and assessment of fracture risk is the measurement of bone mineral density (BMD), using dual energy X-ray absorptiometry (DXA) and FRAX[®] respectively. In the developed world, the lifetime risk of a fracture is believed to be between 30% and 40% (Kanis, 2007). A hip fracture has the highest mortality rate of any type of fragility fracture (Johnell and Kanis, 2004), with incidence increasing exponentially with age (Kanis *et al.*, 2002; Boonen *et al.*, 2004). Responding to the rising population burden of the disease, the World Health Organization (WHO) has identified a need for improved prognostic indicators and alternatives to BMD-based diagnostic tools to assess the fracture risk (Kanis, 2007; Kanis *et al.*, 2008b).

Diagnosis and treatment for osteoporosis is currently focused on women over the age of 65. However, it is recognised that the onset of the menopause is a crucial age at which to make long-term treatment decisions, as the most rapid bone loss takes place at this time (MacDonald *et al.*, 2001; Braga *et al.*, 2000). Unfortunately, there are only a few prospective cohort studies with BMD data at baseline that follow women from the menopause until fracture for time periods of over five years compared with the number carried out on women over the age of 65 to assess the predictive power of diagnostic tools. One cohort study from Finland (n=3068) showed that DXA measurements in perimenopausal women predicted forearm fracture risk after five years of follow-up (Huopio *et al.*, 2000). More recently, as part of the Aberdeen Prospective Osteoporosis Screening Study (APOSS) involving 4,800 women at 45-54 years of age at baseline, Stewart *et al.* found that both DXA and quantitative ultrasound (QUS) were predictive of osteoporotic fracture risk at all sites including hip, wrist and humerus over 10 years of follow-up in a subgroup of 1000 women (Stewart *et al.*, 2006). A study in the Study of Women's Health Across the Nation (SWAN) has shown that women with diabetes are at higher risk of fractures compared to healthy women across the menopause in a study with eight years of follow up (Khalil *et al.*, 2010).

Both DXA and QUS require direct patient measurement and regular equipment calibration by trained professionals. The introduction of more convenient tests for both the patient and physician that are capable of accurately assessing long-term fracture risk would be a significant advance in the osteoporosis field. The ideal test would be amenable to a high

sample throughput operation in clinical reference labs, and would work in conjunction with clinical risk factors to enable the screening of a broader population base and aid preventative treatment decisions. Current markers of bone resorption, amenable to central lab testing, include collagen type 1 cross-linked C-telopeptide (CTX) and N-telopeptide (NTX) measured using either serum or urine samples. These biomarkers are used primarily to assess the effects of drug therapy but are also considered to predict hip fracture in the elderly (Garnero *et al.*, 1996a). To date, however, clinical studies have not provided information on the performance of these biochemical markers for prediction of hip fracture in women under 65 years of age over long follow-up periods.

Recently, there has been a trend towards the use of BMD measurements in combination with other clinical risk factors in order to improve overall predictive performance (Cefalu, 2004; Miller *et al.*, 1996). A clinically accepted risk calculator is the Fracture Risk Assessment Tool, FRAX[®] (McCloskey *et al.*, 2009). Clinical risk factors evaluated in FRAX include age, sex, weight, height, previous fracture, parental hip fracture, current smoking, glucocorticoids, rheumatoid arthritis, secondary osteoporosis, alcohol consumption >3 drinks per day and femoral neck BMD. In the United States, postmenopausal women identified as being at risk using FRAX[®] are recommended some form of treatment along National Osteoporosis Foundation (NOF) therapy guidelines (NOF, 2008). This re-definition of fracture risk and diagnosis of osteoporosis is a departure from the historical use of BMD T-score alone and creates opportunities for new, clinically-validated independent risk factors and prognostic tests to be identified and incorporated into risk calculators.

Two previous clinical studies have suggested a relationship between human nail structure, osteoporosis and fracture risk (Pillay *et al.*, 2005; Towler *et al.*, 2007) using Raman spectroscopy, an optical analytical technique used to obtain vibrational information on molecules in a sample as they are excited by a laser source. The spectrum, obtained using a charge-coupled device camera, provides information on the intensity of the signal at particular wavelengths, with signal peaks being associated with particular biochemical structures. Each molecule has specific vibrational modes, depending on the structure and

conformation of that molecule, resulting in a Raman spectrum that is specific for that molecule.

Towler *et al.* hypothesised that nail structure is a surrogate marker for bone structure due to a relationship between the proteins keratin (nail) and collagen (bone) (Towler *et al.*, 2007), which express the same characteristic peaks in key spectral regions (Matousek *et al.*, 2006; Moran *et al.*, 2007). An observation from a physician linking changes in bone from treatment to nail changes led to the initial research. While providing preliminary support for a potential new biomarker of fracture risk, each of these studies was conducted retrospectively and included few hip fractures.

Therefore, the aim of this study was to assess the ability of Raman spectroscopy of nail samples to discriminate between women with and without hip fracture over a time period of up to 20 years in a nested case-control study of postmenopausal women less than 65 y, using nail samples archived in the Nurses' Health Study (NHS). The study also explored whether superior performance of the Raman-based BQT could be obtained in combination with clinical risk factors over risk factors alone.

A key assumption in the use of archived nails for a cohort study using nails to predict future fracture risks is that the nails remain representative of their state when initially collected after over 20 years in storage. It is, therefore, important to be able to demonstrate that archived nails tested many years after collection are representative of the nails at baseline in order to avoid the prohibitive costs of a long-term prospective study. A literature review using the terms keratin, archive, sample, stability and nail identified a number of relevant papers in the field. A number of papers by Edwards *et al.* have demonstrated that the key peaks in keratin bearing materials, hair and nail experience limited degradation over decades and centuries when stored in non-humid and non-light-exposed environments (Akhtar and Edwards, 1997; Williams *et al.*, 1994; Edwards *et al.*, 2002; Edwards *et al.*, 1998). These studies used Raman spectroscopy to compare historical keratin-bearing materials with modern samples, but have not directly compared the results from the same samples.

Toenail samples have been collected in a number of cohorts to evaluate cancer risks from selenium levels (Michaud *et al.*, 2005). The rationale justifying the long-term use of nail clippings is, however, unclear (Landi and Caporaso, 1997). Other major cohorts have evaluated the use of nails; for example, UK Biobank, who rejected it for operational reasons (Elliott and Peakman, 2008). Only one study by Fraser *et al* investigated the long-term (up to six months storage) variability in nails and hair examined using isotope ratio mass spectrometry. They found in the majority of measurements no difference between results at six months and six weeks, with some of the identified variability attributed to the different sample containers used (Fraser *et al.*, 2008). Discussions with a principal investigator of a leading cohort has revealed that their rationale for using nails to measure selenium was that selenium was a material unlikely to evaporate, but no specific research had been undertaken.

Fraser states that nails are suitable for long-term studies since they can be stored for long periods of time at room temperature (He, 2011). He also recommends that they are stored in the driest conditions possible but provides no quantified reason for this assertion. Some recent work by Xun *et al.* has quantified the changes in levels of trace elements in toenails over time. Using toenail clippings collected from a number of subjects, the average coefficient of variation in nail samples was 2.5% for selenium and 6.8% for mercury (Xun *et al.*, 2011). In the CARDIA study, the Spearman correlation coefficient between measures of trace elements collected with a time gap of 20 years was 0.56 for selenium and 0.6 for mercury in a randomly selected group of 64 subjects (Xun *et al.*, 2010). These values indicate that there is a strong positive correlation between the two values separated by a substantial time period.

The key challenge with nail samples collected over two decades ago is the absence of baseline data using the equivalent testing modality. In the case of this specific application, Raman spectroscopy allows a direct comparison of the same nails over an extended period. The CARDIA study does, however, provide an indication of what might be expected in terms of the level of correlation. In terms of demonstrating long-term stability, a well-known method is the Arrhenius method based on temperature change, which serves to accelerate the chemical reactions of aging (Waterman and Adami, 2005). The use of the Arrhenius equation relies on the assumption that the rate of deterioration in the materials follows the Arrhenius reaction rate function. This function states that a 10 degree increase leads to a two times increase in the rate of change. Based on the findings from the literature review, exposure to

temperature was considered important to assess the long-term stability of the nails. Ultraviolet (UV) light is also a known aging contributor for skin and exposing the nail to UV was also considered an important consideration.

2.3 METHODS

2.3.1 Nurses' Health Study

The Nurses' Health Study (NHS) is a cohort study involving 121,700 women, which commenced in 1976. When the study began, the participants were registered nurses aged between 30 and 55 years of age living in one of eleven US states; approximately 96% of the participants are Caucasian. The women have been followed-up by questionnaire every two years. Data collected on the biennial questionnaires that are relevant to this investigation include smoking status, weight and height (from which body mass index was calculated), thyroid hormone use, stroke, cancer, rheumatoid arthritis, osteoporosis diagnosed by a physician, age at menopause, and use of postmenopausal hormones. Frequency of alcohol consumption was assessed every four years as part of a larger dietary questionnaire.

With respect to hip fracture outcomes, the women were asked to report all previous fractures in 1982 and subsequent fractures were recorded on later questionnaires. Reports of hip fractures can be expected to be accurate in a population of trained nurses. Validity was examined in a 1986 study that confirmed thirty reports of hip fracture, all of which were present in the medical records (Colditz *et al.*, 1986). A number of studies focused on hip fracture risk have already been conducted in this cohort (Feskanich *et al.*, 2002b; Feskanich *et al.*, 2004; Feskanich *et al.*, 2003; Feskanich *et al.*, 2002a).

2.3.2 Archived toenail clippings

Toenail clippings were collected from 62,865 participants in the NHS cohort between December 1982 and July 1984 from participants between 58 and 63 years of age. The nail clippings have been stored in a dry environment since collection. The toenails were originally collected to investigate the relationship between selenium levels in the nail and cancer risk

(Hunter *et al.*, 1990; Garland *et al.*, 1995) and have since been used in a number of published studies (Al-Delaimy *et al.*, 2008; Garland *et al.*, 1995; Al-Delaimy and Willett, 2008; Al-Delaimy *et al.*, 2002; Feskanich *et al.*, 1998). The toenail samples used in these investigations were often destroyed in processing; 51,430 samples remained in storage at the commencement of this study.

There are number of alternative approaches that can be used to demonstrate the long-term viability of nails in storage when measures taken today cannot be compared with those taken at baseline. Firstly, a comparison can be made between nails from the archive and recently collected nails. An accelerated aging study using temperature, water, dehydration and UV exposure can be conducted; alternatively, a real-time study commencing in the current time period can be carried out.

A key component of these approaches is the comparison of the nails using a relevant technique. Since the nails are to be analysed using Raman spectroscopy, this technique needs to be used on the nails in all the scenarios. The nails were analysed using the same techniques used to analyse the nail for the assessment of fracture risk. By assessing the variation between the full spectra of the nails, the aim is to demonstrate that there are no statistically significant differences between the nail spectra. In order to be able to operate within a reasonable time frame, an accelerated aging study was conducted by Dr Tibebe Lemebe at Dublin City University and reported in an internal document (Lemma, 2011). This work is described in the methods and description section.

A HORIBA Jobin-Yvon Labram HR 1000 spectrometer was used to make the measurements. The instrument used a Peltier cooled ($-70\text{ }^{\circ}\text{C}$) charge-coupled device (CCD) camera and a 785 nm wavelength laser. The diameter of the laser area was $1\mu\text{m}$ and the device had a spectral resolution of 2 cm^{-1} . Acquisition and exposure times were 11 seconds and 6 minutes respectively. A laser power of 60-80 mW was used over a spectral range of $200\text{-}2000\text{ cm}^{-1}$. Three or four Raman spectra were taken from each nail.

Nail samples were collected from young healthy female subjects at Dublin City University. The nails were washed using soap (Alconox, powdered precision Cleaner) and subsequently washed in de-ionised water. This was followed by sonication in 10mL of acetone to eliminate possible surface contaminants. Acetone had previously been investigated for its suitability as a solvent for cleaning nails and found to be acceptable (Cummins, 2009). The room used to store the nails was at 18-20 °C temperature.

In order to evaluate the effect of hydration on the nails, Raman spectra were collected from the nails after immersion in de-ionised water for three periods, 24hrs, 48hrs and 72hrs.

To evaluate the effects of UV, one sample was irradiated using a UV lamp which emitted light at between 200 and 240 nm. This was at a distance of 1 cm from the sample, which was irradiated for up to 25 minutes.

To evaluate the effects of dehydration, a number of previously unused samples were dried in a vacuum desiccator and compared with samples stored at room temperature by collecting Raman spectra after six days duration. Spectra were collected from the top and underside of the nail.

To evaluate the effects of temperature, one nail was exposed to increasing temperatures from room temperature up to 190 °C and spectra were collected every 20 °C increment. The temperature was held steady at each increment for five minutes before the spectra were collected.

2.3.3 Study population and design

A nested case-control design was selected to establish the ability of the BQT to differentiate between women with a history of hip fracture and control subjects. The study population from which the sample was drawn consisted of women who, at the time of their toenail return, were postmenopausal, between the ages of 50 and 63, not currently using postmenopausal hormones, and without a hip fracture, stroke, or cancer history (n=13,312). From this population, we identified 279 cases who had a hip fracture from 3 to 20 years after

toenail return (median 14.5 y). Hip fractures due to traumatic events (e.g. motor vehicle accident, skiing, and horseback riding) were excluded. Of the remaining 13,033 women in the study population who did not report a hip fracture through 2004, we randomly selected one control per case matched on month and year of birth. One case per control was selected based on a requirement to ensure sufficient unused samples would be available for future analysis.

For the part of the study described in this article in which we developed the test algorithm, we randomly selected 82 case-control pairs from the available 279 pairs to ensure that we would have sufficient remaining data for a later validation study. Approval for the study was obtained from the Institutional Review Board of the Brigham and Women's Hospital; the subjects provided informed consent when the nail samples were collected and the study was conducted in accordance with the declaration of Helsinki.

2.3.4 Nail spectral measurement and processing

Raman spectral analysis of the toenail samples was conducted without blinding as to case or control status. A Model 3510 Skin Composition Analyser (River Diagnostics, Rotterdam, NL), using 785 nm laser light of 60 ± 5 mW (at the sample location) with a spectral range of 300-2200 cm^{-1} and spectral resolution 4-5 cm^{-1} , was used for data collection. In order to account for the heterogeneity of the nail, multiple spectra were taken from areas at 2-6 different locations on the underside of the nail until an average nail spectrum with a 1.5% confidence interval was obtained. The nails were pressed on the measurement window (flattened) using a weight to ensure good contact with the window. The (confocal) Raman measurements were performed at a depth of 30 microns below the surface to minimise possible artefacts due to surface contamination from dirt or surface coats. All nail samples were sufficient for Raman analysis to be conducted. A spectroscopy reading was not determined for one control subject due to the loss of the toenail samples in transit to the testing lab.

An EMSC-scaling procedure (Extended Multiplicative Signal Correction) was used to remove the influence of slowly varying backgrounds and the known interfering signal from the sample carrier, and to scale the spectra to equal Raman spectral contributions (Martens *et*

al., 2003). The EMSC-procedure uses a linear-fitting model to estimate the contribution of polynomial backgrounds and known interfering signals. It subsequently scales the spectra in such a way that the contributions of both the background and interfering signals are equal in all spectra and that the spectra are scaled to Raman spectral contributions only.

2.3.5 Prediction model development

Principal Component Analysis (PCA) and Linear Discrimination Analysis (LDA) were used to develop a predictive multivariate statistical model (Tabachnick and Fidell, 1996). The EMSC-scaled Raman spectra were used to identify the key Principal Components (PCs) of variation across the full spectrum range, representing peak shifts and peak shape alterations. The significance of each PC in discriminating between the two predefined groups, hip fractures and controls, was then determined using a Student's t-test. Finally, an LDA model that best discriminated between the fracture and the control group was developed, using only the two most significant PCs to reduce the risk of over-fitting the LDA model. The scores of the spectra on the LDA discriminant were used to calculate the probability of belonging to the fracture group.

A leave one subject out (LOSO) cross-validation was performed to validate the prediction model internally. For each subject, a model was developed as described above, using all spectra except the spectra from that subject. Subsequently, the average spectrum of that subject was projected on the multivariate statistical model to yield the probability for that subject belonging to the fracture group. These probability values were used in the subsequent statistical analysis. The scientific computing program Matlab (R2007b) was used to develop all signal processing and multivariate modelling algorithms and the test scores were generated in a range between 0.0 and 1.0.

2.3.6 The Raman test and risk of hip fracture

We examined the test scores in conjunction with several CRFs, including age, BMI, smoking status, thyroid hormone use, early age at menopause (<45y), high alcohol consumption (>3

drinks/day), rheumatoid arthritis and diagnosis of osteoporosis. Many of these CRFs are evaluated in the FRAX tool; however, not all FRAX variables could be included due to lack of data (i.e. glucocorticoid use, parental history of hip fracture, secondary osteoporosis, femoral neck BMD). A paired t-test was used to determine significant differences between these CRFs in the hip fracture cases and their matched controls.

We examined the potential of the Raman-based test score to predict hip fractures within the same case/control samples used to develop the test algorithm. Odds ratios (ORs) for hip fracture per 1 standard deviation increase in the test were first calculated using both a conditional regression analysis, with an unconditional analysis adjusted for age at toenail collection (matching factor for cases and controls). The results from these analyses were very similar and therefore the unconditional model was selected.

For identification of significant factors within multivariate analysis, the test score, age, and all CRFs, as well as their first order interactions, were initially included in a regression model. Non-significant factors ($p > 0.1$) were then eliminated starting with those with the largest p-values. Once a final model was established, the performances of the test alone, the CRFs alone, and the test in combination with the CRFs were evaluated. To compare the models which contain both continuous and categorical data, predictive values were generated by logistic regression and standardised to a distribution with mean=0 and standard deviation=1. The distribution of predictive values was checked for normality. The analyses to compare models used these standardised predictive values, yielding an odds ratio (OR) per 1 standard deviation increase. As a secondary approach, the AUCs were calculated. Sensitivity and specificity at different cut-off values was also explored. Statistical significance was based on two-sided tests with a significance level of $p < 0.05$. All statistical analyses were carried out using SAS release 9.2 and Medcalc version 11.3. A sub-group analysis of up to 16 years was selected to allow a shorter time period to be analysed without having too small a sample group.

2.4 RESULTS

2.4.1 Nail Stability

Dehydration

Figure 2.1 shows the spectra for the room temperature and dehydrated nails. The main structural features can be seen in all the different collected spectra. The absolute changes in height relate to intensity which does not affect the diagnostic evaluation, the peak to trough measurements for each peak are the key measure.

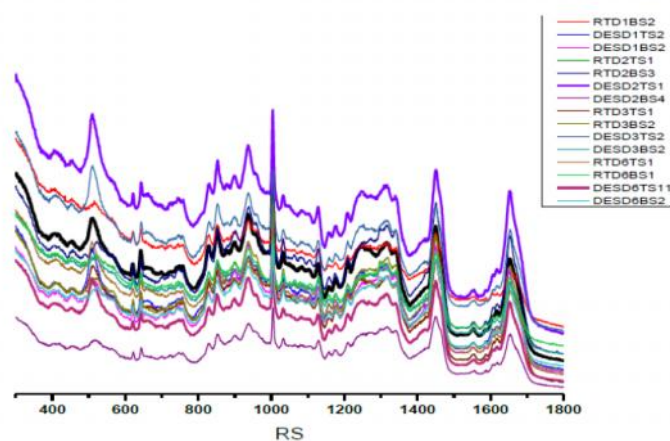


Figure 2.1: Raman spectra from a female toenail stored in desiccator and room temperature for six days. The abbreviations of the file names: RTD1BS2 (Room Temperature, Day 1, Bottom, Spot2), DESD1TS2 (DESiccator, Day 1, Top, Spot2)

Hydration

Figure 2.2 shows that hydration of the nail over long periods results in changes in absolute intensity of the Raman signal, but not in changes in frequency of the peaks.

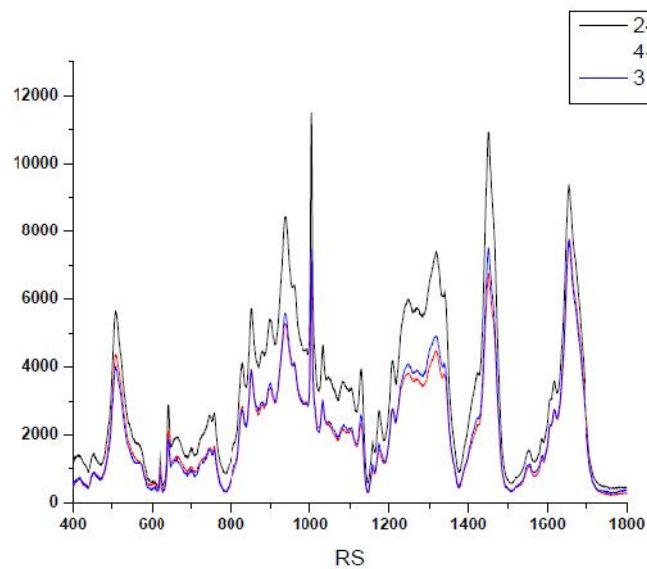


Figure 2.2: Raman spectra of toenail soaked in de-ionized water for three time periods, 24hrs, 48hrs and 72 hrs.

UV aging

Figure 2.3 shows some significant changes in protein structure. These were observed at the S-S bond (510cm^{-1}) and at two other important peaks CH₂ scissoring (1447 cm^{-1}) and amide I (1653cm^{-1}). This indicates that UV aging caused free radical damage or photolysis of the bonds. The nails also physically changed colour to brown under long exposures.

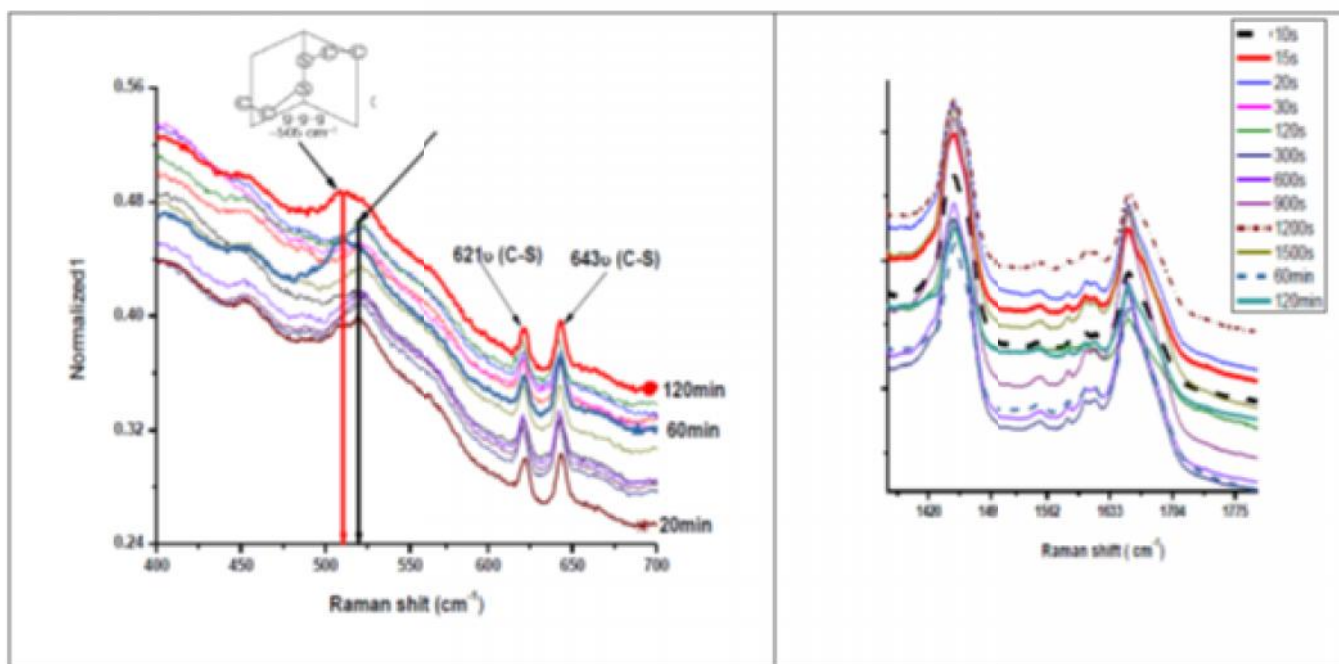


Figure 2.3: Raman spectra of nail following UV light exposure at durations of up to 120 minutes. The chart on the left shows the normalised spectra and the chart on the right the unmodified spectra.

Temperature based changes

Figure 2.4 depicts the Raman spectra at various temperature points. As can be seen, there are changes in intensity but no changes in the position of the peaks in terms of frequency.

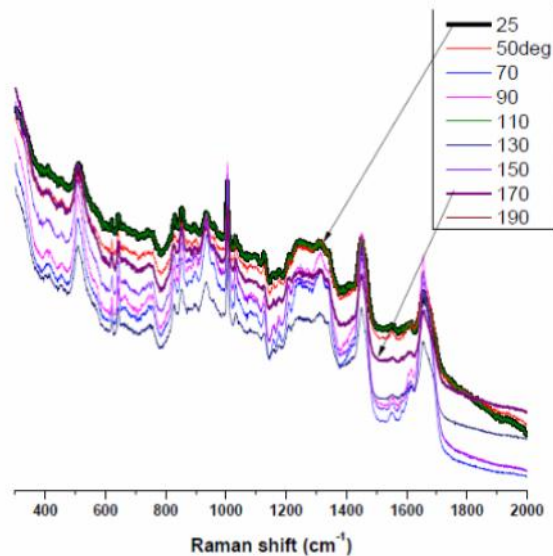


Figure 2.4: Raman spectra of nail following temperature-based changes at temperatures between 25 deg C and 190 deg C.

2.4.2 Characteristics of the study population

Characteristics of the hip fracture cases and controls at the time of toenail collection are shown in Table 2.1. As cases and controls were matched on month and year of birth, the mean age (57) was identical in both groups. Thyroid hormone use was more prevalent in cases than controls ($p=0.02$) and was the only factor that was of any significant difference. None of the cases or controls had a diagnosis of rheumatoid arthritis when the toenails were collected; therefore, this factor could not be considered in further analyses. Ninety-nine per cent of both cases and controls were Caucasian, with one Black subject among the cases and one Asian subject among the controls. The results exclude the subject for which Raman spectra could not be taken.

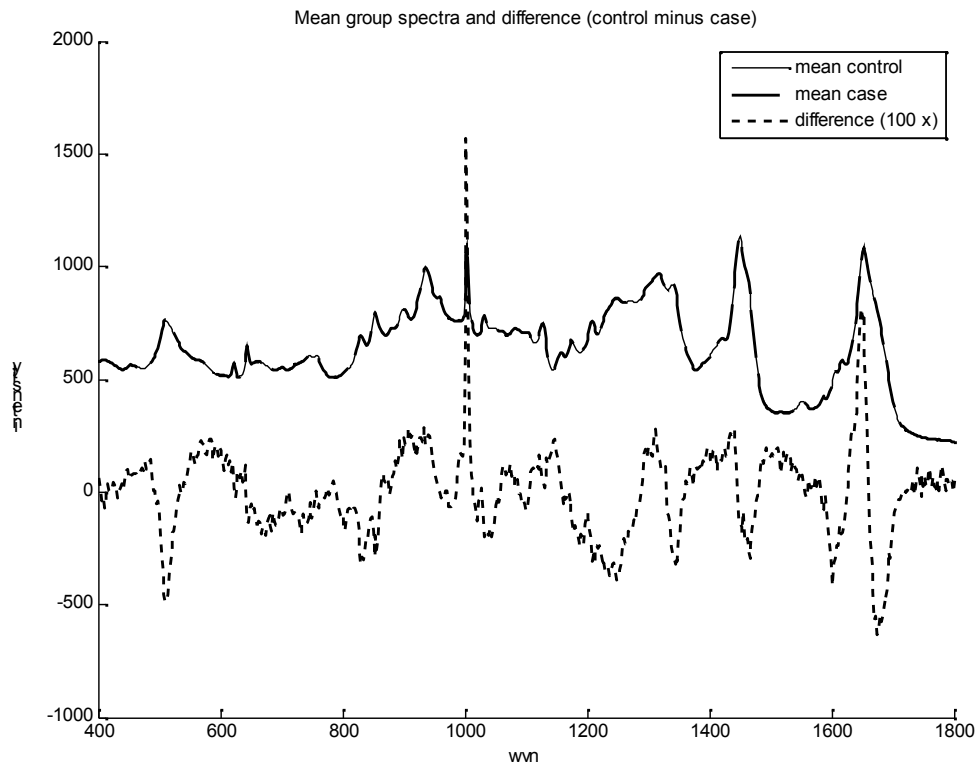
Table 2.1: Clinical risk factors at time of toenail collection in hip fracture cases and controls among postmenopausal women in the Nurses' Health Study.

	Cases	Controls
	(n=82)	(n=81)
Age ^a , y (mean)	57.6	57.6
BMI, kg/m ² (mean)	24.4	25.5
Current Smoker (n)	28	21
Osteoporosis (n)	2	4
Rheumatoid Arthritis (n)	0	0
Age at Menopause < 45 y (n)	8	8
Thyroid Hormone User (n)	16	5
≥ 3 Alcoholic Drinks/day (n)	1	3

^a controls were matched to cases on month and year of birth

2.4.3 Spectral analysis and test algorithm model

Figure 2.5 illustrates the mean spectra grouped into case and control groups. The case and control lines overlap to the naked eye because the differences are too small to be observed without magnification. The mean spectra differences are, however, reliably detectable using the equipment selected and they are magnified 100 times for illustration in the figure. The ratio of inter group coefficient of variation (CV) to inter-subject CV was greater than 2.0 in the Raman spectra collected. With the resulting algorithm for the test, the mean score was significantly higher in hip fracture cases than controls (0.59 vs 0.51, $p=0.004$).



wn: wavenumber cm^{-1}

intensity: arbitrary units

The difference is magnified 100x

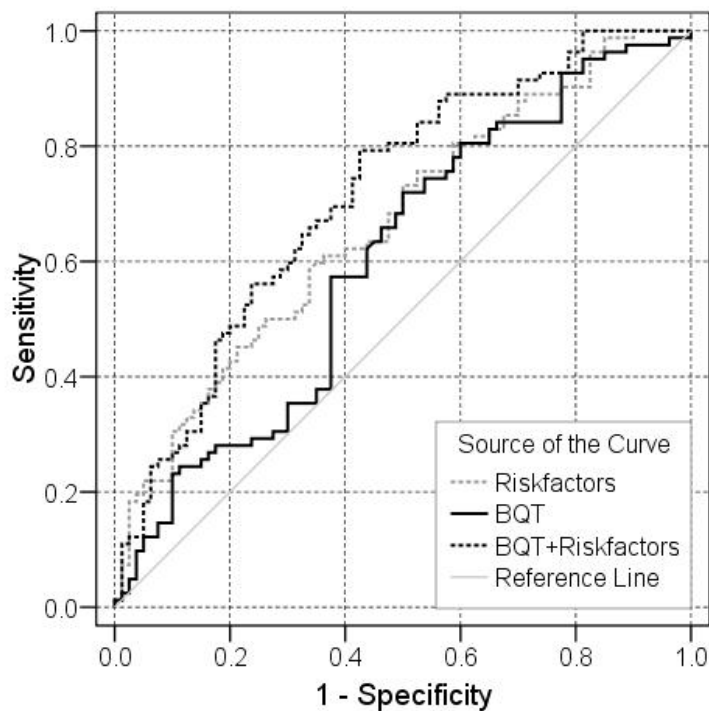
Figure 2.5: Mean case and control spectra and the difference between cases and controls magnified one hundred times. The case and control spectra are visible as a single line because they overlap to the naked eye.

2.4.4 Hip Fracture prediction models

Backward stepwise elimination model

Backward stepwise elimination identified the following CRFs as adding significantly to the test score in the prediction of hip fracture: age, BMI, and the interactions age*BMI, age*thyroid hormone use, and BMI*thyroid hormone use. In a model with only a test score as the predictor, the OR for a hip fracture was 1.59 (95% CI 1.15-2.21) with an AUC=0.61

(Table 2.2). A model with only the significant CRFs as predictors performed somewhat better (OR=2.06, 95% CI 1.40-3.03, AUC=0.66). The predictive ability of this model, however, was significantly improved ($p=0.002$) when the BQT score was added to the CRFs (OR=2.50, 95% CI 1.69-3.70, AUC=0.72). A comparison of the area under the ROC curves for these models is shown in Figure 2.6. The figure illustrates that the curve in which BQT and CRFs are combined had the best performance followed by CRFs on their own.



BQT

Figure 2.6: Areas under receiver operator characteristic curve for prediction models for all 82 hip fracture cases and 81 controls

The predictive ability of the Raman-based test and CRFs improved in subgroup analyses that limited hip fracture cases and their matched controls to less than 16 years post nail collection ($n=99$) (Table 2.2). The OR for the model with both test score and CRFs was 3.56 (95% CI 2.04-6.23) with an AUC of 0.79, which was a significant improvement ($p=0.0009$) over the model with CRFs alone. The mean age of the women in this subgroup (57.7) was very similar to the mean age of the total sample (57.6). A comparison of the ROC curves is shown in figure 2.7, again the BQT and CRFs in combination perform best.

Table 2.2: Prediction of hip fracture by Raman test score and clinical risk factors ^a (CRF) among postmenopausal women in the Nurses' Health Study

Cases/controls		OR (95% CI) ^b	AUC (95% CI) ^c
Total sample	82/81		
Raman test		1.59 (1.15-2.21)	0.61 (0.53-0.68)
CRFs		2.06 (1.40-3.03)	0.66 (0.59-0.74)
Raman test + CRFs		2.50 (1.69-3.70)	0.72 (0.64-0.78)
< 16 y to Hip Fracture ^d	51/48		
Raman test		2.01 (1.28-3.16)	0.67 (0.57-0.76)
CRF only		2.69 (1.54-4.72)	0.73 (0.63-0.82)
Raman test + CRF s		3.56 (2.04-6.23)	0.79 (0.70-0.87)

^a clinical risk factors (assessed at toenail collection) include age, BMI, and interactions age*BMI, age*thyroid hormone use, and BMI*thyroid hormone use

^b odds ratio and 95% confidence interval for risk of hip fracture per 1 standard deviation increase in standardised predicted values of factors in the model

^c area under the Receiver Operating Characteristic curve and 95% confidence interval

^d years from toenail collection to hip fracture of case and its matched control (range is 3-20 y in the total sample)

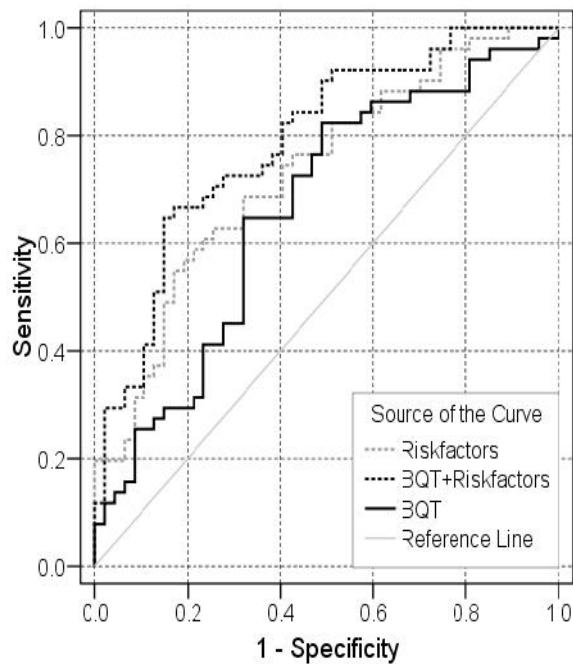


Figure 2.7: Area under receiver operator characteristic curve for prediction models in 51 hip fracture cases and 48 controls up to 15 years post nail collection

Significant risk factor model

Adjusting the model for the significant clinical risk factors (age, smoking status, thyroid hormone use) increased the OR to 2.22 (95% CI 1.53 – 3.20) with an AUC of 0.70. A model based only on the influential clinical risk factors had an OR of 1.71 (95% CI 1.20 – 2.42) and an AUC of 0.62. A statistically significant difference was observed between the ORs for clinical risk factors alone and clinical risk factors in combination with the BQT ($p=0.004$). A comparison of the ROC curves is shown in figure 2.8.

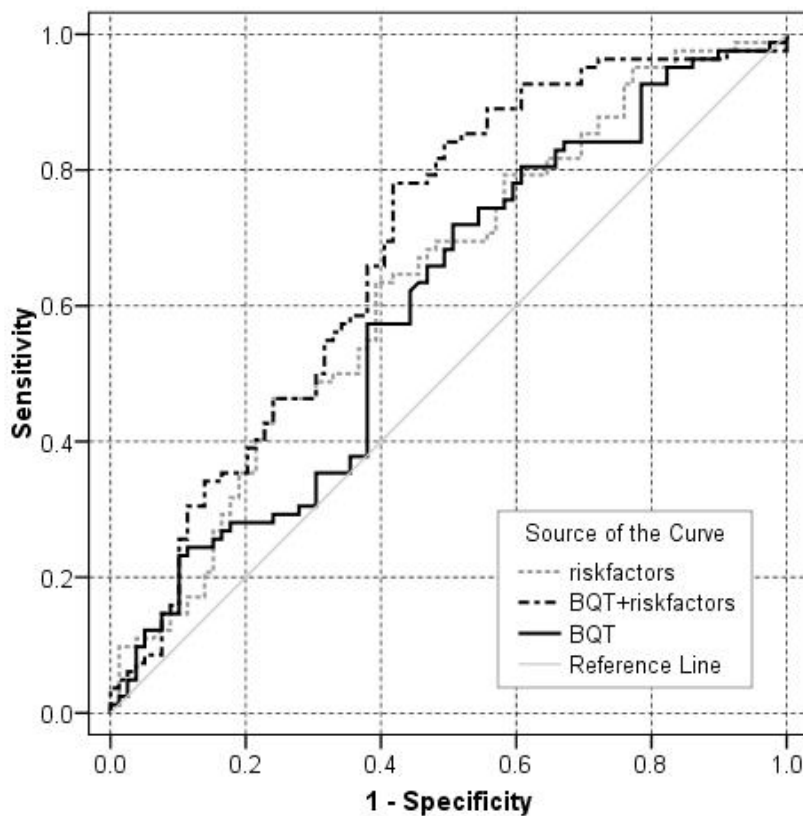


Figure 2.8: Areas under receiver operator characteristic curve for prediction models for all 82 hip fracture cases and 81 controls

In a subgroup analysis that limited hip fracture cases to 15 years post nail collection (n=98), the predictive ability of the Raman test improved, with an OR of 2.0 (95% CI 1.28 – 3.16, $p < 0.01$) and area under the curve (AUC) of 0.67. In combination with the clinical risk factors, the figures for OR and AUC were 2.60 (95% CI 1.57 – 4.30) and 0.73 respectively. Using clinical risk factors alone, an OR of 1.73 (95% CI 1.09 – 2.76) and an AUC of 0.60 was achieved. The use of the BQT significantly improved the model over CRFs alone ($p = 0.0037$). A comparison of the ROC curves is shown in figure 2.9.

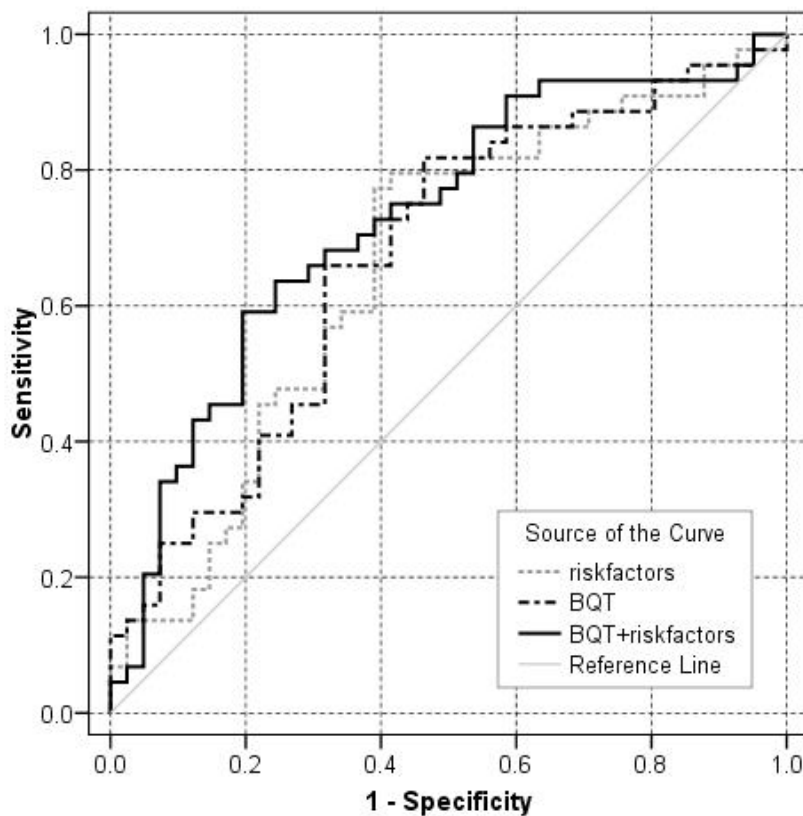


Figure 2.9: Area under receiver operator characteristic curve for prediction models in hip fracture cases up to 15 years post nail collection and their matched controls.

Table 2.3: Diagnostic performance for the different diagnostic scenarios based on alternative cut-off values

Predictor	sens at 80% spec	sens at 90% spec	AUC	OR at 80%spec	OR at 90%spec
BQT	28.0%	23.2%	0.61	1.58 (0.77 – 3.28)	2.75 (1.15 -6.58)
BQT + smkst	28.0%	19.5%	0.65	1.56 (0.75 – 3.23)	2.18 (0.89 – 5.37)
BQT + BMI	36.6%	13.4%	0.65	2.31 (1.14 – 4.66)	1.39 (0.53 – 3.67)
BQT + all	41.5%	20.7%	0.68	2.79 (1.39 – 5.58)	2.32 (0.95 – 5.66)

As can be seen in table 2.3, when all the risk factors are combined, increasing the specificity reduces the odds ratio in the case of BQT and all risk factors available. This table allows a

comparison between the performance of BQT with and without associated risk factors at high specificities which are similar to the specificities of DXA to be made. DXA is considered to have a sensitivity of less than 30% at specificities in excess of 80%.

2.5 DISCUSSION

Archived samples

Previous studies have suggested a relationship between human nail structure, osteoporosis and fracture risk, providing preliminary supporting evidence for a new biomarker of fracture risk (Pillay *et al.*, 2005; Towler *et al.*, 2007). Taking advantage of archived nail samples, this study examined the ability of Raman spectroscopy of nail samples to identify fracture risk over periods of 3 to 20 years in postmenopausal women aged less than 65 years. The use of archived toe nail samples enabled this longitudinal study to be conducted in a resource-efficient manner.

The use of archived nails for this type of analysis is based on the assumption that they are not substantially changed from collection at baseline, and this appears to be valid from the literature. Previous studies have demonstrated consistent Raman spectral signatures over time, e.g. a comparison of 500-year-old mummified baby nails with the nails of modern day babies showed little variation in key structures (Edwards *et al.*, 2002). Processes such as freezing and drying can be harmful to the overall integrity of the nail, but keeping the nail dry and at a consistent temperature will support its preservation over potentially hundreds of years and enable representative Raman spectra to be collected decades after the original sample was stored without degradation (Gniadecka *et al.*, 1998). The nails in the Nurses' Health Study have been kept for over 25 years in a manner consistent with these findings. Increasing the temperature from 25 to 170°C resulted in only minor changes and no observable structural degradation in the characteristic proteins bands, which suggests the robustness of the keratin structure, which was unexpected. Further increasing the temperature to 190°C resulted in significant structural changes to the sample. As this high temperature is well outside normal storage ranges, the samples will produce a high structural alteration and the sample will undergo thermal decomposition. A study of a number of months would be required to fully mimic the full archive duration of 23 years.

Dehydration and hydration led to changes in peak intensity, but not to the magnitude of the Raman shift related to the key structural peaks. Exposure to UV is clearly damaging to the nail samples and, therefore, it is essential that nails archived are stored in a dark environment.

In order to maintain good structural integrity that is suitable for high quality Raman spectroscopy on nail samples stored over the long term, keeping the nails at room temperature in a dark environment should be sufficient.

Prediction model

A prediction model was developed using the multivariate analysis techniques PCA and LDA, which have produced more predictive results compared with those achieved using simple peak height measurement methods to separate the hip fracture cases and control groups (Towler *et al.*, 2007). The prediction model was cross-validated using a leave one sample out cross-validation technique. The amount of data did not permit the use of separate training and validation sets. However, the same two principal components were identified in each iteration, demonstrating that the model was not sensitive to any single spectra and, therefore, robust.

In this study, we demonstrate a significant 59% increased risk of hip fracture per 1 SD change in the test score compared with age-matched controls for postmenopausal women between 50 and 63 years of age, from which the algorithm for the test score was developed. The BQT may be a more predictive marker of fracture risk for periods of less than 20 years, as evidenced by the higher 100% increased risk in the first 16 years after the nail clipping was collected. It can also be seen that the test score in combination with clinical risk factors is more predictive than risk factors alone, which suggests that the Raman-based test is providing new information and could provide a significant performance enhancement.

DXA and QUS have been shown to be predictive of fracture risk, particularly in the elderly over short time frames (Hans *et al.*, 1996; Guglielmi *et al.*, 2003; Khaw *et al.*, 2004). In perimenopausal women between the ages of 45 and 54, who were followed for an average of

10 years, Stewart *et al.* reported hazard ratios (HR) of 1.90 (95% CI 1.54-2.32) for osteoporotic fractures per 1 SD decrease in spine BMD and 1.78 (95% CI 1.43-2.20) per 1 SD decrease in femoral neck BMD when measured by DXA (Stewart *et al.*, 2006). The results were similar in the subgroup with BMD measurements using QUS. The majority of the osteoporotic fractures in this population were wrist fractures, a less serious and costly outcome than the hip fractures assessed in the current study. If we find similar results to those shown in this study, when the Raman-based test is validated in an independent sample, its performance in combination with risk factors would be at least comparable to those reported by Stewart *et al.* for DXA and QUS.

A relationship between nail structure and various features of BMD has previously been hypothesised. Ohgitani *et al.* reported a significant positive correlation between the calcium content of finger nails and lumbar BMD in a group of postmenopausal women ($p=0.0016$). A similar relationship was also observed for toenail samples ($p=0.0215$) (Ohgitani *et al.*, 2005). Collagen and keratin are fibrous proteins that serve structural and mechanical roles in the body, providing a framework for the support of cells and tissues. Both proteins consist of polypeptide chains formed by amino acid condensation (Branden and Tooze, 1999) and express the same characteristic bands (CH_2 and amide I) in key regions of Raman spectra collected from biological samples (Matousek *et al.*, 2006; Moran *et al.*, 2007). By-products of bone remodelling are evident in serum and urine; therefore, it is conceivable that this evidence of changes in bone chemistry could also be detected in nails, a continually growing material which is in direct contact with the periosteum of the phalangeal bone during its growth (Ohgitani *et al.*, 2005).

There have been calls to implement population-based screening protocols for osteoporosis, supported by recent fracture risk data around the time of the menopause and the demonstrated improvement in outcomes based on the earlier implementation of treatment, when appropriate (Barr *et al.*, 2009). In contrast to DXA or QUS, tests amenable to high throughput testing by clinical reference labs could facilitate screening a broader population base in combination with clinical risk factor based screening tools. A number of markers of bone resorption are offered as central lab tests, although to-date clinical studies have not provided information on

the biochemical marker fracture prediction performance in women under 65 over long follow-up periods. The simplicity of preparing, transporting and testing nail clippings may enable central lab testing for fracture risk assessment. A simple to use test, which could accurately identify high-risk women who would most benefit from increased monitoring and treatment at an earlier stage, could be a significant public health advancement.

In the short-term, decisions on treatment prescription will continue to be made using DXA and new predictive markers will act as an effective pre-screen in this context. In the near future, predictive markers may come into their own as individual countries implement the new WHO guidelines advising prescriptions based on absolute future fracture risk rather than T-scores (Kanis, 2007).

A limitation of this study was that the BQT scores were derived from an algorithm which was validated using leave-one-out techniques rather than using separate training and validation sample sets. This approach was selected to preserve the limited resource of samples from fracture cases available in the Nurses' Health Study for future validation analyses. A direct comparison between DXA and the test could not be made since DXA data was not collected on the NHS cohort at baseline.

In conclusion, this preliminary study suggests that Raman spectroscopy of the human nail, as captured by the BQT score, could be a clinical tool to use, in combination with other risk factors, for identifying postmenopausal women less than 65 years of age who are at increased risk of hip fracture over a period of up to 20 years. The preliminary results compare favourably with existing technologies and support the view that the prediction model provides a platform for a confirmatory study in an independent population to validate the prognostic marker. Further work to understand the underlying mechanism of action linking bone and nail is also required.

Chapter 3 : A Comparison of FRAX and QFractureScores in a case-control study in Ireland and the UK

1. Modified and expanded¹ version of a manuscript published in collaboration with Dr Niamh Cummins (University of Limerick) and Professor Stuart Ralston (School of Molecular and Clinical Medicine, Western General Hospital, Edinburgh, UK in *Calcified Tissue International* 89 [2] 172-7 (2011) (Appendix A). The dataset was also used as the basis for an MSc research project by Joan Quigley, the abstract is provided in Appendix D.

¹ Modifications include: title, cross-referencing; removing redundancies; and a greater level of detail

3.1 ABSTRACT

Introduction: The systematic use of clinical risk factors (CRFs) in algorithms derived from cohort studies to estimate a 10-year probability of fracture is a significant advance in the management of osteoporosis. The aim of this study was to compare the performance of the FRAX and QFractureScores algorithms in a patient group using their individual CRFs. The secondary aims were to evaluate the prevalence of the relevant clinical risk factors, evaluate the clinical utility of the recommendations using the UK National Osteoporosis Guideline Group (NOGG) recommendations and compare the predicted fracture rates in Ireland and the UK.

Methods: Postmenopausal women (584) between the ages of 50 and 85 participated in this retrospective case-control study. Demographic, anthropometric and lifestyle data were collected during a single clinic visit. Femoral bone mineral density (BMD) was measured by dual energy X-ray absorptiometry (DXA) at the same visit. An absolute fracture risk prognosis was calculated using both the FRAX and QFractureScores online algorithms.

Results: Mean major fracture risk for QFractureScores was 9.5% vs. 15.2% for FRAX. Similarly, hip fracture risk values were 2.9% and 4.7% respectively. The correlation between FRAX and QFractureScores was $R=0.803$ for major fracture and $R=0.857$ for hip fracture ($p \leq 0.0001$). Regional differences in risk factors and fracture risk were also observed, using an AUC comparison for major fracture against actual fractures. QFractureScores was 0.67, FRAX including BMD was 0.82, while FRAX excluding BMD was 0.87.

Conclusions: The higher FRAX calculated absolute risks are more in line with the observed outcomes in published cohort studies, indicating that this algorithm may be more predictive than QFractureScores when the same clinical risk factors are utilised. The level of correlation between the FRAX and QFractureScores algorithms may be considered satisfactory for clinicians involved in the management of osteoporosis. The FRAX and QFractureScores algorithms could be modified to include other relevant CRFs in the future, if the factors can provide the clinical data required. Regional differences in the prevalence of clinical risk factors and fracture risk may warrant further investigation to introduce specific lifestyle advice and reduce fracture risk in targeted geographical regions.

3.2 INTRODUCTION

The clinical significance of osteoporosis has been described in Chapter 1.4. The development of FRAX and QFractureScores is described in Chapter 1.6. A list of the clinical risk factors used by the two algorithms is shown in Table 3.1 below. The link between the 10-year probability and the clinical utility of the FRAX algorithm is shown in Figure 3.1 taken from a Kanis *et al.* paper (Kanis *et al.*, 2009). The chart comes from the National Osteoporosis Guideline Group (NOGG) guidelines which were developed to add clinical utility to the FRAX absolute risk scores (Compston *et al.*, 2009). Two charts are shown because one is used when BMD data (the chart on the right) is available and the other when it is not. When BMD data is available for the assessment of fracture risk, a cross is shown on the second chart enabling a decision on whether to treat or not. When the data input into the algorithm lacks BMD information, then a recommendation to measure BMD can be made based on the chart on the left.

Absolute risk predictions for major fracture (leg, arm, and wrist) or hip fracture are both available whether BMD is available or not. However, NOGG guidelines are only available for the following situations on the FRAX website:

- Major Fracture (no BMD)
- Major Fracture (with BMD)
- Hip Fracture (with BMD).

A hip fracture NOGG recommendation is not provided by the FRAX website when BMD data is not supplied.

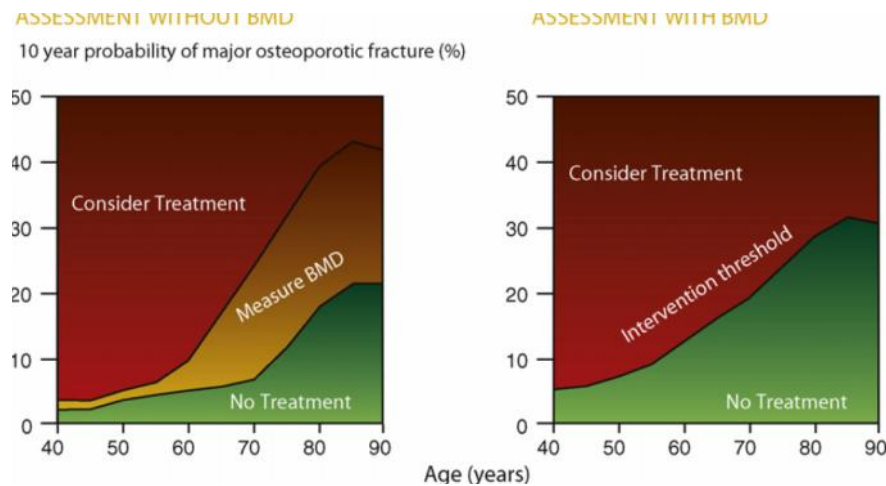


Figure 3.1: NOGG assessment and treatment thresholds in relation to fracture probability. The first chart is used when BMD is included in FRAX and the second chart is used when BMD information is available.

This study was undertaken using data available from the FRAX study, which assessed a new diagnostic test for osteoporosis. The focus of the FRAX study was to evaluate the link between fingernail structure, osteoporosis, fracture history and DXA scores. Subsequently, it was realised that the resulting data, as well as the introduction of FRAX and QFractureScores into clinical use, presented an opportunity to conduct original research to compare the algorithms, investigate the relationship between them and the nature of clinical risk factors across different regions of the Britain and Ireland.

The primary aim of this study was to compare the performance of the FRAX and QFractureScores algorithms, for estimation of absolute osteoporotic fracture risk in an independent female postmenopausal group. The secondary aims were to evaluate the diagnostic power of the individual clinical risk factors compared with use in combination within the algorithms, evaluate the clinical utility of the recommendations made using the UK National Osteoporosis Guideline Group (NOGG) guidelines and compare the predicted fracture rates and NOGG recommendations between Ireland and the countries of the United Kingdom.

Table 3.1: CRFs evaluated by the FRAX and QFractureScores Algorithms

Clinical Risk Factor	FRAX	QFractureScores
Age	X	X
Sex	X	X
Weight	X	X
Height	X	X
Previous fracture	X	
Parental hip fracture/osteoporosis	X	X
Smoking	X	X
Glucocorticoids*	X	
Rheumatoid arthritis	X	X
Secondary osteoporosis**	X	
Alcohol intake	X	X
Femoral neck BMD	X	
Asthma		X
Steroids		X
Heart attack/Stroke		X
Falls		X
Chronic liver disease		X
Tricyclic antidepressants		X
Type 2 diabetes		X
HRT		X
Endocrine problem		X
Gastrointestinal Malabsorption		X
Menopausal symptoms		X

* Use of steroids is a CRF in QFractureScores but it does not specifically question the use of Glucocorticoids (a class of steroid).

** Similarly some QFractureScores questions relate to causes of secondary osteoporosis (e.g. endocrine problem) but do not question the actual presence of secondary osteoporosis.

3.3 METHODS

3.3.1 Study Design

The study was co-led by Professor Stuart Ralston (Western General Hospital, Edinburgh) and E. Poku. The study investigated the ability of the BQT and Dual energy X-ray Absorptiometry (DXA) to discriminate between subjects with and without a history of low trauma fracture. Subjects were recruited from osteoporosis clinics in England (2), Ireland (1), Wales (1) and Scotland (2). Table 3.2 provides details of the participating centres.

Table 3.2: Participating centres in the FRAN study

Principal Investigator	Centre
Dr. Gill Pearson	Southampton University Hospitals NHS Trust
Prof. Declan Lyons	Mid-Western Regional Hospital, Limerick, Ireland
Prof. Stuart Ralston	Western General Hospital, Edinburgh
Prof. Richard Eastell	Northern General Hospital, Sheffield
Dr. Michael Stone	Cardiff University Academic Centre, Llandough Hospital
Dr. Alastair McLellan	Greater Glasgow and Clyde NHS trust

Ethical approval was obtained for each of the participating hospitals (REC 07/Q1704/1). Written informed consent was obtained from all subjects prior to enrolment and the initiation of study procedures. Eligible subjects were Caucasian women aged between 50 and 85 years (inclusive) and at least five years post-menopausal based on self-report. Subjects prescribed bone-active medication or corticosteroids were ineligible, as were those with known bone metabolic disease (except osteoporosis). Subjects with gastrointestinal malabsorption, liver, kidney and malignant disease were also excluded. This was a retrospective case-control study and all study requirements, including the collection of demographic, anthropometric and lifestyle data, were completed during a single clinic visit. The Clinical Report Form was completed by the attending nurse. The nurse also confirmed that relevant fractures were low impact fractures. Subjects were recruited at the time of their DXA scan or within 12 months

of their scan. Overall recruitment took place over a thirteen-month period. Upon entry into the study, subjects were assigned to the non-fracture or fracture study arm as follows: the non-fracture group consisted of subjects who had never sustained a fracture in adulthood (>18y), while the fracture group consisted of subjects with a history of low-trauma fracture at the hip, spine, humerus, pelvis or wrist, after the age of 45.

DXA scans were performed as per local hospital protocol. The resulting BMD g/cm² and T-scores for antero-postero lumbar vertebrae, total hip and femoral neck were recorded. Subjects with falsely elevated (FE) lumbar spine BMD/T-scores were recorded as such and analysed separately, as this artefact is known to impact on data validity. Quality assurance (QA) of DXA equipment was performed by the study monitors on a regular basis. At the hospitals both Hologic and GE machines were used.

The Clinical Report Form for the original study was developed prior to the introduction of FRAX to clinical practice and the publication of the QFractureScores algorithm. The questions on the form were selected based on the most widely known risk factors to influence fracture risk; additionally, some of the data used in the algorithms were exclusion factors in the original study. As can be seen in the following table, the original study provided most of the data required by FRAX, but less of the data required by QFractureScores.

The data collected from the nurse used to complete the Case Report Form is shown in table 3.3:

Table 3.3: A comparison of the CRFs used by FRAX and QFracture against the data available in the study group

Clinical Risk Factor	FRAX	QFractureScores	Data Available
Age	X	X	X
Sex	X	X	X
Weight	X	X	X
Height	X	X	X
Previous fracture	X		X
Parental hip	X	X	X
Smoking	X	X	X
Glucocorticoids	X		Exclusion factor
Rheumatoid arthritis	X	X	Exclusion factor
Secondary osteoporosis	X		
Alcohol Intake	X	X	X
Femoral neck BMD	X		X
Asthma		X	
Steroids		X	
Heart attack/Stroke		X	
History of Falls		X	X
Chronic liver disease		X	
Tricyclic antidepressants		X	
Type 2 diabetes		X	
HRT		X	
Endocrine problem		X	
Gastrointestinal Malabsorption		X	
Menopausal symptoms		X	X

3.3.2 Data-Entry

FRAX is currently available online at www.shef.ac.uk/FRAX. The underlying algorithm (web version 3) is not publicly available and therefore data entry using the UK version was

conducted manually with a double data-entry rate of 10%. The 10-year probabilities (%) of major osteoporotic and hip fracture with and without BMD were recorded, as were the corresponding NOGG recommendations. Femoral neck BMD was selected from the DXA data available for data entry as per FRAX guidelines, while vertebral data was also available from the original study.

The QFractureScores algorithm was accessed at www.qfracture.org. The underlying weightings and coefficients which make up the algorithm (web version 1) are in the public domain and are available on the website. This enabled the data entry procedure to be automated using an algorithm embedded into an MS Excel spreadsheet. The macro within the spreadsheet provided both major and hip 10-year percentage absolute risks as outputs. NOGG recommendations are not electronically linked to QFractureScores, the recommendations therefore had to be estimated manually. This was carried out using the published NOGG guideline charts available at the FRAX website. Only major fractures could be evaluated against the NOGG guidelines, since QFractureScores does not consider BMD as a risk factor, and a hip fracture absolute risk estimate is not made by FRAX when BMD data is not available. As a result, a NOGG recommendation was not available in the situation where a potential patient does not provide BMD data.

3.3.3 Statistical Analysis

Statistical analysis was performed using SPSS software V17 (Microsoft, California, USA) and MedCalc V11.3 (Medcalc, Belgium). Descriptive statistics are presented as mean values and standard deviations (SDs). Variables were tested for normality using the Kolmogorov-Smirnov test. Depending on the normality of the distribution, Pearsons or Spearman's correlation coefficients were calculated to examine the relationship between variables. The Students t-test or the Mann-Whitney test (MWU) was used, as appropriate, for comparisons of two groups. In the case of multiple groups, ANOVAs (with Bonferroni post-hoc if necessary) or Kruskal Wallis tests were used, again depending on normality.

3.4 RESULTS

3.4.1 Characteristics of the Study Sample

A total of 640 subjects were originally recruited, 16 of whom were ineligible for further data analysis due to the reasons outlined in Table 3.4. A total of 584 postmenopausal women with complete datasets, including a measurement of femoral neck BMD, were included in this analysis. Characteristics of the study sample are presented in Table 3.5. Fractures occurred in 246 subjects comprising 42% of the sample. Wrist fractures were most common (n=173) with fractures of the hip (n=22), vertebra (n=21) and humerus (n=30) also recorded. There was a significant difference in age ($p \leq 0.01$), history of falls ($p \leq 0.0001$) and femoral BMD/T-score ($p \leq 0.0001$) between the control and fracture groups. A history of falls was recorded in 39% of the total sample (n=227) and 32% of subjects had a family history of osteoporosis (n=188). A diagnosis of osteoporosis was confirmed in 15% of the sample (n=87) and 57% were identified as osteopenic (n=333). Body Mass Index (BMI) ranged from 15.8-49.3 kg/m² with 2% underweight (n=10), 32% normal weight (n=188), 40% overweight (n=234) and 26% categorised as being clinically obese (n=152) based on BMI. Current smokers comprised 15% of the sample (n=87) and 33% were ex-smokers (n=190) with 9% of subjects (n=50) reporting an alcohol intake of ≥ 14 U/wk.

Table 3.4: Ineligible Subjects in the study group

Reason for exclusion	No. of subjects
BQT unobtainable	2
No samples obtained	2
DXA scan >12 months	1
Fracture other than at 5 sites	2
Excluded disease	2
Bone medication or HRT	3
Age or <5 years post-menopausal	4
Total	16

Table 3.5: CRF characteristics of the total study sample split into controls and the fracture group (n=584).

Clinical Risk Factor	Total Sample (n=584)	Controls (n=338)	Fractures (n=246)
Age, mean (SD), Y	67 (7)	66 (7)	68 (8)
Height, mean (SD), M	1.60 (0.06)	1.60 (0.06)	1.59 (0.06)
Weight, mean (SD), Kg	69 (13)	70 (13)	69 (13)
BMI, mean (SD), kg/m ²	27.2 (5.0)	27.3 (5.0)	27.1 (5.0)
Femoral BMD, mean (SD), g/cm ²	0.710 (0.130)	0.739 (0.136)	0.670 (0.109)
Femoral T-score, mean (SD), a.u.	-1.4 (1.0)	-1.2 (1.1)	-1.7 (0.9)
Relative with Fracture, Number	n=188 (32%)	n=113 (33%)	n=75 (30%)
History of Falls, Number	n=227 (39%)	n=69 (20%)	n=158 (64%)
Current Smokers, Number	n=87 (15%)	n=49 (14%)	n=38 (15%)
Alcohol $\geq$ 14 U/wk, Number	n=50 (9%)	n=27 (8%)	n=23 (9%)

Characteristics of the study sample by geographical region are presented in Table 3.6. The number of fracture patients recruited per country was as follows: England n=69, Ireland n=16, Scotland n=139 and Wales n=22. Age, weight and BMI were similar across all the regions, but there was a significant difference in the height of subjects from Scotland, England and Ireland ($p \leq 0.001$). Regional differences in femoral BMD (g/cm²) were observed; however, this is likely to be due to the use of DXA units from different manufacturers, as no such differences were observed in relation to T-scores which are normalised across manufacturers. The number of subjects with a family history of osteoporosis ranged from n=21 (24%) in Ireland to n=35 (55%) in Wales. A history of falls was reported in just 11% of Irish women (n=10) in comparison to 64% of Scottish women

(n=124). The proportion of current smokers was comparable in England n=32, Scotland n=28 and Wales n=8 (14%, 14% and 13% respectively) but was somewhat higher in Ireland (n=19; 22%). The number of subjects reporting an alcohol consumption of greater than 14 units/week was lowest in Ireland (n=0; 0%) and highest in Wales (n=11; 17%).

Table 3.6: Characteristics of the CRFs in the study sample by geographical region.

Clinical Risk Factor	England (n=237)	Ireland (n=88)	Scotland (n=195)	Wales (n=64)
Age, mean (SD), y	66 (7)	68 (6)	67 (8)	66 (7)
Height, mean (SD), m	1.61 (0.06)	1.61 (0.06)	1.58 (0.06)	1.60 (0.06)
Weight, mean (SD), kg	70 (13)	68 (12)	70 (14)	68 (12)
BMI, mean (SD), kg/m²	27.1 (4.8)	26.3 (4.6)	28.0 (5.5)	26.7 (4.6)
Femoral BMD, mean (SD), g/cm²	0.694 (0.115)	0.816 (0.142)	0.684 (0.121)	0.700 (0.119)
Femoral T-score, mean (SD), a.u.	-1.4 (1.0)	-1.4 (1.2)	-1.6 (1.0)	-1.3 (1.1)
Relative with Fracture, number	n=74	n=21	n=58	n=35
History of Falls, number	n=69	n=10	n=124	n=24
Current Smokers, number	n=32	n=19	n=28	n=8
Alcohol ≥14 U/wk, number	n=25	n=0	n=14	n=11

3.4.2. Absolute Fracture Risk Correlation Results

Without BMD Comparison

The absolute fracture risks excluding a BMD measurement (- BMD), calculated by the QFractureScores algorithm, were generally lower than those generated by the FRAX algorithm under the same circumstances. In the total study sample, the mean major fracture risk for QFractureScores was 9.5% (range 1.7-37.0%) vs. 15.2% (3.4-49.0%) for FRAX. Similarly, for hip fracture risks, the values were 2.9% (range 0.2-29.6%) and 4.7% (0.2-36.0%) respectively. However, there was a significant correlation between the risks calculated by QFractureScores and FRAX as demonstrated in Figure 3.2 (n=584). The correlation for major fracture was $R=0.803$, while for hip fracture it was $R=0.857$ ($p \leq 0.0001$). Comparisons between sample groups revealed significant differences between QFractureScores (- BMD) and FRAX (- BMD) for major fracture risk ($p \leq 0.0001$) and hip fracture risk ($p \leq 0.0001$).

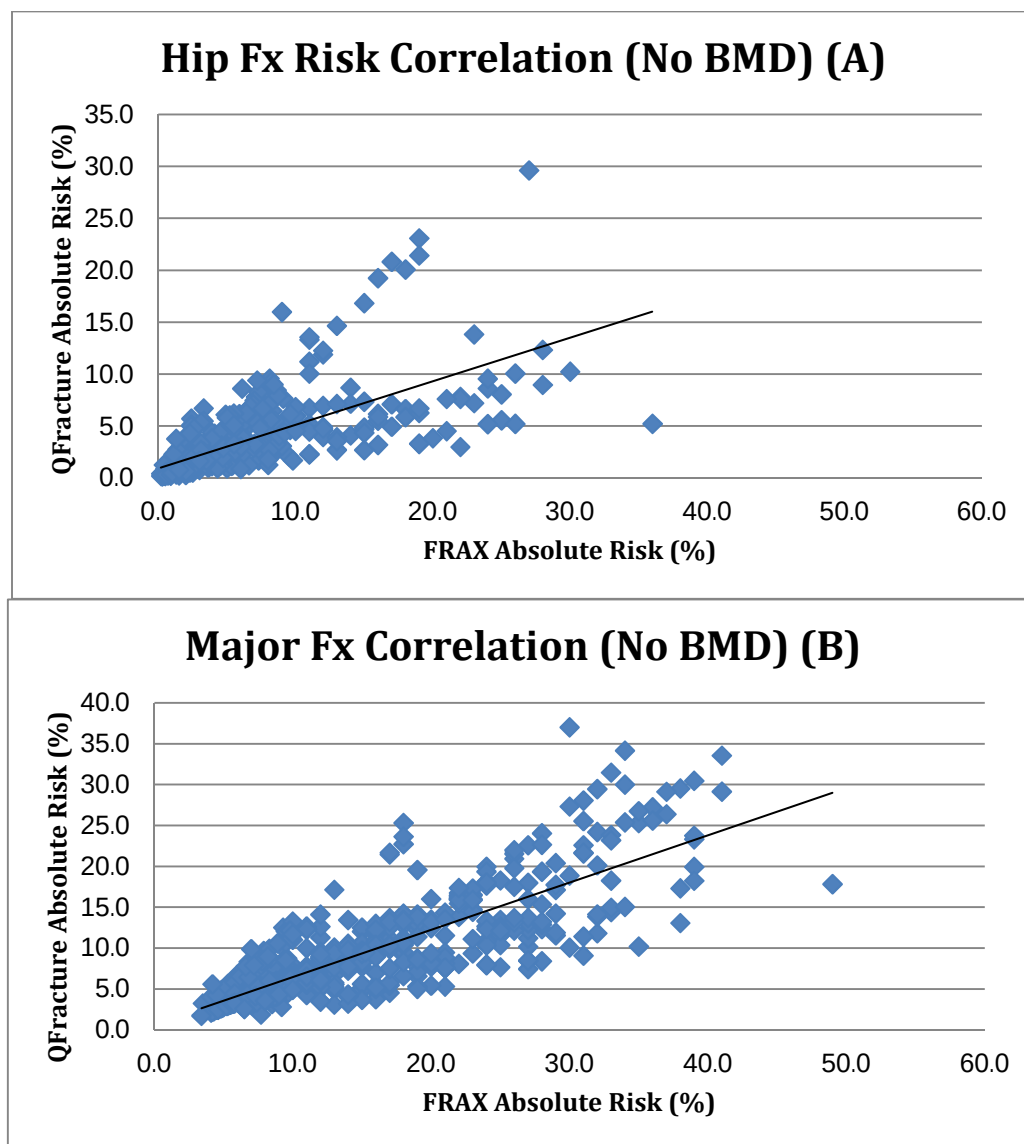


Figure 3.2. Correlation between FRAX and QFractureScores absolute risks (- BMD) for major fracture (A) and hip fracture (B).

Without BMD vs With BMD Comparison

Lesser correlations were observed when the outputs were compared with (+) and without (-) BMD. The correlation between QFractureScores (- BMD) and FRAX (+ BMD) for major fracture was $R=0.657$ and for hip fracture it was $R=0.638$ ($p \leq 0.0001$). The correlation between FRAX (- BMD) and FRAX (+ BMD) was $R=0.862$ for major fracture and $R=0.770$ for hip fracture ($p \leq 0.0001$). Significant differences were also observed between QFractureScores (- BMD) and FRAX (+ BMD) for major fracture risk ($p \leq 0.0001$) and hip fracture risk ($p \leq 0.05$).

FRAX without BMD vs FRAX with BMD

Absolute risks for FRAX (- BMD) vs. FRAX (+ BMD) were also significantly different for major fracture ($p \leq 0.05$) and hip fracture ($p \leq 0.0001$).

Comparison of all algorithms for major fracture

Figure 3.3 allows the trends for absolute risk vs. age to be observed for a major fracture based on a linear interpretation. FRAX without BMD provides the steepest curve which indicates that a woman's risk increases fastest with age using this assessment approach.

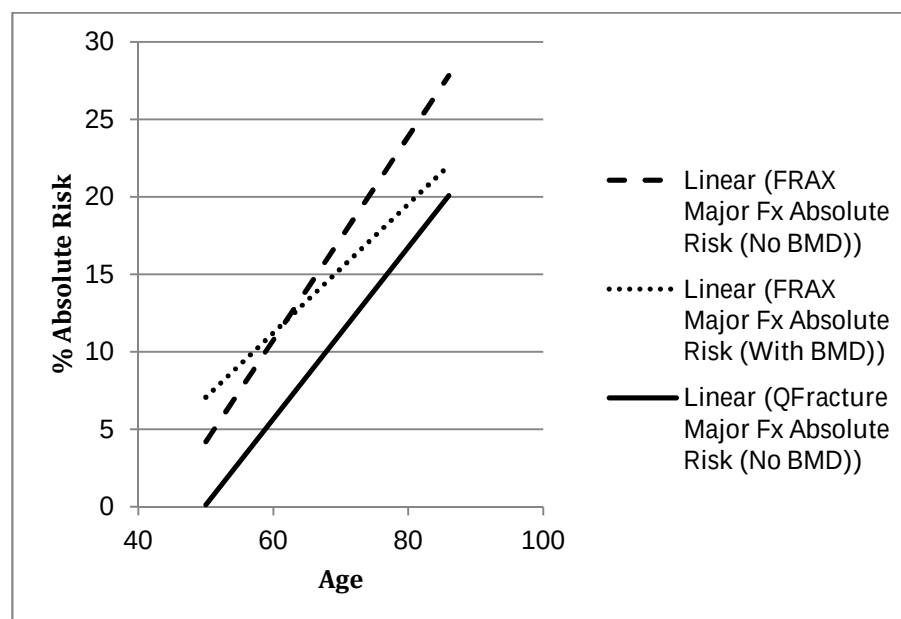


Figure 3.3: Trendline View: Major Fracture for FRAX and QFracture with and without BMD.

Comparison of all algorithms for hip fracture

Figure 3.4 shows that the results for FRAX and QFractureScores excluding BMD are very similar. When BMD is excluded from FRAX, the results are higher in the over 70s. The FRAX without BMD scores have a steeper line than the other two measures.

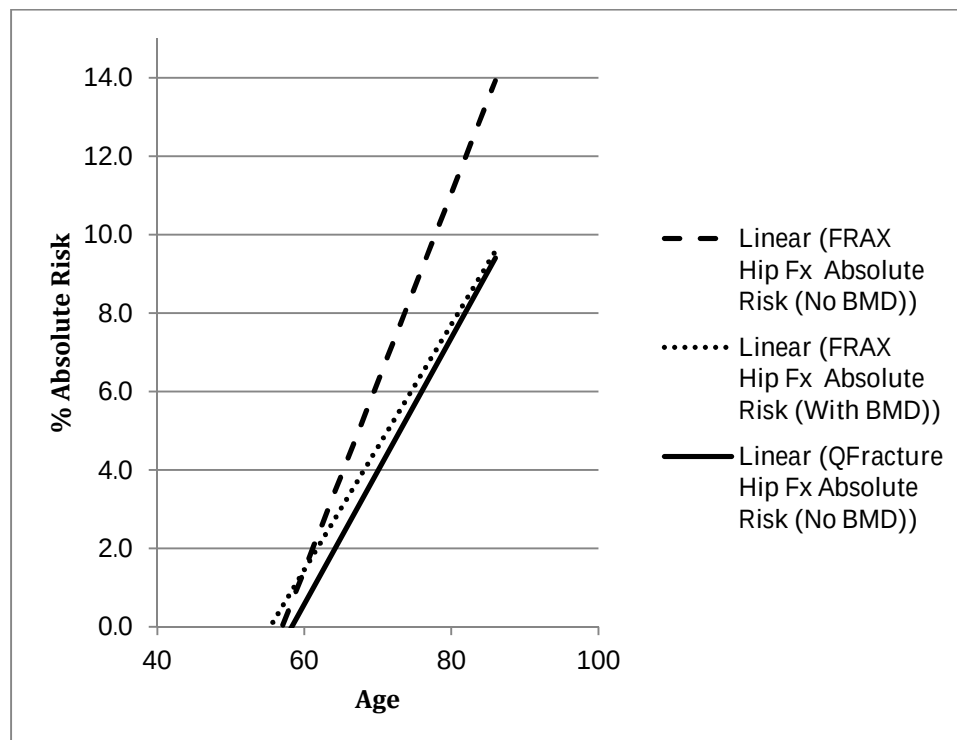


Figure 3.4: Trendline View: Hip Fracture risk for the three treatment scenarios: FRAX with BMD, FRAX without BMD and QFracture.

3.4.3 Regional Variations in Absolute Risk Results

Regional variations in fracture risk were observed, with the highest risks being recorded in the Scottish sample and the lowest risks in the Irish sample. Comparisons between groups revealed no significant difference ($p \leq 0.205$) between regions for FRAX hip fracture risk (+ BMD); however, all other output variables were significantly different across the sample ($p \leq 0.001$).

The absolute risks for major and hip fracture calculated by the FRAX and QFractureScores algorithms are presented in Table 3.7.

Table 3.7: Absolute fracture risks (%) for the total study sample and by geographical region expressed as Mean (SD).

	Group (n=584)	England (n=237)	Ireland (n=88)	Scotland (n=195)	Wales (n=64)
QFractureScores Major Fracture Risk (- BMD)	9.5 (6.0)	8.7 (5.6)	7.9 (3.8)	10.9 (6.7)	10.2 (6.0)
FRAX Major Fracture Risk (- BMD)	15.2 (8.3)	13.7 (7.8)	12.8 (6.1)	17.7 (8.6)	16.7 (9.9)
FRAX Major Fracture Risk (+ BMD)	14.1 (7.8)	12.7 (7.2)	12.9 (7.7)	16.0 (7.8)	14.9 (8.9)
QFractureScores Hip Fracture Risk (- BMD)	2.9 (3.3)	2.5 (2.6)	2.5 (2.8)	3.7 (4.2)	2.6 (2.2)
FRAX Hip Fracture Risk (- BMD)	4.7 (5.5)	4.2 (5.1)	3.4 (3.4)	5.7 (6.0)	5.5 (6.9)
FRAX Hip Fracture Risk (+ BMD)	3.6 (5.0)	3.2 (4.5)	3.3 (5.6)	4.2 (5.0)	3.7 (5.9)

3.4.4 Diagnostic evaluation

A diagnostic performance comparison between the different algorithms and the CRFs can be made using the FRAX scores and the fracture outcome available from the subjects. The diagnostic power of the questionnaires was assessed using a ROC curve. Both major fracture and hip fracture measures were considered in the analysis, figure 3.5 shows the different ROC curves for FRAX with BMD, FRAX without BMD and QFracture for major fracture risk. The outermost line represents FRAX without BMD, the middle line FRAX with BMD and the inner solid line QFracture. The linear line is the reference line representing no discrimination.

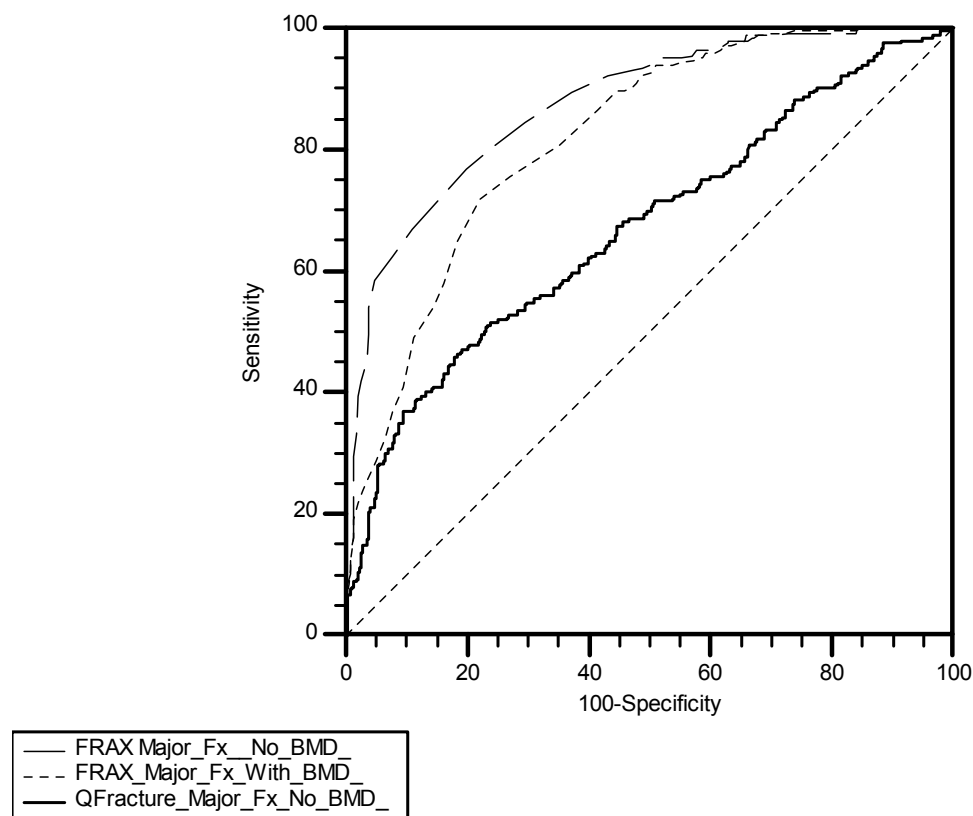


Figure 3.5: ROC curves for the FRAX with BMD (middle line), FRAX without BMD (outer line) and QFracture (inner line) algorithms and major fracture.

Table 3.8: Diagnostic performance of the different algorithms

MAJOR FRACTURE	AUC	P
QFracture	0.67 (0.63-0.72)	<0.001
FRAX + BMD	0.82 (0.79-0.85)	<0.001
FRAX - BMD	0.87(0.85-0.90)	<0.001

Table 3.8 and figure 3.5 show the differences in AUC performance between the three algorithms, it can be seen that FRAX without BMD performed best, followed by FRAX with BMD. Pairwise comparison of the AUC values for the three algorithms revealed that differences between the three measures were statistically significant. These results are shown in Table 3.9.

Table 3.9: Pairwise comparison of the FRAX and QFracture algorithms for major fracture

Measure 1	Measure 2	
QFracture	FRAX + BMD	P<0.001
QFracture	FRAX – BMD	P<0.001
FRAX + BMD	FRAX - BMD	P<0.001

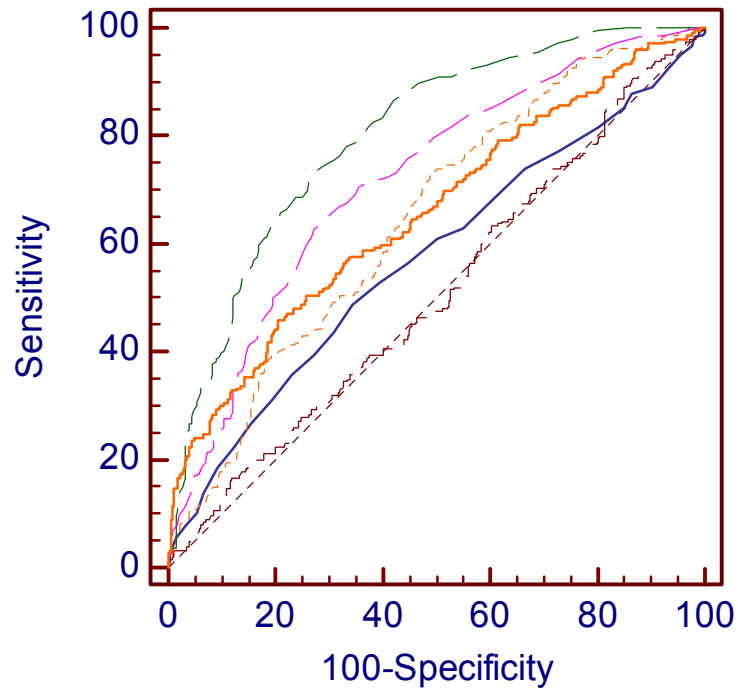


Figure 3.6: A comparison of selected CRFs and algorithm outputs

The outmost line (green) in figure 3.6 represents FRAX without BMD for predicting hip fracture. The orange line represents QFracture, the blue line represents age, while the orange dotted line is Femoral Neck BMD. The purple line represents FRAX with BMD, and the brown line is BMI.

Table 3.10: Comparison of the AUCs of CRFs with the FRAX and QFracture algorithms

	AUC (SD)
Age	0.58 (0.53-0.61)*
BMI	0.51 (0.47-0.55)
History of Falls	0.72 (0.68-0.76)*
Smoking	0.50 (0.46-0.54)
Alcohol	0.51 (0.465 – 0.55)
Femoral Neck g/cm	0.65 (0.61-0.68)*
FRAX Hip No BMD	0.8 (0.77 – 0.83)*
FRAX Hip With BMD	0.72 (0.68-0.75)*
QFractureScores Hip	0.66 (0.62-0.70)*

***Statistically significant**

Table 3.10 provides a comparison of the individual risk factors and the multivariate algorithms; FRAX without BMD performed the second best performer was History of Falls, a single clinical risk factor. Smoking, alcohol consumption and BMI were not statistically significant risk factors. History of falls was a strong diagnostic for fracture.

3.5 NOGG Recommendations

Without BMD comparison FRAX vs QFractureScores

The NOGG recommendations resulting from the major fracture risk results for FRAX, excluding BMD and QFractureScores, were similar. The number of subjects recommended for lifestyle advice was nearly identical for QFractureScores (n=195; 33%) and FRAX (n=197; 34%). However, the number of subjects recommended for a BMD measurement was higher for QFractureScores (n=365; 63%) than for FRAX (n=302; 52%); accordingly, the number recommended for treatment was lower for QFractureScores (n=24; 4%) than for FRAX (n=85; 15%). Overall, there was a significant difference ($p \leq 0.05$) between the NOGG-derived recommendations for QFractureScores and FRAX.

With BMD Comparison

With the inclusion of BMD in the FRAX tool, based on the NOGG major fracture risk thresholds, 73% of subjects (n=425) were recommended for lifestyle advice in comparison to 84% of subjects (n=493) using the NOGG hip fracture risk thresholds. Accordingly, for major fracture risk, 27% of subjects (n=159) were recommended for treatment compared to 16% of subjects (n=91) for hip fracture risk. This equates to a discrepancy in recommendations for 12% of subjects (n=68). In line with this, the FRAX tool borderline thresholds were observed in 10% of subjects (n=56) resulting in NOGG recommendations based on observer judgement.

Regional Variations

Regional variations in NOGG recommendations were also observed. Excluding BMD, for major fracture, significant differences between countries were observed for QFractureScores ($p \leq 0.001$) and FRAX ($p \leq 0.0001$). With the inclusion of BMD in the FRAX tool, there was a significant difference between countries for major fracture ($p \leq 0.001$) but not for hip fracture ($p \leq 0.137$) risk recommendations.

Table 3.11: NOGG Recommendations for the total study sample for the QFracture and FRAX algorithms for hip and major fracture (n=584).

	Lifestyle Advice	Measure BMD	Treat
QFractureScores Major Fracture (- BMD)	n=195 (33%)	n=365 (63%)	n=24 (4%)
FRAX Major Fracture (- BMD)	n=197 (34%)	n=302 (52%)	n=85 (15%)
FRAX Major Fracture (+ BMD)	n=425 (73%)	N/A	n=159 (27%)
FRAX Hip Fracture (+ BMD)	n=493 (84%)	N/A	n=91 (16%)

Table 3.11 provides a comparison of the number of subjects provided with lifestyle advice, how many are recommended for treatment, further BMD assessment or lifestyle advice for the four diagnostic strategies under investigation. Statistical analysis showed that the different measures provide statistically significant differences between QFractureScores Major Fracture and FRAX Major Fracture ($p < 0.001$).

3.6 DISCUSSION

The systematic use of clinical risk factors (CRFs) through algorithms derived from cohort studies to estimate the 10-year probability of a fracture is a significant advance in the management of osteoporosis.

3.6.1 Demographic Data

The aim of this study was to compare the performance of the FRAX and QFractureScores algorithms and to evaluate the prevalence of risk factors and predicted fracture rates in Ireland and the UK. In this study, there was a significant difference in age (66 vs. 68y; $p \leq 0.01$), history of falls (69 vs. 158; $p \leq 0.0001$) and femoral T-score (-1.2 vs. -1.7; $p \leq 0.0001$) between the control and fracture groups respectively. This was not unexpected as age, falls history and BMD are among the strongest risk factors for fracture (Kanis, 2007). However, the history of falls data was striking, as a history of falls was recorded in 39% of subjects. Falls history is included as a CRF in the QFractureScores tool (Hippisley-Cox and Coupland, 2009), but not in the FRAX tool, as patients selected on the basis of risk factors for falling may not respond well to agents that preserve bone mass (Kanis *et al.*, 2008a). The debate over whether falls should be included in these studies is an area which needs more research. However, the risk of falling may be reduced by a non-pharmaceutical intervention such as physical activity, which has been shown to improve balance and muscle strength, in addition to having positive effects on bone density (Nelson *et al.*, 1994). Therefore, in the authors' opinion, falls history is an important factor for inclusion in CRF based algorithms.

A family history of osteoporosis was reported in 32% of subjects and low BMD was recorded in 72% of the sample (15% osteoporotic and 57% osteopenic). Although this percentage is certainly higher than would be found in the general population, our intention was to evaluate the algorithms using the most relevant subjects that would be targeted for CRF screening. This approach was regardless of osteoporosis status as this is not material in the comparison. In relation to BMI, there was a wide range (15.8-49.3 kg/m²) across the sample. Although 66% of subjects were categorised as being overweight (40%) or obese (26%), this is in line with the increasing prevalence of obesity in Western countries. BMI will become a more important factor influencing fracture risk in the future as a result of this trend (World Health

Organization, 2000). Fracture risk is markedly higher at lower values of BMI, particularly with a BMI of less than 20kg/m². However, the relationship between BMI and fracture risk is non-linear and from 25-35kg/m² the differences in fracture risk are quite small (De Laet *et al.*, 2005). Therefore, obesity should not be regarded as a protective factor against fractures (World Health Organization, 2000). With regard to regional variations in CRFs, as expected the groups were quite similar. There were some minor differences between the regions, however, these may have been due to a bias in recruitment. For example, a smaller number of fracture patients were recruited in Ireland in comparison to Scotland and this, in turn, most likely influenced the data collected on history of falls (11% vs. 64% respectively). Regional rates of smoking and alcohol intake are generally in line with those found in the literature (Robinson and Bugler, 2008; Morgan *et al.*, 2008). The finding that none of the Irish women in this study reported an alcohol consumption of greater than 14 units/week was somewhat unexpected, particularly considering that Ireland has previously been reported to have the highest proportion of regular drinkers in the European Union (European Commission, 2002). However, it should be noted that the validity of self-reported alcohol consumption in general is notoriously unreliable (Midanik, 1988).

3.6.2. Absolute risks

The absolute fracture risk values calculated in this study are broadly in agreement with previous published findings, where QFractureScores predicted lower risks than FRAX (Hippisley-Cox and Coupland, 2009; Kanis *et al.*, 2008a). The mean major fracture risk for QFractureScores was 9.5 vs. 15.2% for FRAX. Similarly, for hip fracture risk the values were 2.9% and 4.7% respectively. Significant differences were observed between the values calculated by the different algorithms, with values generated by QFractureScores being consistently lower than those generated by FRAX. However, the conformity between the methods was still strong. The correlation for major fracture was R=0.803, while for hip fracture it was R=0.857 (p≤0.0001). This conformity between FRAX and QFractureScores could generally be considered satisfactory for clinicians involved in the management of osteoporosis in the absence of BMD data. The linear graphs show a marked difference in output when BMD data is included in the FRAX model, while risks rise significantly in the over-70s compared with QFractureScores and FRAX without BMD. This correlates with the WHO report which estimates that most women have a 50% risk of fracture over a lifetime

(Kanis, 2007). The AUCs for the different approaches are widely different, but all are statistically significant. Major fracture risk calculated using FRAX (FRMJR) appears to be the most predictive, which is surprising considering it lacks BMD information. This finding has to be tempered by the fact that the comparison being made is retrospective and the risk factors are designed to identify prospective fracture risk. However, the result may indicate that BMD is not the key factor in determining fracture risk in these subjects and that some of the other factors may be more predictive in this particular study. As can be seen from the AUCs for the various risk factors, in this study, history of falls was a very strong factor and its use in combination with other non-significant risk factors diluted its influence. Clearly, the algorithms provide a superior systematic performance to any individual CRF.

3.6.3. Regional variations

Regional variations in fracture risk were observed which, again, may be due to the different number of fracture patients recruited in each country. In order to fully evaluate the regional variations in fracture risk between Ireland, England, Scotland and Wales, a larger population-based study may be appropriate to evaluate which CRFs are more significant in different countries, the contributing factors and how these risks may be reduced. As a starting point, the QResearch database could be broken down into its constituent countries and analysed appropriately.

3.6.4 NOGG recommendations

The NOGG recommendations were developed as a result of a requirement for assessment and intervention thresholds based on clinical considerations and cost-effectiveness (Kanis *et al.*, 2008c; Compston *et al.*, 2009). In this study, for major fracture risk the number of subjects recommended for lifestyle advice was very similar to that of QFractureScores and FRAX (33 vs. 34%). However, the number of subjects recommended for a BMD measurement was higher for QFractureScores than for FRAX (63 vs. 52%) and, in line with this, the number recommended for treatment was lower for QFractureScores than for FRAX (4 vs. 15%). Overall, there was a statistically significant difference in the NOGG recommendations between FRAX and QFractureScores. This can be attributed to the lower absolute risk values calculated by QFractureScores in comparison to FRAX.

Consistency in recommendations is very important for primary care physicians. Additionally, with the inclusion of BMD in the FRAX tool, there can be disagreement in the recommendations for major and hip fracture. For example, in this study, based on major fracture risk results, 73% of subjects were recommended for lifestyle advice in comparison to 84% based on the hip fracture risk results. Accordingly, for major fracture risk, 27% of the subjects were recommended for treatment in comparison to 16% for hip fracture risk. This is equivalent to a discrepancy in recommendations for 12% of subjects. Borderline thresholds can also be observed, which occurred in 10% of subjects in this study, resulting in subjective, observer-based recommendations. However, it should be noted that these recommendations are a useful guide for physicians, but are not intended to replace clinical judgement (Kanis *et al.*, 2008c; Compston *et al.*, 2009). The large number of people recommended for BMD testing (over 50%) when BMD data was not entered indicates that the clinical community will continue to rely heavily on BMD testing to assess fracture risk, despite the introduction of these tools. It is clear that the risks increase significantly when BMD information is included.

Regional variations in NOGG recommendations were also observed, which was not unexpected as the recommendations are based directly on the absolute fracture risk values, which also varied between regions.

3.6.5 Diagnostic performance

QFractureScores is clearly inferior to FRAX in terms of diagnostic performance. This may indicate that FRAX is a more robust tool, since it has been developed using multiple clinical databases. This level of performance difference is clearly clinically relevant if it can be repeated in other studies. The superior performance of FRAX without BMD compared with FRAX when BMD is included is surprising and the reasons for this may be the types of fracture in the study group. The diagnostic performance of the individual risk factors is also interesting; history of falls, for example, had a performance of 0.72, a performance which is superior to the algorithms. Only FRAX excluding BMD provided superior results. This suggests that, in this particular study, the other risk factors, which were not statistically significant, were diluting the performance of the history of fractures in the algorithms. In this study group, the majority of CRFs were not clinically significant for diagnosing fracture; this may be caused by the fact that the study uses a case-control design rather than a large cohort-

based prospective design, but suggests these risk factors are less influential compared with BMD and major risks, such as history of falls.

3.6.6 Impact of the new algorithms on clinical care

At present, there is no universally accepted UK policy for screening to identify individuals with osteoporosis or those at high risk of fracture (Compston *et al.*, 2009). This situation has been driven by the cost-benefit equation using the current diagnostic tools available, which is not favourable for mass screening from a health economic viewpoint. The main advantage of QFractureScores is that it does not require any laboratory testing or clinical measurements, which influences the cost-benefit equation significantly. All of the variables used in the algorithm will either be known by the patient or can easily be collected by a clinician during a standard consultation (Hippisley-Cox and Coupland, 2009). Due to its simplicity, QFractureScores could be used at a population level to identify high-risk patients that might benefit from a more detailed assessment, including a measurement of BMD. The FRAX model can be used with or without the inclusion of a femoral neck BMD measurement (Kanis *et al.*, 2008a). Some studies have indicated that FRAX does not offer significant advantages over less complex models that also incorporate BMD (Ensrud *et al.*, 2009; Pluskiewicz *et al.*, 2010). However, at the moment it is certainly the most complete CRF-based algorithm available, considering the extensive data collected during its development and the validation involved (Kanis *et al.*, 2008a) (Kanis *et al.*, 2008c).

The advantage of FRAX is that it has been designed as a platform technology, which can be upgraded as new validated risk indicators become available (Kanis *et al.*, 2009). The absolute risk evidence requirements mean that long-term studies and significant investments will be required for a new clinical risk factor to meet the FRAX inclusion criteria. CRFs that may be considered for FRAX in the future include some of those already evaluated in QFractureScores; for example, history of falls and dose response effects for CRFs, such as smoking and alcohol consumption. Other behavioural factors such as diet (in particular, calcium and vitamin D) and physical activity levels may also be relevant. Medications, aside from steroids, that influence bone health may be worth inclusion; for example, aromatase inhibitors and even pharmaceutical treatments for osteoporosis such as bisphosphonates and teriparatide. Diagnostic measurements, such as BMD at other skeletal sites, quantitative ultrasound, and biochemical markers of bone remodelling may also be appropriate for

inclusion in future CRF tools. The implementation of a more detailed CRF-based algorithm may be more time-consuming for clinicians, but ultimately improvement of patient care is the most important consideration and costs overall would be saved in the long term. The major barrier to the introduction of wider CRFs into FRAX is the lack of large prospective studies for new CRFs, which can improve the overall predictive power of these algorithms, further research of which is needed in this area.

3.6.7 Study Limitations

To the best of the authors' knowledge, this is the first study to compare the FRAX and QFractureScores tools in an independent sample. Limitations of this study include its retrospective case-control design and relatively small sample size. However, this study did recruit subjects from Ireland and across the UK, and the prevalence of risk factors for fracture have not previously been well-documented in Ireland.

3.7 CONCLUSIONS

The level of conformity between the FRAX and QFractureScores algorithms observed in this study when BMD is absent could generally be considered satisfactory enough for clinicians involved in the management of osteoporosis to consider using both algorithms. When the NOGG guidelines are used, the large number of subjects recommended for DXA testing is an issue for the practical application of these algorithms. The clinical utility of these techniques, when BMD information is not available, may be limited when over half the population is recommended for treatment. Both of these CRF-based algorithms could have an application in primary care practice, depending on access to DXA and the clinical setting. In relative terms, these CRF algorithms are in their infancy and will most likely be developed to include other relevant CRFs in the future, if the evidence hurdle is lowered. Regional differences in clinical risk factors and fracture risk were observed across Ireland and the UK; this may warrant further investigation in prospective studies, in order to ensure the most appropriate regional fracture prevention strategies are in place. The diagnostic performance observed suggests that FRAX is a superior algorithm based on high AUCs observed for both major and hip fracture when compared with QFractureScores.

Chapter 4 : A comparison of the QFractureScores and FRAX risk calculators and a regression model with the actual incidence of fracture in a nested case-control within the Nurses' Health Study cohort

4.1 ABSTRACT

Introduction: A number of new online risk calculators for osteoporotic fracture have been introduced in recent years. Understanding their relative performance at predicting outcomes in different patient populations is important in order to understand the best situations in which to use them. An analysis of the relative performance of the QFractureScores and FRAX risk calculators and an internally developed regression model has been conducted by analysing questionnaire-derived non-BMD risk predictors for healthy and at-risk women in the Nurses' Health Study database. This study is original for two reasons. First, an analysis of the clinical risk factors in combination with the Nurses' Health Study database using online risk calculators has not yet been conducted. Second, a comparison between the FRAX[®] and QFractureScores calculators in a US-based longitudinal study has also not been conducted.

Methods: Risk factors from the NHS were manually entered into the FRAX online calculator and an MS Excel spreadsheet was used to generate QFractureScores. Logistic regression was used to develop a model and generate risk scores for each subject. The results for each individual were then compared with the actual outcomes recorded for each subject based on an established high risk/low risk cut-off.

Results: The FRAX calculator produced an AUC of 0.52 ($p=0.65$) and the QFractureScores produced an AUC result of 0.554 ($p=0.65$). A logistic regression based on the available data produced an AUC of 0.65 ($p=0.15$), while the BMI alone produced an AUC of 0.58 ($p=0.057$).

Conclusions: We conclude that tools developed in other populations perform less well than logistic regression performed in the population of interest, and may perform less well than individual risk factors like BMI. Neither QFractureScores nor FRAX performed better than BMI as a single risk factor, suggesting that simple case finding approaches may be appropriate.

4.2 INTRODUCTION

As the population ages, the number of women at risk of osteoporotic fracture continues to increase. This situation results in significant healthcare system costs associated with the disease. Published clinical studies have shown that early identification of at-risk women using BMD and clinical risk factors and intervention with appropriate pharmaceutical treatments can reduce the risk of hip fracture (Barr *et al.*, 2009; Johansson *et al.*, 2009). Early identification of at-risk individuals is, therefore, crucial.

The historical gold standard for evaluating risk of fracture is dual x-ray assessment (DXA); however, the low levels of bone mineral density required for a DXA diagnosis can occur years after the original onset of the condition when treatment would have been beneficial. In order to improve predictive power to identify women at risk of fracture, the World Health Organization (WHO) developed FRAX (World Health Organization Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK, 2010). This tool was developed using a meta-analysis of nine international cohorts with data on fractures and clinical risk factors in a large number of women.

Recently, another tool QFractureScores (Hippisley-Cox and Coupland, 2009) has become available as an online tool. This tool was developed using data from a single large UK-based cohort accessed through primary care physicians in the UK.

No US-based cohort data was used in the development of either risk calculator. Data available within the Nurses' Health Study, therefore, provides a unique opportunity to compare the predictive performance of the FRAX and QFractureScores in a US population. Previously, we have compared the performance of the two risk calculators in a retrospective study which has highlighted some differences in their performance between the two risk calculators (Chapter 3). While this study was useful in comparing some characteristics of the two calculators, a clear limitation in this study was the inability to compare the capacity of the two calculators to predict future fracture based on risk factors collected at baseline.

The primary objective of this study was to address this limitation by assessing the comparative performance of FRAX, QFractureScores and a logistic regression model developed specifically using the available dataset. The results were then evaluated to assess how well actual fractures were predicted. This study is original for two reasons. Firstly, no

predictive study using any of the bone fracture predictors in the Nurses' Health Study database has been conducted. Secondly, a comparison between these factors in a US-based longitudinal study has not been conducted over the observed timeframe of over 20 years.

4.3 METHODS

Subjects

The subjects were derived from the Nurses' Health Study. The Nurses' Health study is an on-going prospective study based in the United States at the Brigham and Women's Hospital and has been described previously in Chapter 2. A nested case-control group consisting of 164 women aged between 50 and 63, 82 hip fracture cases and 82 age-matched controls was selected for the analysis and derived from previous work carried out to evaluate a novel biomarker for fracture risk. There was an average time to fracture of 13 years in the group. A number of risk factors were available in the database:

age at toenail return, smoking status at toenail return, diagnosis of osteoporosis at toenail return, BMI at toenail return, months from baseline to case hip fracture, months of hormone use from toenail return to case hip fracture, diagnosis of osteoporosis at time of case hip fracture, year of hip fracture, month of hip fracture, alcohol consumption in grams/day, confirmed cases of rheumatoid arthritis, and steroid use data collected at multiple time points between 1994 and 2002. However, most hip fractures occurred before steroid use was reported. Only three cases and one control reported steroid use before the hip fracture. As a result, the decision was taken to omit steroid use from the analysis.

The following CRFs had multiple options:

Smoking status at toenail return

- 1=never
- 2=former
- 3=current

Alcohol in grams/day, which was measured in 1980 and 1984. 1980 was before toenail collection and 1984 was after toenail collection.

In the confirmed cases of rheumatoid arthritis, the options were to verify a subject was a case before toenail collection, or after toenail collection and before hip fracture.

Oral steroid use in 1994, 1996, 1998, 2000 and 2002 was recorded based on whether a subject had never used steroids, had used steroids, whether a hip fracture occurred before the date in question, or whether the data was missing.

Table 4.1 provides a comparison of the data available from the Nurses' Cohort Study with that required by the FRAX and QFractureScores algorithms.

Table 4.1: Clinical Risk Factor data available from the NHS Cohort

Clinical Risk Factor	FRAX	QFractureScores	NHS Cohort
Age	X	X	X
Sex	X	X	X
Weight	X	X	X
Height	X	X	X
Previous fracture	X		Exclusion factor
Parental hip	X	X	X
Smoking	X	X	X
Glucocorticoids	X		X
Rheumatoid arthritis	X	X	X
Secondary osteoporosis	X		
Alcohol Intake	X	X	X (grams)
Femoral neck BMD	X		
Asthma		X	
Steroids		X	X
Heart attack/Stroke		X	
History of Falls		X	
Chronic liver disease		X	
Tricyclic antidepressants		X	
Type 2 diabetes		X	
HRT		X	X
Endocrine problem		X	
Gastrointestinal Malabsorption		X	
Menopausal symptoms		X	X

There were differences between reported data type and required data entry type in a number of instances.

Incident fracture rate

The Nurses' Health Study has 415 incident fractures between 1986 and 1998 in a group of 61,200 women (Feskanich *et al.*, 2002b). This is a prevalence of 0.67% over the period. Since only hip fracture data was available, only hip fracture predictions were recorded from the risk calculators FRAX and QFracture.

Data Entry

QFracture

The data entered into the QFractureScores calculator was as shown in the table above. A number of conversions had to be made for particular options from the NHS data available to make them compatible with the online calculator. Smoking options in QFractureScores have five states: non-smoker, ex-smoker, light smoker (<10 per day), moderate smoker (10-19 per day) and heavy smoker (20+ per day). Since information on different smoking levels was not available, current smokers in the NHS cohort were all assumed to be moderate smokers.

For alcohol consumption, five options were also available: trivial (<1 day), light (1-2 per day), moderate (3-6 units per day), heavy (7-9 units per day) and very heavy (> 9 units per day). The alcohol consumption data was collected in grams of alcohol consumed per day on the Nurses' Health Study questionnaire. Since the QFractureScores questionnaire measures alcohol consumption in units of alcohol, a conversion rate of 8 grams per unit of alcohol has been used.

The QFractureScores questionnaire also required the weight and height for each woman rather than allowing BMI to be entered, so the height was back-calculated from a standard weight of 100 kg for the women using the following equation:

$$\text{BMI} = \left(\frac{\text{Weight in Kilograms}}{(\text{Height in centimetres}) \times (\text{Height in centimetres})} \right) \times 10,000$$

HRT was considered to be a relevant risk factor when taken for at least two periods (Hippisley-Cox and Coupland, 2009). The dose selected was CYCLICAL LOW DOSE NON-EQUINE, a mid-range risk factor in the algorithm.

FRAX

FRAX Web Version 3.2, US (Caucasian) was used to generate the risk data. This model was developed by adjusting the model for national epidemiological data on osteoporotic fracture (Kanis *et al.*, 2010). The data entered into the FRAX calculator was based entirely on yes/no answers. As a result, thresholds were set for the selection of positive and negative answers in FRAX. Any use of hormone or steroid treatment was considered a 'yes' for the relevant risk factors. The FRAX questionnaire also required the weight and height for each woman rather than allowing BMI to be entered, so height was back-calculated from a standard weight of 100 kg for the women using the equation shown above. In eighteen subjects, the calculated height exceeded 220 cm and then the calculation was made assuming a woman was 220 cm tall. Former smokers were considered alongside women who never smoked to be negative for smoking, as no allowance for former smokers is made in FRAX. The recorded hormone use in the Nurses' Health Study questionnaire was used as the input item in the glucorticosteroid risk factor component of the FRAX questionnaire. Based on the threshold of 3 units of alcohol per day, a threshold of 24 grams of alcohol per day was selected for a positive answer to the alcohol consumption question in FRAX in either 1980 or 1984.

Conditional logistic regression model

A prediction model based on the clinical risk factor data samples available in the nested case-control study was developed using entry-based conditional logistic regression. The conditional approach was selected to account for the age-matched cases and controls. The factors included in the model were BMI, smoking status, osteoporotic status, alcohol consumption in 1980 and 1984, and HRT use.

Missing data

Where data was unavailable, the risk factor was entered as a negative value to exclude the risk factor from the calculation. In one case, BMI was unknown and was chosen based on an average of the results in the dataset. There were 11 subjects that lacked alcohol consumption

data; for these subjects, consumption was assumed to be zero. Smoking data was missing for one subject and they were assumed to be a non-smoker in the analysis. Assuming the missing data in the negative serves to under-estimate the risks generated by the algorithms which, in the absence of the data, this is the most appropriate approach.

Data analysis

Statistical analysis was conducted using MedCalc software v11.5.1 and SPSS v19. Firstly, Kolmogorov-Smirnov was used to evaluate each risk factor for normality, the Student t-test was used to determine significance in normally distributed (parametric) data and Mann-Whitney was used in non-normally distributed (non-parametric) data, then the associations between the clinical risk factors themselves, and then fracture were evaluated.

The relative performance of the risk calculators, the regression model and the most important single clinical risk factor was evaluated. The predictive performance of the test was assessed using the AUC, and whether a particular model was significantly superior to the other models was also evaluated. In order to carry out this assessment, risks of fracture in excess of 3% were considered sufficient to identify a woman at risk of a hip fracture. A threshold of 3% was selected based on current US National Osteoporosis Foundation guidelines (NOF, 2008).

The mean predicted fracture rates were derived to enable comparison with the observed rates in the NHS and other studies.

In a subgroup analysis, the analyses were repeated in subjects with a fracture up to 16 years from baseline.

4.4 RESULTS

Table 4.2 describes the characteristics of the study population, there were 82 cases and 82 controls and six clinical risk factors were available for analysis. Table 4.3 describes the smoking characteristics in detail, this information was required by the QFracture algorithm.

Table 4.2: CRF data available for the study population

	Cases	Controls
	(n=82)	(n=82)
Age ^a , y (mean)	57.6	57.6
BMI, kg/m ² (mean)	24.4	25.6
Current Smoker (n)	28	22
Osteoporosis (n)	2	4
Thyroid Hormone User (mean)	9.73	17.12
Alcohol grams/day (mean)	5.88	7.24

Table 4.3: Smoking characteristics in the NHS population

	Never	Former	Current	Totals
Control	34	26	22	82
Case	39	15	28	82
	73	41	50	164

None of the nominal factors under consideration in the study were normally distributed, as shown in the Kolmogorov-Smirnov tests for normality in Table 4.4.

Table 4.4: Tests of Normality used to evaluate the key CRFs

	Kolmogorov-Smirnov (Lillefors Significance Correction)		
	Statistic	Df	Sig
BMI	0.135	164	0.00
HRTUSE	0.414	164	0.00
ALCOHOL80	0.280	164	0.00
ALCOHOL84	0.275	164	0.00

As a result, non-parametric tests were used to assess whether statistically significant differences between the means were present and none were observed for any of the risk factors. Whether there were correlations between these risk factors was also investigated using Spearman's Rho. Significant correlations were observed between the Age and HRT Use ($p=0.05$), Age and Alcohol consumption in 1980 ($p=0.03$), BMI and Alcohol consumption in 1980 ($p=0.02$) and Alcohol Consumption in 1980 and 1984 ($p<0.01$).

FRAX Model

The prediction results of the FRAX model were compared with the actual outcomes in the study. An AUC of 0.52 (0.44-0.59), $p=0.65$ was achieved. The AUC graph for the FRAX algorithm is shown in Figure 4.1

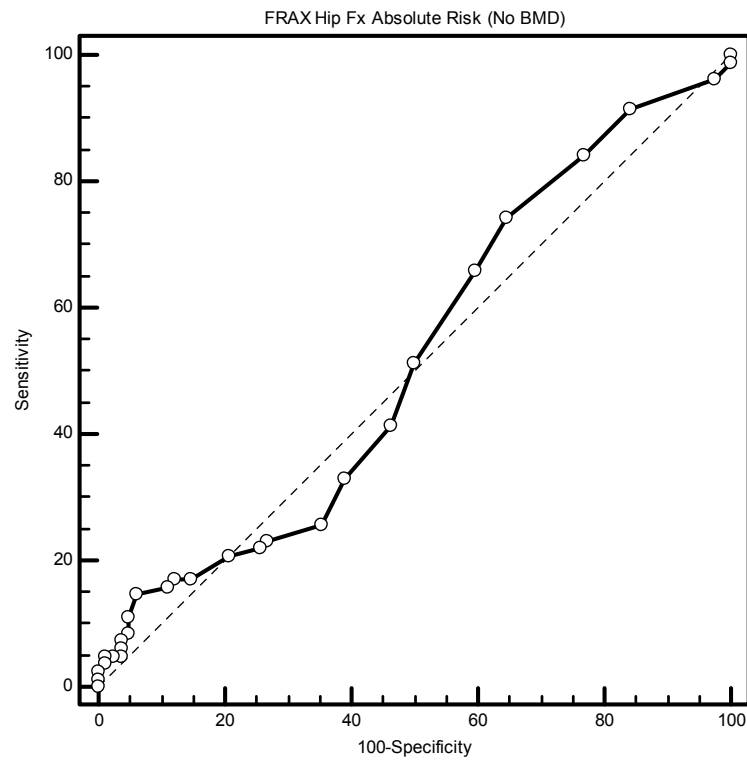


Figure 4.1: AUC graph for the FRAX algorithm.

QFractureScores Model

The prediction results of the QFractureScores model was compared with the actual outcomes in the study. An AUC of 0.554 (0.47-0.63) was generated, $p=0.23$, a graphical depiction of the model is shown in Figure 4.2.

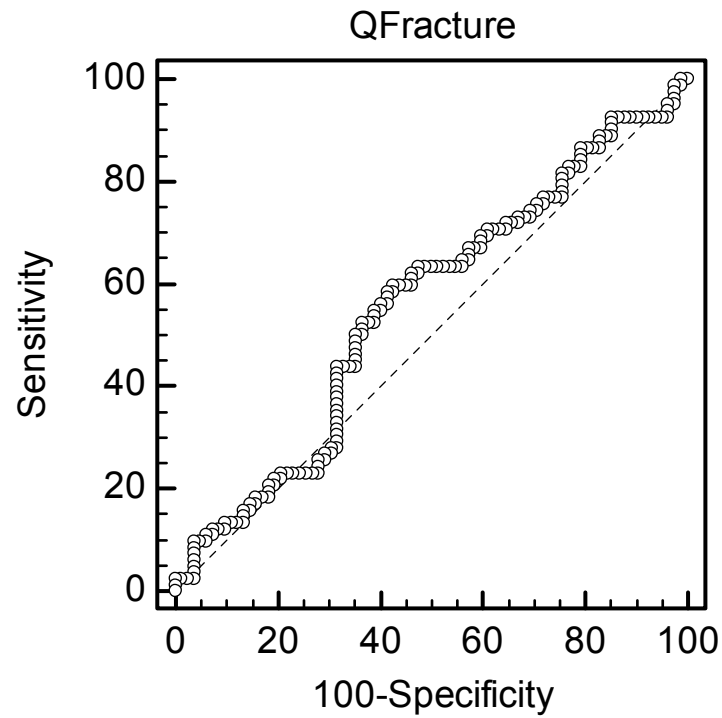


Figure 4.2: AUC graph for the QFractureScores Model

Body Mass Index

BMI was compared with actual fracture outcomes in the study, an AUC of 0.58(0.50-0.66), $p=0.057$ was achieved. Figure 4.3 depicts the AUC graph for the BMI risk factor.

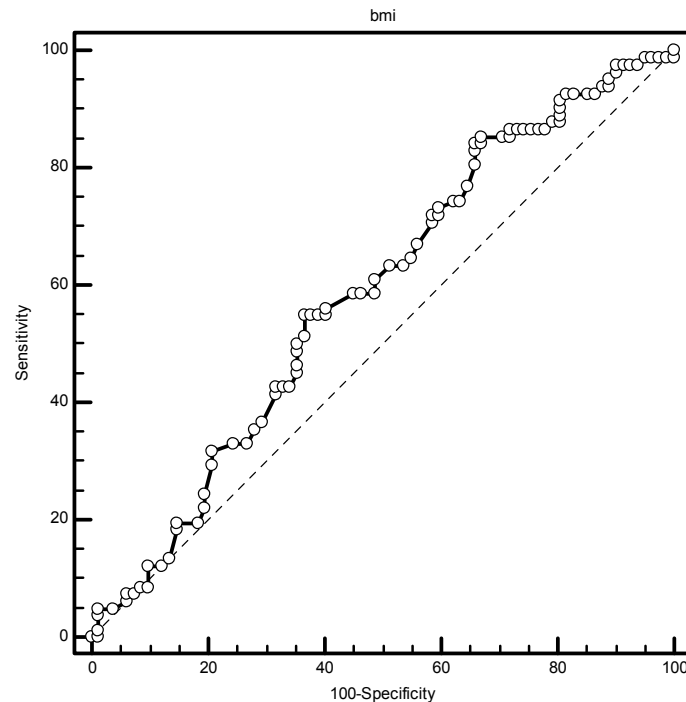


Figure 4.3: AUC graph for BMI risk factor

Conditional logistic regression

The logistic regression showed that none of the risk factors were significant explanatory factors in fracture risk, either by using the likelihood ratio tests or the parameter estimates. BMI came closest to significance ($p=0.07$) but did not achieve it. Former smokers were identified as a factor but, again, this was not significant. The R^2 show that less than 10% of the difference is explained by the factors. The Odds Ratio for BMI was 1.07 (0.994-1.16).

A combination model performed better than any individual factors achieving an AUC of 0.65 (0.57-0.72), but again did not achieve significance ($p=0.15$). The Hosmer and Lemeshow chi square for model fit was 11.90.

Comparison of ROC curves

In order to provide a comprehensive comparison of the Area under the ROC curves for the different diagnostic models a multiple ROC curve graph was plotted. It can be seen in the graph in Figure 4.4 and Table 4.5 that logistic regression had the best performance of the models but that none were statistically significant.

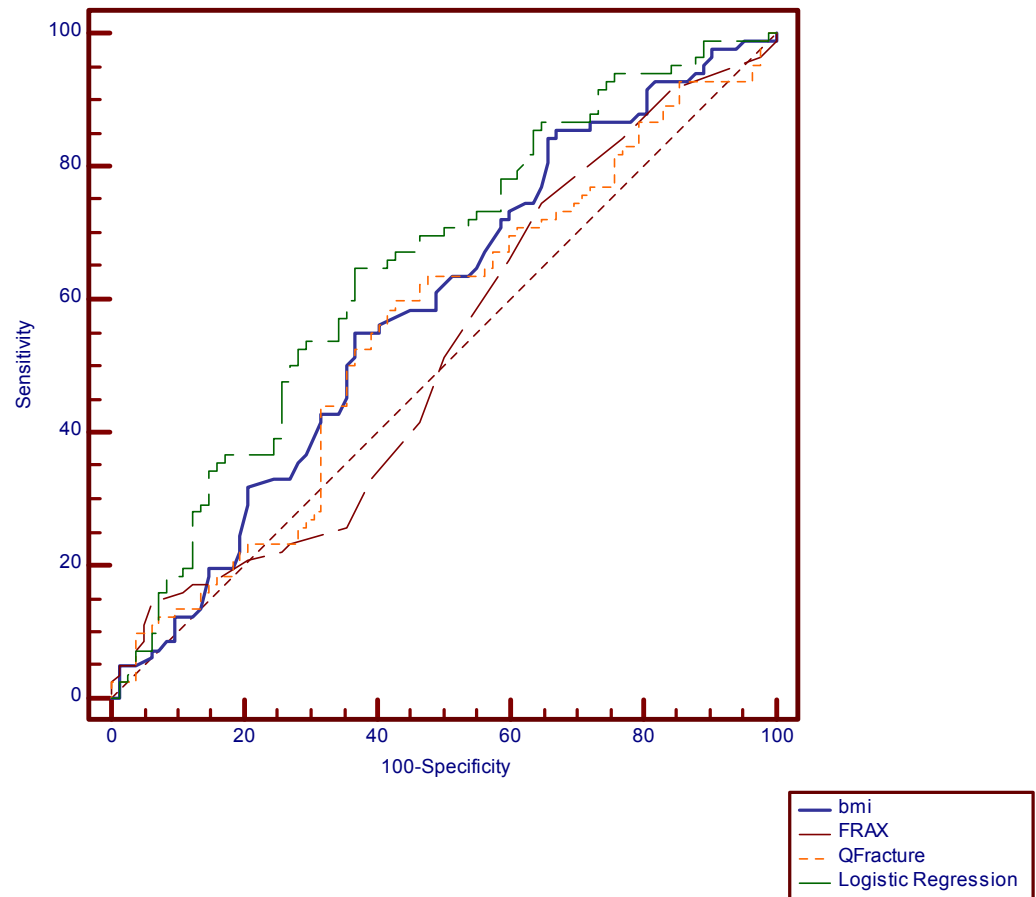


Figure 4.4: Comparison of ROC Curves for QFracture, FRAX, logistic regression and BMI in the NHS Cohort

Table 4.5: Summary of results for QFracture, FRAX, logistic regression and BMI in the NHS Cohort

Method	AUC	p-value
QFracture	0.554 (0.47 – 0.63)	0.23
FRAX	0.52 (0.44 – 0.60)	0.65
Logistic regression	0.65 (0.57 – 0.72)	0.15
BMI	0.585 (0.50 – 0.66)	0.057

In order to assess whether any of the methods were a significant improvement over another method, a pairwise comparison of the ROC curves was carried out.

Table 4.6: Pairwise comparison of four ROC curves developed in the NHS Cohort

Method 1	Method 2	P-value
BMI	FRAX	0.08
BMI	QFracture	0.47
BMI	Regression	0.12
QFracture	FRAX	0.32
FRAX	Regression	0.012*
QFracture	Regression	0.046*

These results, shown in Table 4.6, demonstrated that the conditional logistic regression model developed was significantly better than either FRAX or QFractureScores at predicting the fracture risk, but was not statistically significantly superior to BMI alone.

Predicted Fracture Rates

FRAX and QFractureScores produced the average fracture risk data presented in Table 4.7. The correlation between the two was found to be good, $R^2=0.67(0.57-0.74)$, $p<0.0001$.

Table 4.7: A comparison of predicted fracture rates in QFracture and FRAX models

	FRAX	QFracture
Highest value	5.7	3.23
Lowest value	0.1	0.1
Mean	0.96 (0.84 – 1.07)*	0.58 (0.52 – 0.64)*

***Non-normal**

Sub-group analysis

Subgroup analysis was conducted in 99 subjects with fractures which occurred up to 16 years after baseline. There were 50 cases and 49 controls in the subgroup, the results are depicted in Figure 4.5

The logistic regression produced the best prediction of the four models with a statistically significant AUC of 0.66 (0.56-0.76). In the logistic regression, BMI was a statistically significant risk factor, though the overall model fit was still not statistically significant, with the chi square being 9.4 ($p=0.30$). A summary of these results is shown in Table 4.8.

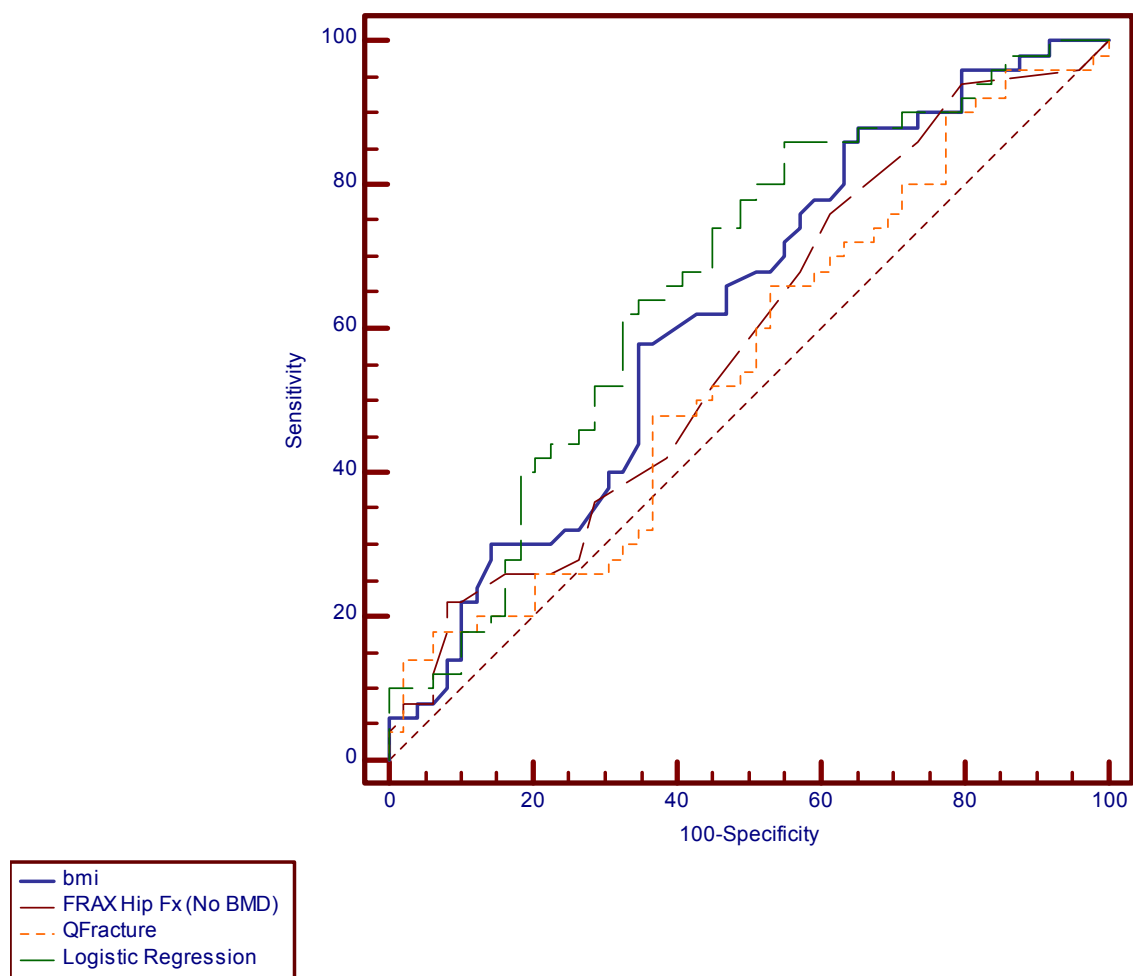


Figure 4.5: Comparison of ROC Curves in NHS subgroup of fracture up to 16 years from baseline

Table 4.8: AUC values for the NHS subgroup of fracture up to 16 years from baseline.

Method	AUC	p-value
QFracture	0.55 (0.45 – 0.65)	0.371
FRAX	0.58 (0.47-0.68)	0.16
Logistic regression	0.66 (0.56 – 0.76)	0.0025*
BMI	0.62 (0.52 – 0.72)	0.029*

All the models increased in AUC compared with the full group analysis. Pairwise analysis showed that none of the prognostic models were significantly superior to the others.

4.5 DISCUSSION

Introduction

Online risk calculators like FRAX and QFractureScores are becoming increasingly important for ascertaining absolute risk of fracture. These calculators take significantly different approaches to calculating absolute fracture risk. The QFractureScores based on a UK population uses a more detailed set of inputs and aims to provide a more accurate result based on the additional information provided when compared with FRAX.

Regression model development

The inclusion of non-statistically significant risk factors in a regression model has been debated in the literature (Steyerberg *et al.*, 2010). Steyerberg came out in support of its use and non-significant risk factors have been included in the developed regression model for these reasons. This would suggest that both BMI and smoking are relevant factors in long-term fracture risk. The advantage of Steyerberg's approach is that more risk factors are included which makes for a more clinically relevant model from a clinician's viewpoint and a model can be supported by the general body of literature; the downside is that the model is likely to be more vulnerable when used in a different population due to the number of variables in the model.

Predictive performance, AUC

The comparison made in this study provides evidence that the use of these risk calculators in new patient groups can lead to a reduction in performance compared with the level of

performance seen in the training and test cohorts used to develop the algorithms. In this study, the regression model specifically developed was found to be significantly better than either FRAX or QFracture. Additionally, neither FRAX nor QFractureScores offered significantly more information than BMI alone. The approach of combining risk factors appears to be valid in principle based on the superior performance of the regression model over the use of single risk factors. The regression model has effectively combined a number of risk factors to produce a prediction model with significantly better accuracy than that which is offered by either FRAX or QFractureScores. In sub-group analysis, this ultimately led to a statistically significant result being observed.

Other studies evaluating FRAX in national cohorts have reported similar findings. A recent Japanese study evaluating the local version of FRAX found that a logistic regression model of age, body weight and BMD produced similar AUC prediction rates to FRAX, 0.71 and 0.69 respectively (Tamaki *et al.*, 2011). The near significance of BMI ($p=0.057$) in this study would suggest that the Nurses' Health Study confirms the finding that BMI is a very important factor in fracture risk and one that outweighs many of the other incorporated factors. These AUCs in the Japanese study are better than those observed in the Nurses' Health Study data. It is reasonable to assume that the presence of BMD data in the FRAX calculations was an important factor in its superior performance.

The original QFractureScores paper reported an AUC for hip fracture in the test set of 89%; the results achieved in this study do not come close to this level of performance (Hippisley-Cox and Coupland, 2009). Additionally, in the previous study conducted among UK/Irish subjects, FRAX produced higher AUCs than QFracture. This may indicate that the country of origin is influential in the performance of the calculators.

Observed vs actual fracture rates

The Hippisley-Cox paper observed that the FRAX algorithm tended to overestimate fracture risk compared with their approach. The results presented here confirm this finding. QFractureScores was found to produce lower percentage risks than FRAX and were more in line with the actual results in the cohort.

A recent CaMOS cohort-based paper observed good correlations between the observed fracture rate in women and the predicted fracture rate, namely 12% (11.0%-12.9%) observed and 10.8% predicted (Fraser *et al.*, 2011).

The Tamaki study in Japan evaluating the local version of FRAX found that both predicted hip fracture and major osteoporotic event rates were not significantly different from observed rates and, therefore, had clinical utility.

In the QFractureScores and FRAX results, we observed significantly correlated mean predictions of fracture rate, 0.58% and 0.96% respectively. The QFractureScores figures are more in line with the actual fracture rate of 0.67% observed over the duration of the follow-up, which was 20 years rather than the expected 10 years in the cohorts used to develop the models. Over a 10-year time frame, QFractureScores appear to provide a more accurate prediction of the results in this cohort. They may be due to the larger number of younger women available in their study data.

Statistically significant AUCs may have been achieved with a larger sample size, but sub 60% values for all the results, excluding the regression model, would have limited clinical utility on their own.

Applicability of nested case-control approach

The 164 samples used in this study were nested case-control samples from a large cohort of samples in the Nurses' Health Study. Recent work in the diagnostic area has shown the applicability of using nested case-control approaches to estimate the diagnostic performance of tests in the wider population (Biesheuvel *et al.*, 2008). This study showed similar results for nested case-control approaches and the full cohort with differences in confidence intervals between the different designs. The variability decreased with increasing sample size. Increasing the sample size in a follow-up study in the Nurses' Health Study cohort might, therefore, result in FRAX and QFractureScores providing statistically significant results.

Sub-group analysis

All the prognostic models improved with the reduction in time frame used in sub-group analysis. In particular, BMI and the logistic regression model improved sufficiently to achieve statistically significant AUCs. The improvement in FRAX and QFractureScores performance was smaller, but FRAX now became a superior predictor to QFracture. Once again, the logistic regression model provided superior performance over BMI alone, but not significantly. This may be because the factors explained in the models are more influential in fractures over a shorter follow-up period.

Study limitations

The data collected in the NHS at baseline was decided before the major clinical studies that led to the development of FRAX and QFractureScores were carried out. This means that a number of relevant risk factors were not collected at baseline, which limits the ability to draw conclusions on the absolute performance of these prediction models based on this study. Data on steroid use were also collected in the study, but later on in the study, and the majority of fractures occurred prior to the collection of this data. Only three cases and one control reported steroid use before hip fracture, it has therefore been assumed that steroid use occurred at baseline. Data on rheumatoid arthritis status was also collected, but only 10% of the data was confirmed resulting in only four cases of rheumatoid arthritis being confirmed, which was too few to enable the data to be included. No DXA data was available for the cohort due to the era when it was collected; few DXA machines were available in the US, thus limiting the overall performance of FRAX in this study.

Other limitations in this study were the number of subjects available for analysis. The case-control approach utilised yields odds ratios, which have been assumed to provide similar results to relative risks. It is appropriate to consider odds ratios to be equivalent to relative risks due to the rare nature of the events under investigation (Steyerberg *et al.*, 2010). This approach and has become a standard tool for the analysis of case-control data within a cohort context.

The underlying equations used in the FRAX model are unknown. While the data for the QFractureScores and regression model were available, it was not possible to make a direct comparison of the influence of the risk factors between the three models.

Conclusions

This study indicates that FRAX and QFractureScores are not unequivocally superior to the use of a more limited number of risk factors. Their performance appears to decline in new populations compared with published performance values (Hippisley-Cox and Coupland, 2009) but this may be due to the limited data available from the Nurses' Health Study. The performance of the prognostic models in the Nurses' Health Study may have been influenced by the relative youth of the subjects and the length of the follow-up times. The average age of these women was 57; in FRAX, the majority of the studies evaluated comprised women in their 60s and 70s with shorter follow-up times. The improved performance in subgroup analysis shows the challenges of making long term (>10 year predictions) using questionnaire-based risk factors alone. It was only when the time to fracture limit was reduced that some of the prognostic models achieved statistical significance. The nested case-control approach is valid for the comparison of prognostic measures in a cohort study and can provide insights into the relative cost-effective performance of prognostic models. The study also confirms the finding that QFractureScores does not offer significantly superior performance over FRAX based on the additional data included in the QFractureScores model, as observed in Chapter 3, and can be considered to offer similar results.

It is possible to conclude that tools which are developed in other populations perform less well than logistic regression specifically developed and performed on the population of interest, and may perform less well than individual markers.

Chapter 5 : Advanced and novel methods for evaluating biomarker added prognostic performance in the Nurses' Health Study

5.1 ABSTRACT

Introduction: This chapter analyses the performance of a test based on Raman spectroscopy of nail samples to significantly add to the performance of clinical risk factors for osteoporotic fracture. Previous analytic work explored the performance of the biomarker using logistic regression, but this does not take time into consideration. Additionally, new work on alternative approaches to AUC for evaluating prognostic performance offered the opportunity for an alternative assessment of the performance of the new biomarker.

Methods: Data from the Nurses' Health Study and an analysis of the nail samples was used to conduct Kaplan-Meier and Cox's proportional hazard analysis. This compared the nail test alone, CRFs alone and the two in combination. Reclassification analysis was conducted to assess the improvement over CRFs introduced by the BQT.

Results: Body Mass Index as a risk factor was statistically significant using Kaplan-Meier analysis ($p=0.049$), while two other clinical risk factors, alcohol and use of hormone replacement therapy (HRTUSE) were not statistically significant. The nail test was a statistically significant predictor ($p=0.018$). The BQT hazard ratio (HR) of 3.2 is higher than the odds ratio of 1.59 for hip fracture found for using logistic regression. A CRF-only Cox model was statistically significant ($p=0.012$), while a combined CRFs and BQT model was also statistically significant ($p=0.008$). The Cook method showed that using the BQT model led to the reclassification of 24% of subjects. Using the Pepe method, an NRI of 12.1% was observed, compared with an AUC improvement between CRFs alone and BQT + CRFs of 8%. The improvement in NRI was not statistically significant ($p=0.07$).

Conclusions: Time is a significant factor in the prediction of fracture risk and the performance of the BQT benefits from the consideration of time as a measure. The introduction of new classification methods like NRI and IDI provide additional performance information over measures like AUC.

5.2 INTRODUCTION

This chapter analyses the performance of the BQT biomarker based on the Raman spectroscopy of nail samples and its ability to add significantly to the existing clinical risk factors available. It used a number of traditional and novel statistical methodologies, which may yield more information than considering fracture as a binary event without time as a factor or accounting for censored data.

The initial work (Chapter 2) evaluated the performance of the clinical risk factors and the BQT using binary logistic regression, which does not take time-based factors into consideration. Additional analysis using Kaplan-Meier and Cox's proportional hazards techniques could provide more information on whether considering time and censored data increases predictive performance in terms of time-to-fracture, as defined by avoiding a hip fracture rather than death. The clinical data available from the NHS includes time to fracture data, which can be used to demonstrate additional discrimination using these advanced statistical techniques.

There has been increased recognition in the academic literature in recent years that there is a need for additional measures to assess the performance of different prognostic risk factors and multivariate models beyond what is offered by the ROC curve. This is because the ROC curve has been observed to perform poorly as a measure of diagnostic performance in population-based cohorts in which the disease has a low prevalence (Cook, 2007). This is the situation in osteoporotic hip fractures where the prevalence is low, but the consequences for health are very serious. Additionally, some work by Keller *et al* indicated that there are limitations in the use of the ROC curve when the clinical utility of the test is being taken into consideration (Keller *et al.*, 2005). McClish also identified issues related to the ROC curve and suggested solutions to improve the clinical utility of the ROC curve by analysing a portion of the ROC curve (Katzman McClish, 1989). He criticises the full area under the ROC curve approach for equally weighting false-positive rates. His concern is that it may not reflect the clinical outcomes in a number of conditions.

Both papers identify issues with the analysis of the entire ROC curve to assess diagnostic performance, the primary one being that it is not representative of a number of clinical situations. The two papers propose methods to enable the analysis of portions of the ROC curve which help address the issue.

Nancy Cook of Harvard University has led the way by developing diagnostic measures that look beyond the ROC curve in recent years. In a landmark paper in 2007, she observed the following issue with the ROC curve: “a model which that assigns all cases a value of 0.52 and all controls a value of 0.51 would have perfect discrimination, although the probabilities it assigns may not be helpful. The actual predicted probabilities do matter, however, in clinical risk prediction models such as those commonly used for the assessment of global cardiovascular risk these factors are not considered” (Cook, 2007).

A number of researchers are now addressing this issue, including Cook, Pencina and Steyerberg. They are developing tools to move beyond whether a new prognostic test offers good discrimination as evaluated by the AUC and the other traditional factors, the Hosmer-Lemeshow test, Brier Score and Nagelkerke R^2 test.

The traditional measures evaluate the following in logistic models:

- Is the model well-calibrated to actual percentage outcomes? (Hosmer-Lemeshow test). Large chi-square values and significant p-values indicate poor fit.
- Is the model accurate on an individual basis based on the boundary outcomes from a particular cut-off? (Brier Score)
- What is the level of explained variability (Nagelkerke R^2 test)? Values in excess of 30% are preferred for a viable model.

Cook has noted that the addition of risk factors into prognostic models can result in small changes in AUC, which do not reflect the changes in risk category that result from the new information (Cook, 2008). She notes that many new biomarkers may have clinical relevant ORs (between 1.5 and 2.0) but which will have only a modest impact on the ROC curves. New reclassification metrics are able to address the weaknesses of these measures in terms of perfect discrimination using ROC curves. These authors argue that these measures are more important to prognostic models than the traditional ROC curve and c-statistic. These measures have been used to explore the performance of cardiac markers, but have only just started to be used to assess osteoporosis risk factors (Chan *et al.*, 2011).

5.2.1 The new measures

Cook developed the reclassification table measure by using a set cut-off value to assess whether a patient is at high, medium or low risk. These cut-offs are required for a prognostic model to have clinical utility based on its projected event rate. The table compares the numbers classified into each risk category before and after the addition of the new predictive factor into the prognostic model (Cook, 2008).

Pepe *et al.* (Pepe *et al.*, 2008) proposed a method based on comparing the number above and below treatment thresholds based on whether they are a case or non-case.

Pencina advanced the concept by considering the impact of the direction of the movement. Upward movement is considered positively in patients who had the outcome under investigation, and downward movement negatively; this can be summarised in a net reclassification index, or NRI measure (Pencina *et al.*, 2008). The authors also proposed a measure which integrates all the probability values, known as the Integrated Discrimination Improvement (IDI) measure. The advantage of the IDI measure is that it is independent of any cut-offs and, therefore, avoids any biasness that might arise from the cut-off.

The reclassification methods described are used in the following papers to evaluate prognostic models based on the addition of new markers (Steyerberg *et al.*, 2010; Cook, 2007; Cook, 2008; Cook, 2010). One clear advantage of these approaches is the introduction of time-to-event information, which is lacking in the traditional AUC measure.

An important factor is how the percentage risk of fracture is selected. New decision curve analysis tools are also being proposed in the literature (Vickers and Elkin, 2006) to address this question.

5.2.2 The potential impact of these new measures in osteoporosis

These measures are spreading beyond their original application in cardiovascular risk markers to other disease areas. One example of these methods being used on clinical data from an osteoporosis study (Donaldson *et al.*, 2011) is where they conclude these methods provide additional information over the c-statistic alone.

In order to assess the potential impact of these new reclassification approaches in the field of osteoporotic fracture prevention; a relevant reclassification level based on decision analysis is

required. The national organisations responsible for osteoporosis treatment guidelines have provided percentage risks above which intervention should be considered. In the US, the National Osteoporosis Foundation recommends action when the 10-year hip fracture risk exceeds 3%. In the UK, the National Osteoporosis Guideline Group (NOGG) has proposed a different approach, which links the relevant percentage risk of fracture for treatment to the age of the subject.

This study will build on the original binary logistic regression work developed in the NHS to create advanced models using Kaplan-Meier and Cox's PH approaches. The performance of the logistic regression and Cox's PH prediction model will be compared using the novel evaluation methods, using a risk factor alone and risk factor + nail Raman model and then evaluated.

5.3 METHODS

5.3.1 Nurses' Health Study

Data collected as part of a nested control study from the Nurses' Health Study cohort, the characteristics of which have been previously described, were used. As a retrospective case control study, the 82 age-matched controls would be censored in any time -to-fracture analysis, leaving 82 events to be analysed.

5.3.2 Models developed

Kaplan-Meier Analysis

Initially, Kaplan-Meier analysis was used to assess the following risk factors based on a pre-determined cut-off to separate the events into two groups. The cut-offs selected are detailed below:

- Nail Raman, BQT Score (Cut-off > 0.5)
- BMI (Cut-off >25): To assess BMI as a risk factor, a cut-off was set at BMI >25 compared with BMI < 25

- Alcohol consumption in 1980: A threshold of 24 grams of alcohol per day was used to separate the groups for the Kaplan-Meier analysis initially; additionally, the results were evaluated for any alcohol consumption
- HRTUSE: HRTUSE was evaluated based on “any use” and “no use” as the two groups.

Binary Logistic Regression Predictive model

Predictive values for each subject were generated using backward unconditional binary logistic regression for both models, one with CRFs alone and one with CRFs and Nail. The performance of the CRFs in the prediction model was weak, therefore the cut-off for inclusion into the model was raised to $p=0.1$ from the standard $p=0.05$ to enable more variables to be included in the final predictive model. These values were converted from relative risks into estimates of absolute risks by using the prevalence rate in the full cohort, which was 2.8% and the probability generated by SPSS using the equation, odds ratio = $1/(1-p)$.

Cox's PH model

A Cox's proportional hazards model was developed based on month of fracture post baseline data collection, which was available from the NHS dataset.

Statistical analysis

Logrank analysis was used to provide probability, ROC curve, Chi-squared and R^2 Nagelkerke values. SPSS v18 was used to calculate all results.

Treatment Thresholds for Reclassification

To evaluate the reclassification methodologies, we used the National Osteoporosis Foundation (USA) treatment guidelines. These recommend treatment when a subject has a hip fracture risk in excess of 3% since the data was derived from a US population.

Reclassification Methods

In order to evaluate the effect of reclassification, a number of comparisons were made:

- a) CRFs + BQT vs CRFs alone in Binary Logistic regression model: Cook Method(Cook, 2008)
- b) CRFs + BQT vs CRFs alone in Binary Logistic regression model: Pepe Method (Pepe *et al.*, 2008).

Binary logistic regression was used to develop the predictive values used in the NRI and IDI calculations. The NRI and IDI were calculated based on a comparison of CRFs alone and CRFs in combination with the BQT to assess the additional performance provided by the addition of the biomarker. No single function is available in the established statistical packages at this time to evaluate the NRI and IDI, so they must be calculated in a number of steps manually.

A reclassification table was then constructed based on a cut-off of 0% up to 3% absolute risk of fracture over the study period, and over 3% risk for high-risk patients, creating two groups. The approach used to calculate the NRI was, as Cook states, that “The NRI is the difference in proportions moving up and down among cases vs controls, or $NRI = [Pr(up | case) - Pr(down | case)] - [Pr(up | control) - Pr(down | control)]$.” The p-value was calculated from the test-statistic; this was calculated by dividing the NRI by its standard error. The standard error was calculated using the following equation:

$$SQRT ([Pr(up | case) + Pr(down | case)]/\#cases + [Pr(up | control) + Pr(down | control)]/\#controls).$$

$Pr(up | case)$ represents the number of cases which move up in classification when the new marker is added.

$Pr(down | case)$ represents the number of cases which down up in classification when the new marker is added.

$Pr(up | control)$ represents the number of controls which move up in classification when the new marker is added.

$Pr(down | control)$ represents the number of controls which down up in classification when the new marker is added.

#case represents the overall number of cases, and #controls represents the overall number of controls.

The IDI was calculated from the mean values of the probabilities of the two models, and is the difference in the estimated probabilities between the two models for cases and controls. In order to assess whether it was statistically significant, the standardised IDI was established by scaling the IDI by its standard error (SD of the mean). A table containing the difference in probabilities for each subject was created and compared against the outcome fracture to establish the standard error of the mean. Using the standard errors for cases and controls, the pooled estimate was calculated by squaring the value for each, adding them together and taking the square root. This pooled estimate was compared with the z-distribution in SPSS to obtain the 2-sided p-value. Relative IDI was calculated by dividing the estimated probabilities and subtracting 1; this allows the IDI to be expressed as a percentage.

5.4 RESULTS

5.4.1 Traditional Measures: Logistic Regression Model

In order to assess the performance of the logistic regression model its calibration and goodness-of-fit were assessed. The Hosmer-Lemeshow results for calibration assessment were a chi-square of 2.76, $p=0.948$, while the Nagelkerke R^2 goodness-of-fit score was 11% for a model containing all the CRFs and the nail Raman score.

5.4.2 Kaplan-Meier analysis of the cohort

Overall, there were 82 events which occurred over the time period, as shown in Figure 5.1. The Kaplan-Meier function allows the rate at which events have occurred to be seen visually.

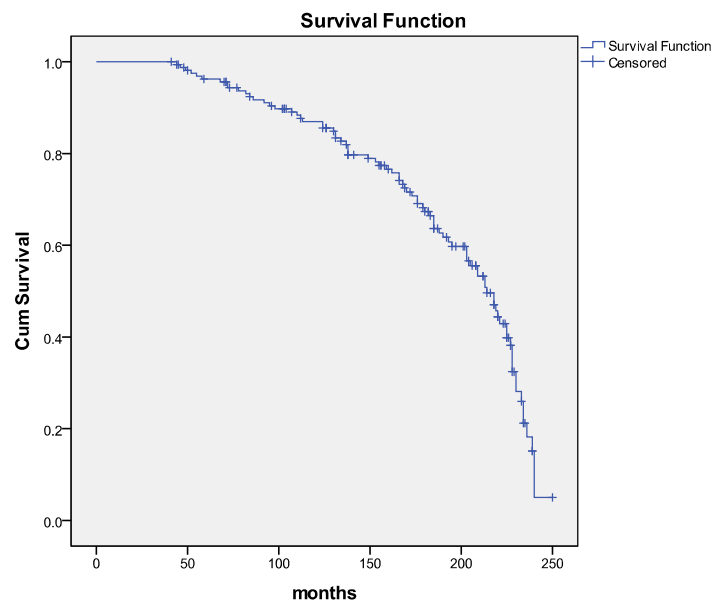


Figure 5.1: Kaplan-Meier Curve for the 82 events and 82 censored events

BMI Clinical Risk factor Time to event analysis

There were 27 events in women with a BMI over 25 out of a total of 63 women 57.1%. There were 55 events out of a total of 101 women with a BMI under 25, a percentage of 45.5%. The median time-to-fracture period for the women with a BMI over 25 was 228 months, 95% CI (217-238). For women with a BMI under 25, the median was 205 months (183-226). The logrank test had a chi-square of 3.874, which was statistically significant ($p=0.049$). Table 5.1 describes how the data was processed, the majority of censored subjects were in the

control group. Figure 5.2 depicts how the fracture events occurred over the time period. It can be seen that the cases represented by “1” have events more rapidly than the controls.

Table 5.1: BMI Case Summary for Kaplan-Meier Analysis

Case Processing Summary				
1	Total N	N of Events	Censored	
			N	Percent
.00	63	27	36	57.1%
1.00	101	55	46	45.5%
Overall	164	82	82	50.0%

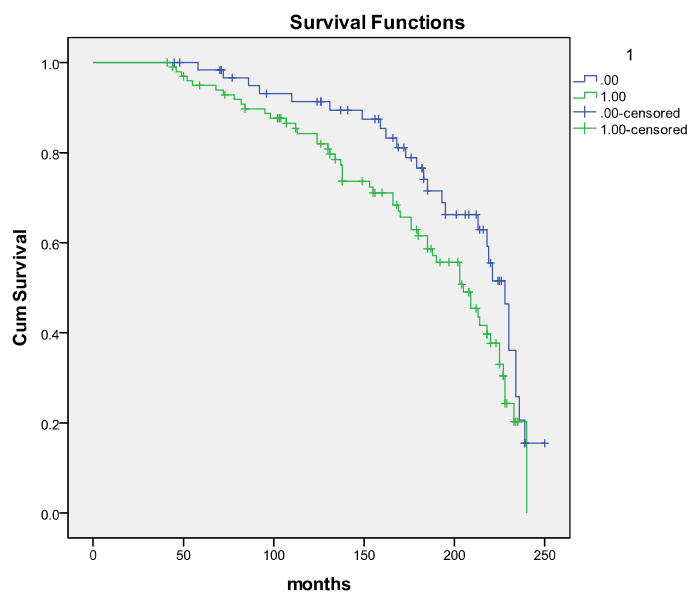


Figure 5.2: Event timing for those above and below a BMI of 25

Alcohol risk factor: Time to event analysis

Alcohol (> 24 grams per day) yielded only five events. The case processing summary shown in Table 5.2 shows that the majority of events occurred in those with consumption lower than the cut-off.

Table 5.2: Case summary for alcohol consumption (>24 grams)

Case Processing Summary				
alcoholbinary	Total N	N of Events	Censored	
			N	Percent
.00	153	77	76	49.7%
1.00	11	5	6	54.5%
Overall	164	82	82	50.0%

The effect of no alcohol use was also investigated, the results are shown in Table 5.3. There were 78 women who drank no alcohol and had 39 events (an incidence rate of 50%), while there were 86 women who drank any alcohol and had 43 events (an incidence rate of 50%).

Table 5.3: Case summary for Kaplan-Meier analysis of subject divided by Alcohol (any use) and no alcohol consumption

Case Processing Summary				
alcoholbinarytt	Total N	N of Events	Censored	
			N	Percent
.00	78	39	39	50.0%
1.00	86	43	43	50.0%
Overall	164	82	82	50.0%

The median time-to-fracture period for women who had no alcohol was 219 months 95%CI (197-240), while for women who drank alcohol it was 214 months 95%CI (198-229). The logrank test yielded a Chi-Square of 0.535, which was not statistically significant (p=0.465).

Age: Clinical risk factor: Time to event analysis

Age at baseline was evaluated as a risk factor based on being 58 years of age or over and under 58. No significant difference between the two groups was observed based on the logrank test ($p=0.61$).

Use of HRT: Time to event analysis

There were 40 subjects who used HRT within the group with 17 events, an incidence level of 42.5%. The median time-to-fracture period was 209 months 95%CI (191-226) was non-users, and 221 months 95%CI (212-229) for users, this is depicted graphically in figure 5.3. The logrank test provided a Chi-Square of 3.48 and a borderline significant result, ($p=0.062$).

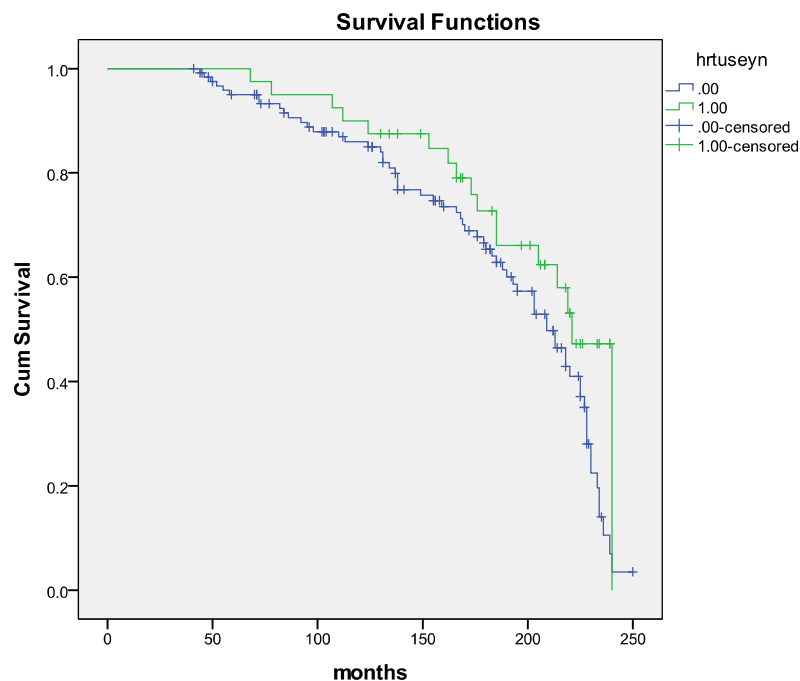


Figure 5.3: Time to Fracture based on dividing the study group based on HRT use

BQT Raman Score: Time to event analysis

Only 163 subjects were analysed because one subject did not have Raman data available. A cut-off was selected at 0.5, where scores could lie anywhere between 0.0 and 1.0. This led to a total of 63 subjects in the low risk group and 23 events, and 100 subjects in the high risk group and 59 events. The median time-to-fracture period in the low risk group was 228 months 95%CI (220-235) and 195 months in the high risk group (175-215) as shown in figure 5.4. The logrank test provided a Chi-Square test of 5.569 ($p=0.018$).

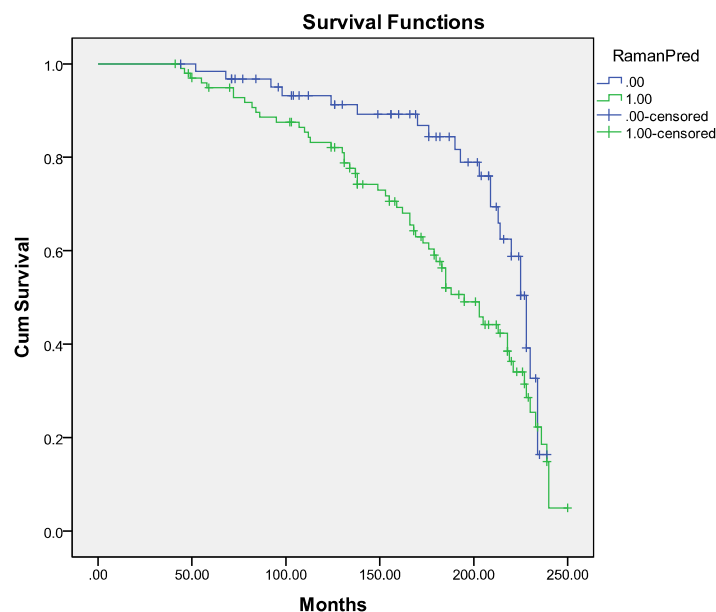


Figure 5.4: Time to fracture based on dividing the study group by Raman score

5.4.3 Cox's proportional hazards

Model 1: All Clinical Risk Factors

An analysis using all clinical risk factors, including the full dataset, was carried out. Using a backward conditional logistic regression analysis for model development, a Chi-Square of 8.78 was achieved, which was statistically significant ($p=0.012$). BMI ($p=0.045$) and HRT ($p=0.029$) were statistically significant and both offered a protective effect with increase shown by the hazard ratio. BMI yielded a HR of 0.959 95%CI (0.92-0.99) and HRT use yielded a HR of 0.91 95%CI (0.984-0.999). The hazard function is shown graphically in figure 5.5.

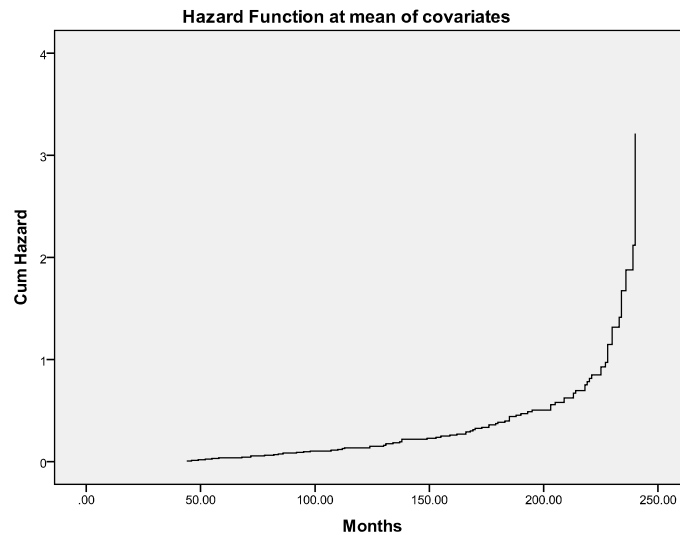


Figure 5.5: Hazard Function: Cox's model clinical risk factors in the NHS group

Model 2: BQT score + CRFs

Using a backward conditional logistic regression approach, a Chi-Square of 11.96 was achieved which was significant ($p=0.008$). BMI ($p=0.022$), HRT use ($p=0.028$) and Raman score ($p = 0.042$) were statistically significant. The Hazard Ratios were BMI: 0.948 (0.905-0.992), HRT use: 0.991 (0.983-0.999), Raman Score: 3.2 (1.043-9.95). The hazard function for the complete Cox model is shown in figure 5.6. Table 5.4 describes the associated chi-square and p-values for each component of the Cox model.

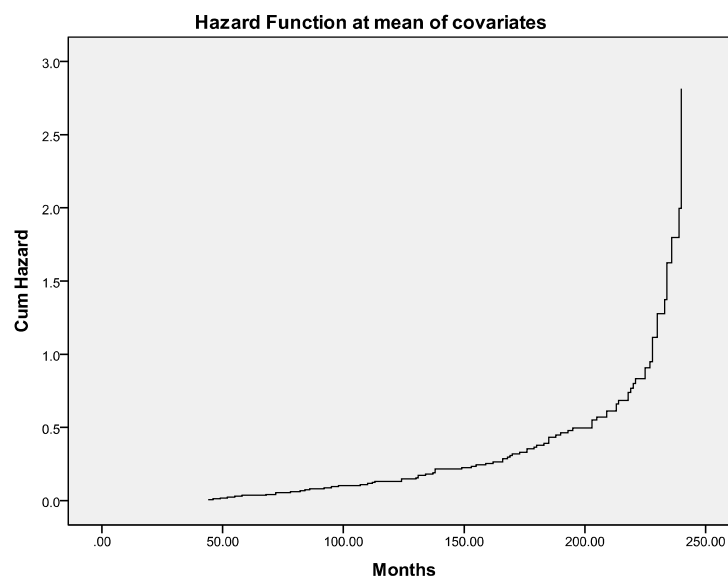


Figure 5.6: Hazard Function: Cox's model clinical risk factors

Table 5.4: Chi-Square and p-values for each function of the Cox model

	Chi-Square	P-value
BMI	3.874	0.049
Alcohol (any use)	0.535	0.465
HRT Use	3.48	0.062
Raman Score	5.57	0.018
All CRFs	8.78	0.012
BQT Score + CRFs	11.96	0.008

5.4.4 Reclassification methods

Binary Logistic Regression Model

The reclassification models were developed using absolute risks for each subject developed in models. These were created in the original Nurses' Health Study work and detailed in a previous chapter. Table 5.5 describes the Cook method, the numbers in BOLD and underlined represent the number of subjects who have been reclassified as a result of the addition of BQT to the CRF model. Table 5.6 depicts the Pepe method, again, the numbers in BOLD and underlined represent the number of subjects who have been reclassified by the addition of BQT to the CRF model. The Pepe method provides more information because it considers whether a subject is a case or not and therefore allows different values of outcome to be attached to the cases and controls.

Table 5.5: Cook Method

	CRFs + BQT		% Reclassified
CRFs	0– 3%	>3%	
0-3%			
No of subjects	68	<u>26</u>	
% classified by CRFs + BQT	72%	28%	28%
>3%			
No of subjects	<u>16</u>	51	
% classified by CRFs + BQT	24%	76%	24%

Table 5.6: Pepe Method

	CRFs + BQT				Total
	Cases		Non-cases		
CRFs	0– 3%	>3%	0– 3%	>3%	
0-3%	23	<u>17</u>	45	<u>9</u>	94
>3%	<u>7</u>	35	<u>9</u>	17	68
Total	30	52	54	26	162

Net Reclassification Index

The Net Reclassification Index (NRI) which provides a percentage value for the benefit provided by the addition of the BQT is calculated by subtracting the number of incorrectly reclassified cases from the number of correctly reclassified cases and the number of misclassified reclassified non-cases from the number of correctly reclassified non-cases. Using the standard error and the Z-statistic a p-value can also be assigned. The standard error (standard deviation of the mean) is calculated separately for cases and controls and then one is subtracted from the other. The p-value can be calculated from the z-statistic in SPSS.

$$\text{NRI} = (17/82 - 7/82) - (9/80 - 9/80) = 12.1\%$$

Calculating the p-value for the NRI

Standard Error = 0.07998

Z statistic = $0.121/0.07998 = 1.512$

P = 0.07

Integrated Discrimination Improvement

The IDI is calculated using the difference in risk probabilities rather than categories. The average risk of new model cases is subtracted from the average probability of old model cases. This value is then subtracted from the difference in the average probability of new model controls and old model controls. A p-value can be calculated for the IDI by using the following steps; the standard error (standard deviation of the mean) is calculated using SPSS and a standardised IDI is calculated by dividing the IDI by the standard error. This p-value can be calculated by comparing the standardised IDI result with the z-distribution in SPSS.

IDI = (av prob new model cases – av prob old model cases) – (av prob new model control – av prob old model controls)

= $(0.049 - 0.044) - (0.027 - 0.028) = 0.005 + 0.001 = 0.006$

Calculating the p-value

Standard Error = 0.02613

Standardised IDI = $0.006/0.02613 = 0.2296$

P-value = 0.41

Relative IDI = (av prob new model cases – av prob old model cases)/ (av prob new model control – av prob old model controls) – 1

= $(0.049 - 0.044)/(0.027-0.028) - 1 = 4\%$.

Reclassification Results Summary

The Cook method showed that using the BQT model led to the reclassification of 24% of subjects. Using the Pepe method, a NRI of 12.1% was observed, compared with an AUC improvement between CRFs alone and BQT + CRFs of 8%. The improvement in NRI and

IDI were not statistically significant ($p=0.07$) and 0.39 respectively. Table 5.7 provides a summary of the NRI, IDI, Relative IDI and AUC improvement which results from the addition of BQT to the CRF model.

Table 5.7: Results summary of the improvement gain from adding BQT to the CRF model.

	NRI	IDI	Relative IDI	AUC Improvement
Logistic Regression	12.1% ($p=0.07$)	0.006 ($p=0.39$)	4%	8% ($p=0.002$)

5.5 DISCUSSION

5.5.1 Traditional measures

The traditional measures provide information on the statistical validity of the models, but have limited value as clinical utility evaluation tools. The results using the Hosmer-Lemeshow test show that the model incorporating all clinical risk factors was a good fit, but that the model only explained a limited amount of the variability based on the Nagelkerke R^2 result. These results conflict each other in terms of the overall performance of the model, but the Hosmer-Lemeshow test could be considered more relevant based on knowledge of the performance of the test using other evaluation measures.

5.5.2 Kaplan-Meier analysis

High BMI was shown to have a statistically significant protective effect compared with low BMI, which is consistent with the results seen in other studies of BMI in osteoporotic fracture (LaFleur *et al.*, 2008). The reduction in time to fracture of more than a year may be shown to have a health economic, as well as a quality of life, impact in a study which considers these factors.

The results for alcohol consumption, whether limited at the FRAX mandated cut-off or assuming no alcohol consumption at all, did not result in a statistically significant result; this

is despite the fact that it is a risk factor strong enough to justify inclusion in the FRAX risk assessment tool (Kanis *et al.*, 2009). The reasons for the discrepancy are unclear, but possibly the large number of people who drink no alcohol in the investigated population could be a factor.

HRT use has been shown to be protective with a borderline significant result. This is consistent with work done in the UK by Hippisley-Cox (Hippisley-Cox and Coupland, 2009).

The performance of the BQT score based on a cut-off was very good, being $p=0.018$, the best of any of the risk factors assessed. This is concurrent with the logistic regression analysis, which had a p-value of 0.004 and also shows that the BQT score has a time-based factor in its predictive ability.

The introduction of time has strengthened the statistical power of the study, with significant results for a number of risk factors. This widened the understanding of the predictive ability of the BQT to include time-to-fracture as a factor which can be evaluated.

5.5.3 Cox's Proportional Hazards analysis

Using Cox's proportional hazards analysis has changed the variables selected as significant in the model and has improved the model's predictive power compared with the results achieved with conditional logistic regression previously developed. The HR of 3.2 for the BQT score would indicate that the biomarker is a strong risk factor for fracture risk when time is considered, compared with the odds ratio of 1.59 observed in the logistic regression. Another osteoporosis study in women of this age range has shown HRs of between 1.61 and 2.25 for ultrasound and DXA to predict hip fracture up to 10 years in advance (Stewart *et al.*, 2006).

The proportional hazards model with the BQT score included performed better than the model without it, with a Chi-squared 11.96 vs 8.78.

The BQT HR of 3.2 is higher than the odds ratio of 1.59 (95% CI 1.15-2.21) for hip fracture found for the nail Raman score using logistic regression. This suggests that time is a useful factor in the prediction of fracture risk, when possible hazard ratios should be used for analysis.

A limitation of this study was the use of age-matched case-control data, meaning that age could not be evaluated as a risk factor in this study.

This study shows that, whenever possible, time-based data should be used to analyse a prognostic indicator because time-to-event can significantly improve the predictive power of a model.

5.5.4 Reclassification Methods

Reclassification methods were used to evaluate whether the NRI would demonstrate an improvement in the performance of the marker when the BQT is added to the predictive logistic regression model. The increase in NRI (12.1%) when the BQT was added into the binary logistic model shows that the BQT had a positive impact on classification of at-risk patients. This value was, however, not significant but could potentially become significant with a larger sample size. The increase in NRI was larger than the increase in AUC, which was an increase of 8%, from 66% to 72%. Pencina *et al.* have stated that the p-values in NRI assessments are of limited value and that the magnitude of the improvement in the context of the clinical environment is of more significance. On this basis, more work is required to evaluate the importance of the NRI in the context of the clinical need and cost considerations. They proposed a weighted NRI approach to consider cost considerations (Pencina *et al.*, 2011).

The IDI was 0.006 and not statistically significant. This indicates that the improvement brought by BQT was not a statistically significant addition based on this measure. The IDI values are difficult to understand in their raw format. The relative IDI values provide a clearer value, which can be related to other measures of added predictive ability. The AUC increase was larger than the relative IDI increase of 4%.

At this time, methods to calculate NRI and IDI for Cox's proportional hazards models with censored data are in the process of development (Pencina *et al.*, 2011; Steyerberg and Pencina, 2010). The challenge with applying the current methods to censored data is that all subjects are assumed to be present for the entire duration of the study, which is incorrect in time-to-fracture analysis.

When these new methods are developed fully it will be interesting to evaluate the performance of the Cox's models with and without BQT data and also with the logistic regression models to enhance classification.

The reclassification methods rely heavily on the selection of appropriate cut-offs by clinical experts for the test. For example, the National Osteoporosis Foundation uses a cut-off of 3%, while the UK-based National Osteoporosis Guideline Group uses an age-based approach which increases the cut-off required for treatment significantly with age. The current literature does not address the question as to whether a version of these techniques could be developed to enable optimal cut-offs to be selected for classification to maximise reclassification, or whether that would have an impact on the overall clinical utility of a test. The IDI does not rely on cut-offs and an approach which generates an IDI curve may aid the selection of cut-offs. Steyerberg *et al.* suggest that decision analytics tools are used (Steyerberg and Pencina, 2010).

5.6 CONCLUSIONS

Time is a significant factor in the prediction of fracture risk for the BQT measure. This has been shown both using Kaplan-Meier and Cox's proportional hazards censored approaches.

The new reclassification methods have distinct advantages over AUC and other traditional methods because they allow increased discrimination in a manner which is clinically relevant, based on clinically-assigned cut-offs. This allows assessment of the measures in a clinical context in a way that is not practical with AUCs.

This study shows that the use of Raman spectroscopy of the nail has increased clinical utility for the prediction of fracture risk using the new reclassification measures, compared with the more established methodologies of AUC, and the other traditional measures. The results were not statistically significant, indicating that further work to improve the performance of the BQT would be beneficial.

The use of time-based data shows improved performance over non-time influenced data. This shows that the new methods demonstrate the clinical utility of BQT analysis and that prospective studies are required to fully demonstrate the performance of BQT to predict osteoporosis.

The introduction of new measures, like NRI and IDI, is the start of a process that will lead to the closer integration of statistical and clinical measures in the evaluation of new prognostic tools. The true value of the performance numbers can only be truly assessed in the context of clinical practice and costs.

Additional work to evaluate these results in the context of the health economic environment is required to assess the true clinical utility of the biomarker. This may also lead to methods which use NRI and IDI to actually help select the optimal cut-off for whether a patient should be selected for treatment.

Chapter 6 : Traditional and alternative health economic measures to evaluate the cost-effectiveness of osteoporotic prognostic biomarkers

6.1. ABSTRACT

Introduction: When evaluating a new prognostic test, it is important to consider its impact from a health economic viewpoint by modelling its cost-effectiveness; this approach is applied in this chapter to osteoporotic biomarkers and clinical risk models. This chapter explores different screening strategies and modelling techniques using patient data from the Nurses' Health Study and performance data from DXA, BQT and clinical risk factors (CRFs).

Methods: In order to explore the cost-effectiveness of DXA, BQT and CRFs relative to each other, three different evaluations were undertaken. First, a decision tree model, and then a Markov model were developed to explore their relative performance; finally, a model based on weighted NRI, a potential alternative health economic measure, was developed. Two different BQT cut-offs were evaluated to explore whether a high specificity or high sensitivity strategy was beneficial.

Results: BQT low cut-off strategies were consistently on the cost-effective frontier for the Markov and decision tree models. In some scenarios, sensitivity analysis showed the results were highly sensitive to PPV in the decision tree analysis. NNT to prevent a fracture was in a range between 29 and 510, depending on the treatment scenario. Weighted NRI provided an ICER of \$10,975. The Markov model identified that DXA was on the cost-effective frontier and produced higher effectiveness levels than the decision tree model.

Conclusions: The consistency of the relative positioning of the BQT low cut-off strategy in the different model approaches show that it is the most viable strategy. The ability of the weighted NRI to consider the effects of reclassification adds another dimension to cost-effectiveness evaluation when considering incorporating new biomarkers into a prognostic model. The consideration of drug and diagnostic cost, sensitivity, specificity and predictive performance is essential when assessing the clinical usefulness of the addition of new markers into prognostic models and are useful approaches during the development of new markers.

6.2. INTRODUCTION

In order to evaluate the potential usefulness of the BQT to reduce the costs of fracture in the population, a health economic analysis is required. The National Institute of Clinical Excellence (NICE) has recently issued a Diagnostics Assessment Programme Manual to bring a consistent approach to the health economic evaluation of diagnostics (National Institute for Health and Clinical Evidence, December 2011). The tests for osteoporosis used in clinical evaluation today are either laboratory or imaging tests and they can be used for diagnosis, monitoring, screening, prognosis or a combination of these approaches. While interim outcome measures like DXA T-scores can be used to evaluate health economics, the ideal outcome measure is the clinical endpoint, namely bone fracture. A key component in establishing the characteristics of the test's accuracy and a common feature used in health economic models is the AUC. Recently-published health economic studies mostly use odds ratio, relative risk or AUC to evaluate the economic performance of osteoporosis tests (Schwenkglenks and Lippuner, 2007). It is clear from recent work in the cardiovascular field that the limited impact of additive tests to multivariate models on AUC measures limits the ability of clinical practitioners to evaluate the cost-effectiveness of using different tests in combination. AUC measures do not take the reclassification of subjects into account (Cook, 2010). This issue has led to the proposal of a new measure of prognostic performance when new markers are added to pre-existing prognostic models: the net reclassification index (NRI) (Cook, 2007; Pencina *et al.*, 2008). The authors of these papers state that these new measures can be used to evaluate whether an expensive new biomarker should be introduced, but the methodologies to determine this have not yet been published in detail. The Pencina *et al.* paper offers a methodology to evaluate cost-effectiveness, which weights the NRI based on the cost saving when a person moves up in classification compared to incurred costs when they move down in classification to the non-case arm caused by misclassification. For example, when a person no longer receives unnecessary treatment as a result of reclassification, there is a cost saving; they do not, however, offer a practical example in their paper.

This chapter evaluates the cost-effectiveness of a new prognostic test for osteoporosis. The BQT[®] is shown in comparison and in conjunction with established tests for the condition within the framework of the NICE guidelines using three different modelling approaches, to

assess whether clinicians would benefit from using an alternative test to current established technologies.

Work in previous chapters has established that there is a time-based component to the BQT. Cox's proportional hazard-based analysis led to superior predictive results compared with logistic regression analysis. This time aspect was evaluated for cost-effectiveness by comparing a non-time influenced modelling approach, decision tree analysis with a time influenced approach, and Markov analysis.

6.3 METHODS

6.3.1 Nurses' Health Study

As per the NICE guidelines, a scoping exercise was undertaken to establish the parameters of the economic evaluation study.

The BQT test was developed for screening and prognosis applications to identify women at risk of fracture who are asymptomatic. Other tests widely used for this application are DXA and CRF-based models, such as FRAX[®]. These additional tests were selected for inclusion in the study, while QUS was rejected for inclusion due to the similarity of the measure to DXA and its lower use in clinical practice. All tests used solely for diagnostic or monitoring applications, like blood-based tests and biopsies, were rejected. The features of the three tests, DXA, BQT[®] in combination with CRFs and a CRF-based model, have been described in previous chapters.

The evaluation focussed on post-menopausal women up to the age of 69. This limitation was imposed by the dataset available for evaluation, namely the Nurses' Health Study (Feskanich *et al.*, 2002b). This population age is considered relevant for preventative treatment in scenarios when long-term follow-up is possible (Kanis, 2007).

NICE guidelines recommend the use of quality-of-life measures as the final outcome. Since this particular study was intended to evaluate the relative prognostic performance of tests in the same clinical area, the decision was made to use fracture, the relevant clinical outcome as the measure of interest. In order to explore whether the cost-effectiveness of DXA, BQT and

CRFs, relative to each other, was influenced by the modelling methodology, three different economic evaluations were undertaken. First, a decision tree model, then a Markov model, and finally a new model based on weighted NRI, a potential alternative health economic measure. DXA screening was assumed to be the standard of care. The modelling analysis was carried out using TreeAge Pro 2011 release 2.0 (TreeAge Software, Williamstown, MA, USA).

Preventative Strategies

Data from a real cohort, the Nurses' Health Study, was utilised in the study. A detailed description of this cohort is provided in Chapter 2.

The following screening strategies were evaluated:

- DXA based on measures derived from Kanis *et al.* (Kanis, 2007; Kanis *et al.*, 2009)
- BQT + CRFs based on NHS Study with a high BQT[®] cut-off and fewer people treated
- BQT+CRFs based on the NHS study with a low BQT[®] cut-off and more people treated
- CRFs alone based on NHS study.

In this study, we considered the DXA-based approach to be the current standard of care. The alternative strategies were evaluated over the 23 years of the Nurses' Health Study follow-up period, a long enough time horizon to ensure that a substantial number of relevant clinical events will have occurred and which exceeds the standard 10-year time horizon used in osteoporosis studies. Due to the novel nature of the BQT[®], it was felt during the scoping stage that incorporating two tests with alternative cut-offs would be beneficial in order to explore the importance of sensitivity and specificity; a high cut-off was selected to maximise specificity, while a low cut-off was selected to increase sensitivity.

The modelling software was also used to evaluate whether the tests could perform better in a number of combinations. The timing of when the test was to be taken was not evaluated because the available dataset only had one collection point at study commencement. Also, the subsequent care pathway was not modified based on the diagnostic method because data was not available on potential alternative care pathways driven by diagnostic type.

Treatment criteria

The US NOF foundation treatment guidelines recommend treatment when a fracture risk in excess of 3% over a ten-year period exists. This criterion was used to select patients for treatment (National Osteoporosis Foundation, 2008). This value was assumed to apply over the full 23-year period of the study.

Denosumab was selected as the treatment of choice due to its best-in-class efficacy, which would offer the best impact on patients identified as high risk (Cummings SR *et al.*, 2008).

Cost data

Fracture cost and long-term care costs resulting from hip fracture were derived from the Mayo Cost Data Warehouse (Gabriel *et al.*, 2002). This study demonstrates the good data integrity associated with the dataset and it was necessary to use US cost data with clinical samples from a US study. The paper identified direct medical costs for subjects prior to hip fracture of \$761 per annum; after hip fracture, the median costs went up to \$11,756 for fracture cases. These figures are for 2002 and a 30% increase was included to allow for inflation over the past decade; a value of \$14,293 was therefore used.

The current Medicare reimbursement for DXA-based bone densitometry is \$172.88 based on the national global average for Medicare code 77080, while the cost of a physician's visit is \$97.81 based on national global average reimbursement to Medicare code 99213. These figures were taken from the CodeMap website (www.codemap.com). The cost of assessing CRFs is assumed to be the cost of a physician's visit at \$97.81 and the cost of running the BQT[®] is assumed to be \$100 in the study.

Preventative treatment with Denosumab is assumed to cost €425 (\$552) per year based on the Jonsson *et al.* paper (Jonsson *et al.*, 2011). Based on treatment covering 23 years, this is a cost of \$12,696, substantially more than the cost of diagnosis using DXA or the other treatment modalities.

Non-medical and indirect costs have been excluded from this model.

6.3.2 Traditional modelling methodologies

The specific methods used to develop each of the three modelling methodologies are now described.

Decision Tree Model

Parameters of the model

The hip fracture prevalence in the modelled cohort of 2.8% was derived from the Nurses' Health Study dataset. This incidence rate used in the model was not increased as the patients aged.

Test accuracy data in terms of sensitivity and specificity have been taken from Kanis *et al.* for the DXA from the NHS data set for the BQT[®] option (Kanis, 2007; Kanis *et al.*, 2009).

Denosumab (Prolia) is modelled as offering a risk reduction of 40% in hip fracture patients based on data in the FREEDOM trial (Cummings *et al.*, 2009). The costs of the drug were assumed to run for all 23 years, for each subject identified for treatment, making a total cost of \$12,696. The decision tree was modelled with fracture costs taken into consideration.

In the NHS treatment model developed using a logistic regression model, 50% of women were found to be above the 3% risk threshold for treatment using BQT[®] and CRFs combined with a low cut-off, 10% were above the risk threshold for treatment in the combined model with a high cut-off, and 37% of women were found to be above the threshold for treatment using CRFs alone. The prevalence of osteoporosis as defined by DXA in women between the ages of 60 and 69 is 22%; this figure was used to identify women for treatment in the DXA arm. The treatment thresholds used for DXA + CRFs and BQT was, therefore, different, as in the case in clinical practice.

Table 6.1 provides a summary of the costs and other parameters used in the model, the costs were derived from a number of sources as shown.

Table 6.1: Table of parameters used in base case and their ranges used in the decision tree model

Parameter	Value	Range	Data Source
Incidence Rate	2.8%	NA	NHS Study
Hip Fracture Cost	\$14,293	\$7,146 - \$28,587	Gabriel <i>et al.</i>
DXA Cost	\$172.88	\$86.44 - \$345.76	Codemap.com
Physician visit cost	\$97.81	\$48.9 - \$195.62	Codemap.com
BQT Cost	\$100	\$50 - \$150	Assumption
Denosumab Cost	\$12,696		Jonsson <i>et al.</i>
Denosumab risk reduction	0.60	0.2 – 0.8	Cummings <i>et al.</i>
DXA RR per SD	1.54	1.34 – 1.75	Stewart <i>et al.</i>
BQT + CRF RR per SD	2.50	1.69 – 3.7	NHS data
BQT + CRF RR per SD (High cut-off)	2.32	0.95 – 5.66	NHS data
CRF RR per SD	2.06	1.4 – 3.03	NHS data
Discount rate	3%	0 – 5%	Nagata-Kobayashi <i>et al.</i> , NICE guidelines
DXA Treatment Prevalence	22%		Kanis <i>et al.</i>
BQT Treatment Prevalence	50%		NHS Data
CRF Treatment Prevalence	37%		NHS data
DXA Sensitivity	30%		Kanis <i>et al.</i>
DXA Specificity	80%		Kanis <i>et al.</i>
BQT Sensitivity	72%		NHS data
BQT Specificity	52%		NHS data
BQT Sensitivity: High cut-off	20.7%		NHS data
BQT Specificity: High Cut-off	90%		NHS data
CRF Sensitivity	12%		Kanis <i>et al.</i>
CRF Specificity	95%		Kanis <i>et al.</i>

Decision Tree Development

The decision tree for each strategy was developed using two branches:

1. Treat or not treat based on the threshold criteria for each diagnostic test.
2. The True Positive, False Positive, True Negative and False Negative values for each test were calculated using the following equations:

- Treatment Arm

- Positive Predictive Value; PPV evaluates the percentage of subjects with positive test results correctly diagnosed.
- 1-PPV evaluates the percentage of subjects with positive test results incorrectly diagnosed.
- $PPV = (sensitivity) * (prevalence) / ((sensitivity) * (prevalence) + (1 - specificity) * (1 - prevalence))$.

- Non-Treatment Arm

- Negative Predictive Value; NPV evaluates the percentage of subjects with negative test results correctly diagnosed.
- 1-NPV evaluates the percentage of subjects with positive test results incorrectly diagnosed.
- $NPV = (specificity) * (1 - prevalence) / ((specificity) * (1 - prevalence) + (1 - sensitivity) * (prevalence))$.

Table 6.2 shows the sensitivity, specificity, positive predictive values and negative predictive values for the four different diagnostic strategies, DXA, BQT with a high value cut-off, BQT with a low value cut-off and CRFs.

Table 6.2: Base case predictive value data for the four diagnostic strategies

	Sensitivity	Specificity	PPV	NPV
DXA	30	80	0.04	0.98
BQT low cut-off	72	52	0.04	0.98
BQT high cut-off	20	90	0.06	0.98
CRFs	12	95	0.06	0.97

Calculation of incremental cost effectiveness ratio (ICER)

The interventions were considered mutually exclusive, meaning that implementing one strategy will affect or prevent the implementation of other strategies. Incremental cost effectiveness ratios (ICERs) were therefore used to evaluate the results. The results were first ranked based on their costs and ICERs were then calculated. The output of this analysis allows the different approaches to be compared using marginal cost effectiveness. The cost-effectiveness measure being evaluated was intervention cost per hip fracture prevented.

Effectiveness data

The effectiveness of each testing methodology is measured in terms of their ability to minimise the number of fractures which occur. This can be measured using relative risk reduction (RRR), absolute risk reduction (ARR), and/or number needed to treat (NNT). The ARR was calculated according to the change in risk between high risk patients and low risk patients based on multiplying the RRR by the prevalence rate. NNT was calculated as $(1/ARR)$.

The RRR used to evaluate the effect of the outcome was calculated based on a combination of the relative risk, drug-mitigating effect when treated, and standard deviation based on whether they were in the high risk or low risk group. The RRR was calculated by multiplying the risk of a subject in the relevant arm by their ability to reduce their risk by treatment intervention. Subjects who were not treated, therefore, did not benefit from the intervention.

A subject identified as a true positive or false negative was given a risk based on a multiple of the RR of the test and the standard risk of 2.8%. A subject who was considered to be low risk, i.e. a true negative or false positive, was given the standard risk minus the increased risk of the high risk subject, thereby modelling the level of risk in the normal population.

The following relative risks were utilised to determine the effectiveness of the intervention:

- DXA: 1.54 (1.34-1.75) per SD (Stewart *et al.*, 2006)
- CRFs alone: 2.06 (1.4-3.03)
- BQT + CRFs: 2.50 (1.69-3.70) Low cut-off
- BQT + CRFs: 2.32 (0.95-5.66) High cut-off.

The Stewart *et al.* paper was selected for the DXA performance data because it provides data on the performance of DXA over a long time frame in women at the menopause, which is a close approximation of the condition in the NHS (Stewart *et al.*, 2006).

In order to incorporate the data into the model, different high and low risk values were required for each test methodology. These were developed as shown below:

$$RRR = (RR \text{ per SD}) \times SD \times \text{Treatment effect}$$

A high risk subject was considered to have a risk 2.5 times that of a low risk patient.

Table 6.3 shows the relative risk per SD for the high risk and low risk patients in each of the four diagnostic strategies.

Table 6.3: Relative Risks per SD for the four diagnostic strategies for high risk and low risk patients

	Arm	RR per SD	SD	Treatment effect	Effect:RRR
DXA: high risk	Treatment	1.54	2.5	40% reduction	1.54
DXA: low risk	Treatment	1.54	0.40	40% reduction	0.246
BQT+CRF: high risk	Treatment	2.50	2.5	40% reduction	2.50
BQT+CRF: low risk	Treatment	2.50	0.40	40% reduction	0.4
BQT+CRF High cut-off: high risk	Treatment	2.32	2.5	40% reduction	2.32
BQT+CRF High cut-off: low risk	Treatment	2.32	0.40	40% reduction	0.368
CRF: high risk	Treatment	2.06	2.5	40% reduction	2.06
CRF: low risk	Treatment	2.06	0.40	40% reduction	0.329

Table 6.4 provides information on the relative risks used for the DXA, BQT+CRFs and CRF strategies for high risk and low risk subjects and also explains in which diagnostic scenario (FP, TP, FN, TN) they were used.

Table 6.4: Relative Risk factors used in the different diagnostic scenarios

	Status	DXA	BQT/CRFs	CRF
High Relative Risk (2.5SD)	FN, TP	3.85	6.25	5.15
RR		1.54 (1.34-1.75)	2.5 (1.69-3.70)	2.06 (1.4-3.03)
Low Relative Risk	TN, FP	0.64	0.4	0.48

Figure 6.1 provides a visual representation of the decision tree, there are four arms for each of the treatment scenarios. Each treatment arm has a treat or no treat option which then results in either a true positive fracture case who has correctly received treatment, a false positive non-fracture case who has incorrectly received treatment, a true negative non-fracture case who has correctly not received treatment or a false negative fracture case who has incorrectly not received treatment.

Sensitivity analysis

A one-way sensitivity analysis was conducted to assess the strength of the results. The impact of a change in the cost and performance of Denosumab was assessed by changing the value 50% in each direction, while a change in the PPV and NPV performance of the BQT and CRFs was assessed using a 50% change in each direction. Variations in sensitivity and specificity were not modelled because alternative sensitivity and specificity scenarios were selected for the BQT[®] and alternative relative risks were used in the models.

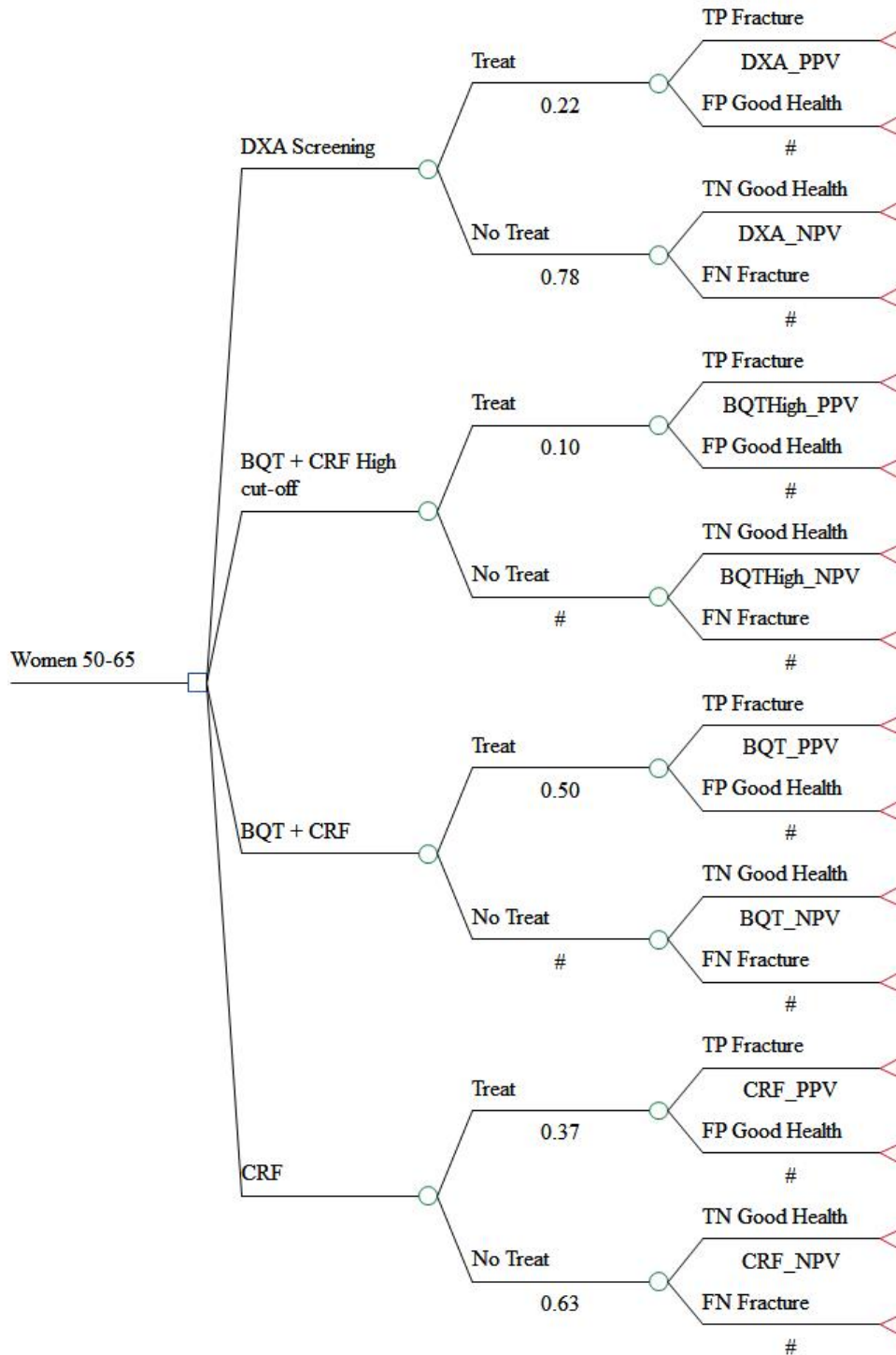


Figure 6.1: Decision tree diagram for the four diagnostic strategies

Markov model development

A Markov cost-effectiveness model was developed to compare the cost-effectiveness of the four strategies using a time-to-event based approach. The Markov model was based on the decision tree previously developed.

Base case

The analysis was based on a cohort of healthy postmenopausal women, starting at 58 years of age with no history of fracture and continuing for 23 years as per the real cohort. The Markov decision tree is shown in figure 6.2, the two health states selected were no-fracture (Well) and post-hip fracture. All women start in the no-fracture (Well) state and then transition into the post-fracture (Fracture) state once a fracture has occurred, or they continue in the Well state.



Figure 6.2: Markov decision tree with two states well and fracture

Transition Probabilities

The same percentages for treatment thresholds were used as the decision tree model to select subjects for treatment. The PPV and NPV of each test used in the decision tree analysis was used to determine the probability split between true positives and false positives and between true negatives and false negatives for the Markov state transitions. An incidence rate of 0.12% was used to represent the annual incidence rate based on the 2.8% incidence of fractures over the entire 23-year period of the study. In the initial state of the analysis, all members of the cohort were assumed to be well. The PPVs and NPVs are previously shown in Table 6.2. A cycle length of one year was used and a stopping rule was set up after 23 years. Half-cycle correction has been used to account for the fact that fractures occur continuously over time; if fractures were only accounted for at the end of each system, it would overestimate the survival rate. Half-cycle correction is a mathematical which shifts all the completed Markov cycles $\frac{1}{2}$ a cycle to the right to reduce the discrepancy caused by the use of the assumption that the events occur at the end of the cycle to simplify the software

analysis; this correction is automatically carried out in the analytical software when the option is selected once the analytical run has been completed. It is recommended when using Markov models to evaluate continuous events accurately.

Cost-effectiveness data

The same costs as detailed in the decision tree model were used, these are shown again in table 6.5 as they were used in the Markov model. The effectiveness was measured based on number of hip fractures prevented. A discount rate of 3% was used for the transition costs, which is within the range of the NICE guidelines. The additional costs shown in the table below were also required to build the model.

Table 6.5: Cost data used at the different states of the Markov model

Parameter	Initial State	Transition	Final State
Physician Visit	\$97.81	0	0
DXA	\$172.88	0	0
CRFs alone	97.81	0	0
BQT	\$100	0	0
Denosumab Cost	0	\$552	0
Hip Fracture Cost	0	\$14,293	0
Treatment Scenario (TP)	Physician visit + diagnostic cost	\$552	0
No Treatment Scenario (FN)	Physician visit + diagnostic cost	0	0

Outcome effectiveness data

The same approach was taken as in the decision tree model; a relative risk-reduction strategy in high risk patients was the focus of the treatment strategy. The same effectiveness data was used from the decision tree analysis, as shown in Table 6.3, except for the BQT effectiveness,

which was increased to reflect the increased performance of the BQT when time was considered as a factor. The hazard ratio for the BQT was 3.2 (1.043-9.95) without CRFs; a 10% sacrifice in predictive power was assumed to raise the cut-off to increase specificity to 90%, reducing its value to 2.88 HR per SD for the alternative scenario. The calculated values for hazard ratios for the high risk and low risk cases for the four diagnostic scenarios are shown in table 6.6.

Table 6.6: Diagnostic test hazard ratios for the high risk and low risk cases in the four diagnostic scenarios

	Arm	HR per SD	SD	Treatment effect	Effect:RRR
DXA: high risk	Treatment	1.54	2.5	40% reduction	1.54
DXA: low risk	Treatment	1.54	0.40	40% reduction	0.246
BQT: high risk	Treatment	3.2	2.5	40% reduction	3.2
BQT: low risk	Treatment	3.2	0.40	40% reduction	0.512
BQT High cut-off: high risk	Treatment	2.88	2.5	40% reduction	2.88
BQT High cut-off: low risk	Treatment	2.88	0.40	40% reduction	0.46
CRF: high risk	Treatment	2.06	2.5	40% reduction	2.06
CRF: low risk	Treatment	2.06	0.40	40% reduction	0.3296

Cost-effectiveness frontier

In order to analyse the performance of the different treatment strategies, the results were displayed graphically using cost on the y-axis and effectiveness on the x-axis. In health

economic analysis a line is drawn between the options which are the most cost effective, this line forms the cost-effectiveness frontier. Options to the left of this line are less effective and either the same cost or more costly, options to the right of this line are more effective and either the same cost or less costly. This graphical approach helps assess the potential of alternative approaches.

Figure 6.3 provides a graphical illustration of the Markov model. There are four arms to the model, for each arm there is a fracture arm which is terminal, once a subject reaches the fracture arm they leave the study. There is a treatment arm which in each annual cycle leads to a true positive fracture at which point they leave the study or a false positive good health result which will move them onto another cycle in the study. In the no treatment arm, a true negative result which does not lead to a fracture will result in the subject continuing in the study, a false negative result which leads to a fracture will result in the subject leaving the study.

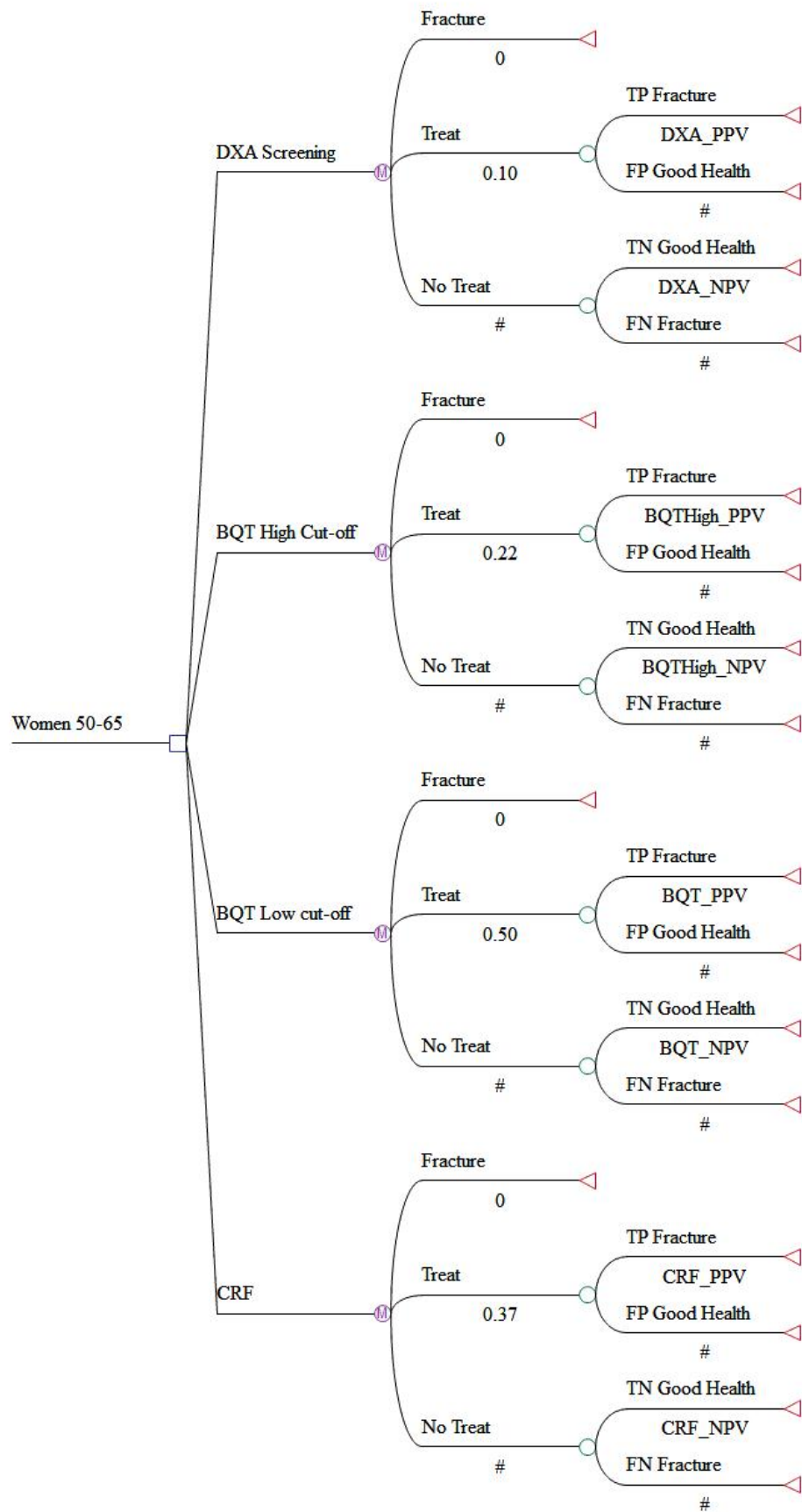


Figure 6.3: Markov Model Tree

6.3.3 Novel approaches

Using reclassification data from the Nurses' Health Study and appropriate cost data, an assessment of the incremental cost-effectiveness for the addition of the BQT[®] over a CRF-only strategy was conducted using the weighted NRI approach.

The details of this measure are described in Chapter 5. The equation of interest is:

$$\text{SQRT} \left(\frac{[\text{Pr}(\text{up} \mid \text{case}) + \text{Pr}(\text{down} \mid \text{case})]/\# \text{cases} + [\text{Pr}(\text{up} \mid \text{control}) + \text{Pr}(\text{down} \mid \text{control})]/\# \text{controls}}{2} \right)$$

This allows the benefit of patients moving up and down in classification relative to the previous model to be calculated.

Weighted costs

This new cost-effectiveness measure proposed by Pencina *et al.* based on calculating weighted costs is a 'cost per patient correctly classified' measure. The measure evaluates the relationship between the cost of the new test and the NRI and is calculated by dividing the NRI by the cost of the new test. The new measure takes account of the number of patients that are newly-correctly classified and assigns a cost to it, as well as the patients falsely classified and assigns a cost to them.

Using the approach suggested by Pencina *et al.* of a weighted value, two values are needed:

- S₁: Saving associated with upward classification of a person who was going to fracture
- S₂: Saving associated with downward classification of a person who was not going to fracture

$$\text{Total savings} = S_1 * ([\text{Pr}(\text{up} \mid \text{case}) + \text{Pr}(\text{down} \mid \text{case})] + S_2 * [\text{Pr}(\text{up} \mid \text{control}) + \text{Pr}(\text{down} \mid \text{control})])$$

A table of reclassification was developed to obtain the numbers used in the equation.

6.4 RESULTS

6.4.1 Decision Tree Analysis

Figure 6.4 shows that DXA screening is dominated by a combination of the BQT with a high cut-off and a clinical risk factor-based approach. The two BQT scenarios and CRFs form the cost-effectiveness frontier for testing.

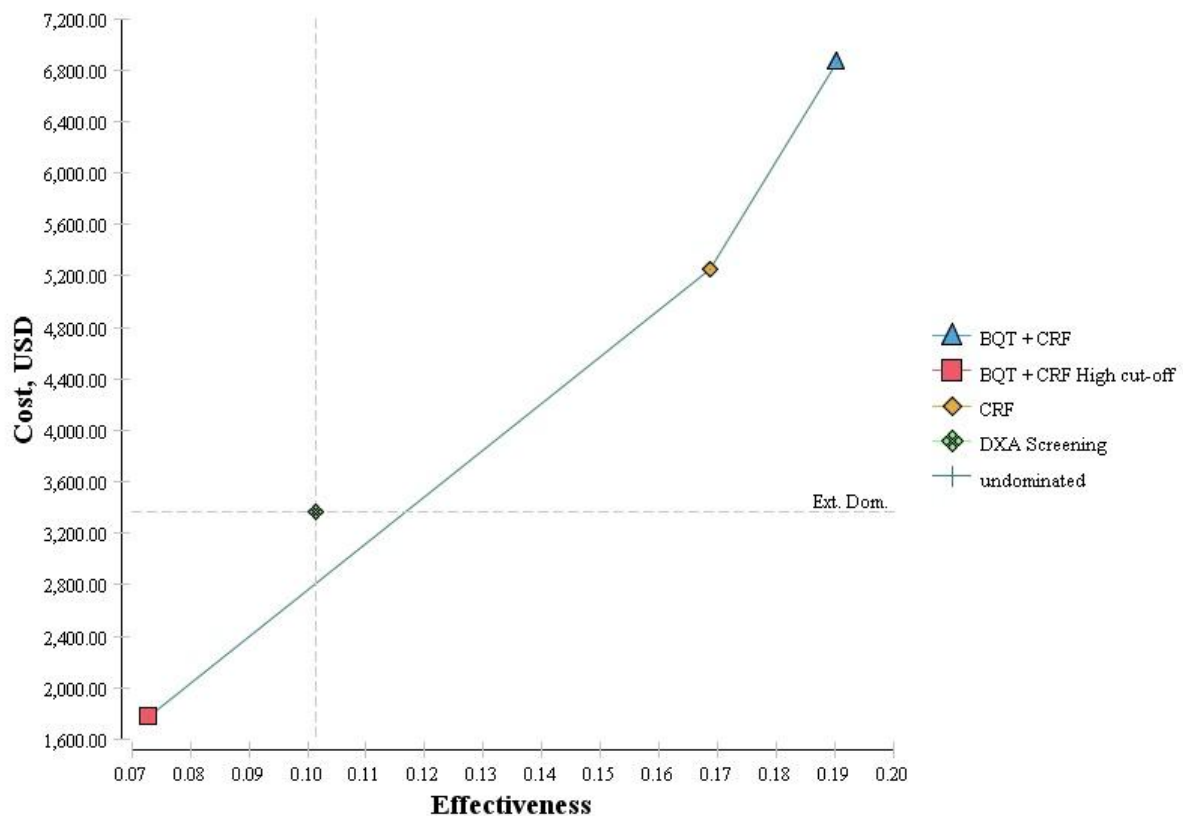


Figure 6.4: Cost-effectiveness chart: Decision Tree: Four diagnostic scenarios are shown.

The results of the decision tree analysis are shown in tabular form in table 6.7, the combined BQT and CRF strategy using a high cut-off was the lowest cost, the only dominated strategy was DXA. DXA is shown separately since it was not included in the cost-effective frontier. Table 6.8 shows the resulting RRR, ARR and NNT for the four scenarios, a combined BQT and CRF strategy with a low cut-off had the highest risk reduction (0.19) and the lowest number needed to treat to prevent a fracture at 187 patients.

Table 6.7: Cost-effectiveness results table for the four diagnostic options

Rank by cost	Strategy	Cost	Eff	Incr Cost	Incr Eff	ICER	Comment
1	BQT High	\$1,776.14	0.07	0	0	0	Undominated
3	CRF	\$5,225.85	0.17	3,479.71	0.10	36,248.36	Undominated
4	BQT Low	\$6,860.26	0.19	1,604.41	0.02	74,847.71	Undominated
2	DXA	\$3,362.25	0.10	1,586.11	0.03	55,340.71	Ext Dominated

Table 6.8: RRR, ARR and NNT results for the four diagnostic scenarios

	RRR	ARR	NNT
BQT High	0.07	0.00196	510
DXA	0.10	0.0028	357
CRF	0.17	0.00476	210
BQT Low	0.19	0.00534	187

A number of sensitivity analyses were carried out. BQT cost-sensitivity was analysed in ten increments between \$50 and \$150, while the overall positioning of the tests was unaffected by costs in this range. The BQT PPV was also analysed, as shown in figure 6.5; a PPV in excess of 0.06 resulted in removing a CRF-only strategy from the cost-effective frontier, leaving the best strategy to be based solely on selecting an appropriate cut-off for the

BQT+CRFs test. Further increasing the BQT PPV above six would result in the CRF-only strategy moving further and further away from the cost-effectiveness frontier.

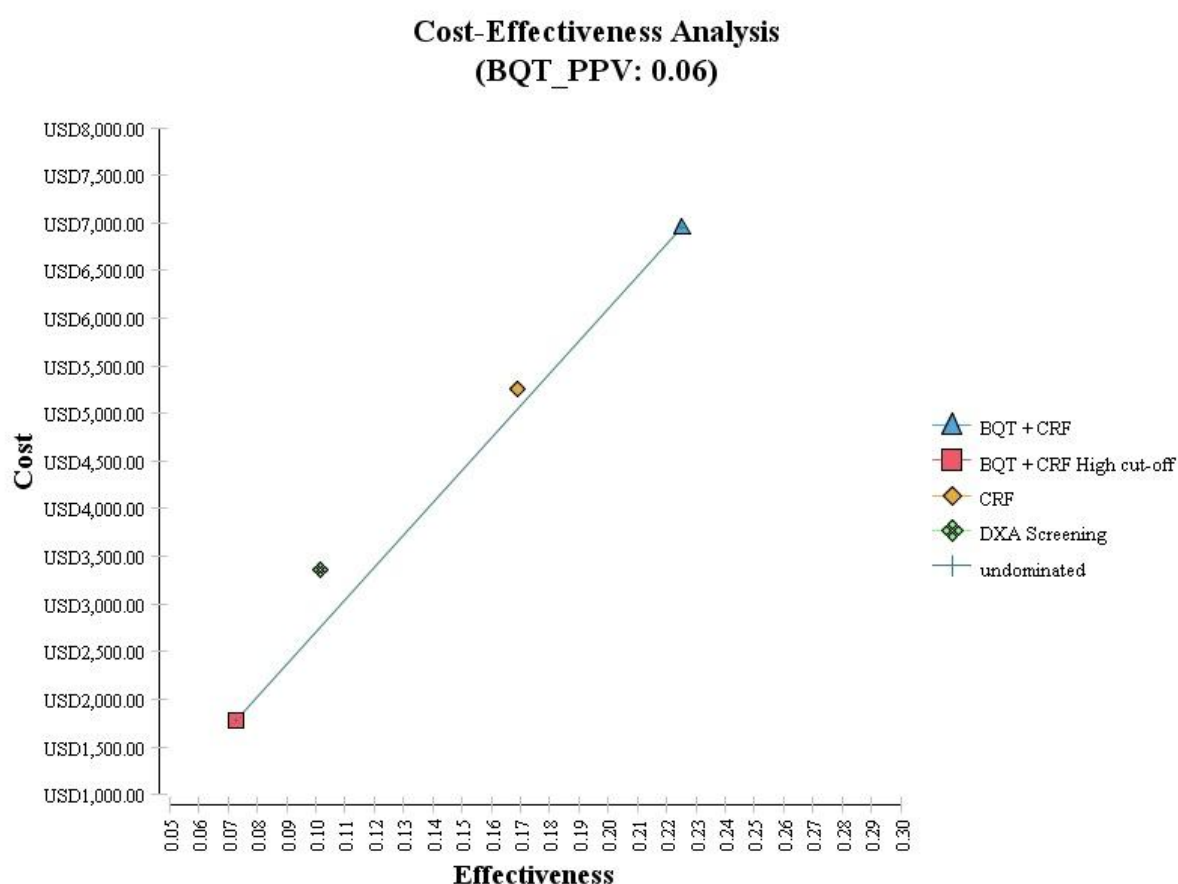


Figure 6.5: BQT PPV Sensitivity Analysis: Increasing PPV

Modifying the NPV of the BQT in a range between 0.7 and 1.0 did not affect the positioning of the cost-effectiveness graph. Sensitivity analysis showed that, if the PPV of the CRF-based strategy could be raised to 0.09, a CRF-only strategy in combination with the BQT high cut-off strategy would dominate the BQT low cut-off strategy and would also be superior to DXA.

Figure 6.6 shows the cost effectiveness chart when the CRF PPV was increased to 0.09; the combined BQT and CRF with a low cut-off and the DXA strategy become dominated by both the combined BQT and CRF with a low cut-off and the CRF-only strategy.

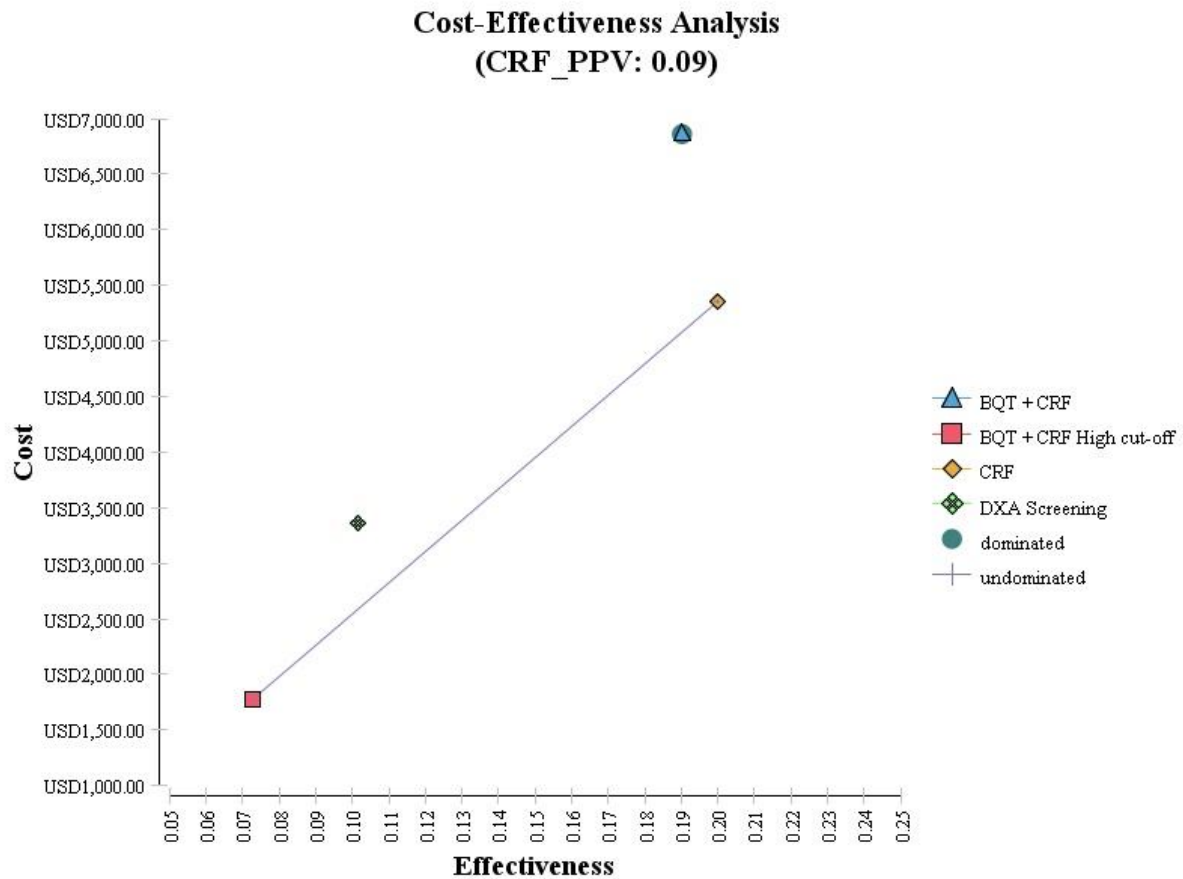


Figure 6.6: CRF PPV Sensitivity Analysis with increasing PPV

Sensitivity analysis showed that, if the NPV of the CRF-only strategy were to fall below 0.90, it would no longer be on the efficient frontier and, if it fell below 0.79, it would be dominated by a BQT-based strategy. Varying the drug cost or risk-reduction performance did not have an effect on the relative positioning of the alternate strategies.

6.4.2 Markov Model

Cost effectiveness analysis

Figure 6.7 shows that both the CRF-only and combined BQT and CRF with high cut-off scenarios are dominated by the DXA and BQT low cut-off scenarios. These two scenarios form the cost-effectiveness frontier for testing.

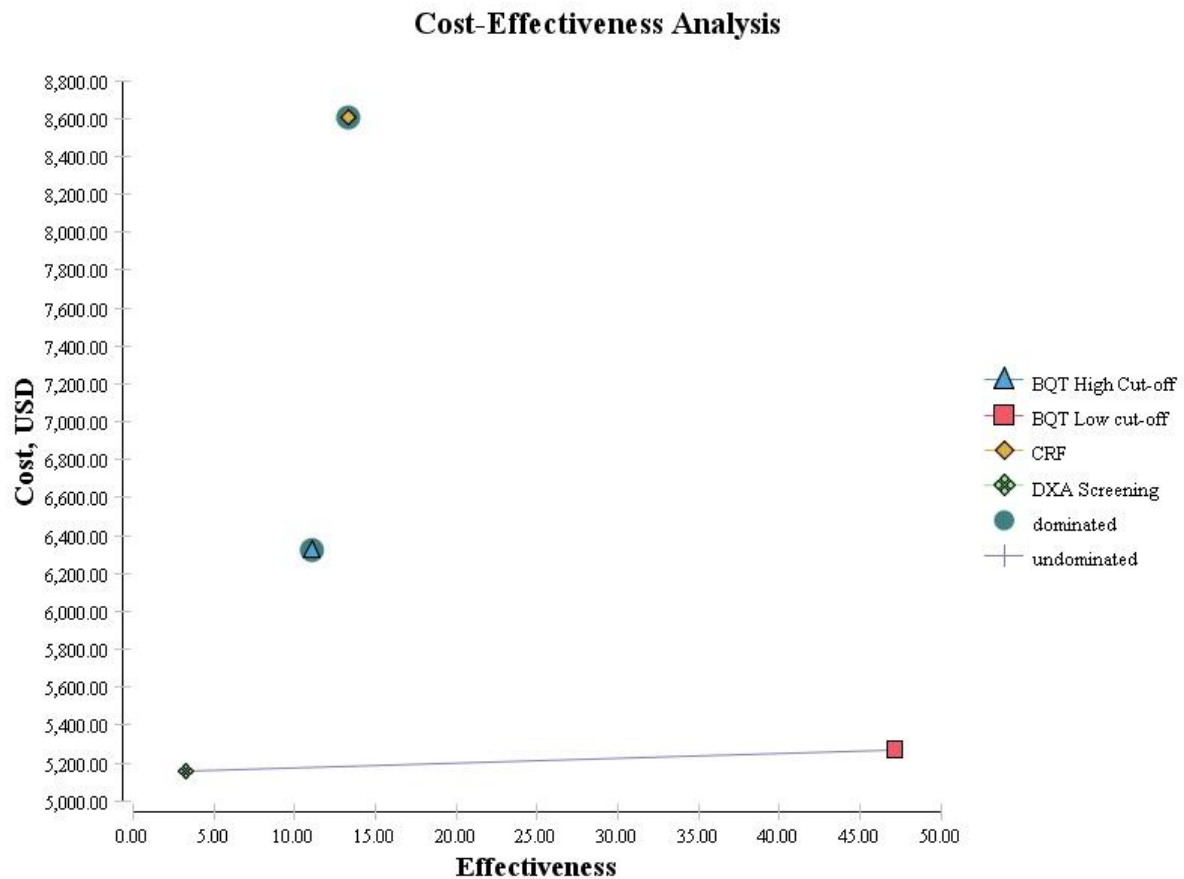


Figure 6.7: Markov cost-effectiveness based on the four diagnostic scenarios

Table 6.9 shows the incremental cost-effectiveness of the four different strategies, DXA is the lowest cost but a combined BQT and CRF strategy using the low-cut off approach is the most efficient in terms of ICER. Table 6.10 provides the RRR, ARR and NNT results, the combined BQT and CRF strategy with a low cut-off provided the highest risk reduction and lowest number needed to treat to prevent a fracture.

Table 6.9: Cost-effectiveness values for the four diagnostic scenarios

Strategy	Cost	Eff	Incr Cost	Incr Eff	ICER	Comment
DXA	5,153.83	3.29	0	0	0	Undominated
BQT Low cut-off	5,271.20	47.13	117.37	43.85	2.68	Undominated
BQT High cut-off	6,322.90	11.07	1,051.70	-36.06	-29.17	Dominated
CRF only	8,604.81	13.32	3,333.61	-33.81	-98.59	Dominated

Table 6.10: ARR, RRR and NNT results for the four diagnostic scenarios

	RRR	ARR	NNT
BQT High cut-off	5.03	0.0061	163
DXA	7.23	0.0088	113
CRF	13.32	0.016	62
BQT Low cut-off	27.96	0.034	29

Sensitivity Analysis

A number of sensitivity analyses were carried out on the BQT and CRF-only approaches as alternatives to DXA. BQT cost was adjusted in a range between \$50 and \$150; however, the relative positioning of the strategies remained unchanged. Changing the BQT hazard showed no change in the relative status of the strategies in a range of 4.0 to 10.0. Changing the CRF hazard ratios did not alter the relative positioning of the strategies in a range between 0.0 and 10.0. Modifying the cost of Denosumab between \$276 and \$1104 showed that, at a lower drug cost (\$276), the BQT low cut-off strategy dominated all other strategies.

6.4.3 Novel measures results

Table 6.11 is the table of reclassification used to determine the savings and effectiveness of using a BQT+CRFs strategy over a CRF strategy alone. The numbers in bold and underlined represent the subjects who were reclassified as a result of the addition of BQT to the CRF model.

Table 6.11: Reclassification Table for the CRF only model and with BQT added to the model.

	CRFs + BQT				Total
	Cases		Non-cases		
CRFs	0– 3%	>3%	0– 3%	>3%	
0-3%	23	<u>17</u>	45	<u>2</u>	94
>3%	<u>7</u>	35	<u>2</u>	17	68
Total	30	52	54	26	162

The weighted costs results were as follows:

- S_1 was estimated as $\$14,293 \times 0.4$ (based on the effectiveness of the treatment)
- S_2 was estimated as \$2760 based on the cost of a 5-year course of treatment

In order to estimate the total savings, 163 subjects was considered:

- Cases up = $10 (17 - 7)$
- Cases down = $0 (9 - 9)$.

The move to CRFs + BQT has resulted in 10 new cases being identified, which has an impact on both cost and effectiveness measures.

Total savings in the 163 subjects = $10 \times \$5,717.20 = \$57,172$

Average savings are then: $\$57,172/163 = \350 per person

New costs: BQT tests = $163 \times 100 + \text{Treatment}, 10 \times \$2760 = \$43,900$

The additional costs for introducing the BQT are \$269 per person.

Introducing BQT is, therefore, a cost-effective strategy, since it saves \$81 per person in this scenario compared with a CRF-only strategy.

The novel approach developed for this thesis provides an alternative ICER-based comparison:

The incremental cost to additional apply the BQT test is: \$43,900

The incremental effectiveness which results from the addition of the BQT: 4 fractures saved

By dividing the incremental cost by the incremental effectiveness an ICER of \$10,975 per fracture saved is achieved.

6.5 DISCUSSION

Decision Tree Analysis

The strategic option with the lowest ICER was the BQT+CRF low cut-off scenario. The good performance of the BQT+CRF option resulted from moving the cut-off to increase the specificity of the test. This appears to have been offset by the improved performance of the BQT resulting from the higher specificity. The combination of the sens/spec mix, the cost of the test and the number of patients selected for treatment are all factors that need to be scrutinised carefully in any change in screening strategy based around a new test. It is clear from these differences that a higher specificity resulting from a higher cut-off is preferable to higher sensitivity in the ideal prognostic test due to the low incidence of hip fracture in the population of interest. The best strategy to exceed the performance of DXA was a combination of the CRFs and the BQT strategy; this suggests that careful identification of the cut-off for high risk patients is required and resides somewhere between 20% and 50% of patients when used in conjunction with CRFs to maximise the cost-effectiveness. The sensitivity analysis is also useful in understanding the clinical impact of potential alternative scenarios; this shows that small increases in PPV can have a significant impact on the performance of the strategies and effect changes in costs of the underlying diagnostic tests.

TreeAge Pro calculates ICER in relation to the lowest cost option which interestingly was not DXA, the base case option which was planned to be measured against, BQT+CRF high cut-off was the lower cost option.

Markov Model

The Markov model identified BQT low cut-off and DXA as lying on the cost-effective frontier. This result indicates that using a combination of DXA and BQT could be an effective strategy. The extremely high cost of the CRF-only and BQT high cut-off strategies are driven by the large number of women treated and not by the cost of the tests; this indicates that the cost of the test overall is not relevant within normal test boundaries, which is supported by the evidence of the sensitivity analysis. The use of DXA and BQT without CRFs is a surprising result and may result from the similar results observed from DXA and CRFs in terms of sensitivity (Kanis, 2007). When DXA is available, it appears to be the

higher performing alternative to CRFs, since the additional cost of the test is negligible compared with the treatment costs.

New measures

The results show that the new weighted NRI measure can evaluate the cost-effectiveness of expanding a predictive model to include a new test. The new measure takes account of the number of patients that are correctly classified and assigns a cost to it. The weighted NRI can be used to compare the relative effectiveness of additional markers in prognostic models. The ICER calculated based on this approach adds another dimension to the cost-effectiveness evaluation. The calculation is simplistic and does not incorporate time-based factors and, therefore, enhanced versions of this approach are needed. Staa *et al.* report that costs per fracture avoided of up to £15k (\$24k) can be considered acceptable (van Staa *et al.*, 2007). The value generated falls below this threshold, indicating there is value in treating patients based on the addition of BQT to the model.

NNT values in both models

The NNT value in the various screening scenarios results in values in a range between 29 at the low end to 510 at the high end for the various scenarios. The lowest NNT of 29 was observed with the BQT in the Markov model based on its combination of good discrimination and high effectiveness at reducing fractures; it does not take cost into consideration.

This indicates that, based on Denosumab performance, diagnostic interventions that have NNTs below eight (8) are acceptable. Today, there are drugs with much lower costs but lower efficacy which could offer NNT values of around 30. A key challenge, therefore, is getting the pricing of the drug right to ensure that, based on cost-effectiveness levels, currently acceptable new interventions are justified.

Comparison of methodologies

The results between the Markov and decision tree models were similar in that BQT low cut-off was consistently selected on the cost-efficient frontier, though the decision tree model did include CRFs in with the BQT. The BQT performance value used in the Markov model did not include CRFs; this was so as to provide a clear evaluation of whether adding CRFs improved performance without further enhancing the hazard ratio. The costs were similar, but with larger performance differences between the strategies in the Markov model. The Markov

model offers the strong advantage of allowing hazard ratio performance to be evaluated. This could be a key factor in determining the relative performance of a test.

The much higher effectiveness payoffs in the Markov model indicate that the potential benefits of using alternative approaches to DXA are significant and that relying on decision tree analysis alone would underestimate them. The weighted NRI approach also misses out on the time-based element of the analysis. The conclusions to be drawn from each model were sufficiently different to confirm the requirement for care when selecting a model for cost-effectiveness analysis.

The approach taken by the TreeAge Pro software made comparisons of the ICERs difficult. This was because the base case selected for each model was not the same; however, graphically one can see that the overall strategy remains the same. The efficient frontier lies at a level of cut-off between the low and high values for BQT+CRF provided.

The order of magnitude of the NNT was equivalent in the decision tree analysis and the Markov model. This suggests that NNT may be helpful in evaluating diagnostic methodologies. The NNT, however, does not take cost into consideration and therefore it can only be used to evaluate effectiveness. There are examples in the literature of researchers that attempt to develop cost-effectiveness approaches that utilise NNT. A review by Kristiansen *et al.* identified some of these studies and concluded that using NNT was in-appropriate for events which are delayed rather than completely avoided, since NNT does not adequately consider time (Kristiansen and Gyrd-Hansen, 2004).

Overall conclusions

The alternative screening strategies were only able to improve effectiveness by increasing screening costs. This may be acceptable in some healthcare environments where an explicit commitment to treatment effectiveness without regard to cost is made, but is unlikely to be acceptable in most jurisdictions.

The actual measures used to evaluate diagnostic measures are highly influential to the outcome of the study. The key driver of performance in a treatment scenario is the PPV, which incorporates features of disease prevalence, sensitivity and specificity.

The two traditional approaches, namely decision tree analysis and Markov modelling, yielded differing results, but the BQT low-cut off strategy was a common theme. The Markov model resulted in lower NNT values, indicating that considering time in the analysis leads to better performance and can assist in more accurate modelling of the real situation.

The use of a simple reclassification approach led to a valid value for ICER. This approach can be used as a rapid evaluation of whether the addition of an additional biomarker is cost-effective. It does not, however, take time into account in the manner of the Markov analysis.

Going forward, there is a need for increased consideration of the health economic effects of reclassification and the incorporation of these tools into commercial programs like TreeAge Pro.

In the clinical community, it is becoming less and less credible to evaluate the performance of a new biomarker without due consideration to cost issues. Therefore, the use of cut-offs, costs and measures like NRI will lead to a more complete evaluation of new biomarkers and their relevance to clinical practice.

The development of new economic models for the use of these new prognostic measures is an area in which significant development will take place over the coming years. There is the potential to develop new frameworks and approaches for cost-effectiveness and cost-utility using NRI and other potential measures in all areas of health.

The use of ICER, NNT and other measures when evaluating new diagnostics tools needs to be carefully considered in the context of the modelling technique used. This study shows that minor changes in value and modelling approaches can lead to significantly different results. In this case, however, the diagnostic regime recommended remained the same, being a combined BQT and CRF strategy with a low cut-off.

Clinicians may find NNT measures more accessible than measures such as ICER to assist them in which technology to use. The underlying model used will not, however, yield fundamentally different information, which will provide the clinician with additional information quickly. The NRI and IDI information may provide better discrimination, but will still need to be converted into an ICER or NNT number that is accessible to the clinician. It is possible, based on these health economic measures, that different diagnostics tests will be used to evaluate whether a particular treatment is justified. Potentially, tests used in

combination may give the additional discrimination required to justify using an extremely expensive, but highly effective, drug on a particular patient.

Limitations

The FREEDOM trial design involved the selection of osteoporotic women and their random assignment to a treatment or placebo arm. A key assumption made is the use of the 40% fracture risk assumption, in that women who are at high risk for fracture will benefit at the same rate. The cut-off selected was 3% over a 23-year time frame, which is a lower risk than the NOF recommended value of 3% over a 10-year time frame. We have used US data and figures within a UK-developed framework. The two countries have fundamentally different approaches to costing healthcare. The Mayo data warehouse provides a fully-costed approach and this is assumed to closely resemble the NHS cost structure for fracture events. This chapter uses DXA data from published figures rather than values derived from the Nurses' Health Study. The sensitivity and specificity of the tests is assumed to remain unchanged, whether relative risks or hazard ratios are used, allowing the same PPVs and NPVs to be used.

Chapter 7 OVERALL CONCLUSIONS

7.1 THE NEED

The literature review has provided compelling evidence for the need to improve fracture risk prediction, the drivers include the increasing prevalence of the disease in the population globally, its negative impact on quality of life, the effectiveness of the current drugs and the weaknesses of a bone mineral density focussed diagnostic approach. It is clear that physicians feel they have the drugs to make a difference in the incidence levels of fracture but lack the prognostic tools to identify the women truly at risk.

7.2 BQT DEVELOPMENT

This thesis documents work conducted to develop a novel prognostic marker for osteoporotic fracture risk. The literature review identified the fact that there is a requirement to collect patient samples at baseline and follow the patient for a number of years to adequately validate a new marker. Due to the low incidence rate of hip fractures in postmenopausal women (less than 5%), cohorts in excess of 10,000 subjects are required to ensure sufficient events have occurred over a ten year study, making the costs and time commitment for new studies substantial. This thesis has described an alternative, more cost-effective approach in which samples used were collected in the past and re-analysed. This approach has been effective, but has limitations. The retrospective cohort approach restricts the researcher to existing samples and associated data already collected which may be sub-optimal for the marker of interest. This limitation can mean that data required to evaluate a particular predictive multivariate algorithm may not have been collected at baseline, either for an individual patient or for the entire group and this was a challenge with the Nurses' Health Study data and its analysis using FRAX and QFracture.

This research has shown that it is possible to conduct a prospective study in a timely and cost-effective manner without having to test tens of thousands of samples. The approach taken to develop the BQT is novel and may become more commonplace for diseases with long durations between when prognosis is required and the event occurring. It is clear from the research however that BQT on its own cannot supersede DXA but needs to be applied in

conjunction with other risk factors, this increases the importance of multivariate algorithms to finding an effective solution to the need for better fracture prediction.

The importance of using cohort based approaches is highlighted by the superior predictive performance of the BQT in the Cox model (HR 3.2) compared to its performance when evaluated using a logistic regression model (OR 1.59), time has been shown to be an important factor in predictive performance that would have been missed in a retrospective case-control study.

7.3 MULTIVARIATE ALGORITHMS

Online fracture risk assessment tools offer significant opportunities for novel biomarkers to be used for fracture risk prediction. The research in this thesis and recent work with the QResearch database indicates that better predictive performance can be achieved by the addition of more risk factors, if appropriately validated. The multivariate algorithms, FRAX and QFractureScores were compared in case-control and cohort models in this thesis and found to be predictive, but the relative performance of the two algorithms was highly dependent on the input data. In the UK and Ireland focused retrospective study described in chapter 3, FRAX was superior to QFracture and in the NHS study focused cohort study QFracture was superior to FRAX in terms of AUC but the differences were small and unlikely to be statistically significant.

7.4 PROGNOSTIC MARKER PERFORMANCE EVALUATION

There has been increased recognition in the recent academic literature that there is a need for additional measures to assess the performance of different prognostic risk factors and multivariate risk models, beyond what is offered by the receiver operating characteristic (ROC) curve. The ROC curve has been observed to perform poorly as a measure of prognostic performance in population based cohorts in which the disease has a low prevalence. This is the situation in the osteoporosis disease category, with hip fractures where the incidence rate is low, but the consequences for health are very serious in terms of increased mortality and lower quality of life after an event.

New approaches to evaluate the performance of additional clinical markers in a multivariate model has been proposed, which progress beyond whether a new prognostic test offers good discrimination between the cases and controls (as evaluated by the ROC curve) to whether it significantly changes the classification of the subject who may be at risk. The new reclassification metrics used in this thesis are able to address the weaknesses of ROC curves. Two new measures of prognostic performance, in particular, are gaining popularity: the net reclassification index (NRI) and integrated discrimination improvement (IDI) and these have been evaluated in this thesis. Based on data from the NHS study and comparing the addition of BQT to a clinical risk factor model over a clinical risk factor model alone, NRI showed an improvement of 12.1% when BQT was used in conjunction with CRFs over CRFs alone. This compares with an improvement of 8% when comparing AUCs indicating that NRI offers better discrimination. The relative IDI did also produce an improvement but it was only 4%, lower than the AUC improvement.

7.5 COST-EFFECTIVENESS ANALYSIS

Cost-effectiveness analysis using Markov and decision tree approaches showed that the BQT with a low cut-off in combination with DXA was consistently on the cost-effective frontier, indicating that this new biomarker would be an integral part of any mass screening strategy. The attractiveness of using cost-effectiveness analysis is that it incorporates all aspects of the patient journey from initial diagnosis to treatment and incidence rate and therefore is the most complete way to evaluate any new prognostic marker.

7.6 FURTHER WORK

Work to expand the number of prognostic markers available to clinicians is required and the QFractureScores team are leading this work using a retrospective cohort approach. Additional work to compare QFracture and FRAX using health economic and reclassification techniques is required to establish whether one approach is superior to the other. More work to establish widely acceptable performance measures is also required, NRI is a good step forward but additional work may identify a more discriminatory tool. Finally, there is a need for better health economic tools for evaluating the addition of new prognostic markers into multivariate algorithms, the weighted NRI is only a step in the right direction.

7.7 RECOMMENDATIONS

The use of the BQT may enhance fracture prediction in conjunction with clinical risk factors and, with further spectral collection and algorithm improvements in the BQT, the combination may offer a highly cost-effective alternative to the use of DXA for mass screening in multivariate algorithms. Further work to enable clinical acceptance of this novel prognostic marker is recommended.

The analysis of any new prognostic marker should include a prospective evaluation that is appropriate for the disease category, retrospective cohort study designs may assist in achieving this goal in a timely and cost-effective manner. Reclassification metrics and health economic evaluations are critical to getting a real understanding of the clinical utility and cost-effectiveness of any new prognostic marker before it is recommended for use with patients and these techniques should therefore become standard evaluation tools.

REFERENCES

- Akhtar, W. and Edwards, H. G. M. (1997), "Fourier-transform Raman spectroscopy of mammalian and avian keratotic biopolymers", *Spectrochimica Acta - Part A Molecular Spectroscopy*, vol. 53, no. 1, pp. 81-90.
- Al-Delaimy, W. K., Mahoney, G. N., Speizer, F. E. and Willett, W. C. (2002), "Toenail nicotine levels as a biomarker of tobacco smoke exposure", *Cancer Epidemiology Biomarkers and Prevention*, vol. 11, no. 11, pp. 1400-1404.
- Al-Delaimy, W. K., Stampfer, M. J., Manson, J. E. and Willett, W. C. (2008), "Toenail nicotine levels as predictors of coronary heart disease among women", *American Journal of Epidemiology*, vol. 167, no. 11, pp. 1342-1348.
- Al-Delaimy, W. K. and Willett, W. C. (2008), "Measurement of tobacco smoke exposure: Comparison of toenail nicotine biomarkers and self-reports", *Cancer Epidemiology Biomarkers and Prevention*, vol. 17, no. 5, pp. 1255-1261.
- Arden, N. K., Baker, J., Hogg, C., Baan, K. and Spector, T. D. (1996), "The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: A study of postmenopausal twins", *Journal of Bone and Mineral Research*, vol. 11, no. 4, pp. 530-534.
- Augat, P., Reeb, H. and Claes, L. E. (1996), "Prediction of fracture load at different skeletal sites by geometric properties of the cortical shell", *Journal of Bone and Mineral Research*, vol. 11, no. 9, pp. 1356-1363.
- Augat, P. and Schorlemmer, S. (2006), "The role of cortical bone and its microstructure in bone strength", *Age and Ageing*, vol. 35, no. SUPPL.2.
- Banse, X., Sims, T. J. and Bailey, A. J. (2002), "Mechanical properties of adult vertebral cancellous bone: Correlation with collagen intermolecular cross-links", *Journal of Bone and Mineral Research*, vol. 17, no. 9, pp. 1621-1628.
- Barr, R. J., Stewart, A., Torgerson, D. J. and Reid, D. M. (2009), "Population screening for osteoporosis risk: a randomised control trial of medication use and fracture risk", *Osteoporosis International*, pp. 1-8.
- Bauer, D. C. (2001), "Chapter 21: Biochemical markers of turnover: the Study of Osteoporotic fracture. Bone Markers, Biochemical and Clinical Perspectives", in Martin Duntz, UK.
- Bauer, D. C., Gluer, C. C., Cauley, J. A., Vogt, T. M., Ensrud, K. E., Genant, H. K. and Black, D. M. (1997), "Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women: A prospective study", *Archives of Internal Medicine*, vol. 157, no. 6, pp. 629-634.

- Bayraktar, H. H., Morgan, E. F., Niebur, G. L., Morris, G. E., Wong, E. K. and Keaveny, T. M. (2004), "Comparison of the elastic and yield properties of human femoral trabecular and cortical bone tissue", *Journal of Biomechanics*, vol. 37, no. 1, pp. 27-8.
- Bean, W. B. (1953), "A note on fingernail growth", *The Journal of investigative dermatology*, vol. 20, no. 1, pp. 27-31.
- Ben Greener (2009), *Pipeline and Commercial Insight: Osteoporosis*, DMHC2511, Datamonitor.
- Bianchi, M. L., Orsini, M. R., Saraifoger, S., Ortolani, S., Radaelli, G. and Betti, S. (2005), "Quality of life in post-menopausal osteoporosis", *Health and Quality of Life Outcomes*, vol. 3.
- Biesheuvel, C. J., Vergouwe, Y., Oudega, R., Hoes, A. W., Grobbee, D. E. and Moons, K. G. M. (2008), "Advantages of the nested case-control design in diagnostic research", *BMC Medical Research Methodology*, vol. 8.
- Black, D. M., Delmas, P. D., Eastell, R., Reid, I. R., Boonen, S., Cauley, J. A., Cosman, F., Lakatos, P., Ping, C. L., Man, Z., Mautalen, C., Mesenbrink, P., Hu, H., Caminis, J., Tong, K., Rosario-Jansen, T., Krasnow, J., Hue, T. F., Sellmeyer, D., Eriksen, E. F. and Cummings, S. R. (2007), "Once-yearly zoledronic acid for treatment of postmenopausal osteoporosis", *New England Journal of Medicine*, vol. 356, no. 18, pp. 1809-1822.
- Boivin, G. and Meunier, P. J. (2002), "The degree of mineralization of bone tissue measured by computerized quantitative contact microradiography", *Calcified tissue international*, vol. 70, no. 6, pp. 503-511.
- Boonen, S., Autier, P., Barette, M., Vanderschueren, D., Lips, P. and Haentjens, P. (2004), "Functional outcome and quality of life following hip fracture in elderly women: a prospective controlled study", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 15, no. 2, pp. 87-94.
- Borah, B., Gross, G. J., Dufresne, T. E., Smith, T. S., Cockman, M. D., Chmielewski, P. A., Lundy, M. W., Hartke, J. R. and Sod, E. W. (2001), "Three-dimensional microimaging (MRmul and muCT), finite element modeling, and rapid prototyping provide unique insights into bone architecture in osteoporosis", *The Anatomical Record*, vol. 265, no. 2, pp. 101-110.
- Borgström, F., Ström, O., Coelho, J., Johansson, H., Oden, A., McCloskey, E. V. and Kanis, J. A. (2009), "The cost-effectiveness of risedronate in the UK for the management of osteoporosis using the FRAX®", *Osteoporosis International*, pp. 1-11.
- Boskey, A. L., Spevak, L. and Weinstein, R. S. (2009), "Spectroscopic markers of bone quality in alendronate-treated postmenopausal women", *Osteoporosis International*, vol. 20, no. 5, pp. 793-800.

- Bradburn, M. J., Clark, T. G., Love, S. B. and Altman, D. G. (2003), "Survival Analysis Part III: Multivariate data analysis - Choosing a model and assessing its adequacy and fit", *British journal of cancer*, vol. 89, no. 4, pp. 605-611.
- Braga, V., Mottes, M., Mirandola, S., Lisi, V., Malerba, G., Sartori, L., Bianchi, G., Gatti, D., Rossini, M., Bianchini, D. and Adami, S. (2000), "Association of CTR and COL1A1 alleles with BMD values in peri-and postmenopausal women", *Calcified tissue international*, vol. 67, no. 5, pp. 361-366.
- Braithwaite, R. S., Col, N. F. and Wong, J. B. (2003), "Estimating hip fracture morbidity, mortality and costs", *Journal of the American Geriatrics Society*, vol. 51, no. 3, pp. 364-370.
- Branden, C. and Tooze, J. (1999), *Introduction to protein structure*, Garland Publishing, New York.
- Brismar, T., Tofts, L. and Jansky, I. (2010), "**Calcaneal BMD Obtained by Dual X-ray and Laser (DXL) Predicts Future Hip Fractures – A Prospective Study on 4 398 Swedish Women**", *Journal of Osteoporosis*, vol. In Press.
- Brown, M. A., Bradlow, J. and Gray, A. M. (2001), "Cost effectiveness of bone density measurements", *Journal of the British Menopause Society*, vol. 7, no. 3, pp. 130-135.
- Burge, R., Dawson-Hughes, B., Solomon, D. H., Wong, J. B., King, A. and Tosteson, A. (2007), "Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025", *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, vol. 22, no. 3, pp. 465-475.
- Burge, R. T., Worley, D., Johansen, A., Bhattacharyya, S. and Bose, U. (2001), "The cost of osteoporotic fractures in the UK: Projections for 2000-2020", *Journal of Medical Economics*, vol. 4, no. 51-62, pp. 51-62.
- Caspers, P. J., Lucassen, G. W., Carter, E. A., Bruining, H. A. and Puppels, G. J. (2001), "In vivo confocal raman microspectroscopy of the skin: Noninvasive determination of molecular concentration profiles", *Journal of Investigative Dermatology*, vol. 116, no. 3, pp. 434-442.
- Caspers, P. J., Lucassen, G. W. and Puppels, G. J. (2003), "Combined in vivo confocal Raman spectroscopy and confocal microscopy of human skin", *Biophysical journal*, vol. 85, no. 1, pp. 572-580.
- Cefalu, C. A. (2004), "Is bone mineral density predictive of fracture risk reduction?", *Current medical research and opinion*, vol. 20, no. 3, pp. 341-349.
- Chan, M. Y., Nguyen, N. D., Center, J. R., Eisman, J. A. and Nguyen, T. V. (2011), "Absolute Fracture-Risk Prediction by a Combination of Calcaneal Quantitative Ultrasound and Bone Mineral Density", *Calcified tissue international*, pp. 1-9.

- Chapurlat, R. D., Garnero, P., Brárt, G., Meunier, P. J. and Delmas, P. D. (2000), "Serum type I collagen breakdown product (serum CTX) predicts hip fracture risk in elderly women: The EPIDOS study", *Bone*, vol. 27, no. 2, pp. 283-286.
- Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D. G. (2003), "Survival Analysis Part I: Basic concepts and first analyses", *British journal of cancer*, vol. 89, no. 2, pp. 232-238.
- Colditz, G. A., Martin, P., Stampfer, M. J., Willett, W. C., Sampson, L., Rosner, B., Hennekens, C. H. and Speizer, F. E. (1986), "Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women", *American Journal of Epidemiology*, vol. 123, no. 5, pp. 894-900.
- Compston, J. (2009), "Clinical and therapeutic aspects of osteoporosis", *European Journal of Radiology*, vol. 71, no. 3, pp. 388-391.
- Compston, J., Cooper, A., Cooper, C., Francis, R., Kanis, J. A., Marsh, D., McCloskey, E. V., Reid, D. M., Selby, P. and Wilkins, M. (2009), "Guidelines for the diagnosis and management of osteoporosis in postmenopausal women and men from the age of 50 years in the UK", *Maturitas*, vol. 62, no. 2, pp. 105-108.
- Compston, J. E. (2000), "Pharmacological interventions for post-menopausal osteoporosis: An evidence-based approach", *Rheumatology*, vol. 39, no. 12, pp. 1309-1312.
- Conroy, R. M., Pyörälä, K., Fitzgerald, A. P., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetière, P., Jousilahti, P., Keil, U., Njølstad, I., Oganov, R. G., Thomsen, T., Tunstall-Pedoe, H., Tverdal, A., Wedel, H., Whincup, P., Witheimsen, L. and Graham, I. M. (2003), "Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project", *European heart journal*, vol. 24, no. 11, pp. 987-1003.
- Cook, N. R. (2007), "Use and misuse of the receiver operating characteristic curve in risk prediction", *Circulation*, vol. 115, no. 7, pp. 928-935.
- Cook, N. R. (2008), "Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve", *Clinical chemistry*, vol. 54, no. 1, pp. 17-23.
- Cook, N. R. (2010), "Assessing the Incremental Role of Novel and Emerging Risk Factors", *Current Cardiovascular Risk Reports*, pp. 1-8.
- Cox, D. R. (1972), "Regression Models and Life Tables", *Journal of the Royal Statistical Society Series B*, vol. 34 (2) 187 - 220, pp. 187-23.
- Cranney, A. and Adachi, J. D. (2005), "Benefit-risk assessment of raloxifene in postmenopausal osteoporosis", *Drug Safety*, vol. 28, no. 8, pp. 721-730.
- Cummings, S. R., McClung, M. R. and Christiansen, C. E. A. (2008), "A phase III study of the effects of Denosumab on vertebral, nonvertebral, and hip fracture in women with osteoporosis: Results from the FREEDOM trial.", Vol. Presentation 1286, 15th September 2008, Montreal, Canada, .

- Cummings, S. R., Black, D. M., Nevitt, M. C., Browner, W., Cauley, J., Ensrud, K., Genant, H. K., Palermo, L., Scott, J. and Vogt, T. M. (1993), "Bone density at various sites for prediction of hip fractures. The Study of Osteoporotic Fractures Research Group", *Lancet*, vol. 341, no. 8837, pp. 72-75.
- Cummings, S. R., Karpf, D. B., Harris, F., Genant, H. K., Ensrud, K., LaCroix, A. Z. and Black, D. M. (2002), "Improvement in spine bone density and reduction in risk of vertebral fractures during treatment with antiresorptive drugs", *The American Journal of Medicine*, vol. 112, no. 4, pp. 281-289.
- Cummings, S. R., Martin, J. S., McClung, M. R., Siris, E. S., Eastell, R., Reid, I. R., Delmas, P., Zoog, H. B., Austin, M., Wang, A., Kutilek, S., Adami, S., Zanchetta, J., Libanati, C., Siddhanti, S. and Christiansen, C. (2009), "Denosumab for prevention of fractures in postmenopausal women with osteoporosis", *New England Journal of Medicine*, vol. 361, no. 8, pp. 756-765.
- Cummins, N. M. (2009), *Raman Spectroscopy of Keratotic Tissues: A Novel Tool for Evaluation of Bone Quality ?* (Ph.D. thesis), University of Limerick, .
- Dawson-Hughes, B., Harris, S. S., Dallal, G. E., Lancaster, D. R. and Zhou, Q. (2002), "Calcium supplement and bone medication use in a US Medicare Health Maintenance Organization", *Osteoporosis International*, vol. 13, no. 8, pp. 657-662.
- Dawson-Hughes, B., Looker, A. C., Tosteson, A. N. A., Johansson, H., Kanis, J. A. and Melton III, L. J. (2009), "The potential impact of new National Osteoporosis Foundation guidance on treatment patterns", *Osteoporosis International*, pp. 1-12.
- De Laet, C., Odén, A., Johansson, H., Johnell, O., Jönsson, B. and Kanis, J. A. (2005), "The impact of the use of multiple risk indicators for fracture on case-finding strategies: A mathematical approach", *Osteoporosis International*, vol. 16, no. 3, pp. 313-318.
- Deal, C. (2004), "The use of intermittent human parathyroid hormone as a treatment for osteoporosis.", *Current rheumatology reports*, vol. 6, no. 1, pp. 49-58.
- Diarra, D., Stolina, M., Polzer, K., Zwerina, J., Ominsky, M. S., Dwyer, D., Korb, A., Smolen, J., Hoffmann, M., Scheinecker, C., Van Der Heide, D., Landewe, R., Lacey, D., Richards, W. G. and Schett, G. (2007), "Dickkopf-1 is a master regulator of joint remodeling", *Nature medicine*, vol. 13, no. 2, pp. 156-163.
- Donaldson, M. G., Cawthon, P. M., Schousboe, J. T., Ensrud, K. E., Lui, L. -, Cauley, J. A., Hillier, T. A., Taylor, B. C., Hochberg, M. C., Bauer, D. C. and Cummings, S. R. (2011), "Novel methods to evaluate fracture risk models", *Journal of Bone and Mineral Research*, vol. 26, no. 8, pp. 1767-1773.
- Draper, E. R. C., Morris, M. D., Camacho, N. P., Matousek, P., Towrie, M., Parker, A. W. and Goodship, A. E. (2005), "Novel assessment of bone using time-resolved transcutaneous Raman spectroscopy", *Journal of Bone and Mineral Research*, vol. 20, no. 11, pp. 1968-1972.

- Drummond, M. (2005), *Methods for the Economic Evaluation of Health Programmes*, Oxford University Press, Oxford.
- Eddy, D. (2004), "How to think about Screening", in Vincenza Snow (ed.) *Screening for Diseases: Prevention in Primary Care*, American College of Physicians, USA.
- Edwards, H. G. M., Gniadecka, M., Petersen, S., Hart Hansen, J. P., Faurskov Nielsen, O., Christensen, D. H. and Wulf, H. C. (2002), "NIR-FT Raman spectroscopy as a diagnostic probe for mummified skin and nails", *Vibrational Spectroscopy*, vol. 28, no. 1, pp. 3-15.
- Edwards, H. G. M., Hunt, D. E. and Sibley, M. G. (1998), "FT-Raman spectroscopic study of keratotic materials: Horn, hoof and tortoiseshell", *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 54, no. 5, pp. 745-757.
- Elliott, P. and Peakman, T. C. (2008), "The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine", *International journal of epidemiology*, vol. 37, no. 2, pp. 234-244.
- Ensrud, K. E., Lui, L. -, Taylor, B. C., Schousboe, J. T., Donaldson, M. G., Fink, H. A., Cauley, J. A., Hillier, T. A., Browner, W. S. and Cummings, S. R. (2009), "A comparison of prediction models for fractures in older women: Is more better?", *Archives of Internal Medicine*, vol. 169, no. 22, pp. 2087-2094.
- Ettinger, B., Black, D. M., Mitlak, B. H., Knickerbocker, R. K., Nickelsen, T., Genant, H. K., Christiansen, C., Delmas, P. D., Zanchetta, J. R., Stakkestad, J., Glüer, C. C., Krueger, K., Cohen, F. J., Eckert, S., Ensrud, K. E., Avioli, L. V., Lips, P. and Cummings, S. R. (1999), "Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: Results from a 3-year randomized clinical trial", *Journal of the American Medical Association*, vol. 282, no. 7, pp. 637-645.
- European Commission (2002), "Health Statistics-Key data on Health", *Panorama of the European Union*, .
- Farren, L., Shayler, S. and Ennos, A. R. (2004), "The fracture properties and mechanical design of human fingernails", *Journal of Experimental Biology*, vol. 207, no. 5, pp. 735-741.
- Felsenberg, D. and Boonen, S. (2005), "The bone quality framework: Determinants of bone strength and their interrelationships, and implications for osteoporosis management", *Clinical therapeutics*, vol. 27, no. 1, pp. 1-11.
- Ferretti, J. L., Spiaggi, E. P., Capozza, R., Cointry, G. and Zanchetta, J. R. (1992), "Interrelationships between geometric and mechanical properties of long bones from three rodent species with very different biomass: Phylogenetic implications", *Journal of Bone and Mineral Research*, vol. 7, pp. S433-S435.
- Feskanich, D., Owusu, W., Hunter, D. J., Willett, W., Ascherio, A., Spiegelman, D., Morris, S., Spate, V. L. and Colditz, G. (1998), "Use of toenail fluoride levels as an indicator for

- the risk of hip and forearm fractures in women", *Epidemiology*, vol. 9, no. 4, pp. 412-416.
- Feskanich, D., Singh, V., Willett, W. C. and Colditz, G. A. (2002a), "Vitamin A intake and hip fractures among postmenopausal women", *Journal of the American Medical Association*, vol. 287, no. 1, pp. 47-54.
- Feskanich, D., Willett, W. and Colditz, G. (2002b), "Walking and leisure-time activity and risk of hip fracture in postmenopausal women", *Journal of the American Medical Association*, vol. 288, no. 18, pp. 2300-2306.
- Feskanich, D., Willett, W., Colditz, G. and Macera, C. A. (2004), "Exercise and Risk of Hip Fracture in Postmenopausal Women", *Clinical Journal of Sport Medicine*, vol. 14, no. 2, pp. 103-104.
- Feskanich, D., Willett, W. C. and Colditz, G. A. (2003), "Calcium, vitamin D, milk consumption, and hip fractures: A prospective study among postmenopausal women", *American Journal of Clinical Nutrition*, vol. 77, no. 2, pp. 504-511.
- Fraser, I., Meier-Augenstein, W. and Kalin, R. M. (2008), "Stable isotope analysis of human hair and nail samples: The effects of storage on samples", *Journal of forensic sciences*, vol. 53, no. 1, pp. 95-99.
- Fraser, L. -, Langsetmo, L., Berger, C., Ioannidis, G., Goltzman, D., Adachi, J. D., Papaioannou, A., Josse, R., Kovacs, C. S., Olszynski, W. P., Towheed, T., Hanley, D. A., Kaiser, S. M., Prior, J., Jamal, S., Kreiger, N., Brown, J. P., Johansson, H., Oden, A., McCloskey, E., Kanis, J. A. and Leslie, W. D. (2011), "Fracture prediction and calibration of a Canadian FRAX® tool: A population-based report from CaMos", *Osteoporosis International*, vol. 22, no. 3, pp. 829-837.
- Frost, H. M. (2001), "From Wolff's law to the Utah paradigm: Insights about bone physiology and its clinical applications", *Anatomical Record*, vol. 262, no. 4, pp. 398-419.
- G Russell Warnick (2001), "Chapter 13: Variation in biochemical markers of bone turnover: sources, quantification, minimization. Bone Markers, Biochemical and Clinical Perspectives", in Martin Duntz, UK.
- Gabriel, S. E., Gabriel, S. E., Tosteson, A. N. A., Leibson, C. L., Crowson, C. S., Pond, G. R., Hammond, C. S. and Melton, I., L.J. (2002), "Direct medical costs attributable to osteoporotic fractures", *Osteoporosis International*, vol. 13, no. 4, pp. 323-330.
- Gabriel, S. E., Kneeland, T. S., Melton III, L. J., Moncur, M. M., Ettinger, B. and Tosteson, A. N. A. (1999), "Health-related quality of life in economic evaluations for osteoporosis: Whose values should we use?", *Medical Decision Making*, vol. 19, no. 2, pp. 141-148.
- Galimont-Collen, A. F. S., Vos, L. E., Lavrijsen, A. P. M., Ouwerkerk, J. and Gelderblom, H. (2007), "Classification and management of skin, hair, nail and mucosal side-effects of epidermal growth factor receptor (EGFR) inhibitors", *European journal of cancer*, vol. 43, no. 5, pp. 845-851.

- Garland, M., Morris, J. S., Stampfer, M. J., Colditz, G. A., Spate, V. L., Baskett, C. K., Rosner, B., Speizer, F. E., Willett, W. C. and Hunter, D. J. (1995), "Prospective study of toenail selenium levels and cancer among women", *Journal of the National Cancer Institute*, vol. 87, no. 7, pp. 497-505.
- Garnero, P., Dargent-Molina, P., Hans, D., Schott, A. M., Bréart, G., Meunier, P. J. and Delmas, P. D. (1998), "Do markers of bone resorption add to bone mineral density and ultrasonographic heel measurement for the prediction of hip fracture in elderly women? The EPIDOS prospective study", *Osteoporosis International*, vol. 8, no. 6, pp. 563-569.
- Garnero, P., Hausherr, E., Chapuy, M. -, Marcelli, C., Grandjean, H., Muller, C., Cormier, C., Bréart, G., Meunier, P. J. and Delmas, P. D. (1996a), "Markers of bone resorption predict hip fracture in elderly women: The EPIDOS prospective study", *Journal of Bone and Mineral Research*, vol. 11, no. 10, pp. 1531-1538.
- Garnero, P., Sornay-Rendu, E., Chapuy, M. - and Delmas, P. D. (1996b), "Increased bone turnover in late postmenopausal women is a major determinant of osteoporosis", *Journal of Bone and Mineral Research*, vol. 11, no. 3, pp. 337-349.
- Genant, H. K., Engelke, K., Fuerst, T., Glüer, C. -, Grampp, S., Harris, S. T., Jergas, M., Lang, T., Lu, Y., Majumdar, S., Mathur, A. and Takada, M. (1996), "Noninvasive assessment of bone mineral and structure: State of the art", *Journal of Bone and Mineral Research*, vol. 11, no. 6, pp. 707-730.
- Geyer, A. S., Onumah, N., Uyttendaele, H. and Scher, R. K. (2004), "Modulation of linear nail growth to treat diseases of the nail", *Journal of the American Academy of Dermatology*, vol. 50, no. 2, pp. 229-234.
- Glüer, C. C., Eastell, R., Reid, D. M., Felsenberg, D., Roux, C., Barkmann, R., Timm, W., Blenk, T., Armbrrecht, G., Stewart, A., Clowes, J., Thomasius, F. E. and Kolta, S. (2004), "Association of five quantitative ultrasound devices and bone densitometry with osteoporotic vertebral fractures in a population-based sample: The OPUS Study", *Journal of Bone and Mineral Research*, vol. 19, no. 5, pp. 782-793.
- Gniadecka, M., Nielsen, O. F., Christensen, D. H. and Wulf, H. C. (1998), "Structure of water, proteins, and lipids in intact human skin, hair, and nail", *Journal of Investigative Dermatology*, vol. 110, no. 4, pp. 393-398.
- Griffith, J. F. and Genant, H. K. (2008), "Bone mass and architecture determination: state of the art", *Best Practice and Research: Clinical Endocrinology and Metabolism*, vol. 22, no. 5, pp. 737-764.
- Guglielmi, G. and de Terlizzi, F. (2009), "Quantitative Ultrasound in the assessment of Osteoporosis", *European Journal of Radiology*, vol. 71, no. 3, pp. 425-431.
- Guglielmi, G., Njeh, C. F., De Terlizzi, F., De Serio, D. A., Scillitani, A., Cammisa, M., Fan, B., Lu, Y. and Genant, H. K. (2003), "Phalangeal quantitative ultrasound, phalangeal morphometric variables, and vertebral fracture discrimination", *Calcified tissue international*, vol. 72, no. 4, pp. 469-477.

- Hans, D., Dargent-Molina, P., Schott, A. M., Sebert, J. L., Cormier, C., Kotzki, P. O., Delmas, P. D., Pouilles, J. M., Breart, G. and Meunier, P. J. (1996), "Ultrasonographic heel measurements to predict hip fracture in elderly women: The EPIDOS prospective study", *Lancet*, vol. 348, no. 9026, pp. 511-514.
- Harrington, J. T., Broy, S. B., Derosa, A. M., Licata, A. A. and Shewmon, D. A. (2002), "Hip fracture patients are not treated for osteoporosis: A call to action", *Arthritis Care and Research*, vol. 47, no. 6, pp. 651-654.
- He, K. (2011), "Trace elements in nails as biomarkers in clinical research", *European journal of clinical investigation*, vol. 41, no. 1, pp. 98-102.
- Henry, Y. M., Fatayerji, D. and Eastell, R. (2004), "Attainment of peak bone mass at the lumbar spine, femoral neck and radius in men and women: Relative contributions of bone size and volumetric bone mineral density", *Osteoporosis International*, vol. 15, no. 4, pp. 263-273.
- Hippisley-Cox, J. and Coupland, C. (2009), "Predicting risk of osteoporotic fracture in men and women in England and Wales: Prospective derivation and validation of QFractureScores", *BMJ*, vol. 339, no. 7733, pp. 1291-1295.
- Hsu, H., Lacey, D. L., Dunstan, C. R., Solovyev, I., Colombero, A., Timms, E., Tan, H. -, Elliott, G., Kelley, M. J., Sarosi, I., Wang, L., Xia, X. -, Elliott, R., Chiu, L., Black, T., Scully, S., Capparelli, C., Morony, S., Shimamoto, G., Bass, M. B. and Boyle, W. J. (1999), "Tumor necrosis factor receptor family member RANK mediates osteoclast differentiation and activation induced by osteoprotegerin ligand", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 7, pp. 3540-3545.
- Humphrey, L., Chan, S., Helfand, M. and Woolf, S. H. (2004), "Breast Cancer Screening", in Vincenza Snow (ed.) *Screening for Diseases: Prevention in Primary Care*, American College of Physicians, USA.
- Hunter, D. J., Morris, J. S., Stampfer, M. J., Colditz, G. A., Speizer, F. E. and Willett, W. C. (1990), "A prospective study of selenium status and breast cancer risk", *Journal of the American Medical Association*, vol. 264, no. 9, pp. 1128-1131.
- Huopio, J., Kröger, H., Honkanen, R., Saarikoski, S. and Alhava, E. (2000), "Risk factors for perimenopausal fractures: A prospective study", *Osteoporosis International*, vol. 11, no. 3, pp. 219-227.
- Huybrechts, K. F., Ishak, K. J. and Caro, J. J. (2006), "Assessment of compliance with osteoporosis treatment and its consequences in a managed care population", *Bone*, vol. 38, no. 6, pp. 922-928.
- Jablecka, A., Olszewski, J. and Marzec, E. (2009), "Dielectric properties of keratin-water system in diabetic and healthy human fingernails", *Journal of Non-Crystalline Solids*, vol. 355, no. 50-51, pp. 2456-2460.

- Johansson, H., Kanis, J. A., Oden, A., Johnell, O. and McCloskey, E. (2009), "BMD, clinical risk factors and their combination for hip fracture prevention", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 20, no. 10, pp. 1675-1682.
- Johnell, O. and Kanis, J. A. (2004), "An estimate of the worldwide prevalence, mortality and disability associated with hip fracture", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 15, no. 11, pp. 897-902.
- Johnell, O. and Kanis, J. A. (2006), "An estimate of the worldwide prevalence and disability associated with osteoporotic fractures", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 17, no. 12, pp. 1726-1733.
- Jönsson, B., Ström, O., Eisman, J. A., Papaioannou, A., Siris, E. S., Tosteson, A. and Kanis, J. A. (2011), "Cost-effectiveness of Denosumab for the treatment of postmenopausal osteoporosis", *Osteoporosis International*, vol. 22, no. 3, pp. 967-982.
- Kanis, J. A., Johnell, O., De Laet, C., Jonsson, B., Oden, A. and Ogelsby, A. K. (2002), "International variations in hip fracture probabilities: Implications for risk assessment .", *Journal of Bone and Mineral Research*, vol. 17, no. 7, pp. 1237-7.
- Kanis, J. (2007), *World Health Organisation: Assessment of Osteoporosis at the Primary Health Care Level*, .
- Kanis, J. A., Borgstrom, F., Zethraeus, N., Johnell, O., Oden, A. and Jonsson, B. (2005), "Intervention thresholds for osteoporosis in the UK", *Bone*, vol. 36, no. 1, pp. 22-32.
- Kanis, J. A., Johansson, H., Oden, A., Dawson-Hughes, B., Melton III, L. J. and McCloskey, E. V. (2010), "The effects of a FRAX® revision for the USA", *Osteoporosis International*, vol. 21, no. 1, pp. 35-40.
- Kanis, J. A. and Johnell, O. (2005), "Requirements for DXA for the management of osteoporosis in Europe", *Osteoporosis International*, vol. 16, no. 3, pp. 229-238.
- Kanis, J. A., Johnell, O., Oden, A., Johansson, H. and McCloskey, E. (2008a), "FRAX and the assessment of fracture probability in men and women from the UK", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 19, no. 4, pp. 385-397.
- Kanis, J. A., McCloskey, E. V., Johansson, H., Oden, A., Melton, L. J., 3rd and Khaltayev, N. (2008b), "A reference standard for the description of osteoporosis", *Bone*, vol. 42, no. 3, pp. 467-475.
- Kanis, J. A., McCloskey, E. V., Johansson, H., Strom, O., Borgstrom, F., Oden, A., Compston, J. E., Cooper, A. L., Cooper, C., Francis, R., Marsh, D., Reid, D., Selby, P. and Wilkins, M. (2008c), "Case finding for the management of osteoporosis with

- FRAX® - Assessment and intervention thresholds for the UK", *Osteoporosis International*, vol. 19, no. 10, pp. 1395-1408.
- Kanis, J. A., Oden, A., Johansson, H., Borgström, F., Ström, O. and McCloskey, E. (2009), "FRAX® and its applications to clinical practice", *Bone*, vol. 44, no. 5, pp. 734-743.
- Kanis, J. A., Oden, A., Johnell, O., Johansson, H., De Laet, C., Brown, J., Burckhardt, P., Cooper, C., Christiansen, C., Cummings, S., Eisman, J. A., Fujiwara, S., Gluer, C., Goltzman, D., Hans, D., Krieg, M. A., La Croix, A., McCloskey, E., Mellstrom, D., Melton, L. J., 3rd, Pols, H., Reeve, J., Sanders, K., Schott, A. M., Silman, A., Torgerson, D., van Staa, T., Watts, N. B. and Yoshimura, N. (2007), "The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 18, no. 8, pp. 1033-1046.
- Kaplan, E. L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations", *Journal of the American Statistical Association*, vol. 53, pp. 457-24.
- Karasik, D., Myers, R. H., Cupples, L. A., Hannan, M. T., Gagnon, D. R., Herbert, A. and Kiel, D. P. (2002), "Genome screen for quantitative trait loci contributing to normal variation in bone mineral density: The Framingham study", *Journal of Bone and Mineral Research*, vol. 17, no. 9, pp. 1718-1727.
- Karsten Juhl Jørgensen and Peter C. Gøtzsche (2010), **"Who evaluates public health programmes? A review of the NHS Breast Screening Programme"** *J R Soc Med* 103(1): 14—20; doi:10.1258/jrsm.2009.090342", *Journal Royal Society Medicine*, vol. J R Soc Med 103(1): 14—20; doi:10.1258/jrsm.2009.090342.
- Katz, M. (2001a), "Setting up a multivariable analysis", in *Multivariable analysis: A practical guide for clinicians*, Cambridge University Press, USA.
- Katz, M. (2001b), "Special Topics", in *Multivariable analysis: A practical guide for clinicians*, Cambridge University Press, USA.
- Katzman McClish, D. (1989), "Analyzing a portion of the ROC curve", *Medical Decision Making*, vol. 9, no. 3, pp. 190-195.
- Keller, T., Butz, H., Lein, M., Kwiatkowski, M., Semjonow, A., Luboldt, H. -, Hammerer, P., Stephan, C. and Jung, K. (2005), "Discordance analysis characteristics as a new method to compare the diagnostic accuracy of tests: Example of complexed versus total prostate-specific antigen", *Clinical chemistry*, vol. 51, no. 3, pp. 532-539.
- Khalil, N., Sutton-Tyrrell, K., Strotmeyer, E. S., Greendale, G. A., Vuga, M., Selzer, F., Crandall, C. J. and Cauley, J. A. (2010), "Menopausal bone changes and incident fractures in diabetic women: a cohort study", *Osteoporosis International*, pp. 1-10.
- Khaw, K., Reeve, J., Luben, R., Bingham, S., Welch, A., Wareham, N., Oakes, S. and Day, N. (2004), "Prediction of total and hip fracture risk in men and women by quantitative

- ultrasound of the calcaneus: EPIC-Norfolk prospective population study", *Lancet*, vol. 363, no. 9404, pp. 197-202.
- Kristiansen, I. and Gyrd-Hansen, D. (2004), "Cost-effectiveness analysis based on the number-needed-to-treat: common sense or non-sense ?", *Health Economics*, vol. 13, pp. 9-10.
- LaFleur, J., McAdam-Marx, C., Kirkness, C. and Brixner, D. I. (2008), "Clinical risk factors for fracture in postmenopausal osteoporotic women: a review of the recent literature", *The Annals of Pharmacotherapy*, vol. 42, no. 3, pp. 375-386.
- Landi, M. T. and Caporaso, N. (1997), "Sample collection, processing and storage.", *IARC scientific publications*, no. 142, pp. 223-236.
- Langholz, B. and Borgan, O. (1997), "Estimation of absolute risk from nested case-control data", *Biometrics*, vol. 53, no. 2, pp. 767-774.
- Langton, C. M. and Langton, D. K. (2000), "Comparison of bone mineral density and quantitative ultrasound of the calcaneus: Site-matched correlation and discrimination of axial BMD status", *British Journal of Radiology*, vol. 73, no. 865, pp. 31-35.
- Lawson, E. E., Barry, B. W., Williams, A. C. and Edwards, H. G. M. (1997), "Biomedical applications of Raman spectroscopy", *Journal of Raman Spectroscopy*, vol. 28, no. 2-3, pp. 111-117.
- Lee, E., Wutoh, A. K., Xue, Z., Hillman, J. J. and Zuckerman, I. H. (2006), "Osteoporosis management in a Medicaid population after the Women's Health Initiative study", *Journal of Women's Health*, vol. 15, no. 2, pp. 155-161.
- Leeming, D. J., Henriksen, K., Byrjalsen, I., Qvist, P., Madsen, S. H., Garnero, P. and Karsdal, M. A. (2009), "Is bone quality associated with collagen age?", *Osteoporosis International*, vol. 20, no. 9, pp. 1461-1470.
- Lemma. (2011), *10 Week Project Report: Transformational Diagnostic Test for Osteoporosis Based on Raman Spectroscopy* .
- Lindsay, R., Scheele, W. H., Neer, R., Pohl, G., Adami, S., Mautalen, C., Reginster, J. -, Stepan, J. J., Myers, S. L. and Mitlak, B. H. (2004), "Sustained vertebral fracture risk reduction after withdrawal of teriparatide in postmenopausal women with osteoporosis", *Archives of Internal Medicine*, vol. 164, no. 18, pp. 2024-2030.
- Lips, P., Cooper, C., Agnusdei, D., Caulin, F., Egger, P., Johnell, O., Kanis, J. A., Kellingray, S., Leplege, A., Liberman, U. A., McCloskey, E., Minne, H., Reeve, J., Reginster, J. -, Scholz, M., Todd, C., De Vernejoul, M. C. and Wiklund, I. (1999), "**Quality of life in patients with vertebral fractures: Validation of the quality of life questionnaire of the European Foundation for Osteoporosis (QUALEFFO)**", *Osteoporosis International*, vol. 10, no. 2, pp. 150-10.
- Lips, P. (1997), "Epidemiology and predictors of fractures associated with osteoporosis", *The American Journal of Medicine*, vol. 103, no. 2A, pp. 3S-8S; discussion 8S-11S.

- Lips, P. and van Schoor, N. M. (2005), "Quality of life in patients with osteoporosis.", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 16, no. 5, pp. 447-455.
- Looker, A. C., Orwoll, E. S., Johnston, C. C., Jr, Lindsay, R. L., Wahner, H. W., Dunn, W. L., Calvo, M. S., Harris, T. B. and Heyse, S. P. (1997), "Prevalence of low femoral bone density in older U.S. adults from NHANES III", *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, vol. 12, no. 11, pp. 1761-1768.
- Lyles, K. W., Gold, D. T., Shipp, K. M., Pieper, C. F., Martinez, S. and Mulhausen, P. L. (1993), "Association of osteoporotic vertebral compression fractures with impaired functional status", *American Journal of Medicine*, vol. 94, no. 6, pp. 595-601.
- MacDonald, H. M., McGuigan, F. A., New, S. A., Campbell, M. K., Golden, M. H. N., Ralston, S. H. and Reid, D. M. (2001), "COL1A1 Sp1 polymorphism predicts perimenopausal and early postmenopausal spinal bone loss", *Journal of Bone and Mineral Research*, vol. 16, no. 9, pp. 1634-1641.
- Majumdar, S. (1998), "A review of magnetic resonance (MR) imaging of trabecular bone micro- architecture: Contribution to the prediction of biomechanical properties and fracture prevalence", *Technology and Health Care*, vol. 6, no. 5-6, pp. 321-327.
- Mann, V., Hobson, E. E., Li, B., Stewart, T. L., Grant, S. F. A., Robins, S. P., Aspden, R. M. and Ralston, S. H. (2001), "A COL1A1 Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality", *Journal of Clinical Investigation*, vol. 107, no. 7, pp. 899-907.
- MARCUS, R., WANG, O., SATTERWHITE, J. and MITLAK, B. (2003), "The Skeletal Response to Teriparatide Is Largely Independent of Age, Initial Bone Mineral Density, and Prevalent Vertebral Fractures in Postmenopausal Women With Osteoporosis", *Journal of Bone and Mineral Research*, vol. 18, no. 1, pp. 18-23.
- Marie, P. J., Ammann, P., Boivin, G. and Rey, C. (2001), "Mechanisms of action and therapeutic potential of strontium in bone", *Calcified tissue international*, vol. 69, no. 3, pp. 121-129.
- Marshall, D., Johnell, O. and Wedel, H. (1996), "Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures", *BMJ (Clinical research ed.)*, vol. 312, no. 7041, pp. 1254-1259.
- Martens, H., Nielsen, J. P. and Engelsen, S. (2003), "Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures", *Analytical Chemistry*, vol. 75, no. 3, pp. 394-404.
- Matousek, P., Draper, E. R. C., Goodship, A. E., Clark, I. P., Ronayne, K. L. and Parker, A. W. (2006), "Noninvasive Raman spectroscopy of human tissue in vivo", *Applied Spectroscopy*, vol. 60, no. 7, pp. 758-763.

- Matousek, P. and Stone, N. (2009), "Emerging concepts in deep Raman spectroscopy of biological tissue", *Analyst*, vol. 134, no. 6, pp. 1058-1066.
- McCloskey, E. V., Johansson, H., Oden, A. and Kanis, J. A. (2009), "From relative risk to absolute fracture risk calculation: the FRAX algorithm", *Current osteoporosis reports*, vol. 7, no. 3, pp. 77-83.
- McClung, M. R., Geusens, P., Miller, P. D., Zippel, H., Bensen, W. G., Roux, C., Adami, S., Fogelman, I., Diamond, T., Eastell, R., Meunier, P. J., Reginster, J. Y. and Hip Intervention Program Study Group (2001), "Effect of risedronate on the risk of hip fracture in elderly women. Hip Intervention Program Study Group", *The New England journal of medicine*, vol. 344, no. 5, pp. 333-340.
- McClung, M. R., Michael Lewiecki, E., Cohen, S. B., Bolognese, M. A., Woodson, G. C., Moffett, A. H., Peacock, M., Miller, P. D., Lederman, S. N., Chesnut, C. H., Lain, D., Kivitz, A. J., Holloway, D. L., Zhang, C., Peterson, M. C. and Bekker, P. J. (2006), "Denosumab in postmenopausal women with low bone mineral density", *New England Journal of Medicine*, vol. 354, no. 8, pp. 821-831.
- Meunier, P. J., Roux, C., Seeman, E., Ortolani, S., Badurski, J. E., Spector, T. D., Cannata, J., Balogh, A., Lemmel, E. -, Pors-Nielsen, S., Rizzoli, R., Genant, H. K. and Reginster, J. -. (2004), "The Effects of Strontium Ranelate on the Risk of Vertebral Fracture in Women with Postmenopausal Osteoporosis", *New England Journal of Medicine*, vol. 350, no. 5, pp. 459-468.
- Michaëlsson, K., Melhus, H., Ferm, H., Ahlbom, A. and Pedersen, N. L. (2005), "Genetic liability to fractures in the elderly", *Archives of Internal Medicine*, vol. 165, no. 16, pp. 1825-1830.
- Michaud, D. S., De Vivo, I., Morris, J. S. and Giovannucci, E. (2005), "Toenail selenium concentrations and bladder cancer risk in women and men", *British journal of cancer*, vol. 93, no. 7, pp. 804-806.
- Midanik, L. T. (1988), "Validity of self-reported alcohol use: A literature review and assessment", *British journal of addiction*, vol. 83, no. 9, pp. 1019-1029.
- Miller, P. D., Bonnick, S. L., Rosen, C. J., Altman, R. D., Avioli, L. V., Dequeker, J., Felsenberg, D., Genant, H. K., Gennari, C., Harper, K. D., Hodsman, A. B., Kleerekoper, M., Mautalen, C. A., McClung, M. R., Meunier, P. J., Nelson, D. A., Peel, N. F. A., Raisz, L. G. and Recker, R. R. (1996), "Clinical utility of bone mass measurements in adults: Consensus of an international panel", *Seminars in arthritis and rheumatism*, vol. 25, no. 6, pp. 361-372.
- Moayyeri, A., Kaptoge, S., Dalzell, N., Bingham, S., Luben, R. N., Wareham, N. J., Reeve, J. and Khaw, K. T. (2009), "Is QUS or DXA better for predicting the 10-year absolute risk of fracture?", *Journal of Bone and Mineral Research*, vol. 24, no. 7, pp. 1319-1325.
- Moon, G. and Gould, M. (2000), "Observational studies and design choices", in *Epidemiology: An introduction*, Open University Press, UK.

- Moran, P., Towler, M. R., Chowdhury, S., Saunders, J., German, M. J., Lawson, N. S., Pollock, H. M., Pillay, I. and Lyons, D. (2007), "Preliminary work on the development of a novel detection method for osteoporosis", *Journal of Materials Science: Materials in Medicine*, vol. 18, no. 6, pp. 969-974.
- Morgan, K., McGee, H., Watson, D., Perry, I., Barry, M., Shelley, J., Harrington, M., Molcho, R., Tully, E., van Lente, E., Ward, M., Lutonski, J., Conroy, R. and Brugha, R. (2008), "Survey of Lifestyle, Attitudes & Nutrition in Ireland", .
- Mueller, D., Econ, H. and Gandjour, A. (2009), "Cost-effectiveness of using clinical risk factors with and without DXA for osteoporosis screening in postmenopausal women", *Value in Health*, vol. 12, no. 8, pp. 1106-1117.
- Nagata-Kobayashi, S., Shimbo, T. and Fukui, T. (2002), "Cost-effectiveness analysis of screening for osteoporosis in postmenopausal Japanese women", *Journal of bone and mineral metabolism*, vol. 20, no. 6, pp. 350-357.
- National Institute for Health and Clinical Evidence (December 2011), *Diagnostics Assessment Programme manual*, National Institute for Health and Clinical Evidence, London.
- National Osteoporosis Foundation (2008), *Clinician's Guide to Prevention and Treatment of Osteoporosis*, National Osteoporosis Foundation.
- National Osteoporosis Guideline Group (2009), *Osteoporosis Clinical Guidelines for Prevention and Treatment*, Sheffield WHO Collaborating Centre for Osteoporosis.
- National Screening Committee (2002), "Osteoporosis and criteria for screening", .
- Neer, R. M., Arnaud, C. D., Zanchetta, J. R., Prince, R., Gaich, G. A., Reginster, J. -, Hodsmann, A. B., Eriksen, E. F., Ish-Shalom, S., Genant, H. K., Wang, O., Mitlak, B. H., Mellström, D., Oefjord, E. S., Marciniowska-Suchowierska, E., Salmi, J., Mulder, H., Halse, J. and Sawicki, A. Z. (2001), "Effect of parathyroid hormone (1-34) on fractures and bone mineral density in postmenopausal women with osteoporosis", *New England Journal of Medicine*, vol. 344, no. 19, pp. 1434-1441.
- Nelson, M. E., Fiatarone, M. A., Morganti, C. M., Trice, I., Greenberg, R. A. and Evans, W. J. (1994), "Effects of high-intensity strength training on multiple risk factors for osteoporotic fractures: A randomized controlled trial", *Journal of the American Medical Association*, vol. 272, no. 24, pp. 1909-1914.
- Neuprez, A. and Reginster, J. (2008), "Bone-forming agents in the management of osteoporosis", *Best Practice and Research: Clinical Endocrinology and Metabolism*, vol. 22, no. 5, pp. 869-883.
- Nicholls, S. J. and Hazen, S. L. (2005), "Myeloperoxidase and cardiovascular disease", *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 25, no. 6, pp. 1102-1111.
- NOF (2008), *Clinician's Guide to Prevention and Treatment of Osteoporosis*, National Osteoporosis Foundation, Washington DC, USA.

- Ohgitani, S., Fujita, T., Fujii, Y., Hayashi, C. and Nishio, H. (2005), "Nail calcium and magnesium content in relation to age and bone mineral density", *Journal of bone and mineral metabolism*, vol. 23, no. 4, pp. 318-322.
- Ohmi, M., Ohnishi, Y., Yoden, K. and Haruna, M. (2000), "In vitro simultaneous measurement of refractive index and thickness of biological tissue by the low coherence interferometry", *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1266-1270.
- Ottani, V., Raspanti, M. and Ruggeri, A. (2001), "Collagen structure and functional implications", *Micron*, vol. 32, no. 3, pp. 251-260.
- Oxlund, H., Mosekilde, L. and Ørtoft, G. (1996), "Reduced concentration of collagen reducible cross links in human trabecular bone with respect to age and osteoporosis", *Bone*, vol. 19, no. 5 SUPPL., pp. 479-484.
- Pandey, G., Nakamura, T. and Singh, R. P. (2008), "Investigation into the correlation between viscoelastic properties of fingernails and osteoporosis", Vol. 4, pp. 1911.
- Panichkul, S., Panichkul, P., Sritara, C. and Tamdee, D. (2006), "Cost-effectiveness analysis of various screening methods for osteoporosis in perimenopausal Thai women", *Gynecologic and obstetric investigation*, vol. 62, no. 2, pp. 89-96.
- Parfitt, A. M., Mathews, C. H. E. and Villanueva, A. B. (1983), "Relationships between surface, volume, and thickness of iliac trabecular bone in aging and in osteoporosis. Implications for the microanatomic and cellular mechanisms of bone loss", *Journal of Clinical Investigation*, vol. 72, no. 4, pp. 1396-1409.
- Paschalis, E. P., Shane, E., Lyritis, G., Skarantavos, G., Mendelsohn, R. and Boskey, A. L. (2004), "Bone fragility and collagen cross-links", *Journal of Bone and Mineral Research*, vol. 19, no. 12, pp. 2000-2004.
- Pasco, J. A., Seeman, E., Henry, M. J., Merriman, E. N., Nicholson, G. C. and Kotowicz, M. A. (2006), "The population burden of fractures originates in women with osteopenia, not osteoporosis", *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 17, no. 9, pp. 1404-1409.
- Pencina, M. J., D'Agostino Sr., R. B., D'Agostino Jr., R. B. and Vasan, R. S. (2008), "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond", *Statistics in medicine*, vol. 27, no. 2, pp. 157-172.
- Pencina, M. J., D'Agostino, R. B. and Steyerberg, E. W. (2011), "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers", *Statistics in medicine*, vol. 30, no. 1, pp. 11-21.
- Pepe, M. S., Feng, Z. and Gu, J. W. (2008), "Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M.J. Pencina *et al.*, *Statistics in Medicine* (DOI: 10.1002/sim.2929)", *Statistics in medicine*, vol. 27, no. 2, pp. 173-181.

- Petry, R., Schmitt, M. and Popp, J. (2003), "Raman spectroscopy - A prospective tool in the life sciences", *ChemPhysChem*, vol. 4, no. 1, pp. 14-30.
- Pillay, I., Lyons, D., German, M. J., Lawson, N. S., Pollock, H. M., Saunders, J., Chowdhury, S., Moran, P. and Towler, M. R. (2005), "The use of fingernails as a means of assessing bone health: A pilot study", *Journal of Women's Health*, vol. 14, no. 4, pp. 339-344.
- Pluskiewicz, W., Adamczyk, P., Franek, E., Leszczynski, P., Sewerynek, E., Wichrowska, H., Napiorkowska, L., Kostyk, T., Stuss, M., Stepień-Kłos, W., Golba, K. S. and Drozdowska, B. (2010), "Ten-year probability of osteoporotic fracture in 2012 Polish women assessed by FRAX and nomogram by Nguyen *et al.*-Conformity between methods and their clinical utility", *Bone*, .
- Qu, Y., Wong, M., Thiebaud, D. and Stock, J. L. (2005), "The effect of raloxifene therapy on the risk of new clinical vertebral fractures at three and six months: A secondary analysis of the MORE trial", *Current medical research and opinion*, vol. 21, no. 12, pp. 1955-1959.
- Ralston, S. H. (2003), "Genetic determinants of susceptibility to osteoporosis", *Current Opinion in Pharmacology*, vol. 3, no. 3, pp. 286-290.
- Ralston, S. H. and De Crombrughe, B. (2006), "Genetic regulation of bone mass and susceptibility to osteoporosis", *Genes and Development*, vol. 20, no. 18, pp. 2492-2506.
- Raman, c. v. and Krishnan, K. S. (1928), "A new type of secondary radiation", *Nature*, vol. 121, pp. 501-2.
- Randell, A. G., Nguyen, T. V., Bhalerao, N., Silverman, S. L., Sambrook, P. N. and Eisman, J. A. (2000), "Deterioration in quality of life following hip fracture: A prospective study", *Osteoporosis International*, vol. 11, no. 5, pp. 460-6.
- Reeve, J., Meunier, P. J. and Parsons, J. A. (1980), "Anabolic effect of human parathyroid hormone fragment on trabecular bone in involutional osteoporosis: A multicentre trial", *British medical journal*, vol. 280, no. 6228, pp. 1340-1344.
- Reginster, J. Y., Seeman, E., De Vernejoul, M. C., Adami, S., Compston, J., Phenekos, C., Devogelaer, J. P., Curiel, M. D., Sawicki, A., Goemaere, S., Sorensen, O. H., Felsenberg, D. and Meunier, P. J. (2005), "Strontium ranelate reduces the risk of nonvertebral fractures in postmenopausal women with osteoporosis: Treatment of Peripheral Osteoporosis (TROPOS) study", *Journal of Clinical Endocrinology and Metabolism*, vol. 90, no. 5, pp. 2816-2822.
- Reppe, S., Refvem, H., Gautvik, V. T., Olstad, O. K., Høvring, P. I., Reinholt, F. P., Holden, M., Frigessi, A., Jemtland, R. and Gautvik, K. M. "Eight genes are highly associated with BMD variation in postmenopausal Caucasian women", *Bone*, 46(3):604-12.
- Rho, J., Tsui, T. Y. and Pharr, G. M. (1997), "Elastic properties of human cortical and trabecular lamellar bone measured by nanoindentation", *Biomaterials*, vol. 18, no. 20, pp. 1325-1330.

- Rimes, K. A., Salkovskis, P. M. and Shipman, A. J. (1999), "Psychological and behavioural effects of bone density screening for osteoporosis", *Psychology and Health*, vol. 14, no. 4, pp. 585-608.
- Robinson, S. and Bugler, C. (2008), *Smoking and Drinking among Adults*, Office for National Statistics.
- Roche (2011), *Roche Annual Report 2010*, Roche, Switzerland.
- Rosen, C. J. and Bilezikian, J. P. (2001), "Clinical review 123: Hot topic - Anabolic therapy for osteoporosis", *Journal of Clinical Endocrinology and Metabolism*, vol. 86, no. 3, pp. 957-964.
- Rosen, H. N., Moses, A. C., Garber, J., Iloputaife, I. D., Ross, D. S., Lee, S. L. and Greenspan, S. L. (2000), "Serum CTX: A new marker of bone resorption that shows treatment effect more often than other markers because of low coefficient of variability and large changes with bisphosphonate therapy", *Calcified tissue international*, vol. 66, no. 2, pp. 100-103.
- Royal College of Physicians (1999), *Osteoporosis: Clinical guidelines for the prevention and treatment*, Royal College of Physicians.
- Ruiz, C., Abril, N., Tarn, J. J., García-Pérez, M. Á. and Cano, A. (2009), "The new frontier of bone formation: A breakthrough in postmenopausal osteoporosis", *Climacteric*, vol. 12, no. 4, pp. 286-300.
- Russell, G. (2001), "Chapter 1: Introduction: Bone metabolism and regulation: Bone Markers, Biochemical and Clinical Perspectives", in Martin Duntz, UK.
- Saito, M., Fujii, K. and Marumo, K. (2006a), "Degree of mineralization-related collagen crosslinking in the femoral neck cancellous bone in cases of hip fracture and controls", *Calcified tissue international*, vol. 79, no. 3, pp. 160-168.
- Saito, M., Fujii, K., Mori, Y. and Marumo, K. (2006b), "Role of collagen enzymatic and glycation induced cross-links as a determinant of bone quality in spontaneously diabetic WBN/Kob rats", *Osteoporosis International*, vol. 17, no. 10, pp. 1514-1523.
- Saito, M. and Marumo, K. (2009), "Collagen cross-links as a determinant of bone quality: a possible explanation for bone fragility in aging, osteoporosis, and diabetes mellitus", *Osteoporosis International*, pp. 1-20.
- Sassi, F. (2006), "Calculating QALYs, comparing QALY and DALY calculations", *Health policy and planning*, vol. 21, no. 5, pp. 402-408.
- Schousboe, J. T. (2008), "Cost effectiveness of screen-and-treat strategies for low bone mineral density: How do we screen, who do we screen and who do we treat?", *Applied Health Economics and Health Policy*, vol. 6, no. 1, pp. 1-18.
- Schousboe, J. T., Bauer, D. C., Nyman, J. A., Kane, R. L., Melton, L. J. and Ensrud, K. E. (2007), "Potential for bone turnover markers to cost-effectively identify and select post-

- menopausal osteopenic women at high risk of fracture for bisphosphonate therapy", *Osteoporosis International*, vol. 18, no. 2, pp. 201-210.
- Schousboe, J. T., Ensrud, K. E., Nyman, J. A., Melton III, L. J. and Kane, R. L. (2005), "Universal bone densitometry screening combined with alendronate therapy for those diagnosed with osteoporosis is highly cost-effective for elderly women", *Journal of the American Geriatrics Society*, vol. 53, no. 10, pp. 1697-1704.
- Schuit, S. C., van der Klift, M., Weel, A. E., de Laet, C. E., Burger, H., Seeman, E., Hofman, A., Uitterlinden, A. G., van Leeuwen, J. P. and Pols, H. A. (2004), "Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam Study", *Bone*, vol. 34, no. 1, pp. 195-202.
- Schwenkglenks, M. and Lippuner, K. (2007), "Simulation-based cost-utility analysis of population screening-based alendronate use in Switzerland", *Osteoporosis International*, vol. 18, no. 11, pp. 1481-1491.
- Scott, I. (2008), "Interpreting risks and ratios in therapy trials", *Australian Prescriber*, vol. 31, no. 1, pp. 12-16.
- Sculpher, M., Torgerson, D., Goeree, R. and O'Brien, B. (1999), "A Critical Structured Review of Economic Evaluations of Interventions for The Prevention and Treatment of Osteoporosis", .
- Shinzawa, H., Awa, K., Kanematsu, W. and Ozaki, Y. (2009), "Multivariate data analysis for Raman spectroscopic imaging", *Journal of Raman Spectroscopy*, vol. 40, no. 12, pp. 1720-1725.
- Sim, M. F. V., Stone, M. D., Phillips, C. J., Cheung, W. Y., Johansen, A., Vasishta, S., Pettit, R. J. and Evans, W. D. (2005), "Cost effectiveness analysis of using quantitative ultrasound as a selective pre-screen for bone densitometry", *Technology and Health Care*, vol. 13, no. 2, pp. 75-85.
- Simon, J. A., Lewiecki, E. M., Smith, M. E., Petruschke, R. A., Wang, L. and Palmisano, J. J. (2002), "Patient preference for once-weekly alendronate 70 mg versus once-daily alendronate 10 mg: A multicenter, randomized, open-label, crossover study", *Clinical therapeutics*, vol. 24, no. 11, pp. 1871-1886.
- Slotnick, M. J. and Nriagu, J. O. (2006), "Validity of human nails as a biomarker of arsenic and selenium exposure: A review", *Environmental research*, vol. 102, no. 1, pp. 125-139.
- Solomon, D. H., Avorn, J., Katz, J. N., Finkelstein, J. S., Arnold, M., Polinski, J. M. and Brookhart, M. A. (2005), "Compliance with osteoporosis medications", *Archives of Internal Medicine*, vol. 165, no. 20, pp. 2414-2419.
- Stevenson, M., Jones, M. L., De Nigris, E., Brewer, N., Davis, S. and Oakley, J. (2005), "A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis", *Health technology assessment*, vol. 9, no. 22.

- Stewart, A., Kumar, V. and Reid, D. M. (2006), "Long-term fracture prediction by DXA and QUS: A 10-year prospective study", *Journal of Bone and Mineral Research*, vol. 21, no. 3, pp. 413-418.
- Stewart, A. and Reid, D. M. (2000), "Precision of quantitative ultrasound: comparison of three commercial scanners", *Bone*, vol. 27, no. 1, pp. 139-143.
- Steyerberg, E. W. and Pencina, M. J. (2010), "Reclassification calculations for persons with incomplete follow-up", *Annals of Internal Medicine*, vol. 152, no. 3, pp. 195-196.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. and Kattan, M. W. (2010), "Assessing the performance of prediction models: A framework for traditional and novel measures", *Epidemiology*, vol. 21, no. 1, pp. 128-138.
- Stone, N., Baker, R., Rogers, K., Parker, A. W. and Matousek, P. (2007), "Subsurface probing of calcifications with spatially offset Raman spectroscopy (SORS): Future possibilities for the diagnosis of breast cancer", *Analyst*, vol. 132, no. 9, pp. 899-905.
- Tabachnick, B. and Fidell, L. (1996), *Using Multivariate Statistics*, 3rd ed, HarperCollins.
- Tamaki, J., Iki, M., Kadowaki, E., Sato, Y., Kajita, E., Kagamimori, S., Kagawa, Y. and Yoneshima, H. (2011), "Fracture risk prediction using FRAX®: a 10-year follow-up survey of the Japanese Population-Based Osteoporosis (JPOS) Cohort Study", *Osteoporosis International*, pp. 1-9.
- The Lewin Group (2005), *The Value of Diagnostics Innovation, Adoption and Diffusion into Health Care*.
- Tosteson, A. N. A., Melton III, L. J., Dawson-Hughes, B., Baim, S., Favus, M. J., Khosla, S. and Lindsay, R. L. (2008), "Cost-effective osteoporosis treatment thresholds: The United States perspective", *Osteoporosis International*, vol. 19, no. 4, pp. 437-447.
- Tosteson, A. N. A. and Weinstein, M. C. (1991), "Cost-effectiveness of hormone replacement therapy after the menopause", *Bailliere's Clinical Obstetrics and Gynaecology*, vol. 5, no. 4, pp. 943-959.
- Towler, M. R., Wren, A., Rushe, N., Saunders, J., Cummins, N. M. and Jakeman, P. M. (2007), "Raman spectroscopy of the human nail: A potential tool for evaluating bone health?", *Journal of Materials Science: Materials in Medicine*, vol. 18, no. 5, pp. 759-763.
- United Nations Population Information Network (2009), *Females over 50 years of age*, .
- Urdea, M., Kolberg, J., Wilber, J., Gerwien, R., Moler, E., Rowe, M., Jorgensen, P., Hansen, T., Pedersen, O., Jørgensen, T. and Borch-Johnsen, K. (2009), "Validation of a multimarker model for assessing risk of type 2 diabetes from a five-year prospective study of 6784 Danish people (Inter99)", *Journal of diabetes science and technology*, vol. 3, no. 4, pp. 748-755.

- van Staa, T. P., Geusens, P., Zhang, B., Leufkens, H. G. M., Boonen, A. and Cooper, C. (2007), "Individual fracture risk and the cost-effectiveness of bisphosphonates in patients using oral glucocorticoids", *Rheumatology*, vol. 46, no. 3, pp. 460-466.
- Vashishth, D. (2007), "The role of the collagen matrix in skeletal fragility", *Current Osteoporosis Reports*, vol. 5, no. 2, pp. 62-66.
- Vickers, A. J. and Elkin, E. B. (2006), "Decision curve analysis: A novel method for evaluating prediction models", *Medical Decision Making*, vol. 26, no. 6, pp. 565-574.
- Viguet-Carrin, S., Garnero, P. and Delmas, P. D. (2006), "The role of collagen in bone strength", *Osteoporosis International*, vol. 17, no. 3, pp. 319-336.
- Wainwright, S. A., Marshall, L. M., Ensrud, K. E., Cauley, J. A., Black, D. M., Hillier, T. A., Hochberg, M. C., Vogt, M. T., Orwoll, E. S. and Study of Osteoporotic Fractures Research Group (2005), "Hip fracture in women without osteoporosis", *The Journal of clinical endocrinology and metabolism*, vol. 90, no. 5, pp. 2787-2793.
- Wang, H. and Zheng, H., (2008), *An improved support vector machine for the classification of imbalanced biological datasets*.
- Wang, X., Shen, X., Li, X. and Mauli Agrawal, C. (2002), "Age-related changes in the collagen network and toughness of bone", *Bone*, vol. 31, no. 1, pp. 1-7.
- Waterman, K. C. and Adami, R. C. (2005), "Accelerated aging: Prediction of chemical stability of pharmaceuticals", *International journal of pharmaceutics*, vol. 293, no. 1-2, pp. 101-125.
- Watts, N. B. (1999), "Clinical utility of biochemical markers of bone remodeling", *Clinical chemistry*, vol. 45, no. 8 II, pp. 1359-1368.
- WHO (1994), *Assessment of Fracture Risk and its Application to Screening for Postmenopausal Osteoporosis*, 843, World Health Organisation.
- Widjaja, E. and Seah, R. K. H. (2006), "Use of Raman spectroscopy and multivariate classification techniques for the differentiation of fingernails and toenails", *Applied Spectroscopy*, vol. 60, no. 3, pp. 343-345.
- Williams, A. C., Edwards, H. G. M. and Barry, B. W. (1994), "Raman spectra of human keratotic biopolymers: Skin, callus, hair and nail", *Journal of Raman Spectroscopy*, vol. 25, no. 1, pp. 95-98.
- Woo, S. -, Hellstein, J. W. and Kalmar, J. R. (2006), "Systematic review: Bisphosphonates and osteonecrosis of the jaws", *Annals of Internal Medicine*, vol. 144, no. 10, pp. 753-761.
- World Health Organisation (2000), *Obesity: Preventing and Managing the Global Epidemic*, World Health Organisation, Geneva.

- World Health Organization Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK (2010), *FRAX Website*, available at: <http://www.sheffield.ac.uk/FRAX/> (accessed July 16).
- Xun, P., Hou, N., Daviglus, M., Liu, K., Morris, J. S., Shikany, J. M., Sidney, S., Jacobs, D. R. and He, K. (2011), "Fish oil, selenium and mercury in relation to incidence of hypertension: A 20-year follow-up study", *Journal of internal medicine*, vol. 270, no. 2, pp. 175-186.
- Xun, P., Liu, K., Steven Morris, J., Daviglus, M. L., Stevens, J., Jacobs Jr., D. R. and He, K. (2010), "Associations of toenail selenium levels with inflammatory biomarkers of fibrinogen, high-sensitivity C-reactive protein, and interleukin-6", *American Journal of Epidemiology*, vol. 171, no. 7, pp. 793-800.
- Yin, L., Venkatesan, S., Webb, D., Kalyanasundaram, S. and Qin, Q. (2009), "Effect of cryo-induced microcracks on microindentation of hydrated cortical bone tissue", *Materials Characterization*, vol. 60, no. 8, pp. 783-791.
- Zhu, K., Devine, A. and Prince, R. L. (2009), "Quantitative ultrasound measurements predict fracture in older women: A 10-year longitudinal study", *Bone*, vol. 44, no. Supplement 1, pp. S118-S118.
- Žižak, I., Paris, O., Roschger, P., Bernstorff, S., Amenitsch, H., Klaushofer, K. and Fratzl, P. (2000), "Investigation of bone and cartilage by synchrotron scanning-SAS and -WAXD with micrometer spatial resolution", *Journal of Applied Crystallography*, vol. 33, no. 3 I, pp. 820-823.

APPENDIX A

Clinical Risk Factors for Osteoporosis in Ireland and the UK: A Comparison of FRAX and QFractureScores

N. M. Cummins · E. K. Poku · M. R. Towler ·
 O. M. O'Driscoll · S. H. Ralston

Received: 18 March 2011 / Accepted: 14 May 2011
 © Springer Science+Business Media, LLC 2011

Abstract Recently two algorithms have become available to estimate the 10-year probability of fracture in patients suspected to have osteoporosis on the basis of clinical risk factors: the FRAX algorithm and QFractureScores algorithm (QFracture). The aim of this study was to compare the performance of these algorithms in a study of fracture patients and controls recruited from six centers in the United Kingdom and Ireland. A total of 246 postmenopausal women aged 50–85 years who had recently suffered a low-trauma fracture were enrolled and their characteristics were compared with 338 female controls who had never suffered a fracture. Femoral bone mineral density was measured by dual-energy X-ray absorptiometry, and fracture risk was calculated using the FRAX and QFracture algorithms. The FRAX algorithm yielded higher scores for fracture risk than the QFracture

algorithm. Accordingly, the risk of major fracture in the overall study group was 9.5% for QFracture compared with 15.2% for FRAX. For hip fracture risk the values were 2.9% and 4.7%, respectively. The correlation between FRAX and QFracture was $R = 0.803$ for major fracture and $R = 0.857$ for hip fracture ($P \leq 0.0001$). Both algorithms yielded high specificity but poor sensitivity for prediction of osteoporosis. We conclude that the FRAX and QFracture algorithms yield similar results in the estimation of fracture risk. Both of these tools could be of value in primary care to identify patients in the community at risk of osteoporosis and fragility fractures for further investigation and therapeutic intervention.

Keywords Clinical risk factor · Osteoporosis · Fracture · FRAX · QFractureScores

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

N. M. Cummins (✉)
 Clinical Materials Unit, Materials and Surface Science Institute,
 University of Limerick, National Technological Park, Limerick,
 Ireland
 e-mail: niamh.cummins@ul.ie

E. K. Poku
 Cranfield Health, Cranfield University, Cranfield, UK

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

O. M. O'Driscoll
 Medical Engineering Design and Innovation Centre, Cork
 Institute of Technology, Cork, Ireland

S. H. Ralston
 School of Molecular and Clinical Medicine, Western General
 Hospital, Edinburgh, UK

Osteoporosis is a common condition characterized by low bone mass and an increased risk of low-trauma fracture [1, 2]. Currently, the diagnosis of osteoporosis is based on bone densitometry, and the disease is defined to exist when bone mineral density (BMD) values at the spine or hip fall 2.5 standard deviations (SD) or more below normal values in young healthy individuals. Although BMD values can be measured conveniently and noninvasively by dual-energy X-ray absorptiometry (DXA), not all physicians have access to this test. In addition, many patients who suffer fragility fractures do not have osteoporosis as defined by DXA [3], demonstrating that complementary approaches are required to develop new techniques to better identify patients at risk of fragility fractures. Reflecting this fact, a recent report by the World Health Organization (WHO) has recommended that research be conducted into the use of alternative technologies to DXA [4].

Osteoporosis is a multifactorial disease, and many clinical risk factors (CRFs) for susceptibility to the disease have been identified [5–11], some of which increase the risk of fracture independently of BMD [12]. This has led to the development of algorithms to assess fracture risk on the basis of CRFs in the absence of BMD measurements. The first of these to be developed was the FRAX algorithm, which used data from nine prospective population-based cohorts (190,000 patient-years) from Europe, North America, Australia, and Japan [13] and validated the performance in 11 independent population-based cohorts (1.2 million person-years) [14]. The FRAX algorithm is country-specific as fracture rates vary considerably in different countries [15].

A limitation of the FRAX tool is that for several of the CRFs, such as corticosteroid use, alcohol, and smoking, no account is taken of the magnitude of exposure and that no information is collected on falls, an important risk factor for fragility fracture [16, 17]. In order to address this and other issues, the QFractureScores algorithm (QFracture) was developed to estimate fracture risk based on CRFs alone [18]. The QFracture tool was developed using data from a prospective cohort study of 1,183,663 females and 1,174,232 males in the United Kingdom (15.9 million person-years), and the validation cohort was composed of 642,153 females and 633,764 males [18]. QFracture uses many of the CRFs included in FRAX in addition to other variables that influence fracture risk (Table 1). Unlike FRAX, the QFracture algorithm does not incorporate BMD or previous fractures but does include more detailed information on dose response for variables like alcohol intake and smoking habit. The age range has also been extended in the QFracture algorithm (30–85 years) to allow for assessment of younger patients [18].

The aim of this study was to compare the performance of the FRAX and QFracture algorithms in identifying patients who suffered fractures in a case–control study of postmenopausal women recruited from the United Kingdom and Ireland.

Materials and Methods

Cases and controls were recruited as part of a multicenter study in Ireland and the United Kingdom. Subject recruitment by geographical region was as follows: England $n = 237$, Ireland $n = 88$, Scotland $n = 195$, and Wales $n = 64$. The study received ethical approval (MREC 07/Q1704/1), and all subjects gave written informed consent to participate. All participants were Caucasian women aged 50–85 years who were at least 5 years postmenopausal. To fully assess the performance of both tools in the identification of patients at risk of future fracture,

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

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

^a In QFracture the use of “steroids” is recorded rather than glucocorticoids

^b In QFracture secondary causes of osteoporosis are not recorded as a single entity but are recorded separately as shown above

participants included subjects who had recently suffered a fracture (cases) as well as individuals who had never suffered a fracture (controls). All fracture cases had suffered a low-trauma fracture at the hip, spine, humerus, pelvis, or wrist after the age of 45 years, whereas controls were subjects who had never sustained a fracture during adulthood (age > 18 years). We excluded subjects who were receiving treatment for osteoporosis, those on corticosteroids, and those with a secondary cause of osteoporosis such as malabsorption, chronic liver disease, renal failure, and malignant disease.

FRAX scores were calculated manually from the FRAX Web site (www.shef.ac.uk/FRAX), with double data entry in 10% of subjects. The UK version of FRAX was used for all subjects as an Irish version of FRAX is not currently available. The 10-year probabilities of major osteoporotic and hip fracture with and without BMD were recorded for FRAX. Values for QFracture were assessed using the published algorithm (Web version 1) as implemented at www.qfracture.org, and the 10-year probabilities of major osteoporotic and hip fracture were recorded.

Statistical analysis was performed using SPSS software V17 (Microsoft, San Diego, CA). Descriptive statistics are presented as mean values and SDs. Variables were tested for normality using the Kolmogorov-Smirnov test. Depending on the normality of the distribution, Pearson's or Spearman's correlation coefficients were calculated to examine the relationship between variables. Student's *t*-test or the Mann-Whitney U-test was used, as appropriate, for comparisons of two groups. Receiver operating curve (ROC) analysis was performed to calculate sensitivity, specificity, and area under the curve (AUC) values.

Results

Characteristics of the Study Sample

Characteristics of the study population are summarized in Table 2. There was a significant difference in age ($P \leq 0.01$), history of falls ($P \leq 0.0001$), and femoral BMD/T score ($P \leq 0.0001$) between the fracture cases and controls. The most common fracture type was wrist fracture ($n = 173$, 70%), followed by fractures of the humerus ($n = 30$, 12%), hip ($n = 22$, 9%), and spine ($n = 21$, 9%).

A history of falls was recorded in 39% of the total sample ($n = 227$), and 32% of subjects had a family history of osteoporosis ($n = 188$). A diagnosis of osteoporosis was confirmed in 15% of the sample ($n = 87$), and 57% were identified as osteopenic ($n = 333$). Body mass index (BMI) ranged from 15.8–49.3 kg/m² with 2% underweight ($n = 10$), 32% normal weight ($n = 188$), 40% overweight ($n = 234$), and 26% obese ($n = 152$). Current smokers comprised 15% of the sample ($n = 87$), and 33% were ex-smokers ($n = 190$), with 9% of subjects ($n = 50$) reporting an alcohol intake of ≥ 14 U/week.

Absolute Fracture Risk

The estimated absolute fracture risks are summarized in Table 3. The risks generated by the QFracture algorithm were consistently lower than those calculated by the FRAX algorithm (excluding a BMD measurement). In the total study sample the mean major fracture risk for QFracture was 9.5% (range 1.7–37.0%) vs. 15.2% (3.4–49.0%) for FRAX. Similarly, for hip fracture risk the values were 2.9% (range 0.2–29.6%) and 4.7% (0.2–36.0%), respectively. Overall, however, there was a significant correlation between the risks calculated by QFractureScores and FRAX, as demonstrated in Fig. 1. The correlation was $r = 0.803$ for major fracture and $r = 0.857$ for hip fracture ($P \leq 0.0001$).

The FRAX algorithm can also produce absolute risks including a BMD measurement; therefore, these risks were also calculated in this data set. Major fracture risk for the fracture cases was 18.7 ± 8.2 compared with 10.7 ± 5.3 for the control subjects ($P < 0.0001$). Hip fracture risk for the cases was 5.3 ± 6.2 compared with 2.4 ± 3.4 for the control subjects ($P < 0.0001$). Overall, when the risks calculated for FRAX (excluding BMD) were compared with FRAX (including BMD) ($n = 584$), a significant difference was observed for estimated major fracture ($P \leq 0.05$) and hip fracture ($P \leq 0.0001$) risks.

Sensitivity and specificity for prediction of osteoporosis (as defined by femoral BMD) were calculated for QFracture and FRAX (excluding BMD). Both algorithms yielded high specificity but poor sensitivity (Table 4, Fig. 2) for prediction of osteoporosis as defined by DXA.

Discussion

The use of CRFs to estimate 10-year probability of fracture is a significant advance in the management of osteoporosis. The aim of this study was to compare the performance of

Table 2 Characteristics of the study population

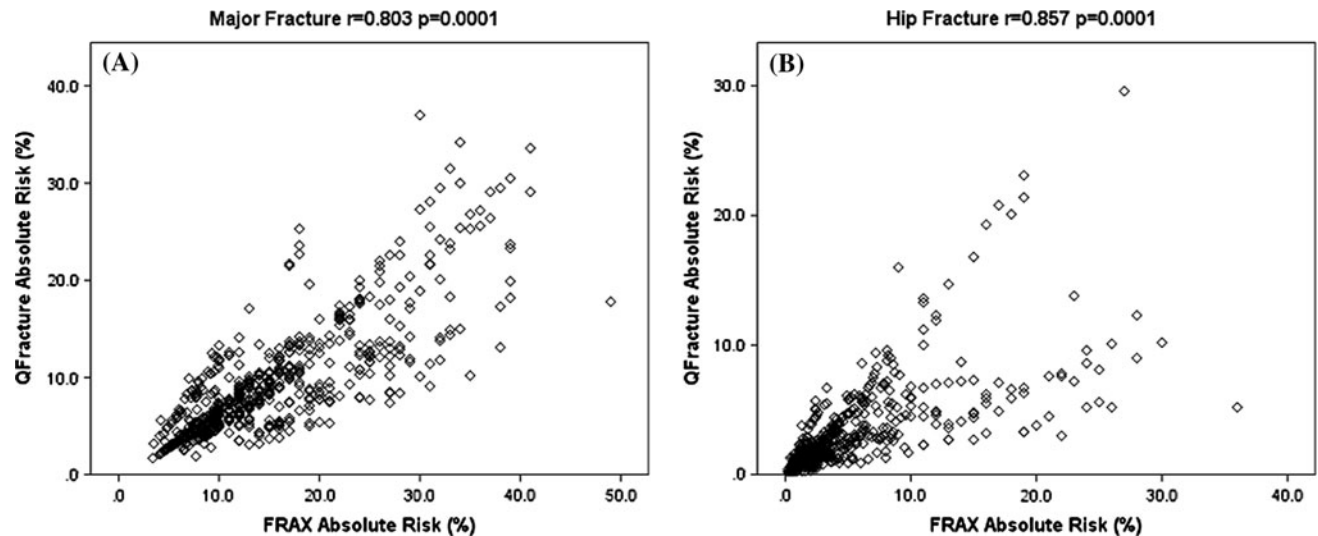
	Fractures ($n = 246$)	Controls ($n = 338$)	<i>P</i>
Age (years)	68 ± 8	66 ± 7	0.002
Height (m)	1.59 ± 0.06	1.60 ± 0.06	0.051
Weight (kg)	69 ± 13	70 ± 13	0.296
BMI (kg/m ²)	27.1 ± 5.0	27.3 ± 5.0	0.683
Femoral BMD (g/cm ²)	0.670 ± 0.109	0.739 ± 0.136	0.0001
Femoral BMD (T score)	-1.7 ± 0.9	-1.2 ± 1.1	0.0001
Relative with fracture (<i>n</i>)	75 (30)	113 (33)	0.452
History of falls (<i>n</i>)	158 (64)	69 (20)	0.0001
Current smoker (<i>n</i>)	38 (15)	49 (15)	0.992
Alcohol ≥ 14 U/week (<i>n</i>)	23 (9)	27 (8)	0.562

Values are mean $\pm$ SD or numbers (%)

Table 3 Estimated fracture risk in cases and controls as calculated by FRAX and QFracture

	Fractures (<i>n</i> = 246)	Controls (<i>n</i> = 338)	<i>P</i>
QFracture major fracture	11.7 ± 7.0	7.9 ± 4.5	<0.0001
FRAX major fracture	21.3 ± 8.2	10.8 ± 4.9	<0.0001
QFracture hip fracture	4.0 ± 4.3	2.1 ± 2.0	<0.0001
FRAX hip fracture	7.3 ± 6.4	2.8 ± 3.7	<0.0001

Values are mean ± SD risk for major osteoporotic fracture and hip fracture as estimated by the FRAX and QFracture algorithms

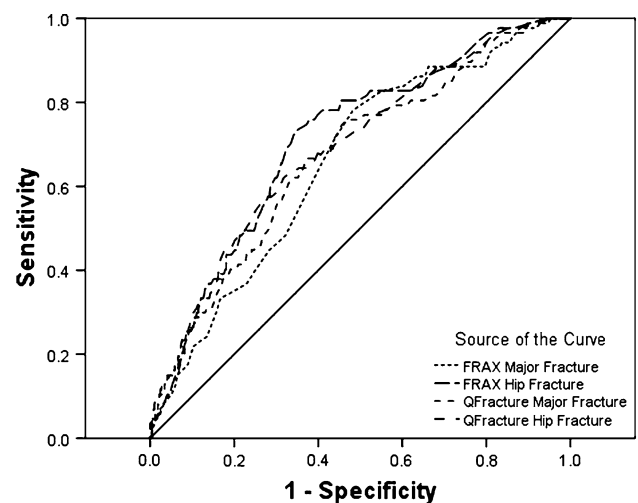
**Fig. 1** Correlation between FRAX and QFracture scores and absolute risks for **a** major fracture and **b** hip fracture (*n* = 584)**Table 4** Sensitivity, specificity, and AUC values for prediction of osteoporosis using QFracture and FRAX

	Sensitivity (%)	Specificity (%)	AUC
QFracture major fracture	16	95	0.668
FRAX major fracture	37	77	0.665
QFracture hip fracture	3	99	0.637
FRAX hip fracture	7	97	0.710

Values are derived from the whole study population of cases and controls

AUC area under the curve, as determined by receiver-operator curve analysis

the FRAX and QFracture algorithms in a case-control study of postmenopausal women. In this study there was a significant difference in age (66 vs. 68 years, $P \leq 0.01$), history of falls (69 vs. 158, $P \leq 0.0001$), and femoral T score (-1.2 vs. -1.7 , $P \leq 0.0001$) between the control and fracture groups, respectively. This was not unexpected as these risk factors are among the strongest for fracture [19, 20]; however, the history of falls data was striking. A history of falls was recorded in 39% of the total study sample. Fall history is included in the QFracture tool [18] but not in the FRAX tool. This might explain why

**Fig. 2** ROC analysis for prediction of osteoporosis by FRAX and QFracture

QFractureScores had greater specificity for prediction of major fractures and hip fractures than the FRAX score, although this was counterbalanced by a poorer sensitivity.

The absolute fracture risk values calculated in this study are broadly in agreement with previous published findings,

where QFracture predicted lower risks than FRAX [18, 21]. We think that the lower scores with the QFracture algorithm are likely due to the fact that this tool does not take previous fracture into account, which is a strong risk factor for future fracture [22]. Reflecting this fact, the differences in estimates of fracture risk between FRAX and QFracture were much greater in the cases than in the controls (Table 3). It is not possible to determine from this cross-sectional study, however, whether fracture risk is underestimated by QFracture or overestimated by FRAX. Although the algorithms differed, there were similarities between estimates, with an overall correlation between FRAX and QFracture of $r = 0.803$ for major fracture and $r = 0.857$ for hip fracture ($P \leq 0.0001$). Although both algorithms yielded high specificity for the detection of osteoporosis as defined by DXA, sensitivity was poor.

At present, there is no universally accepted policy for population screening in the United Kingdom to identify individuals with osteoporosis or those at high risk of fracture [23], which has been driven by the cost–benefit equation using the current diagnostic tools available. A potential advantage of QFracture is the collection of data on many more risk factors than FRAX, but an important limitation is that it does not take previous fracture into account. Most of the variables assessed in the QFracture algorithm may have already been gathered in patients' electronic general practice record in the United Kingdom or could be collected by a clinician during a standard consultation [18]. The FRAX score is also easy and quick to calculate and can be used with or without the inclusion of a femoral neck BMD measurement [21]. Some studies have indicated that FRAX does not offer significant advantages over less complex models that also incorporate BMD [24, 25].

Both FRAX and QFracture are platform technologies, which theoretically could be upgraded as new validated risk indicators become available [26]. However, these additional risk factors would need to be validated on the original populations.

This study has some limitations, including its retrospective nature, case–control design, and relatively small sample size. However, to the best of the authors' knowledge, this is the first study to compare the FRAX and QFracture tools in an independent sample.

There has been considerable debate as to how FRAX should be used in routine clinical practice. The procedure implemented on the UK National Osteoporosis Guideline Group Web site suggests that patients at high risk of fracture on the basis of FRAX should be treated without recourse to DXA, although this remains controversial since the vast majority of randomized controlled trials of osteoporosis therapies have focused on patients with osteoporosis as defined by DXA [27–29]. Until further evidence

emerges to demonstrate that targeting patients for therapy on the basis of absolute fracture risk is effective, it could be that the optimal use of FRAX and/or QFracture might be as a prescreening tool to identify patients who should be referred for DXA.

Acknowledgments The authors gratefully acknowledge the financial support of the Enterprise Ireland Innovation Partnership grant board. The authors also sincerely acknowledge the assistance of all the participating subjects.

References

1. Melton LJ, Atkinson EJ, O'Connor MK, O'Fallon WM, Riggs BL (1998) Bone density and fracture risk in men. *J Bone Miner Res* 13:1915–1923
2. Kanis JA, Johnell O, Oden A, Sembo I, Redlund-Johnell I et al (2000) Long-term risk of osteoporotic fracture in Malmö. *Osteoporos Int* 11:669–674
3. Schuit SC, Van der Klift M, Weel AE, De Laet C, Burger H et al (2004) Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam study. *Bone* 34:195–202
4. Organisation World Health (2004) WHO Scientific Group on the Assessment of Osteoporosis at Primary Health Care Level. WHO, Brussels
5. Kanis JA, Johnell O, De Laet C, Johansson H, Oden A et al (2004) A meta-analysis of previous fracture and subsequent fracture risk. *Bone* 35:375–382
6. Kanis JA, Johansson H, Oden A, Johnell O, De Laet C et al (2004) A family history of fracture and fracture risk: a meta-analysis. *Bone* 35:1029–1037
7. Kanis JA, Johansson H, Oden A, Johnell O, de Laet C et al (2004) A meta-analysis of prior corticosteroid use and fracture risk. *J Bone Miner Res* 19:893–899
8. Kanis JA, Johnell O, Oden A, Johansson H, De Laet C et al (2005) Smoking and fracture risk: a meta-analysis. *Osteoporos Int* 16:222–228
9. Kanis JA, Johansson H, Johnell O, Oden A, De Laet C et al (2005) Alcohol intake as a risk factor for fracture. *Osteoporos Int* 16:737–742
10. Kanis JA, Johansson H, Oden A, De Laet C, Johnell O et al (2005) A meta-analysis of milk intake and fracture risk: low utility for case finding. *Osteoporos Int* 16:799–804
11. De Laet C, Kanis JA, Odén A, Johansson H, Johnell O et al (2005) Body mass index as a predictor of fracture risk: a meta-analysis. *Osteoporos Int* 16:1330–1338
12. De Laet C, Odén A, Johansson H, Johnell O, Jönsson B et al (2005) The impact of the use of multiple risk indicators for fracture on case-finding strategies: a mathematical approach. *Osteoporos Int* 16:313–318
13. Kanis JA, Borgstrom F, De Laet C, Johansson H, Johnell O et al (2005) Assessment of fracture risk. *Osteoporos Int* 16:581–589
14. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C et al (2007) The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 18:1033–1046
15. Kanis JA, Johnell O, De Laet C, Jonsson B, Oden A et al (2002) International variations in hip fracture probabilities: implications for risk assessment. *J Bone Miner Res* 17:1237–1244
16. Ward KD, Klesges RC (2001) A meta-analysis of the effects of cigarette smoking on bone mineral density. *Calcif Tissue Int* 68:259–270

17. Berg KM, Kunins HV, Jackson JL, Nahvi S, Chaudhry A et al (2008) Association between alcohol consumption and both osteoporotic fracture and bone density. *Am J Med* 121:406–418
18. Hippisley-Cox J, Coupland C (2009) Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 339:b4229
19. Johnell O, Kanis JA, Oden A, Johansson H, De Laet C et al (2005) Predictive value of BMD for hip and other fractures. *J Bone Miner Res* 20:1185–1194
20. Cummings SR, Nevitt MC (1994) Non-skeletal determinants of fractures: the potential importance of the mechanics of falls. Study of Osteoporotic Fractures Research Group. *Osteoporos Int* 4:67–70
21. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E (2008) FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 19:385–397
22. Black DM, Arden NK, Palermo L, Pearson J, Cummings SR (1999) Prevalent vertebral deformities predict hip fractures and new vertebral deformities but not wrist fractures. *J Bone Miner Res* 14:821–828
23. Compston J, Cooper A, Cooper C, Francis R, Kanis JA et al (2009) Guidelines for the diagnosis and management of osteoporosis in postmenopausal women and men from the age of 50 years in the UK. *Maturitas* 62:105–108
24. Ensrud KE, Lui LY, Taylor BC, Schousboe JT, Donaldson MG et al (2009) A comparison of prediction models for fractures in older women: is more better? *Arch Int Med* 169:2087–2094
25. Pluskiewicz W, Adamczyk P, Franek E, Leszczynski P, Sewerynek E et al (2010) 10-year probability of osteoporotic fracture in 2012 Polish women assessed by FRAX and nomogram by Nguyen et al.—conformity between methods and their clinical utility. *Bone* 46:1661–1667
26. Kanis JA, Oden A, Johansson H, Borgstrom F, Strom O et al (2009) FRAX and its applications to clinical practice. *Bone* 44:734–743
27. Siris ES, Simon JA, Barton IP, McClung MR, Grauer A (2008) Effects of risedronate on fracture risk in postmenopausal women with osteopenia. *Osteoporos Int* 19:681–686
28. Rogers A, Glover SJ, Eastell R (2009) A randomised, double-blinded, placebo-controlled, trial to determine the individual response in bone turnover markers to lasofoxifene therapy. *Bone* 45:1044–1052
29. Finkelstein JS, Wyland JJ, Lee H, Neer RM (2010) Effects of teriparatide, alendronate, or both in women with postmenopausal osteoporosis. *J Clin Endocrinol Metab* 95:1838–1845

APPENDIX B

Raman spectroscopy of toenail as prognostic indicator of future hip fracture risk.

J Renwick Beattie¹, Mark R Towler², Diane Feskanich³, John Armhein⁴, Niamh M Cummins⁵, Tom C Bakker Schut⁶ and Ernest Poku¹

Keywords: *Keratin, osteoporosis, protein structure, clinical prognosis, bone quality*

Diagnosis and treatment for osteoporosis is focused on women over 65 year of age, but the most rapid bone loss occurs earlier, at the onset of menopause. The WHO has identified a need for improved prognostic indicators of fracture risk [1], supplementing the use of clinical risk factors in FRAX® or the use of DEXA scans. Bone is comprised of a mineral phase (primarily calcium and phosphate ions) and a protein scaffold round which the mineral is deposited. The protein scaffold provides flexibility that allows the bone to absorb impact energy, but this flexibility declines with age and is believed to be an important factor in bone quality independent of the mineral component. It has previously been reported that spectral features of the Raman spectrum of nails were correlated with the occurrence of a fracture in post-menopausal women [2,3]. The objective of this study was to investigate if the Raman spectra of nail samples could be used as an early stage predictive tool for hip fracture, many years before the occurrence of the fracture.

METHODS

Archived toenail samples from postmenopausal women aged 50 to 63y in the Nurses' Health Study: 82 women with a hip fracture up to 20 years after nail collection and 81 age-matched controls. The samples were analysed on a Skin Composition Analyzer (River Diagnostics, Rotterdam, NL; 785 nm excitation, 60 mW, spectral resolution 4 cm⁻¹), measuring spectra at 2-6 locations on the underside of the nail. The data was standardized by EMSC, the spectral variable reduced by PCA before developing a linear discriminant analysis model. The discriminant scores were used along with clinical risk factors (CRF) to develop clinical risk of fracture models for 5, 10 and 15 year intervals. The ability of the test score to predict hip fracture was tested within these same women in models with and without CRF by comparing the odds ratios (OR) per 1 SD increase in standardized predictive values.

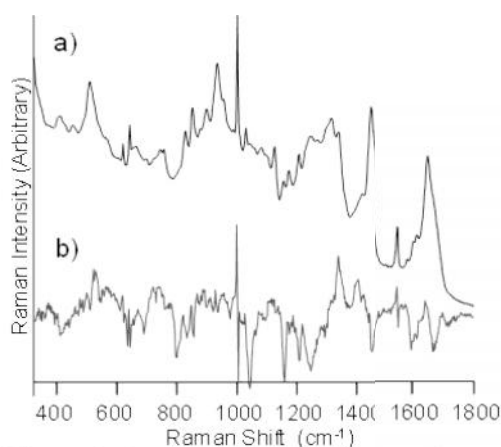


Figure 1 a) average Raman spectra of human toenails. b) discriminant function between fracture and non fracture

RESULTS

The Raman spectrum of toenails, Figure 1a, is dominated by the signature of keratin, the dominant constituent of the nail matrix. The spectrum exhibits very strong contributions from alpha helical proteins (peaks at 1651 and 935 cm^{-1}), phenylalanine (1003 cm^{-1}), tyrosine (doublet at 830 and 850 cm^{-1}), cysteine (640 and 620 cm^{-1}) and cystine (510 cm^{-1}). All of these modes are found to contribute to the discriminant function (Figure 1b), although the cystine mode is at the higher position of 520 cm^{-1} , reflecting an earlier report that the width of the disulphide mode was an important indicator of fracture risk [3].

The mean discriminant score for the cases was significantly higher than the controls ($p = 0.004$) and at all Raman score values the cumulative proportion of controls exceeded that of the cumulative proportion of the cases (Figure 2). Using the Raman discriminant score alone the odds ratio (OR) for hip fracture was 1.59 with an area under the curve (AUC) of 0.61. Combining scores with CRFs improved performance to $\text{OR} = 2.5$, $\text{AUC} = 0.75$, which was a significant improvement over using the CRFs alone ($p = 0.002$).

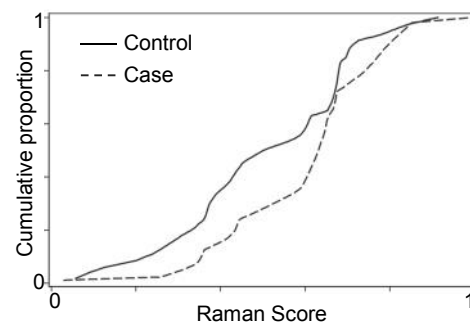


Figure 2 Cumulative proportion of case and controls with increasing Raman score.

However, this analysis included samples from patients that fractured up to 20 years after the sample was taken, considerably longer than would clinically relevant to detect. Excluding fractures over 15 years post sampling (and their matching controls) obtained an OR of 3.56 and AUC of 0.79. The improvement in the prediction compared to using CRFs alone indicates that the Raman method adds significant new information to the assessment of fracture risk and provides a significant performance enhancement. The mean age of the women in the study was 57.7 years of age and is able to provide a prognostic indicator of fracture risk up to 15 years in advance. It is hoped that such early detection of patients most at risk of fracture, will provide substantial time for appropriate interventions to prevent further bone quality deterioration.

CONCLUSION

The study demonstrated Raman spectroscopic evaluation of human toenails provides early prognostic indication of fracture risk, up to 15 years before the event and the Raman information improves the identification of at risk patients. Earlier and more reliable detection of at-risk patients allows appropriate and targeted intervention against age related fractures.

AUTHOR AFFILIATIONS

¹. Crescent Diagnostics, NOVA-UCD, Belfield, Dublin 4, Ireland. ². Inamori School of Engineering, Alfred University, NY, USA ³. Women's Hospital and Harvard Medical School, Boston MA, USA ⁴. McDougall Scientific Ltd. Toronto, Canada ⁵. Materials & Surface Science Institute, University of Limerick, Ireland ⁶. River Diagnostics BV, Rotterdam, Netherlands.

Email: rene.beattie@crescentdx.com

REFERENCES

1. "Assessment of Osteoporosis at the Primary Health Care Level", J. Kanis, WHO (2007)
2. "The use of fingernails as a means of assessing bone health: A pilot study", I. Pillay, D. Lyons, M.J. German, N.S. Lawson, H.M. Pollock, J. Saunders, S. Chowdhury, P. Moran, M.R. Towler, Journal of Women's Health **14**, 339 (2005)

3. “Raman spectroscopy of the human nail: A potential tool for evaluating bone health?”, M.R. Towler, A. Wren, N. Rushe, J. Saunders, N.M. Cummins, P.M. Jakeman, *J Mater Sci* **18**, 759 (2007)

APPENDIX C

Developing Novel Prognostic Biomarkers for Multivariate Fracture Risk Prediction Algorithms

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

Received: 25 April 2012 / Accepted: 15 June 2012
© Springer Science+Business Media, LLC 2012

Abstract Multivariate prediction algorithms such as FRAX® and QFractureScores provide an opportunity for new prognostic biomarkers to be developed and incorporated, potentially leading to better fracture prediction. As more research is conducted into these novel biomarkers, a number of factors need to be considered for their successful development for inclusion in these algorithms. In this review, we describe two well-known multivariate prediction algorithms for osteoporosis fracture risk applicable to the UK population, FRAX and QFractureScores, and comment on the current prognostic tools available for fracture risk; dual X-ray assessment, quantitative ultrasound, and genomic/biochemical markers. We also highlight the factors that need to be considered in the development of new biomarkers. These factors include the requirement for prospective data, collected in new cohort studies or using archived samples; the need for adequate stability data to be provided; and the need for appropriate storage methods to be used when retrospective data are

required. Area under the receiver operating characteristic curve measures have been found to have limited utility in assessing the impact of the addition of new risk factors on the predictive performance of multivariate algorithms. New performance evaluation measures, such as net reclassification index and integrated discrimination improvement, are increasingly important in the evaluation of the impact of the addition of new markers to multivariate algorithms, and these are also discussed.

Keywords Algorithms · DXA · Fracture · FRAX · Prognostic

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

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

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

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

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

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

Traditionally, osteoporosis has been defined using bone mineral density (BMD) as measured by T-scores. In recent years, there has been a move away from T-scores as the operating definition of osteoporosis, in favor of the use of absolute risk of fracture and risk calculators that are based on algorithms that estimate those risks [3]. This movement has led to an evolved definition of the disease that incorporates more clinical risk factors (CRF) and that bases treatment decisions on absolute risk of fracture thresholds over a 10 year period rather than T-scores [3]. This

paradigm shift brings osteoporosis into line with other conditions, such as heart disease, where patient risks are assessed on an absolute risk-of-event basis over a 10 year period [4]. This major change in the definition of osteoporosis creates an opportunity for new prognostic biomarkers to be identified and incorporated into risk assessment quickly and efficiently via the absolute risk approach once they meet the appropriate clinical evidence requirements.

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

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

Current Prognostic Fracture Risk Calculators

The increased interest in regression model-based risk calculators developed from established cohort studies has been driven by the need to develop more accurate models for who is likely to experience a fracture and when the fracture will occur. Another important issue is the lack of availability of dual X-ray assessment (DXA) machines in many countries. Also, DXA performance is not optimal for detecting osteoporotic fracture risk as a result of poor predictive sensitivity, and therefore the use of additional CRFs in combination with DXA could help increase the sensitivity of diagnosis without impairing specificity [8].

Health economic evaluations have indicated that it is most effective to implement mass screening programs using an initial assessment with CRFs followed by DXA evaluation in high-risk subjects [9]. Mass screening can therefore be justified with the support of non-BMD prognostic markers to enhance overall prognostic performance in combination with DXA. Early work to combine CRFs into prediction models for fracture risk to supplement DXA was conducted by Black et al. [10]. Subsequent work has resulted in three validated fracture risk prediction models that are currently available online: FRAX, QFractureScores, and the Garvan model. The Garvan and Black models were developed in Australian and U.S. populations, respectively. The FRAX model is currently the most widely used. In order to provide an illustrative comparison, two of these models, FRAX and QFractureScores, both of which are available for UK populations, are described in more detail.

FRAX

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

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

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

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

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

^a In QFractureScores, the use of “steroids” is recorded rather than glucocorticoids

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

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

hip fracture rises significantly in women aged between 70 and 90, and clinical studies indicate that between these ages, the prognostic performance of BMD as determined by DXA falls by more than the performance of CRFs [7]. There may therefore be an argument to focus on CRFs and exclude BMD as a risk factor when identifying elderly women who would benefit from treatment.

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

QFractureScores

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

The hazard ratios (HR) and coefficients in the model were derived by the Cox proportional hazard regression model. To validate the QFractureScores model, hip fracture prognostic performance in a separate defined QResearch group was compared with the actual events over a 10 year period and with the predictions generated by FRAX in the same cohort. The validation group contained 653,789 women, and the average hip fracture incidence rate was 1.15 % (range 1.13–1.17 %) [15]. QFractureScores has also been externally validated in a UK-based population using records in The Health Improvement Network (THIN) database (<http://www.thin-uk.com>), which added

an additional 13 million observation years. The observed results closely matched those observed in the internal validation study, adding further evidence for the integrity of the QFractureScores approach [16]. The developers of QFracture have recently released a new algorithm incorporating additional risk factors such as ethnicity and previous fracture on the basis of their analysis of the prospective cohort study, QResearch, which has improved predictive performance over the original QFracture algorithm [17].

A Comparison of FRAX and QFractureScores

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

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

Table 2 Comparison of outcomes

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

Current Prognostic Biomarkers

DXA

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

Quantitative Ultrasonography

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

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

Biochemical Markers

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

Genomic Markers

Osteoporosis is a polygenic disease, involving a large variety of gene products implicated in both bone modeling

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

Development of New Prognostic Biomarkers

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

Importance of Cohort Studies

In order to develop a completely novel prognostic marker for osteoporotic fracture risk, there is a requirement to collect patient samples at baseline and then to follow the patient for a number of years. Because of the low incidence rate of hip fractures in postmenopausal women (less than 5 %), cohorts in excess of 10,000 subjects could be required

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

In order to enable a retrospective study to be carried out in a timely and cost-effective manner without having to test tens of thousands of archived samples, nested case-control designs are attractive. Sample types previously collected in published osteoporosis studies have included DXA scores as well as bone, blood, urine, and skin samples [40–42]. Because the incidence rate of hip fracture is less than 3 % in the age range with most clinical utility, 50–70 years of age, the use of case-to-control ratios of 1:3 or more is recommended [43]. As an example of this nested case-control approach, envisage a scenario where a new

technique has been developed that can extract bone quality information from X-ray images. If we assume that a 1:3 case-to-control ratio is sufficient, rather than conducting a completely new study, we instead could retrospectively examine archived X-ray images from 100 fracture events and 300 controls. These 400 data points could then be evaluated more cost- and time-effectively than the traditional prospective approach.

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

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

Sample Stability Considerations

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

Performance Evaluation Measures

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

New approaches to evaluate the performance of additional clinical markers in a multivariate model have been proposed that move beyond whether a new prognostic test offers good discrimination between the cases and controls (as evaluated by the ROC curve) to whether it significantly changes the classification of the subject who may be at risk [4]. The addition of risk factors into prognostic models can result in small changes in the AUC, which do not reflect the

changes in risk category that result from the new information [46]. Cook notes that many new biomarkers may have clinically relevant odds ratios (between 1.5 and 2.0), but these will have only a modest impact on the ROC curves [46]. New reclassification metrics are able to address the weaknesses of these measures in terms of perfect discrimination using ROC curves. It is now being argued that these new measures are more important to prognostic models than the traditional ROC curve and AUC measure.

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

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

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

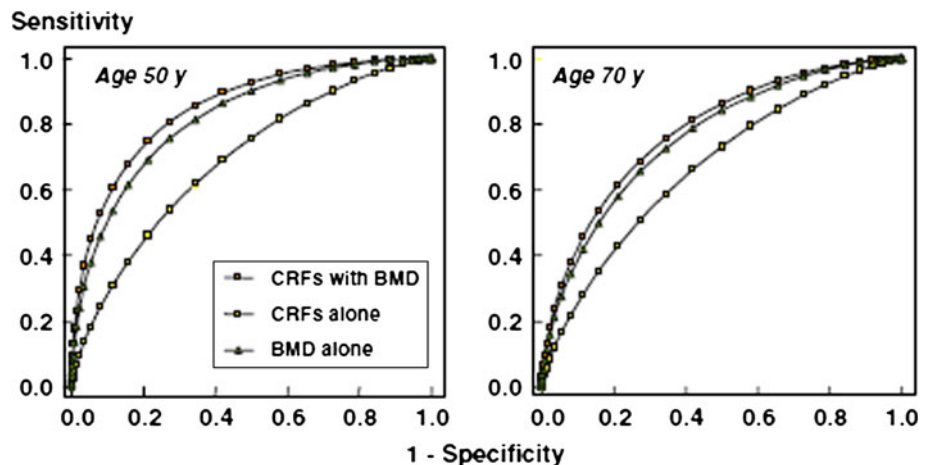


Fig. 2 Predicted to observed risk of hip fracture using QFractureScores [15]

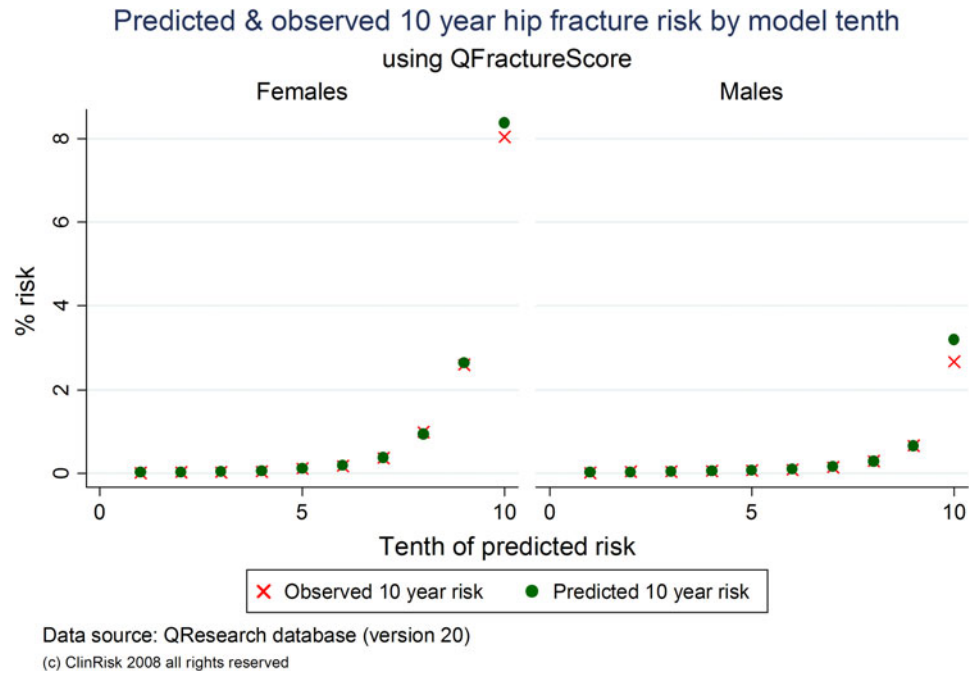
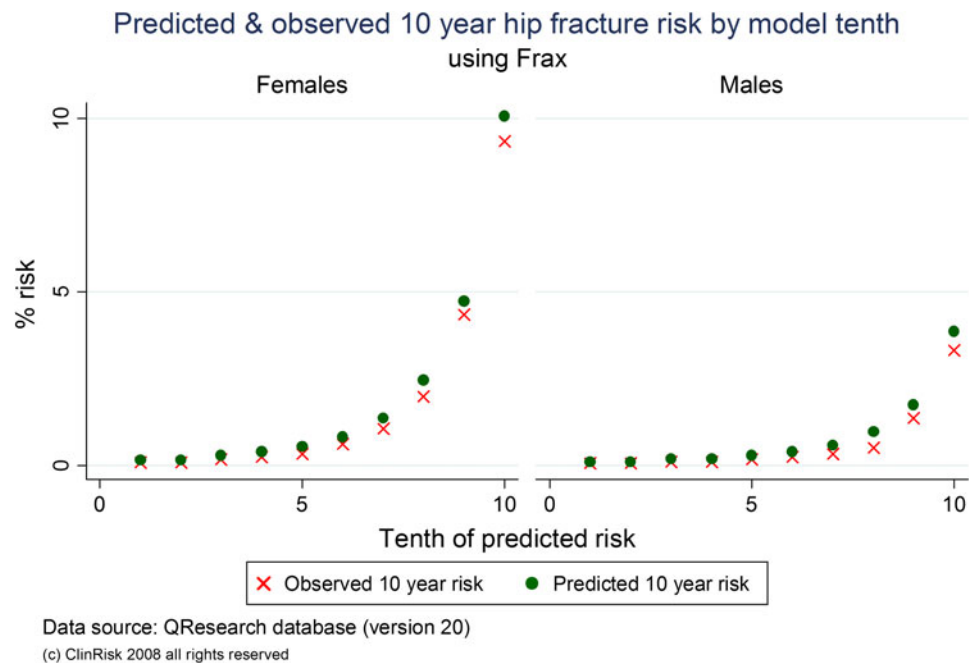


Fig. 3 Predicted to observed risk of hip fracture using the FRAX® [15]



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

It is notable that the authors of these articles state that these new measures can be used to evaluate whether an expensive new biomarker should be introduced from an economic standpoint. However, the criteria to evaluate this have not yet been published in detail. Published health

economic studies typically use odds ratio, relative risk, or AUC to evaluate the economic performance of tests [51]. It is clear from recent work that the limited impact of some new additive tests on AUC measures restrains the ability of clinical practitioners to evaluate cost-effectiveness because the AUC is not taking the reclassification of subjects into account. A recent study has explored the possibility of evaluating cost-effectiveness using NRI as an alternative to traditional relative risk based approaches and have

investigated how measures of discrimination, classification, and costs can be linked [52]. Pencina et al. [50] offer a process that weighs the NRI on the basis of the cost saving when a person moves up in classification compared to incurred costs when they move down in classification, caused by misclassification. An example would be when a person no longer receives unnecessary treatment as a result of reclassification.

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

Conclusions

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

Several alternative prognostic biomarkers have been evaluated to date, but as yet none has provided the evidence base to supersede DXA. It may be that the way forward now is to use these tools in combination with DXA and, where cost-effective, as a prescreening tool to select subjects for DXA testing. This review has set out some of the considerations required by researchers seeking to

incorporate new risk factors into the existing prediction algorithms. There is significant scope in the field of osteoporosis for the following: increased use of real patient data, increased use of archived samples, increased use of end points with real clinical utility, i.e., hip fracture, and for prognostic-based end points like NRI and IDI to be applied. These new techniques could ultimately lead to the development of a new generation of prognostic tools to improve patient care for people with osteoporosis.

References

1. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E (2008) FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 19:385–397
2. Bouxsein ML (2003) Bone quality: Where do we go from here? *Osteoporos Int* 14(suppl 5):S118–S127
3. McCloskey EV, Johansson H, Oden A, Kanis JA (2009) From relative risk to absolute fracture risk calculation: the FRAX algorithm. *Curr Osteoporos Rep* 7:77–83
4. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 115:928–935
5. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A (2007) Incidence and economic burden of osteoporosis-related fractures in the United States, 2005–2025. *J Bone Miner Res* 22:465–475
6. Abrahamsen B, van Staa T, Ariely R, Olson M, Cooper C (2009) Excess mortality following hip fracture: a systematic epidemiological review. *Osteoporos Int* 20:1633–1650
7. Kanis J (2007) Assessment of osteoporosis at the primary health care level. World Health Organization, Lyon
8. Kanis JA, Oden A, Johansson H, Borgström F, Ström O, McCloskey E (2009) FRAX® and its applications to clinical practice. *Bone* 44:734–743
9. Kanis JA, Johnell O (2005) Requirements for DXA for the management of osteoporosis in Europe. *Osteoporos Int* 16: 229–238
10. Black DM, Steinbuch M, Palermo L, Dargent-Molina P, Lindsay R, Hoseney MS, Johnell O (2001) An assessment tool for predicting fracture risk in postmenopausal women. *Osteoporos Int* 12:519–528
11. Johansson H, Kanis JA, Oden A, Johnell O, McCloskey E (2009) BMD, clinical risk factors and their combination for hip fracture prevention. *Osteoporos Int* 20:1675–1682
12. Johnell O, Kanis JA, Oden A, Johansson H, De Laet C, Delmas P, Eisman JA, Fujiwara S, Kroger H, Mellstrom D, Meunier PJ, Melton LJ 3rd, O'Neill T, Pols H, Reeve J, Silman A, Tenenhouse A (2005) Predictive value of BMD for hip and other fractures. *J Bone Miner Res* 20:1185–1194
13. Garton MJ, Cooper C, Reid D (1997) Perimenopausal bone density screening—Will it help prevent osteoporosis? *Maturitas* 26:35–43
14. Barr RJ, Stewart A, Torgerson DJ, Reid DM (2010) Population screening for osteoporosis risk: a randomised control trial of medication use and fracture risk. *Osteoporos Int* 21:561–568
15. Hippisley-Cox J, Coupland C (2009) Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 339:1291–1295
16. Collins GS, Mallett S, Altman DG (2011) Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective

- independent and external validation of QFractureScores. *BMJ* (in press)
17. Hippisley-Cox J, Coupland C (2012) Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* (in press)
18. Kanis JA, Oden A, Johansson H, McCloskey E (2012) Pitfalls in the external validation of FRAX. *Osteoporos Int* 23:423–431
19. Cummins NM, Poku EK, Towler MR, O'Driscoll OM, Ralston SH (2011) Clinical risk factors for osteoporosis in Ireland and the UK: a comparison of FRAX and QFractureScores. *Calcif Tissue Int* 89:172–177
20. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, Burckhardt P, Cooper C, Christiansen C, Cummings S, Eisman JA, Fujiwara S, Gluer C, Goltzman D, Hans D, Krieg MA, La Croix A, McCloskey E, Mellstrom D, Melton LJ 3rd, Pols H, Reeve J, Sanders K, Schott AM, Silman A, Torgerson D, van Staa T, Watts NB, Yoshimura N (2007) The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 18: 1033–1046
21. Schousboe JT (2008) Cost effectiveness of screen-and-treat strategies for low bone mineral density: How do we screen, who do we screen and who do we treat? *Appl Health Econ Health Policy* 6:1–18
22. Stewart A, Kumar V, Reid DM (2006) Long-term fracture prediction by DXA and QUS: a 10-year prospective study. *J Bone Miner Res* 21:413–418
23. Hans D, Krieg MA (2009) Quantitative ultrasound for the detection and management of osteoporosis. *Salud Publica Mex* 51(suppl 1):S25–S37
24. Bauer DC, Gluer CC, Cauley JA, Vogt TM, Ensrud KE, Genant HK, Black DM (1997) Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women: a prospective study. *Arch Intern Med* 157:629–634
25. Langton CM, Langton DK (2000) Comparison of bone mineral density and quantitative ultrasound of the calcaneus: site-matched correlation and discrimination of axial BMD status. *Br J Radiol* 73:31–35
26. Zhu K, Devine A, Prince RL (2009) Quantitative ultrasound measurements predict fracture in older women: a 10-year longitudinal study. *Bone* 44:S118
27. Moayyeri A, Kaptoge S, Dalzell N, Bingham S, Luben RN, Wareham NJ, Reeve J, Khaw KT (2009) Is QUS or DXA better for predicting the 10-year absolute risk of fracture? *J Bone Miner Res* 24:1319–1325
28. Garnero P, Hausherr E, Chapuy M-, Marcelli C, Grandjean H, Muller C, Cormier C, Bréart G, Meunier PJ, Delmas PD (1996) Markers of bone resorption predict hip fracture in elderly women: the EPIDOS prospective study. *J Bone Miner Res* 11:1531–1538
29. Vasikaran S, Eastell R, Bruyère O, Foldes AJ, Garnero P, Griesmacher A, McClung M, Morris HA, Silverman S, Trenti T, Wahl DA, Cooper C, Kanis JA (2011) Markers of bone turnover for the prediction of fracture risk and monitoring of osteoporosis treatment: a need for international reference standards. *Osteoporos Int* 22:391–420
30. Garnero P, Sornay-Rendu E, Chapuy MC, Delmas PD (1996) Increased bone turnover in late postmenopausal women is a major determinant of osteoporosis. *J Bone Miner Res* 11:337–349
31. Ralston SH (2005) Genetic determinants of osteoporosis. *Curr Opin Rheumatol* 17:475–479
32. Arden NK, Baker J, Hogg C, Baan K, Spector TD (1996) The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. *J Bone Miner Res* 11:530–534
33. Mann V, Ralston SH (2003) Meta-analysis of *COL1A1* Sp1 polymorphism in relation to bone mineral density and osteoporotic fracture. *Bone* 32:711–717
34. Navarro MC, Sosa M, Del Pino-Montes J, Torres A, Salido E, Saavedra P, Corral-Gudino L, Montilla CA (2007) Collagen type 1 (*COL1A1*) Sp1 binding site polymorphisms is associated with osteoporotic fractures but not with bone density in post-menopausal women from the Canary Islands: a preliminary study. *Aging Clin Exp Res* 19:4–9
35. Klee EW, Hoppman-Chaney NL, Ferber MJ (2011) Expanding DNA diagnostic panel testing: Is more better? *Exp Rev Mol Diagn* 11:703–709
36. Brown MA (2005) Genetic studies of osteoporosis—a rethink required. *Calcif Tissue Int* 76:319–325
37. Barr R, Macdonald H, Stewart A, McGuigan F, Rogers A, Eastell R, Felsenberg D, Glüer C, Roux C, Reid DM (2010) Association between vitamin D receptor gene polymorphisms, falls, balance and muscle power: results from two independent studies (APOSS and OPUS). *Osteoporos Int* 21:457–466
38. Robbins JA, Schott AM, Garnero P, Delmas PD, Hans D, Meunier PJ (2005) Risk factors for hip fracture in women with high BMD: EPIDOS study. *Osteoporos Int* 16:149–154
39. Shepstone L, Fordham R, Lenaghan E, Harvey I, Cooper C, Gittos N, Heawood A, Peters TJ, O'Neill T, Torgerson D, Holland R, Howe A, Marshall T, Kanis JA, McCloskey E (2012) A pragmatic randomised controlled trial of the effectiveness and cost-effectiveness of screening older women for the prevention of fractures: rationale, design and methods for the SCOOP study. *Osteoporos Int* (in press)
40. Elliott P, Peakman TC (2008) The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 37:234–244
41. Patel R, Blake GM, Fogelman I (2007) Evaluation of osteoporosis using skin thickness measurements. *Calcif Tissue Int* 81:442–449
42. Saito M, Marumo K (2010) Collagen cross-links as a determinant of bone quality: a possible explanation for bone fragility in aging, osteoporosis, and diabetes mellitus. *Osteoporos Int* 21:195–214
43. Langholz B, Thomas DC (1990) Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol* 131:169–176
44. Kenis G, Teunissen C, De Jongh R, Bosmans E, Steinbusch H, Maes M (2002) Stability of interleukin 6, soluble interleukin 6 receptor, interleukin 10 and CC16 in human serum. *Cytokine* 19:228–235
45. McClish DK (1989) Analyzing a portion of the ROC curve. *Med Decis Making* 9:190–195
46. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 54:17–23
47. Rapsomaniki E, White IR, Wood AM, Thompson SG, Emerging Risk Factors Collaboration (2012) A framework for quantifying net benefits of alternative prognostic models. *Stat Med* 31:114–130
48. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 26:565–574
49. Pepe MS, Feng Z, Gu JW (2008) Comments on “Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond,” by M. J. Pencina et al. *Stat Med* 27:173–181. doi:10.1002/sim.2929
50. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 27:157–172
51. Schwenkglenks M, Lippuner K (2007) Simulation-based cost-utility analysis of population screening-based alendronate use in Switzerland. *Osteoporos Int* 18:1481–1491

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

APPENDIX D

Osteoporotic genetic markers: can they improve individualised fracture risk prediction?

J.Quigley. MSc Thesis

Osteoporosis is a metabolic bone disorder characterised by compromised bone strength that predisposes a person to an increased risk of fragility fracture. A significant proportion of patients who suffer osteoporotic fractures are not diagnosed until after the event has occurred. Therefore, one of the foremost priorities in osteoporotic research is to develop valid models for identifying individuals at high risk of fracture so preventive measures can be taken. The risk of fracture is determined by genetic and non-genetic clinical risk factors. The aim of this study was to quantify the contribution of genetic profiling to fracture prognosis, over and above the contribution of clinical risk factors. The study was built on data previously collected in Ireland and the UK. Clinical data from one hundred women aged fifty to eighty five years who had recently suffered a low-trauma fracture was used, as well as data from one hundred equivalent controls. Genotypes for thirteen independent single nucleotide polymorphisms (SNP) were simulated from population level data obtained from distinct published sources and using two different methods; HAP-SAMPLE simulation and allele frequency simulation. Five models were built to test the potential; (1) clinical risk factors alone, (2) genes alone in the HAP-SAMPLE population, (3) genes alone in the frequency population, (4) genes and clinical risk factors in the HAP-SAMPLE population and (5) genes and clinical risk factors in frequency population. The area under the curve (AUC) for model 4 and model 5 showed improvements of 4.63% and 8.56% respectively, from the AUC for model 1. Net reclassification improvements (NRI) of 0.03 and 0.06 respectively, were also recorded. These results suggest that genetic profiling could enhance the predictive accuracy of fracture prognosis models.

Keywords:

Osteoporosis, fragility fracture, risk prediction model, absolute risk, single nucleotide polymorphism (SNP), clinical risk factor