Bone Ageing and Osteoporosis: Automated DXA Image Analysis for Population Imaging

By:

M. Farzi

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

The University of Sheffield
Faculty of Medicine
Department of Oncology and Metabolism

September, 2018

Bone Ageing and Osteoporosis: Automated DXA Image Analysis for Population Imaging

By:

Mohsen Farzi

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

The University of Sheffield
Faculty of Medicine
Department of Oncology and Metabolism

September, 2018

Abstract

Osteoporosis is an age-associated bone disease characterised by low bone mass. The consequent fragility fractures with increased follow-up mortality and morbidity underlie the clinical significance of osteoporosis in public health. However, current diagnostic criteria using bone mineral density (BMD) at the femoral neck at most can identify half of the fragility fractures, and thereby the ability to provide new metrics capturing the bone strength beyond neck BMD remains of interest in osteoporosis research.

This study aims to, first, quantify pixel BMD at anatomically corresponding locations in the femur; second, model the evolution of spatial BMD patterns with ageing; and third, characterise how trabecular and cortical bone arrangements change at different stages of osteoporosis progression.

To construct the atlas, a novel cross-calibration procedure is proposed to integrate data from different DXA manufacturers into an amalgamated large-scale dataset ($n > 13000$). A new technique, termed region free analysis (RFA), is proposed to eliminate morphological variation between scans by warping each image into a reference template. This image warping establishes a correspondence between pixel coordinates that allows modelling pixel BMD evolution with ageing using smooth quantile curves. Given access to large-scale datasets, automatic quality control of DXA scans has been identified as an emerging challenge to the community for which an unsupervised, non-distortion-specific, opinion-free framework was proposed.

The developed atlas usefully added to our understanding of spatial BMD patterns and their relationship with osteoporosis. The concept of osteoporosis progression is introduced by proposing bone age as the age at which an individual bone map best fits the constructed atlas. Normalising BMD maps for bone age, local fracture-specific patterns were identified. The proposed framework in this thesis constitutes a first step toward modelling osteoporosis progression to identify better bone-based risk factors for prediction of fragility fractures.

Key Words: Dual-energy X-ray Absorptiometry (DXA), Region Free Analysis (RFA), Osteoporosis, Disease Progression Estimation, Atlas Development

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Mark Wilkinson and Prof. Alejandro Frangi for their generous support, guidance, and contributions without them I would not have this achievement. Mark, your attention to details and strict practice for keeping deadlines helped me to stay focus throughout the PhD. Alex, your passion and inspiration to go beyond classical image processing tasks kept me motivated during my PhD. With this thesis at the interface of clinical medicine and engineering, it was a privilege and honour for me to work with you both.

I would like to thank Prof. Eugene McCloskey and Prof. Richard Eastell for their insightful advice on the development of the bone ageing atlas. Eugene, thank you for providing access to the MRC-Hip database and your time and expertise to discuss DXA artefacts. Richard, thank you for providing access to the OPUS database and your time and expertise to discuss DXA calibration. I would like to thank Dr. Lang Yang without his help I could not demystify pixel BMD extraction. I would like to thank Dr. Margaret Paggiosi for her expertise to identify various types of DXA artefacts and her help to schedule the productive bone imaging meetings.

I would like to express my special gratitude to Dr. Jose Pozo for his excellent co-supervision in this project. Jose, it was always a pleasure to go through technical details with you. Your thoughtful and critical reviews helped me to develop a deeper understanding of the algorithms.

During my PhD, I had the privilege to be part of two top-leading research group centres: The Mellanby Centre for bone research and CISTIB for computational imaging and simulation technologies in biomedicine. I am grateful to my colleagues at both centres for their insights and help.

Collection of more than 13000 scans to build the ageing atlas was a difficult task which would not have happened without the contribution of several people. Thank you to Diane Charlesworth at Northern General Hospital who helped to restore archived Hologic files for the MRC-Hip database. Thank you to the OPUS Steering Committee for permission to use this dataset. Thank you to Steve Garratt who facilitated access to the UK Biobank dataset.

This work was kindly funded by Medical Research Council-Arthritis Research UK Centre for Integrated research into Musculoskeletal Ageing (CIMA) through a PhD Fellowship. I am grateful for this generous funding which made this project possible.

Last but not the least, I would like to thank my parents and my two beloved brothers for supporting me spiritually throughout writing this thesis and my life in general.

Publications

International Journals

1. **Mohsen Farzi**, Richard M. Morris, Jeannette Penny, Lang Yang, Jose M. Pozo, Soren Overgaard, Alejandro F. Frangi, and J. Mark Wilkinson, "Quantitating the effect of prosthesis design on femoral remodelling using high-resolution region-free densitometric analysis (DXA-RFA)," *Journal of Orthopaedic Research*, vol. 35, no. 10, pp. 2203–2210, 2017.

2. Andrew M.Parker, Lang Yang, **Mohsen Farzi**, Jose M. Pozo, Alejandro F. Frangi, and J. Mark Wilkinson, "Quantifying pelvic periprosthetic bone remodeling using dual-energy X-ray absorptiometry region free analysis," *Journal of Clinical Densitometry*, vol. 20, no. 4, pp. 480–485, 2017.

Peer-Reviewed Conferences

1. **Mohsen Farzi**, Jose M. Pozo, Eugene McCloskey, J. Mark Wilkinson, and Alejandro F. Frangi, "Automatic Quality Control for Population Imaging: A Generic Unsupervised Approach," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI2016). Springer, pp. 291–299, 2016.

2. **Mohsen Farzi**, Jose M. Pozo, Eugene McCloskey, Richard Eastell, J. Mark Wilkinson, and Alejandro F. Frangi, "Spatio-Temporal Atlas of Bone Mineral Density Ageing," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI2018). Springer, pp. 720–728, 2018.

Selected Peer-Reviewed Abstracts

1. **Mohsen Farzi**, Jose M. Pozo, Lang Yang, Alejandro F. Frangi, and J. Mark Wilkinson, *Quantitating prosthesis design effect on femoral remodelling using DXA-RFA with FDR analysis*, British Orthopaedic Research Society (BORS) annual meeting, Liverpool, UK, September 2015. **(oral)**

2. **Mohsen Farzi**, Jose M. Pozo, Eugene McCloskey, Alejandro F. Frangi, and J. Mark Wilkinson, *Automatic quality control for population imaging*, British Orthopaedic Research Society (BORS) annual meeting, Glasgow, UK, September 2016. **(poster)**

3. **Mohsen Farzi**, Jose M. Pozo, Eugene McCloskey, Alejandro F. Frangi, and J. Mark Wilkinson, *Age-specific bone density distribution in the femur*, British Orthopaedic Research Society (BORS) annual meeting, London, UK, September 2017. (**oral**)

4. **Mohsen Farzi**, Jose M. Pozo, Eugene McCloskey, Richard Eastell, Alejandro F. Frangi, and J. Mark Wilkinson, *Bone ageing atlas development: the UK Biobank study*, British Orthopaedic Research Society (BORS) annual meeting, Leeds, UK, September 2018. (**oral**)

Contents

1 Introduction 1

1.1 Importance of Osteoporosis in Clinical Practice 1

1.2 Bone Quality 3

1.3 DXA in Clinical Practice 5

1.3.1 Principles of DXA 6

1.3.2 Volumetric BMD with DXA 8

1.3.3 Shape and Geometry 9

1.3.4 Trabecular Bone Microarchitecture 12

1.3.5 DXA Region Free Analysis 14

1.4 Rationale and Motivation 15

1.5 Objectives of This Thesis 16

1.6 Thesis Structure 17

2 False Discovery Rate Integration with Region Free Analysis 19

2.1 Introduction 20

2.2 Multiple Comparison Problem 21

2.2.1 Bonferroni Correction 23

2.2.2 Random Field Theory 24

2.2.3 False Discovery Rate Analysis 29

2.3 Periprosthetic BMD Analysis 30

2.3.1 Motivation 30

2.3.2 Study Populations and Scan Acquisitions 31

2.3.3 Study Design 32

2.4 Results 33

2.4.1 FDR Validation 33

2.4.2 Clinical Trial Subject Characteristics 34

2.4.3 Post-Operative Baseline Mean BMD Distribution 35

2.4.4 Effect of Cemented Stem Design on Bone Remodelling . 36

2.4.5 Effect of Hip Resurfacing Versus Cementless THA on Bone Remodelling 38

2.5 Discussion 40
2.6 Conclusion 42

3 Development of a Spatio-Temporal Atlas of Ageing Bone in the Native Proximal Femur 43
3.1 Introduction 44
3.2 Bone Ageing Analysis Pipeline 46
  3.2.1 Pre-Processing and Data Organisation 47
  3.2.2 Region Free Analysis 49
  3.2.3 Comparative Calibration 55
  3.2.4 Quantile Regression 61
3.3 Results 66
  3.3.1 Datasets 66
  3.3.2 Precision Analysis 67
  3.3.3 Parameter Estimation for Comparative Calibration between DXA Systems 69
  3.3.4 Bilateral Calibration 70
  3.3.5 The Spatio-Temporal Atlas 72
3.4 Validation of the Atlas Construction Steps 74
  3.4.1 Segmentation Accuracy 74
  3.4.2 Point Localisation Accuracy 76
  3.4.3 Experimental Validation for the Quantile Matching Regression Technique 77
  3.4.4 Compliance with Normality after the Box-Cox Transformation 78
3.5 Atlas Validation using Longitudinal Data 79
3.6 Discussion 83
3.7 Conclusion 86

4 Application of Bone Atlas to Understand Ageing and Osteoporosis 87
4.1 Introduction 88
4.2 Cortical and Trabecular BMD Variation 90
4.3 Osteoporosis Progression Index 93
  4.3.1 Consistency with Current Diagnostic Guidelines 96
  4.3.2 Bone Ageing Precision 97
  4.3.3 Distinction between a Young Healthy Cohort and an Elderly Population 98

4.3.4 Fracture Risk and Bone Age 98
4.3.5 Fracture Prediction 100
4.4 Localised Fracture-Specific Bone Patterns 101
4.5 BMI Impact on BMD Distribution 105
4.6 Discussion 105
4.7 Conclusion 108

5 Automatic Quality Control of DXA Images 111
5.1 Introduction 112
5.2 Unsupervised Non-Distortion-Specific Image Quality Control Framework 114
5.2.1 Background 115
5.2.2 Robust Dictionary Learning 115
5.2.3 Optimal Image Coverage 116
5.2.4 Artefact Detection 117
5.3 DXA Artefacts 118
5.4 Experiments and Results 121
5.5 Conclusion 123

6 Conclusions and Future Work 125
6.1 Overview 125
6.2 Thesis Contributions 128
6.3 Limitations 129
6.4 Future Work 130
6.4.1 Extending the Bone Ageing Atlas 130
6.4.2 Osteoporosis Progression Model 131
6.4.3 Predicting Local Fracture Patterns 132
6.5 Conclusion 133

Bibliography 134

List of Figures

1.1 Bone strength determinants and fragility fracture risk factors. Several non-BMD clinical risk factors are identified for fragility fractures [17]. Bone strength has an important influence on the chance of developing a new fracture, but could be targeted as the important outcome by itself. Bone strength is the result of an inextricable relationship between both architectural and material properties [20, 21]. 3

1.2 A schematic stress-strain curve of bone in tension. The initial part of the curve is almost linear where the slope of the line defines the Young's modulus (stiffness). The area under the curve equals the amount of energy used to deform the bone. The blue dashed area defines toughness (the amount of energy absorbed before fracture) and the orange shaded area defines resilience (the amount of energy absorbed before the yield point). The height of the curve at the fracture point defines strength (the maximum stress bone sustains before fracture). 4

1.3 Geometrical hip indices. (A-C) Hip axis length is the linear distance from the inner acetabulum surface to the lateral margin of the femoral shaft below the greater trochanter drawn through a midpoint in the femoral neck parallel to the cortices of the femoral neck. (B-C) Femoral neck axis length is the linear distance from the apex of the femoral head to the base of the greater trochanter. This index is similar to the hip axis length but does not include the acetabular portion. (θ) Neck-shaft angle is the angle between the shaft axis and the femoral neck axis. (E-F) Femoral neck diameter is measured as the length of the line perpendicular to the femoral neck axis passing through the centre of the femoral neck. (F-G) Diaphysis diameter is the width of the femoral shaft below the lesser trochanter. 10

1.4 DXA region free analysis (DXA RFA) pipeline [36]. Bone maps are restored semi-automatically from the raw files (with .p and .r extensions) using a Matlab script. Sixty three control points are automatically selected on the bone contour. A mean shape is generated using generalised Procrustes analysis to serve as the reference template. Each bone map is then individually warped into this template to remove the morphological variation between scans. This image warping establishes a correspondence between pixel BMD values across the population. 14

1.5 Seven Gruen zones commonly used as ROIs for analysis of periprosthetic BMD after total hip arthroplasty. 15

2.1 (a) A typical Z-map of size 100 × 100 with FWHM = 20. (b) The family-wise error rate is plotted as a function of the height threshold t_α. To control FWE rate at 0.05, a height threshold of 3.38 should be chosen, which is smaller than 4.42 in the Bonferroni correction. 26

2.2 The Euler characteristic (EC) for two arbitrary excursion sets. Panel (a) shows an excursion set with 4 blobs without any hole, and so EC = 4 - 0 = 4. Panel (b) shows an excursion set with 4 blobs and 1 hole, and so the EC $= 4 - 1 = 3$. 27

2.3 For a three-dimensional search volume, the Euler characteristic is computed by summing the contributions from all cubes in the region. An arbitrary cube with eight voxels is plotted to illustrate this calculation. Each solid circle represents one voxel inside the excursion set, while each hollow circle represents a voxel outside the set. For this example, $V = 5$, $E = 5$, $F = 1$, and $C = 0$ (see Eq. 2.12). Hence, $\chi = \frac{-1}{8}$. 28

2.4 The expected Euler characteristic $E(\chi)$ is plotted against the height threshold t for an SPM of square shape (100 × 100 pixels) with FWHM = 6. At large threshold values, $E(\chi)$ would approximate FWE rate. 28

2.5 Theoretical versus experimental computation of family-wise error rate. The blue solid line represents the estimated FWE rate using Eq. 2.7 versus the experimental value (red dashed line) computed based on 10,000 SPMs with a normal distribution. FWHM = 6 and the search region is assumed to be an square of size 100 × 100 pixels. Note that the expected EC gives an upper bound on the actual FWE rate. 29

2.6 P-P plot for the FDR analysis. (a) The P-P plot for the set of 29 repositioned pairs of scans. As shown, the blue line almost perfectly follows the diagonal line of identity indicating that the null hypothesis of no change is valid in all pixels. (b) The P-P plot for Charnley prosthesis after 24 months. The blue line deviates below the line of identity, indicating the rejection of the null hypothesis. (c) All pixels below the slope-α line corresponding with the p-value less than 0.012 are statistically significant at $\alpha = 0.05$. 34

2.7 Mean pixel BMD distribution. The mean distribution of pixel BMD values at baseline measurement is shown for (a) composite-beam (Charnley), (b) double-taper (Exeter), (c) triple-taper (C-stem), (d) Bi- Metric total hip replacement, and (e) ASR hip resurfacing prosthesis designs, respectively. 35

2.8 Longitudinal pixel BMD change over 24 months for the cemented stem designs 36

2.9 FDR q-value maps after 24 months for the cemented stem designs 36

2.10 Cemented composite beam (Charnley). The first row shows the pixel-level percentage change in BMD with respect to baseline at 3, 6, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar. 37

2.11 Cemented double-taper slip (Exeter). The first row shows the pixel-level percentage change in BMD with respect to baseline at 3, 6, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar. 38

2.12 Cemented triple-taper slip (C-Stem). The first row shows the pixel-level percentage change in BMD with respect to baseline at 3, 6, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar. 38

2.13 Longitudinal mean pixel BMD change and the corresponding FDR q-value maps after 24 months are shown for Bi-Metric total hip replacement and ASR hip resurfacing prosthesis designs. BMD change is expressed as a percentage of the baseline measurement. All pixels with $q \leq 0.05$ are declared as significant events. 39

2.14 Cementless hip replacement (Bi-Metric). The first row shows the pixel-level percentage change in BMD with respect to baseline at 2, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar. 40

2.15 Cemented hip resurfacing (Articular Surface Replacement). The first row shows the pixel-level percentage change in BMD with respect to baseline at 2, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar. 40

3.1 Femoral regions of interest (ROIs). The neck, trochanteric, and intertrochanteric regions are shown in red, blue, and green, respectively. The aggregation of these three regions comprises the total hip. 45

3.2 Bone ageing analysis pipeline. Scans are automatically organised into sub-folders according to the study ID, geographic location, subject ID, anatomic site, and follow-up time points. Each scan is then warped into a reference domain to eliminate morphological variations. Pixel BMD values are calibrated across different centres such that the probability density functions match one another for a subset of samples matched for gender, age, body mass index, ethnicity, scan side, and geographic location. Finally, a set of smooth quantile curves is fitted to the standardised pixel BMD values for each pixel coordinate. 46

3.3 Median pixel-wise SNR for the Hologic QDR4500A versus the Lunar iDXA systems. The x-axis shows the standard deviation for the smoothing Gaussian kernel (σ) deployed to reduce the noise level. The y-axis shows the variation in the signal-to-noise ratio (SNR). For the Hologic system, pixel-level SNR was computed using a set of 25 scan pairs, each pair collected on the same day from the same subject with repositioning between scans (section 3.3.2). For the Lunar system, pixel-level SNR was computed using a random selection of 100 bilateral hip scans. Deming regression analysis was deployed to compute SNR for both systems (see section 3.2.3.(B)). 48

3.4 Conceptual illustration of region free analysis. Sixty-five landmark points are automatically selected around the bone contour. A reference shape is learned by averaging over all the scans after being aligned to a common position, scale, and orientation. A thin plate spline (TPS) deformation function is fitted for each individual scan such that the controlling landmark points are mapped to the corresponding reference landmark points in the template. Given the warp in the space, pixel intensities are estimated using a linear interpolation technique. 50

3.5 The first mode of shape variation in the proximal femur. 51

3.6 The check function for quantile regression with $u = 0.5$ and $u = 0.9$. To compute the u^{th} quantile of the random variable Y, $Q_Y(u)$, the expected loss $E(\rho_u(Y - \xi))$ is minimised with respect to ξ. Observe that $u = 0.5$ is equivalent to the median. . 62

3.7 DXA RFA precision analysis. (a) The pixel level CV (%) is visualised using a heat-map. Precision is worse around the bone contour. This may be due to the inaccuracy in placing controlling landmark points at the bone surface. (b) The distribution of pixel-level CV values in the femur. The median is 7.96% and the interquartile range is 6.69% – 10.05%. 69

3.8 Estimated cross-calibration parameters between the Lunar iDXA and the Hologic QDR 4500A systems. The quantile matching regression technique was applied using the Hologic system as the reference, i.e. [Lunar] = a [Hologic] + b. The average and standard deviation of the estimated parameters over all the pixel coordinates within the femur were 1.019 (SD, 0.140) for the slope a and 0.170 (SD, 0.130) for the intercept b. 69

3.9 Bilateral hip comparison. (a) The left and the right hips are highly correlated inside the femur, but the correlation is worse at the boundary. (b) Average right-left differences in pixel BMD values normalised to the population mean for the left side. (c) Localised regions with a statistically significant difference in BMD were observed between the bilateral sides using FDR analysis. (d) The pp-plot deviated from the identity line (dashed red line) demonstrating a significant difference between the bilateral sides. 70

3.10 Estimated cross-calibration parameters between the left and the right hips. The Deming regression technique with $\delta = 1$ was applied taking the right hip as the reference, i.e. [left] = a [right] + b. 71

3.11 The ability of the deployed calibration procedure to cancel the observed difference between the right and the left hips. Here, the right hip is mapped to the left side. Panel (a) shows the average differences in BMD between the left and the calibrated right hips normalised to the population mean for the left side. (b, c) No statistically significant difference in BMD was observed between the two sides following the bilateral calibration using FDR analysis at $q \leq 0.05$. 72

3.12 The Bone ageing atlas. The median together with the first and the third quartiles at each pixel coordinate is visualised using heat-maps for 20, 35, 50, 65, 80, and 95 years of age. The atlas is shown for the Lunar system at the left hip. 73

3.13 Quantile curves fitted at the femoral neck. The solid, dashed, and dotted lines show the median, the 50% and the 90% quantile ranges, respectively. The green shadow shows the 95% confidence interval. The curves are shown for the Lunar system at the left hip. 73

3.14 Quantile curves fitted at the intertrochanteric region. The solid, dashed, and dotted lines show the median, the 50% and the 90% quantile ranges, respectively. The green shadow shows the 95% confidence interval. The curves are shown for the Lunar system at the left hip. 74

3.15 The graphical user interface (GUI) developed in Matlab to facilitate manual segmentation of femoral scans. The user would select a number of control points around the bone and the software computes a smooth contour passing through the selected points. The toolkit allows the user to move the control points, delete them, or insert new ones if required. 75

3.16 The best and the worst femoral segmentation among 32 randomly selected scans from the database using the dice coefficient index as the evaluation metric. The green and the red contours show the ground truth and the automatic segmentation, respectively. 75

3.17 Point localisation error. Five landmark points were selected on the template at the centre of the femoral head; the centre, superior, and inferior positions at the femoral neck; and the apex at the greater trochanter (the red cross-marks). To assess the point localisation error, thirty-two scans were randomly selected. For each image, five landmark points were selected manually at anatomically correspondent locations and then mapped to the reference domain using the estimated TPS transformations for each image (the blue dots). The average error was 1.57 mm. The space is shown in millimetre. 76

3.18 Estimated cross-calibration parameters between the left and the right hips. The quantile matching regression technique was applied taking the right hip as the reference, i.e. [left] $= a$ [right] $+ b$. The estimated parameters are similar to those computed using the Deming regression (cf. Fig. 3.10). Over the region with a high correlation between the left and the right hips ($r^2 \geq 0.5$), the RMS error was 0.013 for the slope a and 0.017 for the intercept b, respectively 78

3.19 FDR analysis to identify pixels where the distribution of the transformed pixel BMD values using the estimated LMS model significantly deviates from a normal distribution. The learned LMS models are valid in the majority of pixels except for regions at the rim of the femoral head, and at the bone margin next to the lesser trochanter. 79

3.20 Longitudinal atlas validation (sub-group 1: 55-60 years, $n = 120$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively. . . 80

3.21 Longitudinal atlas validation (sub-group 2: 60-65 years, $n = 99$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively. . . 81

3.22 Longitudinal atlas validation (sub-group 3: 65-70 years, $n = 82$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively. . . 81

3.23 Longitudinal atlas validation (sub-group 4: 70-75 years, $n = 63$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively. . . 82

3.24 Longitudinal atlas validation (sub-group 5: 75-80 years, $n = 36$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively. . . 82

4.1 Ultra-structure arrangements of cortical and trabecular bone in the proximal femur. Cortical bone (the outer highly mineralised shell) is seen at the shaft and the inferior neck, whereas trabecular bone with the sponge-like structure resides inside the cortical shell. (Adapted from [158]). 90

4.2 Yearly percentage BMD change in the proximal femur. Pixel BMD change rates are normalised to the median BMD at 25 years. BMD was mostly preserved in the cortical bone till 60^{th} year following a consistent decrease with approximate annual rate of 0.5%. Variation in the arrangement of trabecular architecture looks spatially complex. In the early adulthood, BMD reduction in trabecular bone was faster at the femoral neck. In the middle adulthood, BMD reduction accelerated throughout the femur. In the advanced adulthood, BMD reduction was more dominant at the femoral shaft and greater trochanter. . . . 91

4.3 Proximal femoral bone density profiles at two cross-sections. (a) BMD profiles at the femoral shaft demonstrated an M-shape graph with peak BMD at the outer cortex and lower trabecular BMD in the middle. On each graph, the two peak BMD values in the interior and exterior cortex are marked with asterisks. Peak cortical thickness is defined as the width of the cross-section line minus the distance between the two peaks in the BMD profile. (b) BMD profile at the femoral neck demonstrated only one distinct local peak BMD at the inferior cortex where cortical bone is present (see Fig. 4.1). Ageing is associated with a decrease in peak BMD. 92

4.4 Average peak cortical thickness variation with ageing. Average peak cortical thickness at the diaphysis was linearly decreased with ageing from 6.65 mm at 20 years to 5.55 mm at 80 years. . 92

4.5 Schematic bone ageing trajectory in a 2-dimensional space. The solid black line represents the median ageing trajectory using BMD at two pixel coordinates. One pixel is selected from the femoral neck, and the other one is selected from the cortex near the lesser trochanter. For a given bone map (green dot), its bone age is estimated by mapping the given bone map to the closest point on the trajectory. Note that the actual bone ageing trajectory lies on an N-dimensional space where N equals the number of pixels in the template. 94

4.6 Median bone ageing trajectory. The median bone ageing trajectory is a 1D graph in the N-dimensional space where $N = 16035$ is the number of pixels in the template. 94

4.7 Intuitive illustration of bone age potential to differentiate between fracture and control subjects with similar neck BMD values. The top row shows the bone map for a woman of aged 75.8 years with femoral neck BMD of 0.5860 g/cm^2 who experienced a hip fracture following the baseline measurement. The bottom row shows the bone map for a non-fracture subject with similar age (75.9 years) and femoral neck BMD (0.5900 g/cm^2). Despite similar age and femoral neck BMD, the spatial texture pattern varies between the two subjects. The associated bone age was 80 and 62 years for the top and bottom rows, respectively. 95

4.8 Relationship between bone age and femoral neck BMD. Bone age is linearly correlated with neck BMD ($r = -0.86$; $p < 0.001$). The blue and red dots represent the fracture-free controls ($n = 4249$) and the hip fracture cases ($n = 178$), respectively. The density of red dots increases at the bottom-right corner, consistent with both a decrease in neck BMD and an increase in bone age. 96

4.9 Relationship between bone age and FRAX. The FRAX score consistently increases with bone ageing with an exponential pattern. The blue and red dots represent the fracture-free controls ($n = 4249$) and the hip fracture cases ($n = 178$), respectively. The density of red dots increases with an increase in bone age and FRAX score. Note that FRAX is reported with BMD as a risk factor. 97

4.10 Bone ageing precision analysis. Twenty-five subjects were scanned two times on the same day with patient repositioning between scans. For each subject, bone age is estimated independently for each of the two collected scans. The SD for precision error was 1.4 years. No significant difference was observed using a paired t-test ($p = 0.54$). 97

4.11 The ROC curve for ability of bone age versus neck BMD to classify between young and old populations. The young cohort ($n = 284$) includes all white women aged 40 years or less selected from the OPUS dataset. The old cohort ($n = 2165$) includes all white women aged 80 years or more selected from the OPUS or the MRC-Hip datasets. The AUC for bone age and neck BMD was 0.94 (95% CI=0.930-0.954) and 0.89 (95% CI=0.874-0.906). 98

4.12 Stratified fracture risk based on the bone age (the first row), neck BMD (the second row), FRAX score (the third Row), and chronological age (the forth row) in the MRC-Hip study. The first column shows the fracture risk at the hip while the second column shows the risk for any fragility fractures occurred at the hip, spine, pelvis, lower limb, or the upper limb. The total number of fractures was 684 out of which 178 occurred at the hip. The number of control cases was 4249. 99

4.13 The ROC curve for prediction of fragility fractures. The AUC for prediction of hip fractures (n=178) was 0.731 (95% CI=0.689-0.761), 0.723 (95% CI=0.690-0.754), 0.660 (95% CI=0.619-0.694), and 0.719 (95% CI=0.682-0.755) for neck BMD, FRAX with BMD, FRAX without BMD, and bone age, respectively. The AUC for prediction of any major osteoporotic fractures (n=684) was 0.632 (95% CI=0.609-0.651), 0.636 (95% CI=0.613-0.656), 0.590 (95% CI=0.569-0.613), and 0.639 (95% CI=0.618-0.661) for neck BMD, FRAX with BMD, FRAX without BMD, and bone age, respectively. The number of fracture-free controls was 4249. 100

4.14 Bone-age normalised BMD map. Panel (a) shows the bone map for a woman of age 77 years who sustained a follow-up hip fracture. Panel (b) shows the median atlas at the estimated bone age 83 years for this subject. Panel (c) shows the normalised BMD map or the quantile map. Pixel quantiles reflect the rank of the given pixel BMD values among the population with a similar bone age. 102

4.15 Localising fracture-specific patterns using bone-age normalised BMD maps. Panel (a) shows the difference in mean quantile maps between the fracture and the fracture-free control groups. Panel (b) shows the corresponding statistical significance map using a two-sample t-test followed by FDR analysis. A local pattern of bone loss was observed in the same orientation as principal tensile curves characterised in the radiography scans [51]. Panel (c) shows the trabecular arcades in the proximal femur deployed for assessing the Singh index. The image is adapted from [163]. 102

4.16 Localising fracture-specific patterns using raw BMD maps. Panel (a) shows the difference in mean BMD maps between the fracture and the fracture-free control groups. Panel (b) shows the corresponding statistical significance map using a two-sample t-test followed by FDR analysis. Raw BMD unlike the quantiles cannot localise fracture-specific patterns (cf. Fig. 4.15). 102

4.17 The ability of f-score to predict fractures independent of bone age or neck BMD. (a) Relationship between f-score and bone age. (b) Relationship between f-score and neck BMD. The blue and red dots represent the fracture-free controls and the hip fracture cases, respectively. (c) The ROC curve for classification between hip fractures ($n = 178$) and controls ($n = 4249$). The AUC was 0.731 (95% CI=0.689-0.761) and 0.736 (95% CI=0.694-0.769) for neck BMD and f-score, respectively. The low correlation between f-score and either bone age or neck BMD and similar power of f-score versus neck BMD for hip fracture prediction suggest the potential to enhance fracture prediction by combining f-score and bone age (see Fig. 4.18). . . 104

4.18 Ability of combined f-score and bone age versus neck BMD to predict hip fractures. Logistic regression analysis was deployed to find appropriate weights to combine f-score and bone age. (a) ROC curve for classification between hip fractures ($n = 178$) and controls ($n = 4249$). (b) Box-plot for the estimated AUC using 1000 different iterations. A five-fold cross-validation technique was deployed; each time one fold was left out for testing and the combination weights were learned on the remaining data. I repeated this procedure 1000 times and the distribution of the average AUC values on all 5 folds are reported. 104

4.19 Spatial BMD variation with Age and BMI. The median bone maps are visualised for 20, 35, 50, 65, 80, and 95 years of age and different BMI values of 15, 20, 25, 30, 35, 40, and 45 kg/m^2. The atlas is shown for the Hologic system at the left hip. 106

4.20 Various types of fracture patterns observed in the proximal femur. DXA RFA would allow correlation of local BMD patterns with actual fracture patterns in the femur to enhance hip fracture prediction. The image is adapted from [165]. 109

5.1 Image representation and visual word segmentation. The input image is divided into a set of overlapping image patches quantised to the best-matched visual words from a learned dictionary. Each visual word is a representative patch for a cluster of similar image patches. Visual words only account for frequent texture patterns; the abnormal key-shape object is eliminated in the reconstructed image. 113
5.2 Unsupervised non-distortion-specific image quality framework. The proposed method works on image patches. In the first step, a dictionary of visual words is learned by grouping similar image patches together using the fixed-width clustering algorithm [175]. Each visual word is simply the centroid of each cluster. In the second step, each image is divided into a set of patches paired with their best-matched visual words from the dictionary. A set of dissimilarity scores is then computed per each pair. Finally, a probability distribution is established for each dissimilarity metric over the full dataset. Artefacts are detected as outliers of these distributions. 114
5.3 Various types of femoral DXA artefacts. DXA artefacts can be broadly classified into three categories: errors due to poor calibration of the instrument prior to scan collection, imaging artefacts during scan acquisition, and errors due to poor analysis following the scan collection. Imaging Artefacts can be further divided into six categories. 119
5.4 Examples of various imaging artefacts in DXA. The red contour overlaid the image shows the location of each artefact. 121
5.5 Examples of successful and unsuccessful DXA artefact detection using the proposed algorithm. The red square, if present in the image, shows the location of detected artefacts. 123

6.1 Basic types of fracture patterns in the proximal femur. A single pattern or a combination of these patterns may be observed in practice. DXA RFA would allow correlation of local BMD patterns with actual fracture patterns in the femur to enhance hip fracture prediction. The image is adapted from [165]. 132

List of Tables

1.1 Mean (SD) for five frequently reported femoral geometrical indices, and significance of the difference measured in women with a hip fracture (F) and age-matched control (C) group. 11

2.1 Possible outcomes when testing N hypotheses. 21
2.2 Characteristics of the patient populations participating in the DXA RFA analyses. 35
2.3 Area size of regions with significant pixel BMD change ($q \leq 0.05$) with corresponding mean BMD change for three cemented prosthesis designs over 24 months. 36
2.4 Area size of regions with significant pixel BMD change ($q \leq 0.05$) with corresponding mean BMD change for a conventional cementless femoral prosthesis (Bi-Metric) versus a hip resurfacing femoral prosthesis (ASR) over 24 Months. 39

3.1 Coefficient of variation (%) at four common conventional ROIs. Top row shows scans measured using DXA RFA with pixels were aggregated to reproduce the conventional ROIs. Lower rows show comparison with published precision data from other investigators. 68
3.2 Point localisation error at five prominent anatomic locations in millimetre (mm). 76
3.3 The comparison between the proposed quantile matching regression versus the Deming regression using synthetic samples at different noise levels ($\sigma_2 = \sigma_1$). 77

4.1 Yearly percentage BMD reduction in women at four femoral ROIs from selected publications. 88

List of Abbreviations

ASR	Articular Surface Replacement
AUC	Area Under the Curve
BMD	Bone Mineral Density
BMI	Body Mass Index
BOWV	Bag of Visual Words
CDF	Cumulative Distribution Function
CI	Confidence Interval
CLM	Constrained Linear Model
CSMI	Cross-Sectional Moment of Inertia
CV	Coefficient of Variation
DRR	Digitally Reconstructed Radiograph
DSC	Dice Similarity Coefficient
DXA	Dual-energy X-ray Absorptiometry
EC	Euler Characteristic
ESP	European Spine Phantom
FDR	False Discovery Rate
FE	Finite Element
FWE	Family-Wise Error
FWHM	Full Width at Half Maximum
GUI	Graphical User Interface
HR-pQCT	High-Resolution peripheral Quantitative Computed Tomography
HSA	Hip Strength Analysis
IDSC	International DXA Standardisation Committee
IQA	Image Quality Assessment
MRI	Magnetic Resonance Imaging
OPUS	Osteoporosis and Ultrasound Study
PAR	Population Attributable Risk
PCA	Principal Component Analysis
PCE	Per-Comparison Error
PDM	Point Distribution Function
PRDS	Positive Regression Dependency on Subset
QCT	Quantitative Computed Tomography
QUS	Quantitative Ultra Sound
RFA	Region Free Analysis
RFT	Random Field Theory
ROC	Receiver Operator Characteristic
ROI	Region of Interest
SAM	Statistical Appearance Model
SD	Standard Deviation
SNR	Signal-to-Noise Ratio
SOF	Study of Osteoporotic Fractures
SPM	Statistical Parametric Map
SSM	Statistical Shape Model
SSR	Sum of Squared Residuals

TBS	Trabecular Bone Score
THA	Total Hip Arthroplasty
TPS	Thin Plate Spline
VGAM	Vector Generalised Additive Model
VGLM	Vector Generalised Linear Model
VBA	Voxel-Based Analysis
WHO	World Health Organisation

Chapter 1

Introduction

1.1 Importance of Osteoporosis in Clinical Practice

The human skeletal is composed of two types of bone structure, termed cortical (80%) and trabecular (20%) [1]. Trabecular bone, also known as cancellous bone, has a spongy architecture forming a scaffolding network. Cortical bone forms the hard exterior cortex giving the smooth white appearance to the bones. Bone is a living tissue composed of two main components: an organic collagen matrix and an inorganic phase composed of mostly calcium phosphate crystallised as an apatite [2]. To maintain homoeostasis, bone undergoes a lifetime process of resorption and formation called *bone remodelling* [1]. Following bone resorption by osteoclasts, osteoblasts first synthesise the collagen matrix laid down at specific sites. Next, the new matrix starts a *primary mineralisation step* after about 5-10 days. After completion of this step, a *secondary mineralisation step* begins. This step is slower than the primary step and gradually increases both the amount and the size of mineralised crystals in bone. Imbalanced regulation of bone remodelling can lead to erosion of the arches and trusses in trabecular bone making it weaker and more prone to fracture. Osteoporosis, which literally means porous bone, is a bone disease characterised by low bone mass and micro-architectural deterioration leading to an increased risk of fractures. [3, 4].

Osteoporosis is an age-associated disease caused by gradual bone loss, and is asymptomatic unless a fracture occurs. In fact, the consequent fragility fractures underlie the clinical significance of osteoporosis in public health [5]. The lifetime risk of suffering from a fracture in the forearm, hip, or vertebra is

40% in women and 15% in men from the age of 50 onwards [6]. Osteoporotic fractures are associated with increased mortality, morbidity, and reduced quality of life [7]. These fractures can bring a heavy burden on community and healthcare systems. In Europe, the total number of osteoporotic fractures in 2000 was estimated at 3.79 million (1.05 million in men and 2.74 million in women) with estimated total direct costs at £21.2 billion which is expected to be doubled by 2050 (£51.1 billion) [8]. However, effective treatment options available for osteoporosis patients could prevent up to a quarter of all osteoporotic-related fractures [8]. With these treatments that favourably alter the natural osteoporosis progression, development of accurate quantifiable techniques for fracture prediction and diagnosis of osteoporosis is a crucial step so that treatments can be targeted efficiently to high-risk individuals [9].

An operational definition for osteoporosis is based on bone mineral density (BMD) measured at the proximal femoral neck [9]. It is well established that bone mass is inversely related to the fracture risk [10, 11, 12]. The World Health Organisation (WHO) definition of osteoporosis is a BMD that lies 2.5 standard deviations or more below the mean for young healthy women [13, 4]. With this definition, however, a majority of individuals who will experience fractures are not identified [14, 15, 16]. Almost half of all fragility fractures occur in subjects with a normal BMD at the femoral neck [15, 16]. A meta-analysis of eleven separate studies showed that methods based on BMD analysis can be used for predicting fracture trends in large populations but cannot assess individual fracture risk accurately [10]. To address this limitation, researchers have suggested deploying easily obtained clinical risk factors other than BMD to enhance fracture prediction (Fig. 1.1). Currently, FRAX is the leading toolbox for fracture risk assessment based on a combination of clinical risk factors and BMD at the femoral neck [17]. FRAX takes into account several clinical risk factors such as age, low body mass index (BMI), previous fragility fractures, parental history of hip fracture, long-term use of oral glucocorticoids, smoking, rheumatoid arthritis, secondary osteoporosis, and alcohol consumption [17]. The FRAX algorithm can be used with or without BMD as a risk factor, and so prior estimation of risk can save the added expense of bone densitometry when a DXA scan is not required.

The introduction of the FRAX algorithm has facilitated the assessment of fracture risk but it does not capture other skeletal determinants of bone strength beyond the femoral neck BMD. Arguing that femoral neck BMD is not an optimal surrogate for bone strength, researchers have tried to provide alternative valid metrics for bone quality assessment. In the following sections, first, an explanation of bone quality is presented (section 1.2) and then the recent advances in bone quality assessment using DXA is reviewed (section 1.3). Several other imaging techniques also exist for quantitative bone quality assessment, including Quantitative Ultra Sound (QUS), Quantitative Computed Tomography (QCT), and Magnetic Resonance Imaging (MRI) [18]. All techniques have advantages and disadvantages (see [18, 19] for a review) but DXA is the most widely available method in clinical practice and will remain so for the foreseeable future. This thesis mainly focuses on the application of DXA and other imaging modalities are not further discussed here. Section 1.5 reviews the current limitations in the literature and presents the main objectives of this study. Finally, the outline of this thesis is presented in section 1.6.

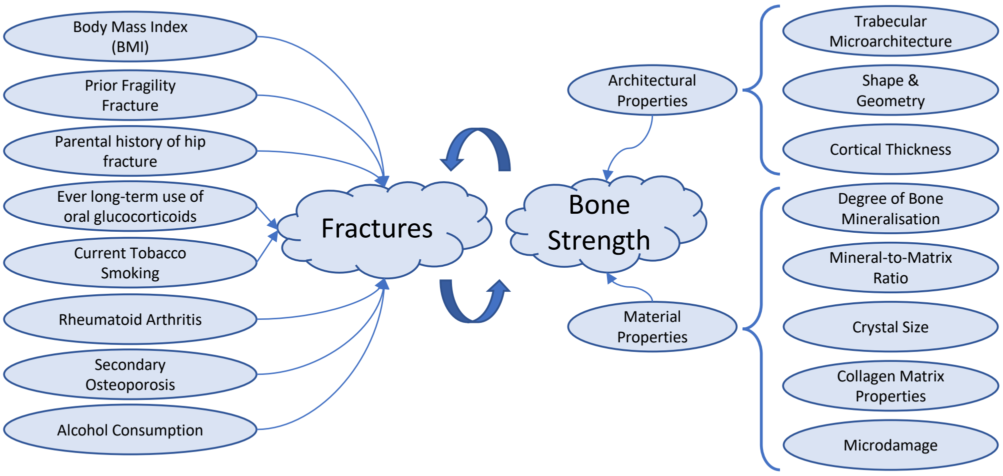

Body Mass Index (BMI)
Prior Fragility Fracture
Parental history of hip fracture
Ever long-term use of oral glucocorticoids
Current Tobacco Smoking
Rheumatoid Arthritis
Secondary Osteoporosis
Alcohol Consumption
Fractures
Bone Strength
Architectural Properties
Material Properties
Trabecular Microarchitecture
Shape & Geometry
Cortical Thickness
Degree of Bone Mineralisation
Mineral-to-Matrix Ratio
Crystal Size
Collagen Matrix Properties
Microdamage

Figure 1.1: Bone strength determinants and fragility fracture risk factors. Several non-BMD clinical risk factors are identified for fragility fractures [17]. Bone strength has an important influence on the chance of developing a new fracture, but could be targeted as the important outcome by itself. Bone strength is the result of an inextricable relationship between both architectural and material properties [20, 21].

1.2 Bone Quality

When bone is loaded, it deforms in response to the force. The intensity of the force divided by the cross-sectional area where it acts is called *stress*. The proportional change in length due to the applied force is called *strain*. The stress-strain curve is a useful tool to assess the mechanical properties of bone (Fig. 1.2). Four quantities are used for this purpose: strength, toughness, resilience, and stiffness. Strength can be defined as the maximum stress that bone can sustain before it breaks when loaded slowly. Toughness is defined as the amount of energy bone can absorb before it breaks measured as the area under the curve (the blue dashed area in Fig. 1.2). This is specifically important during falls as bone should absorb a huge amount of energy in a short time. Note that strength is different from toughness; for example, glass is quite strong but brittle. Resilience is defined as the amount of elastic energy before the yield point (the orange shaded area in Fig. 1.2). Finally, stiffness is the ratio of stress to strain when the stress-strain curve is still linear. This is also known as *Young's modulus*.

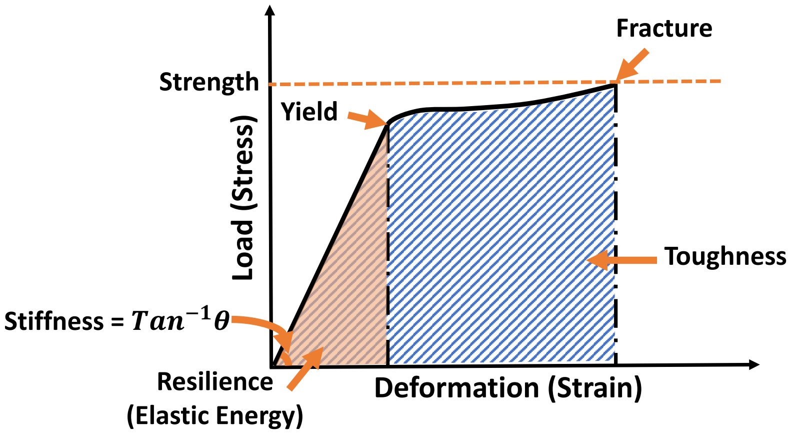


Figure 1.2: A schematic stress-strain curve of bone in tension. The initial part of the curve is almost linear where the slope of the line defines the Young's modulus (stiffness). The area under the curve equals the amount of energy used to deform the bone. The blue dashed area defines toughness (the amount of energy absorbed before fracture) and the orange shaded area defines resilience (the amount of energy absorbed before the yield point). The height of the curve at the fracture point defines strength (the maximum stress bone sustains before fracture).

Bone strength can be measured destructively in a laboratory by gradually increasing the loading until the bone breaks [22]. For *in vivo* measurements of bone deformation in response to various loading conditions, two types of engineering models are broadly deployed: finite element (FE) and beam models [23]. The latter assumes the femur as a supporting beam and stresses are computed at three different cross-sections across the neck, intertrochanter, and the shaft [22]. A software called hip strength analysis (HSA) was developed by Beck and colleagues [22] to measure the bending strength reported as the cross-sectional moment of inertia (CSMI). HSA combines the BMD measurements and the hip geometry to predict bone strength but its precision is sensitive to femur positioning [22]. Scaling for body size is also critically important in this technique [22]. FE modelling provides a means for more sophisticated simulation of the load behaviour in the femur by breaking it up into small elements [24]. For an accurate model it is necessary to have information on the specific loading conditions, bone geometry, and the material properties distributed in the femur. Despite the frequent application of these techniques in research, their usage in clinical practice is not established [18].

Fragility fractures are indeed the result of decreased bone strength, but interpreting *sufficiently strong* may be different, depending on the specific event leading up to the fracture [21]. For example, a stress applied repeatedly may break the bone while a similar impact applied only once may not break the bone. Each of the mechanical properties described above could potentially have a different impact on the overall bone competence and the relative importance of each of these properties needs to be better understood [25]. Therefore, the ability to measure a biomechanical parameter like strength may not provide a comprehensive picture of bone competence. Here, I use the term *bone quality* to emphasise the difference between bone strength as the maximum sustainable stress and the soundness of bone in general. However, whenever clear from the context, bone quality and bone strength are used interchangeably in this thesis.

Bone quality is known to be the result of an inextricable relationship between the bone architecture including bone geometry, trabecular microarchitecture, and cortical thickness and its material properties including the degree of mineralisation of bone, crystal size, the mineral-to-matrix ratio, and microdamage (Fig. 1.1) [18, 20, 21]. The ability to derive a single ideal measure for bone quality is still unmet [21]; however, any tools that can reflect one or more determinants of bone quality independent of femoral neck BMD can potentially have an added value in the prediction of fractures. In the next section, I review the recent advances in bone quality assessment using DXA.

1.3 DXA in Clinical Practice

DXA is the WHO gold standard tool used to measure areal BMD, and is defined as the amount of bone mass per unit area (g/cm^2). Femoral neck BMD is a simple metric that has been shown to be predictive of approximately 60-70% of the variance in the bone mechanical properties and has been used widely as a surrogate for bone strength in clinical practice [18, 26, 22]. Several studies have shown a robust relationship between BMD and fracture risk [10, 11, 12]. However, given that DXA provides a 2D projection map of the BMD distribution in the whole femur (see section 1.3.1), it can be postulated that DXA scans contain much more information beyond which femoral neck BMD could represent. A typical DXA scan can reflect the femoral shape and geometry [27] as well as trabecular microarchitectural [28]. This extra information, if manipulated effectively, can potentially improve our understanding of bone quality in the femur.

Recent advances in image processing and statistical analysis techniques have opened new opportunities for the development of novel frameworks for DXA analysis providing a more comprehensive picture of bone quality in the femur [29, 30]. The central question of data mining in DXA has been approached from different angles: section 1.3.2 reviews attempts to restore 3D spatial distribution of BMD values from 2D DXA scans; section 1.3.3 reviews the literature on extracting new geometrical indices for enhanced fracture risk assessment; and finally section 1.3.4 reviews techniques to represent spatial texture patterns of BMD distribution extracted from DXA scans.

1.3.1 Principles of DXA

The fundamental principle underlying DXA systems is based on the amount of attenuation for two X-rays with different energy levels traversing through the body. When an X-ray beam passes through a tissue with mass attenuation coefficient μ (cm^2/g) and areal density $\mathcal{M}$ (g/cm^2), the incident radiation intensity I_0 is attenuated due to photoelectric and Compton effects [31, 32]. The pattern of attenuation for a homogeneous material can be described according to the formula:

$$I = I_0\, exp\left[-\mu\mathcal{M}\right], \tag{1.1}$$

where I is the transmitted intensity. In practice, the tissue that the X-ray beams are transmitted through it is not homogeneous. However, for BMD computation, it is sufficient to assume the traversing medium as a two-compartment model of bone mineral and soft tissue. Then, Eq. 1.1 can be rewritten as:

$$I = I_0\, exp\left[-(\mu_s\mathcal{M}_s + \mu_b\mathcal{M}_b)\right], \tag{1.2}$$

where the subscripts s and b stand for soft tissue and bone mineral, respectively. Taking the natural logarithm of Eq. 1.2 gives:

$$-\ln(\frac{I}{I_0}) = \mu_s\mathcal{M}_s + \mu_b\mathcal{M}_b. \tag{1.3}$$

To differentiate between soft tissue and bone mineral, two X-rays one with high-energy and the other with low-energy are required. Tow different ways are deployed for generating the required spectrum with an X-ray tube: the energy switching technique or the rare-earth K-edge filtering approach. In Hologic

DXA scanners (Hologic In, Waltham, MA), the X-ray tube potential is rapidly switching between 100 and 140 kVp [31]. Alternatively, Lunar DXA scanners (GE Healthcare, Madison, WI) use an X-ray generator with a highly stable constant potential (100 kVp). The X-ray beam is then passed through a K-edge filter that divides the spectrum into the high- and low- energy components.

The high- and low- energy X-rays provide two independent equations

$$J = \mu_s \mathcal{M}_s + \mu_b \mathcal{M}_b, \tag{1.4}$$

$$J' = \mu'_s \mathcal{M}_s + \mu'_b \mathcal{M}_b, \tag{1.5}$$

where the primed variables are associated with the low-energy radiation. For ease of notation, $-\ln(\frac{I}{I_0})$ in Eq. 1.3 is replaced with J in Eqs. 1.4 and 1.5. Given the value of the four attenuation coefficients, areal densities for both bone mineral and soft tissue can be computed as:

$$\mathcal{M}_b = \frac{J' - (\mu'_s/\mu_s)J}{\mu'_b - (\mu'_s/\mu_s)\mu_b} \tag{1.6}$$

$$\mathcal{M}_s = \frac{J' - (\mu'_b/\mu_b)J}{\mu'_s - (\mu'_b/\mu_b)\mu_s} \tag{1.7}$$

Eq. 1.6 can be used for BMD computation only if the attenuation coefficients for soft tissue and bone are known. In practice, attenuation coefficients are not known *a priori* given the variation in the composition of soft tissues and bones. For an accurate assessment of bone density, Hologic scanners pass the generated high- and low- energy X-ray beams through a proprietary automatic internal reference system [33, 31]. In this system, patient's bone is continuously compared against a known value contained in the internal reference standard. Let $\mathcal{M}_{cs}$ and $\mathcal{M}_{cb}$ denote the areal densities of the known calibration filters for soft tissue and bone, respectively. Therefore, Hologic scanners provide six transmission measurements through the air (no filter), bone, and soft tissue filters at both high and low energies. Let J, J_b, and J_s denote the received logarithmic transmission factors for air, bone, and soft tissue, respectively. The increments in attenuation when the calibration filters are interposed can be computed for bone and soft tissue.

$$\Delta J_{cb} = J_b - J \quad \text{and} \quad \Delta J'_{cb} = J'_b - J', \tag{1.8}$$

$$\Delta J_{cs} = J_s - J \quad \text{and} \quad \Delta J'_{cs} = J'_s - J'. \tag{1.9}$$

Given these values, the effective values of attenuation coefficients can be computed as follows:

$$\mu_b = \Delta J_{cb}/\mathcal{M}_{cb} \quad \text{and} \quad \mu'_b = \Delta J'_{cb}/\mathcal{M}_{cb}, \tag{1.10}$$

$$\mu_s = \Delta J_{cs}/\mathcal{M}_{cs} \quad \text{and} \quad \mu'_s = \Delta J'_{cs}/\mathcal{M}_{cs}. \tag{1.11}$$

Substituting the estimated attenuation coefficients in Eq. 1.6, the areal density for bone mineral can be computed.

The bone density is computed per each pixel of the area being scanned. This results in a bone map consisting of many thousands of pixel BMD values. In conventional analysis, these pixel values are averaged in *a priori* identified regions of interest (ROIs) including the femoral neck. This data pooling has been established as the standard protocol for DXA in clinical practice. However, pixel BMD information was deployed for research purposes in several previous studies [34, 35, 36].

1.3.2 Volumetric BMD with DXA

DXA is a 2D modality that can provide projected BMD measurements on a plane perpendicular to the X-ray beams. This can raise three limitations: first, slight variation in the angle of the scan could potentially lead to a significant change in the measured BMD values [37]. For example, ten degrees of internal rotation from the customary position decreased the average BMD by 0.009, 0.005, and 0.006 g/cm^2 in the femoral neck, trochanter, and Ward's area (p-value: <0.001, 0.008, and <0.001), respectively [37]. Second, areal BMD cannot account for the variation in bone size; given the same volumetric BMD, larger bones tend to have a higher areal BMD value [18]. Third, 3D spatial BMD information is lost in DXA.

To address these limitations, several techniques have been proposed in the literature for 3D reconstruction of shape and volumetric BMD information using one or more DXA scans collected at different angels [38, 39]. These techniques, albeit different in implementation details, conceptually follow three main steps: first, construction of a 3D statistical atlas from a subset of QCT scans. Second, construction of a digitally reconstructed radiograph (DRR) by projecting an instance from the atlas to the plane where DXA scan was collected. Various parameters including shape modes, scale, rotation, and translation are modified iteratively so that the constructed DRR resembles the actual DXA scan. Third, calibration of the volumetric BMD values using areal BMD measurements from DXA scans collected at the anteroposterior position. For each pixel areal BMD, all voxels in the 3D reconstructed image contributing to that projected pixel are linearly scaled.

Accurate 3D reconstructed scans can potentially improve the diagnosis of osteoporosis and fracture risk estimation. For example, Whitmarsh et al. [40] reported an enhancement in the area under the receiver operator characteristic (ROC) curve for discrimination between a group of 80 patients with a contra-lateral femur fracture and a control group of the same size. A two-fold cross-validation technique was applied in this study. The average area under the ROC curve was increased from 0.60 using only the femoral neck BMD as the feature to 0.68 when adding the first mode of variation and the scaling parameter as new features. Despite this marginal improvement, 2D-3D techniques have other limitations as well. The 3D atlases are often trained based on a small set with a few hundred QCT images. Therefore, this atlas may not account for the full population variation including the effects of ageing. Further research is required to demonstrate the added value of 3D spatial BMD information in comparison to 2D planar BMD measurements [41]. For example, Goodyear et al. [42] also reported similar enhancement in discrimination between fracture and non-fracture groups using statistical shape and appearance models directly applied to DXA scans. The second limitation with bone size does not seem to be addressed with 2D-3D techniques as the final calibration is still based on planar BMD values from DXA.

1.3.3 Shape and Geometry

The geometric structure of the femur is known to be an important risk factor for hip fracture. However, characterising the shape of the femur and its association with fracture is a non-trivial task. Several hand-crafted geometrical indices have been suggested in the literature. For example, Michelotti el at. [43] investigated 15 different indices to identify their contribution to hip fracture risk using logistic regression analysis. Table 1.1 shows the mean (SD) for five different indices frequently reported in the literature: the hip axis length, the femoral neck axis length, the neck-shaft angle, the femoral neck diameter, and the diaphysis diameter. Fig. 1.3 demonstrates the definition for each term.

The literature does not present a clear consensus (Table 1.1); Faulkner et al. [27] suggested that the hip axis length could be predictive of hip fractures independent of age and femoral neck BMD where one standard deviation increase in the hip axis length almost doubled the hip fracture risk (odds ratio = 1.8; 95% confidence interval = 1.3-2.5). While several studies confirmed sim-

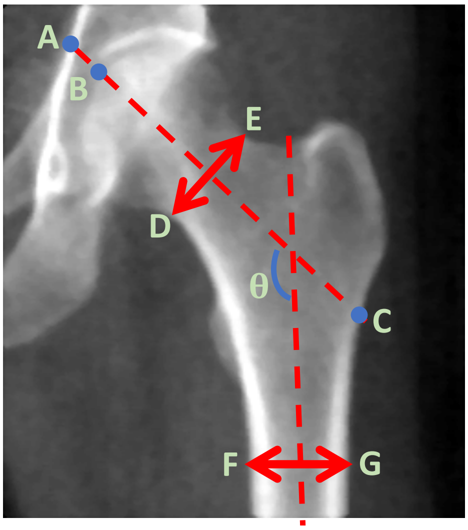


Figure 1.3: Geometrical hip indices. (A-C) Hip axis length is the linear distance from the inner acetabulum surface to the lateral margin of the femoral shaft below the greater trochanter drawn through a midpoint in the femoral neck parallel to the cortices of the femoral neck. (B-C) Femoral neck axis length is the linear distance from the apex of the femoral head to the base of the greater trochanter. This index is similar to the hip axis length but does not include the acetabular portion. (θ) Neck-shaft angle is the angle between the shaft axis and the femoral neck axis. (E-F) Femoral neck diameter is measured as the length of the line perpendicular to the femoral neck axis passing through the centre of the femoral neck. (F-G) Diaphysis diameter is the width of the femoral shaft below the lesser trochanter.

ilar findings [44, 45, 46], others do not show a significant difference between fracture and control groups for the hip axis length or the femoral neck axis length [47, 43, 48, 49]. Despite this discrepancy in results, all studies reported a higher length for the fracture group in comparison to the control group. The neck-shaft angle was the second most frequently measured geometrical index. The neck-shaft angle was consistently larger for the fracture group. Again, the significance of this larger neck-shaft angle was confirmed in some studies [45, 47, 46] and rejected in others [27, 44, 43].

Evaluation of the published data is complicated for three reasons. First, the measurements are not standardised. For example, the hip axis length varies from 3.71 cm in [45] to 12.96 cm in [44]. Second, the effect of patient positioning on the geometrical indices has not been addressed. For example, the reported 3.4% increase in the hip axis length by Faulkner et al. [27] could be created by about 5° of hip abduction [43]. Third, the measurement error using manual or automatic techniques should be taken into account for each study. Given the discrepancies in the published studies, Michelotti et al. [43] suggested that the common belief that the femoral neck length is an independent risk factor for the hip fractures remains unproven.

Table 1.1: Mean (SD) for five frequently reported femoral geometrical indices, and significance of the difference measured in women with a hip fracture (F) and age-matched control (C) group.

	nC:nF	Hip Axis Length (cm)			Neck Axis Length (cm)			Neck-Shaft Angel (degree)			Neck Width (cm)			Shaft Width (cm)		
		C	F	p-value	C	F	p-value	C	F	p-value	C	F	p-value	control	fracture	p-value
Faulkner et al. [27]	134:64	6.70 (0.38)	6.93 0.40	0.0001	-	-	-	126 (4.7)	127 (5.5)	0.42	1.98 (0.16)	2.0 (0.16)	0.41	-	-	-
Peacock et al. [44]	43:22	12.96 (0.68)	13.31 (0.54)	0.04	11.38 (0.55)	11.63 (0.41)	0.082	122.91 (2.95)	123.82 (3.76)	0.211	3.23 (0.36)	3.25 (0.31)	0.74	-	-	-
Boonen et al. [45]	75:30	3.71 (0.27)	3.91 (0.17)	0.0003	-	-	-	122.4 (2.4)	126.7 (3.3)	< 0.0001	1.04 (0.09)	1.14 (0.13)	< 0.0001	-	-	-
Michelotti et al. [43]	41:43	-	-	-	11.48 (0.79)	-	-	127.1 (6.3)	128.0 (6.5)	0.52	3.65 (0.28)	3.75 (0.29)	0.11	3.65 (0.30)	3.58 (0.30)	0.29
Center et al. [49]	100:23	-	-	-	8.97 (0.54)	9.15 (0.54)	0.15	-	-	-	-	-	-	-	-	-
Alonso et al. [47]	310:295	6.39 (0.37)	6.41 (0.40)	0.52	-	-	-	124.6 (4.2)	129.6 (5.3)	< 0.0001	3.27 (0.26)	3.48 (0.37)	< 0.0001	-	-	-
Testi et al. [46]	166:106	10.7 (0.6)	10.9 (0.7)	0.01	-	-	-	122.5 (5.0)	125.3 (6.2)	< 0.0001	3.2 (0.2)	3.3 (0.2)	< 0.0001	3.3 (0.2)	3.4 (0.3)	0.001
Gregory et al. [48]	24:26	-	-	-	10.97 (0.76)	11.02 (0.53)	0.76	-	-	-	3.78 (0.28)	3.82 (0.32)	0.67	-	-	-

With developments in statistical shape models, rather than manually crafting discriminant geometrical features, Gregory et al. [48] deployed active shape models to quantify the morphological variation in the proximal femur. Although the sample size was small ($n = 26$ fracture cases and 24 control subjects), the results look promising. While active shape models have the potential to correct for image magnification and poor patient positioning, these issues have not been addressed in [48]. Further research is required to investigate this technique in a larger population.

1.3.4 Trabecular Bone Microarchitecture

Trabecular bone microarchitecture is a key determinant of bone strength [1]. DXA cannot measure bone microarchitecture directly but has the potential to provide reliable information on the overall status of bone microarchitecture in patients [50]. Each DXA scan typically yields over 10,000 pixel BMD measurements depending on the scan site and resolution. This information, if quantified properly, can reflect on both material and architectural properties concerning the bone strength [2, 28]. However, BMD values extracted from DXA scans are manipulated inefficiently by pooling pixels into *a priori* identified ROIs.

Since the introduction of DXA in the late 1960s, little effort has been made to quantify BMD deficit patterns using DXA scans. The earliest attempt to deploy BMD information to quantify spatial texture patterns was likely in 1996 by Berry et al. [34]. Motivated by the Singh index [51], Berry et al. [34] proposed a semi-automatic grading system to classify observed spatial patterns into 5 groups. In another study, Boehm et al. [52] introduced a new index called MF2D based on the Minkowski functions as follows: the bone map is binarised using various threshold values. For each supra-threshold map, the area, perimeter, and Euler characteristic (EC) number are computed. The EC number is defined as the number of connected components in the supra-threshold map minus the number of holes (see section 2.2.2.(C)). This procedure gives three different functionals representing the texture pattern in bone. The final MF2D score is proportional to the integral sum of these profiles such that the area under the ROC curve is maximised for classification between fracture and control groups.

More recently, trabecular bone score (TBS) was developed to provide insight into the overall status of trabecular microarchitecture [28]. TBS is a manually-handcrafted feature of bone texture representing the variation in the bone density distribution in the femur. TBS can be explained using the *variogram* [28]. For a given Bone map $M(\boldsymbol{x})$, the variogram $V(k)$ is defined as the half of the expected squared differences of BMD between any two points with the lag distance k.

$$V(k) = \frac{1}{2}E[(M(\boldsymbol{x}) - M(\boldsymbol{x} + k\boldsymbol{u}_\theta))^2], \tag{1.12}$$

where $\boldsymbol{u}_\theta$ is the unit vector in the θ direction. The variogram can be computed experimentally by averaging over a large number of random initialisation for the location $\boldsymbol{x}$ and direction θ. Given the experimental variogram with a log-log representation, TBS is then defined as the slope of the variogram at the origin. The conceptual interpretation for TBS is as follows: the greater the TBS, the greater the degree of variation between adjacent pixels in the bone map, and so the more dense the trabecular architecture in the 3D volume. Therefore, an elevated TBS would be associated with a better bone quality resistant to fractures whereas a low TBS would be associated with a more fragile bone.

The importance of TBS in clinical practice has been reviewed previously [53, 50]. TBS has been shown to be correlated with microarchitectural parameters including trabecular number ($r = -0.84$), trabecular spacing ($r = 0.73$), and connectivity ($r = -0.85$) but not with the trabecular thickness ($r = 0.23$) [28]. Several studies suggest that TBS can enhance fracture risk estimation independent of BMD [54, 55, 53].

TBS, despite its prospective advantages, has also a number of limitations. First, the physical meaning of TBS is not clear as grey-level values rather than actual pixel BMD measurements are deployed in the computation algorithm. Doge et al. [35] addressed this issue by using actual BMD measurements. Dong et al. [35], arguing that TBS does not capture the global trend of the variogram, suggested to fit an exponential function to the variogram:

$$V(k) = c_0 + c[1 - \exp(-k/L)], \tag{1.13}$$

where c, sill variance, c_0, nugget variance, and L, correlation length, are the model parameters. Dong et al. [35] showed that using the three parameters together with the BMD can increase the area under the receiver-operating characteristic (ROC) curve for classification between fracture and control groups from 0.625 (when only neck BMD is deployed) to 0.748 (p-value = 0.001).

TBS or other similar scores presented in the literature provide a means of global texture characterisation in bone with limited ability to analyse localised BMD deficits. To address this limitation, Morris et al. [36] presented a framework called *region free analysis* (RFA) to allow localised statistical inference in the proximal femur.

1.3.5 DXA Region Free Analysis

DXA RFA is an innovative toolkit developed for periprosthetic BMD variation analysis in a longitudinal study (Fig.1.4) [36]. To render pixel-wise bone maps, a Matlab script was developed by the authors based on the proprietary algorithm Apex v3.2 (Hologic Inc, Waltham, MA) as follows: the raw scan files (with .p and .r extensions) were read to restore 6 image maps associated with the X-ray transmitted intensities through the air (no filter), bone, and soft tissue density measurements at both high and low energies. The maps were used to semi-automatically segment the prosthesis, bone, and soft tissue using thresholding and morphological operations. Finally, the BMD map was computed as discussed in section 1.3.1. Given the segmented bone maps, sixty-three controlling landmark points were selected automatically on the bone contour. Next, a reference template was generated as the mean of selected control points after being corrected for translation, rotation, and scaling using Generalised Procrustes analysis [56, 57]. Finally, each scan was warped to the mean shape using a deformable thin plate spline (TPS) registration technique [58]. This image alignment enforces pixel correspondence between scans.

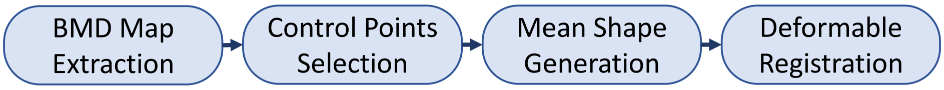


Figure 1.4: DXA region free analysis (DXA RFA) pipeline [36]. Bone maps are restored semi-automatically from the raw files (with .p and .r extensions) using a Matlab script. Sixty three control points are automatically selected on the bone contour. A mean shape is generated using generalised Procrustes analysis to serve as the reference template. Each bone map is then individually warped into this template to remove the morphological variation between scans. This image warping establishes a correspondence between pixel BMD values across the population.

Morris et al. [36] applied DXA RFA for longitudinal analysis of periprosthetic BMD changes in the femur after total hip arthroplasty (THA). In conventional DXA analysis, seven Gruen zones ROIs are usually selected (Fig. 1.5) [59]. Use of predefined ROIs imposes a potential bias on the BMD analysis and may lead to a masking of the true BMD changes [60]. Moreover, due to the pooling of pixel values, global analysis of BMD maps within the whole imaging site is not possible. DXA RFA allows global pixel-level inferences by removing the morphological variation between scans. A two-tailed paired t-test was deployed in [36] to test the significance of BMD change after 12 months at each pixel coordinate. The derived pixel p-values were visualised using a heat-map. Though the heat-map is a useful tool to visualise the p-values within the femur, making a sound global statistical inference requires p-value corrections for a large number of tests. This problem known as multiple comparisons problem is not addressed in [36].

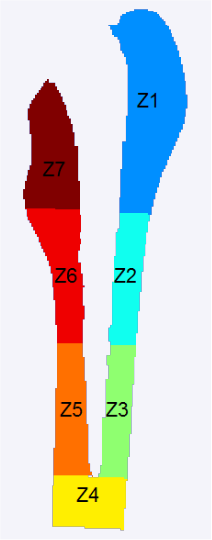


Figure 1.5: Seven Gruen zones commonly used as ROIs for analysis of periprosthetic BMD after total hip arthroplasty.

1.4 Rationale and Motivation

DXA is an inexpensive, widely available clinical tool with excellent precision and extremely low radiation exposure. Conventional DXA analysis by means of averaging pixel BMD information in *a priori* identified ROIs is long-established in clinical practice as the standard tool for osteoporosis management and fracture risk prediction. However, advances in medical image analysis have promoted the extraction of additional information from DXA scans in clinical research. TBS is a recently-developed tool for representing texture patterns in DXA scans, and thereby provide insight on trabecular microarchitecture. Low TBS is consistently associated with an increase in the prevalence of fragility fractures. Moreover, TBS has been shown to be predictive of fracture independent of FRAX scores. These findings suggest that exploring spatial texture variation in BMD maps can potentially have an independent role from neck

BMD and other clinical risk factors to enhance fracture prediction. TBS, however, cannot capture localised BMD variation. More specifically, to generate the variogram, local grey-level variation is averaged over the whole femur, and so TBS provides an insight on the overall texture patterns rather than local BMD deficits.

To capture localised BMD variation in bone, Morris et al. [36] proposed a technique called DXA RFA. DXA RFA allows pixel-level BMD analysis by removing morphological variation between scans. However, it does not address multiple comparisons problem over a large number of pixels. To address this limitation, I integrated the False Discovery Rate (FDR) analysis into DXA RFA. This integration would allow global inference over the whole femur, localising significant BMD variations in the femur (chapter 2).

To date, little research has been done to explore site-specific BMD variation patterns in the native femur. Bone turnover events may lead to areas of both high and low mineral density with complex spatial patterns [20]. Understanding which spatial distribution pattern has a detrimental effect on bone strength is an important consideration in the management of osteoporosis. More specifically, it is of interest to know what changes osteoporosis produces in the trabecular arrangement of the proximal femur at different stages of the disease. Given the close relationship between involutional bone loss and the underlying mechanism of osteoporosis, one can postulate that age-related bone loss patterns, if quantified properly, may help explain why bones get weaker with ageing.

1.5 Objectives of This Thesis

The overall aims of this study are, firstly, to extend the RFA framework into a fully automatic pipeline in the setting of the native femur, and, secondly, to apply this technique to a large cohort of Caucasian women to examine site-specific patterns of involutional bone loss in the femur and its relationship with osteoporosis. To this end, a spatio-temporal atlas of ageing bone in the femur is developed, and some initial clinical observations made exploring its potential future clinical utility.

In summary, the objectives of this thesis are as follows:

1. Integrate False Discovery Rate analysis into DXA RFA to localise pixels with significant BMD variation.

2. Extend the DXA RFA technique into a fully automatic pipeline applicable to large-scale population analysis.
3. Construct a reference spatio-temporal atlas of ageing bone in the femur.
4. Examine site-specific patterns of involutional bone loss in the femur and its relationship with osteoporosis.
5. Explore the feasibility of 6-year bone loss prediction based on the baseline scans in the cohorts from the OPUS study
6. Develop an automatic quality control framework of femoral DXA scans.

1.6 Thesis Structure

The remainder of this thesis is organised as follows:

Chapter 2 demonstrates the integration of FDR analysis into DXA RFA framework to address the multiple comparisons problem over a large number of pixels. I applied the technique to quantitate the magnitude and areal size of periprosthetic BMD changes using scans acquired during two previous randomised clinical trials (2004 to 2009); one comparing three cemented prosthesis design geometries [61], and the other comparing a hip resurfacing versus a conventional cementless prosthesis [62]. DXA RFA resolved subtle differences in magnitude and area of bone remodelling between prosthesis designs not previously identified in conventional DXA analyses.

Chapter 3 presents the development of a reference spatio-temporal atlas of ageing bone in the native femur using a large cohort of North Western European Caucasian women (n=13,338). To this end, the RFA framework, initially developed for periprosthetic BMD analysis, is extended to the native femur. The extended RFA framework is fully automatic applicable to large-scale datasets with thousands of images. To integrate data from different densitometer manufacturer technologies, a novel cross-calibration procedure termed quantile matching regression technique is proposed.

Chapter 4 presents four potential applications of the developed atlas in osteoporosis management. First, the delineation between trabecular and cortical bone architecture in the femur and how these patterns evolve with ageing is presented, for which conventional region-based analysis would be insensitive. Second, a new index called *bone age* is introduced to reflect the overall evolution of spatial BMD variation with ageing. Bone age aims to estimate the actual progression, rather than the chronological age, of each subject along the median bone ageing trajectory with potential to serve as an alternative for progression estimation in osteoporosis. Third, a new index called *f-score* is introduced to reflect the localised fracture-specific patterns in the femur. Integration of bone age (the global metric) and f-score (the local metric) to improve hip fracture prediction is discussed here. Finally, the potential to extend the proposed bone ageing analysis framework for other explanatory variables including body mass index (BMI) is presented.

Chapter 5 presents an emerging challenge for retrospective quality control of large-scale DXA scans. Subjective quality assessment would require a considerable amount of time and expertise, making it unfeasible for use in population imaging studies with thousands of scans. This chapter presents the first automatic quality control framework for use in femoral DXA scans. The proposed framework would be fully unsupervised; it does not require any prior information on the anticipated artefact types or the subjective ground truth labels for each scan. This framework, despite its limitations, is the first attempt toward automatic quality control of DXA images.

Finally, chapter 6 concludes this thesis by summarising the key contributions this work makes to the field. Current limitations are discussed and suggestions are provided for future work on this tool.

Chapter 2

False Discovery Rate Integration with Region Free Analysis

DXA region free analysis (DXA RFA) allows pixel-level quantitation of BMD change within a longitudinal study where a paired t-test could be deployed between the baseline and the follow-up at each pixel coordinate to determine a statistically significant difference in BMD. However, to make a global statistical inference, determining regions with a significant BMD change, the false positive rate should be controlled over the whole femoral area rather than a single pixel. This chapter demonstrates the integration of False Discovery Rate (FDR) analysis with DXA RFA to address the multiple comparisons problem over the group of pixel p-values. Here, I applied the technique to quantitate the magnitude and areal size of periprosthetic BMD changes using scans acquired during two previous randomised clinical trials (2004 to 2009); one comparing three cemented prosthesis design geometries, and the other comparing a hip resurfacing versus a conventional cementless prosthesis. DXA RFA resolved subtle differences in magnitude and area of bone remodelling between prosthesis designs not previously identified in conventional DXA analyses.

The content of this chapter is adapted from the following publication:

Mohsen Farzi, Richard M. Morris, Jeannette Penny, Lang Yang, Jose M. Pozo, Soren Overgaard, Alejandro F. Frangi, and J. Mark Wilkinson, "Quantitating the effect of prosthesis design on femoral remodelling using high-resolution region-free densitometric analysis (DXA-RFA)," *Journal of Orthopaedic Research*, vol. 35, no. 10, pp. 2203–2210, 2017.

2.1 Introduction

In conventional DXA analysis, BMD values are averaged in pre-defined regions of interest. This data averaging aims to account for morphological variation due to differences in patient anatomy and positioning during scan acquisition, but limits our understanding of more focal BMD deficits. Recently, Morris et al. [36] reported a high-resolution computational method for DXA analysis, termed DXA region free analysis (RFA). DXA RFA maps each scan into a standardised coordinate space to account for the morphological variation between scans [36]. This image warping establishes a correspondence between pixel coordinates across the subjects. This correspondence allows the deployment of appropriate statistical tests per each pixel coordinate to compare pixel-level BMD values between different groups. The objective of a statistical test is to evaluate the probability that the observed effect (or a more extreme one) has occurred by chance given the null hypothesis is true. This probability is called the *p-value.* Morris et al.[36] rendered the computed p-values per each pixel as a heat-map to visualise the statistical significance of the observed BMD change patterns in the femur.

In statistical hypothesis testing, it is often of interest to declare a non-zero effect or equivalently reject the null hypothesis. To draw such an inference, the computed p-value is compared against a cut-off level known as the *significance level* or the *alpha level* (α). If the p-value is below α, the observed effect is declared as *statistically significant.* Note that declaring an effect as significant is based on an arbitrary selection of the α-level as part of the study design. If the null hypothesis is true and a significant effect is declared, it is said to be a *false positive.* In a single hypothesis test, the probability of committing a false positive, known as *type one error*, would be less than α. However, increasing the number of performed tests will increase the chance of committing a false positive given the same α-level used in a single hypothesis test. This increased false positive rate hampers sound statistical inference over a group of p-values known as the *multiple comparisons problem.*

The multiple comparisons problem is not addressed in the proposed DXA RFA framework by Morris et al. [36]. In this chapter, I extend the DXA RFA method to allow global statistical inference within the femur by integrating the false discovery rate (FDR) analysis into the toolkit. This extension enables quantitation of the areal size and the anatomic position of regions with statistically significant BMD change without imposing any *a priori* assumptions on the analysis region of interest. The results are usable in a wide range of applications, including localising fracture-specific patterns (see section 4.4), but are discussed with particular reference to periprosthetic BMD change in this chapter. More specifically, I analyse scans collected during two previous randomised clinical trials conducted between 2004 and 2009; one comparing three cemented prosthesis design geometries [61], and the other comparing a hip resurfacing versus a conventional cementless prosthesis [62].

The remainder of this chapter is organised as follows. Section 2.2 reviews different techniques to address the multiple comparisons problem. Section 2.3 presents the motivation, the study population, and the study design for analysing periprosthetic BMD change after total hip arthroplasty (THA). Results and discussion are presented in sections 2.4 and 2.5. Section 2.6 concludes this chapter.

2.2 Multiple Comparison Problem

Notation: Consider the problem of testing N hypotheses $H_1, \cdots, H_N$, of which N_0 are true and $N_1 = N - N_0$ are false. Table 2.1 summarises the required notations as follows: N_{0n}, N_{0p}, N_{1n}, and N_{1p} are representing the number of true negatives, false positives, false negatives, and true positives, respectively. $N_p = N_{0p} + N_{1p}$ is the total number of tests that are declared as positive, i.e. an effect exists and the null hypothesis is rejected. $N_n = N_{0n} + N_{1n}$ is the total number of tests that are declared negative, i.e. the null hypothesis cannot be rejected.

Table 2.1: Possible outcomes when testing N hypotheses.

	Declared non-significant	Declared significant	Total
True null hypothesis	N_{0n}	N_{0p}	N_0
False null hypothesis	N_{1n}	N_{1p}	N_1
	N_n	N_p	N

In a single hypothesis testing procedure, a *test statistic*, i.e. a numerical value derived from a sample data-set, is selected such that it measures the distance between the data and the predictions under the null hypothesis, for example, t-statistic and χ^2-statistic. Each test statistic is a random variable itself and so is denoted by a capital letter. For example, T denotes a t-statistic and the scalar t denotes the computed statistic on the sample data-set. The p-value is then defined as the probability of observing an effect at least as large as the one computed on the sample data-set when the null hypothesis is true, i.e. $p = \mathcal{P}(T > t | H = 1)$. Conclusions based on the statistical tests are uncertain and, in general, an acceptable maximum probability of committing a false positive, known as the alpha level, is selected. The null hypothesis is rejected if $p < \alpha$. It is important to note that the computed p-value is itself a random variable with uniform distribution under the null hypothesis, i.e. $P|H = 1 \sim \mathcal{U}(0, 1)$. Therefore, rejecting a test with a p-value less than α ensures controlling the false positive rate at the level α, i.e. $\mathcal{P}(\text{reject } H|H = 1) = \mathcal{P}(P \leq \alpha|H = 1) = \alpha$.

Multiple comparisons refer to the testing of more than one hypothesis at a time. In multiple hypothesis testing, a variety of generalisations are possible to control the false positive rate. *Per-comparison error* (PCE) rate is defined for each test as the probability of the type one error. The average PCE rate can be defined as the expected value of the number of false positives divided by the total number of tests.

$$\text{PCE} = \alpha_n = \mathcal{P}(\text{reject } H_n|H_n = 1) \quad \text{for } n = 1, \cdots, N \tag{2.1}$$

$$\text{average PCE} = E(\frac{N_{0p}}{N}) = \frac{1}{N}\sum_{n=1}^{N} \alpha_n \tag{2.2}$$

Family-wise error (FWE) rate is defined as the probability of committing at least one false positive in the family.

$$\text{FWE} = \mathcal{P}(N_{0p} \geq 1) \tag{2.3}$$

False discovery rate (FDR) is the expected proportion of false positives among all detected pixels.

$$\text{FDR} = E(\frac{N_{0p}}{N_p}) \tag{2.4}$$

In Eq. 2.4, the ratio $\frac{N_{0p}}{N_p}$ is defined zero when $N_p = 0$.

When controlling the PCE rate in a multiple hypotheses testing problem, the FWE rate increases, often sharply, with the number of tests. This may pose serious consequences if the set of tests must be considered as a whole. Numerous techniques have been proposed to address this issue. Below, two common approaches based on the Bonferroni method (section 2.2.1) and the random field theory (section 2.2.2) are presented to control the FWE rate. Section 2.2.3 reviews the Benjamini and Hochberg technique [63] to control the FDR. For explanation purposes, I assumed a z-test wherever a test statistic should be considered. However, the discussion is valid for any other test statistics.

2.2.1 Bonferroni Correction

Calculation of the exact FWE rate is not trivial. The Bonferroni inequality provides a conservative upper bound on the FWE rate. Assume $\{z_i\}_{i=1}^N$ are the corresponding Z-statistics per each hypothesis test H_n. Then, the FWE rate is bounded by $N\alpha$ where α is the PCE rate per each test;

$$\begin{aligned} \text{FWE} &= \mathcal{P}\{\cup_n (Z_n \geq z_n) | H_1 = 1, \cdots, H_N = 1\} \\ &\leq \sum_n \mathcal{P}(Z_n \geq z_n) \leq N\alpha. \end{aligned} \tag{2.5}$$

(A) The Simple Bonferroni Method

Given the Bonferroni inequality (Eq. 2.5), rejecting each hypothesis H_n with $\alpha_n = \frac{\alpha}{N}$ will control the FWE rate at the α-level. When all α_n are chosen to be equal, the method is called the *unweighted* Bonferroni method. However, as long as the sum of α_n equals α, the method still controls the FWE rate at the desired α-level.

This simple method is an example of a *single-stage* testing procedure. One drawback with this technique is the low average power for testing the individual hypotheses; the larger the number of hypotheses, the smaller the average power. To partially overcome this limitation, *multi-stage* testing procedures based on sorted p-values are deployed in the literature [64].

(B) Holm's Sequential Procedure

This method is deployed in multiple stages as follows [65]: sort the p-values increasingly so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ with arbitrary ordering in case of ties. At the first stage, reject $H_{(1)}$ if $p_{(1)} \leq \frac{\alpha}{N}$; otherwise, all hypotheses are accepted and the method terminates. If $H_{(1)}$ is rejected, continue the procedure with the remaining hypotheses; at any stage n, reject $H_{(n)}$ if all previous hypotheses $H_{(n')}$ with $n' < n$ are rejected and $p_n \leq \frac{\alpha}{N-n+1}$.

(C) Hochberg's Sequential Procedure

This method is based on the Simes equality [66]; if all hypotheses are true and independent, then with probability $1-\alpha$, $p_{(n)} \geq \frac{n\alpha}{N}$. $p_{(n)}$ are increasingly sorted p-values. Hochberg [67] utilised the Simes equality to modify the Holm's procedure as follows: at the first stage, check if $p_{(N)} \leq \alpha$, then reject all hypotheses; otherwise, continue the procedure with the remaining hypotheses. At any stage $n = 1, \cdots, N$, if $p_{(n)} \leq \frac{\alpha}{N-n+1}$, then reject all hypotheses $H_{(n')}$

with $n' \leq n$ and terminate the procedure. If no n exists such that $p_{(n)} \leq \frac{\alpha}{N-n+1}$, all hypotheses are accepted.

(D) Hommel's Sequential Procedure

Hommel [68] utilised the results of the Simes equality [66] to modify the Bonferroni method. This procedure is more powerful than the Hochberg's method [67]. Let n be the largest integer for which $p_{(N-n+n')} > \frac{n'\alpha}{n}$ for all $n' = 1, \cdots, n$. If no such n exists, reject all hypotheses; otherwise, reject all $H_{(n'')}$ with $p_{(n'')} \leq \frac{\alpha}{n}$.

2.2.2 Random Field Theory

In medical imaging applications, it is of interest to make an inference over a search volume where the set of test statistics for each voxel are rendered as a *statistical parametric map* (SPM) [69]. For example, detection of activated regions in the brain in response to a certain condition [70] or identifying regions with thinner femoral cortex in a fracture group in comparison to a non-fracture group [71]. To address the multiple comparisons problem over the search volume, the conventional Bonferroni-based correction methods are often too conservative for use in the imaging data; the average power in localising an effect at each individual voxel is too small. This can be explained by the fact that the number of independent tests is much fewer than the number of voxels in the image due to the spatial correlation between voxels [69]. In fact, attributing any effect to the voxels is ill-posed as the number of voxels is more or less arbitrary in an image. Therefore, instead of controlling false positive voxels, a statistical map can be seen as a *random field* where inferences are made based on the topological features of an SPM [69].

A random field can be simply defined as a *stochastic process* over a parameter space of dimensionality $D \geq 1$ [72, 73]. Random field theory (RFT) allows analysing SPMs to identify local extremes that are unlikely to happen under the null hypothesis. Assume $X(\boldsymbol{s})$ denotes an SPM where $\boldsymbol{s} \in \mathbb{R}^D$ characterises the parameter space. The FWE rate defined in Eq. 2.3 can be then interpreted as the probability of observing a local maximum in the SPM;

$$\text{FWE} = \mathcal{P}(N_{0p} > 0) = \mathcal{P}(X(\boldsymbol{s}_i) \geq t_\alpha; \text{for some } i) = \mathcal{P}(\max_i X(\boldsymbol{s}_i) \geq t_\alpha), \tag{2.6}$$

where i indexes the pixels in the image and t_α is the height threshold corresponding to the significance level α. Given the spatial correlation between pixels in an SPM, when one pixel passes the threshold t_α, adjacent pixels that are close enough to this pixel will pass the threshold as well. Thus, deriving an expression for $\mathcal{P}(\max_i X(\boldsymbol{s}_i) \geq t_\alpha)$ is equivalent to the probability that a local maximum appears on the map. Since the number of peaks is less than the number of voxels, RFT yields a less conservative threshold for smoothed maps in comparison with the Bonferroni-based methods.

To localise significant regions in a statistical map X, a suitable height threshold t_α is calculated and pixels above this threshold are selected as significant events. This method is called *height thresholding* in the literature [69]. How RFT selects this threshold is discussed in section 2.2.2.(A). Other topological features including the area of local peaks or the number of local peaks are also suggested in the literature [74].

(A) Peak level analysis

Deriving an expression for the probability of local peaks is not trivial from the statistical point of view [73]. The way RFT solves this problem is by using results that give the expected *Euler Characteristic* for a smooth SPM that has been thresholded. The EC is discussed in details in section 2.2.2.(C), but for now, it is important to note that the expected EC leads to the expected number of local peaks, and so it can be used to approximate the FWE rate [70].

$$\text{FWE} \approx E(\text{EC}) = \sum_{d=0}^{D} R_d\ \rho_d(t). \tag{2.7}$$

In Eq. 2.7, D is the dimension of the statistic map. For example, for 2D maps $D = 2$. R_d is the number of d-dimensional resolution elements or concisely *resels*, which is explained in section 2.2.2.(B). The R_d would be a constant number in Eq. 2.7 that is dependent on the shape and smoothness of the map. $\rho_d(t)$ is the EC density function which is only dependent on the distribution function of the statistic map. For example, the EC density functions for the student-t distribution with ν degree of freedom are given below [70]:

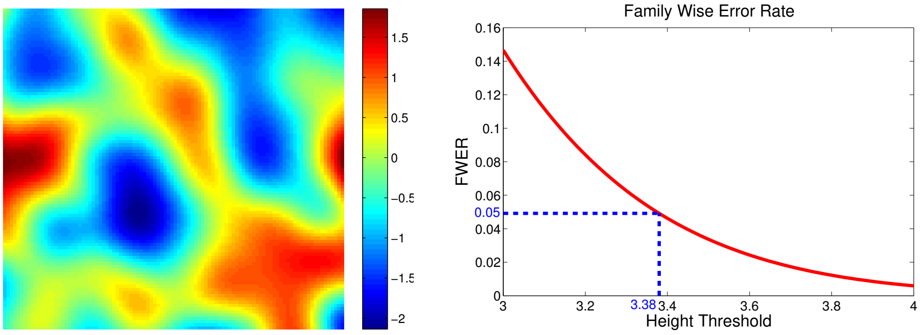
$$\rho_0(t) = \int_t^\infty \frac{\Gamma(\frac{\nu+1}{2})}{(\nu\pi)^{1/2}\Gamma(\frac{\nu}{2})(1+\frac{u^2}{\nu})^{-1/2(\nu+1)}}du \tag{2.8}$$

$$\rho_1(t) = \frac{\sqrt{4\ln 2}}{2\pi}(1+\frac{t^2}{\nu})^{-1/2(\nu-1)} \tag{2.9}$$

$$\rho_2(t) = \frac{4\ln 2}{(2\pi)^{3/2}}\frac{\Gamma(\frac{\nu+1}{2})}{(\frac{\nu}{2})^{1/2}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-1/2(\nu-1)} \tag{2.10}$$

where $\Gamma(.)$ is the gamma function.

The application of RFT proceeds in the following stages: first, the smoothness, i.e. the number of resels, of the SPM is estimated. Second, FWE rate is computed for each threshold value t using Eq. 2.7. Finally, the corresponding threshold t_α is calculated and all supra-threshold pixels are marked as significant events. For example, Fig. 2.1(a) shows a smoothed Z-map of size 100×100 with *Full Width at Half Maximum* (FWHM) 20. FWHM is a metric to express the smoothness level of a smoothing kernel $g(x)$ defined as the difference between the two data points x_1 and x_2 at which the kernel function is equal to half of its maximum value. Fig. 2.1(b) shows the experimental estimation of the FWE rate. RFT gives the height threshold of 3.38 that is lower than the Bonferroni counterpart (4.42).



(a) Z-map **(b)** Selection of suitable height threshold

Figure 2.1: (a) A typical Z-map of size 100×100 with FWHM = 20. (b) The family-wise error rate is plotted as a function of the height threshold t_α. To control FWE rate at 0.05, a height threshold of 3.38 should be chosen, which is smaller than 4.42 in the Bonferroni correction.

(B) Resolution Elements

Worsley et al. [75] introduced the term resolution element or briefly resel to capture the concept of smoothness in an SPM. The number of resels can be thought as an analogy to the number of independent observations in an SPM, but it is not the same as the height threshold t_α also depends on the EC density functions (Eq. 2.7). Note that the number of resels is only dependent on the smoothness level and the geometry of the search volume. For a Gaussian random field of dimension D with the covariance matrix of partial derivatives Λ, FWHM at direction $i = 1, \cdots, D$ can be derived as [76]:

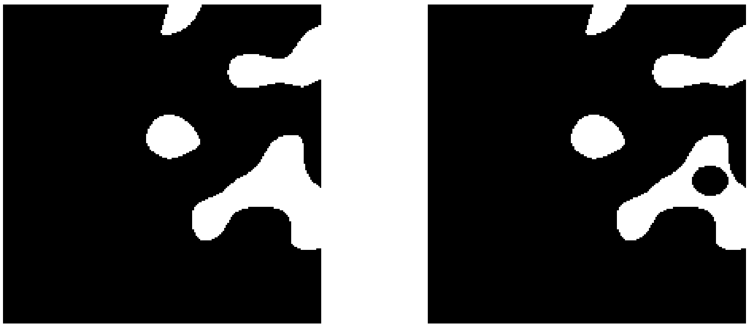
$$\begin{aligned} \mathrm{FWHM}_i &= \sqrt{8\ln(2) W_{ii}}, \\ \text{where} \qquad W &= (2\Lambda)^{-1}. \end{aligned} \tag{2.11}$$

Following estimation of the FWHM at each voxel, transferring to the resel space is simply achieved by division of voxel dimensions $d_x \times d_y \times d_z$ with FWHM_x, FWHM_y, and FWHM_z, respectively. The number of resels R_d is defined as follows [75]:

- R_3: It is defined as the volume of the search region in the resel space.
- R_2: It is defined as half of the surface area of the search region in the resel space. If the map is 2D, then R_2 is simply the area of the search region.
- R_1: It is defined as twice the average width of all bounding boxes of the search region in the resel space. If the map is 2D, then R_1 is half of the perimeter length of the search region in the resel space. If the map is 1D, then R_1 is simply the length of the search region in the resel space.
- R_0: It is the same as the Euler characteristic (EC) of the search region (see section 2.2.2.(C)).

(C) Euler Characteristic

For an SPM X, the *excursion set* is defined as the supra-threshold voxels where $X(\boldsymbol{s}) > t_\alpha$. It is clear that for any arbitrary height threshold t_α, a unique excursion set can be defined. In the special case of 2D maps, any excursion set would be comprised of some blobs and holes (Fig. 2.2). Then, EC can be defined over an excursion set as the number of connected blobs above the threshold value t_α minus the number of holes.



(a) EC = 4 **(b)** EC = 3

Figure 2.2: The Euler characteristic (EC) for two arbitrary excursion sets. Panel (a) shows an excursion set with 4 blobs without any hole, and so EC = 4 - 0 = 4. Panel (b) shows an excursion set with 4 blobs and 1 hole, and so the EC = 4 − 1 = 3.

Worsley et al. [75] provided an exact mathematical definition for the EC likewise the classical definition for the polyhedrons, i.e. $\chi = V - E + F$. Here,

V, E, and F denote respectively the number of vertices, edges, and faces. The formulation proposed by Worsley et al. [75] is as follows. Assume a cube of $2 \times 2 \times 2$ with 8 adjacent vertices (Fig. 2.3). Let V be the number of vertices above threshold t_α, E be the number of edges connecting the vertices inside the excursion set, and F be the number of faces that all 4 vertices belong to the excursion set. In this way, EC is defined as:

$$\chi = V/8 - E/4 + F/2 - C, \tag{2.12}$$

where C is a binary variable indicating whether all vertices of the cube lie inside the excursion set or not. For a specific search volume, EC is defined as the sum of χ in Eq. 2.12 for all the cubes in the region.

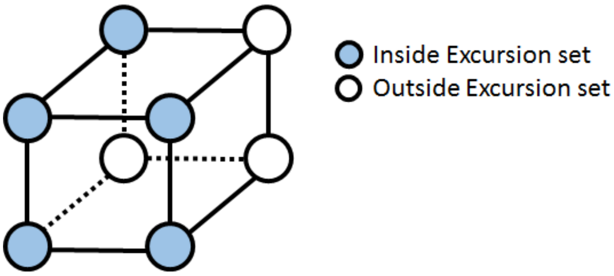


Figure 2.3: For a three-dimensional search volume, the Euler characteristic is computed by summing the contributions from all cubes in the region. An arbitrary cube with eight voxels is plotted to illustrate this calculation. Each solid circle represents one voxel inside the excursion set, while each hollow circle represents a voxel outside the set. For this example, $V = 5$, $E = 5$, $F = 1$, and $C = 0$ (see Eq. 2.12). Hence, $\chi = \frac{-1}{8}$.

Since EC is defined based on the excursion set of a random field, it is a random variable itself with a specific probability density function $\rho(t)$. In [70], it is calculated for Gaussian, Student-t, and Fisher random fields for dimensions $d = 0, 1, 2, 3$. Eq. 2.8-2.10 present density functions for t-distribution with ν degrees of freedom. The expected EC can be computed using Eq. 2.7. Fig. 2.4 shows the expected EC for an SPM of square shape (100×100 pixels) with FWHM $= 6$.

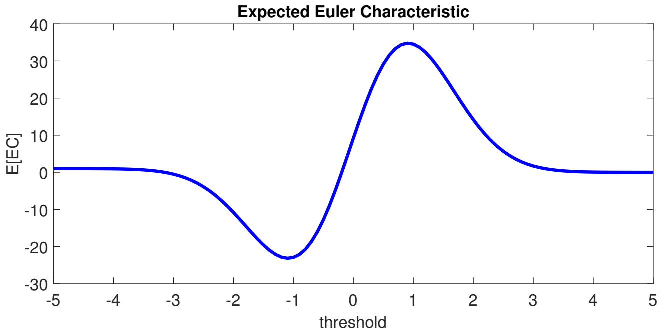


Figure 2.4: The expected Euler characteristic $E(\chi)$ is plotted against the height threshold t for an SPM of square shape (100×100 pixels) with FWHM = 6. At large threshold values, $E(\chi)$ would approximate FWE rate.

For high thresholds t_α, a blob may appear or not; thus, the EC would be one or zero. With this intuition, the expected value of EC, $E(\chi)$, can be interpreted as the probability that a local peak appears on the map. More specifically,

$$\mathcal{P}(\max_i X(\boldsymbol{s}_i) \geq t_\alpha) \approx \mathcal{P}(\chi \geq 1) \approx E(\chi) \quad \text{as} \quad \mathcal{P}(\chi > 1) \to 0 \quad \text{for } t_\alpha \to \infty \tag{2.13}$$

This explanation justifies the estimation of the FWE rate in Eq. 2.7. Fig. 2.5 shows the FWE rate for different thresholds based on the random field theory (blue solid line) and the experimental simulations (red dashed line). Please note that the expected EC gives an upper bound on the actual FWE rate.

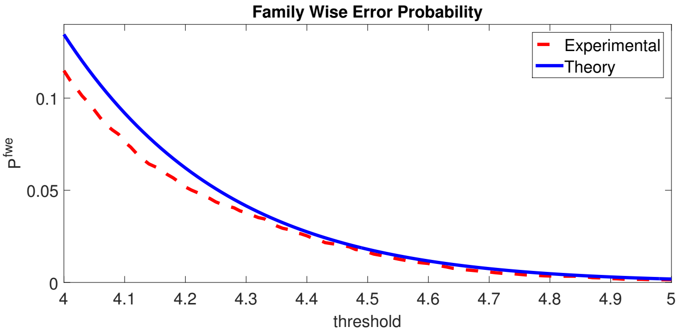


Figure 2.5: Theoretical versus experimental computation of family-wise error rate. The blue solid line represents the estimated FWE rate using Eq. 2.7 versus the experimental value (red dashed line) computed based on 10,000 SPMs with a normal distribution. FWHM = 6 and the search region is assumed to be an square of size 100×100 pixels. Note that the expected EC gives an upper bound on the actual FWE rate.

2.2.3 False Discovery Rate Analysis

In some scenarios, it is of interest to control the number of erroneous rejections of null hypothesis rather than whether any error was made at all. In this case, Benjamini and Hochberg [63] introduced FDR as an alternative to the FWE rate to address the multiple comparisons problem. FDR is the expected proportion of false positives among all detected pixels (Eq. 2.4).

FDR can be interpreted as a weak controller of the FWE rate [69]; given that all null hypotheses are true ($N_1 = 0$), FDR controls the false positives exactly in the same manner as the FWE rate. In case of an arbitrary mixture of null hypotheses ($N_1 > 0$), it can be shown that the FDR is always less than or equal to the FWE rate; if a procedure controls the FWE rate at the level α, it also controls the FDR at the level α. However, the reverse is not true.

Benjamini and Hochberg [63] introduced a simple procedure to control the FDR at a predefined level α. The BH-FDR procedure is as follows: sort the p-values increasingly so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ with arbitrary ordering in case of ties. Let k be the largest n for which $p_{(n)} \leq \frac{n}{N}\alpha$. Then, reject all hypotheses $H_{(n)}$ for $n = 1, \cdots, k$. This procedure guarantees control of the FDR at the level α [63].

FDR does not provide any corrected p-values. However, q-value, as an analogy to the p-value, can be defined as the minimum FDR level α for which the test hypothesis $H_{(n)}$ would be rejected. The mapping from p-values to q-values is obtained as follows. First, sort the p-values increasingly as mentioned above. The corresponding q-values are then given by $q_{(n)} = \min(p_{(n)}\frac{N}{n}, 1)$. With this interpretation, to control the FDR at the level α, all tests with $q_n \leq \alpha$ are rejected.

Although FDR does not explicitly use the notion of smoothness in SPMs deployed in RFT, it is still independent of the number of pixels in the map. Since FDR controls the expected proportion of false positives among all detected pixels, doubling the number of pixels, for example, would double both the number of detected pixels as well as the number of false positive pixels, and so the ratio would remain unchanged.

The primary BH-FDR algorithm requires the independence of the test statistics corresponding to the true null hypotheses. However, Benjamini and Yekutieli [77] showed that the method is still valid to be used on correlated tests under the Positive Regression Dependency on Subset (PRDS) condition. PRDS seems a reasonable condition for many medical imaging applications. For example, Li et al. [78] used BH-FDR analysis to identify fracture-critical regions inside the proximal femur. Glickman et al. [79] also suggested the BH-FDR method as an alternative to Bonferroni corrections in medical applications.

2.3 Periprosthetic BMD Analysis

2.3.1 Motivation

Prosthesis design influences the local mechanical environment of the proximal femur after THA, resulting in strain-adaptive bone remodelling [80, 81, 82]. Several factors influence the extent of bone loss that occurs around different prosthesis types; including prosthesis geometry, material stiffness, method of fixation, and surface coating [83, 84, 85, 86, 87, 88, 89]. Periprosthetic bone loss is a risk factor for fracture and causes reconstruction challenges at revision surgery [90, 91].

There is a need for high-resolution, low-radiation exposure technologies for evaluating the bone architectural changes associated with different biomaterial designs and implant geometries [92]. Such technologies would facilitate the non-invasive clinical assessment of novel prostheses that aim to better mimic the natural loading environment, or have surface coatings that aim to modulate the biology of the local bone environment [93]. Here, I applied the DXA RFA integrated with FDR analysis to examine the impact of prosthesis design on strain-adaptive bone remodelling in the setting of two previously reported clinical trials using substantially different femoral prosthesis designs [61, 62]. In one trial, I compared three different geometries of cemented femoral prosthesis, the Charnley (DePuy International, Leeds, UK), Exeter (Stryker, Newbury, UK), and the C-Stem (DePuy International, Leeds, UK). These prostheses may be classified as shape-closed or force-closed designs [94, 95]. Shape-closed designs, like the Charnley, use a bonded prosthesis-cement interface to fix the stem within the cement mantle, acting as a composite-beam, and transfer the load to the femur mainly at the level of femoral diaphysis. Force-closed designs, such as the double-tapered (Exeter) and triple-tapered (C-Stem) prostheses, have a non-bonded prosthesis-cement interface, where the stem acts as a mobile wedge within the cement mantle [94, 96]. This allows initial distal migration to set up hoop stresses in the proximal cement mantle resulting in more proximal load transfer between the femoral prosthesis and the host bone [97]. In the other trial, I compared bone remodelling around a hip resurfacing prosthesis versus a conventional cementless total hip replacement. The load transfer pattern in hip resurfacing occurs directly from the femoral head to the metaphysis, and is thought to be more representative of that found in the native proximal femur than that for a conventional stemmed prosthesis [98, 99, 100, 101, 102].

2.3.2 Study Populations and Scan Acquisitions

Anonymised DXA scans from two previous ethically approved clinical trials, for which written, informed consent was provided, were examined using DXA RFA [61, 62]. All subjects underwent surgery for idiopathic or secondary osteoarthritis, and were free from use of drugs known to affect BMD. All scans were acquired using a Hologic QDR 4500A fan-beam densitometer (Hologic Inc., Waltham, MA), using the *metal removal hip* scanning mode with a point resolution of 0.6 mm and a line spacing of 1.1 mm. Scans were performed with the subject in the supine position with the legs in neutral rotation and full extension. Scan acquisition was started approximately 2.5 cm distal to the tip of the femoral prosthesis, with the longitudinal axis of the prosthesis shaft vertical and occupying the centre of the scan field. The scan was continued proximally until 2 cm above the tip of the greater trochanter [103].

2.3.3 Study Design

(A) FDR Validation

To investigate the reliability of FDR algorithm incorporation into the DXA RFA framework, I examined sequential DXA scans taken on the same day after repositioning in 17 men (mean age 50 years, range 33–67) and 12 women (mean age 53 years, range 35–61). Scans were acquired a mean of 6 months (SD 3) after THA [103]. The hypothesis tested here was that no significant differences are expected in measured pixel-level BMD between the individual scan pairs at FDR level of 0.05.

(B) The Effect of Cemented Stem Design on Bone Remodelling

The subjects in this study were randomised at a ratio of 1:1:1 to receive either a cemented composite-beam prosthesis (Charnley, DePuy Synthes Ltd, $n = 35$), a double-tapered prosthesis (Exeter, Stryker UK Ltd, $n = 38$), or a triple-tapered prosthesis (C-stem, DePuy Synthes Ltd, $n = 38$) [61]. All patients were mobilised with unrestricted weight bearing on the first or second postoperative days. BMD was measured at postoperative baseline within 1 week of surgery, and at 3, 6, 12, and 24 months later using the same Hologic densitometer.

(C) Effect of Hip Resurfacing Versus Cementless THA on Bone Remodelling

The subjects in this study were randomised at a ratio of 1:1 to receive either a hip resurfacing prosthesis (Articular Surface Replacement (ASR) total femoral prosthesis, DePuy Synthes Ltd, $n = 13$) or THA using a cementless, proximally plasma-coated, titanium femoral component (Bi-metric, Biomet, Bridgend, UK, $n = 17$) [62]. All patients were mobilised full weight bearing on the first or second postoperative days. BMD was measured at postoperative baseline within 1 week of surgery, and at 2, 12, and 24 months later using the same Hologic densitometer.

(D) Baseline Analysis

The baseline demographic characteristics of the subjects between each of the prosthesis groups were compared using the χ^2 test, Fisher's exact test, the Mann–Whitney U test, or Student's t-test, as appropriate. The mean distribution of pixel BMD values among the post-operative baseline scans was computed for each prosthesis.

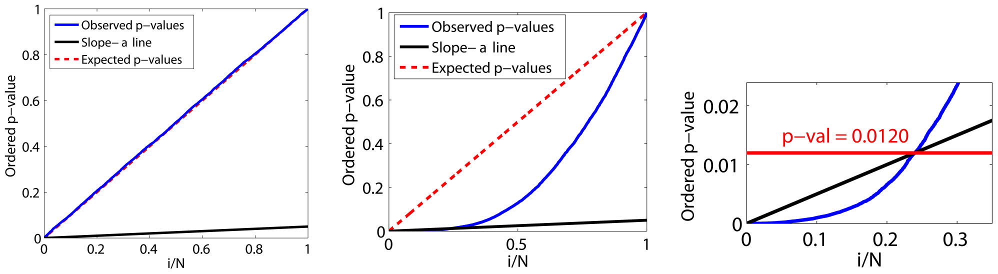
(E) Follow-up analysis

For each Hologic prosthetic hip scan, the pixel-level BMD map was extracted from the two archived Hologic scan files (.p and .r files) using DXA RFA based upon a proprietary DXA bone map extraction algorithm APEX 3.2 (Hologic Inc, Waltham, MA) [36]. DXA RFA rendered a bone map with approximately 14,000 pixels per scan where the pixel size was $0.56 \times 0.56\,\text{mm}^2$. Next, a reference template was learned per each prosthesis design and scans were warped into their corresponding template to remove shape variability between scans as described previously (see section 1.3.5) [36]. The pixel-level BMD change with respect to the baseline measurement was examined using a paired t-test at each time-point. To address the multiple comparisons problem, I deployed the Bonferroni correction, RFT, and FDR techniques. However, the statistical power was limited with the Bonferroni and RFT techniques and so here I only report results for the FDR approach [63]. All pixels with $q \leq 0.05$ were selected as statistically significant. The areal size of regions with significant BMD change was quantitated as the fraction of the periprosthetic bone area, i.e. the number of pixels with $q \leq 0.05$ divided by the number of all pixels in the template. The pixel-level FDR q-values were also rendered as heat-maps to identify the anatomic location of significant BMD change events within the bone.

2.4 Results

2.4.1 FDR Validation

Figure 2.6(a) shows the P-P plots for the repositioned scans examined here. A P-P plot is a diagram of increasingly sorted observed p-values against the $i/(N+1)$ quantile of the uniform distribution, where N is the total number of observed p-values. Under the null hypothesis, the expected curve in the P-P plot is the diagonal line of identity. Large deviations from this diagonal have



(a) **(b)** **(c)**

Figure 2.6: P-P plot for the FDR analysis. (a) The P-P plot for the set of 29 repositioned pairs of scans. As shown, the blue line almost perfectly follows the diagonal line of identity indicating that the null hypothesis of no change is valid in all pixels. (b) The P-P plot for Charnley prosthesis after 24 months. The blue line deviates below the line of identity, indicating the rejection of the null hypothesis. (c) All pixels below the slope-α line corresponding with the p-value less than 0.012 are statistically significant at $\alpha = 0.05$.

lower probability given the null hypothesis is true. As shown in Fig. 2.6(a), the P-P plot follows the line of identity. This means that no pixels with significant BMD change were identified across all pixels in the bone as expected. In comparison, Fig. 2.6(b) shows the P-P plot for the Charnley prosthesis after 24 months as an example where the null hypothesis is rejected, since the P-P plot deviates below the slope-α line (Fig. 2.6(c)).

2.4.2 Clinical Trial Subject Characteristics

The participants within each clinical trial were of similar age, sex distribution, and body mass index (Table 2.2). The subjects participating in the cemented stem geometry trial were older than those participating in the conventional cementless femoral prosthesis versus hip resurfacing trial (71±6 vs. 57±6, $p < 0.001$), and a greater proportion were female (53:58 vs. 22:8, $p = 0.013$). The BMI of participants in each study was 29.2 ± 4.4 versus 28.3 ± 4.4, respectively ($p = 0.397$).

Table 2.2: Characteristics of the patient populations participating in the DXA RFA analyses.

Cemented Femoral Stem Geometry Study

Characteristic	Charnley ($n = 35$)	C-Stem ($n = 38$)	Exeter ($n = 38$)	p-value
Age at surgery (years)	70 ± 6	71 ± 7	71 ± 6	^{a}0.929
Sex (M:F)	14:21	19:19	20:18	^{c}0.527
BMI (kg/m^2)	28.9 ± 4.6	29.2 ± 4.8	29.3 ± 3.9	^{a}0.914

Cementless Stemmed Versus Hip Resurfacing Study

Characteristic	Hip Resurfacing ($n = 13$)	Cementless Stem ($n = 17$)	p-value
Age at surgery (years)	57 ± 6	56 ± 6	^{b}0.320
Sex (M:F)	8:5	14:3	^{d}0.201
BMI (kg/m^2)	28.0 ± 5.9	28.6 ± 3.0	^{b}0.680

Continuous data are presented as mean±standard deviation, and analysis is between groups within each study using ^{a}ANOVA or ^{b}Mann–Whitney test. Categorical data were analysed using the ^{c}chi-squared or ^{d}Fisher's exact test.

2.4.3 Post-Operative Baseline Mean BMD Distribution

Baseline scans for all prosthesis groups showed a pattern of mean BMD distribution consistent with proximal femoral architecture with differentiation of cancellous versus cortical bone (Fig. 2.7). Areas of lowest BMD (approximately, 0.5–1 g/cm^2) were observed in the cancellous bone within the greater and lesser trochanter. BMD was highest (2–3 g/cm^2) in the cortical bone of the femoral diaphysis. Subjects with cemented prostheses showed the highest bone mass in the region of cementation, with a measured BMD of up to 4 g/cm^2.

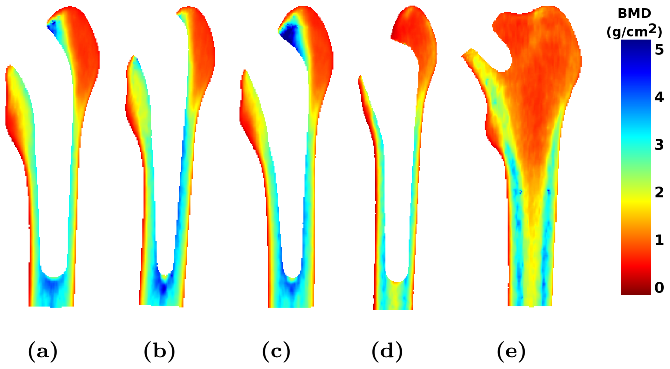


Figure 2.7: Mean pixel BMD distribution. The mean distribution of pixel BMD values at baseline measurement is shown for (a) composite-beam (Charnley), (b) double-taper (Exeter), (c) triple-taper (C-stem), (d) Bi- Metric total hip replacement, and (e) ASR hip resurfacing prosthesis designs, respectively.

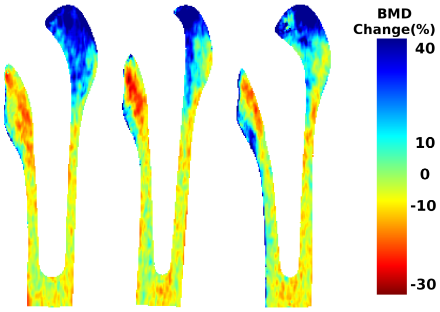
2.4.4 Effect of Cemented Stem Design on Bone Remodelling

The areal size of regions with significant BMD change and the corresponding BMD change in that regions are reported in Table 2.3. Fig. 2.8 shows the magnitude of pixel BMD change (%) at 24 months. Fig. 2.9 shows the corresponding FDR q-maps.

Table 2.3: Area size of regions with significant pixel BMD change ($q \leq 0.05$) with corresponding mean BMD change for three cemented prosthesis designs over 24 months.

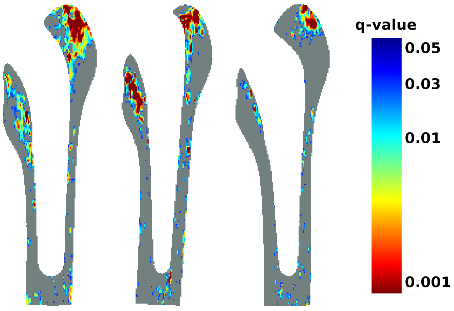
	Total		Increased BMD		Decreased BMD	
	Area(%)	Average BMD(%)	Area(%)	Average BMD(%)	Area(%)	Average BMD(%)
Charnley	31.4	12.2	16.6	32.1	14.8	-10.3
Exeter	24.1	5.3	9.7	31.2	14.4	-12.1
C-stem	12.7	12.1	6.5	34.5	6.2	-11.1

The area sizes are expressed as a percentage of the total area of periprosthetic bone in the template image. The average BMD change values are also expressed as a percentage of the baseline BMD value.



(a) (b) (c)

Figure 2.8: Longitudinal mean pixel BMD change over 24 months expressed as a percentage of the baseline measurement for (a) composite-beam (Charnley), (b) double-taper (Exeter), and (c) triple-taper (C-stem), respectively.



(a) (b) (c)

Figure 2.9: FDR q-value maps after 24 months are shown for (a) composite-beam (Charnley), (b) double-taper (Exeter), and (c) triple-taper (C-stem), respectively. All pixels with $q \leq 0.05$ are declared as significant events.

BMD change events occurred in discrete focal areas. An increase in bone mass was observed consistently in the greater trochanter area, a site of multiple tendinous attachments. Here, an average BMD increase of 32.1% within 16.6% of the periprosthetic bone area was observed for the cemented composite beam (Charnley) prosthesis, 31.2% within 9.7% of the area for the cemented sliding double-taper (Exeter) prosthesis, and 34.5% within 6.5% of the area for the cemented sliding triple-taper (C-stem) prosthesis was observed at 24 months ($q \leq 0.05$ for all comparisons).

An average bone loss of 10.3%, 12.1%, and 11.1% within an area of size 14.8%, 14.4%, and 6.2% was observed for the Charnley, Exeter, and C-stem prostheses, respectively ($q \leq 0.05$), mostly at the lesser trochanter. The greatest BMD changes occurred in the metaphyseal region for all cemented prosthesis designs, with relatively less change at the femoral diaphysis.

Bone remodelling patterns were both rate and location specific to each prosthesis design (Fig. 2.10–2.12). No significant BMD change was observed at any pixel at 3 months for the Charnley prosthesis. However, an average BMD increase of 12.7% was observed within a small fraction (0.7%) of the periprosthetic bone area for the C-stem prosthesis at this time-point ($q \leq 0.05$), and bone loss of 6.8% over 7% of the bone area medial to the Exeter prosthesis ($q \leq 0.05$).

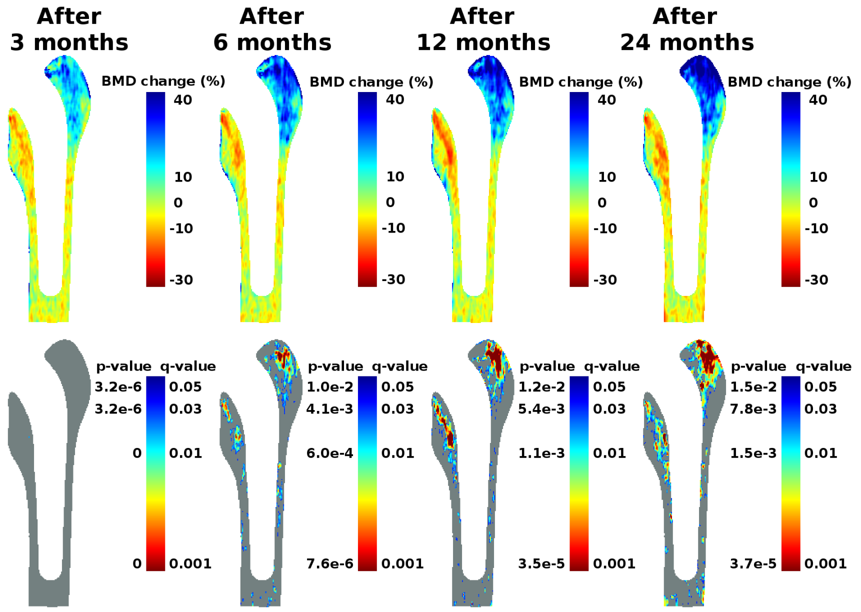

After 3 months
After 6 months
After 12 months
After 24 months
BMD change (%)
40
10
0
-10
-30
p-value q-value
3.2e-6 0.05
3.2e-6 0.03
0 0.01
0 0.001
1.0e-2 0.05
4.1e-3 0.03
6.0e-4 0.01
7.6e-6 0.001
1.2e-2 0.05
5.4e-3 0.03
1.1e-3 0.01
3.5e-5 0.001
1.5e-2 0.05
7.8e-3 0.03
1.5e-3 0.01
3.7e-5 0.001

Figure 2.10: Cemented composite beam (Charnley). The first row shows the pixel-level percentage change in BMD with respect to baseline at 3, 6, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar.

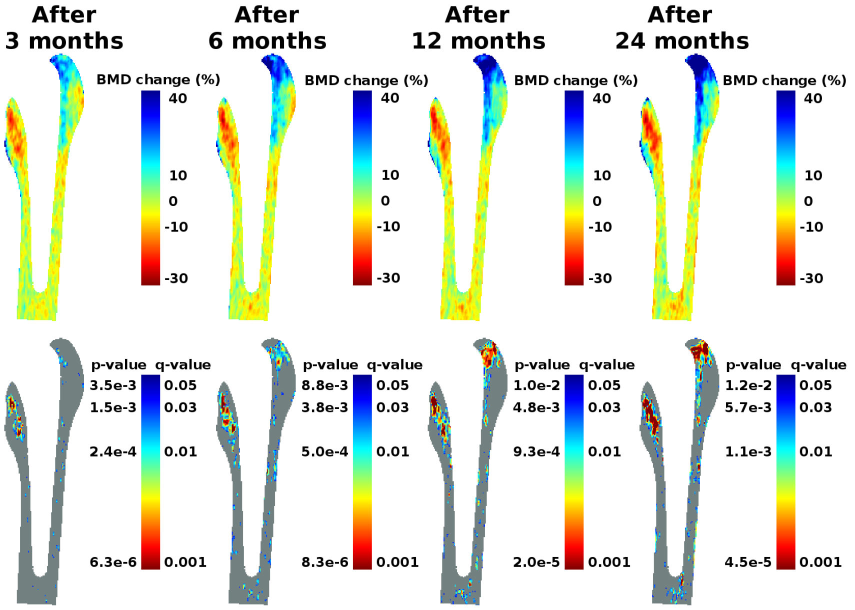

After 3 months
After 6 months
After 12 months
After 24 months
BMD change (%)
40
10
0
-10
-30
p-value q-value
3.5e-3 0.05
1.5e-3 0.03
2.4e-4 0.01
6.3e-6 0.001
p-value q-value
8.8e-3 0.05
3.8e-3 0.03
5.0e-4 0.01
8.3e-6 0.001
p-value q-value
1.0e-2 0.05
4.8e-3 0.03
9.3e-4 0.01
2.0e-5 0.001
p-value q-value
1.2e-2 0.05
5.7e-3 0.03
1.1e-3 0.01
4.5e-5 0.001

Figure 2.11: Cemented double-taper slip (Exeter). The first row shows the pixel-level percentage change in BMD with respect to baseline at 3, 6, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar.

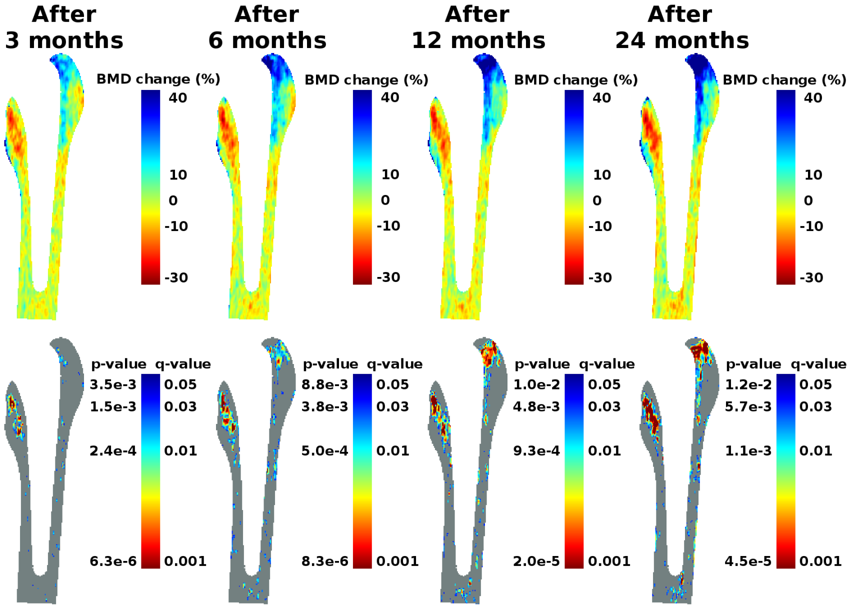

After 3 months
After 6 months
After 12 months
After 24 months
BMD change (%)
40
10
0
-10
-30
p-value q-value
3.5e-3 0.05
1.5e-3 0.03
2.4e-4 0.01
6.3e-6 0.001
p-value q-value
8.8e-3 0.05
3.8e-3 0.03
5.0e-4 0.01
8.3e-6 0.001
p-value q-value
1.0e-2 0.05
4.8e-3 0.03
9.3e-4 0.01
2.0e-5 0.001
p-value q-value
1.2e-2 0.05
5.7e-3 0.03
1.1e-3 0.01
4.5e-5 0.001

Figure 2.12: Cemented triple-taper slip (C-Stem). The first row shows the pixel-level percentage change in BMD with respect to baseline at 3, 6, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar.

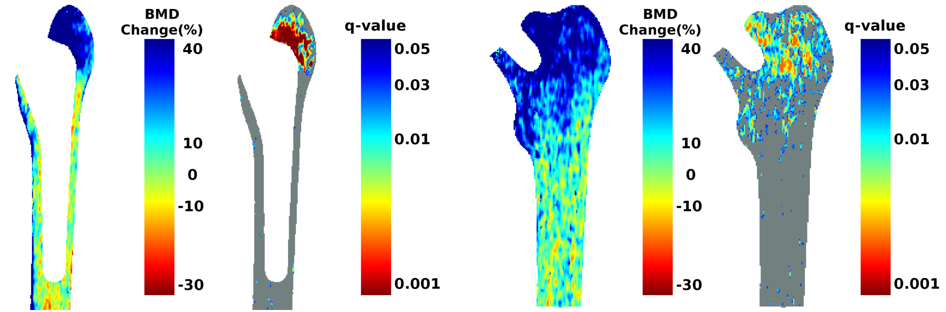
2.4.5 Effect of Hip Resurfacing Versus Cementless THA on Bone Remodelling

The areal size of regions with significant BMD change and the average BMD change in that regions are reported in Table 2.4. Fig. 2.13 shows the magnitude of pixel BMD change (%) and the corresponding FDR q-maps at 24 months. An average BMD increase of 35.9% over an area of 22.3% was observed locally at the greater trochanter for the Bi-metric prosthesis (Figs. 2.13(a) and 2.13(b)). A diffuse pattern of bone loss (14.3%) was also observed at the femoral shaft for the Bi-metric prosthesis at 24 months over a small fraction of the periprosthetic bone area (0.6%). No periprosthetic bone loss was observed around the hip resurfacing prosthesis at 24 months (Figs. 2.13(c) and 2.13(d)). However, an average BMD increase of 34.3% was observed over 30.7% of the proximal femoral metaphysis ($q \leq 0.005$).

Table 2.4: Area size of regions with significant pixel BMD change ($q \leq 0.05$) with corresponding mean BMD change for a conventional cementless femoral prosthesis (Bi-Metric) versus a hip resurfacing femoral prosthesis (ASR) over 24 Months.

	Total		Increased BMD		Decreased BMD	
	Area(%)	Average BMD(%)	Area(%)	Average BMD(%)	Area(%)	Average BMD(%)
Cementless stem	22.9	34.6	22.3	35.9	0.6	-14.3
Hip resurfacing	30.7	34.3	30.7	34.3	0.0	0.0

The area sizes are expressed as a percentage of the total area of periprosthetic bone in the template image. The average BMD change values are also expressed as a percentage of the baseline BMD value.



(a) BMD change **(b)** FDR q-value **(c)** BMD change **(d)** FDR q-value

Bi-Metric total hip replacement **ASR hip resurfacing**

Figure 2.13: Longitudinal mean pixel BMD change and the corresponding FDR q-value maps after 24 months are shown for Bi-Metric total hip replacement and ASR hip resurfacing prosthesis designs. BMD change is expressed as a percentage of the baseline measurement. All pixels with $q \leq 0.05$ are declared as significant events.

The contrasting patterns of focal trochanteric versus a widespread metaphyseal increase in BMD for the Bi-Metric versus ASR prostheses was apparent by 12 months, and persisted at 24 months (Figs. 2.14 and 2.15). The increase in bone mass around the ASR prosthesis was observed over the whole proximal femoral metaphysis, but was most densely concentrated in the bone adjacent to the lateral border of the prosthesis and the greater trochanter.

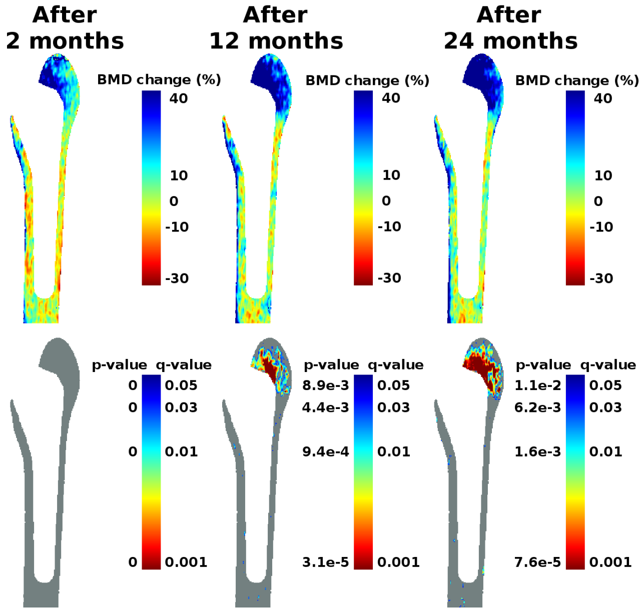

After
2 months
After
12 months
After
24 months
BMD change (%)
40
10
0
-10
-30
p-value q-value
0 0.05
0 0.03
0 0.01
0 0.001
8.9e-3 0.05
4.4e-3 0.03
9.4e-4 0.01
3.1e-5 0.001
1.1e-2 0.05
6.2e-3 0.03
1.6e-3 0.01
7.6e-5 0.001

Figure 2.14: Cementless hip replacement (Bi-Metric). The first row shows the pixel-level percentage change in BMD with respect to baseline at 2, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar.

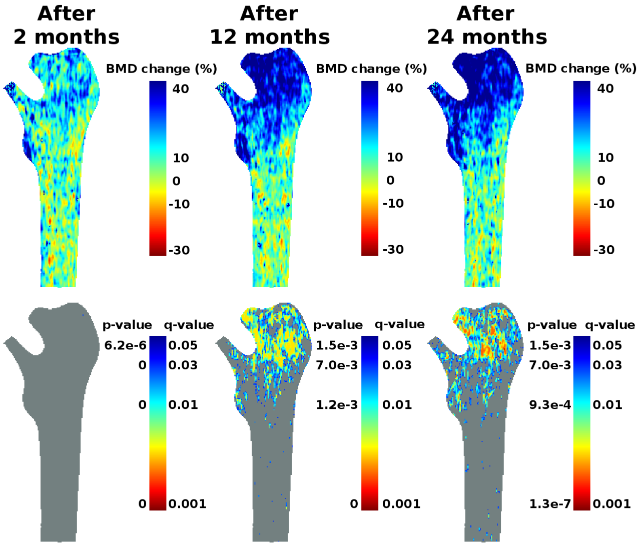

After
2 months
After
12 months
After
24 months
BMD change (%)
40
10
0
-10
-30
p-value q-value
6.2e-6 0.05
0 0.03
0 0.01
0 0.001
1.5e-3 0.05
7.0e-3 0.03
1.2e-3 0.01
0 0.001
1.5e-3 0.05
7.0e-3 0.03
9.3e-4 0.01
1.3e-7 0.001

Figure 2.15: Cemented hip resurfacing (Articular Surface Replacement). The first row shows the pixel-level percentage change in BMD with respect to baseline at 2, 12, and 24 months after the surgery. The second row shows the FDR q-values for longitudinal BMD change versus baseline. Local p-values correspondent to the q-values are also shown on the colour-bar.

2.5 Discussion

I analysed BMD changes around five different prosthesis designs using DXA RFA with FDR to demonstrate in high-resolution the effect that different prosthesis designs have on proximal femoral strain-adaptive remodelling. This approach is widely clinically applicable, non-invasive, and associated with low-radiation exposure. Some remodelling features are observed that were common to all prostheses, and others that were design-specific. Our finding that remodelling events occurred in small but spatially discrete *quanta* is consistent with the concept that post-operative bone remodelling occurs in discrete multicellular units [104, 105]. The observation that periprosthetic bone remodelling events are spatially complex, heterogeneous, and vary in density distribution with prosthesis design supports finite element analysis predictions [106]. It is also consistent with the view that the conventional ROI-based approach results in substantial data loss that impacts interpretation [103].

Consistently across all prosthesis designs, a gain in bone mass was found in the region of the greater trochanter, albeit this increase in bone mass was most widely distributed for the hip resurfacing group. Hip resurfacing was also the only prosthesis design around which increased bone mass occurred within the cortical bone of the proximal medial femur. This aligns with finite element predictions of the stress-redistribution at the femoral neck induced by this prosthesis class [107, 108]. Penny et al [62] have previously identified a similar BMD trend using conventional DXA, however, analysis using DXA RFA enabled precise localisation of the magnitude and area of these events. Although these data support the concept that head resurfacing prosthesis induce load transfer at the metaphyseal level, the approach does not quantitate over the studied time-frame the possible influence of adverse responses to metal debris on the local tissue microenvironment.

Previous conventional analysis using the seven Gruen zones showed that the greatest bone loss occurred in R7 and R6 over 2 years for the three cemented designs [61]. While DXA RFA analysis also showed significant bone loss adjacent to the prosthesis at lesser trochanter (Fig. 2.9), this was more precisely resolved using the RFA technique. Small areas of bone gain at the tendon-bone interface of the lesser trochanter were also observed (Fig. 2.8). In conventional DXA analysis, this spatial information is lost due to the averaging pixels into regions of interest. Moreover, this averaging may cancel out the bone loss with the bone gain in a region. For the hip resurfacing prosthesis, the conventional analysis showed a bone gain in all the Gruen zones [62]. This is compatible with spatial BMD change patterns in Fig. 2.13(c), where these changes are anatomically observed in the femoral shaft.

The incorporation of FDR into the DXA RFA framework enabled quantitation of the architectural details of femoral bone mass distribution and robust statistical analysis of BMD change events. These changes were also rendered as heat-maps for visual assessment. The FDR algorithm was applied to limit the proportion of false positives among statistically significant results. This primary concern is not directly addressed with Bonferroni-type adjustments [79, 63]. Moreover, the FDR approach gives increased statistical power in comparison with the methods that control the FWE rate [79, 63]. The validation of the FDR correction on the set of 29 repositioned scans confirmed the reliability of the method when applied in the DXA RFA framework.

2.6 Conclusion

This chapter presents the importance of deploying appropriate multiple hypothesis testing procedures in the setting of DXA RFA. More specifically, I demonstrated the integration of the FDR analysis with DXA RFA to analyse periprosthetic BMD changes around 5 different prosthesis designs. This integration allows quantification of the areal size of regions with a significant BMD change, which was not possible using the original RFA framework. In the next chapter, I extend the RFA framework to the native femur to analyses age-related bone deficits in the femur.

Chapter 3

Development of a Spatio-Temporal Atlas of Ageing Bone in the Native Proximal Femur

Ageing is associated with a gradual and progressive bone loss, which predisposes to osteoporosis. Given the close relationship between involutional bone loss and the underlying mechanism of osteoporosis, improving the understanding of the bone ageing process could lead to enhanced preventive and therapeutic strategies for osteoporosis. To facilitate this understanding, this chapter presents a method to develop a spatio-temporal atlas of ageing bone in the native proximal femur with a cohort of ~13,000 Caucasian women. To the best of our knowledge, this is the *first* spatio-temporal atlas of ageing bone. To this end, the region free analysis (RFA) framework is extended to the native femur and a fully automatic formulation applicable to large-scale datasets is presented. Furthermore, a novel cross-calibration technique is proposed to homogenise data from different vendors (Hologic and Lunar GE) into a unified multi-centre large-scale dataset.

The content of this chapter is adapted from the following publication:

Mohsen Farzi, Jose M. Pozo, Eugene McCloskey, Richard Eastell, J. Mark Wilkinson, and Alejandro F. Frangi, “Spatio-Temporal Atlas of Bone Mineral Density Ageing,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI2018). Springer, pp. 720–728, 2018.

3.1 Introduction

Ageing is associated with a gradual and progressive bone loss, which predisposes to osteoporosis. Osteoporosis, which literally means porous bone, is a bone disease characterised by low bone mass and micro-architectural deterioration. Given the close relationship between involutional bone loss and the underlying mechanism of osteoporosis, improving the understanding of the bone ageing process has been of interest for the osteoporosis research community [109, 110]. To facilitate this understanding, I propose a method to develop a spatio-temporal atlas of ageing bone in the femur.

Spatio-temporal atlases are useful tools for visualising and accessing a wide range of data in Medical Image Computing [111]. For example, brain atlases demonstrated great potential for visualising age-related pathology in Alzheimer's disease [112]. However, to the best of our knowledge, no bone ageing atlas has been developed in osteoporosis research so far. Developing a comprehensive model of involutional bone loss is a challenging task. Firstly, this requires a robust and accurate quantification technique for bone mineral density (BMD) measurement and its spatial distribution. Dual-energy X-ray Absorptiometry (DXA) is the reference gold standard to measure BMD in clinical practice [18]. In conventional DXA analysis, BMD values are averaged in *a priori* specified regions of interest (ROIs) to compensate for shape variation between scans (Fig. 3.1). This data averaging, however, may reduce our insight on more focal BMD deficits.

The second challenge is the ability to homogenise BMD measurements across different technologies, as a systematic difference exists between different proprietary DXA manufacturers [113, 114, 115]. Two broad cross-calibration procedures are commonly used. In one approach, each scanner is separately calibrated by fitting bone phantom measurements to its nominal density values. Pearson et al. [116] suggested an exponential curve and explored the technique using the European Spine Phantom (ESP) prototype. In the other approach, different scanners are calibrated simultaneously using density values measured on a common group of individuals [113, 114, 115].

Both DXA calibration procedures suffer from a number of key limitations. Cross-calibration using phantom measurements is challenged by a study conducted under the auspices of the International DXA Standardisation Committee (IDSC) [113]. Genant et al. [113] showed a disagreement between regression curves fitted to the phantom measurements and those fitted to the human measurements. On the other hand, the second approach requires re-

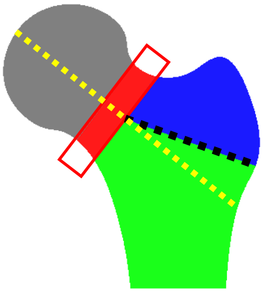


Figure 3.1: Femoral regions of interest (ROIs). The neck, trochanteric, and intertrochanteric regions are shown in red, blue, and green, respectively. The aggregation of these three regions comprises the total hip.

peated measurements of each subject across all machines [113, 114, 115]. This can be a serious limiting constraint in large multi-centre studies, where the first approach may be preferred in practice [117].

I address these challenges as follows: To maintain fidelity to high-resolution pixel BMD values, I have extended the region free analysis (RFA) technique, previously applied to quantitate periprosthetic bone loss [36], to the native femur (section 3.2.2). DXA RFA aligns each individual scan to a reference template and so eliminates the morphological variation between scans. This deformable image alignment establishes a virtual correspondence between pixel coordinates enabling statistical inference at a pixel level. To control the correspondence between scans, the initial RFA technique [36] used a set of anatomical landmark points selected semi-automatically around a prosthesis and the bone contour. Here, I automate the selection of landmark points using constrained linear models (CLM) [118]. This allows application of the toolkit to large-scale datasets with thousands of images.

To amalgamate data from different scanner technologies, I propose a novel cross-calibration technique based on human measurements where the requirement for scanning the same group of subjects on all the machines is moderated. In this method, the patient groups scanned on each machine are assumed to be independent and identically distributed samples of the same population, as all are white North European females from similar geographic latitudes. The proposed method minimises the mutual difference between the probability distributions of BMD values measured by each proprietary DXA scanner (section 3.2.3).

This chapter describes the development of the first spatio-temporal atlas of ageing bone in the femur generated using DXA data from over 13,000 subjects. To this end, I propose a fully automatic bone ageing analysis pipeline to ensure high-throughput computing applicable to large-scale datasets (Fig. 3.2). I also

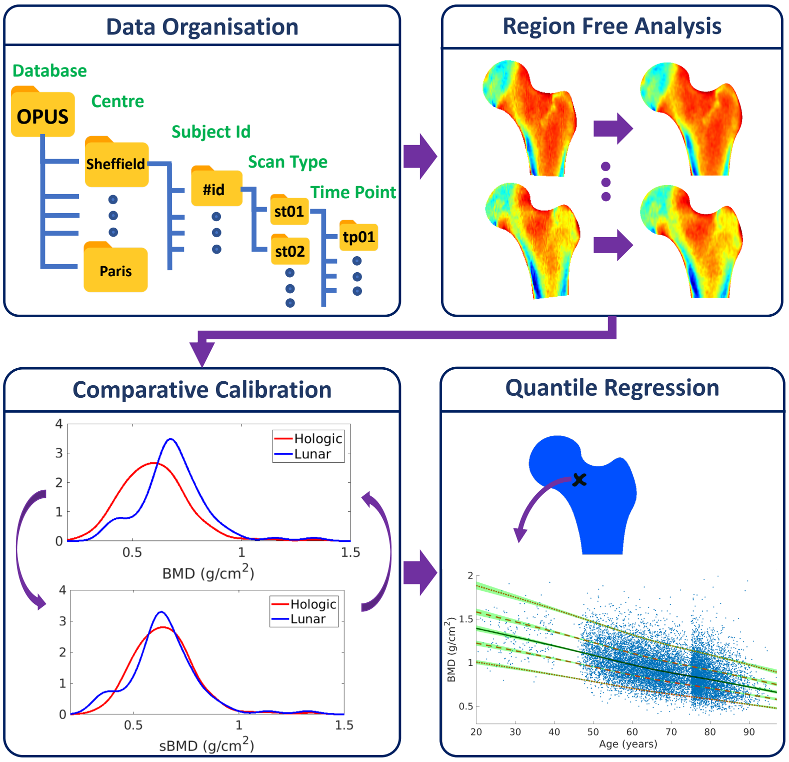

Data Organisation
Database
OPUS
Centre
Sheffield
Paris
Subject Id
#id
Scan Type
st01
st02
Time Point
tp01
Region Free Analysis
Comparative Calibration
Hologic
Lunar
0
1
2
3
4
0.5
1
1.5
BMD (g/cm²)
Hologic
Lunar
sBMD (g/cm²)
Quantile Regression
BMD (g/cm²)
0.5
1
1.5
2
20
30
40
50
60
70
80
90
Age (years)

Figure 3.2: Bone ageing analysis pipeline. Scans are automatically organised into sub-folders according to the study ID, geographic location, subject ID, anatomic site, and follow-up time points. Each scan is then warped into a reference domain to eliminate morphological variations. Pixel BMD values are calibrated across different centres such that the probability density functions match one another for a subset of samples matched for gender, age, body mass index, ethnicity, scan side, and geographic location. Finally, a set of smooth quantile curves is fitted to the standardised pixel BMD values for each pixel coordinate.

derive a set of reference quantile curves per each pixel coordinate to model the temporal BMD evolution as a function of age. I will show that ageing not only affects the amount of bone loss but also the anatomical distribution of bone within the femur. The developed atlas provides new insights into the spatial pattern of bone loss in the femur, for which conventional DXA analysis is insensitive.

3.2 Bone Ageing Analysis Pipeline

I propose an automated image analysis pipeline to construct a spatio-temporal atlas of ageing BMD in the native femur. The generated atlas models the distribution of BMD over the population as a function of age within the femur. Fig. 3.2 shows the conceptual outline of the proposed method. Below, different steps of the proposed framework are explained in detail: pre-processing and data organisation, region free analysis, comparative calibration, and quantile regression.

3.2.1 Pre-Processing and Data Organisation

(A) Pixel BMD Map Extraction

The raw data from the DXA scanner is not immediately usable for analysing BMD maps. To export BMD maps, the raw data requires processing using a proprietary algorithm integrated into a computer software package specific to its vendor. I have used Hologic Apex v3.2 (Hologic Inc, Waltham, MA) and Lunar enCORE v16 (GE Healthcare, Madison, WI) to extract pixel BMD information for scans collected on a Hologic QDR 4500A or a Lunar iDXA densitometer, respectively.

(B) Multi-Scale Analysis and Noise Reduction

Spatial resolution and pixel-wise noise levels vary between different DXA manufacturers. For example, the spatial resolution, expressed as height$\times$width, is 0.50×0.90 mm^2 for a Hologic QDR 4500A scanner and 0.25×0.30 mm^2 for a Lunar iDXA scanner. To compute the signal-to-noise ratio (SNR) at the pixel level, two sets of DXA scans were selected per each vendor. For the Hologic system, $n = 25$ scan pairs were deployed where each pair was collected on the same day from the same subject at the left hip with patient repositioning between scans. For the Lunar system, $n = 100$ scan pairs were selected where each pair was collected on the same day from the same subject at the left and the right hips. All scans were warped to the template space to establish anatomical correspondence between pixel coordinates (see section 3.2.2 for more details). Deming regression was then applied at each pixel to compute the SNR for each system (see section 3.2.3.(B) for more details on Deming regression). Here, the median SNR over all pixels in the template is used to compare the two systems (Fig. 3.3).

While the Lunar system provides a better resolution by a factor of two in height and three in width, pixel-wise SNR is approximately 10 dB higher in the Hologic system compared to the Lunar system (Fig. 3.3); larger pixels often result in higher SNR values. Observe that the lower SNR in the Lunar system versus the Hologic may also be attributed to other factors such as intrinsic bilateral differences between the left and the right hips or the operator proficiency to ensure consistent patient positioning between scans. Nonetheless, to enable pixel-wise comparison between different DXA manufacturers, an appropriate analysis scale should be selected such that both the spatial resolution and the pixel-wise SNR are consistent across the two systems.

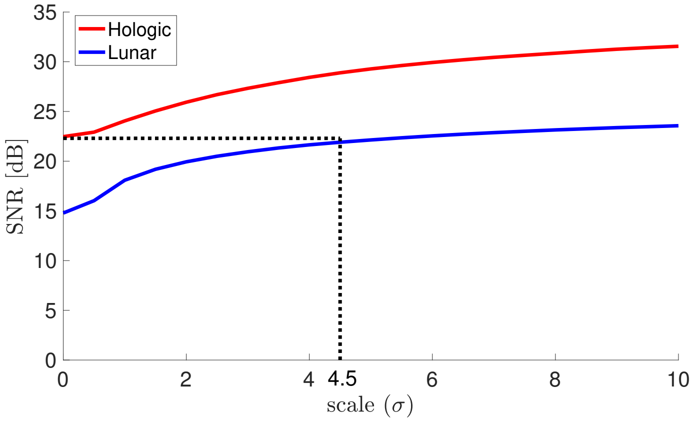


Figure 3.3: Median pixel-wise SNR for the Hologic QDR4500A versus the Lunar iDXA systems. The x-axis shows the standard deviation for the smoothing Gaussian kernel (σ) deployed to reduce the noise level. The y-axis shows the variation in the signal-to-noise ratio (SNR). For the Hologic system, pixel-level SNR was computed using a set of 25 scan pairs, each pair collected on the same day from the same subject with repositioning between scans (section 3.3.2). For the Lunar system, pixel-level SNR was computed using a random selection of 100 bilateral hip scans. Deming regression analysis was deployed to compute SNR for both systems (see section 3.2.3.(B)).

For the selection of an appropriate scale, all scans were resampled at an isotropic spatial resolution of 0.5×0.5 mm^2. Following the resampling, each image was smoothed with a Gaussian kernel to enhance the SNR. Fig. 3.3 shows the median pixel-wise SNR within the whole femur for both the Lunar and the Hologic systems at different scales. Given that the SNR is quite high (22.4 dB) for the Hologic system without any smoothing, I selected $\sigma = 0$ for the Hologic system and then found the σ at which the SNR is equivalent to 22.4 dB for the Lunar system, i.e. $\sigma = 4.5$.

Note that in the conventional region-based analysis where pixel BMD values are averaged in larger ROIs with a few hundreds of pixels, the noise level would be negligible anyway in each ROI and so the choice of the analysis scale does not matter.

(C) Data Organisation

To enable high throughput analysis of the imaging data, a data organisation module is crucial to address the following challenges. First, each DXA manufacturer uses a different format to store the pixel BMD maps. The Apex software uses a proprietary binary file format with a '.b' extension. The enCORE software, however, can export the bone maps in various file formats including the Matlab with a '.mat' extension. Second, the naming scheme is neither consistent between the manufacturers nor informative for image analysis purposes. Third, no meta-information such as the scanner type, the scan type or the pixel size is stored with the bone map. This requires passing the corresponding meta-information as extra variables in the pipeline that would result in an unnecessary increase in the complexity of the pipeline. Fourth, the number of scans per subject is often different; some subjects miss a few scans or have more images than expected.

To address these challenges, the exported bone maps are automatically organised into sub-folders according to the study ID, the geographic location, the anatomic site, and the follow-up ID (Fig. 3.2). A uniform naming scheme is used for the files and directories.

3.2.2 Region Free Analysis

The objective is to find a set of coordinate transformations such that the warped scans are aligned with each other in the template domain. Therefore, each pixel coordinate in the template domain corresponds to the same anatomical location in the image domains. This correspondence allows pixel level inference of the BMD values. The proposed technique has three steps (Fig. 3.4): automatic landmark localisation, template derivation using generalised Procrustes analysis, and pairwise registration between the reference template and each scan.

(A) Automatic Landmark Localisation

To compute the geometrical warp between the image domain and the template (see section 3.2.2.(C)), a set of robust landmark points is required. This section addresses the problem of automatically locating prominent feature points in the femur. A standard approach to this problem is to first build a model of shape and texture variation from a manually labelled training set, and then fit the model to an unseen image [118]. Below, statistical shape models (SSMs) [119] and statistical appearance models (SAMs) [120] are briefly reviewed, and later *constrained local models* (CLMs) [118], an elegant method of combining both shape and appearance models, is presented to find landmark points in the femur [121].

In SSMs, each object is represented by a set of landmark points. Let $\boldsymbol{p}_m = [x_{1,m}, x_{2,m}]^T$ denote the coordinates for the m^{th} landmark point and $\boldsymbol{s} = \left[\boldsymbol{p}_1^T, \cdots, \boldsymbol{p}_M^T\right]^T$ denote a shape in a $2M-$dimensional space $\mathbb{R}^{2M}$. The distribution of this vector, known as *point distribution function* (PDM), can

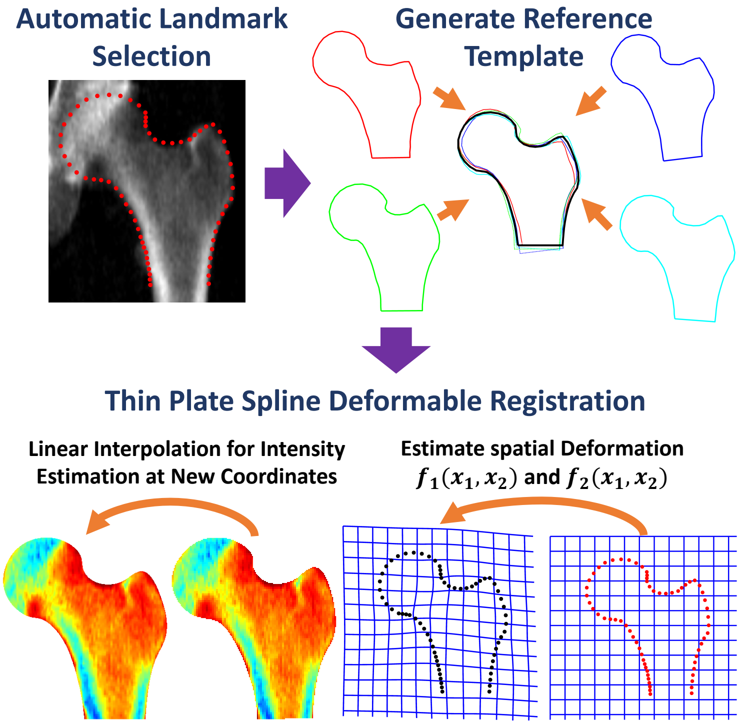

Automatic Landmark Selection
Generate Reference Template
Thin Plate Spline Deformable Registration
Linear Interpolation for Intensity Estimation at New Coordinates
Estimate spatial Deformation $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$

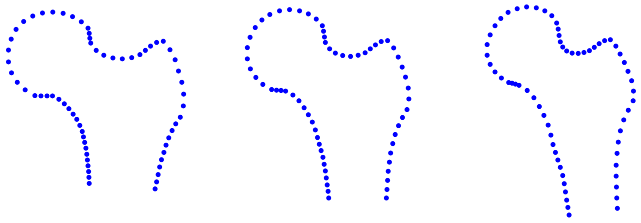
Figure 3.4: Conceptual illustration of region free analysis. Sixty-five landmark points are automatically selected around the bone contour. A reference shape is learned by averaging over all the scans after being aligned to a common position, scale, and orientation. A thin plate spline (TPS) deformation function is fitted for each individual scan such that the controlling landmark points are mapped to the corresponding reference landmark points in the template. Given the warp in the space, pixel intensities are estimated using a linear interpolation technique.

be approximated using the principal component analysis (PCA) [119]:

$$\boldsymbol{s} = \overline{\boldsymbol{s}} + \boldsymbol{\Phi}_s \boldsymbol{b}_s, \tag{3.1}$$

where $\overline{\boldsymbol{s}}$ is the mean shape, $\boldsymbol{\Phi}_s$ is a set of orthogonal modes of shape variation, and $\boldsymbol{b}_s$ is a vector of shape parameters. Fig. 3.5 shows the first mode of variation in the femur. To fit an instance of the model to an unseen image, first, a set of initial landmark points are defined in the image frame. Then, an iterative procedure is applied to improve the quality of fit as follows [119]: for each landmark point, the local image profile perpendicular to the image boundary is searched for an optimal match based on a similarity metric. Next, the new positions are mapped to the model space to avoid individual false detections that are inconsistent with the learned global shape configuration. The algorithm iterates between these two steps until convergence happens. Behiels et al. [122] applied this technique for segmenting the femur in radiographic scans.

In SAMs, both the shape and texture variability are modelled together



(a) $\bar{s} - 3\sigma\phi_1$ (b) $\bar{s}$ (c) $\bar{s} + 3\sigma\phi_1$

Figure 3.5: The first mode of shape variation in the proximal femur.

[120]. The shape model is learned similar to SSMs (Eq. 3.1). To build a statistical model of grey-level appearance, each sample image is warped to the mean shape so that the controlling landmark points in the image frame are matched to the landmark points in the template domain. The grey-level information is then collected over a region covering the landmark points from the warped scans. Let $\boldsymbol{g}$ denotes the vectorised pixel-level intensity information. A linear model of intensity variation can be learned using a PCA transformation as follows:

$$\boldsymbol{g} = \bar{\boldsymbol{g}} + \boldsymbol{\Phi}_g \boldsymbol{b}_g, \tag{3.2}$$

where $\bar{\boldsymbol{g}}$ is the mean grey-level vector, $\boldsymbol{\Phi}_g$ is a set of orthogonal modes of intensity variation, and $\boldsymbol{b}_g$ is a vector of texture parameters. The shape ($\boldsymbol{b}_s$) and texture ($\boldsymbol{b}_g$) parameters are then concatenated into a single vector $\boldsymbol{b}$, and a further PCA transformation is applied to link the shape and texture parameters together.

$$\begin{bmatrix} \boldsymbol{W}\boldsymbol{b}_s \\ \boldsymbol{b}_g \end{bmatrix} = \boldsymbol{b} = \begin{bmatrix} \boldsymbol{Q}_s \\ \boldsymbol{Q}_g \end{bmatrix} \boldsymbol{c} = \boldsymbol{Q}\boldsymbol{c}, \tag{3.3}$$

where $\boldsymbol{W}$ is a diagonal weight matrix for shape parameters accounting for the unit difference between the shape and texture models. The vector $\boldsymbol{c}$ includes the appearance parameters. To fit an instance of the model to an unseen image, an iterative procedure is applied to minimise the texture residual between the model and the target image [120].

CLMs combine the flexibility of appearance models with global shape constraints [118]. A joint shape and texture model is learned in a similar manner to SAMs (Eqs. 3.1, 3.2, and 3.3); however, the texture sampling method is in the form of rectangle patches around landmark points. In the CLM framework, a *response image* is generated per each landmark point independently. To generate a response image for the m^{th} landmark point, random patches at its local neighbourhood are selected and the correlation of each patch with *a priori* trained template is computed. Then, the objective function $\mathcal{J}(\boldsymbol{b}_s)$ is maximised to find the optimal shape parameters [118].

$$\mathcal{J}(\boldsymbol{b}_s) = \alpha \sum_{m=1}^{M} R_m(x'_{1,m}, x'_{2,m}) - \sum_{j=1}^{J} \frac{b_j^2}{\lambda_j}, \tag{3.4}$$

where $[x'_{1,m}, x'_{2,m}]^T$ is the current estimation of landmark point m, R_m is the response image for point m, J is the total number of shape parameters, and λ_j are the corresponding eigenvalues of the shape model. The algorithm iterates until convergence happens.

Lindner et al. [121] applied CLMs in the setting of femur segmentation. However, instead of computing the correlation with a template, random forest voting was deployed to generate the response images where the decision trees voted for the required displacements. To initialise the landmark points, a Hough-like approach was utilised to automatically detect the femur in the scan [123]. Here, I deployed *Bone Finder* [124], a software implementation provided by Lindner et al. [121], to segment the femoral scans using the CLM approach. All parameters were set as explained in [121].

(B) Template Derivation

Assume $\mathcal{S} = \{\boldsymbol{S}_n; n = 1, \cdots N\}$ denotes N shapes where

$$\boldsymbol{S}_n = \begin{bmatrix} x_{1,1} & , \cdots , & x_{1,M} \\ x_{2,1} & & x_{2,M} \end{bmatrix}_{2\times m}, \tag{3.5}$$

represents the set of M landmark point coordinates for the subject n. General Procrustes analysis is adopted to find the reference template $\boldsymbol{T}$ [56]. First, all scans are aligned to a common position, scale, and orientation. Next, the reference template is updated as the average of the aligned shapes. The algorithm iterates between these two steps until convergence as detailed below.

Let $\mathcal{T}_n(\boldsymbol{S}_n) = k_n\boldsymbol{R}_n\boldsymbol{S}_n + \boldsymbol{c}_n$ denote the geometric transformation that aligns the subject n to the template, where the scalar k is the scaling factor, $\boldsymbol{R}$ is the rotation matrix, and the vector $\boldsymbol{c}$ represents the translation. The objective is to find the reference template $\boldsymbol{T}$ such that

$$\mathcal{J}(\mathcal{S}) = \sum_{n=1}^{N} \|\boldsymbol{T} - \mathcal{T}_n(\boldsymbol{S}_n)\|_F^2, \tag{3.6}$$

is minimised (Algorithm 1). $\|\boldsymbol{S}\|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{S}$ defined as the square root of the sum of the absolute squares of its elements.

Let $\boldsymbol{T}_i$ be the estimated template for the iteration i. Then, a closed-form solution exists to map each shape $\boldsymbol{S}_n$ to the template $\boldsymbol{T}_i$ such that $\|\boldsymbol{T}_i - \mathcal{T}_n(\boldsymbol{S}_n)\|_F^2$ is minimised (the first step).

$$\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V} \leftarrow \text{SVD transform of } \widetilde{\boldsymbol{T}}_i \widetilde{\boldsymbol{S}}_n^T \tag{3.7}$$

$$\boldsymbol{R}_n = \boldsymbol{U}\boldsymbol{V}^T \tag{3.8}$$

$$k_n = \frac{tr(\widetilde{\boldsymbol{S}}_n^T \boldsymbol{R}_n^T \widetilde{\boldsymbol{T}}_i)}{tr(\widetilde{\boldsymbol{S}}_n^T \widetilde{\boldsymbol{S}}_n)} \tag{3.9}$$

$$\boldsymbol{c}_n = \overline{\boldsymbol{T}}_i - k_n \boldsymbol{R}_n \overline{\boldsymbol{S}}_n \tag{3.10}$$

$\overline{\boldsymbol{S}}_n$ and $\overline{\boldsymbol{T}}_i$ represent the column-wise average of matrices $\boldsymbol{S}_n$ and $\boldsymbol{T}_i$, respectively. $\widetilde{\boldsymbol{S}}_n = \boldsymbol{S}_n - \overline{\boldsymbol{S}}_n$ and $\widetilde{\boldsymbol{T}}_i = \boldsymbol{T}_i - \overline{\boldsymbol{T}}_i$. $tr(.)$ denotes the trace of a matrix defined as the sum of its diagonal elements. Given the geometrical transformations $\mathcal{T}_n(\boldsymbol{S}_n)$, the template is updated to the average of the transformed shapes (the second step).

$$\boldsymbol{T}_{i+1} = \frac{1}{N} \sum_{i=1}^{N} k_n \boldsymbol{R}_n \boldsymbol{S}_n + \boldsymbol{c}_n \tag{3.11}$$

It can be shown that generalised Procrustes analysis converges to a unique solution except for a scaling, rotation, or translation factor. To cancel out the arbitrary scaling of the template, the converged template was normalised with the scale $k = \left[\frac{1}{N}\left(k_1^* + \cdots + k_N^*\right)\right]^{-1}$ where k_n^* is the final scale factor after convergence. To cancel out the arbitrary rotation of the template, the template was rotated such that the bottom cross-section at the femoral shaft is parallel to the horizontal axis. To cancel out the arbitrary translation, the centre of gravity, i.e. the average of all landmark points on the template, is shifted to the origin at the $[0, 0]^T$ coordinate.

(C) Image Registration

To eliminate morphological variation between scans, each individual scan is warped to the template domain using a thin plate spline (TPS) registration technique [58]. In this technique, a geometrical transformation is found such that the landmark points in the source image are exactly mapped to their corresponding landmark points in the reference template. To do this mapping,

Algorithm 1 General Procrustes Analysis

```
Input: $\{\boldsymbol{S}_n\}_{n=1}^{N}$
Parameter: $\epsilon$
Output: $\boldsymbol{T}$
procedure GeneralProcrustes
    $i \leftarrow 0$
    $\boldsymbol{T}_i \leftarrow \boldsymbol{S}_n$                                        ▷ for a random $n$
    while $\|\boldsymbol{T}_{i-1} - \boldsymbol{T}_i\|_F^2 \leq \epsilon$ do
        for $n = 1 : N$ do
            $\mathcal{T}_n \leftarrow \text{argmin}_{\mathcal{T}} \|\boldsymbol{T}_i - \mathcal{T}(\boldsymbol{S}_n)\|_F^2$ (Eq. 3.7)
        $i \leftarrow i + 1$
        $\boldsymbol{T}_{i+1} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{T}_n(\boldsymbol{S}_n)$
```

the transformation function involves two components: an affine transformation to compensate for the global scale, translation, and rotation variation; and a radial basis function, i.e. $g(r) = r^2 \log r^2$, to compensate for the local variation around each control point.

$$y_d = f_d(x_1, x_2) = \underbrace{a_{d,0} + a_{d,1}x_1 + a_{d,2}x_2}_{\text{Affine Transformation}} + \sum_{m=1}^{M} \omega_{d,m} g\left(\sqrt{(x_1 - x_{1,m})^2 + (x_2 - x_{2,m})^2}\right), \tag{3.12}$$

where $d = 1, 2$ represents the horizontal and vertical axes in a 2D image. Note that $[y_1, y_2]^T$ and $[x_1, x_2]^T$ denote the coordinate space in the template and image domains, respectively. $y_{d,m}$ and $x_{d,m}$ also denote the coordinates of the m^{th} landmark point in the template and image domains, respectively.

Given the transformation function $f_d(x_1, x_2)$ (Eq. 3.12), the total number of parameters is $2(M + 3)$. For an exact solution, $f_d(x_1, x_2)$ should map each landmark point in the image domain to the corresponding landmark point in the template domain:

$$y_{d,m} = f_d(x_{1,m}, x_{2,m}) = a_{d,0} + a_{d,1}x_{1,m} + a_{d,2}x_{2,m} + \sum_{m'=1}^{M} \omega_{d,m'} g_{m,m'}, \tag{3.13}$$

where

$$g_{m,m'} = g\left(\sqrt{(x_{1,m} - x_{1,m'})^2 + (x_{2,m} - x_{2,m'})^2}\right). \tag{3.14}$$

Eq. 3.13 provides $2M$ constraints. The other 6 constraints are suggested by

Bookstein [58] as follows:

$$\sum_{m=1}^{M} \omega_{1,m} = \sum_{m=1}^{M} \omega_{2,m} = 0 \tag{3.15}$$

$$\sum_{m=1}^{M} x_{1,m}\omega_{1,m} = \sum_{m=1}^{M} y_{1,m}\omega_{1,m} = 0 \tag{3.16}$$

$$\sum_{m=1}^{M} x_{2,m}\omega_{2,m} = \sum_{m=1}^{M} y_{2,m}\omega_{2,m} = 0 \tag{3.17}$$

Using Eq. 3.13 and Eqs. 3.15-3.16, I have $2(M+3)$ linear equations that can be solved for the computation of the parameters in the model. Given the transformation functions, the whole image space is warped to the template domain and the intensity values are interpolated using a linear interpolation.

Note that the linear interpolation preserves the average BMD measured at conventional ROIs after warping each scan to the template. An alternative would be to preserve the average bone mineral content, i.e. BMD multiplied by the area, by scaling the pixel BMD values in proportion to the areal size of each scan with respect to the template. This property is not of interest in this study and so no calibration for bone size is made here.

3.2.3 Comparative Calibration

Assume C systems each used to measure the same characteristics on a common set of N subjects. Each system may not be consistent in the repeated measurements of the same patient resulting in a *within-patient* sampling variation. However, I assume this variation is consistent for different patients. Ignoring this sampling fluctuation, I refer to the mean of repeated measurements as *true* values. Note that the true underlying values are not directly observable. Furthermore, I assume that a linear relationship exists between each pair of systems given the true underlying measurements. Then, *comparative calibration* refers to the problem of simultaneous estimation of the pairwise relationships between these systems [125, 115].

Let the latent random variable X represent the underlying true value and the random variable Y^c represents the observed value measured on the machine c. Barnett [125] proposed a linear model for comparative calibration between the systems.

$$Y^{(c)} = a_c X + b_c + E^{(c)}, \qquad \text{for } c = 1, \cdots, C. \tag{3.18}$$

$E^{(c)} \sim \mathcal{N}(0, \sigma_c^2)$ represents the measurement noise for each system and $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ represents the distribution of the population. Given the observed measurements $\boldsymbol{y}_n = \left[y_n^1, \cdots, y_n^C\right]^T$ for the subject n, the objective is to estimate the model parameters $\{a_c, b_c, \sigma_c\}_{c=1}^C$. This model is overparametrised and to resolve this identifiability problem, it is common to take one system, e.g. c^*, as the reference. For this system, then, it is assumed that $a_{c^*} = 1$ and $b_{c^*} = 0$ [125]. Alternatively, Lu et al. [115] added two extra linear equations:

$$\frac{1}{C}\sum_c b_c = B_0 \qquad \text{and} \qquad \frac{1}{C}\sum_c a_c = A_0, \tag{3.19}$$

where B_0 and A_0 are two constants defined based on either hypothetical assumptions or phantom measurements. Barnett [125] presented the solution for $C = 3$ using the second order moment estimates. Lu et al. [115] presented an expectation maximisation (EM) approach to estimate the model parameters for $C > 3$. For $C = 2$, the problem is known as *Deming Regression* problem (section 3.2.3.(B)).

Here, I deployed this method in the settings of two calibration problems: cross-calibration between DXA manufacturers (section 3.2.3.(C)) and calibration between the left and the right hips (section 3.2.3.(D)). In these settings, it is common not to access multiple measurements of one subject on all different systems. In the extreme case, only one sample measurement is available for each subject. In these scenarios, I propose a new technique based on pairwise quantile matching between different systems (section 3.2.3.(A)). I will show that this technique provides a reliable alternative for parameter estimation when multiple measurements are not accessible.

(A) Quantile Matching Technique

I propose a novel quantile matching technique for the comparative calibration problem (Eq. 3.18) when only one single measurement is available for each subject (cf. [115] and [125]). The new technique is developed based on two assumptions: First, a unique distribution of the latent variable X exists independent of the measurement systems. Second, the SNR is sufficiently large such that

$$Q_{Y^{(c)}}(u) \approx a_c Q_X(u) + b_c, \tag{3.20}$$

where $Q_X(u)$ and $Q_{Y^{(c)}}(u)$ denote the quantile functions. For a random variable X, the quantile function $u \to Q_X(u)$ is defined as

$$Q_X(u) := \inf \{x : u \leq \mathcal{P}(X \leq x)\} . \tag{3.21}$$

Note that quantiles are invariant to monotone transformations; if ψ is a monotone function, then

$$Q_{\psi(X)}(u) = \psi(Q_X(u)). \tag{3.22}$$

Therefore, if the noise power is zero, then the approximation would be replaced with equality in Eq. 3.20. With this assumption, estimation of the model parameters $\Theta = \{a_c, b_c\}$ can be decoupled from the estimation of noise variances, i.e. $\{\sigma_c^2\}$. This technique cannot estimate the noise variances because of insufficient statistics due to missing multiple measurements. However, this technique can provide reliable estimations for the slope a_c and intercepts b_c as detailed below.

The parameters Θ are estimated by minimising the cost function

$$\mathcal{J} = \frac{1}{2} \sum_{c=1}^{C} \int_0^1 (Q_{Y^{(c)}}(u) - a_c Q_X(u) - b_c)^2 du, \tag{3.23}$$

$$\text{subject to} \quad \sum_c b_c = 0 \quad \text{and} \quad \frac{1}{C} \sum_c a_c = 1. \tag{3.24}$$

To set the constants in Eq. 3.19, I assume that the true value X equals the average of the expected observations given the latent variable X, i.e. $X = \frac{1}{C} \sum_c E(Y^{(c)}|X)$. This results in $B_0 = 0$ and $A_0 = 1$ (Algorithm 2).

Optimisation: To convert the constrained optimisation problem into an unconstrained one, I can simply express the parameters a_C and b_C based on the other parameters:

$$a_C = C - \sum_{c \neq C} a_c \quad \text{and} \quad b_C = - \sum_{c \neq C} b_c \tag{3.25}$$

To estimate the parameters, an alternating minimisation technique is adopted: Given the model parameters, the latent variable x_n for each of N subjects can be estimated as (step 1),

$$x_n = E(X|y_n^{(c_n)}; a_{c_n}, b_{c_n}) \approx \frac{1}{a_{c_n}} (y_n^{(c_n)} - b_{c_n}), \tag{3.26}$$

where c_n is the corresponding system for subject n. To update the model parameters, the gradients $\frac{\partial}{\partial a_c}\mathcal{J}$ and $\frac{\partial}{\partial b_c}\mathcal{J}$ are set to zero.

$$\begin{aligned}\frac{\partial}{\partial a_c}\mathcal{J} = (a_c + \sum_{c' \neq C} a_{c'} - C)\int_0^1 Q_X(u)^2 du + (b_c + \sum_{c' \neq C} b_{c'})\int_0^1 Q_X(u)du \\ + \int_0^1 Q_X(u)(Q_{Y^{(C)}}(u) - Q_{Y^{(c)}}(u))du = 0, \qquad (3.27)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial b_c}\mathcal{J} = (a_c + \sum_{c' \neq C} a_{c'} - C)\int_0^1 Q_X(u)du + (b_c + \sum_{c' \neq C} b_{c'}) \\ + \int_0^1 (Q_{Y^{(C)}}(u) - Q_{Y^{(c)}}(u))du = 0. \qquad (3.28)\end{aligned}$$

Computing $Q_X(u)$ using the estimated latent variables, Eq. 3.27 and Eq. 3.28 are linear with respect to the model parameters. Therefore, I have $2(C-1)$ linear equations with $2(C-1)$ parameters for which a closed-form solution exists (step 2). The algorithm iterates between these two steps until the root mean square of the difference between estimated parameters at two consecutive iterations is less than a user-defined tolerance ϵ.

Algorithm 2 Quantile Matching Technique for Comparative Calibration

```
Input:  $\mathcal{Y} = \{y_n^{(c_n)}\}_{n=1}^N$
Parameters: $\epsilon$                                   ▷ Convergence tolerance
Output: $\Theta = \{a_c, b_c\}_{c=1}^C$
procedure QUANTILE-MATCHING($\mathcal{Y}$)
    for $c = 1 : C$ do
        $Q_{Y^{(c)}}(u) \leftarrow$ Estimate quantile values for $\mathcal{Y}^c = \{y_n^{(c_n)} : c_n = c\}$
    $i \leftarrow 1$                                       ▷ Number of iterations
    $a_c \leftarrow 1$ and $b_c \leftarrow 0$    for $c = 1, \cdots, C$.
    while $\sum_{c=1}^C (a_c^{(i)} - a_c^{(i-1)})^2 + (b_c^{(i)} - b_c^{(i-1)})^2 \leq \epsilon^2$ do
        $i \leftarrow i + 1$
        for $n = 1 : N$ do
            $x_n \leftarrow E(X|y_n^{(c_n)}; \Theta_i)$ (Eq. 3.26)
        $Q_X(u) \leftarrow$ Estimate quantile values for $\mathcal{X} = \{x_1, \cdots, x_N\}$  ▷ (step 1)
        $a_c^{(i)}, b_c^{(i)} \leftarrow \underset{a_c, b_c}{\text{argmin}} \frac{1}{2}\sum_{c=1}^C \int_0^1 (Q_{Y^{(c)}}(u) - a_c Q_X(u) - b_c)^2 du$  ▷ (step 2)
```

(B) Deming Regression Problem

If $C = 2$, the comparative calibration formulation (Eq. 3.18) reduces to a single regression problem. Assuming the first system as the reference, i.e. $a_1 = 1$

and $b_1 = 0$, I have six parameters to be estimated: a_2, b_2, σ_1, σ_2, μ, and σ_x. Maximum likelihood estimation of the parameters results in minimising the weighted sum of squared residuals (SSR) of the model [126]:

$$SSR = \sum_{n=1}^{N} \frac{\epsilon_n^{(1)2}}{\sigma_1^2} + \frac{\epsilon_n^{(2)2}}{\sigma_2^2} = \sum_{n=1}^{N} \frac{1}{\sigma_1^2}(y_n^1 - x_n)^2 + \frac{1}{\sigma_2^2}(y_n^2 - a_2 x_n - b_2)^2, \tag{3.29}$$
$$\text{subject to} \qquad \sigma_1^2 \leq S_{Y^{(1)},Y^{(1)}} \quad \text{and} \quad \sigma_2^2 \leq S_{Y^{(2)},Y^{(2)}}.$$

Minimising the SSR (Eq. 3.29) is indefinite as the number of sufficient statistics is 5 and one extra constraint is required. The most common constraint is to fix the variance ratio $\delta = \frac{\sigma_2^2}{\sigma_1^2}$. Then, the intercept b_2 and the slope a_2 can be estimated using the second order moments:

$$a_2 = \begin{cases} \frac{(S_{Y^{(2)},Y^{(2)}} - \delta S_{Y^{(1)},Y^{(1)}}) + \sqrt{(S_{y^{(2)},Y^{(2)}} - \delta S_{Y^{(1)},Y^{(1)}})^2 + 4\delta S^2_{Y^{(1)},y^{(2)}}}}{2S_{Y^{(1)},y^{(2)}}}, & \text{if } S_{Y^{(1)},Y^{(2)}} \geq 0, \\ \frac{(S_{Y^{(2)},Y^{(2)}} - \delta S_{Y^{(1)},Y^{(1)}}) - \sqrt{(S_{Y^{(2)},Y^{(2)}} - \delta S_{Y^{(1)},Y^{(1)}})^2 + 4\delta S^2_{Y^{(1)},Y^{(2)}}}}{2S_{Y^{(1)},Y^{(2)}}}, & \text{if } S_{Y^{(1)},Y^{(2)}} < 0, \end{cases} \tag{3.30}$$

$$b_2 = \overline{Y}^{(2)} - a_2 \overline{Y}^{(1)}, \tag{3.31}$$

where

$$\overline{Y}^{(1)} = \frac{1}{N} \sum_{n=1}^{N} y_n^{(1)} \tag{3.32}$$

$$\overline{Y}^{(2)} = \frac{1}{N} \sum_{n=1}^{N} y_n^{(2)} \tag{3.33}$$

$$S_{Y^{(1)},Y^{(1)}} = \frac{1}{N-1} \sum_{n=1}^{N} (y_n^{(1)} - \overline{Y}^{(1)})^2 \tag{3.34}$$

$$S_{Y^{(1)},Y^{(2)}} = \frac{1}{N-1} \sum_{n=1}^{N} (y_n^{(1)} - \overline{Y}^{(1)})(y_n^{(2)} - \overline{Y}^{(2)}) \tag{3.35}$$

$$S_{Y^{(2)},Y^{(2)}} = \frac{1}{N-1} \sum_{n=1}^{N} (y_n^{(2)} - \overline{Y}^{(2)})^2. \tag{3.36}$$

Note that the Deming regression problem is different from the simple linear regression. When the ratio of the standard deviation of the measurement error to the standard deviation of the population exceeds 0.2, i.e. $\frac{\sigma_1}{\sigma_x} > 0.2$, the simple linear regression results in a significant error in the estimation of parameters and the Deming regression should be adopted instead [127].

(C) Comparative Calibration Between DXA Manufacturers

Systematic differences in BMD measurements exist between DXA manufacturers [113, 114, 115]. Discussing the biological or technical reasons for this discrepancy is not the purpose of this study, but to provide a universal standardisation of BMD values. The first attempt at cross-calibration between DXA scanners, sponsored by IDSC, showed that measurements across different machines are highly correlated [113]. Later, Lu et al. [115] formulated this as a comparative calibration problem and proposed a fully statistical framework for cross-calibration between DXA manufacturers. This method cannot be used if any given subject is scanned only once on each machine. Requiring multiple measurements of each subject across all machines is an implausible assumption in large-scale multi-centre studies. Alternatively, calibration against phantom measurements is a common pragmatic approach [117]. However, using human measurements is preferred for calibration purposes as a significant disagreement exists between the model parameters fitted to the phantom measurements and those fitted to the human measurements [113].

In this study, I used the proposed quantile matching regression technique (section 3.2.3.(A)) to address missing multiple scans. To deploy this calibration technique, one should ensure that the population distribution of BMD values is identical on different machines (cf. assumption 1 in section 3.2.3.(A)). To this end, I selected a prospective cohort from those scanned on each machine such that they were matched for gender, age, body mass index (BMI), scan side, ethnicity, and the geographical location.

(D) Bilateral Calibration

Scanning only one side (left or right) has become the standard procedure in bone densitometry [128]. Good correlation between BMD of the left and the right hip subregions and small absolute differences reported in the literature may have reinforced unilateral hip measurements [129, 130]. However, to amalgamate data from both sides to construct the atlas, a calibration procedure is still required as a statistically significant difference in BMD exists between the bilateral hip measurements [128, 131]. Here, given access to $n = 6916$ bilateral hip measurements, I applied the Deming regression to estimate the calibration parameters.

3.2.4 Quantile Regression

Regression analysis allows statistical modelling of the relationship between a response variable Y and a set of explanatory covariates X [132]. The ordinary least squares regression can capture how the mean of Y changes with X, i.e. estimates $E(Y|X)$; however, this method does not provide a complete picture by considering the conditional distribution of Y given X, i.e. $\mathcal{P}(y|x)$. Quantile regression is a type of regression analysis where conditional quantiles of the response variable are estimated. These quantile curves show the distribution of Y as it changes according to the given covariates and so no information is lost. Here, the objective was to model the evolution of pixel BMD values as a function of age using quantile regression.

(A) Notation and Background

Assume the real-valued random variable Y with *cumulative distribution function* (CDF) $F_Y(y) = P(Y \leq y)$ represents a response variable of interest, e.g. BMD values at a single pixel coordinate, and the multivariate random variable $\boldsymbol{X} = [X_1, \cdots, X_p]^T$ represents an explanatory covariate vector, e.g. age, BMI, etc. Then, the conditional quantile function $(u, \boldsymbol{x}) \longmapsto Q_{Y|\boldsymbol{X}=\boldsymbol{x}}(u, \boldsymbol{x})$ is defined as

$$Q_{Y|\boldsymbol{X}}(u, \boldsymbol{x}) := \inf \left\{ y : u \leq F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) \right\}, \tag{3.37}$$

where $0 < u < 1$. The main objective is to estimate $Q_{Y|\boldsymbol{X}}(u, \boldsymbol{x})$ from N observed scattered points $(y_n, \boldsymbol{x}_n)$.

(B) Classical Quantile Regression

Formulating conditional quantile functions in terms of a regression problem was introduced by Koenker and Bassett [133]. The u^{th} quantile of the random variable Y can be found by minimising the $E(\rho_u(Y - \xi))$ with respect to ξ, where

$$\rho_u(x) = x(u - \mathbf{1}(x < 0)), \tag{3.38}$$

is known as the *check* function. This is plotted in Fig. 3.6 for $u = 0.5$ and 0.9. Therefore, given the observed samples of Y, the u^{th} quantile can be found by solving

$$Q_Y(u) = \min_{\xi \in \mathbb{R}} \sum_{n=1}^{N} \rho_u(y_n - \xi). \tag{3.39}$$

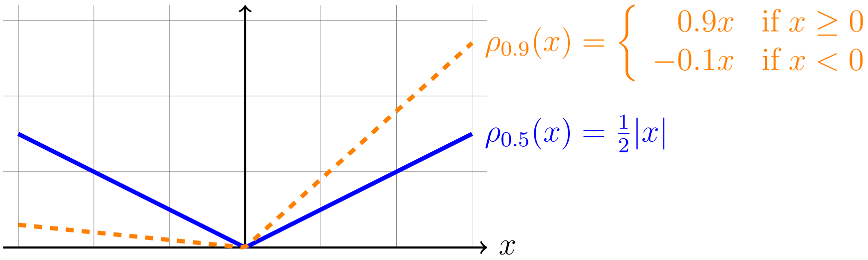


Figure 3.6: The check function for quantile regression with $u = 0.5$ and $u = 0.9$. To compute the u^{th} quantile of the random variable Y, $Q_Y(u)$, the expected loss $E(\rho_u(Y - \xi))$ is minimised with respect to ξ. Observe that $u = 0.5$ is equivalent to the median.

Given the explanatory random variable $\boldsymbol{X}$, the linear conditional quantile function can be estimated as follows:

$$Q_{Y|\boldsymbol{X}}(u, \boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}(u), \tag{3.40}$$

where $\hat{\beta}(u)$ is the solution of the following optimisation problem:

$$\hat{\boldsymbol{\beta}}(u) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{n=1}^{N} \rho_u(y_n - \boldsymbol{x}_n^T \boldsymbol{\beta}). \tag{3.41}$$

Eq. 3.41 is solved by linear programming as details are explained in the work by Koenker and Bassett [133].

(C) The LMS Technique

Classical quantile regression can lead to the quantile crossing problem. One way to avoid this problem is to enforce *commonality* between adjacent quantile curves, i.e. the spacings between quantiles are constrained to be related to each other [134]. To establish commonality, some forms of a probability distribution are assumed for the measurements. Cole and Green [134] assumed an underlying skewed normal distribution so that a suitable Box-Cox transformation (Eq. 3.42) would render a normal distribution. For ease of explanation here, I assumed a scalar covariate denoted by t.

$$Z = \begin{cases} \frac{(\frac{Y}{\mu(t)})^{\lambda(t)} - 1}{\sigma(t)\lambda(t)}, & \lambda(t) \neq 0; \\ \\ \frac{1}{\sigma(t)} \ln(\frac{Y}{\mu(t)}), & \lambda(t) = 0. \end{cases} \tag{3.42}$$

The Box-Cox transformation has three parameters: the power λ, the mean μ, and the coefficient of variation σ. For fitting the quantile curves, the three parameters are first estimated as a smooth function of the covariate t, and then the quantile curves can be simply computed as:

$$Q_{Y|T}(u,t) = \begin{cases} \mu(t)\left[1+\lambda(t)\sigma(t)Q_Z(u)\right]^{1/\lambda(t)}, & \lambda(t) \neq 0; \\ \\ \mu(t)\exp\left(\sigma(t)Q_Z(u)\right), & \lambda(t) = 0. \end{cases} \tag{3.43}$$

The initial letters of the Roman transcriptions of the Greek letters λ, μ, and σ give the name *LMS* to this method. Note that a key assumption of the LMS method is that the Box-Cox transformation with appropriate parameters exists such that the random variable Y can be mapped to a standard normal distribution $Z \sim \mathcal{N}(0,1)$, for which $Y > 0$ is required.

A penalised maximum log-likelihood is presented by Cole and Green [134] to estimate the smooth parameter curves. Given the scattered points for N measurements (y_n, t_n), the log-likelihood function ℓ is given by:

$$\ell = \sum_{n=1}^{N}\left[\lambda(t_n)\ln\frac{y_n}{\mu(t_n)} - \ln\sigma(t_n) - \frac{1}{2}z_n^2\right], \tag{3.44}$$

where z_n is the corresponding mapped score of y_n. The curves $\lambda(t)$, $\mu(t)$, and $\sigma(t)$ are estimated by maximizing the penalized likelihood:

$$\mathcal{J} = \ell - \frac{1}{2}\alpha_\lambda\int\{\lambda''(t)\}^2dt - \frac{1}{2}\alpha_\mu\int\{\mu''(t)\}^2dt - \frac{1}{2}\alpha_\sigma\int\{\sigma''(t)\}^2dt \tag{3.45}$$

The pseudo-code to implement this technique is detailed in [134].

(D) Vector Generalised Additive Models

Different LMS-type techniques could arise depending on the choice of the underlying probability distribution. Yee [135] proposed a unified framework for this class of techniques by formulating the regression problem using vector generalised additive models (VGAMs). Below, the underlying concepts for VGAMs are reviewed briefly and later the application of the R-package `VGAM` for fitting quantile curves using the LMS technique is presented.

To facilitate the explanation of VGAMs, let's start with *vector generalised linear models* (VGLMs). VGLMs are defined as a framework to model the conditional distribution of a response variable Y given the explanatory P-vector $\boldsymbol{X}$ as:

$$\mathcal{P}(y|\boldsymbol{x};\boldsymbol{\beta}_0,\boldsymbol{B}) = h(y,\eta_1,\cdots,\eta_M), \tag{3.46}$$

where $h(.)$ is a known function, coefficients $\boldsymbol{B} = [\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_M] \in \mathbb{R}^{P\times M}$ and the intercepts $\boldsymbol{\beta}_0 = \left[\beta_{(1)0}, \cdots, \beta_{(M)0}\right]^T$ are unknown regression coefficient, and $\boldsymbol{\eta} = [\eta_1, \cdots, \eta_M]^T$ are *linear predictors.* For example, for a normal distribution, $M = 2$ and $\eta_1 = \mu$ and $\eta_2 = \log \sigma$. Note that log(.) is a *link function* to ensure estimated standard deviation is always positive. The m^{th} linear predictor is then estimated as:

$$\begin{aligned} \eta_m = \eta_m(\boldsymbol{x}) &= \beta_{(m)0} + \sum_{p=1}^{P} \beta_{(m)p} x_p, \\ \Rightarrow \boldsymbol{\eta} &= \boldsymbol{\beta}_0 + \boldsymbol{B}^T \boldsymbol{x}. \end{aligned} \tag{3.47}$$

VGAMs are an extension to VGLMs in the sense that predictors η_m are estimated as the sum of smooth functions of the individual covariates x_p rather than being a linear function of the covariates. So, Eq. 3.47 is generalised to

$$\begin{aligned} \eta_m = \eta_m(\boldsymbol{x}) &= \beta_{(m)0} + \sum_{p=1}^{P} f_{(m)p}(x_p), \\ \Rightarrow \boldsymbol{\eta} &= \boldsymbol{\beta}_0 + \sum_{p=1}^{P} \boldsymbol{f}_p(x_p), \end{aligned} \tag{3.48}$$

where $\boldsymbol{f}_p = \left[f_{(1)p}, \cdots, f_{(M)p}\right]^T$. Given the log-likelihood of the parameters ℓ, $\boldsymbol{f}_p(x_p)$ are estimated simultaneously using *vector smoothers* in a VGAM by maximising the penalised likelihood below.

$$\mathcal{J} = \ell - \frac{1}{2} \sum_{p=1}^{P} \sum_{m=1}^{M} \alpha_{(m)p} \int f''_{(m)p}(x_p)^2 dx_p. \tag{3.49}$$

Eq. 3.49 (cf. Eq. 3.45) naturally formulates the LMS quantile regression for $M = 3$.

The smoothness of the fitted parameter curves, e.g. $\boldsymbol{f}(t) = [\lambda(t), \sigma(t), \mu(t)]^T$ in the LMS quantile regression, is controlled using a vector smoothing spline. Assuming a scalar explanatory variable t and letting $t_1 < t_2 < \cdots < t_n$ be the given knots, each parameter curve $f_{(m)}(t)$ is estimated using piece-wise smooth polynomials at each interval $t_n \leq t < t_{n+1}$ as follows:

$$f_{(m)}(t) = a_{(m)}(n) + b_{(m)}(n)(t - t_n) + c_{(m)}(n)(t - t_n)^2 + d_{(m)}(n)(t - t_n)^3. \tag{3.50}$$

Substituting the log-likelihood ℓ (Eq. 3.44) and smooth functions $f_{(m)}(t)$ (Eq. 3.50) into the cost function $\mathcal{J}$ given in Eq. 3.49, one should maximise $\mathcal{J}$ with respect to the parameters $a_n(m)$, $b_n(m)$, $c_n(m)$, and $d_n(m)$.

In this study, I deployed the R-package `VGAM` (version 1.0.3) to fit the quantile curves [135] using the LMS technique. I have modelled the two parameters λ and σ as intercepts. To control the smoothness of the parameter μ, the equivalent degree of freedom (edf) was set to 3.

(E) Numerical Stability and Range-Restriction Problems

The VGAM implementation for the LMS technique has two limitations [135]: first, the optimisation procedure is numerically complex; the second derivatives are approximate and the algorithm fails to converge for a fraction of pixels. Second, $1+\lambda(t)\sigma(t)Q_Z(u) > 0$ is required to compute the quantile values using the Eq. 3.43. Hence, the range of the transformation depends on λ.

To address the first problem, outliers were removed and the LMS technique was applied to the cleaned data. Algorithm 3 shows the outlier removal procedure: given the pixel BMD value y_n and the age a_n, subjects are divided into different sub-groups based on their ages. Next, the first and the third quartiles, denoted by q_1 and q_3, are estimated at each sub-group. Subjects with a pixel BMD value above $q_3 + w_2(q_3 - q_1)$ or below $q_1 - w_1(q_3 - q_1)$ are marked as outliers.

Algorithm 3 Outlier Selection Scheme for the LMS Quantile Regression

```
Input: $\mathcal{D} = \{(y_n, a_n)\}_{n=1}^N$
Parameters: $w_1 = 1.5, w_2 = 2$
Output: $\boldsymbol{o} = [o_1, \cdots, o_n]^T$        ▷ Binary vector; one indicates the outliers
procedure OUTLIER-IDENTIFICATION($\mathcal{D}$)
    $a_{\min} \leftarrow \lfloor \min_n a_n \rfloor$
    $a_{\max} \leftarrow \lceil \max_n a_n \rceil$
    $o_n \leftarrow 0$ for $n = 1, \cdots, N$.
    for $c = a_{\min} : a_{\max}$ do
        $\mathcal{Y} \leftarrow \{y_n : c - 0.5 \leq a_n \leq c + 0.5\}$
        $q_1, q_3 \leftarrow$ estimate the first and the third quartiles of $\mathcal{Y}$
        $\mathcal{J} \leftarrow n : [y_n \in \mathcal{Y}]$ & $[y_n < q_1 - w_1(q_3 - q_1) \,\|\, y_n > q_3 + w_2(q_3 - q_1)]$
        $o_n \leftarrow 1$ for $n \in \mathcal{J}$.
```

Following outlier removal, the LMS technique converged for a majority of pixels (> 99%) in the template. For those pixel coordinates with the range-restriction problem, an offset was added to the pixel BMD values to shift the original BMD range to extremely positive values. The original quantiles were then computed by subtracting the offset term from the estimated quantiles on the shifted data. In this study, I used an offset value of 20 for all the pixels where the LMS technique was crashed.

(F) Confidence Intervals

Assume $\mathcal{Y}(i) = \{y_n(i) : n = 1, \cdots, N\}$ represents the BMD values from N different subjects measured at an individual pixel i. It is of interest to compute a confidence interval for the estimated quantile curves using the observed dataset $\mathcal{Y}(i)$. Here, a bootstrapping procedure was deployed [136]: at each pixel coordinate, N observations are randomly sampled with replacement from $\mathcal{Y}(i)$ to obtain a bootstrap dataset denoted $\mathcal{Y}^*(i)$. Note that each sample from $\mathcal{Y}(i)$ may contribute more than once in the set $\mathcal{Y}^*(i)$ as selected samples are replacing in the original set $\mathcal{Y}(i)$ during the sampling procedure. Next, quantile curves are re-estimated using the bootstrap dataset $\mathcal{Y}^*(i)$. This procedure was repeated 1000 times collecting a distribution of possible quantile curves. From these observations, the 95% confidence intervals were estimated.

3.3 Results

3.3.1 Datasets

To generate the spatio-temporal bone ageing atlas over the adulthood age range (20-95 years), I integrated data from three North Western European population studies: The UK Biobank [137], the Osteoporosis and Ultrasound Study (OPUS) [138], and the MRC-Hip study [139]. The UK Biobank covers the middle-age range (45-80 years); The OPUS covers the younger age range (20-39 years) and the older age range (55-79 years); and finally, the MRC-Hip covers the elderly population (75-95 years).

(A) The UK Biobank Dataset

UK Biobank is a prospective study with over 500,000 participants recruited in middle-age during 2006-2010 across the UK [137]. UK Biobank aims to provide an extensive source of phenotypic and genotypic information about its participants to facilitate the investigation of a wide range of life-threatening diseases, e.g. heart diseases, stroke, diabetes, arthritis, osteoporosis, etc. As part of data collection within UK Biobank, a multi-organ, multi-modality imaging study aims to acquire and store imaging data from 100,000 participants [140]. Here, I used a cohort of 6,918 white women aged 45-80 years at the time of scan acquisition. DXA scans are available for left and right hips, left and right knees, spine (both lateral and anterior-posterior (AP) views), and the whole body using a Lunar iDXA densitometer.

(B) The OPUS Dataset

The OPUS study was a multi-centre European study [138]. Five centres were involved: Sheffield ($n = 535$), Aberdeen ($n = 161$), Berlin ($n = 189$), Kiel ($n = 399$), and Paris ($n = 468$). All participants were women recruited at two different age segments: 20-39 and 55-79 years of age. Scans were acquired using either a Hologic QDR4500 Acclaim densitometer (Sheffield, Paris, and Kiel) or a Lunar Prodigy scanner (Aberdeen and Berlin). In this study, only scans ($n = 1402$) collected on the Hologic system were used.

(C) The MRC-Hip Dataset

The MRC-Hip study was a randomized pharmaceutical clinical trial to examine the effect of clodronate on the incidence of hip fractures [139]. An elderly population cohort of 5018 White women ($\geq$ 75 years) living in the general community in South Yorkshire and North Derbyshire was included in this study. BMD was measured at the hip using a Hologic QDR4500 Acclaim densitometer.

3.3.2 Precision Analysis

Precision or reproducibility of a quantitative measurement technique describes the ability of that technique to produce consistent results when measuring the same quantity repeatedly. In other words, precision is a description of random errors in the system. In DXA bone densitometry, three sources of error exist [141]: the machine (e.g., the scanner noise), the operator (e.g., patient positioning), and the software (e.g., femur segmentation and deformable image alignment).

To assess the overall precision of the RFA technique, 25 Caucasian women (mean age = 70.1 ± 6.2 years) were scanned on the same day twice with repositioning between the scans. This data had been collected as part of the OPUS study in Sheffield. In conventional DXA analysis, precision is reported as the coefficient of variation (CV), i.e. the root mean square standard deviation divided by the mean of paired measurements, for the selected ROIs [142].

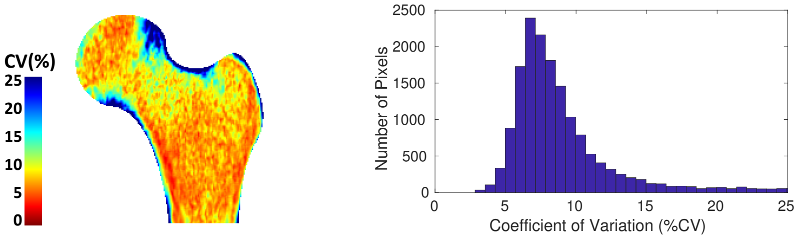
$$CV(\%) = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{(y_n - y'_n)^2}{2}}}{\frac{1}{N}\sum_{n=1}^{N}\frac{(y_n + y'_n)}{2}} \tag{3.51}$$

Here, $N = 25$ is the number of paired measurements; y and y' are the measured BMD values at the two independent positions.

Table 3.1 reports the precision of conventional region-based DXA analysis at four common ROIs. To use RFA to recreate conventional region-based analysis, pixel BMD values of the warped scans were averaged at each ROI in the template domain. RFA resulted in similar precision scores to those reported in the literature at these ROIs (Table 3.1). However, imaging at a finer spatial resolution results in a worse precision at the pixel level in the RFA technique. Fig. 3.7(a) shows the distribution of pixel-level CV values at the proximal femur. Precision was worse around the bone contours. This may be explained due to the inaccuracy in placing controlling landmark points around the bone. Fig. 3.7(b) shows the histogram of pixel-level CV values where the median is 7.96% and the interquartile range is 6.69% − 10.05%. Note that the worse precision in comparison to conventional region-based analysis is a compromise that offers a higher spatial resolution which is necessary for characterising spatially complex bone remodelling events.

Table 3.1: Coefficient of variation (%) at four common conventional ROIs. Top row shows scans measured using DXA RFA with pixels were aggregated to reproduce the conventional ROIs. Lower rows show comparison with published precision data from other investigators.

method	scanner	subjects No.×scans No.	CV%			
			total hip	neck	trochanter	intertrochanter
RFA	Hologic QDR 4500A	25 × 2	1.05	1.73	1.87	1.14
[143]	Hologic QDR 2000	71 × 2	1.2	1.7	1.4	1.7
[144]	Hologic QDR 4500A	27 × 2	1.69	1.11	1.27	-
[145]	Lunar Prodigy	6 × 6	0.65	1.66	1.16	-
[142]	Hologic QDR 2000	95 × 2	1.59	2.25	-	-



(a) Heat-map **(b)** Histogram

Figure 3.7: DXA RFA precision analysis. (a) The pixel level CV (%) is visualised using a heat-map. Precision is worse around the bone contour. This may be due to the inaccuracy in placing controlling landmark points at the bone surface. (b) The distribution of pixel-level CV values in the femur. The median is 7.96% and the interquartile range is $6.69\% - 10.05\%$.

3.3.3 Parameter Estimation for Comparative Calibration between DXA Systems

In this study, scans were collected either on a Hologic QDR 4500A or a Lunar iDXA scanner. To integrate data from both scanners, the proposed quantile matching regression technique is deployed to cross-calibrate the BMD maps between the two systems. For each scanner, $n = 406$ white British women matched for age and BMI with an scan on the left side were selected. No significant difference in age or BMI distribution was observed between the two groups using a two-sample Kolmogorov–Smirnov test (p-value = 0.9). Note that this cohort selection step is a prerequisite for the quantile matching regression technique to ensure that any variation in the BMD distributions between the two groups is only associated with the imaging systems. Fig. 3.8 shows the estimated calibration parameters per each pixel coordinate, i.e. the slope a and the intercept b, taking the Hologic system as the reference. Given these parameters, the calibration parameters for mapping the Lunar system to the Hologic system can be simply estimated as the slope $\frac{1}{a}$ and the intercept $\frac{-b}{a}$.

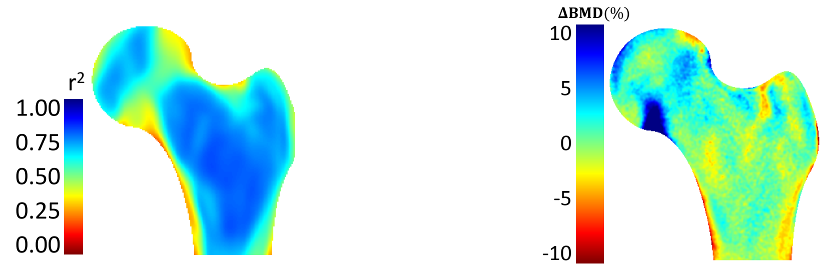
1.4
1.2
1.0
0.8
0.6
0.8
0.6
0.4
0.2
0.0
-0.2

(a) The slope a **(b)** The intercept b

Figure 3.8: Estimated cross-calibration parameters between the Lunar iDXA and the Hologic QDR 4500A systems. The quantile matching regression technique was applied using the Hologic system as the reference, i.e. [Lunar] = a [Hologic] + b. The average and standard deviation of the estimated parameters over all the pixel coordinates within the femur were 1.019 (SD, 0.140) for the slope a and 0.170 (SD, 0.130) for the intercept b.

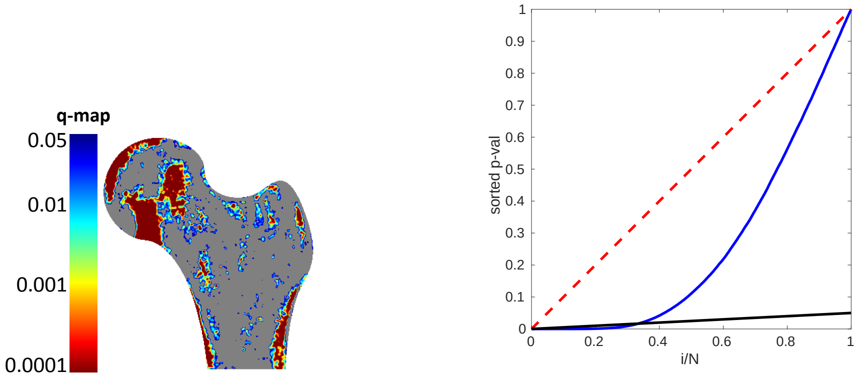
3.3.4 Bilateral Calibration

A total number of 6,916 bilateral hip scans were available in the UK Biobank dataset. Pixel level analysis of BMD values between the right and the left hips confirmed a high correlation between the two sides (Fig. 3.9(a)). However, regions with a statistically significant difference in BMD were observed in the femur (Fig. 3.9(b),(c), and (d)). Given the high correlation between the two sides, it is possible to calibrate the BMD maps for the lateral side. This calibration would enable integration of data from both sides to generate a single bone ageing atlas.



(a) Correlation (r^2)

(b) Normalised right-left difference in the population mean



(c) FDR q-map

(d) PP-plot

Figure 3.9: Bilateral hip comparison. (a) The left and the right hips are highly correlated inside the femur, but the correlation is worse at the boundary. (b) Average right-left differences in pixel BMD values normalised to the population mean for the left side. (c) Localised regions with a statistically significant difference in BMD were observed between the bilateral sides using FDR analysis. (d) The pp-plot deviated from the identity line (dashed red line) demonstrating a significant difference between the bilateral sides.

To test the validity of the estimated calibration parameters, I randomly selected 2000 scans for testing and the remaining scans were used to learn the calibration parameters. The Deming regression technique with $\delta = 1$ was deployed here as both the left and the right scans were available from the same subject. Figs. 3.10(a) and 3.10(b) show the estimated calibration parameters, i.e. the slope a, the intercept b, respectively.

1.2
1.1
1.0
0.9
0.8
0.2
0.1
0.0
-0.1
-0.2

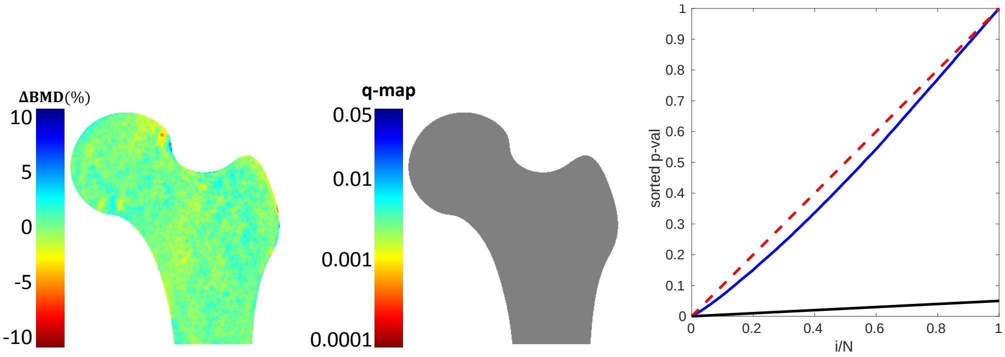
(a) The slope a

(b) The intercept b

Figure 3.10: Estimated cross-calibration parameters between the left and the right hips. The Deming regression technique with $\delta = 1$ was applied taking the right hip as the reference, i.e. [left] = a [right] + b.

Inside the femur, the slope a and the intercept b are approximately one and zero, respectively (see the shade of green in Fig. 3.10). This observation suggests a close relationship between the bilateral sides within the femur and away from the bone contour. At the bone boundary where the correlation between the bilateral sides was low (see Fig. 3.9(a)), the calibration parameters a and b deviated from one and zero, respectively.

To test the validity of the proposed calibration technique, I deployed the estimated calibration parameters to map the pixel BMD measured at the right side to the left side, i.e. computing the expected pixel BMD at the left side given the measurements at the right hip. Note that the test set used in this experiment was not deployed during the estimation of calibration parameters. Fig. 3.11 shows the normalised difference in population mean between the two sides, the corresponding FDR q-map, and the pp-plot for the FDR analysis following calibration for the scan side. No statistically significant difference in BMD was observed between the two sides confirming the validity of the proposed calibration technique.



(a) Normalised difference in BMD **(b)** FDR q-map **(c)** PP-plot

Figure 3.11: The ability of the deployed calibration procedure to cancel the observed difference between the right and the left hips. Here, the right hip is mapped to the left side. Panel (a) shows the average differences in BMD between the left and the calibrated right hips normalised to the population mean for the left side. (b, c) No statistically significant difference in BMD was observed between the two sides following the bilateral calibration using FDR analysis at $q \leq 0.05$.

3.3.5 The Spatio-Temporal Atlas

Fig. 3.12 shows the constructed Atlas; visualising the median, the first and the third quartiles of BMD values at different ages using heat-maps. An overall decline in BMD with increasing age was observed throughout the proximal femur. However, the observed bone loss patterns were site-specific and spatially-complex. Cortical thinning was observed consistently with ageing around the femoral shaft from the 60^{th} decade onwards. A widespread bone loss was also observed in the trochanteric area.

Quantile regression curves demonstrated different rates of bone loss at different anatomic locations within the proximal femur (Figs. 3.13 and 3.14). For example, the decrease in BMD at the superior femoral neck cortex was bimodal; the bone loss slowed down from the 70s onwards (Fig. 3.13(a)). BMD at the mid-femoral neck showed a steady decrease throughout the whole age range (Fig. 3.13(b)), whilst bone mass was preserved the most in the inferior femoral neck cortex (Fig. 3.13(c)). Fig. 3.14 shows quantile regression curves at the intertrochanteric region. Bone mass at the superior trochanteric region was preserved until the just before 70 years, and was followed by a decline with a similar slope to the other trochanteric regions (Fig. 3.14(a)). Bone loss was observed at a consistent rate at the mid trochanteric region throughout the whole age range (Fig. 3.14(b)). BMD in the inferior cortex close to the lesser trochanter was maintained until the age of 60^{th} years, following which point BMD showed a steady decline (Fig. 3.14(c)).

Note that the inflection point observed at age 75 years in Fig. 3.14 is indeed due to ageing rather than the integration of the MRC-Hip dataset (age range: 75-97 years). Repeating the same analysis using only the UK Biobank dataset (age range: 45-80 years) demonstrated similar ageing trends (data not shown). Here, I present the results for the integration of all datasets together.

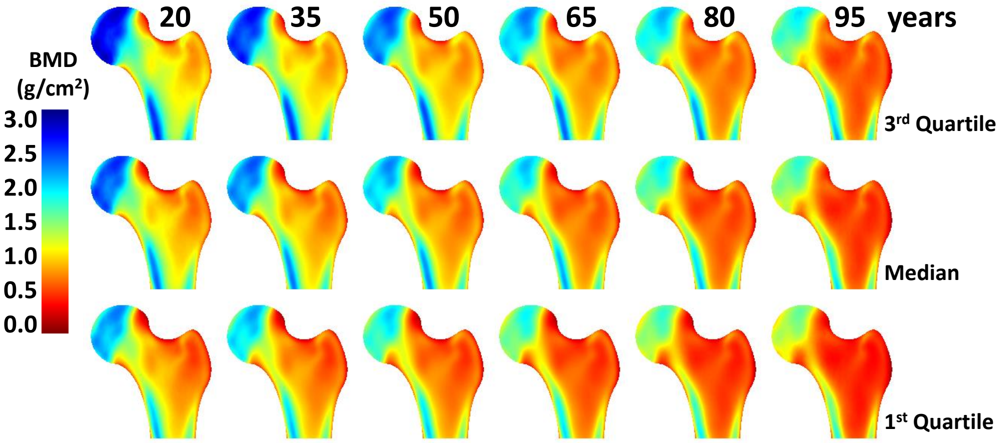
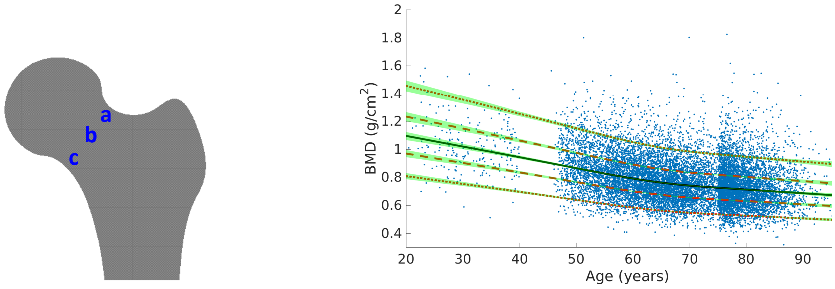
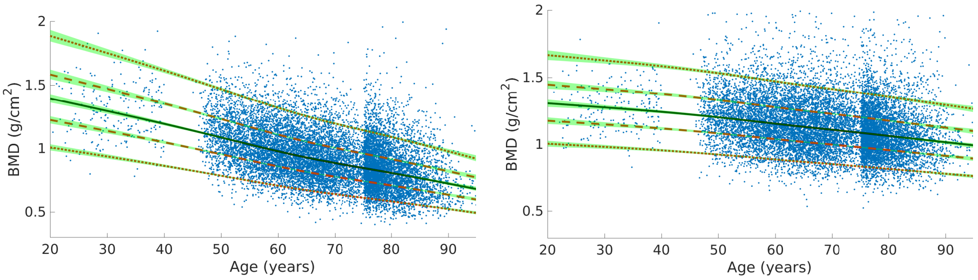


Figure 3.12: The Bone ageing atlas. The median together with the first and the third quartiles at each pixel coordinate is visualised using heat-maps for 20, 35, 50, 65, 80, and 95 years of age. The atlas is shown for the Lunar system at the left hip.

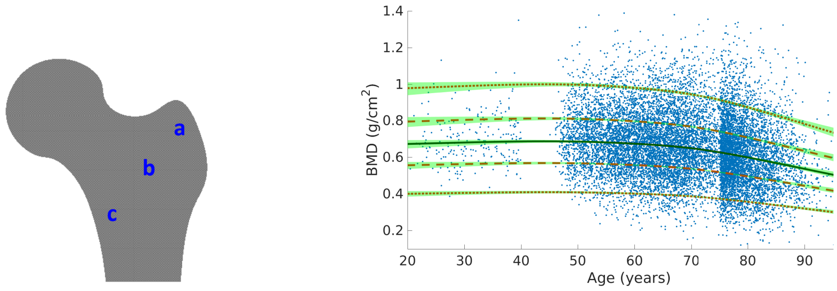


(a) Superior Femoral Neck

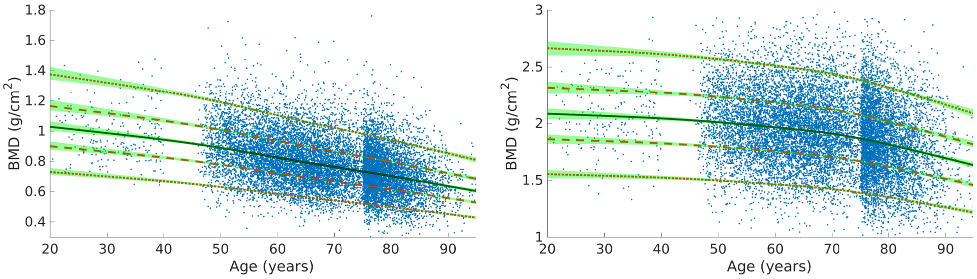


(b) Mid Femoral Neck **(c)** Inferior Femoral Neck

Figure 3.13: Quantile curves fitted at the femoral neck. The solid, dashed, and dotted lines show the median, the 50% and the 90% quantile ranges, respectively. The green shadow shows the 95% confidence interval. The curves are shown for the Lunar system at the left hip.



(a) Superior Trochanteric Region



(b) Mid Trochanteric Region

(c) Inferior Trochanteric Region

Figure 3.14: Quantile curves fitted at the intertrochanteric region. The solid, dashed, and dotted lines show the median, the 50% and the 90% quantile ranges, respectively. The green shadow shows the 95% confidence interval. The curves are shown for the Lunar system at the left hip.

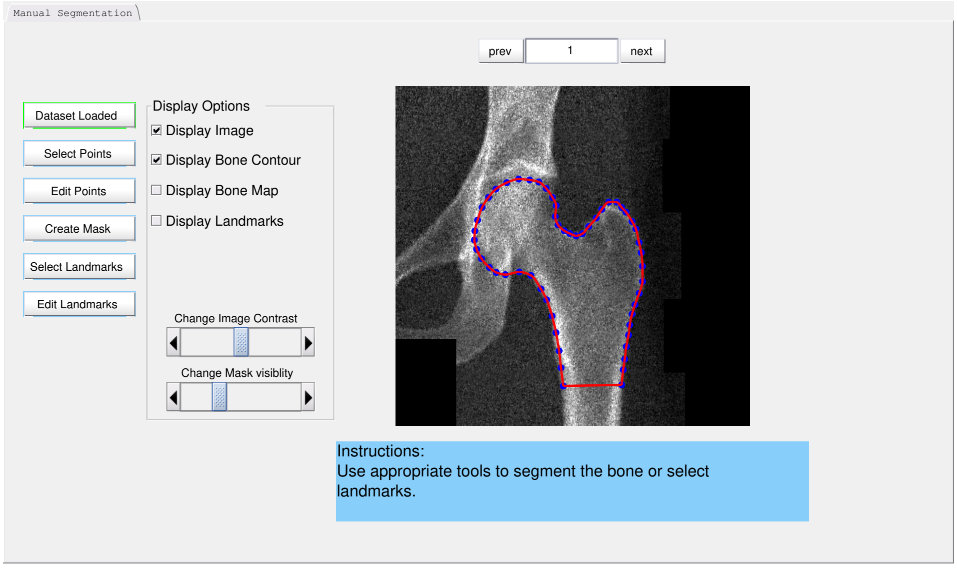
3.4 Validation of the Atlas Construction Steps

3.4.1 Segmentation Accuracy

To evaluate the segmentation accuracy, I manually annotated a subset of scans ($n = 32$) randomly selected from the database; 16 scans (8 from each side) from the Lunar iDXA system and 16 scans (8 from each side) from the Hologic QDR 4500A system. For this purpose, an interactive toolkit was developed in Matlab (Fig. 3.15). The user should select a number of control points around the bone and the software computes a smooth contour passing through the selected points. The toolkit allows the user to move the control points, delete them, or insert new ones if required. The segmentation accuracy was evaluated using the Dice similarity coefficient (DSC). DSC is defined as the twice the areal size of the overlap between two binary masks divided by the sum of the areal size of each mask;

$$\mathrm{DSC} = \frac{2|A \cap M|}{|A| + |M|}, \tag{3.52}$$

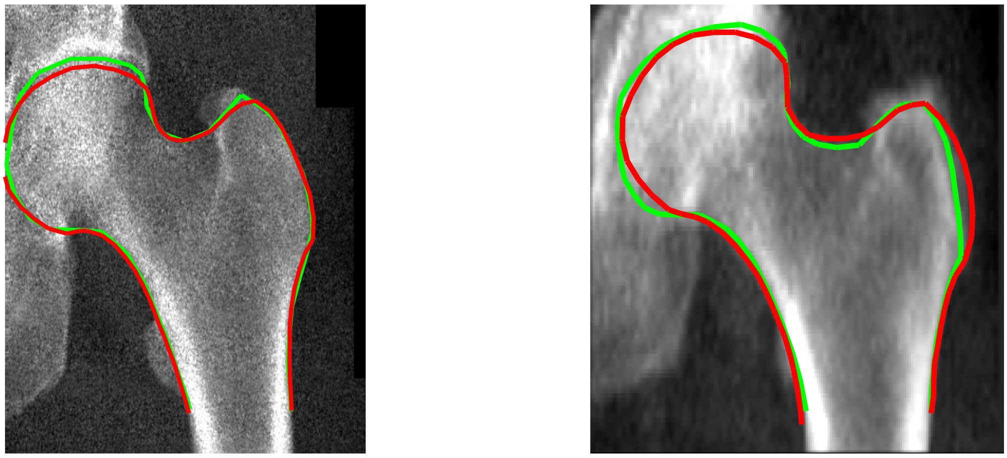
where A and M represents the automatic and the manual segmentation masks, respectively. DSC ranges between 0 and 1 with 1 representing a perfect agreements between the two masks. The mean and the standard deviation for DSC over the 32 selected scans were 0.9698 and 0.0048, respectively. Fig. 3.16 shows the worst and the best segmentation results based on the DSC metric. Observe that since the cut-off point at the femoral shaft is arbitrary, the shorter distal cut-off point between the manual and the automated masks is used to cancel out the variation in the shaft before computing the DSC metric.


Manual Segmentation
prev
1
next
Dataset Loaded
Select Points
Edit Points
Create Mask
Select Landmarks
Edit Landmarks
Change Image Contrast
Change Mask visiblity

Display Options
☑ Display Image
☑ Display Bone Contour
☐ Display Bone Map
☐ Display Landmarks

Instructions:
Use appropriate tools to segment the bone or select landmarks.

Figure 3.15: The graphical user interface (GUI) developed in Matlab to facilitate manual segmentation of femoral scans. The user would select a number of control points around the bone and the software computes a smooth contour passing through the selected points. The toolkit allows the user to move the control points, delete them, or insert new ones if required.



(a) DSC = 0.9801 **(b)** DSC = 0.9620

Figure 3.16: The best and the worst femoral segmentation among 32 randomly selected scans from the database using the dice coefficient index as the evaluation metric. The green and the red contours show the ground truth and the automatic segmentation, respectively.

3.4.2 Point Localisation Accuracy

To evaluate the point localisation accuracy, five landmark points were selected manually at key prominent geometrical locations: centre of the femoral head; the centre, superior, and inferior positions at the femoral neck; and finally the apex at the greater trochanter. To reduce the observer error in placing landmark points, the femoral hip axis is first selected semi-automatically. Next, the femoral head centre and the femoral neck centre are selected on this axis. Then, the user is asked to select the upper and the lower margins on an axis perpendicular to the femoral hip axis passing through the neck centre.

The same dataset ($n = 32$) used for the evaluation of segmentation accuracy were deployed here. For each image, the landmarks are then transferred to the template using the same TPS warping transformation computed per each image (Fig. 3.17).

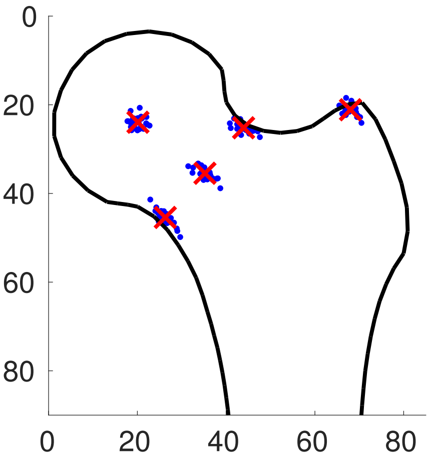


Figure 3.17: Point localisation error. Five landmark points were selected on the template at the centre of the femoral head; the centre, superior, and inferior positions at the femoral neck; and the apex at the greater trochanter (the red cross-marks). To assess the point localisation error, thirty-two scans were randomly selected. For each image, five landmark points were selected manually at anatomically correspondent locations and then mapped to the reference domain using the estimated TPS transformations for each image (the blue dots). The average error was 1.57 mm. The space is shown in millimetre.

Table 3.2 shows the mean and the standard deviation of the distance between each landmark point and its corresponding point on the template. The overall error was 1.57 mm (cf. [146]). Given the space resolution of 0.5 × 0.5 mm in the template domain, the average error was 3.15 pixels.

Table 3.2: Point localisation error at five prominent anatomic locations in millimetre (mm).

	Centre of the Femoral Head	Centre of the Femoral Neck	Superior Cortex of the Femoral Neck	Inferior Cortex of the Femoral Neck	Apex of the Greater Trochanter
mean (SD)	1.64 (0.67)	1.63 (1.10)	1.54 (0.94)	1.67 (1.34)	1.39 (0.82)

3.4.3 Experimental Validation for the Quantile Matching Regression Technique

The proposed quantile matching regression technique provides a means of comparative calibration when multiple measurements of a subject on different systems are not available. I validated this technique using synthetic numerical values at different noise levels. Furthermore, I validated this technique in the setting of bilateral calibration where paired measurements are available.

(A) Synthetic Data

The experimental set-up is as follows: the number of different systems was set to 2, i.e. $C = 2$ in Eq. 3.18. I randomly sampled $N = 5000$ observations for the latent random variable X from a skewed normal distribution; Each observation was sampled from a standard normal distribution $\mathcal{N}(1, 0)$ and then transformed using an inverse Box-Cox transformation (Eq. 3.42) with parameters $\lambda = 0.4$, $\mu = 1.3$, and $\sigma = 0.5$. The nominal values for the model parameters were set to $a_1 = 1$, $b_1 = 0$, $a_2 = 0.8$, and $b_2 = 0.1$. The Gaussian noises $E^{(c)}$ were selected independently from $\mathcal{N}(0, \sigma_c^2)$ for $c = 1, 2.$. I tested the performance of the proposed quantile matching regression technique for various noise levels (Table 3.3). A Monte Carlo procedure with 1000 iterations was conducted and the mean and the standard deviation of the estimated parameters are reported.

Table 3.3: The comparison between the proposed quantile matching regression versus the Deming regression using synthetic samples at different noise levels ($\sigma_2 = \sigma_1$).

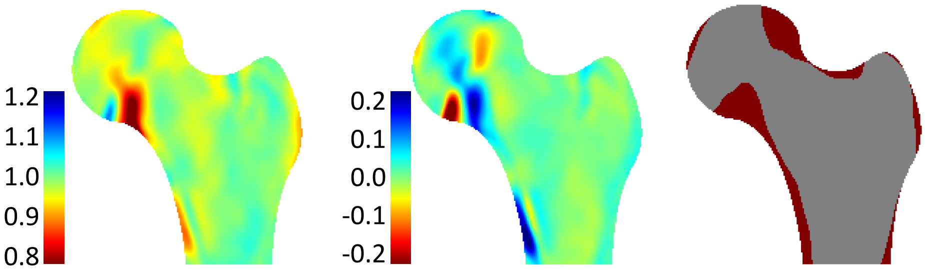
	GT[a]	$\sigma_1 = 0.1$ ($r^2 = 0.94$)		$\sigma_1 = 0.2$ ($r^2 = 0.81$)		$\sigma_1 = 0.4$ ($r^2 = 0.48$)	
		QMR[b]	DR[c]	QMR[b]	DR[c]	QMR[b]	DR[c]
a_2	0.8	0.80(0.003)	0.80(0.003)	0.82(0.006)	0.80(0.005)	0.86(0.011)	0.80(0.012)
b_2	0.1	0.09(0.005)	0.10(0.004)	0.08(0.009)	0.10(0.008)	0.02(0.017)	0.10(0.018)

[a] The Ground Truth, [b] the proposed Quantile Matching Regression technique, and [c] the Deming Regression with $\delta = 1$ [126]. r^2 is the squared correlation value between the two systems. Estimated values for the parameters are reported as mean (standard deviation) of 1000 Monte Carlo repetitions.

Deming regression is a reliable tool for comparative calibration when paired measurements of each subject on both systems are available. When paired measurements are not available, the proposed quantile matching regression resulted in good approximations when the noise level was low ($r^2 = 0.94$) but as the noise power increased, the estimated parameters started to deviate from the ground truth (Table 3.3).

(B) Clinical Data

To validate the performance of the proposed quantile matching regression technique in a clinical setting, I applied this technique to the bilateral calibration problem addressed previously in section 3.3.4 using the Deming regression. Here, I did not use the fact that the left and the right hip measurements are collected from the same subject. Figs. 3.18(a) and 3.18(b) show the slope and the intercept, respectively. The overall pattern is similar to the results for the Deming regression (c.f. Figs. 3.10). However, subtle differences were observed at the bone margin near the contours where the correlation between the left and the right hips were low (Fig. 3.18(c)). In the grey zone with $r^2 \geq 0.5$ (Fig. 3.18(c)), however, the estimated parameters using quantile matching technique perfectly matches the results from the Deming regression with a root mean square (RMS) error of 0.013 and 0.017 for the slope a and the intercept b, respectively.

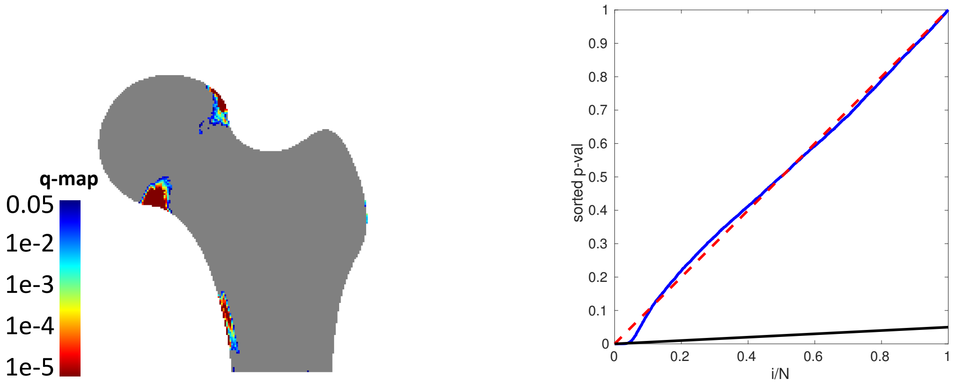


(a) The slope a **(b)** The intercept b **(c)** $r^2 < 0.5$

Figure 3.18: Estimated cross-calibration parameters between the left and the right hips. The quantile matching regression technique was applied taking the right hip as the reference, i.e. [left] = a [right] + b. The estimated parameters are similar to those computed using the Deming regression (cf. Fig. 3.10). Over the region with a high correlation between the left and the right hips ($r^2 \geq 0.5$), the RMS error was 0.013 for the slope a and 0.017 for the intercept b, respectively

3.4.4 Compliance with Normality after the Box-Cox Transformation

The LMS quantile regression technique assumes that a Box-Cox transformation of the response variable with appropriate parameters exists such that the mapped values are normally distributed. Given this assumption, the penalised log-likelihood of the parameters is minimised. However, minimising the cost function given in Eq. 3.45 does not guarantee that the transformed BMD values using the Box-Cox function with the estimated parameters indeed follows a normal distribution. To test this hypothesis, I applied the Kolmogorov–Smirnov test to the transformed pixel BMD values using the learned parameters λ, μ, and σ of the constructed atlas. To account for the multiple comparisons problem, FDR analysis is applied to the computed p-values at each pixel coordinate (Fig. 3.19). The learned LMS models are valid in the majority of pixels except for regions at the rim of the femoral head, and at the bone margin next to the lesser trochanter (Fig. 3.19(a)). Otherwise, as shown in Fig. 3.19(b), the pp-plot (the solid blue line) follows the identity (the red dashed line) confirming the validity of the null hypothesis over a large portion of the proximal femur.



(a) FDR q-map

(b) pp-plot

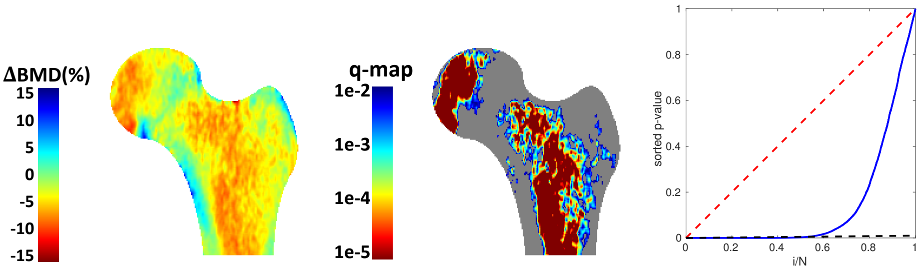
Figure 3.19: FDR analysis to identify pixels where the distribution of the transformed pixel BMD values using the estimated LMS model significantly deviates from a normal distribution. The learned LMS models are valid in the majority of pixels except for regions at the rim of the femoral head, and at the bone margin next to the lesser trochanter.

3.5 Atlas Validation using Longitudinal Data

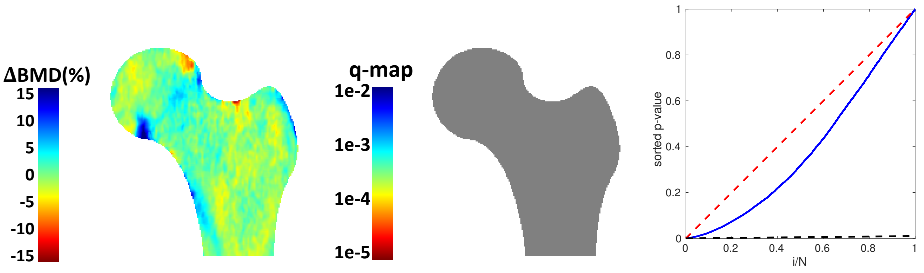
The bone ageing atlas was developed based on cross-sectional data from a large cohort of women ($n = 13,338$). To show that the cross-sectional atlas generation fits with the actual longitudinal BMD change, a subset of scans from the OPUS dataset ($n = 400$; mean age, 64.7 years; range, $55 - 80$ years) is deployed here for which follow-up measurements at 6 years (mean time lapse, 70.9 months; standard deviation, 1.2 months) were available. The hypothesis tested here is that no significant BMD change should be observed between the expected BMD values at 6 years based on the projected BMD atlas and the actual measurements at 6 years. To project the BMD values at six years at each pixel coordinate, firstly the quantile value for the given pixel BMD at the baseline age is read from the atlas. Next, the corresponding BMD value at the follow-up age is read from the same quantile trajectory.

For this analysis, I divided the data into 5 intervals: $55-60$ years ($n = 120$); $60 - 65$ years ($n = 99$); $65 - 70$ years ($n = 82$); $70 - 75$ years ($n = 63$); and $75 - 80$ years ($n = 36$). In each sub-group, a paired t-test proceeded by the FDR analysis was used once between the baseline and the actual follow-up measurements, and another time between the projected and the actual follow-up BMD values. No change would be expected in the latter.

Figs. 3.20-3.24 show the results for each sub-group, respectively. Significant bone loss was observed in the trochanteric region and the medial femoral shaft in all the groups except the last one (Fig. 3.24). This can be explained by the small number of samples in this sub-group ($n = 36$). The projected BMD values using the constructed atlas fits the actual measurements where no significant BMD change was observed between the projected and the actual BMD values.

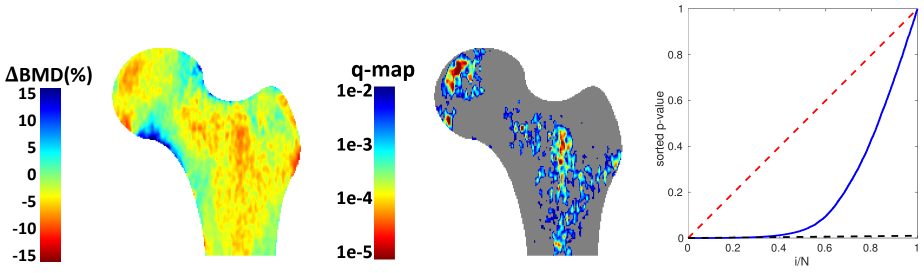


(a) BMD change **(b)** FDR q-map **(c)** PP plot

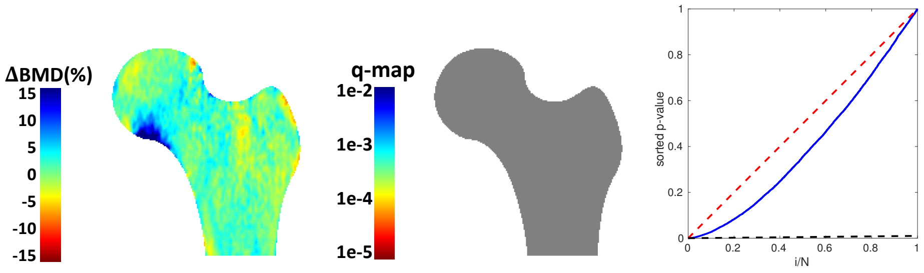


(d) BMD change **(e)** FDR q-map **(f)** PP plot

Figure 3.20: Longitudinal atlas validation (sub-group 1: 55-60 years, $n = 120$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively.

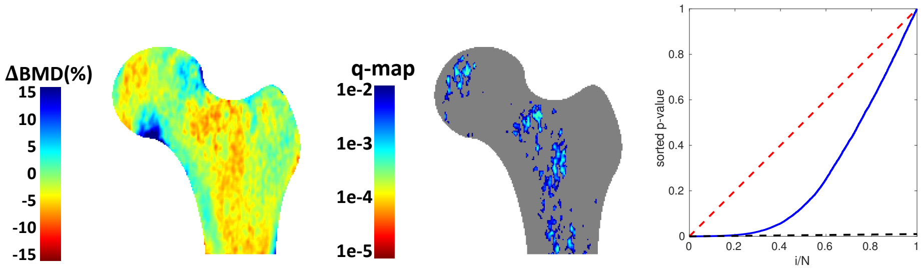


(a) BMD change **(b)** FDR q-map **(c)** PP plot

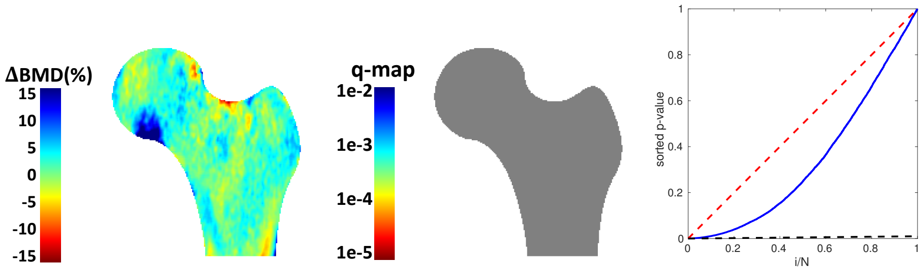


(d) BMD change **(e)** FDR q-map **(f)** PP plot

Figure 3.21: Longitudinal atlas validation (sub-group 2: 60-65 years, $n = 99$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively.

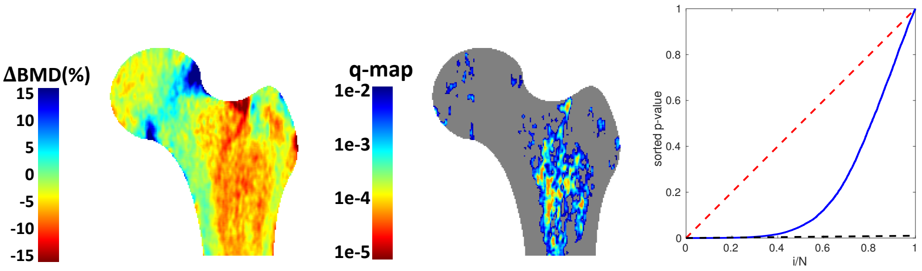


(a) BMD change **(b)** FDR q-map **(c)** PP plot

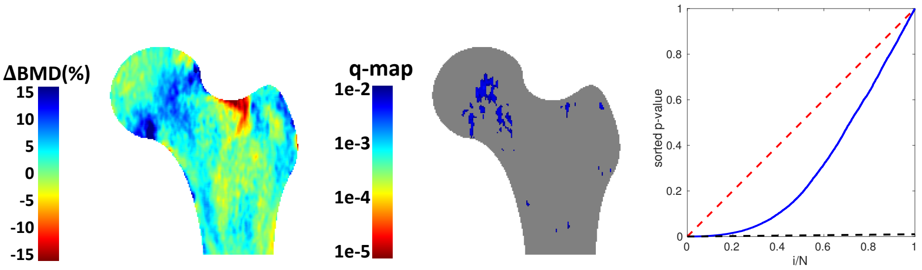


(d) BMD change **(e)** FDR q-map **(f)** PP plot

Figure 3.22: Longitudinal atlas validation (sub-group 3: 65-70 years, $n = 82$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively.

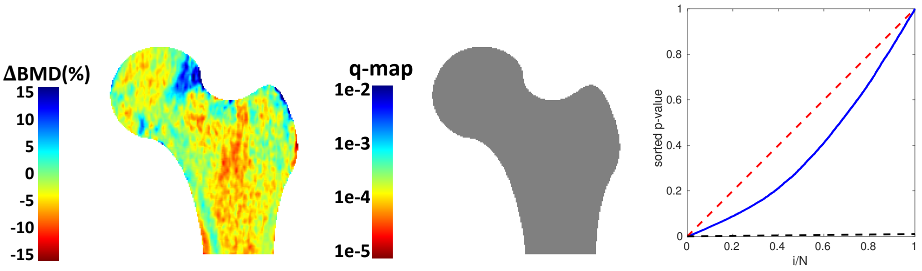

ΔBMD(%)
15
10
5
0
-5
-10
-15
q-map
1e-2
1e-3
1e-4
1e-5
sorted p-value
i/N
0
0.2
0.4
0.6
0.8
1

(a) BMD change **(b)** FDR q-map **(c)** PP plot

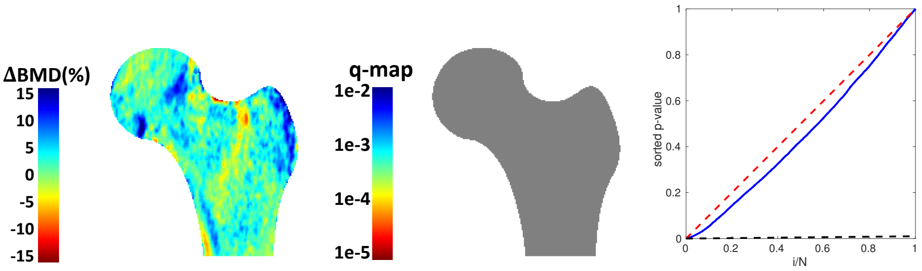

ΔBMD(%)
15
10
5
0
-5
-10
-15
q-map
1e-2
1e-3
1e-4
1e-5
sorted p-value
i/N
0
0.2
0.4
0.6
0.8
1

(d) BMD change **(e)** FDR q-map **(f)** PP plot

Figure 3.23: Longitudinal atlas validation (sub-group 4: 70-75 years, $n = 63$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively.


ΔBMD(%)
15
10
5
0
-5
-10
-15
q-map
1e-2
1e-3
1e-4
1e-5
sorted p-value
i/N
0
0.2
0.4
0.6
0.8
1

(a) BMD change **(b)** FDR q-map **(c)** PP plot


ΔBMD(%)
15
10
5
0
-5
-10
-15
q-map
1e-2
1e-3
1e-4
1e-5
sorted p-value
i/N
0
0.2
0.4
0.6
0.8
1

(d) BMD change **(e)** FDR q-map **(f)** PP plot

Figure 3.24: Longitudinal atlas validation (sub-group 5: 75-80 years, $n = 36$). The top row compares actual baseline and follow-up measurements at 6 years while the bottom row compares the projected BMD values at 6 years versus the actual follow-up measurements. The first column shows the normalised BMD change between the follow-up and either the baseline or the projected values. The second and the third columns show the FDR significance q-map and the corresponding PP plot, respectively.

3.6 Discussion

Osteoporosis is an age-associated disease caused by gradual deterioration of bone tissue with ageing. While ageing is associated with significant bone loss, its effect on bone strength still needs to be better understood [147]. Conventional DXA analysis has provided important insights into the bone loss patterns in different ROIs including the femoral neck; however, its utility is limited by the fact that spatial BMD information is lost by pooling pixels into larger ROIs with a few hundreds of pixels. This data averaging limits our understanding of more focal BMD deficits. To resolve this issue, this work presents the development of a reference spatio-temporal atlas of ageing bone in the proximal femur using cross-sectional data from a large cohort of North Western European Caucasian women (n=13,338). Atlas development is a complex task with a number of sophisticated steps including data organisation, image segmentation and alignment, inter-scanner calibration, and quantile regression analysis. A fully automatic pipeline applicable to large-scale population analysis was proposed to streamline the process.

I validated the methodology for the creation of the atlas using both experimental and synthetic datasets. Each module in the pipeline was evaluated separately. To evaluate the segmentation accuracy, 32 scans were randomly selected and manually annotated. The average segmentation accuracy expressed as the Dice index was 0.97. To assess the overall registration error, five control points were manually selected at the femoral head centre; inferior, mid, and superior femoral neck; and the apex of the greater trochanter. The same 32 scans used for the segmentation evaluation were also deployed here. Warping each individual image to the mean template, the landmark points were mapped onto the template. The mean point to point distance was 1.57 mm equivalent to 3.15 pixels.

The extended RFA framework precision was evaluated using a set of 25 scan pairs, each pair collected on the same day from the same subject with repositioning between scans. All subjects were women (mean age $= 70.1 \pm 6.2$ years) and scanned using a Hologic QDR 4500A densitometer as part of the OPUS study [138]. RFA precision was comparable to conventional DXA analysis when measured on the same ROIs, but worse at the pixel level. However, this compromise offers a higher spatial resolution which is necessary for characterising spatially complex bone remodelling events. The median for the pixel-level coefficient of variation within the femur was 7.96%. The precision was worse at the bone margins ($\geq 15\%$). This lower precision could be explained with landmark localisation error, but this error does not affect the bone distribution patterns within the femur since all landmarks were selected on the bone contour (Fig. 3.7(a)).

The validity of the proposed quantile matching regression technique for comparative calibration was tested using both synthetic and experimental datasets. I tested the viability of the quantile matching regression for bilateral calibration using a subset of scans from the UK Biobank study where both hips were scanned. The estimated calibration parameters were consistent with the results from the Deming regression analysis. However, at regions close to the bone margin where the correlation between the left and the right hips were low, the estimation error increased for the quantile matching regression technique. Since no paired measurements on the Hologic and the Lunar systems were available, I could not test the viability of the proposed technique directly for DXA cross-calibration. However, numerical experiments with synthetic datasets supported the validity of the proposed framework in this context as well.

The precision of the LMS quantile regression for modelling the temporal BMD evolution was tested using a bootstrapping procedure. The overall uncertainty was sufficiently small so the ageing effect was observable (Figs. 3.13 and 3.14). However, the uncertainty was higher for the young age group, i.e. age < 40 years, and the elderly population, i.e. age > 90 years, due to the small number of samples. The validity of the Box-Cox transformation for mapping the skewed pixel BMD distributions to a Normal distribution was tested using a Kolmogorov-Smirnov test proceeded with the FDR analysis to correct for multiple comparisons. Except for three small blobs at the inferior and superior femoral head; and next to the lesser trochanter at the bone margin, the test was not rejected in the majority of pixels in the femur confirming the validity of the LMS technique for this application.

The new technique presented three key contributions. First, RFA allowed high-resolution pixel level BMD analysis. The increased spatial resolution made it possible to observe spatially-complex bone ageing patterns for which conventional region-based bone densitometry routine is insensitive. The validity of the observed ageing patterns was tested using a subset of scans ($n = 400$) with follow-up measurements at 6 years. The data were divided into five groups based on the age at the baseline measurement. No statistically significant difference was observed between the atlas-based projected BMD values and the actual BMD measurements at 6 years using a paired t-test except for group

4 (age, 70-75 years, $n = 63$). Individual analysis of these scans suggested the development of osteoarthritis between the baseline and the follow-up time points for a subset of scans ($n = 5$). Removing these scans and repeating the FDR analysis, no significant difference was observed. Second, the proposed calibration technique allowed the integration of data from different DXA manufacturers. The new method does not require multiple scans from the same subject and so is applicable to large multi-centre studies where every subject is often scanned only on one system. Third, a fully automatic bone ageing analysis pipeline was proposed that would streamline the atlas generation process. This automation would facilitate population-specific atlas generation from other ethnic libraries in the same way that population-specific z scores are computed.

This technique also had limitations. To observe the ageing effect, the variability due to RFA inaccuracy should be smaller than the BMD variation in the population so that its impact on the estimated quantiles is negligible. Pixel-level noise has two effects on the estimated quantile curves: increasing the inter-quartile range, and increasing the confidence interval around each estimated quantile curve. The latter is included in the estimated confidence intervals shown in Figs. 3.13 and 3.14. Since the sample size was sufficiently large, the uncertainty around each curve was sufficiently small and the ageing effect was observable. However, the bias effect leading to an increase in the inter-quartile range cannot be easily observed in these plots; it is difficult to attribute the observed variability to either the population or the measurement noise. However, since noise only affects the inter-quartile ranges and the estimated median curves are unbiased regardless of the noise power, the ageing trend visualised in Fig. 3.12 is still valid in all pixel coordinates.

The areal BMD measured by DXA does not represent the true volumetric BMD, and so the constructed atlas is a 2D projection of the actual 3D patterns. A 2D/3D approach could address this issue [148, 39]. These techniques are often based a 3D statistical shape/appearance model learned from a small subset of QCT images, for example, $n = 57$ (all highly osteoporotic women) [148]. Hence, the learned atlas cannot account for the full population variation (cf. $n = 13,338$ in this study). If a large QCT dataset was available, the ageing atlas could have been directly developed from them where the principle applied here can be readily transferred to 3D imaging.

This technique shows promise in characterising spatially-complex BMD changes with ageing. These patterns were visualised using heat-maps. Furthermore, quantile curves plotted at different pixel coordinates showed consistently different rates of bone loss at different regions of the femoral neck. Our future work aims at improving fracture risk assessment using the developed atlas to determine whether this increased resolution enhances the fracture predictive ability of DXA.

3.7 Conclusion

This work presented the development of a reference spatio-temporal model of ageing bone in the femur using a large cohort of North Western European Caucasian women (n=13,338). I have presented a technique, termed region free analysis (RFA), to eliminate morphological variation between DXA scans by warping each image into a reference template. This image warping establishes a virtual correspondence between pixel coordinates enabling sound statistical inference at the pixel level. I have also presented a novel cross-calibration procedure, termed quantile matching regression technique, to integrate data from different studies into an amalgamated large-scale dataset. Unlike previous techniques, no multiple measurements of each subject on different scanners are required. DXA RFA has the potential to transform conventional bone densitometry routine where spatial resolution is limited due to pooling pixels in pre-defined regions of interest. In the next chapter, I will explore the new insights taken from the developed atlas into the osteoporosis research.

Chapter 4

Application of Bone Atlas to Understand Ageing and Osteoporosis

Chapter 3 presented the development of a reference spatio-temporal atlas of bone mineral density (BMD) ageing in the proximal femur. Heat-maps were deployed to visualise the evolution of spatial BMD patterns with ageing. This chapter presents four key contributions. First, the added value of the developed atlas to delineate between trabecular and cortical bone architecture in the femur is presented, for which conventional region-based analysis is insensitive. Second, a new index called *bone age* is introduced to reflect the evolution of bone microarchitecture with ageing. Bone age aims to model the actual progression, rather than the chronological age, of each subject along the median bone ageing trajectory. I will demonstrate the ability of bone age to serve as a metric for estimation of the progression of osteoporosis. Third, normalising BMD maps for bone age revealed subtle localised fracture-specific patterns that would not be identifiable without excluding the ageing effect. A new index called *f-score* is introduced to quantify these patterns. Integrating bone age and f-score together enhanced hip fracture prediction by 3% measured as the area under the receiver operative characteristic (ROC) curve. Finally, the ability of the proposed pipeline to support extra explanatory variables beyond age as the primary covariate is demonstrated by modelling the impact of body mass index (BMI) on spatial BMD distribution.

4.1 Introduction

Ageing is associated with an increased rate of fracture in the elderly population [149]. Both bone quality deterioration and other extraosseous age-related factors such as reduced proprioceptive efficiency or impaired reflexes also contribute to this elevated fracture risk [25]. The first factor concerning the detrimental effects of ageing on bone strength underlies the mechanism for osteoporosis development [110]. The details of osteoporosis progression with ageing needs to be better understood [150, 147], for which, development of a comprehensive model of ageing bone has been of great interest in osteoporosis research.

Dual-energy X-ray absorptiometry (DXA) has been widely used to model age-related changes in bone mineral density (BMD) measured at different regions of interest (ROIs) in the femur [151, 152, 150, 153, 154]. Despite an overall decline in BMD, the pattern is site-specific; the decrease in BMD measured at femoral neck, trochanter, and Ward's triangle is almost linear while the observed pattern at total hip is bimodal with a steeper drop at advanced ages [151, 152, 153, 154]. Table 4.1 shows the yearly percentage BMD reduction for women at four ROIs commonly reported in the literature. BMD reduction rates varied at different sites within the femur. The steepest reduction in BMD occurred in Ward's triangle whereas the modest decrease occurred in great trochanter. At total hip, BMD was mostly preserved before the 60^{th} year followed by a steep decrease afterwards.

Table 4.1: Yearly percentage BMD reduction in women at four femoral ROIs from selected publications.

	N	Age Range (years)	Neck	Ward's triangle	Trochanter	Total Hip	
						<60	>50
Warming et al. [151]	398	20-89	0.42	-	-	0.24	0.37
Melton et al. [152]	351	21-93	0.50	0.88	-	0.27	0.54
Beck et al. [150]	2904	20-99	0.45	-	-	0.11	0.51
Burger et al.[153]	1084	>55	0.39	0.51	0.25	-	-
Aloia et al.[154]	257	20-80	0.40	0.71	0.20	-	-

Due to variation in methodologies or unreported site-specific BMD change rates, all rates were recomputed from the raw data presented in the manuscripts using linear regression analysis. For total hip BMD, data is divided into two overlapping segments: those who aged 60 years or less in one group, and those who aged 50 years or more in another one.

Conventional DXA analysis has facilitated our understanding of ageing bone in the femur, but has two limitations. First, conventional DXA analysis has limited ability to characterise local ageing patterns in spatial BMD distribution. The reported BMD reduction rates at different ROIs suggests that averaging pixel BMD values at larger ROIs could mask more localised BMD patterns in the femur (Table 4.1); while a steep decrease is observed consistently at Ward's triangle throughout the adulthood age range, total hip BMD appears to be less sensitive to ageing especially at the younger age groups. Second, conventional DXA analysis cannot reflect age-related changes in the trabecular microarchitecture. Several studies have suggested a relationship between bone microarchitecture and bone strength, independent of site-specific femoral BMD values [155, 156, 150]. Therefore, quantifying the ageing effect on trabecular architecture may further enhance our understanding of ageing bone in the femur.

To address these limitations, other high-resolution imaging modalities including Quantitative Computed Tomography (QCT) [157] and high resolution peripheral QCT (HR-pQCT) [156] have been used. While these techniques provide enhanced spatial resolution comparing with DXA, access to sufficiently large-scale datasets representing the ageing population is difficult; most studies deploying QCT or HR-pQCT were conducted on a few hundred cases (see [156] for a review).

DXA region free analysis (RFA) with an enhanced spatial resolution of $0.5 \times 0.5\,\mathrm{mm}^2$ allows quantification of local BMD variation across large-scale datasets comprising thousands of subjects. In the previous chapter, a spatio-temporal atlas of ageing bone was developed in which spatial BMD distribution was visualised using heat-maps. In this chapter, first, the relationship between the observed atlas-based BMD patterns and the site-specific BMD reduction rates reported in the literature is examined. The added value of the developed atlas to delineate between trabecular and cortical bone arrangements are presented in section 4.2. Second, to quantify the overall change in spatial BMD distribution with ageing, a new index called *bone age* is introduced (section 4.3). Given that osteoporosis is a silent disease in the absence of fracture, the feasibility of using bone age for estimation of osteoporosis progression is discussed. Third, given that age has a dominant impact on spatial BMD distribution, normalising BMD maps for bone age allowed quantifying subtle local fracture-specific patterns in the femur. A new index called *f-score* is introduced to quantify these patterns. F-score, when integrated with bone age, enhanced hip fracture prediction by 3% measured as the area under the receiver operative characteristic (ROC) curve (section 4.4). Fourth, the ability of the proposed pipeline to support further explanatory variables beyond age as the primary covariate is demonstrated by modelling the impact of body mass index (BMI) on spatial BMD distribution. Section 4.6 concludes this chapter.

4.2 Cortical and Trabecular BMD Variation

Fig. 4.1 shows the ultra-structure arrangements of cortical and trabecular bone in the proximal femur. Cortical bone is mainly found in the femoral shaft and inferior neck whereas trabecular bone with the sponge-like structure resides inside the femoral trochanter, neck and head. The turnover of trabecular bone is greater than cortical bone leading to higher BMD change rates inside the femur comparing with the outer cortex (Fig. 4.2).

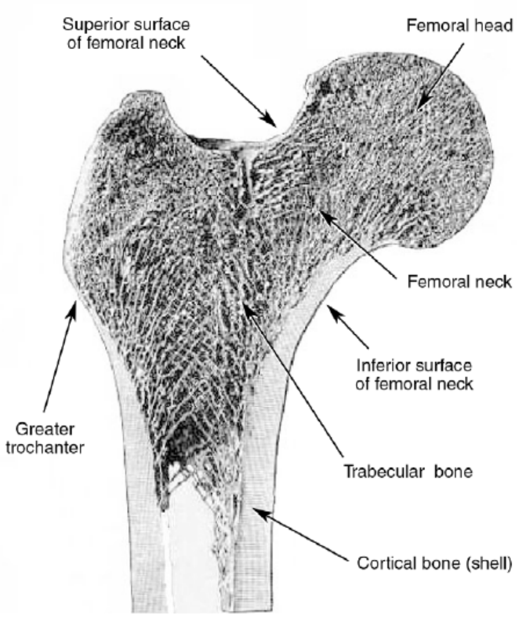


Figure 4.1: Ultra-structure arrangements of cortical and trabecular bone in the proximal femur. Cortical bone (the outer highly mineralised shell) is seen at the shaft and the inferior neck, whereas trabecular bone with the sponge-like structure resides inside the cortical shell. (Adapted from [158]).

For most of adulthood, cortical BMD is mostly preserved with slight increase at the inferior neck until the 7^{th} decade. Following that point, cortical BMD showed a consistent decrease with an approximate annual rate of 0.5%. The enhanced RFA resolution allowed quantification of cortical thickness at the femoral diaphysis. To this end, the BMD profiles at various cross-sections perpendicular to the longitudinal axis of the femur at that given point were analysed for peak bone mass (Fig. 4.3). The width of the femoral shaft at each cross-section minus the distance between the two peak points was then defined as the peak cortical thickness. Fig. 4.4 shows the average cortical thickness variation with ageing at the femoral shaft. The average cortical thickness decreased linearly with ageing from 6.65 mm at 20 years to 5.55 mm at 80 years.

Trabecular BMD patterns inside the femur were spatially complex (Fig. 4.2). For much of adulthood, BMD reduction was more dominant at the

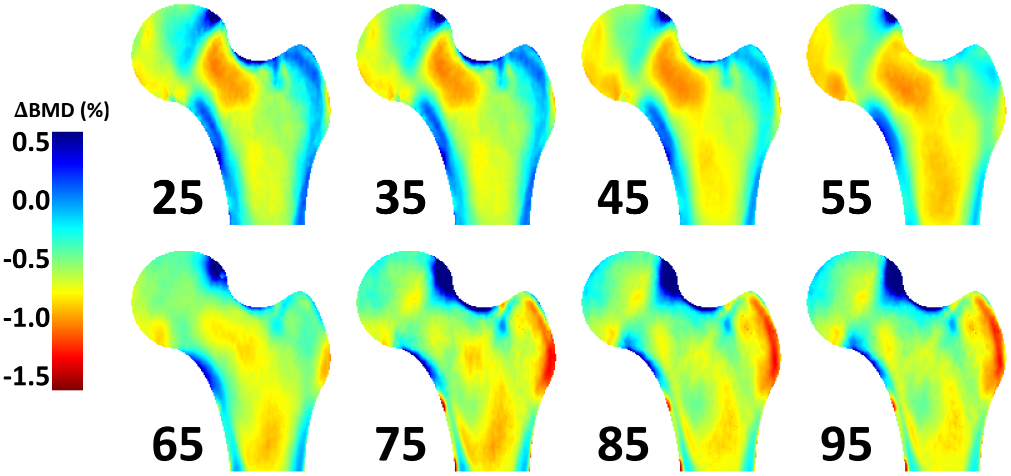
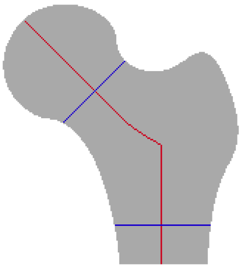
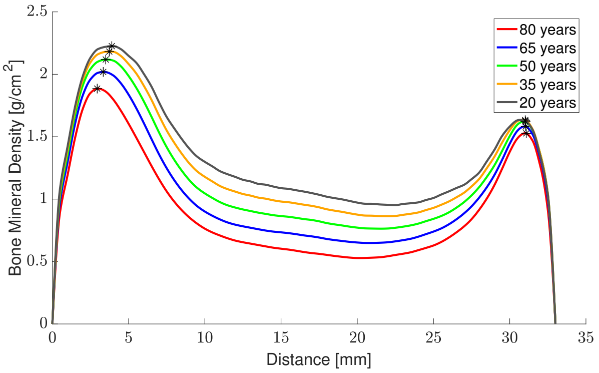


Figure 4.2: Yearly percentage BMD change in the proximal femur. Pixel BMD change rates are normalised to the median BMD at 25 years. BMD was mostly preserved in the cortical bone till 60th year following a consistent decrease with approximate annual rate of 0.5%. Variation in the arrangement of trabecular architecture looks spatially complex. In the early adulthood, BMD reduction in trabecular bone was faster at the femoral neck. In the middle adulthood, BMD reduction accelerated throughout the femur. In the advanced adulthood, BMD reduction was more dominant at the femoral shaft and greater trochanter.

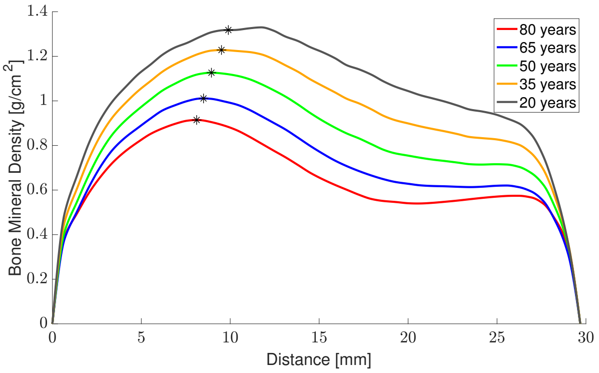
femoral neck whereas, at advanced ages, BMD reduction was more dominant at the femoral shaft and great trochanter. The decrease in trabecular BMD accelerated in the mid-fifties throughout the femur which could be attributed to menopause in women.

Observed spatial BMD patterns were consistent with the reported BMD reduction rates at different ROIs. The sharpest decrease from high to low was reported in the Ward's triangle, femoral neck, and great trochanter, respectively (Table 4.1). This variation in BMD reduction rates can be explained well using the spatial BMD patterns observed in Fig. 4.2; the reddish spot associated with the greatest BMD reduction at the medial femoral neck is consistent with accelerated BMD reduction at the Wards' triangle. Averaging pixel BMD values over the whole neck, however, would make the site-specific BMD change less sensitive to ageing. Pixel BMD change at the great trochanter, visualised with green and blue colours in Fig. 4.2, suggests slower BMD loss at this region comparing with femoral neck. This observation is consistent with the reported BMD reduction rate at the trochanter.





(a) Femoral Diaphysis



(b) Femoral Neck

Figure 4.3: Proximal femoral bone density profiles at two cross-sections. (a) BMD profiles at the femoral shaft demonstrated an M-shape graph with peak BMD at the outer cortex and lower trabecular BMD in the middle. On each graph, the two peak BMD values in the interior and exterior cortex are marked with asterisks. Peak cortical thickness is defined as the width of the cross-section line minus the distance between the two peaks in the BMD profile. (b) BMD profile at the femoral neck demonstrated only one distinct local peak BMD at the inferior cortex where cortical bone is present (see Fig. 4.1). Ageing is associated with a decrease in peak BMD.

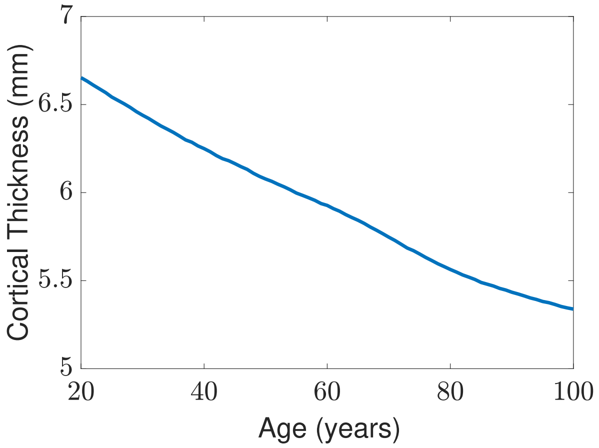


Figure 4.4: Average peak cortical thickness variation with ageing. Average peak cortical thickness at the diaphysis was linearly decreased with ageing from 6.65 mm at 20 years to 5.55 mm at 80 years.

4.3 Osteoporosis Progression Index

Osteoporosis is strongly related to ageing which causes gradual deterioration in bone structure leading to elevated fracture risk in the elderly population. Despite the progressive nature of osteoporosis, currently the degree of progression or the severity of disease is not estimated in clinical practice. Alternatively, bone is categorised as either *normal*, *osteopenia*, or *osteoporosis* based on the BMD measurement at the femoral neck or spine reported as T-score, i.e. the number of standard deviation (SD) from the average for young healthy women in the population. The world health organisation (WHO) criteria for the diagnosis of osteoporosis is based on the T-score cut-off point of -2.5 for osteoporosis and -1 for osteopenia. This definition, however, is recognised as being arbitrary and controversial [159]. Moreover, these discrete categories cannot explain the mechanism by which osteoporosis progressively affects bone architecture in the femur.

Given that osteoporosis is a silent disease unless a fracture occurs, osteoporosis progression estimation is a challenging task due to lack of clinical symptoms. Given the close relationship between osteoporosis and ageing, one can postulate that the *normal* ageing trajectories estimated for the population also represent the disease progression trajectory. Osteoporosis is then modelled as accelerated bone loss on the same trajectory attributed to the natural ageing phenomenon. In this context, the actual progression on the bone ageing trajectory, rather than the chronological age, defines the current osteoporosis severity. Observe that modelling osteoporosis as accelerated normal ageing can only be used for primary age-related/postmenopausal osteoporosis. It has not been tested for causes of secondary osteoporosis such as primary hyperthyroidism or glucocorticoid treatment.

Fig. 4.5 demonstrates the bone ageing concept using a schematic graph in a 2D space. The solid black line represents the median ageing trajectory using BMD at two pixel coordinates. For an individual BMD map represented by a green dot in Fig. 4.5, *bone age* is defined as the actual progression along the ageing trajectory estimated by mapping the given BMD map to the closest point on the graph (red dot in Fig. 4.5). Note that the actual bone ageing trajectory lies on an N-dimensional space with $N = 16035$ is the number of pixels in the template (Fig. 4.6).

Let $\boldsymbol{a}(a)$ denote the median bone map at age a and $\boldsymbol{b}$ denotes the bone map for an individual subject. Both $\boldsymbol{a}$ and $\boldsymbol{b}$ are represented in the vectorised format $\mathbb{R}^{N\times 1}$ where N is the number of pixels in the template. Then, the bone

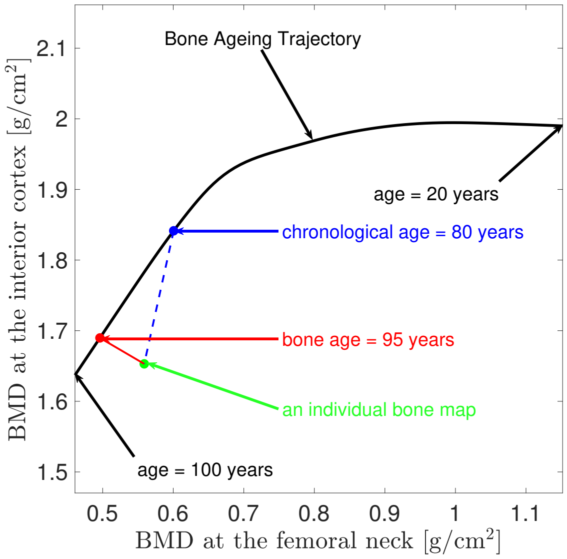


Figure 4.5: Schematic bone ageing trajectory in a 2-dimensional space. The solid black line represents the median ageing trajectory using BMD at two pixel coordinates. One pixel is selected from the femoral neck, and the other one is selected from the cortex near the lesser trochanter. For a given bone map (green dot), its bone age is estimated by mapping the given bone map to the closest point on the trajectory. Note that the actual bone ageing trajectory lies on an N-dimensional space where N equals the number of pixels in the template.

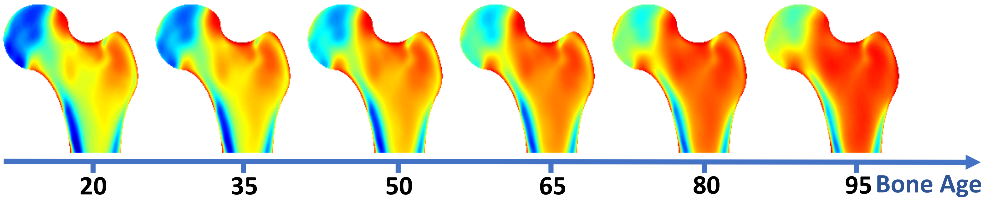
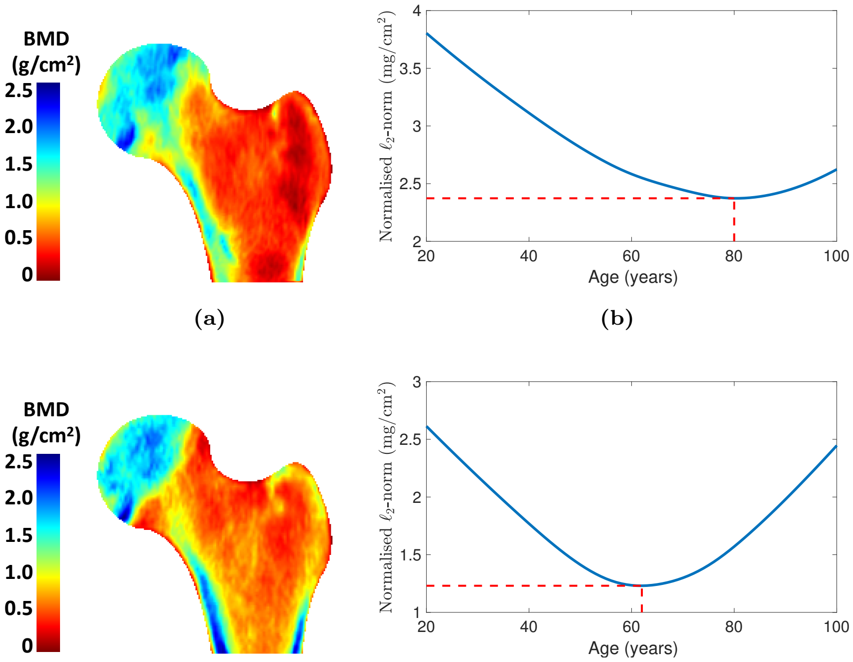


Figure 4.6: Median bone ageing trajectory. The median bone ageing trajectory is a 1D graph in the N-dimensional space where $N = 16035$ is the number of pixels in the template.

age a' is defined as the age for which $\boldsymbol{a}(a')$ best fits the given bone map $\boldsymbol{b}$. Various similarity metrics could be used in practice. Here, the simple L_2-norm is deployed as the dissimilarity metric:

$$a' = \underset{a}{\operatorname{argmin}} \, \|\boldsymbol{b}^T - \boldsymbol{a}(a)\|_{L_2}. \tag{4.1}$$

Fig. 4.7 demonstrates the intuitive rationale behind the bone age. Fig. 4.7 shows two sample BMD maps of women aged 76 years with similar femoral neck BMD of 0.59 g/cm^2 where one sustained a follow-up hip fracture (the top row) and the other did not (the bottom row). Despite similar age and femoral neck BMD, the bone architecture varies between the two subjects. The one with the interval fracture demonstrates widespread bone loss in the trochanteric region and relatively thinner cortical thickness leading to an elevated bone age of 80 years (cf. Fig. 4.6). The bone map shown on the bottom row demonstrates higher densities at the bone cortex in the non-fracture subject giving a bone age of 62 years (cf. Fig. 4.6).



(a) (b)

(c) (d)

Figure 4.7: Intuitive illustration of bone age potential to differentiate between fracture and control subjects with similar neck BMD values. The top row shows the bone map for a woman of aged 75.8 years with femoral neck BMD of 0.5860 g/cm^2 who experienced a hip fracture following the baseline measurement. The bottom row shows the bone map for a non-fracture subject with similar age (75.9 years) and femoral neck BMD (0.5900 g/cm^2). Despite similar age and femoral neck BMD, the spatial texture pattern varies between the two subjects. The associated bone age was 80 and 62 years for the top and bottom rows, respectively.

The term *bone age* has been used for decades in medicine (by paediatricians) to measure the skeletal maturity in a child, and is based on a comparison of a wrist radiograph with atlas patterns to assess the closure of the growth plates [160, 161]. Despite differences in methodologies for estimation of bone age in children and adults, the underlying concept is similar; bone age represents the average age at which a specific degree of maturation/deterioration is expected. Given this analogy, I proposed the term bone age here for assessing the degree of bone deterioration in the adulthood, but one should observe the difference between this technique and the bone age assessment in children in a broader context.

Bone age is an abstract concept for modelling osteoporosis progression. Despite its intuitive perception (see Fig. 4.7), it is difficult to validate bone age directly using clinical criteria as osteoporosis is asymptomatic. However, in order for bone age to usefully facilitate the management of osteoporosis, it must satisfy the following conditions: first, it should be consistent with the current established diagnostic guidelines for osteoporosis, which is based on femoral neck BMD. Second, bone age should be sufficiently precise; bone age measured on repeated DXA measurements on the same day with patient repositioning between scans should be similar. Third, it should demonstrate the ability to distinguish between a young healthy population and a more elderly cohort. Forth, a consistent increase in fracture risk should be observed with more advanced bone ages. Fifth, it should predict fragility fractures at least as well as the current metrics, including FRAX and conventional BMD measurement.

To validate these properties, DXA scans selected from the MRC-Hip dataset and the OPUS study were deployed (see section 3.3.1). The subjects in the MRC-Hip study were followed up for a period of five years after the baseline measurements for any major osteoporotic fractures at the hip, spine, pelvis, upper limb or lower limb sites. The total number of fractures at these sites was 684, of which 178 were reported at the hip. The number of control cases who remained fracture-free was 4249 [139].

4.3.1 Consistency with Current Diagnostic Guidelines

For bone age to serve as a continuum for osteoporosis progression, it should be consistent with currently established metrics for osteoporosis diagnosis including neck BMD and FRAX score. Bone age was highly correlated with neck BMD ($r = -0.87$; $p < 0.001$; Fig. 4.8). The blue and red dots represent the fracture-free controls and the hip fracture cases, respectively. The density of red dots increases at the bottom-right corner of the figure, consistent with both a decrease in neck BMD and an increase in bone age.

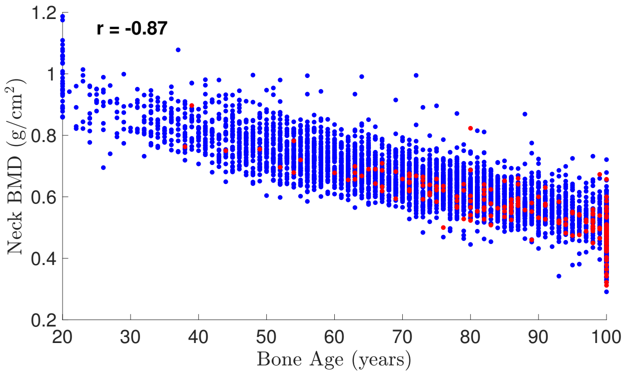


Figure 4.8: Relationship between bone age and femoral neck BMD. Bone age is linearly correlated with neck BMD ($r = -0.86$; $p < 0.001$). The blue and red dots represent the fracture-free controls ($n = 4249$) and the hip fracture cases ($n = 178$), respectively. The density of red dots increases at the bottom-right corner, consistent with both a decrease in neck BMD and an increase in bone age.

Fig. 4.9 shows the relationship between bone age and FRAX estimated with BMD as a risk factor. FRAX score increased with bone ageing with an almost exponential pattern. This pattern is consistent with the established exponential relationship between FRAX and neck BMD.

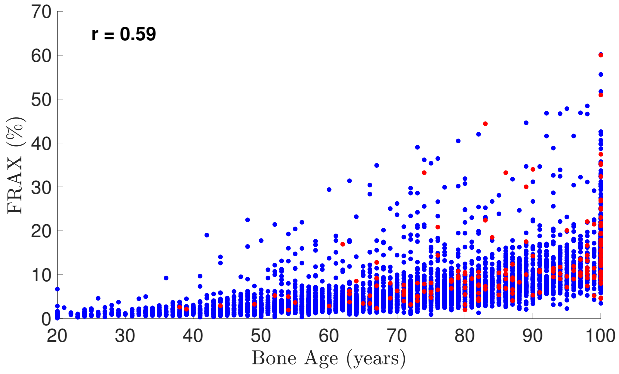


Figure 4.9: Relationship between bone age and FRAX. The FRAX score consistently increases with bone ageing with an exponential pattern. The blue and red dots represent the fracture-free controls ($n = 4249$) and the hip fracture cases ($n = 178$), respectively. The density of red dots increases with an increase in bone age and FRAX score. Note that FRAX is reported with BMD as a risk factor.

4.3.2 Bone Ageing Precision

For bone age to be useful in clinical practice, it should be precise when measured repeatedly with a short time lapse between scans. Here, precision was validated using 25 pairs of scans collected on the same day with patient repositioning between scan collections. The standard deviation (SD) of error was 1.4 years. No significant difference was observed between groups using a paired t-test ($p = 0.54$). Fig. 4.10 shows the scatter plot of the estimated bone ages for each subject. Using Deming regression analysis (see section 3.2.3.(B)), the slope and the intercept were estimated as 1.00 and 0.51, respectively.

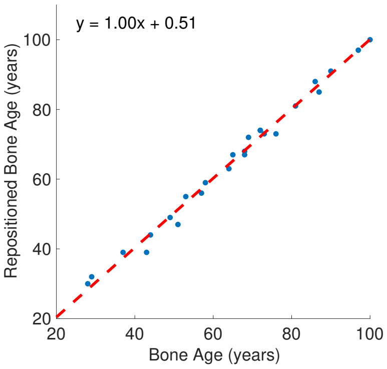
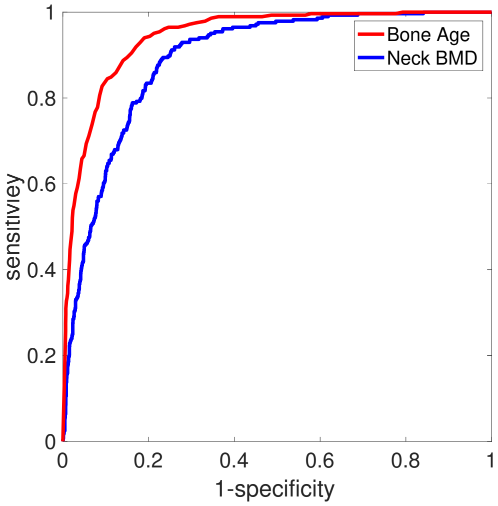


Figure 4.10: Bone ageing precision analysis. Twenty-five subjects were scanned two times on the same day with patient repositioning between scans. For each subject, bone age is estimated independently for each of the two collected scans. The SD for precision error was 1.4 years. No significant difference was observed using a paired t-test ($p = 0.54$).

4.3.3 Distinction between a Young Healthy Cohort and an Elderly Population

For bone age to serve as a useful metric to assess gradual deterioration of bone structure with ageing, it should be able to distinguish between a young healthy cohort with strong bones and an older cohort with excess fragility. For this analysis, a young healthy cohort ($n = 284$; age<40 years) and an older one ($n = 2165$; age>80 years), selected from the OPUS or MRC-hip study, were compared against each other. Fig. 4.11 shows the receiver operative characteristic (ROC) curve for the ability of bone age versus neck BMD to classify between these two groups. The area under the curve (AUC) was 0.89 (95% confidence interval (CI)=0.874-0.906) and 0.94 (95% CI=0.930-0.954) for neck BMD and bone age, respectively. The AUC for bone age was found to be 5% higher than neck BMD alone ($p < 0.001$; Fig. 4.11).



(a) Hip Fractures

Figure 4.11: The ROC curve for ability of bone age versus neck BMD to classify between young and old populations. The young cohort ($n = 284$) includes all white women aged 40 years or less selected from the OPUS dataset. The old cohort ($n = 2165$) includes all white women aged 80 years or more selected from the OPUS or the MRC-Hip datasets. The AUC for bone age and neck BMD was 0.94 (95% CI=0.930-0.954) and 0.89 (95% CI=0.874-0.906).

4.3.4 Fracture Risk and Bone Age

As bone ages, it gets weaker and thereby more likely to sustain a fragility fracture. In line with this hypothesis, fracture risk increased consistently with bone age (Figs. 4.12(a) and (b)). Fracture risk was computed as the number of fracture cases divided by the number of subjects at each age band. Similar trends were also observed with neck BMD and the FRAX score (Fig. 4.12).

However, the increase in fracture risk the chronological age was only consistent at the hip, and the chance of suffering from a fracture at any site was broadly similar across age bands (Fig. 4.12(h)).

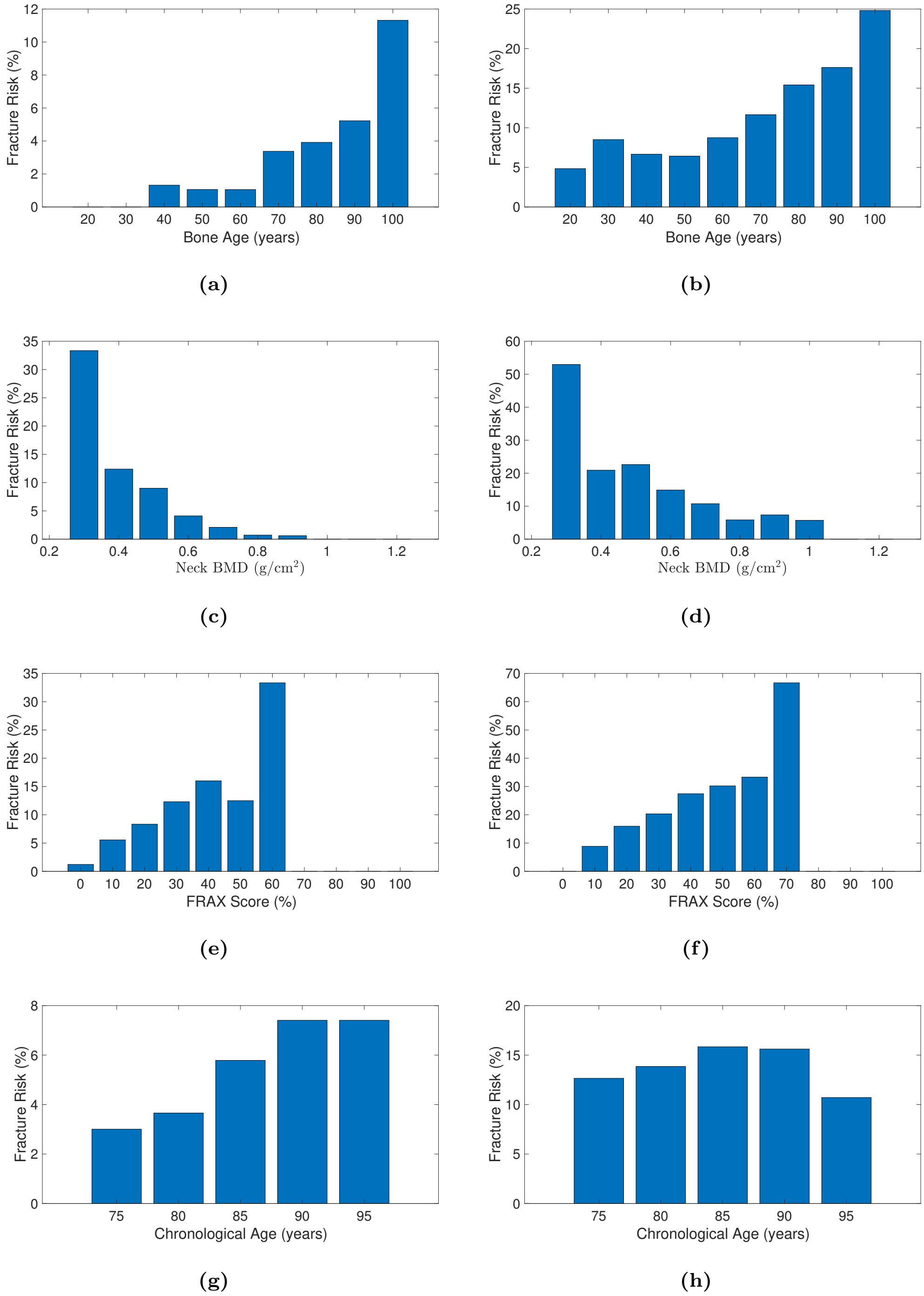
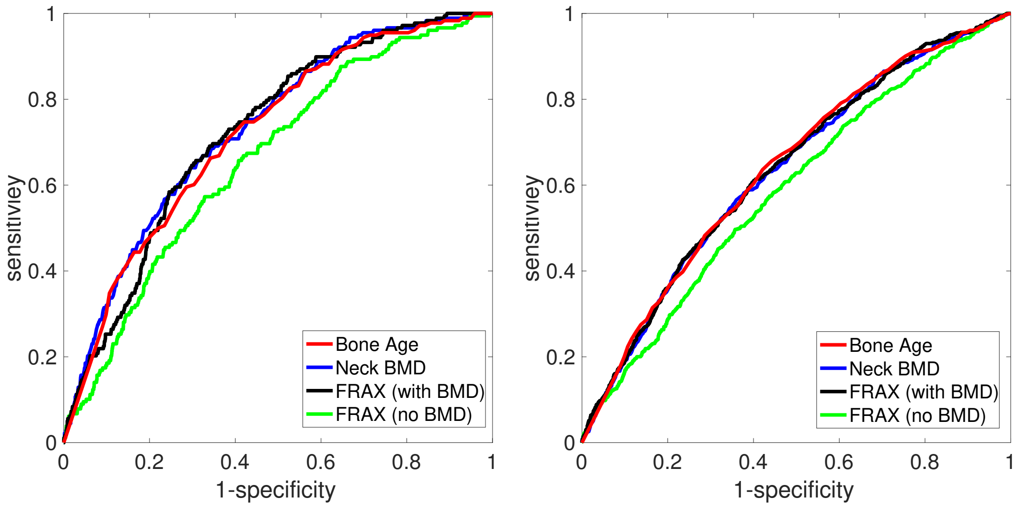


Figure 4.12: Stratified fracture risk based on the bone age (the first row), neck BMD (the second row), FRAX score (the third Row), and chronological age (the forth row) in the MRC-Hip study. The first column shows the fracture risk at the hip while the second column shows the risk for any fragility fractures occurred at the hip, spine, pelvis, lower limb, or the upper limb. The total number of fractures was 684 out of which 178 occurred at the hip. The number of control cases was 4249.

4.3.5 Fracture Prediction

Fragility fractures are the result of reduced bone strength but also other extraosseous factors such as the likelihood of falling. Therefore, population attributable risk (PAR) attributable to low bone mass is modest (approximately 50% as reported in the study of osteoporotic fractures (SOF) [15]) due to significant overlap in BMD between fracture and fracture-free control groups. However, any metric that can improve on the discrimination between the fracture and the control groups should, in practice, further facilitate osteoporosis management. Fig. 4.13 shows the ROC curves for prediction of follow-up fragility fractures in the MRC-Hip study. The AUC for hip fracture prediction was 0.731 (95% CI=0.689-0.761), 0.723 (95% CI=0.690-0.754), 0.660 (95% CI=0.619-0.694), and 0.719 (95% CI=0.682-0.755) for neck BMD, FRAX with BMD, FRAX without BMD, and bone age, respectively. The AUC for prediction of any major osteoporotic fractures was 0.632 (95% CI=0.609-0.651), 0.636 (95% CI=0.613-0.656), 0.590 (95% CI=0.569-0.613), and 0.639 (95% CI=0.618-0.661) for neck BMD, FRAX with BMD, FRAX without BMD, and bone age, respectively.



(a) Hip Fractures **(b)** All Major Fragility Fractures

Figure 4.13: The ROC curve for prediction of fragility fractures. The AUC for prediction of hip fractures (n=178) was 0.731 (95% CI=0.689-0.761), 0.723 (95% CI=0.690-0.754), 0.660 (95% CI=0.619-0.694), and 0.719 (95% CI=0.682-0.755) for neck BMD, FRAX with BMD, FRAX without BMD, and bone age, respectively. The AUC for prediction of any major osteoporotic fractures (n=684) was 0.632 (95% CI=0.609-0.651), 0.636 (95% CI=0.613-0.656), 0.590 (95% CI=0.569-0.613), and 0.639 (95% CI=0.618-0.661) for neck BMD, FRAX with BMD, FRAX without BMD, and bone age, respectively. The number of fracture-free controls was 4249.

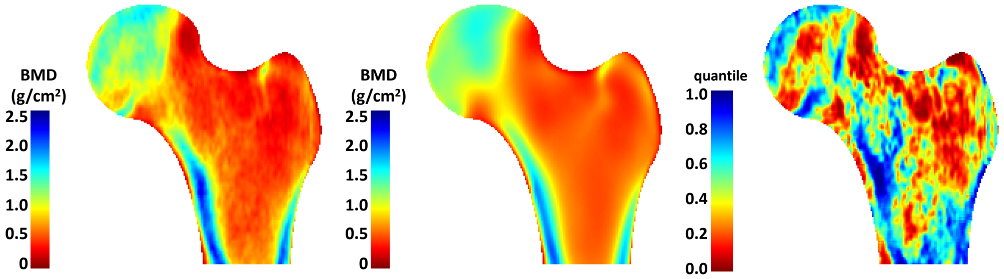
No statistically significant difference in the measured AUC was observed between neck BMD, FRAX with BMD, and bone age (Fig. 4.13). Note that using neck BMD, the PAR for osteoporotic fractures in the MRC-Hip cohort is relatively high (75%) in comparison to the SOF study [15]. This may be due to the extremely old cohort recruited in the MRC-hip study; the average age for the MRC-Hip study [139] was approximately ten years older than the SOF study [15]. Nonetheless, the PAR between neck BMD and osteoporotic fractures in the MRC-Hip study is already high. For comparison purposes, the PAR between smoking and lung cancer is approximately 80% [162]. The finding of similar performance between bone age and neck BMD in the MRC-Hip cohort is promising. However, applying bone age to a younger cohort may present better performance than neck BMD alone, although this is yet to be tested.

Bone age reflects the overall evolution in trabecular architecture with ageing. However, it does not account for local fracture-specific patterns in the femur. For enhanced fracture prediction, local BMD patterns could be deployed as discussed in the following section.

4.4 Localised Fracture-Specific Bone Patterns

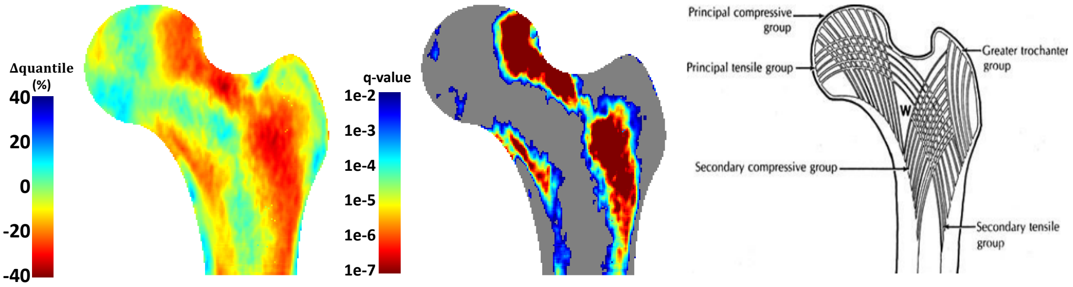
The purpose of this section is to demonstrate the ability of DXA RFA in localising fracture-specific patterns in the femur and to see if these local patterns could potentially enhance hip fracture prediction. To this end, pixel BMD values were first normalised to account for the variation in bone age by mapping each pixel BMD to an appropriate quantile value from the developed atlas. Fig. 4.14 shows the bone map for a woman of age 77 years who sustained an interval hip fracture during follow up, the median atlas at the estimated bone age of 83 years, and the quantile map for this individual. Pixel quantiles reflect the rank of the given pixel BMD values among the population with a similar bone age and thereby, the quantile map could be seen as a normalised BMD map with respect to the estimated bone age.

Normalisation with respect to bone age excludes variation due this variable and allows to identify fracture-specific patterns in the femur. Fig. 4.15 shows the difference in mean between hip fractures ($n = 178$) and fracture-free controls ($n = 4249$) with the statistical significance map reported as a FDR q-value map. The q-map shows a local pattern of bone loss oriented in the same direction as principal tensile curves characterised in the radiography scans using the Singh index [51]. Similar analysis on raw pixel BMD values could not localise fracture patterns (Fig. 4.16).


BMD (g/cm²)
2.5
2.0
1.5
1.0
0.5
0
BMD (g/cm²)
2.5
2.0
1.5
1.0
0.5
0
quantile
1.0
0.8
0.6
0.4
0.2
0.0

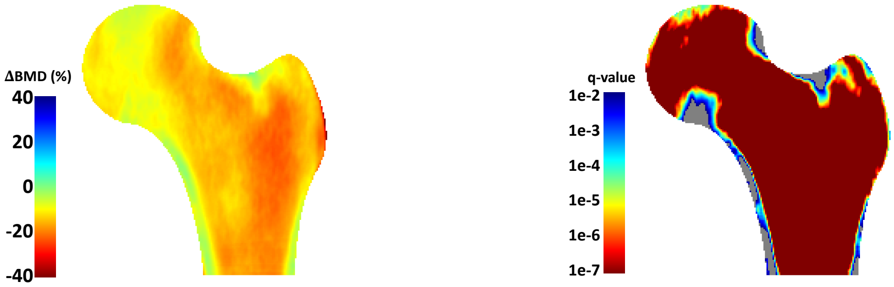
(a) BMD map **(b)** Median atlas **(c)** Quantile map

Figure 4.14: Bone-age normalised BMD map. Panel (a) shows the bone map for a woman of age 77 years who sustained a follow-up hip fracture. Panel (b) shows the median atlas at the estimated bone age 83 years for this subject. Panel (c) shows the normalised BMD map or the quantile map. Pixel quantiles reflect the rank of the given pixel BMD values among the population with a similar bone age.


Δquantile (%)
40
20
0
-20
-40
q-value
1e-2
1e-3
1e-4
1e-5
1e-6
1e-7
Principal compressive group
Principal tensile group
Greater trochanter group
W
Secondary compressive group
Secondary tensile group

(a) (b) (c)

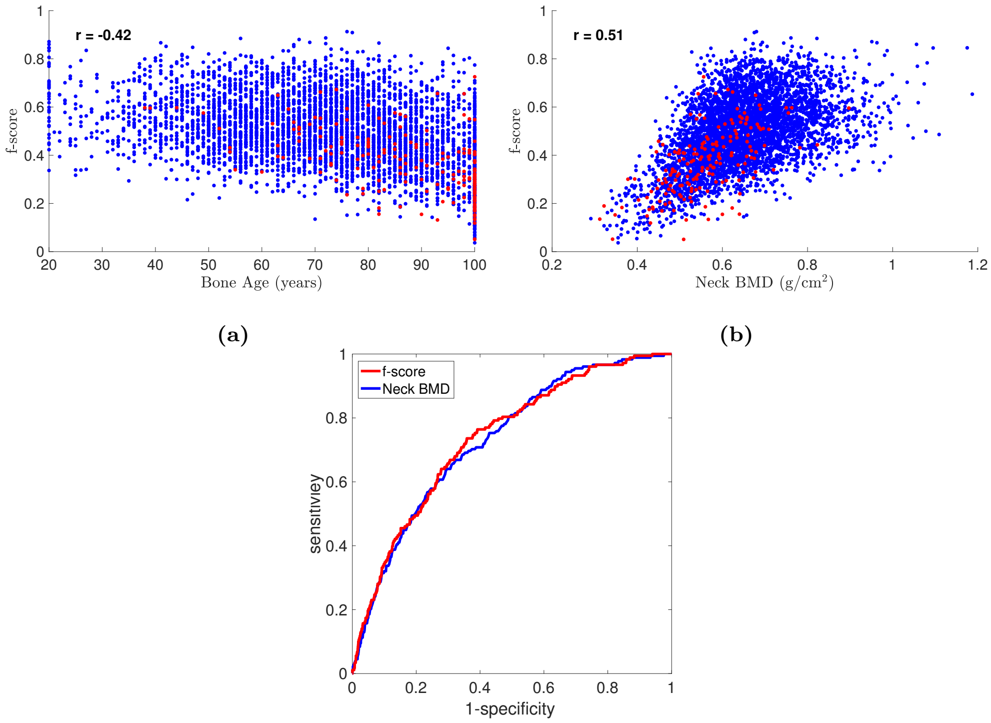
Figure 4.15: Localising fracture-specific patterns using bone-age normalised BMD maps. Panel (a) shows the difference in mean quantile maps between the fracture and the fracture-free control groups. Panel (b) shows the corresponding statistical significance map using a two-sample t-test followed by FDR analysis. A local pattern of bone loss was observed in the same orientation as principal tensile curves characterised in the radiography scans [51]. Panel (c) shows the trabecular arcades in the proximal femur deployed for assessing the Singh index. The image is adapted from [163].


ΔBMD (%)
40
20
0
-20
-40
q-value
1e-2
1e-3
1e-4
1e-5
1e-6
1e-7

(a) (b)

Figure 4.16: Localising fracture-specific patterns using raw BMD maps. Panel (a) shows the difference in mean BMD maps between the fracture and the fracture-free control groups. Panel (b) shows the corresponding statistical significance map using a two-sample t-test followed by FDR analysis. Raw BMD unlike the quantiles cannot localise fracture-specific patterns (cf. Fig. 4.15).

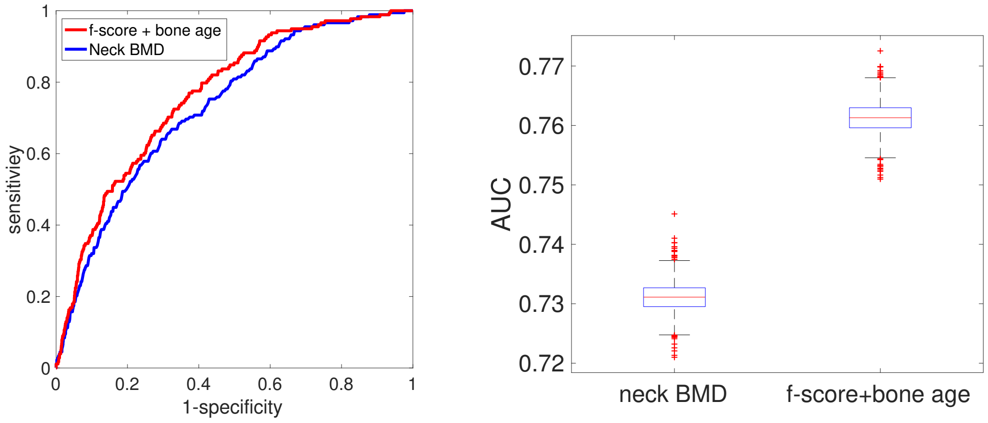
The average pixel quantile values over the region with $q \leq 1e-6$ (the red spot on the q-map; Fig. 4.15(b)) defines a new score, named *f-score*, with power to predict fractures independent of bone age or neck BMD (see Fig. 4.17). Fig. 4.17(a),(b), and (c) show the correlation between f-score and bone age ($r = -0.4$), the correlation between f-score and neck BMD ($r = 0.5$), and the ROC curve for the ability of f-score versus neck BMD to identify hip fractures. The AUC was 0.731 (95% CI=0.689-0.761) and 0.736 (95% CI=0.694-0.769) for neck BMD and f-score, respectively. The low correlation between f-score and either bone age or neck BMD and similar power of f-score to neck BMD for hip fracture prediction may suggest the potential to enhance fracture prediction by combining f-score and bone age. One way to combine bone age and f-score is to deploy a logistic regression technique to compute the appropriate weights between bone age and f-score. Fig. 4.18(a) shows the ROC curve for the combination of f-score and bone age versus neck BMD. Since logistic regression requires training to estimate appropriate weights, I used a 5-fold cross-validation technique; each time one fold was left out for testing and the weights were learned on the remaining data. Therefore, five different values for AUC are computed for each division of the dataset into 5 segments. I repeated this procedure 1000 times and the distribution of the average AUC values for each experiment are reported in Fig. 4.18(b). The new combined score improves the AUC significantly by 3% over the conventional neck BMD.



(a) (b)

(c)

Figure 4.17: The ability of f-score to predict fractures independent of bone age or neck BMD. (a) Relationship between f-score and bone age. (b) Relationship between f-score and neck BMD. The blue and red dots represent the fracture-free controls and the hip fracture cases, respectively. (c) The ROC curve for classification between hip fractures ($n = 178$) and controls ($n = 4249$). The AUC was 0.731 (95% CI=0.689-0.761) and 0.736 (95% CI=0.694-0.769) for neck BMD and f-score, respectively. The low correlation between f-score and either bone age or neck BMD and similar power of f-score versus neck BMD for hip fracture prediction suggest the potential to enhance fracture prediction by combining f-score and bone age (see Fig. 4.18).



(a) (b)

Figure 4.18: Ability of combined f-score and bone age versus neck BMD to predict hip fractures. Logistic regression analysis was deployed to find appropriate weights to combine f-score and bone age. (a) ROC curve for classification between hip fractures ($n = 178$) and controls ($n = 4249$). (b) Box-plot for the estimated AUC using 1000 different iterations. A five-fold cross-validation technique was deployed; each time one fold was left out for testing and the combination weights were learned on the remaining data. I repeated this procedure 1000 times and the distribution of the average AUC values on all 5 folds are reported.

4.5 BMI Impact on BMD Distribution

The purpose of this section is twofold: first, to demonstrate the ability of the proposed framework in chapter 3 to account for new covariates of interest (excluding the primary explanatory variable age); second, to quantify how BMI variation would affect the spatial distribution of BMD in the proximal femur. Fig. 4.19 visualises the spatial variation in BMD for different values of age and BMI using heat-maps. Low BMI was associated with an overall decrease in bone mass. High BMI resulted in increased bone mass especially at the diaphysis and Ward's triangle regions. The observed BMI impact on BMD patterns were stronger at younger ages than more advanced ages (Fig. 4.19). The correlation between BMI and bone age was $r = -0.33$.

4.6 Discussion

In osteoporosis research, bone is commonly categorised as either *normal* (T-score>-1), or *osteopenic* (-2.5<T-score<-1), or *osteoporosic* (T-score<-2.5) using BMD measurement at the femoral neck. Despite widespread use of this definition, it is recognised as being arbitrary and controversial [159]. First, to base the diagnosis criteria on a clear cut-off point could be misleading; the PAR for a fragility fracture in hip was reported as 28% using T-score<-2.5 and 51% for a more relaxed threshold of T-score<-1.5 in the SOF [15]. This means that half of the fragility fractures are attributable to normal bones. Second, osteoporosis is an age-associated disease in which progression with ageing forms a continuum. Classifying bone status into three discrete categories does not precisely reflect the inherently progressive nature of the disease nor allow estimation of osteoporosis progression rate. The ability to estimate the current severity of the disease as well as its progression rate could facilitate the management of age-associated diseases [164]. Third, patients would be interested to know how their bones compare with a similar person who has aged *normally* rather than the relative difference from a healthy young cohort. Z-score may potentially answer this question by comparing the subject with age- and sex-matched cohort, but Z-score still cannot capture the underlying longitudinal ageing effect. More to the point, Z-score accounts for the chronological age rather than the bone age.

Given the close relationship between osteoporosis and ageing, here I assumed a unique underlying mechanism for bone ageing and presented osteoporosis as an inextricable outcome of senescence. However, I recognised the

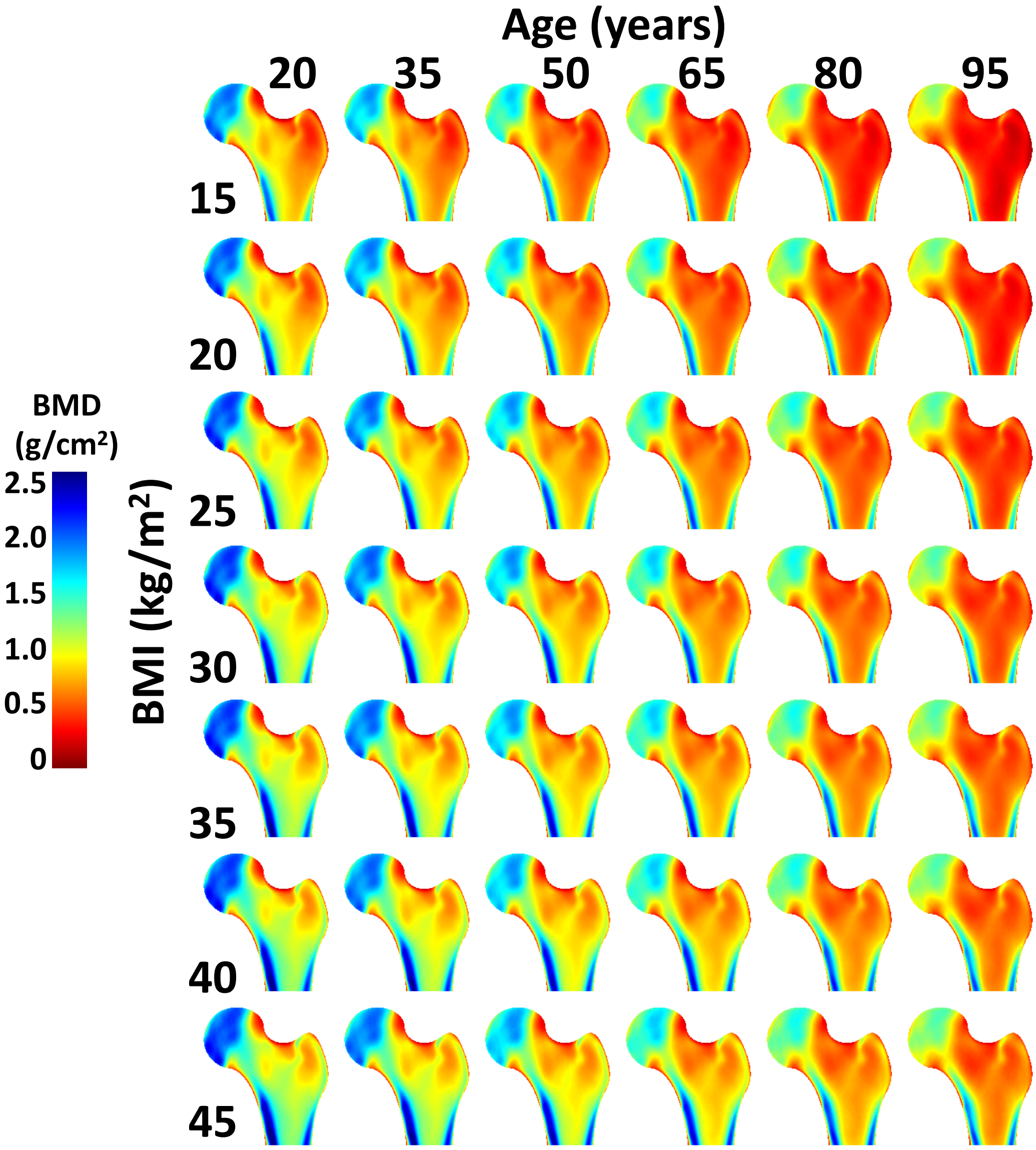


Figure 4.19: Spatial BMD variation with Age and BMI. The median bone maps are visualised for 20, 35, 50, 65, 80, and 95 years of age and different BMI values of 15, 20, 25, 30, 35, 40, and 45 kg/m^2. The atlas is shown for the Hologic system at the left hip.

variation in ageing rates between subjects where osteoporosis could be seen as an accelerated loss in BMD on the same trajectory attributable to the normal ageing process. With this assumption, this chapter presented a technique for estimation of osteoporosis progression based on the trabecular microarchitecture in the femur. A new index, called bone age, was introduced as the age at which the spatial BMD patterns extracted from the atlas best fit the given individual bone map. The terminology *bone age* is not new; it has been used to identify the degree of bone maturation in children. While the methodology for estimation of bone age is different for children than adults, the underlying concept is similar; bone age represents the average age at which a specific degree of maturation/deterioration is expected.

I demonstrated five different properties for bone age that could usefully facilitate the management of osteoporosis. First, bone age is an intuitive metric to reflect the microarchitectural arrangements of trabecular bone in the femur that forms a continuum rather than discrete categories. Second, This continuous scale would potentially allow computation of progression rates as well as the current severity of the disease. The precision error (SD) for bone age estimation was 1.4 years using 25 pairs of scans collected on the same day with patient repositioning between them. Third, bone ageing was associated with a consistent increase in the risk for suffering from an interval fragility fracture. Forth, AUC for the ability of bone age versus neck BMD to discriminate between a young cohort ($n = 284$; age<40 years) and a more older one ($n = 2165$; age>80 years) was 0.94 (95% CI=0.930-0.954) and 0.89 (95% CI=0.874-0.906), respectively. Given that the bone strength would be higher at younger cohorts on average, this enhancement may suggest that bone age could be potentially a better representative of bone strength than neck BMD. Fifth, bone age demonstrated a similar power to neck BMD and FRAX with BMD for prediction of fragility fractures in the MRC-Hip dataset.

Despite the advantages of bone age mentioned above, no improvement over neck BMD was observed for fracture prediction in the MRC-Hip study. One reason for that could be the extremely old population in the MRC-Hip dataset. Using neck BMD, the PAR for fragility fractures was 38% and 75% for a T-score below -2.5 and -1.5, respectively. This means that 75% of fractures are attributable to osteoporotic bones with a more relaxed definition of osteoporosis (cf. PAR of 51% in the SOF study [15]). This is almost comparable with PAR for smoking and lung cancer, which has been estimated to be over 80% [162]. This 25% increase in PAR for MRC-Hip versus SOF could be attributable to the difference in age between the two studies; the MRC-Hip cohort was on average ten years older. The high PAR of 75% makes it difficult to improve fracture prediction over neck BMD as neck BMD already demonstrated good performance in the MRC-Hip cohort. However, given that bone age demonstrated better performance than neck BMD in differentiating between young and old cohorts, deploying bone age in a younger population than the MRC-Hip cohort may potentially demonstrate better performance of bone age versus neck BMD in hip fracture prediction.

The second reason for the limited ability of bone age over neck BMD to improve fracture prediction is that bone age does not *per se* account for localised fracture-specific patterns in the femur. Bone age accounts for age-related changes in the arrangement of trabecular architecture in the femur, but normalising pixel BMD for bone age revealed localised patterns of loss in BMD oriented in the same direction as principal tensile curves characterised on plain radiographs using the Singh index [51]. This BMD normalisation is analogous to Z-scores with two differences: first, bone age is used rather than chronological age; second, since pixel BMD follows a skewed normal distribution, quantile values are reported rather than the difference from the mean expressed in SD. I introduced a new index called f-score by averaging quantile values over the local regions with FDR q-value below $1e-6$. The f-score demonstrated a similar power to neck BMD for fracture prediction independent of bone age or neck BMD. When combining f-score and bone age, the AUC for prediction of hip fractures was significantly increased by 3%. This significant increase, albeit small, may suggest the potential for improving fracture prediction by analysing fracture-specific BMD patterns in bone. This potential may be fully demonstrated once these patterns are correlated with information about the actual fracture patterns in the proximal femur. Fig. 4.20 shows six types of fractures at the subcapital neck, transcervical neck, intertrochanter, subtrochanter, greater trochanter, or lesser trochanter. Although validating this hypothesis requires a large-dataset with fracture data at a wider age range, this is only feasible with the RFA technique and not the conventional region-based analysis.

4.7 Conclusion

This chapter outlines several potential applications for the developed spatio-temporal atlas that could potentially facilitate the management of osteoporosis disease. I demonstrated how the enhanced RFA resolution could be deployed to analyse cortical and trabecular changes in the femur, for which the conventional region-based technique would be insensitive. More specifically, I quantified the average cortical thickness in the diaphysis. A new framework for the progression estimation of osteoporosis was proposed by introducing a new intuitive index called bone age. To improve fracture prediction, a new index called f-score was introduced to reflect localised fracture-specific patterns. Combining f-score and bone age together using logistic regression analysis, the AUC for prediction of hip fractures was significantly increased by 3% over neck BMD alone. Finally, I examined the impact of body mass index (BMI)

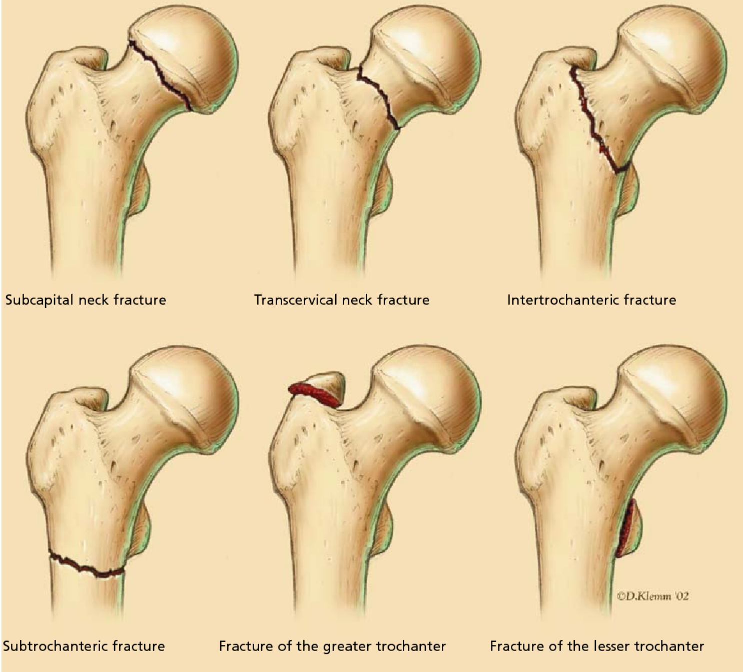

Subcapital neck fracture
Transcervical neck fracture
Intertrochanteric fracture
Subtrochanteric fracture
Fracture of the greater trochanter
Fracture of the lesser trochanter
©D.Klemm '02

Figure 4.20: Various types of fracture patterns observed in the proximal femur. DXA RFA would allow correlation of local BMD patterns with actual fracture patterns in the femur to enhance hip fracture prediction. The image is adapted from [165].

on the spatial distribution of BMD values in the femur to demonstrate the ability of the proposed framework to add new covariates to the developed spatio-temporal atlas.

Chapter 5

Automatic Quality Control of DXA Images

Population imaging studies have opened new opportunities for a comprehensive characterisation of a range of diseases including osteoporosis. Despite strict imaging protocols to ensure consistent high-quality scans, incidental artefacts are inevitable. Detecting these artefacts using a human observer or a panel of experts requires a considerable amount of time and expertise, making it unfeasible for use in large datasets. To address this challenge, methods for automatic image quality control are needed. To date, no standard classification metric exists for the evaluation of DXA artefacts. Here, I first propose a protocol for manual annotation of DXA artefacts. Second, I propose an automatic quality control framework to identify and localise DXA artefacts in large-scale clinical datasets. I tested the proposed method on a subset of scans selected from the MRC-Hip study ($n = 1300$). The sensitivity and specificity are 81.82% and 94.12%, respectively.

The content of this chapter is adapted from the following publication:

Mohsen Farzi, Jose M. Pozo, Eugene McCloskey, J. Mark Wilkinson, and Alejandro F. Frangi, "Automatic Quality Control for Population Imaging: A Generic Unsupervised Approach," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI2016). Springer, pp. 291–299, 2016.

5.1 Introduction

Population imaging studies such as UK Biobank [137] provide large datasets containing prominent imaging components in addition to demographics and other relevant meta-information. Large imaging datasets offer new opportunities for the comprehensive characterisation of the population (see chapter 3), but equally pose new challenges. One main challenge is the ability to control the quality of scans automatically and accurately, given the heterogeneous nature of the image acquisitions. Although strict imaging protocols are deployed to ensure consistent high-quality scans, incidental artefacts are inevitable.

Subjective image quality assessment (IQA) using a human observer is the simplest approach. However, this approach is prone to error. Reliable subjective quality assurance of large-scale datasets would require a prohibitively large amount of time and expertise, making it unfeasible for use in practice. Therefore, an objective technique is preferred for automatic IQA.

Automatic IQA has been explored extensively over the last two decades in the multimedia signal processing community [166]. However, existing algorithms are not directly applicable to medical images mainly for three reasons. Frist, IQA algorithms generally quantify image quality in relationship to human visual perception, whereas in clinical applications, the image quality is defined based on how well the image serves for the intended purpose [167, 168]. For example, in [169, 170], the ability of a human observer to detect lesions in an image, rather than aesthetic considerations, defines the image quality.

Second, current IQA algorithms often require information about the anticipated types of distortions. This information helps to learn a set of relevant features specific to each artefact type. This approach would be practical in the multimedia signal processing community where a limited number of artefacts such as blurring, noise, and JPEG compression are often of interest. On the other hand, artefacts are much more diverse in medical imagery; they are often specific to the imaging modality, the acquisition protocol, or the organ system (cf. [171, 172]). Managing unknown incidental artefacts is a challenging restriction of large medical imaging cohorts. To address this challenge, developing algorithms that are *non-distortion-specific* or *general purpose* is an important consideration.

Third, following extracting relevant quality-aware features from each scan, current IQA algorithms often require supervision to map these features to the subjective human scores. In multimedia signal processing, various benchmarks such as LIVE [173] and TID2008 [174] datasets with manual quality scores

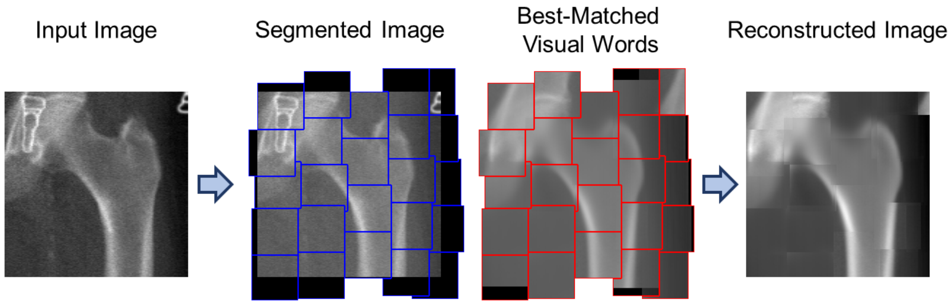


Figure 5.1: Image representation and visual word segmentation. The input image is divided into a set of overlapping image patches quantised to the best-matched visual words from a learned dictionary. Each visual word is a representative patch for a cluster of similar image patches. Visual words only account for frequent texture patterns; the abnormal key-shape object is eliminated in the reconstructed image.

are available for training purposes. However, manual quality annotations in large-scale medical imagery datasets are rare and their creation would require extensive and tedious visual assessment. Therefore, development of an *opinion free* algorithm without access to any reference score would be of great interest.

This chapter proposes an unsupervised, opinion free, general purpose framework to detect and localise artefacts in large-scale medical imaging datasets. In this framework, a novel patch-based image representation technique is proposed to synthesise a reference image corresponding to each image in the dataset (Fig. 5.1). To this end, first, a dictionary of *visual words* is learned (section 5.2.2). Each visual word is an image patch representative that characterises a cluster of similar patches. Second, an optimal coverage algorithm is proposed to cover each image with a subset of visual words. This coverage segments each image into a number of (possibly) overlapping image patches paired with their corresponding visual words (section 5.2.3). Following the image representation, each image is compared against the corresponding reference image and a set of dissimilarity scores are computed. Pooling the computed scores of all the images in the dataset, a probability distribution is then established per each dissimilarity metric and artefacts are detected as outliers to the estimated distributions (Fig. 5.2; section 5.2.4).

The remainder of this chapter is organised as follows. Section 5.2 presents the new framework for automatic image quality control applicable to large-scale clinical studies. Section 5.3 reviews various types of artefacts observed in femoral DXA scans. Section 5.4 demonstrates the application of the proposed technique to identify artefacts in the setting of femoral DXA scans. Section 5.5 concludes this chapter.

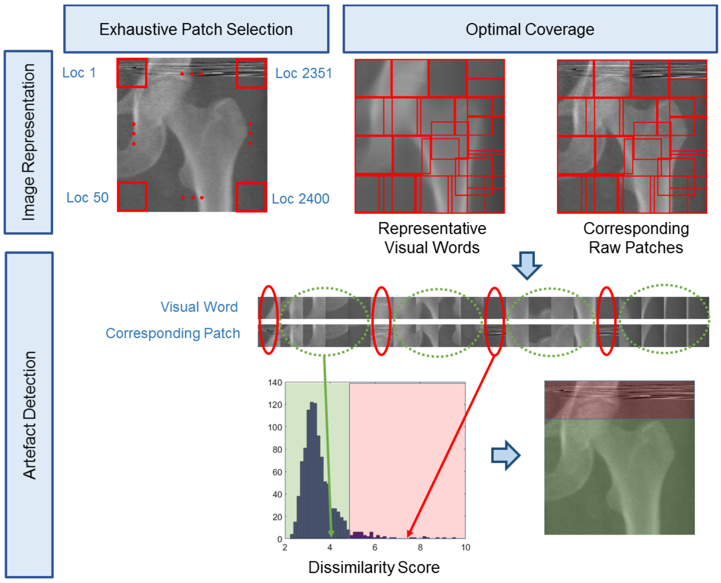

Exhaustive Patch Selection
Optimal Coverage
Image Representation
Loc 1
Loc 2351
Loc 50
Loc 2400
Representative Visual Words
Corresponding Raw Patches
Artefact Detection
Visual Word
Corresponding Patch
140
120
100
80
60
40
20
0
2
4
6
8
10
Dissimilarity Score

Figure 5.2: Unsupervised non-distortion-specific image quality framework. The proposed method works on image patches. In the first step, a dictionary of visual words is learned by grouping similar image patches together using the fixed-width clustering algorithm [175]. Each visual word is simply the centroid of each cluster. In the second step, each image is divided into a set of patches paired with their best-matched visual words from the dictionary. A set of dissimilarity scores is then computed per each pair. Finally, a probability distribution is established for each dissimilarity metric over the full dataset. Artefacts are detected as outliers of these distributions.

5.2 Unsupervised Non-Distortion-Specific Image Quality Control Framework

The proposed framework relies on three main assumptions. First, artefacts have a local nature by observing patches of an appropriate size albeit the extent of such patches could be, in the extreme case, the full image. Second, the image database is large enough so as to capture the statistics of the normal images and that of the artefacts of interest. Third, the incidence of artefacts in the image database should be small enough so that artefacts always remain outliers in the statistical distribution of the database. Under these assumptions, I propose an unsupervised, opinion free, non-distortion-specific framework to detect and localise artefacts in large-scale medical imaging datasets (Fig. 5.2).

The basic idea is to build a model representing the normal characteristics of the image population and detect the artefacts as deviations from this model. Based on the assumptions above, the proposed method works on image patches and comprises three main constituents: robust learning of a dictionary of image patches, an optimal coverage of the images with the learned visual words, and an assessment of the similarity between each covered patch and the corresponding visual word. This assessment allows us to detect outliers, identifying both images with artefacts and their locations in the image.

5.2.1 Background

The term *visual word* was coined in the object recognition literature [176] and later was deployed for image quality assessment [177]. Each visual word is simply a quantised representative for a cluster of similar image patches. Given a dictionary of visual words, each image is modelled as a distribution over the visual words by normalising the histogram of occurrence counts for each visual word. This technique is known as the *bag of visual words* (BOVW) representation in the literature.

The key assumption in the BOVW representation is that the frequency of each word in an image can discriminate between different image classes. However, the document frequency of each word, i.e. the frequency of each word in the whole dataset, can also play a role; the presence of words with a low document frequency in an image might be more informative than words that are quite common in the whole dataset. To address this issue, it is common to weight each bin of the histogram (, i.e. frequency of words in an image) by the inverse frequency of the words in the database in the well-known term frequency-inverse document frequency (tf-idf) scoring technique [178].

To detect image artefacts, however, only the presence of words with low document frequency matters rather than the frequency of each word in an image. However, unlike text documents where words are their intrinsic components, there is no straightforward and natural segmentation of images into visual words (patches). Alternatively, images are first parsed into a set of patches and, next, similar patches are clustered together. Therefore, learning words with low document frequency would be challenging in this approach. Here, only words with a high document frequency are learned as normal patches and then artefact patches are detected as a deviation from this normal representation.

5.2.2 Robust Dictionary Learning

The objective is to learn a dictionary $\mathcal{W} = \{w^1, \cdots, w^N\}$ with N visual words from a large pool of patches, capturing the normal shape and appearance variation in the image class while excluding outlier patches. An outlier patch is expected to lie in a sparse region of the feature space, i.e. raw intensity values here, having few or no neighbours within a typical distance r. Observe that outlier patches detected in this step cannot be used directly to identify image artefacts. Since images are not coregistered and patches are extracted from fixed locations, some proportion of outliers will be due to misalignment not necessarily representing an image artefact.

The proposed robust dictionary learning is as follows. Each image is divided into overlapping square patches of size k for 2D images, or cubic patches for 3D images, with an overlap of size k' between neighbouring patches. The fixed-width clustering algorithm is then applied as follows. All the patches are shuffled randomly and the first patch would be the first visual word. For all the subsequent patches, the Euclidean distance of the patch to each visual word is computed. If a distance is less than r, then the patch is added to the corresponding cluster and its centroid is recalculated with the average of all members. Otherwise, the patch is added as the centroid of a new cluster. Observe that each patch may contribute to more than one cluster. Finally, clusters with only one member are considered as outliers and removed from the dictionary.

5.2.3 Optimal Image Coverage

A coverage of an image I is a selection of visual words placed at different locations in the image so that all pixels are covered at least once. Let us consider that the image I has P pixels and each visual word can be deployed at L locations indexed by $\ell \in [1, L]$, where $L \leq P$ depends on the stride with which the image is swept. The binary mask m_ℓ represents the word location ℓ in the image, d_ℓ^n denotes the word w^n placed at location ℓ with appropriate zero-padding, and the binary variable z_ℓ^n encodes whether d_ℓ^n is selected to cover the image or not. Thus, the binary vector $\boldsymbol{z} = \left[z_1^1, \cdots, z_L^N\right]$ would represent a coverage of the image if

$$\sum_{n,\ell} z_\ell^n m_\ell \geq \mathbf{1}_{P\times 1}, \tag{5.1}$$

where the left-hand side is an integer vector counting how many times each pixel is covered in the image.

The image coverage error is the defined as the L_2-norm of the difference between each selected visual word and the corresponding image patch,

$$E = \sum_{n,\ell} z_\ell^n \left\| d_\ell^n - m_\ell \circ I \right\|^2 . \tag{5.2}$$

Here, $m_\ell \circ I$ denotes the component-wise product between the binary mask m_ℓ and the image I. The optimal image coverage is defined by the minimisation of the coverage error subject to the constraint in Eq. 5.1.

Let us denote by $z^*_\ell = \sum_n z^n_\ell$, the number of visual words placed at location ℓ. If two words, w^{n_1} and w^{n_2}, are used at the same location ℓ ($z^{n_1}_\ell = z^{n_2}_\ell = 1$), then the coverage error will be always larger than using just one of them, without any effect on the constraint. Hence, the optimal solution will place at each location ℓ either one single visual word ($z^*_\ell = 1$) or none ($z^*_\ell = 0$). Therefore, the optimisation can be split into two independent problems. First, for each ℓ, the locally optimal visual word $w^{n(\ell)}$ is selected by minimising the local error,

$$E_\ell = \min_n \|d^n_\ell - m_\ell \circ I\|^2. \tag{5.3}$$

Then, the optimal locations, $\boldsymbol{z}^* = (z^*_1, \cdots, z^*_L)$, are selected by minimising the total coverage error,

$$\boldsymbol{z}^* = \underset{*}{\operatorname{argmin}} \sum_\ell z^*_\ell E_\ell \qquad \text{subject to} \quad \sum_\ell z^*_\ell m_\ell \geq \mathbf{1}_{P\times 1}. \tag{5.4}$$

Eq. 5.4 can be efficiently solved using linear integer programming packages such as Matlab optimisation toolbox (Mathworks Inc, Cambridge, MA).

5.2.4 Artefact Detection

For a given image, a dissimilarity score is computed between each representative visual word and its corresponding image patch. Any image patch with an associated score above an optimal threshold identifies the location of an artefact in the given image. Observe that since matching of the words is local and the best fitting locations are found after an optimal coverage without forcing *a priori* known locations, images do not need to be previously registered.

(A) Dissimilarity score

The local properties of an image can be described by the set of its derivatives, which is named as *local jet* [179]. For a given image I and a scale σ, the local jet of order N at point $\boldsymbol{x}$ is defined as

$$J^N[I](\boldsymbol{x}, \sigma) \triangleq \{L_{i_1,\cdots,i_n}(\boldsymbol{x};\sigma)\}^N_{n=0}, \tag{5.5}$$

where the n^{th}-order derivative tensors are computed by the convolution

$$L_{i_1,i_2,\cdots,i_n}(\boldsymbol{x};\sigma) = \left[G^{(\sigma)}_{i_1,i_2,\cdots,i_n} * I\right](\boldsymbol{x}), \tag{5.6}$$

with the corresponding derivatives of the Gaussian kernel $G^{(\sigma)}_{i_1,i_2,\cdots,i_n}(\boldsymbol{x})$, and $i_k = 1,\ldots,D$, for D-dimensional images. For each derivative order, a complete set of independent invariants under orthogonal transformations can be extracted [179]. For 2D images and second order, for example, this complete set is formed by the intensity, the magnitude of gradients $\sqrt{\sum_i L_i^2}$, the Laplacian $\sum_i L_{ii}$, the Hessian norm $\sum_{i,j} L_{ij}L_{ji}$, and the second derivative along the gradient $\sum_{i,j} L_i L_{ij} L_j$. Multiresolution description can be obtained by changing the scale σ in a convenient set of scale-space representations. For each invariant feature, the Euclidean distance between the visual word and the corresponding image patch is used as the dissimilarity metric.

(B) Optimum threshold

The optimum threshold for each dissimilarity score is computed as follows. For each image in the database, the maximum score among all the representative visual words is computed. The optimum threshold is selected as $q_3 + \nu * (q_3 - q_1)$, where $\nu = 1.5$, and q_1 and q_3 are the first and third quartiles, respectively. An image is then artefact-free only if all the representative visual words have a dissimilarity score below the optimum threshold with respect to all the considered features.

5.3 DXA Artefacts

In bone densitometry using DXA, various types of artefacts may mask the true BMD measurement [171, 180, 181, 182, 37, 183]. Currently, proprietary software packages including Hologic Apex v3.2 (Hologic Inc, Waltham, MA) and Lunar enCORE v16 (GE Healthcare, Madison, WI) used for DXA analysis requires manual interaction and so each individual scan is analysed separately by an expert operator. Hence, artefacts are deemed to be identified online during the analysis step almost accurately. Unlike the common perception, however, a survey of members of the International Society for Clinical Densitometry (ISCD) indicated that errors in DXA acquisition are not rare [180]. Therefore, retrospective quality assessment would be a necessity in DXA analysis. For example, Beck et al. [150] excluded 7% of scans (i.e., 1031 scans out of 14,646) from their proposed hip structural analysis due to various DXA artefacts such as obscured femoral neck margins, incomplete scans, osteoarthritic changes, metal artefacts, prosthetic or calcification, and excessive anteversion.

DXA artefacts can be broadly classified into 3 categories depending on

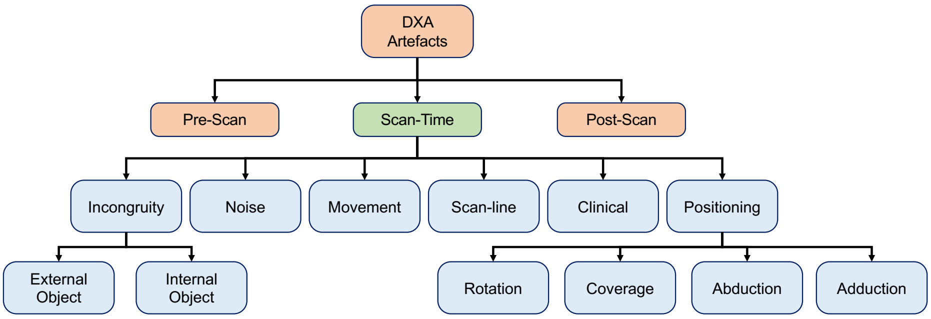


Figure 5.3: Various types of femoral DXA artefacts. DXA artefacts can be broadly classified into three categories: errors due to poor calibration of the instrument prior to scan collection, imaging artefacts during scan acquisition, and errors due to poor analysis following the scan collection. Imaging Artefacts can be further divided into six categories.

the chronological order they may happen (Fig. 5.3). First, prior to scan acquisition, the instrument should be calibrated against a phantom to ensure accurate BMD measurement. This is of great importance in multi-centre studies or longitudinal analysis of individual patients. Second, during scan acquisition, errors may occur due to a fault in the instrument (, e.g., scan-line error), patient (, e.g., metal objects in clothing), or operator (, e.g., poor patient positioning). These artefacts are visible in the collected images and can be identified retrospectively. Third, following scan collection, errors may still happen during the analysis step, e.g. poor segmentation of bone tissue. Software packages for DXA analysis provide some degrees of automation for bone analysis; however, subtle errors should be identified and corrected manually by the operator. For example, the operator should confirm whether the neck ROI is located correctly or not in the hip scan analysis.

The aim of this study is to identify artefacts during the image acquisition step. No standard guideline exists for classification of femoral DXA artefacts. Lack of a standard protocol limits consistent screening for DXA artefacts as different experts may hold different opinions of DXA artefacts. This diversity in opinions may originate from two different perspectives: first, the relative importance of various imaging artefacts in DXA analysis has been rarely studied in the literature and thereby, in extreme, one could be convinced that an error is not indeed an artefact. Second, DXA scans may be deployed for various purposes and so defining artefacts based on the intended purpose could be ambiguous. For example, errors affecting BMD measurement at the trochanter may not be considered as artefacts as long as neck BMD is only of interest. Recognising these challenges, here, I propose to classify artefacts observed in femoral DXA scans into 6 groups following visual assessment of a large number of DXA scans, reading the literature, and consulting with radiology experts in the field (Fig. 5.3).

1. Incongruous objects: This category would capture any localised incongruity in texture patterns that may not represent either bone or soft tissue. Artefacts of this class are often characterised as too bright or too dark spots in the image (Fig. 5.4(a)). Taking into account the potential sources of error, this class could be further divided into three subcategories:

    - External: metal objects in or on clothing, buttons, plastic materials in the pocket (e.g., credit card), sticking plasters.
    - Internal: Surgical clips, implants, body piercing, pacemaker lead.
    - Pharmaceutical: nuclear medicine (e.g., Barium examination), injection of contrast media (e.g., Myelographic contrast agent).

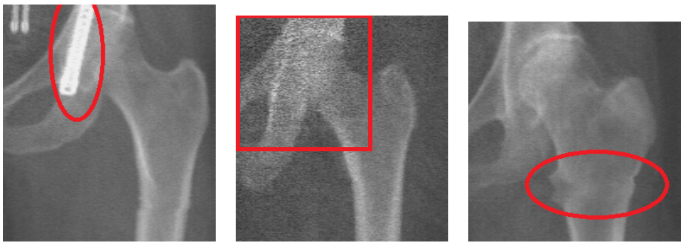
2. Noise: This artefact is characterised either by extra granularity patterns or a dark shadow over the pelvic region (Fig. 5.4(b)). This artefact is often seen in obese subjects and can be attributed to inhomogeneous fat distribution around the bone.

3. Movement: Despite leg fixation during scan acquisition, incidental movements could affect the BMD measurement. This artefact is characterised by an interruption in the continuity of bone tissue that makes the picture looks fuzzy (Fig. 5.4(c)).

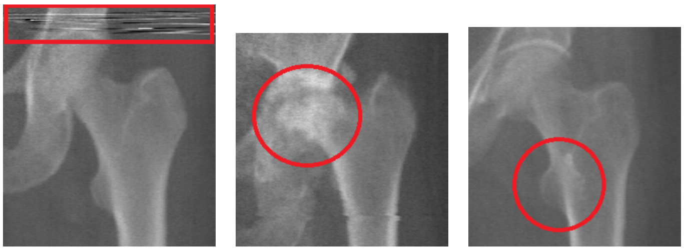
4. Scan-line: This artefact is characterised by distinct horizontal lines with random values in the image, and is attributed to a fault in the scanner (Fig. 5.4(d)).

5. Clinical: This class would capture any medical disorder that may affect the BMD measurement in hip. For example, Enostosis (bone islands), hip osteoarthritis (see Fig. 5.4(e)), Paget disease, calcific tendinitis, vascular calcification, avascular necrosis, developmental dysplasia of the hip, etc [171, 181].

6. Patient positioning: Since DXA is a 2D projection, inconsistency in patient positioning could result in variation in BMD measurements. DXA manufacturers advise a standard protocol for positioning patients in the scanner. Any deviation from this standard positioning is then defined as artefact [37, 183]. This category can be further classified into 4 subcategories:



(a) Metal object on clothing **(b)** Noise **(c)** Movement



(d) Scan-line **(e)** Osteoarthritis **(f)** Rotation

Figure 5.4: Examples of various imaging artefacts in DXA. The red contour overlaid the image shows the location of each artefact.

- Rotation: In this group, the leg is rotated from the standard position either internally or externally. The amount of lesser trochanter observed in the scan may reflect the rotation artefact; if rotated externally (internally) too much (less) of lesser trochanter would be observed. However, due to anatomical variation between subjects, it is difficult to visually assess the presence of this artefact (see Fig. 5.4(f)).
- Coverage. In this group, the scan does not capture the intended ROI advised by DXA manufacturers either by including more parts than the standard or missing a few parts.
- Adduction. The limb is moved toward the mid-line of the body.
- Abduction. The limb is moved away from the mid-line of the body.

5.4 Experiments and Results

(A) Dataset

A subset of 300 hip DXA scans selected randomly from the MRC-Hip study [139] were manually annotated for the following artefacts: incongruous object, noise, scan-line, and movement. These annotations were used as the ground-truth labels for evaluation of the proposed quality control framework. Another random selection of 1000 scans from the same study was used to learn the visual words and to set the optimum thresholds for artefact detection.

(B) Parameter selection

The patch size and the radius r are two parameters for the proposed method. Both parameters would be data dependent. The radius r is automatically selected estimating a typical small distance between patches in the same image: For each image, all pairwise distances between the patches comprising the image are computed. Next, $\frac{1}{n}$-quantile of these distances are computed per image, where n is the total number of patches extracted from each image. Then, the parameter r is selected as the median of the computed quantiles in the image dataset. The patch size could be estimated based on the size of the effect that is measured. For example, in femoral DXA bones, the diameter of the femoral stem is approximately 64 pixels. I have tested the results with patches of size 32 and 64 with 8 pixel overlap. No differences were observed in the sensitivity. I presented the results with patches of size 64. In summary, the total number of 24830 patches were extracted from 1000 images. The radius $r = 3.5$ is estimated for this dataset. The obtained dictionary contained 1146 visual words.

I tested invariant features up to the second order. However, the second order features did not provide any new information. Hence, only intensity and gradient magnitude were finally used as the features. The gradient magnitude for each image patch or visual word is normalised to have Euclidean norm one. Single scale analysis with $\sigma = 0.2$ was used. Optimum thresholds are derived as 0.37 and 4.86 for the gradient magnitude and intensity, respectively.

(C) Results

Sensitivity and specificity of the method are reported on the test data based on a priori manual annotation. The sensitivity is defined as the proportion of images with artefacts that are detected correctly by the algorithm. The specificity is defined as the proportion of normal images that are correctly labelled as artefact-free.

Eleven images out of 300 were manually identified with artefacts. Nine out of eleven are detected using the proposed algorithm. Sensitivity and specificity are 81.82% and 94.12%, respectively. Fig. 5.5 shows normal images and artefacts. Only 2 out of 11 image artefacts are misclassified as normal. These two scans are characterised as movement artefacts that cause subtle vertical displacement in the image. However, the algorithm managed to successfully localise other types of artefacts including the existence of an external object (key-shape object in Fig. 5.5).

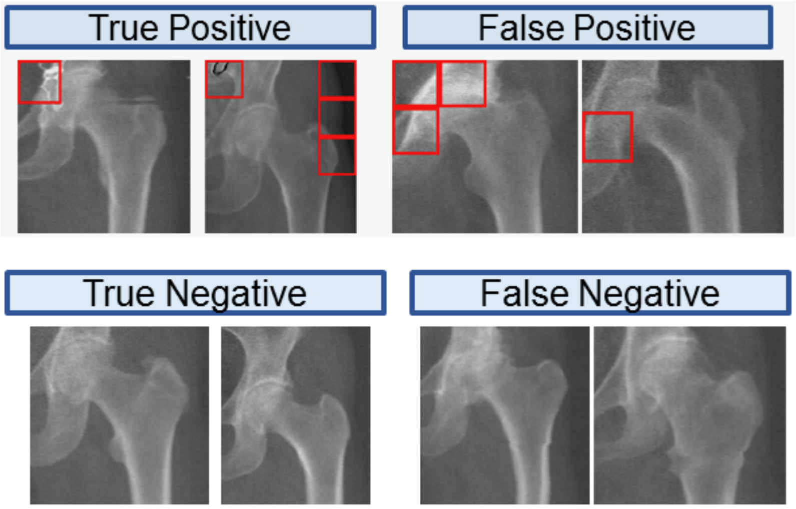


Figure 5.5: Examples of successful and unsuccessful DXA artefact detection using the proposed algorithm. The red square, if present in the image, shows the location of detected artefacts.

5.5 Conclusion

This chapter presented the development of an unsupervised and non-distortion-specific framework to address automatic quality control in large medical imagery datasets. Based on the assumption that artefacts constitute a small proportion of the dataset, a dictionary-based framework based on an optimal coverage of images was introduced to detect and localise image artefacts as outliers to the normal image class. The method computational complexity is linear in the number of input images, providing good scalability to large datasets. I have tested the method on 300 femoral DXA scans and reported good sensitivity and specificity on the dataset.

Chapter 6

Conclusions and Future Work

6.1 Overview

Osteoporosis is an age-associated disease leading to deteriorated bone quality with elevated fracture risk. Despite the progressive nature of osteoporosis, currently, the degree of progression or the severity of disease is not intuitively estimated in clinical practice. Bone status is categorised as *normal*, *osteopenia*, or *osteoporosis* based on the bone mineral density (BMD) measurement at femoral neck or spine reported as T-score, i.e. the number of standard deviation (SD) from the average for young healthy women in the population. The world health organisation (WHO) criteria for the diagnosis of osteoporosis is based on the T-score cut-off point of -2.5 for osteoporosis and -1 for osteopenia. Setting a clear cut-off point could be misleading as half of the fragility fractures at most could be attributed to osteoporotic bones with this definition [15]. Although the population attributable risk (PAR) of 50% is still high when compared with a PAR of 10% between hypertension and congestive heart failure [184], the ability to provide new metrics capturing the bone strength beyond neck BMD remains of interest in osteoporosis research. With these limitations in mind, the overall aim of this study was to model the evolution of spatial BMD distribution within the proximal femur as a function of age. Quantification of age-related architectural changes in the femur would allow the introduction of new indices capturing the gradual deterioration in bone quality with osteoporosis progression.

In Chapter 1, I reviewed three different perspectives on bone quality assessment. In the first approach, assuming fragility fractures are the important outcome rather than the bone strength *per se*, other clinical risk factors together with neck BMD were deployed to improve fracture prediction. In the second approach, defining bone strength as the maximum stress that bone can sustain before breaking, finite element analysis was used to model the stress-strain relationship in bones. In the third approach, given that DXA scans can provide much more information on bone structural properties rather than neck BMD, various techniques were deployed to extract useful geometrical structures or spatial BMD patterns in the setting of osteoporosis management. In line with the third approach, an elegant framework called DXA region free analysis (RFA) was reviewed. In DXA RFA, BMD maps are warped into a reference template to eliminate morphological variation and so an anatomical correspondence between pixel coordinates is established. This pixel correspondence allowed application of statistical tests to quantitate localised remodelling events. However, the initial RFA framework presented in [36] did not account for the known *multiple comparisons problem*.

In Chapter 2, I reviewed various techniques to address the multiple comparisons problem, including the Bonferroni correction, random field theory, and false discovery rate analysis. This chapter presented the integration of false discovery rate analysis with DXA RFA to quantitate the magnitude and areal size of periprosthetic BMD changes using scans acquired during two previous randomised clinical trials.

In Chapter 3, I explored the relationship between osteoporosis and ageing, as understanding the mechanism by which bone gets weaker with ageing might open new perspectives to estimate osteoporosis progression over a continuum. One main contribution of this thesis is to develop the first spatio-temporal atlas of bone ageing in the femur using over 13,000 Caucasian women aged 20-97 years from North Western Europe. Development of a comprehensive model for involutional bone loss is a challenging task: first, a sufficiently large sample size is required to represent the variation in the population; second, a deformable image warping is required to quantify spatial BMD variation at anatomically correspondent sites. To comprise a large-scale dataset capturing the whole adulthood age range, data was integrated from three previous studies: UK Biobank (n=6,918; age range=45-80 years), OPUS (n=1,402; age range=20-40 and 55-79 years), and MRC-Hip (n=5,018; age range=75-97 years). Since a systematic difference in BMD measurement exists between DXA manufacturers, I also proposed and internally validated a new quantile matching regression technique for comparative calibration between different vendors. The new technique applies to large-scale datasets where no repeated measurements of the same subject on different machines would be available, but instead relied on population-based BMD distribution data for calibration. To quantify the BMD variation at anatomically correspondent locations, I extended the region free analysis (RFA) technique developed for peri-prosthetic BMD analysis to a fully automatic framework applicable to the native femur.

In Chapter 4, I explored how the developed atlas usefully adds to our understanding of bone ageing patterns in the femur. The delineation between cortical and trabecular arrangements of bone in the femur was possible with the enhanced RFA resolution. For example, the average cortical thickness was linearly decreased with ageing from 6.65 mm at 20 years to 5.55 mm at 80 years. To reflect the overall effect of ageing on trabecular microarchitecture, *bone age* was introduced as age at which the median bone map best fits the given individual bone map. To find this optimised age, the ℓ_2-norm between the bone map for the given individual and the median map from the atlas at that age was minimised. Given ageing as the underlying mechanism for osteoporosis, here I assumed that all individuals follow a unique bone ageing trajectory but with different speeds. An accelerated loss in BMD is then attributed to osteoporosis.

Promising properties demonstrated by bone age may suggest potentials for better management of osteoporosis using the new metric. First, bone age was highly correlated with neck BMD ($r=-0.87$; $p<0.001$) suggesting consistency with current diagnostic guidelines. Second, bone age was precise; the standard deviation of error estimated on 25 pairs of scans collected on the same day was 1.4 years. Third, its ability to differentiate between a young healthy population ($n = 284$; age<40 years) and a more elderly cohort ($n=2165$; age>80 years) was found to be 5% higher than neck BMD alone ($p<0.001$) measured as the area under the curve (AUC) using receiver operating characteristic (ROC) analysis. Fourth, the risk of sustaining a follow-up fragility fracture increased consistently with bone ageing. Fifth, its ability to predict hip fractures measured as the area under the ROC curve was similar to other metrics including neck BMD and FRAX. When normalising BMD maps for their bone age, localised fracture-specific patterns were observed at the superior femoral neck extended to the trochanter with the same orientation as principal tensile curves. These patterns could not be observed using the raw BMD maps suggesting that excluding the ageing effect may lead to the identification of new scores that can enhance fracture prediction independent of neck BMD or bone age. In this study, I proposed a new metric called f-score by averaging normalised pixel BMD values over the identified localised fracture-specific regions. Integration of f-score and bone age significantly enhanced the AUC by 3% over neck BMD alone.

Population imaging studies such UK Biobank study [137] have opened new opportunities for a comprehensive characterisation of a range of diseases including osteoporosis. However, this poses new emerging challenges in the field. One main challenge is the ability to assess the quality of scans automatically and accurately in large-scale datasets. In Chapter 5, I presented a novel unsupervised, automatic, non-distortion-specific framework for quality control of DXA scans in large-scale studies. The proposed method was tested on a subset of scans selected from the MRC-Hip study ($n = 1300$). The sensitivity and specificity were 81.82% and 94.12%, respectively. This technique is the first step forward toward development of automatic quality control frameworks for DXA images.

6.2 Thesis Contributions

- Integrated FDR analysis to the RFA framework to localise significant bone remodelling events (chapter 2).

- Extended the RFA technique to a fully automatic framework applicable to the native femur for quantification of spatial BMD distribution (chapter 3).

- Developed the first spatio-temporal atlas of bone ageing in the femur with over 13,000 Caucasian women from North Western Europe aged between 20-97 years (chapter 3).

- Proposed a new quantile matching technique for comparative calibration between different DXA manufacturers when multiple measurements of the same subject on different machines are not available (chapter 3).

- Extended the notion of disease progression estimation to osteoporosis by introducing bone age as the age at which bone microarchitecture best fits the developed atlas (chapter 4).

- Characterised local fracture-specific patterns in the femur using enhanced RFA resolution analysis and the introduction of a new metric called f-score with the power to predict fractures independent of neck BMD (chapter 4).

- Developed the first automatic quality control framework of femoral DXA scans (chapter 5).

6.3 Limitations

DXA RFA also has limitations. Although aggregating pixel BMD in *a priori* identified regions of interest (ROIs) results in similar precision to conventional DXA analysis, the pixel level precision in RFA is worse. This is a compromise that offers enhanced resolution analysis. To further improve the pixel level precision, more advanced registration algorithms would be utilised to warp each scan to the template domain. More specifically, the thin plate spline (TPS) technique does not account for uncertainty in the selection of controlling landmark points. Langevin equations may be deployed to address this issue as discussed by Marsland and Shardlow [185]. This would specifically help with the low RFA precision at the bone margins. Another limitation of the TPS technique is that no intensity information is utilised during the image alignment. Due to the random rotation of DXA scans in the population, relying on the intensity variation for image registration could be misleading. Therefore, the TPS was originally chosen for the registration step; however, Kuhnel and colleagues [186] recently proposed a framework to separate deformation from intensity changes in non-rigid registration. Application of this framework would allow benefiting from both warp and intensity variation available in the dataset.

Second, RFA removes potentially useful geometrical properties by aligning all scans to a reference template. This approach, known as voxel-based analysis (VBA), has been widely used in medical imaging community to compare between two groups (e.g., patients and controls [187]) or identifying regions where the disease severity would correlate most with the image features [188]. In this work, RFA resulted in promising results for characterising age-related spatial BMD variation as well as localising fracture-specific patterns in the femur. However, for a comprehensive analysis of bone structural properties, one should explore morphological variation between scans as well. Modelling morphological variation with ageing has not been addressed in the femur yet, but this has been explored in the brain to model the severity of Alzheimer's disease [112]. The method proposed in [112] potentially could also be applied to the femur, but it does not address two main challenges specific to the femoral DXA scans. To model the ageing impact on the femoral morphology, one should first exclude the variation due to poor patient positioning and image magnification in DXA scans.

Third, DXA RFA, despite enhanced spatial resolution analysis, can only offer an approximate reflection on bone structural properties. For a detailed analysis of micro-architectural properties, other advanced imaging techniques including Quantitative Computed Tomography (QCT) [157] and high-resolution peripheral QCT (HR-pQCT) [156] should be utilised in practice. While these techniques provide enhanced spatial resolution comparing with DXA, their application in clinical practice is unlikely in the foreseeable future [18]. On the other hand, DXA is an inexpensive, widely available clinical tool with extremely low radiation exposure. Given that DXA RFA can be applied retrospectively to the original DXA measurements, it has the potential to transform the conventional DXA analysis in future.

6.4 Future Work

The work presented in this thesis has opened new perspectives for understanding the relationship between osteoporosis and the arrangements of trabecular microarchitecture in the proximal femur leading to weaker bones in the elderly population. The proposed framework in this thesis constitutes a first step toward modelling osteoporosis progression to identify better bone-based risk factors for prediction of fragility fractures. This study could potentially contribute to the future research on osteoporosis as discussed below.

6.4.1 Extending the Bone Ageing Atlas

Chapter 3 presented a fully automatic pipeline to streamline the bone ageing analysis. The technique was deployed to generate the atlas for Caucasian women in the proximal femur, but the automated pipeline will facilitate population-specific atlas generation from other ethnic libraries, gender, and anatomic sites.

Since osteoporosis is more prevalent among women than men, the primary concern of this study was women. The same pipeline, however, could be deployed to analyse bone density patterns in men. To construct the ageing atlas for men, a primary challenge would be the collection of a large-scale dataset capturing the adulthood age range. Although the ability of the proposed pipeline for retrospective analysis of DXA scans avoids the necessity for new data collection, clinical studies with a focus on men are still limited. To mention a few ones, Mr Os is an international multicenter study in the United States, Hong Kong, and Sweden with men aged 69 to 80 years [189].

Although age was the primary variable of interest in this study, extending the atlas for other covariates is possible. In chapter 4, I demonstrated how variation due to body mass index (BMI) could be modelled using the flexibility of vector generalised additive models (VGAM) integrated into the pipeline. Other risk factors to be included in the model could include genetic variability between individuals, drug medications, smoking habits, alcohol consumption, physical activity, etc. While the current framework can be extended for additional covariates, missing data could be a new emerging challenge; the more covariates to be included, the more cases with incomplete explanatory variables. Further developments would be required to address the missing data issue in large-scale datasets.

Among all types of fragility fractures, hip fractures are the most menacing complication of osteoporosis with associated disability burden [190]. Therefore, this thesis addressed the development of an ageing atlas in the proximal femur as the first step. However, it is possible to extend the framework to other anatomic sites including the spine. Given the different anatomy of the spine in comparison to the femur, the automatic landmark selection step in the proposed pipeline should be extended to account for spine scans.

6.4.2 Osteoporosis Progression Model

Although disease progression is a new concept in osteoporosis introduced in this thesis, progression estimation has been established as a useful practice in other age-associated diseases such as dementia [164]. In this study, the actual progression on the median ageing trajectory in the original high-dimensional space defined osteoporosis progression. To this end, each subject was mapped into the ageing trajectory using a simple L_2-norm. This similarity metric has two limitations: first, it does not account for the variation in BMD distribution around the median between pixel coordinates. To account for this variation, other metrics such as Mahalanobis distance could be used. This is also known as the generalised least squares minimisation problem in the literature. Second, it does not account for the variation in uncertainty in the estimation of the ageing trajectory at different ages. To account for this variation, a Bayesian framework may be deployed to estimate the bone age as the age with a maximum *a posteriori* probability.

Since only baseline measurements were available in this study, the relation between fracture incidents and the rate of osteoporosis progression has not be analysed. Given longitudinal data measurements, computation of progression rates may lead to a potential added value for fracture prediction beyond the estimated baseline bone status.

6.4.3 Predicting Local Fracture Patterns

Normalising spatial BMD patterns for bone age would exclude ageing effect allowing identification of local fracture patterns in the residual BMD map. I have demonstrated in chapter 4 that these local patterns could enhance fracture prediction beyond bone age or neck BMD alone. Hip fractures may happen either at the subcapital neck, transcervical neck, intertrochanter, subtrochanter, greater trochanter, or lesser trochanter (Fig. 6.1). Knowing the type of fracture, the RFA technique would allow correlating this information with the residual BMD maps to characterise local texture patterns with added predictive value for fracture prediction. This would lead to a more efficient analysis of contained information in BMD maps that could further improve the fracture risk assessment in the setting of osteoporosis disease.

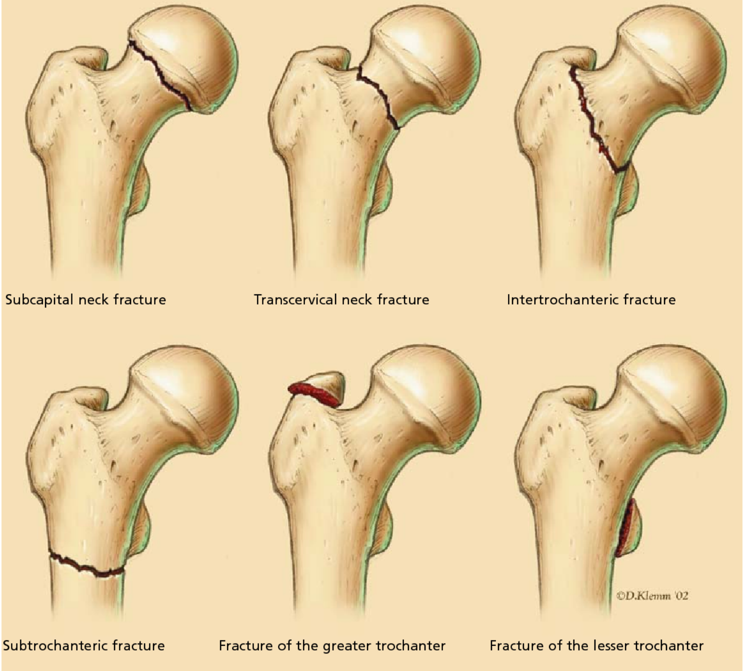

Subcapital neck fracture
Transcervical neck fracture
Intertrochanteric fracture
Subtrochanteric fracture
Fracture of the greater trochanter
Fracture of the lesser trochanter
©D.Klemm '02

Figure 6.1: Basic types of fracture patterns in the proximal femur. A single pattern or a combination of these patterns may be observed in practice. DXA RFA would allow correlation of local BMD patterns with actual fracture patterns in the femur to enhance hip fracture prediction. The image is adapted from [165].

6.5 Conclusion

In this thesis, I developed a region free analysis framework for quantification of pixel BMD at anatomically corresponding locations in the proximal femur. The evolution of spatial BMD distribution was modelled as a function of age. This is, to the best of my knowledge, the first spatio-temporal atlas of BMD ageing in the femur. Bone age was introduced to reflect the overall spatial BMD patterns, estimating the actual progression on the median ageing trajectory in the femur. Given the promising properties of bone age, I proposed its potential to estimate osteoporosis progression as a continuum rather than conventional discrete categories, i.e. normal, osteopenia, and osteoporosis. A new index, called f-score, was introduced to characterise local fracture-specific patterns. Bone age together with f-score demonstrated a statistically significant improvement in hip fracture prediction for the elderly population. Given access to large-scale DXA datasets, automatic quality control of DXA scans has been identified as an emerging challenge to the community for which an unsupervised, non-distortion-specific, opinion-free quality control framework was proposed. The work presented in this thesis has opened new perspectives for better understanding of localised spatial BMD patterns and their relationship with osteoporosis. Future research into the relationship between these spatial patterns and various fracture types in the femur may further enhance hip fracture prediction in the elderly population.

Bibliography

[1] D. Chappard, M. Baslé, E. Legrand, and M. Audran, "Trabecular bone microarchitecture: a review," *Morphologie*, vol. 92, no. 299, pp. 162–170, 2008.

[2] P. Meunier and G. Boivin, "Bone mineral density reflects bone mass but also the degree of mineralization of bone: therapeutic implications," *Bone*, vol. 21, no. 5, pp. 373–377, 1997.

[3] T. D. Rachner, S. Khosla, and L. C. Hofbauer, "Osteoporosis: now and the future," *The Lancet*, vol. 377, no. 9773, pp. 1276–1287, 2011.

[4] J. A. Kanis, E. V. McCloskey, H. Johansson, A. Oden, L. J. Melton, and N. Khaltaev, "A reference standard for the description of osteoporosis," *Bone*, vol. 42, no. 3, pp. 467–475, 2008.

[5] J. Kanis, D. Black, C. Cooper, P. Dargent, B. Dawson-Hughes, C. De Laet, P. Delmas, J. Eisman, O. Johnell, B. Jonsson, *et al.*, "A new approach to the development of assessment guidelines for osteoporosis," *Osteoporosis International*, vol. 13, no. 7, pp. 527–536, 2002.

[6] R. Eastell, "Osteoporosis," *Medicine*, vol. 37, no. 9, pp. 475–480, 2009.

[7] R. S. Braithwaite, N. F. Col, and J. B. Wong, "Estimating hip fracture morbidity, mortality and costs," *Journal of the American Geriatrics Society*, vol. 51, no. 3, pp. 364–370, 2003.

[8] J. Kanis and O. Johnell, "Requirements for DXA for the management of osteoporosis in Europe," *Osteoporosis International*, vol. 16, no. 3, pp. 229–238, 2005.

[9] J. A. Kanis, "Assessment of osteoporosis at the primary health care level. (Technical report)," *WHO Collaborating Centre for Metabolic Bone Diseases*, 2007.

[10] D. Marshall, O. Johnell, and H. Wedel, "Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures," *BMJ*, vol. 312, no. 7041, pp. 1254–1259, 1996.

[11] S. Schuit, M. Van der Klift, A. Weel, C. De Laet, H. Burger, E. Seeman, A. Hofman, A. Uitterlinden, J. Van Leeuwen, and H. Pols, "Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam study," *Bone*, vol. 34, no. 1, pp. 195–202, 2004.

[12] S. R. Cummings, D. B. Karpf, F. Harris, H. K. Genant, K. Ensrud, A. Z. LaCroix, and D. M. Black, "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, 2002.

[13] "Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: report of a WHO study group," Geneva: World Health Organization, 1994.

[14] D. Black, M. Steinbuch, L. Palermo, P. Dargent-Molina, R. Lindsay, M. Hoseyni, and O. Johnell, "An assessment tool for predicting fracture risk in postmenopausal women," *Osteoporosis International*, vol. 12, no. 7, pp. 519–528, 2001.

[15] K. L. Stone, D. G. Seeley, L.-Y. Lui, J. A. Cauley, K. Ensrud, W. S. Browner, M. C. Nevitt, and S. R. Cummings, "BMD at multiple sites and risk of fracture of multiple types: long-term results from the study of osteoporotic fractures," *Journal of Bone and Mineral Research*, vol. 18, no. 11, pp. 1947–1954, 2003.

[16] E. S. Siris, P. D. Miller, E. Barrett-Connor, K. G. Faulkner, L. E. Wehren, T. A. Abbott, M. L. Berger, A. C. Santora, and L. M. Sherwood, "Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal women: results from the national osteoporosis risk assessment," *JAMA*, vol. 286, no. 22, pp. 2815–2822, 2001.

[17] J. A. Kanis, A. Oden, H. Johansson, F. Borgström, O. Ström, and E. McCloskey, "FRAX and its applications to clinical practice," *Bone*, vol. 44, no. 5, pp. 734–743, 2009.

[18] J. E. Adams, “Advances in bone imaging for osteoporosis,” *Nature Reviews Endocrinology*, vol. 9, no. 1, pp. 28–42, 2013.

[19] T. M. Link, “Osteoporosis imaging: state of the art and advanced imaging,” *Radiology*, vol. 263, no. 1, pp. 3–17, 2012.

[20] D. Felsenberg and S. Boonen, “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, 2005.

[21] J. Currey, “Bone strength: what are we trying to measure?,” *Calcified Tissue International*, vol. 68, no. 4, pp. 205–210, 2001.

[22] T. J. Beck, C. B. Ruff, K. E. Warden, J. W. Scott, and G. U. Rao, “Predicting femoral neck strength from bone mineral data. a structural approach.,” *Investigative Radiology*, vol. 25, no. 1, pp. 6–18, 1990.

[23] T. J. Beck, “Extending DXA beyond bone mineral density: understanding hip structure analysis,” *Current Osteoporosis Reports*, vol. 5, no. 2, pp. 49–55, 2007.

[24] L. Lenaerts and G. H. Van Lenthe, “Multi-level patient-specific modelling of the proximal femur. a promising tool to quantify the effect of osteoporosis treatment,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1895, pp. 2079–2093, 2009.

[25] P. Zioupos and J. Currey, “Changes in the stiffness, strength, and toughness of human cortical bone with age,” *Bone*, vol. 22, no. 1, pp. 57–66, 1998.

[26] R. W. Goulet, S. A. Goldstein, M. J. Ciarelli, J. L. Kuhn, M. Brown, and L. Feldkamp, “The relationship between the structural and orthogonal compressive properties of trabecular bone,” *Journal of Biomechanics*, vol. 27, no. 4, pp. 375–389, 1994.

[27] K. G. Faulkner, S. R. Cummings, D. Black, L. Palermo, C.-C. Glüer, and H. K. Genant, “Simple measurement of femoral geometry predicts hip fracture: the study of osteoporotic fractures,” *Journal of Bone and Mineral Research*, vol. 8, no. 10, pp. 1211–1217, 1993.

[28] L. Pothuaud, P. Carceller, and D. Hans, “Correlations between grey-level variations in 2D projection images (TBS) and 3D microarchitecture: applications in the study of human trabecular bone microarchitecture,” *Bone*, vol. 42, no. 4, pp. 775–787, 2008.

[29] W. R. Crum, T. Hartkens, and D. Hill, “Non-rigid image registration: theory and practice,” *The British Journal of Radiology*, vol. 77, no. 2, pp. 140–153, 2004.

[30] J. Ashburner, “Computational anatomy with the SPM software,” *Journal of Magnetic Resonance Imaging*, vol. 27, no. 8, pp. 1163–1174, 2009.

[31] G. M. Blake, H. W. Wahner, and I. Fogelman, *Evaluation Of Osteoporosis.* Taylor & Francis, 1998.

[32] N. J. Crabtree, M. B. Leonard, and B. S. Zemel, “Dual-energy x-ray absorptiometry,” in *Bone Densitometry in Growing Patients*, ch. 3, pp. 41–57, Springer, 2007.

[33] T. Wu and W. Clarke, “Hologic bone densitometry and the evolution of DXA,” *Hologic Inc*, 2012.

[34] E. Berry, J. Truscott, S. Stewart, and M. Smith, “Spatial distribution of femoral bone mineral in dual energy x-ray absorptiometry images: a possible technique to improve discrimination between normal and osteoporotic patients,” *The British Journal of Radiology*, vol. 69, no. 824, pp. 743–750, 1996.

[35] X. N. Dong, R. Pinninti, T. Lowe, P. Cussen, J. E. Ballard, D. Di Paolo, and M. Shirvaikar, “Random field assessment of inhomogeneous bone mineral density from DXA scans can enhance the differentiation between postmenopausal women with and without hip fractures,” *Journal of Biomechanics*, vol. 48, no. 6, pp. 1043–1051, 2015.

[36] R. M. Morris, L. Yang, M. A. Martín-Fernández, J. M. Pozo, A. F. Frangi, and J. M. Wilkinson, “High-spatial-resolution bone densitometry with dual-energy x-ray absorptiometric region-free analysis,” *Radiology*, vol. 274, no. 2, pp. 532–539, 2015.

[37] S. Lekamwasam, R. Sumith, and J. Lenora, “Effect of leg rotation on hip bone mineral density measurements,” *Journal of Clinical Densitometry*, vol. 6, no. 4, pp. 331–336, 2003.

[38] O. Ahmad, K. Ramamurthi, K. E. Wilson, K. Engelke, R. L. Prince, and R. H. Taylor, "Volumetric DXA (VXA): a new method to extract 3D information from multiple in vivo DXA images," *Journal of Bone and Mineral Research*, vol. 25, no. 12, pp. 2744–2751, 2010.

[39] T. Whitmarsh, L. Humbert, M. De Craene, L. M. Del Rio Barquero, and A. F. Frangi, "Reconstructing the 3D shape and bone mineral density distribution of the proximal femur from dual-energy x-ray absorptiometry," *IEEE Transactions on Medical Imaging*, vol. 30, no. 12, pp. 2101–2114, 2011a.

[40] T. Whitmarsh, K. D. Fritscher, L. Humbert, L. M. Del-Rio-Barquero, R. Schubert, and A. F. Frangi, "Hip fracture discrimination using 3D reconstructions from dual-energy x-ray absorptiometry," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 1189–1192, IEEE, 2011b.

[41] A. M. Baker, D. W. Wagner, B. J. Kiratli, and G. S. Beaupre, "Pixel-based DXA-derived structural properties strongly correlate with pQCT measures at the one-third distal femur site," *Annals of Biomedical Engineering*, vol. 45, no. 5, pp. 1247–1254, 2017.

[42] S. Goodyear, R. Barr, E. McCloskey, S. Alesci, R. Aspden, D. Reid, and J. Gregory, "Can we improve the prediction of hip fracture by assessing bone structure using shape and appearance modelling?," *Bone*, vol. 53, no. 1, pp. 188–193, 2013.

[43] J. Michelotti and J. Clark, "Femoral neck length and hip fracture risk," *Journal of Bone and Mineral Research*, vol. 14, no. 10, pp. 1714–1720, 1999.

[44] M. Peacock, C. Turner, G. Liu, A. Manatunga, L. Timmerman, and C. Johnston, "Better discrimination of hip fracture using bone density, geometry and architecture," *Osteoporosis International*, vol. 5, no. 3, pp. 167–173, 1995.

[45] S. Boonen, R. Koutri, J. Dequeker, J. Aerssens, G. Lowet, J. Nijs, G. Verbeke, E. Lesaffre, and P. Geusens, "Measurement of femoral geometry in type I and type II osteoporosis: differences in hip axis length consistent with heterogeneity in the pathogenesis of osteoporotic fractures,"

Journal of Bone and Mineral Research, vol. 10, no. 12, pp. 1908–1912, 1995.

[46] D. Testi, A. Cappello, L. Chiari, M. Viceconti, and S. Gnudi, "Comparison of logistic and bayesian classifiers for evaluating the risk of femoral neck fracture in osteoporotic patients," *Medical and Biological Engineering and Computing*, vol. 39, no. 6, pp. 633–637, 2001.

[47] C. G. Alonso, M. D. Curiel, F. H. Carranza, R. P. Cano, A. D. Pérez, *et al.*, "Femoral bone mineral density, neck-shaft angle and mean femoral neck width as predictors of hip fracture in men and women," *Osteoporosis International*, vol. 11, no. 8, pp. 714–720, 2000.

[48] J. S. Gregory, D. Testi, A. Stewart, P. E. Undrill, D. M. Reid, and R. M. Aspden, "A method for assessment of the shape of the proximal femur and its relationship to osteoporotic hip fracture," *Osteoporosis International*, vol. 15, no. 1, pp. 5–11, 2004.

[49] J. Center, T. Nguyen, N. Pocock, K. Noakes, P. Kelly, J. Eisman, and P. Sambrook, "Femoral neck axis length, height loss and risk of hip fracture in males and females," *Osteoporosis international*, vol. 8, no. 1, pp. 75–81, 1998.

[50] V. Bousson, C. Bergot, B. Sutter, P. Levitz, B. Cortet, *et al.*, "Trabecular bone score (TBS): available knowledge, clinical relevance, and future prospects," *Osteoporosis International*, vol. 23, no. 5, pp. 1489–1501, 2012.

[51] M. Singh, A. Nagrath, and P. Maini, "Changes in trabecular pattern of the upper end of the femur as an index of osteoporosis," *Journal of Bone and Joint Surgery*, vol. 52, no. 3, pp. 457–467, 1970.

[52] H. Boehm, T. Vogel, A. Panteleon, D. Burklein, H. Bitterling, and M. Reiser, "Differentiation between post-menopausal women with and without hip fractures: enhanced evaluation of clinical DXA by topological analysis of the mineral distribution in the scan images," *Osteoporosis International*, vol. 18, no. 6, p. 779, 2007.

[53] N. Harvey, C. Glüer, N. Binkley, E. McCloskey, M.-L. Brandi, C. Cooper, D. Kendler, O. Lamy, A. Laslop, B. Camargos, *et al.*, "Trabecular bone score (TBS) as a new complementary approach for osteoporosis evaluation in clinical practice," *Bone*, vol. 78, pp. 216–224, 2015.

[54] K. Nassar, S. Paternotte, S. Kolta, J. Fechtenbaum, C. Roux, and K. Briot, "Added value of trabecular bone score over bone mineral density for identification of vertebral fractures in patients with areal bone mineral density in the non-osteoporotic range," *Osteoporosis International*, vol. 25, no. 1, pp. 243–249, 2014.

[55] J. Touvier, R. Winzenrieth, H. Johansson, J. Roux, J. Chaintreuil, H. Toumi, R. Jennane, D. Hans, and E. Lespessailles, "Fracture discrimination by combined bone mineral density (BMD) and microarchitectural texture analysis," *Calcified Tissue International*, vol. 96, no. 4, pp. 274–283, 2015.

[56] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53, no. 2, pp. 285–339, 1991.

[57] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[58] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[59] T. A. Gruen, G. M. Mcneice, and H. C. Amstutz, "Modes of failure of cemented stem-type femoral components: a radiographic analysis of loosening," *Clinical Orthopaedics and Related Research*, vol. 141, pp. 17–27, 1979.

[60] R. Huiskes, H. Weinans, H. Grootenboer, M. Dalstra, B. Fudala, and T. Slooff, "Adaptive bone-remodeling theory applied to prosthetic-design analysis," *Journal of Biomechanics*, vol. 20, no. 11, pp. 1135–1150, 1987.

[61] R. L. Jayasuriya, S. C. Buckley, A. J. Hamer, R. M. Kerry, I. Stockley, M. W. Tomouk, and J. M. Wilkinson, "Effect of sliding-taper compared with composite-beam cemented femoral prosthesis loading regime on proximal femoral bone remodeling: a randomized clinical trial," *Journal of Bone and Joint Surgery*, vol. 95A, no. 1, pp. 19–27, 2013.

[62] J. O. Penny, K. Brixen, J. E. Varmarken, O. Ovesen, and S. Overgaard, "Changes in bone mineral density of the acetabulum, femoral neck and femoral shaft, after hip resurfacing and total hip replacement: two-year

results from a randomised study," *Journal of Bone and Joint Surgery*, vol. 94B, no. 8, pp. 1036–44, 2012.

[63] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 1995.

[64] J. P. Shaffer, "Multiple hypothesis testing," *Annual Review of Psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[65] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[66] R. J. Simes, "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 73, no. 3, pp. 751–754, 1986.

[67] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[68] G. Hommel, "A stagewise rejective multiple test procedure based on a modified Bonferroni test," *Biometrika*, vol. 75, no. 2, pp. 383–386, 1988.

[69] K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2011.

[70] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, *et al.*, "A unified statistical approach for determining significant signals in images of cerebral activation," *Human Brain Mapping*, vol. 4, no. 1, pp. 58–73, 1996.

[71] K. E. Poole, G. M. Treece, P. M. Mayhew, J. Vaculík, P. Dungl, M. Horák, J. J. Štěpán, and A. H. Gee, "Cortical thickness mapping to identify focal osteoporosis in patients with hip fracture," *PLoS ONE*, vol. 7, no. 6, p. e38466, 2012.

[72] R. J. Adler and J. E. Taylor, *Random fields and geometry*. Springer Science & Business Media, 2009.

[73] R. J. Adler, *The geometry of random fields*. SIAM, 1981.

[74] K. J. Friston, A. Holmes, J.-B. Poline, C. J. Price, and C. D. Frith, "Detecting activations in PET and fMRI: levels of inference and power," *NeuroImage*, vol. 4, no. 3, pp. 223–235, 1996.

[75] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, “A three-dimensional statistical analysis for CBF activation studies in human brain,” *Journal of Cerebral Blood Flow & Metabolism*, vol. 12, no. 6, pp. 900–918, 1992.

[76] S. J. Kiebel, J.-B. Poline, K. J. Friston, A. P. Holmes, and K. J. Worsley, “Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model,” *NeuroImage*, vol. 10, no. 6, pp. 756–766, 1999.

[77] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[78] W. Li, J. Kornak, T. Harris, J. Keyak, C. Li, Y. Lu, X. Cheng, and T. Lang, “Identify fracture-critical regions inside the proximal femur using statistical parametric mapping,” *Bone*, vol. 44, no. 4, pp. 596–602, 2009.

[79] M. E. Glickman, S. R. Rao, and M. R. Schultz, “False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies,” *Journal of Clinical Epidemiology*, vol. 67, no. 8, pp. 850–7, 2014.

[80] J. Kerner, R. Huiskes, G. H. van Lenthe, H. Weinans, B. van Rietbergen, C. A. Engh, and A. A. Amis, “Correlation between pre-operative periprosthetic bone density and post-operative bone loss in THA can be explained by strain-adaptive remodelling,” *Journal of Biomechanics*, vol. 32, no. 7, pp. 695–703, 1999.

[81] H. M. Frost, “From Wolff’s law to the Utah paradigm: insights about bone physiology and its clinical applications,” *The Anatomical Record*, vol. 262, no. 4, pp. 398–419, 2001.

[82] G. L. Maistrelli, V. Fornasier, A. Binnington, K. McKenzie, V. Sessa, and I. Harrington, “Effect of stem modulus in a total hip arthroplasty model,” *Journal of Bone and Joint Surgery*, vol. 73-B, no. 1, pp. 43–46, 1991.

[83] S. S. Hughes, J. F. Furia, P. Smith, and V. Pellegrini, “Atrophy of the proximal part of the femur after total hip arthroplasty without cement,” *Journal of Bone and Joint Surgery*, vol. 77-A, no. 2, pp. 231–239, 1995.

[84] D. J. Kilgus, E. E. Shimaoka, J. S. Tipton, and R. W. Eberle, “Dual-energy x-ray absorptiometry measurement of bone mineral density around porous-coated cementless femoral implants,” *Journal of Bone and Joint Surgery*, vol. 75-B, no. 2, pp. 279–287, 1993.

[85] L. Rosenthall, D. J. Bobyns, and M. Tanzer, “Periprosthetic bone densitometry of the hip: influence of prosthetic design and hydroxyapatite coating on regional bone remodelling,” *Journal of Musculoskeletal and Neuronal Interactions*, vol. 1, no. 1, pp. 57–60, 2000.

[86] J. Kärrholm, C. Anderber, F. Snorrason, J. Thanner, N. Langeland, H. Malchau, and P. Herberts, “Evaluation of a femoral stem with reduced stiffness: a randomized study with use of radiostereometry and bone densitometry,” *Journal of Bone and Joint Surgery*, vol. 84, no. 9, pp. 1651–1658, 2002.

[87] A. I. Rahmy, T. Gosens, G. M. Blake, A. Tonino, and I. Fogelman, “Periprosthetic bone remodelling of two types of uncemented femoral implant with proximal hydroxyapatite coating: a 3-year follow-up study addressing the influence of prosthesis design and preoperative bone density on periprosthetic bone loss,” *Osteoporosis International*, vol. 15, no. 4, pp. 281–289, 2004.

[88] J. M. Wilkinson and D. G. Little, “Bisphosphonates in orthopedic applications,” *Bone*, vol. 49, no. 1, pp. 95–102, 2011.

[89] P. D. Diegel, A. U. Daniels, and H. K. Dunn, “Initial effect of collarless stem stiffness on femoral bone strain,” *Journal of Arthroplasty*, vol. 4, no. 2, pp. 173–178, 1989.

[90] N. P. Sheth, C. L. Nelson, and W. G. Paprosky, “Femoral bone loss in revision total hip arthroplasty: evaluation and management,” *Journal of the American Academy of Orthopaedic Surgeons*, vol. 21, no. 10, pp. 601–612, 2013.

[91] R. E. Cook, P. J. Jenkins, P. J. Walmsley, J. T. Patton, and C. M. Robinson, “Risk factors for periprosthetic fractures of the hip: a survivorship analysis,” *Clinical Orthopaedics and Related Research*, vol. 466, no. 7, pp. 1652–1656, 2008.

[92] D. a. Puleo and A. Nanci, “Understanding and controlling the bone-implant interface,” *Biomaterials*, vol. 20, no. 23-24, pp. 2311–2321, 1999.

[93] "Long-term radiographic changes in cemented total hip arthroplasty with six designs of femoral components," *Biomaterials*, vol. 24, no. 19, pp. 3351–3363, 2003.

[94] T. Scheerlinck and P. Casteleyn, "The design features of cemented femoral hip implants," *Journal of Bone and Joint Surgery*, vol. 88, no. 11, pp. 1409–1418, 2006.

[95] R. Huiskes, N. Verdonschot, and B. Nivbrant, "Migration, stem shape, and surface finish in cemented total hip arthroplasty," *Clinical Orthopaedics and Related Research*, vol. 355, pp. 103–112, 1998.

[96] J. Alfaro-Adrián, H. S. Gill, and D. W. Murray, "Cement migration after THR. a comparison of charnley elite and exeter femoral stems using RSA," *Journal of Bone and Joint Surgery.*, vol. 81, no. 1, pp. 130–134, 1999.

[97] G. Shen, "Femoral stem fixation. An engineering interpretation of the long-term outcome of Charnley and Exeter stems," *Journal of Bone and Joint Surgery*, vol. 80, no. 5, pp. 754–6, 1998.

[98] Y. Kishida, N. Sugano, T. Nishii, H. Miki, K. Yamaguchi, and H. Yoshikawa, "Preservation of the bone mineral density of the femur after surface replacement of the hip," *Journal of Bone and Joint Surgery*, vol. 86-B, no. 2, pp. 185–189, 2004.

[99] Y. Hayaishi, H. Miki, and T. Nishii, "Proximal femoral bone mineral density after resurfacing total hip arthroplasty and after standard stem-type cementless total hip arthroplasty, both having similar neck preservation and the same articulation type," *Journal of Arthroplasty*, vol. 22, no. 8, pp. 1208–1213, 2007.

[100] J. M. H. Smolders, A. Hol, T. Rijnders, and J. L. C. van Susante, "Changes in bone mineral density in the proximal femur after hip resurfacing and uncemented total hip replacement: A prospective randomised controlled study," *Journal of Bone and Joint Surgery*, vol. 92, no. 11, pp. 1509–14, 2010.

[101] N. J. Cooke, L. Rodgers, D. Rawlings, A. W. McCaskie, and J. P. Holland, "Bone density of the femoral neck following Birmingham hip resurfacing," *Acta Orthopaedica*, vol. 80, no. 6, pp. 660–665, 2009.

[102] R. Cordingley, L. Kohan, and B. Ben-Nissan, “What happens to femoral neck bone mineral density after hip resurfacing surgery?,” *Journal of Bone and Joint Surgery*, vol. 92, no. 12, pp. 1648–53, 2010.

[103] J. Wilkinson, N. Peel, R. Elson, I. Stockley, and R. Eastell, “Measuring bone mineral density of the pelvis and proximal femur after total hip arthroplasty,” *Journal of Bone and Joint Surgery*, vol. 83, no. 2, pp. 283–288, 2001.

[104] J. Kular, J. Tickner, S. M. Chim, and J. Xu, “An overview of the regulation of bone remodelling at the cellular level,” *Clinical Biochemistry*, vol. 45, no. 12, pp. 863–873, 2012.

[105] J. M. Wilkinson, A. C. Eagleton, I. Stockley, N. F. Peel, A. J. Hamer, and R. Eastell, “Effect of pamidronate on bone turnover and implant migration after total hip arthroplasty: a randomized trial,” *The Journal of Orthopaedic Research*, vol. 23, no. 1, pp. 1–8, 2005.

[106] V. B. Shim, R. P. Pitto, and I. A. Anderson, “Quantitative CT with finite element analysis: towards a predictive tool for bone remodelling around an uncemented tapered stem,” *International Orthopaedics*, vol. 36, no. 7, pp. 1363–1369, 2012.

[107] M. Taylor, “Finite element analysis of the resurfaced femoral head,” *The Proceedings of the Institution of Mechanical Engineers, Part H*, vol. 220, no. 2, pp. 289–297, 2006.

[108] Y. Watanabe, N. Shiba, S. Matsuo, F. Higuchi, Y. Tagawa, and A. Inoue, “Biomechanical study of the resurfacing hip arthroplasty: finite element analysis of the femoral component,” *Journal of Arthroplasty*, vol. 15, no. 4, pp. 505–511, 2000.

[109] S. Khosla and B. L. Riggs, “Pathophysiology of age-related bone loss and osteoporosis,” *Endocrinology and Metabolism Clinics*, vol. 34, no. 4, pp. 1015–1030, 2005.

[110] O. Demontiero, C. Vidal, and G. Duque, “Ageing and bone loss: new insights for the clinician,” *Therapeutic Advances in Musculoskeletal Disease*, vol. 4, no. 2, pp. 61–76, 2012.

[111] D. C. Van Essen, "Windows on the brain: the emerging role of atlases and databases in neuroscience," *Current Opinion in Neurobiology*, vol. 12, no. 5, pp. 574–579, 2002.

[112] W. Huizinga, D. Poot, M. Vernooij, G. Roshchupkin, E. Bron, M. Ikram, D. Rueckert, W. Niessen, S. Klein, and Alzheimer's Disease Neuroimaging Initiative, "A spatio-temporal reference model of the ageing brain," *NeuroImage*, vol. 169, pp. 11–22, 2018.

[113] H. K. Genant, S. Grampp, C. C. Gluer, K. G. Faulkner, M. Jergas, K. Engelke, S. Hagiwara, and C. Van Kuijk, "Universal standardization for dual x-ray absorptiometry: patient and phantom cross-calibration results.," *Journal of Bone and Mineral Research*, vol. 9, no. 10, pp. 1503–1514, 1994.

[114] S. L. Hui, S. Gao, X. H. Zhou, C. C. J. Johnston, Y. Lu, C. C. Gluer, S. Grampp, and H. Genant, "Universal standardization of bone density measurements: a method with optimal properties for calibration among several instruments.," *Journal of Bone and Mineral Research*, vol. 12, no. 9, pp. 1463–1470, 1997.

[115] Y. Lu, K. Ye, A. K. Mathur, S. Hui, T. P. Fuerst, and H. K. Genant, "Comparative calibration without a gold standard," *Statistics in Medicine*, vol. 16, no. 16, pp. 1889–1905, 1997.

[116] J. Pearson, J. Dequeker, M. Henley, J. Bright, J. Reeve, W. Kalender, A. Laval-Jeantet, P. Rüegsegger, D. Felsenberg, J. Adams, *et al.*, "European semi-anthropomorphic spine phantom for the calibration of bone densitometers: assessment of precision, stability and accuracy the european quantitation of osteoporosis study group," *Osteoporosis International*, vol. 5, no. 3, pp. 174–184, 1995.

[117] D. M. Reid, I. Mackay, S. Wilkinson, C. Miller, D. Schuette, J. Compston, C. Cooper, E. Duncan, N. Galwey, R. Keen, *et al.*, "Cross-calibration of dual-energy x-ray densitometers for a large, multi-center genetic study of osteoporosis," *Osteoporosis International*, vol. 17, no. 1, pp. 125–132, 2006.

[118] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[119] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[120] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for medical image analysis and computer vision," in *Medical Imaging 2001*, pp. 236–248, International Society for Optics and Photonics, 2001.

[121] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, and T. F. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1462–1472, 2013.

[122] G. Behiels, F. Maes, D. Vandermeulen, and P. Suetens, "Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models," *Medical Image Analysis*, vol. 6, no. 1, pp. 47–62, 2002.

[123] J. Gall and V. Lempitsky, "Class-specific Hough forests for object detection," in *Decision Forests for Computer Vision and Medical Image Analysis*, pp. 143–157, Springer, 2013.

[124] Claudia Lindner, Tim Cootes, and other members of the Centre for Imaging Sciences at The University of Manchester, UK, "Bonefinder (v1.2.0)," 2013. Available at `http://bone-finder.com/`.

[125] V. Barnett, "Simultaneous pairwise linear structural relationships," *Biometrics*, vol. 25, no. 1, pp. 129–142, 1969.

[126] W. E. Deming, "Statistical adjustment of data," 1943.

[127] P. J. Cornbleet and N. Gochman, "Incorrect least-squares regression coefficients in method-comparison analysis.," *Clinical Chemistry*, vol. 25, no. 3, pp. 432–438, 1979.

[128] J. D. Alele, D. L. Kamen, K. L. Hermayer, J. Fernandes, J. Soule, M. Ebeling, and T. C. Hulsey, "The prevalence of significant left-right hip bone mineral density differences among black and white women," *Osteoporosis International*, vol. 20, no. 12, pp. 2079–2085, 2009.

[129] A. D. Rao, S. Reddy, and D. S. Rao, "Is there a difference between right and left femoral bone density?," *Journal of Clinical Densitometry*, vol. 3, no. 1, pp. 57–61, 2000.

[130] R. S. Yang, K. S. Tsai, P. U. Chieng, and T. K. Liu, “Symmetry of bone mineral density at the proximal femur with emphasis on the effect of side dominance,” *Calcified Tissue International*, vol. 61, no. 3, pp. 189–191, 1997.

[131] R. Hamdy, G. M. Kiebzak, E. Seier, and N. B. Watts, “The prevalence of significant left-right differences in hip bone mineral density,” *Osteoporosis International*, vol. 17, no. 12, pp. 1772–1780, 2006.

[132] K. Yu, Z. Lu, and J. Stander, “Quantile regression: applications and current research areas,” *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 3, pp. 331–350, 2003.

[133] R. Koenker and G. Bassett Jr, “Regression quantiles,” *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[134] T. J. Cole and P. J. Green, “Smoothing reference centile curves: the lms method and penalized likelihood,” *Statistics in Medicine*, vol. 11, no. 10, pp. 1305–1319, 1992.

[135] T. W. Yee, “Quantile regression via vector generalized additive models,” *Statistics in Medicine*, vol. 23, no. 14, pp. 2295–2315, 2004.

[136] J. Carpenter and J. Bithell, “Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians,” *Statistics in Medicine*, vol. 19, no. 9, pp. 1141–1164, 2000.

[137] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, “Uk Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age,” *PLoS Medicine*, vol. 12, no. 3, p. e1001779, 2015.

[138] C. C. Glüer, R. Eastell, D. M. Reid, D. Felsenberg, C. Roux, R. Barkmann, W. Timm, T. Blenk, G. Armbrecht, A. Stewart, J. Clowes, F. E. Thomasius, and S. Kolta, “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. 782–93, 2004.

[139] E. V. McCloskey, M. Beneton, D. Charlesworth, K. Kayan, D. DeTakats, A. Dey, J. Orgee, R. Ashford, M. Forster, J. Cliffe, L. Kersh, J. Brazier,

J. Nichol, S. Aropuu, T. Jalava, and J. A. Kanis, "Clodronate reduces the incidence of fractures in community-dwelling elderly women unselected for osteoporosis: results of a double-blind, placebo-controlled randomized study," *Journal of Bone and Mineral Research*, vol. 22, no. 1, pp. 135–141, 2007.

[140] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, *et al.*, "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 46, pp. 1–10, 2013.

[141] S. Baim, C. R. Wilson, E. M. Lewiecki, M. M. Luckey, R. W. Downs, and B. C. Lentle, "Precision assessment and radiation safety for dual-energy x-ray absorptiometry: position paper of the international society for clinical densitometry," *Journal of Clinical Densitometry*, vol. 8, no. 4, pp. 371–378, 2005.

[142] M. Lodder, W. Lems, H. Ader, A. Marthinsen, S. van Coeverden, P. Lips, J. Netelenbos, B. Dijkmans, and J. Roos, "Reproducibility of bone mineral density measurement in daily practice," *Annals of the Rheumatic Diseases*, vol. 63, no. 3, pp. 285–289, 2004.

[143] S. Henzell, S. Dhaliwal, R. Pontifex, F. Gill, R. Price, R. Retallack, and R. Prince, "Precision error of fan-beam dual x-ray absorptiometry scans at spine, hip, and forearm," *Journal of Clinical Densitometry*, vol. 3, no. 4, pp. 359–364, 2000.

[144] S. L. Morgan, W. Abercrombie, and J. Y. Lee, "Need for precision studies at individual institutions and assessment of size of regions of interest on serial DXA scans," *Journal of Clinical Densitometry*, vol. 6, no. 2, pp. 97–101, 2003.

[145] J. White, S. S. Harris, G. E. Dallal, and B. Dawson-Hughes, "Precision of single vs bilateral hip bone mineral density scans," *Journal of Clinical Densitometry*, vol. 6, no. 2, pp. 159–162, 2003.

[146] M. Schröder, H. Gottschling, N. Reimers, M. Hauschild, and R. Burgkart, "Automated morphometric analysis of the femur on large anatomi-

cal databases with highly accurate correspondence detection," *Open Medicine Journal*, vol. 1, no. 1, 2014.

[147] C. D. L. Thomas, P. M. Mayhew, J. Power, K. E. Poole, N. Loveridge, J. G. Clement, C. J. Burgoyne, and J. Reeve, "Femoral neck trabecular bone: loss with ageing and role in preventing fracture," *Journal of Bone and Mineral Research*, vol. 24, no. 11, pp. 1808–1818, 2009.

[148] O. M. Ahmad, K. Ramamurthi, K. E. Wilson, K. Engelke, M. Bouxsein, and R. H. Taylor, "3D structural measurements of the proximal femur from 2D DXA images using a statistical atlas," in *SPIE 7260, Medical Imaging 2009: Computer-Aided Diagnosis*, p. 726005, International Society for Optics and Photonics, 2009.

[149] J. A. Kanis, *Assessment of osteoporosis at the primary health care level.* WHO Collaborating Centre for Metabolic Bone Diseases, University of Sheffield Medical School, 2008.

[150] T. J. Beck, A. C. Looker, C. B. Ruff, H. Sievanen, and H. W. Wahner, "Structural trends in the aging femoral neck and proximal shaft: analysis of the third national health and nutrition examination survey dual-energy x-ray absorptiometry data," *Journal of Bone and Mineral Research*, vol. 15, no. 12, pp. 2297–2304, 2000.

[151] L. Warming, C. Hassager, and C. Christiansen, "Changes in bone mineral density with age in men and women: a longitudinal study," *Osteoporosis International*, vol. 13, no. 2, pp. 105–112, 2002.

[152] L. Melton, S. Khosla, E. Atkinson, M. O'connor, W. O'fallon, and B. Riggs, "Cross-sectional versus longitudinal evaluation of bone loss in men and women," *Osteoporosis International*, vol. 11, no. 7, pp. 592–599, 2000.

[153] H. Burger, P. Van Daele, D. Algra, F. Van den Ouweland, D. Grobbee, A. Hofman, C. Van Kuijk, H. Schütte, J. Birkenhäger, and H. Pols, "The association between age and bone mineral density in men and women aged 55 years and over: the Rotterdam study," *Journal of Bone and Mineral Research*, vol. 25, no. 1, pp. 1–13, 1994.

[154] J. F. Aloia, A. Vaswani, P. Ross, and S. H. Cohn, "Aging bone loss from the femur, spine, radius, and total skeleton," *Metabolism*, vol. 39, no. 11, pp. 1144–1150, 1990.

[155] K. M. Nicks, S. Amin, E. J. Atkinson, B. L. Riggs, L. J. Melton, and S. Khosla, “Relationship of age to bone microstructure independent of areal bone mineral density,” *Journal of Bone and Mineral Research*, vol. 27, no. 3, pp. 637–644, 2012.

[156] A. M. Cheung, J. D. Adachi, D. A. Hanley, D. L. Kendler, K. S. Davison, R. Josse, J. P. Brown, L.-G. Ste-Marie, R. Kremer, M. C. Erlandson, *et al.*, “High-resolution peripheral quantitative computed tomography for the assessment of bone strength and structure: a review by the canadian bone strength working group,” *Current osteoporosis reports*, vol. 11, no. 2, pp. 136–146, 2013.

[157] B. L. Riggs, L. J. Melton, R. A. Robb, J. J. Camp, E. J. Atkinson, J. M. Peterson, P. A. Rouleau, C. H. McCollough, M. L. Bouxsein, and S. Khosla, “Population-based study of age and sex differences in bone volumetric density, size, geometry, and structure at different skeletal sites,” *Journal of Bone and Mineral Research*, vol. 19, no. 12, pp. 1945–1954, 2004.

[158] L. Voo, M. Armand, and M. Kleinberger, “Stress fracture risk analysis of the human femur based on computational biomechanics,” *Johns Hopkins APL Tech Dig*, vol. 25, no. 3, pp. 223–30, 2004.

[159] J. Dequeker, “Osteoporotic fractures, ageing and the bone density T-score,” *Clinical Rheumatology*, vol. 3, no. 19, pp. 171–173, 2000.

[160] A. L. Cole, L. Webb, and T. Cole, “Bone age estimation: a comparison of methods,” *The British Journal of Radiology*, vol. 61, no. 728, pp. 683–686, 1988.

[161] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H. Huang, and V. Gilsanz, “Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal ROI extraction,” *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 715–729, 2001.

[162] R. G. Stevens and S. H. Moolgavkar, “A cohort analysis of lung cancer and smoking in British males,” *American Journal of Epidemiology*, vol. 119, no. 4, pp. 624–641, 1984.

[163] V. Sapthagirivasan, M. Anburajan, and V. Mahadevan, “Bone trabecular analysis of femur radiographs for the assessment of osteoporosis

using DWT and DXA," *International journal of computer theory and engineering*, vol. 5, no. 4, p. 616, 2013.

[164] A. Schmidt-Richberg, C. Ledig, R. Guerrero, H. Molina-Abril, A. Frangi, D. Rueckert, A. D. N. Initiative, *et al.*, "Learning biomarker models for progression estimation of Alzheimer's disease," *PloS ONE*, vol. 11, no. 4, p. e0153040, 2016.

[165] L. C. Brunner, L. Eshilian-Oates, and T. Y. Kuo, "Hip fractures in adults," *American family physician*, vol. 67, no. 3, pp. 537–42, 2003.

[166] R. A. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: A survey," *Information Sciences*, vol. 301, pp. 141–160, 2015.

[167] H. H. Barrett and K. J. Myers, *Foundations of image science.* John Wiley & Sons, 2013.

[168] F. Xie, Y. Lu, A. C. Bovik, Z. Jiang, and R. Meng, "Application-driven no-reference quality assessment for dermoscopy images with multiple distortions," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1248–1256, 2016.

[169] J. G. Brankov, Y. Yang, L. Wei, I. El Naqa, and M. N. Wernick, "Learning a channelized observer for image quality assessment," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 991–999, 2009.

[170] E. C. Frey, K. L. Gilland, and B. M. Tsui, "Application of task-based measures of image quality to optimization and evaluation of three-dimensional reconstruction-based compensation methods in myocardial perfusion SPECT," *IEEE Transactions on Medical Imaging*, vol. 21, no. 9, pp. 1040–1050, 2002.

[171] J. A. Jacobson, D. A. Jamadar, and C. W. Hayes, "Dual X-ray absorptiometry: recognizing image artifacts and pathology," *American Journal of Roentgenology*, vol. 174, no. 6, pp. 1699–1705, 2000.

[172] D. Le Bihan, C. Poupon, A. Amadon, and F. Lethimonnier, "Artifacts and pitfalls in diffusion mri," *Journal of Magnetic Resonance Imaging*, vol. 24, no. 3, pp. 478–488, 2006.

[173] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[174] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008-a database for evaluation of full-reference visual quality assessment metrics,” *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[175] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, “A geometric framework for unsupervised anomaly detection,” in *Applications of data mining in computer security*, pp. 77–101, Springer, 2002.

[176] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering objects and their location in images,” in *10th IEEE International Conference on Computer Vision (ICCV2005)*, vol. 1, pp. 370–377, IEEE, 2005.

[177] P. Ye and D. Doermann, “No-reference image quality assessment based on visual codebook,” in *18th IEEE International Conference on Image Processing (ICIP2011)*, pp. 3089–3092, IEEE, 2011.

[178] G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[179] C. Schmid and R. Mohr, “Local grayvalue invariants for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–534, 1997.

[180] E. M. Lewiecki, N. Binkley, and S. M. Petak, “DXA quality matters,” *Journal of Clinical Densitometry*, vol. 9, no. 4, pp. 388–392, 2006.

[181] R. Lorente-Ramos, J. Azpeitia-Armán, A. Muñoz-Hernández, J. M. García-Gómez, P. Díez-Martínez, and M. Grande-Bárez, “Dual-energy x-ray absorptiometry in the diagnosis of osteoporosis: a practical guide,” *American Journal of Roentgenology*, vol. 196, no. 4, pp. 897–904, 2011.

[182] G. D. Cawkwell, “Movement artifact and dual x-ray absorptiometry,” *Journal of Clinical Densitometry*, vol. 1, no. 2, pp. 141–147, 1998.

[183] A. El Maghraoui and C. Roux, "DXA scanning in clinical practice," *QJM: An International Journal of Medicine*, vol. 101, no. 8, pp. 605–617, 2008.

[184] J. He, L. G. Ogden, L. A. Bazzano, S. Vupputuri, C. Loria, and P. K. Whelton, "Risk factors for congestive heart failure in us men and women: NHANES I epidemiologic follow-up study," *Archives of Internal Medicine*, vol. 161, no. 7, pp. 996–1002, 2001.

[185] S. Marsland and T. Shardlow, "Langevin equations for landmark image registration with uncertainty," *SIAM Journal on Imaging Sciences*, vol. 10, no. 2, pp. 782–807, 2017.

[186] L. Kuhnel, S. Sommer, A. Pai, and L. L. Raket, "Most likely separation of intensity and warping effects in image registration," *SIAM Journal on Imaging Sciences*, vol. 10, no. 2, pp. 578–601, 2017.

[187] X. Hua, A. D. Leow, S. Lee, A. D. Klunder, A. W. Toga, N. Lepore, Y.-Y. Chou, C. Brun, M.-C. Chiang, M. Barysheva, *et al.*, "3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry," *NeuroImage*, vol. 41, no. 1, pp. 19–34, 2008.

[188] E. Salmon, F. Collette, C. Degueldre, C. Lemaire, and G. Franck, "Voxel-based analysis of confounding effects of age and dementia severity on cerebral metabolism in Alzheimer's disease," *Human Brain Mapping*, vol. 10, no. 1, pp. 39–48, 2000.

[189] D. Mellström, L. Vandenput, H. Mallmin, A. H. Holmberg, M. Lorentzon, A. Odén, H. Johansson, E. S. Orwoll, F. Labrie, M. K. Karlsson, *et al.*, "Older men with low serum estradiol and high serum SHBG have an increased risk of fractures," *Journal of Bone and Mineral Research*, vol. 23, no. 10, pp. 1552–1560, 2008.

[190] L. Sànchez-Riera, N. Wilson, N. Kamalaraj, J. M. Nolla, C. Kok, Y. Li, M. Macara, R. Norman, J. S. Chen, E. U. Smith, *et al.*, "Osteoporosis and fragility fractures," *Best Practice & Research Clinical Rheumatology*, vol. 24, no. 6, pp. 793–810, 2010.