Novel Imaging and Processing for Early Musculoskeletal Disease



Yanda Xin

Worcester College

University of Oxford

A thesis submitted to the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Trinity Term 2018

Novel Imaging and Processing for Early Musculoskeletal Disease

A thesis submitted for the degree of

Doctor of Philosophy

Yanda Xin

Worcester College

University of Oxford

Trinity Term 2018

©2019 YANDA XIN

All rights reserved. No part of this thesis may be produced or transmitted in any form or by any means, electronic or mechanical, including photocopying recording or any information storage and retrieval system, without the prior written permission of the author.

Word Count: **46,281** (excluding front matter, figures, tables, program codes, footnotes, Appendices and References)

Submitted to the Examination Schools, University of Oxford on 5 October, Trinity 2018

Defended in viva at the Botnar Research Centre, Institute of Musculoskeletal Sciences, University of Oxford on 7 November, Michaelmas 2018

Minor correction (this edition) submitted on 6 February, Hilary 2019

Minor correction approved by examiners on 8 March, Hilary 2019

Leave to Supplicate granted on 29 March, Hilary 2019

This thesis was typeset by the author using the LaTeX typesetting system originally developed by Leslie Lamport, based on TeX created by Donald Knuth.

The References at the end of the thesis were formatted in accordance with the 2014 *New Hart's Rules* and the 2002 *Oxford Guide to Style*, using the BibLaTeX-oxref (`http://github.com/alex-ball/biblatex-oxref`) package developed by Alex Ball.

Printed at the Botnar Research Centre, University of Oxford

Bound by the Oxford Print Centre, Oxford

To my beloved wife Cheng Li

ABSTRACT

Musculoskeletal disease, especially osteoarthritis (OA) is a major cause of pain and disability. The combination of large burden and limited treatment options for patients leads to a need to improve our understanding of how it manifests and progresses in its early stages, and to link this understanding to individual patients. Quantitative magnetic resonance imaging (MRI) techniques has demonstrated the ability to detect these early changes, among which T_2 mapping is one of the most widely used. This thesis attempts to translate and build a platform technology for the analysis of musculoskeletal MRI, with a focus on T_2 mapping.

By performing a systematic review and meta-analysis, the results of 12 knee studies, and 4 knee studies are combined, from which the difference in T_2 value between OA and healthy tissue are found to be extremely small (ΔT_2^{META} (Knee) = 2.20 ms, ΔT_2^{META} (Hip) = 3.31 ms). Using these values as the bottom line criteria and by performing a Monte Carlo simulation, the performance of 16 T_2 fitting methods are further evaluated. It is found although advanced T_2 fitting methods such as NCEXP (where a simplified equation is reported for the first time) and SQEXP can achieve good trueness performance, the poor precision caused by noise would make absolute measurement of T_2 of little use. A new quantitative approach, which involves transforming T_2 images from different patients to the same space is needed to allow direct comparison and better analysis.

OxteoFlow is the foundation of the analysis platform developed in this thesis, which uses affine transformations followed by a more complex nonlinear transformation to register individuals to a reference scan. As can be seen by visual inspection, OxteoFlow successfully normalised knees and hips of different shapes into their respective reference spaces. Parallel computing and grid search optimisation are enabled in OxteoFlow, making it a suitable tool for musculoskeletal imaging studies.

Using data from the Osteoarthritis Initiative (OAI), the baseline T_2 scans are stratified into four groups—'Both','ROA', 'Pain' and 'None' based on their outcomes. By performing a voxel-based pattern analysis on their baseline T_2 maps, 6 different voxel activation patterns, all relating to different incident/progression outcomes are identified, with their complex relationship illustrated by voxels detected distance maps. Pattern 6, which shows effect of ROA Progression activations is found to be the most significant as it survived the extremely conservative Bonferroni correction at $P < 0.001$, indicating T_2 mapping is a powerful tool to predict ROA progression outcome at the earliest possible time.

Specifically targeting the ROA progression cohort, a deep learning model, OxteoNet, is developed and compared against leading machine learning algorithms. It is found the OxteoNet provides a small, though significant, improvement in AUROC, yet a dramatic improvement in AP score than traditional machine learning methods. The class activation maps (CAMs) of OxteoNet may serve as an indicator for the clinician as sites at which pathology are observed that need further inspection.

In summary, this thesis covered a complete research and development (R&D) life cycle of T_2 imaging and its applications (calculation $\longrightarrow$ preprocessing $\longrightarrow$ analysis $\longrightarrow$ diagnosis/prognosis), answered some fundamental and important questions related to this non-invasive OA biomarker, translated some of the latest and best technical advances from a wide range of medical imaging domains into musculoskeletal, and demonstrated its value as an image analysis platform technology for musculoskeletal studies.



FLOWCHART ABSTRACT OF THE THESIS

Translating fMRI Analysis Methods into a Platform Technology for Musculoskeletal MRI Analysis

Acknowledgements

I would like to thank my supervisor and best friend, Professor Cameron Brown, for his unreserved support and unparalleled trust over the years. Instead of instructing me what to do, Cameron encouraged me to discover and develop research ideas independently. These ideas, however, would not have been developed into the chapters in this thesis without Cameron's thoughtful advice. Almost all the parameters, methods, and analyses presented in this thesis were collective intelligence evaluated, discussed and decided between Cameron and me over our routine coffee breaks in those British rainy afternoons. Very few students were lucky enough to have received as much attention, support and flexibility from their supervisors as I had from Cameron. It was my greatest privilege to have witnessed his academic and professional achievements over the years as his student, from him being promoted to the Associate Professor position at Oxford, to his acceptance of a professorship and directorship at QUT, Australia. It is quite difficult to imagine now what kind of person I would become and what kind of life I would live if he did not offer me the opportunity to read for this DPhil in the first place. In this sense, the lifelong positive impact he has created on me is beyond what one could possibly thank in words.

I would also like to thank my co-supervisor, Professor Siôn Glyn-Jones for his great support and tolerance. Siôn not only offered me great suggestions at all the important milestones throughout my DPhil on the clinical side, but also generously contributed towards half of my first year tuition fees, an offer much more important than he knew, before I secured external funding. Later in my DPhil, the flexibility Siôn allowed me to research beyond the limited number of hip scans we had locally, but to use the much bigger OAI knee image set was a critical boost to the development of my work in Chapter 4 of this thesis. In the same way as I felt for Cameron, I was so honoured to have witnessed Siôn being awarded a Professorship at Oxford in 2014, as one of his DPhil students at that time.

I greatly thank Professor David Beard for participating in all of my DPhil key milestones—as an accessor for both my Transfer of Status and Confirmation of Status examinations, and the internal examiner of my final viva. His insightful advice throughout my DPhil study pushed me to add more clinical 'flavour' in this thesis, a direct result of which was the meta-analysis study in Chapter 2.

I am also very gratefully to my external examiner, Professor Neal Bangerter. I was lucky enough to have my very first lessons in MRI physics taught by Neal, when he was a visiting scholar at Oxford in 2014. He also shared his tutorial notes in MRI physics with me. Neal demonstrated how to code a T_2 fitting program in Matlab, which directly informed my work related to T_2 fitting methods in Chapter 2.

I would also like to thank Professor Afsie Sabokbar for being a great mentor, friend and the departmental Director of Graduate Studies. I can still remember her warm welcome on my first day in the department. She provided excellent references for my funding applications, and I gratefully thank her for taking care of me on many administrative issues of my DPhil, particularly in my final year.

There are many other scholars within and outside of Oxford who contributed to my DPhil research. I thank Drs Minsi Chen and Michael Schlussel for being accessors of my Transfer of Status and Confirmation of Status examinations respectively, both of whom provided valuable guidance on engineering and statistics topics. I thank Professor Mark Jenkinson, developer of FSL, for spending time to discuss with me and Cameron the possibility of using FSL in musculoskeletal MRI. I thank Dr Danny Park for his interesting discussion in T_2 fitting methods and noise in MRI. I thank Dr Ali Zarei for his guidance on wet lab techniques. I also thank my clinician colleagues Antony Palmer and Scott Fernquest for their guidance in manual segmentation of hip MRI, and Luke Jones and Jonathan Palmer for sharing their research insights into the OAI dataset. I thank our research radiologist Claudio Pereira for providing me copies of hip scans. I thank my mentors Jennifer Martay, Stephanie Dakin and Sarah Snelling for their valuable advice at the early stage of my DPhil. I would also like to thank Professors Jamie Warner, Alexander Korsunsky and Richard Todd for allowing me to conduct materials processing and characterisation experiments in their labs, for some of my side projects in biological materials.

I thank Drs Na Wu, Zhiqiao Wang, Ying-chun Chen, Siqi Ying, Peilin Li, Qianqi Lin, Xiaochen Wang and Ms Ran Xu for their great friendships over the years at Oxford.

This DPhil was partially funded by Arthritis Research UK Centre for Sport, Exercise and Osteoarthritis and the State Scholarship Fund of my government. Without the financial support from my funders, it would be impossible for me to complete my DPhil studies at Oxford. I would like to thank the selection panel of my funders whose names were unknown to me. I also thank the administrative staff Lis Ahlström, Bronya Norton, Joanne Bartram, Li Yan and Xiaopeng Hu of my funders, and Cath Fraser and Tom Cooper at Worcester College for their administration of my scholarships.

I thank the open-source communities of neuroimaging, machine learning, and Scientific Python, which provided nearly all the methodologies and tools I would ever need to advance my research. I thank all the developers and contributors of ANTs, FreeSurfer, ITK-SNAP, Nipy, Nipype, scikit-learn, Nilearn, PyTorch, SciPy, NumPy, scikit-image, matplotlib and seaborn. I also thank all the developers of FSL for making such a powerful toolkit a free software.

I thank my parents Dongfang Luo and Xiaoguang Xin for their love. They have provided the best family environment for me to grow up. I am deeply indebted to them for all they have done. They always hoped their son to be academically successful, and I believe no other degree than a doctorate from Oxford can better match their expectation. I also thank my parents-in-law Huijuan Song and Zhi Ma for being supportive of my study. I thank my grandparents Zhongrui Xin, Shuhua Lin and Fulan Liu for their love and encouragement in my whole life. I am sorry that my late grandfather, Jinlong Luo, who greatly influenced me on becoming who I am, cannot witness my achievement. However, as a life-long sufferer of hip conditions he would be proud of his grandson if he knew that I have advanced the musculoskeletal research a little bit. I thank my aunts Li Xin for allowing me to stay at her place in Beijing when I was making PhD applications, and Dongmei Luo for her help in my scholarship application.

Last but definitely not least, my deepest gratitude goes to my wife, Cheng Li. She deserves better than me this degree, as she was involved in every day and everything of this long process, and has devoted as much physical and mental efforts into it as I, if not more. She joined me in the UK immediately after she finished her Master's, and then worked in Bicester as a Sales Assistant to support our young family—she chose to give up her dream of becoming a good scholar for being a good wife. The sacrifice she has made to support my study is something no one other than the two of us can understand.

Of the 5,000 days we have been together since high school, my wife treated my academic pursuit as if it was her own, although I always know she could do better than me. I am grateful that she later worked for the University and still does, and we had our wedding in the middle of my DPhil. Marrying you was the best decision that I have ever made in my life, my love. This degree is for both of us.

Yanda Xin

3 July, Trinity 2019

Oxford

This page intentionally left blank

Contents

Abstract i
Acknowledgements iii
Contents vii
Abbreviations and Acronyms xiii

Chapter 1 Background and Motivation

1.1 Background 1
1.2 General Aims 2
1.3 State of the Art 3
1.3.1 Early Detection of Osteoarthritis 3
1.3.2 Cartilage T_2 Mapping 4
1.4 Narrative Review 5
1.5 Specific Objectives and Approach 6

Chapter 2 T_2 Mapping for Osteoarthritis: Systematic Review, Meta-Analysis and Evaluation of Fitting Methods

2.1 Introduction 9
2.2 Defining the Standard Case-Control T_2 Difference for Osteoarthritis with Systematic Review and Meta-Analysis 11
2.2.1 Methods 11
2.2.1.1 Eligibility Criteria 11
2.2.1.2 Literature Search and Study Selection 11

2.2.1.3 Data Extraction and Processing . . . 12
2.2.1.4 Models for Meta-Analysis . . . 19
2.2.1.5 Mean Difference . . . 21
2.2.1.6 Effect Sizes . . . 21
2.2.1.7 Meta-Analysis in R . . . 23
2.2.1.8 Interpreting Effect Sizes . . . 23
2.2.1.9 Benchmarks for Interpretation . . . 26
2.2.2 Results and Discussion . . . 27
2.2.2.1 ΔT_2^{META}—the Mean Difference in T_2 . . . 27
2.2.2.2 ES^{META}—the Effect Size . . . 29
2.3 Performance Evaluation of Sixteen T_2 Fitting Methods with Monte-Carlo Simulation . . . 35
2.3.1 Theory . . . 35
2.3.1.1 The Noise-Free Signal . . . 35
2.3.1.2 The Rician Distribution of Signal . . . 35
2.3.1.3 The Rayleigh Distribution of Signal . . . 36
2.3.1.4 The Noncentral Chi Distribution of Signal . . . 37
2.3.2 Methods . . . 37
2.3.2.1 Summary of Fitting Methods . . . 37
2.3.2.2 Fitting Methods in Detail . . . 39
2.3.2.3 Monte Carlo Simulation . . . 47
2.3.2.4 Synthetic T_2 signal . . . 49
2.3.2.5 Interpretation of the Synthetic T_2 signal . . . 50
2.3.2.6 Performance Metrics and the Bottom Line Criterion . . . 52
2.3.3 Results and Discussion . . . 59
2.3.3.1 Trueness . . . 59
2.3.3.2 Precision . . . 61
2.3.3.3 Effect Size . . . 62
2.3.3.4 Overview and Recommendations . . . 67

Chapter 3 Automating the Preprocessing of Musculoskeletal MRI: Development of OxteoFlow

3.1 Introduction **69**
3.1.1 Philosophy Behind the OxteoFlow 70
3.2 Theory **71**
3.2.1 Reference Image and Reference Space 71
3.2.2 Linear Transformations 73
3.2.3 Non-linear Transformations 74
3.2.4 Cost Function 74
3.2.5 Optimisation of Cost Function 75
3.2.6 Regularisation 75
3.2.7 Sub-sampling 76
3.2.8 Diffeomorphism 76
3.2.9 Interpolation 77
3.2.10 Intensity Non-uniformity 78
3.3 Methods **79**
3.3.1 Algorithm Layers 79
3.3.1.1 Hip Extraction Layer 79
3.3.1.2 Non-uniformity Correction Layer 82
3.3.1.3 Linear Registration Layer 85
3.3.1.4 Convert Affine Layer 86
3.3.1.5 Non-linear Registration Layer (FSL Branch) 87
3.3.1.6 Non-linear Registration Layer (ANTs Branch) 90
3.3.2 Metrics Layers 92
3.3.2.1 Correlation Coefficient 92
3.3.2.2 Correlation Ratio 93
3.3.2.3 Mutual Information 94
3.3.2.4 Normalised Mutual Information 95
3.3.3 Utility Layers 96
3.3.3.1 Information Source Layer 96
3.3.3.2 Data Source Layer 96
3.3.3.3 Reference Source Layer 97

3.3.3.4 Data Sink Layer . . . 98
3.3.3.5 Assign ID Layer . . . 99
3.3.3.6 Merge Metrics Layer . . . 99
3.3.3.7 Join Metrics Layer . . . 100
3.3.3.8 Save Metrics Layer . . . 100
3.3.4 Implementation . . . 101
3.3.4.1 Interfaces and Workflow Engine . . . 101
3.3.4.2 Multi-processing Support . . . 102
3.3.4.3 Optimisation and Execution Modes . . . 103
3.3.5 MR Imaging Data . . . 106
3.3.5.1 OAI Knee Image Set . . . 106
3.3.5.2 Sibkids Hip Image Set . . . 106
3.3.5.3 Reference Images . . . 107
3.3.5.4 Optimisation and Validation Cohort . . . 107
3.3.6 Optimisation and Validation . . . 108
3.3.6.1 Hyperparameters vs Model Parameters . . . 108
3.3.6.2 Grid search Optimisation of Hyperparameters . . . 109
3.3.6.3 Decision Rule of Similarity Metrics . . . 110
3.3.6.4 Validation . . . 111
3.4 Results and Discussion . . . 111
3.4.1 Architecture of OxteoFlow . . . 111
3.4.1.1 Optimisation Mode . . . 111
3.4.1.2 Execution Mode . . . 112
3.4.2 Optimisation of Regularisation λ and Warp Resolution . . . 112
3.4.2.1 Grid Search Optimisation . . . 112
3.4.2.2 Decision on Parameters . . . 115
3.4.3 Visual Inspection of Preprocessing Outputs . . . 116
3.4.4 Validation . . . 125

Chapter 4 T_2 Patterns Associated with OA Manifestation and Progression: Voxel-based Statistical Analysis, Machine Learning, and Development of OxteoNet

4.1 Introduction 133

4.2 Voxel-Based Statistical Analysis of the Knee T_2 Maps: Data From the Osteoarthritis Initiative 134

4.2.1 Methods 134

4.2.1.1 Hardware and Image Acquisition 134

4.2.1.2 Image Format Conversion 135

4.2.1.3 Data Cleaning 136

4.2.1.4 T_2 Mapping Calculations and Preprocessing 137

4.2.1.5 Definitions of Incident and Progression of Radiographic OA and Pain 138

4.2.1.6 Stratifications of the Outcome Groups 145

4.2.1.7 Voxel-Based T_2 Relaxometry (VBR^{T_2}) 146

4.2.1.8 Voxel-based Statistical Analyses: An Overview 148

4.2.2 Results and Discussion 150

4.2.2.1 Voxel-Based Descriptive Statistics for the Incidence Outcome Groups: Visual Assessment 150

4.2.2.2 VBR^{T_2} of the Incidence Outcome Groups: One-way ANOVA 165

4.2.2.3 VBR^{T_2} of the Incidence Outcome Groups: Two-sample t-tests 166

4.2.2.4 Voxel-Based Descriptive Statistics for the Progression Outcome Groups: Visual Assessment 176

4.2.2.5 VBR^{T_2} of the Progression Outcome Groups: One-way ANOVA . . . 191

4.2.2.6 VBR^{T_2} of the Progression Outcome Groups: Two-sample t-tests . . . 192

4.2.2.7 Summaries and Interpretations 204

4.3 Prognosis with Machine Learning and Deep Learning: Development of OxteoNet 217

4.3.1 Philosophy Behind the OxteoNet 217

4.3.2 Methods 218

4.3.2.1 Problem Formulation 218

4.3.2.2 Study Design 218

4.3.2.3 Data Preparation 218

4.3.2.4 Hardware Preparation 219

4.3.2.5 Traditional Machine Learning Models 219
4.3.2.6 Deep Learning Model—the OxteoNet 220
4.3.2.7 Performance Scores 222
4.3.3 Results and Discussion 223
4.3.3.1 Performance Comparison 223
4.3.3.2 Model Interpretations 224
4.4 Conclusions 230

Chapter 5 General Discussion and Conclusions

5.1 Limitations and Future Work 234

References 239

Abbreviations and Acronyms

2D	Two-dimensional
3D	Three-dimensional
ABC	AdaBoost Classifier
AFNI	Analysis of Functional Neuroimages (Software)
ANTs	Advanced Normalization Tools (Software)
ANOVA	Analysis of Variance
AUC	Area Under the Curve
AUROC	Area Under the Curve of the Receiver Operating Characteristic
B_0	Static Magnetic (Polarising) Field
B_1	Radiofrequency (Excitation) Field
BAYES_{Rice}	Bayesian Estimation of Rician Signal (Algorithm)
BAYES_{ncX}	Bayesian Estimation of Noncentral Chi Signal (Algorithm)
BET	Brain Extraction Tool (Software)
'Both'	Both Incident Radiographic Osteoarthritis and Pain, or Both Radiographic Osteoarthritis and Pain Progression (Outcome)
CC	Correlation Coefficient, or Cross-Correlation
CDF	Cumulative Distribution Function
Chance	A classifier that makes predictions purely based on guess
CheXNet	Deep Learning for Pneumonia Detection on Chest X-Rays (Algorithm based on DenseNet)
CHT	Circle Hough Transforms (Algorithm)
CI	Confidence Interval
CL	Common Language Effect Size
CNN	Convolutional Neural Network, also ConvNet
CPU	Central Processing Unit
CR	Correlation Ratio
CV	Cross-Validation
DAG	Directed Acyclic Graph
DenseNet	Densely Connected Convolutional Networks (Algorithm)

(Continued on the next page)

(Continued)

DICOM	Digital Imaging and Communications in Medicine, or Medial Image stored in DICOM (.dcm) Format
dGEMRIC	Delayed Gadolinium-enhanced MRI of Cartilage
DTs	Decision Trees Classifier
ES	Effect Size
ES^{META}	Effect Size Estimated in Meta-analysis
FLIRT	FMRIB's Linear Image Registration Tool (Software)
fMRI	Functional Magnetic Resonance Imaging
FMRIB	Oxford Centre for Functional MRI of the Brain
FN	False Negative
FNIRT	FMRIB's Non-linear Image Registration Tool (Software)
FOV	Field of View
FP	False Positive
FSL	FMRIB Software Library (Software)
FWE	Family-Wise Error
FWER	Family-Wise Error Rate
GaussianNB	Gaussian Naive Bayes
GE	Gradient Echo
GPU	Graphics Processing Unit
HET	Hip Extraction Tool (Software)
ICBM	International Consortium for Brain Mapping
ICBM152	Same as the 'MNI152' on the next page
ITK	Insight Segmentation and Registration Toolkit (Software)
JSN	Joint Space Narrowing
KNN	k-Nearest Neighbors
KL	Kellgren-Lawrence
KLG	Kellgren-Lawrence Grade
LinearSVC	Linear Support Vector Classification
LLS	Linear Least Squares (Algorithm)

(Continued on the next page)

(Continued)

Logit	Logistic Regression with L2 regularisation
M^1NCM	First-Moment Noise-Corrected Model (Algorithm)
M^1NCM_σ	First-Moment Noise-Corrected Model with Known σ (Algorithm)
M^2NCM	Second-Moment Noise-Corrected Model (Algorithm)
M^2NCM_σ	Second-Moment Noise-Corrected Model with Known σ (Algorithm)
MCID	Minimum Clinically Important Difference
MD	Mean Difference, more correctly Difference in Means
MESE	Multi-echo Spin Echo
MI	Mutual Information
MLE	Maximum Likelihood Estimation
MLE_{ncX}	Maximum Likelihood Estimation of Noncentral Chi Signal (Algorithm)
MLE_{Rice}	Maximum Likelihood Estimation of Rician Signal (Algorithm)
MLP	Multi-layer Perceptron Classifier
MNI	Montreal Neurological Institute
MNI-BIC	McConnell Brain Imaging Centre of MNI
MNI152	MNI Standard-space Brain Template (Average of 152 T_1-weighted Structural Scans)
MINC	Medical Image Network Common Data (.mnc) Format
MNI-ICBM152	Same as the 'MNI152' above
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
N3	Nonparametric Nonuniform Intensity Normalisation (Software)
NCEXP	Noise-Corrected Exponential Model (Algorithm)
$NCEXP_\sigma$	Noise-Corrected Exponential Model with Known σ (Algorithm)
NDORMS	Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
NIfTI	Neuroimaging Informatics Technology Initiative, or Medical Image NIfTI (.nii) and compressed NIfTI (.nii.gz) formats

(Continued on the next page)

(Continued)

Nilearn	Machine learning for Neuroimaging in Python (Software)
Nipy	Neuroimaging in Python (Software)
Nipype	Neuroimaging in Python Pipelines and Interfaces (Software)
NMI	Normalised Mutual Information
NLS	Non-linear Least Squares (Algorithm)
'None'	Neither Incident Radiographic Osteoarthritis Nor Pain, or Neither Radiographic Osteoarthritis Nor Pain Progression (Outcome)
OA	Osteoarthritis
OAI	Osteoarthritis Initiative
OARSI	Osteoarthritis Research Society International
OFFSET	Offset Model (Algorithm)
OxteoFlow	Oxford MRI Preprocessing Workflow for Osteoarthritis (Software/Algorithm developed in Chapter 3)
OxteoNet	Oxford MRI Deep Learning Network for Osteoarthritis Prediction (Software/Algorithm developed in Chapter 4)
'Pain'	Incident Pain Only, or Pain Progression Only (Outcome)
PDF	Probability Density Function
QA	Quality Assurance
qMR	Quantitative Magnetic Resonance
qMRI	Quantitative Magnetic Resonance Imaging
R	A language and environment for statistical computing and graphics
R	Pearson Product-Moment Correlation Coefficient, it is the 'R' in R^2
R^2	Coefficient of Determination, pronounced 'R squared'
R_2	Transverse Relaxation Rate, equals to reciprocal of T_2
R_2^*	Effective Transverse Relaxation Rate, equals to reciprocal of T_2^*
ResNet	Residual Neural Network (Algorithm)
RF	Radiofrequency
RFC	Random Forests Classifier
ROA	Radiographic Osteoarthritis

(Continued on the next page)

(Continued)

'ROA'	Incident Radiographic Osteoarthritis Only, or Radiographic Osteoarthritis Progression Only (Outcome)
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SD	Standard Deviation
SE	Spin Echo, or Standard Error
SEM	Standard Error of the Mean
SMD	Standardised Mean Difference, equivalent to Effect Size
SNR	Signal-to-Noise Ratio
SNR_0	Signal-to-Noise Ratio at Zero Echo
SQEXP	Squared Exponential Model (Algorithm)
$SQEXP_\sigma$	Squared Exponential Model with Known σ (Algorithm)
SyN	Symmetric Normalisation (Algorithm)
SPM	Statistical Parametric Mapping (Method or Software)
SVC (Linear)	Support Vector Classification with a linear kernel function
SVC (Poly)	Support Vector Classification with a polynomial kernel function
SVC (RBF)	Support Vector Classification with a rbf kernel function
SVC (Sigmoid)	Support Vector Classification with a sigmoid kernel function
T	Tesla
T_1	Longitudinal Relaxation Time
T_2	Transverse Relaxation Time
T_2^*	Effective Transverse Relaxation Time
ΔT_2	Difference in T_2 Values
$\Delta T_2^{individual}$	Actual Difference in T_2 Values Between OA and Healthy ROIs
ΔT_2^{META}	Estimated Difference in T_2 Values Between OA and Healthy Samples in Meta-analysis
$\Delta T_2^{population}$	Actual Difference in T_2 Values Between OA and Healthy Populations
TE	Echo Time
TFE	Turbo-field-echo

(Continued on the next page)

(Continued)

TI	Inversion Time
TN	True Negative
TP	True Positive
TR	(Pulse) Repetition Time
TRUNC	Truncation Model (Algorithm)
VBR	Voxel-Based Relaxometry
VBR^{T_2}	Voxel-Based T_2 Relaxometry
WORMS	Whole-Organ Magnetic Resonance Imaging Score
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index

Chapter 1

Background and Motivation

1.1 Background

In terms of disability burden [1] related pain and loss of function [2] and the strong causal link with other major diseases (for example, depression [3] and cardiovascular disease [4], musculoskeletal diseases are a major issue for UK and global health. Musculoskeletal conditions are the largest single cause of years lived with disability, and the third-largest cause of disability adjusted life years [1; 5].

Foremost among musculoskeletal disorders is joint disease, and osteoarthritis (OA) in particular: data from Arthritis Research UK [2] show that one-third of people aged 45 years and over have sought treatment for OA. Effective treatment, however, is struggling to keep pace with the demand for it, particularly in early stages of OA.

In this context, the 'treatment gap' is an emerging concept of considerable importance. A treatment gap is a protracted period in which the patient experiences debilitating pain, but are unresponsive to conservative care and yet are unsuitable for the invasive surgical procedures for end-stage disease. The typical knee osteoarthritis treatment gap extends 20 years although the younger osteoarthritis patient is faced with the treatment

gap throughout the majority of their adult life [6]. Based on Arthritis Research UK data, this equates to approximately 630,000 people in the UK, and is predicted to rise to over 1,000,000 people by 2035 [2].

1.2 General Aims

The combination of large burden and limited treatment options for patients in the treatment gap leads to a need to improve our understanding of how it manifests and progresses in its early stages, and to link this understanding to individual patients.

We expect that a better understanding will lead to better treatment options. If the clinician is able to better predict the progression of both symptoms (primarily pain) and structural damage, we argue that better patient management decisions can be made.

This thesis therefore aims to develop a platform technology for MRI assessment that can be used to explore the manifestation of disease, the relationships between changes within the tissue, and the initiation and progression of disease. Finally, we aim to to use this platform to underpin a new predictive tool to aid clinical decisions and discussions.

Interdisciplinary Writing

This work was undertaken in the Medical Sciences, is written for a readership within, and aims to deliver an impact for, that field. The thesis therefore takes time in describing technical details for that audience. The engineering reader familiar with the theory may wish to skip these explanations. Methods are similarly presented to allow reproduction of the results by researchers within the Medical Sciences field.

1.3 State of the Art

1.3.1 Early Detection of Osteoarthritis

Early detection of OA, where changes to the affected cartilage components may still be reversible, is of great interest to orthopaedic surgeons, physicians, and researchers working in orthopaedics. As can be seen in Fig. 1.1, in early stage of OA, proteoglycans (PGs) are depleted, the collagen matrix begins to break down. As a result, water content increase [7].

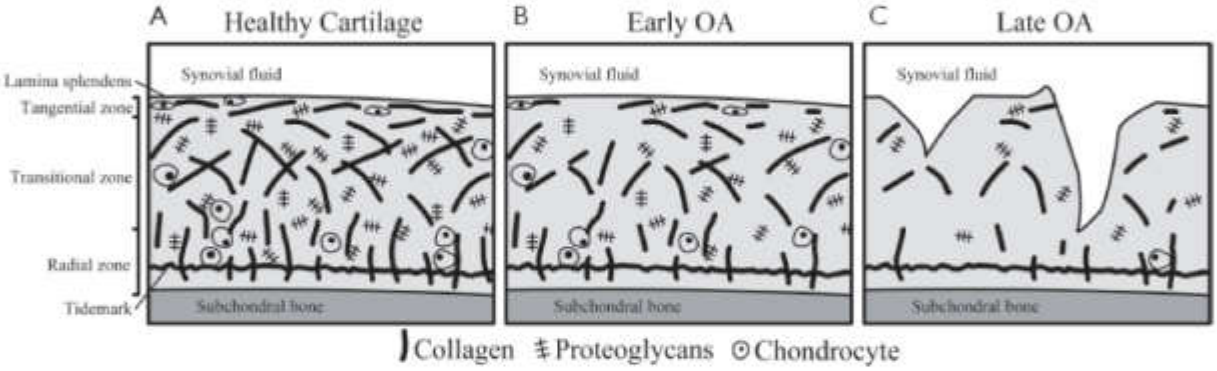


Figure 1.1: Representation of Cartilage Composition at Three Stages of OA. Image Reproduced from [8]

Early detection may allow, as we argued in Section 1.2 on the preceding page, an understanding of how OA is likely to progress in a given patient, which will be critical for clinicians in making better treatment decisions, and also communicating them to the patient [9; 10]. Currently, however, there are limited non-invasive imaging-based clinical tools available which are sensitive enough to assess the changes in cartilage associated with the earliest stages of OA [11].

Arthroscopy is frequently used for in vivo diagnosis of cartilage degeneration as it allows for the direct assessment of the cartilage surface integrity and the qualitative assessment and grading of cartilage stiffness [12]. However, the evidence base for the efficacy and cost-effectiveness of arthroscopic surgery is weak. Clinical trials have found arthroscopic surgery is associated with harms and do not support the practise of it [12–14].

Radiographs are commonly used in initial assessment to evaluate decreased joint space associated with cartilage thinning, but this approach is sensitive to moderate to severe structural cartilage loss, rather than biochemical change earlier than that..

Conventional morphological magnetic resonance imaging (MRI) similarly allows for macroscopic anatomic assessment of the cartilage but is less sensitive to the early biochemical changes illustrated in Fig. 1.1.

In recent years, quantitative MRI techniques, such as T_2, T_2^*, delayed gadolinium-enhanced MRI of cartilage (dGEMRIC), sodium imaging, chemical exchange saturation transfer (CEST), diffusion weighted imaging (DWI) and $T_1\rho$ imaging has demonstrated the ability to detect these changes prior to the morphological degradation of cartilage that follows as OA progresses. Thus, these techniques provide numeric measures that correlate with various physiological properties of cartilage and can be used to detect disease earlier than conventional imaging methods [15].

Overviews of various quantitative MRI techniques have been summarised in various review articles, among which the most informative ones include [8; 15–18].

1.3.2 Cartilage T_2 Mapping

T_2 mapping is currently the most widely used in the clinical setting. Cartilage T_2 mapping uses interstitial water as a probe to study the structural integrity of the extracellular matrix. Changes of the spatial distribution of the T_2 values reflect areas of tightly or loosely constrained water, a major indicator of function [19–21]. OA-related biochemical changes of the cartilage, including PGs loss, increased water content, and deterioration of the collagen network (Fig. 1.1) can in theory be extracted from the T_2 measurement [22].

Cartilage T_2 mapping therefore has considerable potential for practical application. It can be used to diagnose 'at risk' cartilage and cartilage with irreversible damage. Furthermore,

it can be used as a non-invasive tool to study cartilage integrity or post-cartilage repair procedures [23].

1.4 Narrative Review

To better understand how T_2 relates to tissue quality, a narrative literature review aiming to link T_2 signal and cartilage tissue structure has been performed in the initial stage of this DPhil. As this thesis emphasises the systematic approach to review the topic (in Chapter 2), the method and results of the narrative review is only discussed in brief here, but the details can be found in [24].

After a comprehensive search[1] in PubMed, 70 studies were found, then all the manuscripts were screened and irrelevant results removed. As a result, 20 publications were found to be most relevant, and they highlighted a number of key points:

Firstly, reports of efficiency are highly variable, for example:

1) Some authors suggest that $T_1\rho$ is more sensitive than T_2 in terms of detecting early OA [25; 26]. Others found that only certain type of T_2 sequence have the ability to detect early OA [27].

2) When correlating with histology, different centres have obtained contrary results. There were both 'positive' [28] and 'weak' [26; 29] correlations reported.

3) Contrary results were also reported when correlating T_2 values with cartilage's mechanical properties. Again, there were both 'significant' [30] and 'weak' [31] correlations reported.

4) T_2's correlation with sulphated glycosaminoglycans (sGAG) content was not significant [32], but different from what is commonly believed, it can be an indicator of PGs

[1]The search query used was: (((mri[Title]) AND T2[Title/Abstract]) AND cartilage[Title]) AND (immuno* OR mankin OR oars OR hhgs OR (bern score) OR ICRS* OR collagen OR glycosaminoglycan OR GAG OR proteoglycan OR aggrecan OR birefringence OR spectroscop* OR stain* OR stiffness OR *histolog* OR structure OR biochem* OR mechanic*[Title/Abstract])

content [33], although maybe not as sensitive as $T_1\rho$ [26].

Secondly, 'positive' results are predominantly under highly controlled laboratory conditions, specifically using cartilage cores in small-bore, high field scanners, which do not necessarily reflect the reality of clinical scans [30].

Finally, in addition to the inconsistent physical link between T_2 signals and tissue structure, all in-vivo scans in the key papers required manual segmentations. The review highlighted that application of T_2 (and any other quantitative MRI) for OA is limited by the absence of automated and consistent analysis tools.

1.5 Specific Objectives and Approach

The findings of the above narrative review indicate a need to better understand the inconsistency of clinical T_2, and the requirements/limitations for using this technique in decision making. The **Part I (Section 2.2) of Chapter 2** therefore performs a systematic review and meta-analysis on clinical studies to determine the difference in T_2 signals from normal and osteoarthritic tissue.

Among all the factors affecting the diagnostic and prognostic performance of T_2 mapping (for example, magic angle effect, partial volume effect, etc.), one that is most important, yet are most manageable in terms of quality control and computation is the noise in imaging. **Part II (Section 2.3) of Chapter 2** therefore evaluated sixteen different T_2 calculation methods' fitting performance under different noise levels, which showed the small but important ΔT_2 between normal and osteoarthritic tissue cannot be observed without a controlled noise (SNR) level, and a correct modelling of the noise during fitting.

The better fitting in terms of trueness, however, could be meaningless as Chapter 2 shows low SNR in T_2 scan will surely violate the bottom line criteria for precision in a clinical setting. Interpretation of absolute T_2 value, and ΔT_2 could therefore of little use. This

thesis argue that a new approach, which looks for patterns of signals in the T_2 map could provide a better diagnostic performance.

Such a pattern analysis approach requires the ability to compare scans of different anatomical shape in a same space. Translating the analytical methods used by neuroimaging community, this thesis successfully solved this problem by developing OxteoFlow in **Chapter 3**. OxteoFlow is a preprocessing workflow optimised specially for musculoskeletal MRI. It enabled the same-space comparison required for pattern analysis, and laid the foundation of our platform technology for MRI assessment.

Using the outputs of OxteoFlow, the thesis then explores patterns of T_2 associated with the manifestation and progression of OA. **Part I (Section 4.2) of Chapter 4** preprocessed (using OxteoFlow) and stratified the knee MRI data from the Osteoarthritis Initiative (OAI), the largest and most widely studied musculoskeletal database, into the incidence (the OAI way of expressing 'onset', see Table 4.3 on page 138) and progression outcome cohorts, followed by a voxel-based pattern analysis on the baseline T_2 maps of different outcome groups.

Finally, this thesis explored using T_2 mapping to predict ROA progression, in **Part II (Section 4.3) of Chapter 4**(Section 4.3), with traditional machine learning models and a state-of-the-art deep learning model developed in this DPhil—OxteoNet. As another module in the platform technology for MRI assessment, OxteoNet demonstrated great potential to aid radiologists and clinicians in screening and decision making.

In summary, the logical flow of this thesis, and its relation to functional MRI (fMRI) analysis in the neuroscience field, is given in the Flowchart Abstract of the Thesis on page ii as an overview.

This page intentionally left blank

Chapter 2

T_2 Mapping for Osteoarthritis: Systematic Review, Meta-Analysis and Evaluation of Fitting Methods

2.1 Introduction

T_2 is widely used to assess cartilage 'quality'. Conceptually, such measurements of the freedom of water motion should be a useful indicator of cartilage function, which is strongly tied to the control over fluid flow [19]. Strong correlations with disease state have been observed in the highly controlled, small-bore and high field conditions used in vitro [30].

T_2 measurements are widespread in the clinical setting, however performance under these conditions is less clear. Clinical measurements are complicated by the signals' decay into a relatively high-amplitude noise floor (Fig. 2.1 on the following page), making fitting a significant challenge. This is further compounded by the uncertainties introduced in the large volume, heterogeneous, anisotropic and living samples. As a result, the range of T_2 values of normal and diseased cartilage, and their difference, remain undefined and are highly variable in the literature. Interpretation of T_2 mapping, though technically a

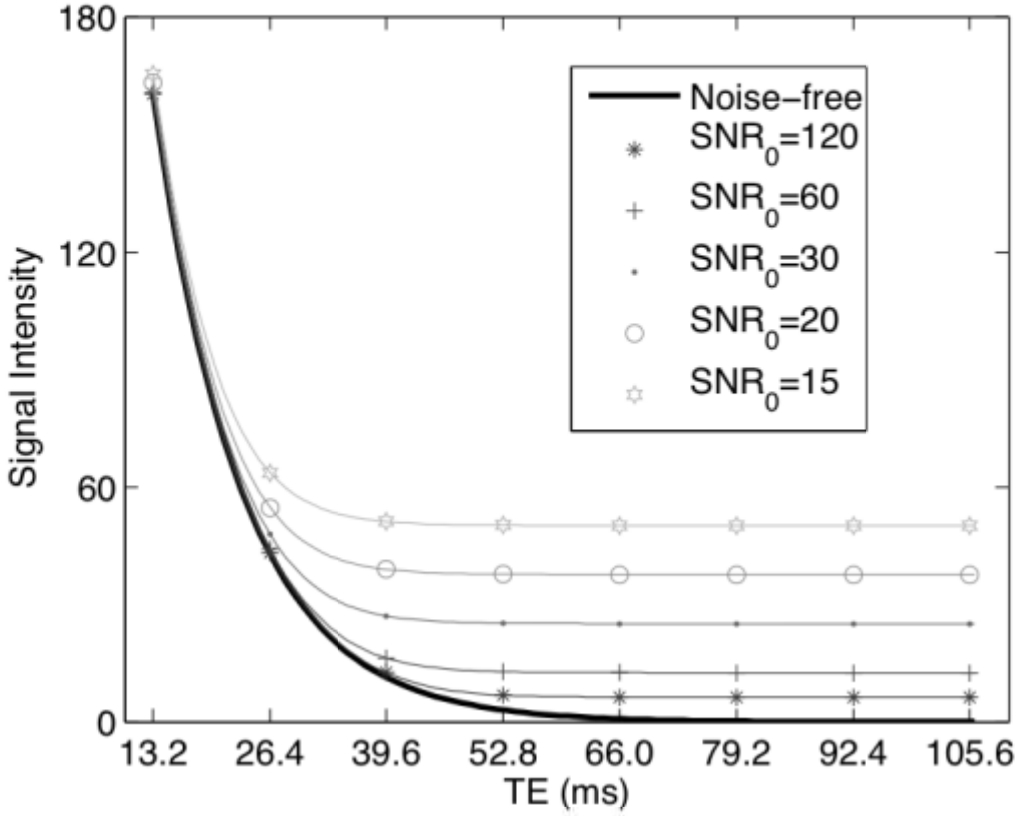


Figure 2.1: T_2 Decay Affected by Noise. Reproduced from [34]

quantitative measurement, are thus in practice qualitative and subjective.

Before the central goals of this thesis, advancing the diagnostic and prognostic value of T_2 mapping for OA, can be achieved, we must first define the clinical difference between normal and osteoarthritic cartilage, and minimise uncertainty in the calculation itself. This chapter therefore examine this difference, and the ability to resolve it, via a systematic review and meta-analysis. It then assess the performance of available fitting method in detecting/resolving differences between normal and OA cartilage, in the context of applications in clinical trials and individual patients.

2.2 Defining the Standard Case-Control T_2 Difference for Osteoarthritis with Systematic Review and Meta-Analysis

2.2.1 Methods

This study followed Cochrane Handbook for Systematic Reviews of Interventions [35] as the guideline for the meta-analysis.

2.2.1.1 Eligibility Criteria

Original studies that conducted T_2 measurements on subjects with and without osteoarthritis of the knee and/or hip, assessed by radiographic Kellgren-Lawrence Grade (KLG) [36] or Whole-Organ Magnetic Resonance Imaging Score (WORMS) [37] were included. To find a clinically relevant T_2 difference between OA and healthy populations, this study included case-control studies in which participants with and without OA were scanned and compared at a point in time or over a short period, and excluded longitudinal studies in which baseline healthy data and the follow-up data of the same group were compared. It also excluded studies in which no healthy subjects were included, or only healthy subjects were included (subjects with OA risk factors but without radiographic or MRI evidence). Studies using other cartilage grading methods were also excluded.

2.2.1.2 Literature Search and Study Selection

Systematic searches in PubMed and Web of Science were performed in April 2016. Due to the advances in MR techniques, only studies published in 2000 or later were included. No search restrictions was set for study size, subject age, follow-up time, imaging protocol, MR field strengths, or language. Although the specific T_2 fitting methods used and SNR of the MRI scans are important information in the context of this thesis, few studies revealed them and there would be a risk of bias to exclude the majority of studies that

concealed these information. To avoid this risk all studies that meet the eligibility criteria were included, regardless of whether information about specific T_2 fitting methods and/or SNR of the MRI scans were available. To identify additional studies, the reference list of the included studies were also reviewed. Full text of all the eligible studies were obtained for data extraction. The eligible studies, scoring methods used and the number of OA and healthy subgroups, are summarised in Table 2.1.

Table 2.1: Summary of Eligible Studies for Meta-analysis

Author(s) and Year	Scoring Method	N of OA Subgroup(s)	N of Healthy Subgroup(s)	Within-study Data Pooling
Knee				
Multanen et al., 2015 [38]	KLG	2	1	Required
Zuo et al., 2014 [39]	KLG	2	1	Required
Liebl et al., 2014 [40]	KLG	1	1	
Prasad et al., 2013 [41]	KLG	1	1	
Jungmann et al., 2013 [42]	WORMS	3	2	Required
Mosher et al., 2011 [43]	KLG	2	1	Required
Zarins et al., 2010 [44]	WORMS	2	1	Required
Bolbos et al., 2008 [25]	KLG	1	1	
Stahl et al., 2007 [45]	KLG	1	1	
Li et al., 2007 [46]	KLG	4	1	Required
Koff et al., 2007 [47]	KLG	4	1	Required
Dunn et al., 2004 [48]	KLG	2	1	Required
Hip				
Wyatt et al., 2015 [49]	KLG	1	1	
Subburaj et al., 2013 [50]	KLG	1	1	
Yamamoto et al., 2011 [51]	KLG	1	1	
Nishii et al., 2008 [52]	KLG	1	1	

2.2.1.3 Data Extraction and Processing

For all eligible studies, mean T_2 , standard deviation (SD) of T_2 and number of subjects (N) in each group (or subgroups) were extracted from the original studies. Where values are not available but a plot was presented, the values were retrieved with GetData Graph Digitizer [53] from the plots.

Within-study data pooling

In this study, the population were divided into two subgroups: healthy and OA. For studies using KLG, healthy controls are defined as subjects with a KLG of 0 or 1, and radiographic OA (ROA) patients as subjects with a KL grade 2 or higher (see also Table 4.3 on page 138). For studies using the WORMS grade, a WORMS grade of 0 or 1 was considered as healthy and grade 2 or higher OA.

As can be seen in Table 2.1, some studies reported the individual KL or WORMS grades separately, resulting in more than one OA and/or healthy subgroups. In these cases, pooled (combined) Mean, SD and N of the OA or healthy subgroups within theses studies are calculated. If the subgroups within a study are indexed $i = 1, 2, \ldots, k$, then the pooled Mean, SD, and N can be computed by [54]:

$$Mean_{pooled} = \frac{\sum_{i=1}^{k} Mean_i \mathrm{N}_i}{\sum_{i=1}^{k} \mathrm{N}_i} \tag{2.1}$$

$$SD^{*}_{pooled} = \sqrt{\frac{\sum_{i=1}^{k}(N_i - 1)SD_i^2}{\sum_{i=1}^{k}(N_i - 1)}} \tag{2.2}$$

$$N_{pooled} = \sum_{i=1}^{k} N_i \tag{2.3}$$

where $Mean_{pooled}$ is the weighted arithmetic mean, SD^{*}_{pooled} the weighted and pooled SD (same as Eq. (2.17) on page 22 when $k = 2$, the $*$ is used to differentiate with SD_{pooled} in Eq. (2.16) on the same page), and N_{pooled} a simple sum of N in the subgroups. The result after the with-in study data pooling is presented in Table 2.2 on the next page, where studies with and without the need of data pooling (see Table 2.1 on the facing page) are shown in compartments. The pooled data are indicated by $(N_1 + N_2 + \ldots + N_k)$ in the N columns.

Table 2.2: With-in Study Data Pooling and Compartmental Data Selection for Meta-analysis

		OA Pooled			Healthy Pooled				
Study	Compartment	Mean	SD	N	Mean	SD	N	*P*	Mean Difference
Knee									
Multanen et al. 2015	MF	50.01	5.70	78(25+53)	50.40	2.80	12	0.709	-0.39
	MT	45.04	4.40	78(25+53)	44.20	5.20	12	0.604	0.83
	LF	50.43	5.73	78(25+53)	50.80	5.40	12	0.829	-0.36
	LT	42.48	7.06	78(25+53)	38.40	2.40	12	**0.0003**	4.08
Zuo et al. 2014	Global	52.43	8.07	42(22+20)	47.30	4.50	50	0.0005	5.13
Liebl et al. 2014	MF	40.38	3.23	50	38.88	2.93	80	0.009	1.5
	MT	32.34	2.39	50	31.4	3.16	80	0.057	0.94
	LF	36.02	2.70	50	34.96	2.37	80	0.025	1.06
	LT	30.43	3.12	50	28.91	2.39	80	**0.004**	1.52
	PAT	35.03	4.57	50	33.45	2.65	80	0.030	1.58
Prasad et al. 2013	MF	33.44	3.18	27	30.43	2.68	28	0.0004	3.01
	MT	30.52	3.93	27	26.17	4.43	28	0.0003	4.35
	LF	35.37	4.26	27	31.69	2.76	28	0.0005	3.68
	LT	28.34	4.77	27	25.00	3.43	28	0.0046	3.34
	PAT	37.46	6.02	27	33.03	2.76	28	0.0013	**4.43**
Jungmann et al. 2013	Global	33.50	1.80	163(135+6+22)	32.83	1.79	82(72+10)	0.0065	0.67
Mosher et al. 2011	F Deep	40.71	2.37	32(16+16)	38.89	2.06	18	0.007	1.82
	F Middle	42.79	2.21	32(16+16)	40.58	1.74	18	0.0003	2.21
	F Superficial	46.79	2.37	32(16+16)	45.59	1.92	18	0.041	1.20
	T Deep	41.54	1.91	32(16+16)	41.54	2.04	18	1.000	0
	T Middle	43.92	2.05	32(16+16)	43.76	1.84	18	0.779	0.16
	T Superficial	48.47	2.15	32(16+16)	47.49	1.64	18	0.216	0.98
	PAT Deep	39.32	2.24	32(16+16)	35.18	1.86	18	**$\ll$ 0.0001**	4.14
	PAT Middle	41.80	2.55	32(16+16)	37.90	2.07	18	$<$ 0.0001	3.90
	PAT Superficial	45.90	3.01	32(16+16)	42.37	2.78	18	0.0002	3.53

(Continued on next page)

(Continued)

Study	Compartment	OA Pooled			Healthy Pooled			P	Mean Difference
		Mean	SD	N	Mean	SD	N		
Zarins et al. 2010	MF	33.13	4.86	40(17+23)	32.14	3.35	23	0.344	0.99
	MT	30.56	4.77	40(17+23)	28.14	4.03	23	**0.036**	2.42
	LF	33.18	2.70	40(17+23)	32.63	2.94	23	0.466	0.55
	LT	29.13	4.89	40(17+23)	27.73	4.53	23	0.257	1.40
Bolbos et al. 2008	MF	35.50	2.23	16	31.89	3.09	16	0.0008	3.61
Stahl et al. 2007	MF	44.35	5.61	8	45.37	6.56	10	0.727	-1.02
	MT	47.49	4.64	8	45.12	5.05	10	0.316	2.37
	LF	45.33	6.03	8	40.56	6.09	10	0.117	4.77
	LT	44.41	7.27	8	39.62	4.38	10	0.129	**4.79**
Li et al. 2007	Global	39.72	1.56	10(2+3+3+2)	34.70	2.50	10	< 0.0001	5.02
Koff et al. 2007	PAT	66.60	9.41	184(89+64+27+4)	65.50	6.10	39	0.361	1.10
Dunn et al. 2004	**MF**	40.25	0.70	48(20+28)	34.90	1.00	7	$\ll$ **0.0001**	5.35
	MT	35.46	0.79	48	32.10	1.40	7	0.0006	3.36
	LF	39.35	0.64	48	35.00	1.10	7	< 0.0001	4.35
	LT	34.46	0.80	48	34.90	1.80	7	0.546	-0.44
Hip									
Wyatt et al. 2015	F	31.60	3.30	30	32.20	3.20	54	0.423	-0.60
	A	28.10	3.50	30	27.30	2.70	54	**0.283**	0.80
Subburaj et al. 2013	**F**	33.90	3.10	9	31.10	1.70	12	**0.031**	2.80
	A	31.7	2.50	9	31.00	2.30	12	0.520	0.70
Yamamoto et al. 2011	F	34.4	3.10	15	30.80	1.20	28	0.0005	3.60
Nishii et al. 2008	F	28	3.70	12	29.40	3.00	14	0.306	-1.4
	A	37.10	12.00	12	33.40	4.50	14	**0.331**	3.7

Knee compartments: F(femoral), T(tibial), PAT(patellar)
MF(medial femoral), LF(lateral femoral), MT(medial tibial), LT(lateral tibial)
Hip compartments: F(femoral head), A(acetabular)

Compartmental Data Selection

Although the the data of different compartments in Table 2.2 on page 14 are all eligible for meta-analysis, this is not considered optimal because data in one compartment of a study might be less significant than other compartments, which itself is one of the reasons why the original studies performed measurements and reported in compartments, rather than using the cartilage as a whole.

Therefore, only one compartment from each of these studies were included in the meta-analysis. One assumption made here is to consider the mean difference (MD, more correctly 'difference in means' [35]) in T_2 between healthy and OA population as a standard value that can be estimated with Eq. (2.4). In a study, MD can be more significant in some compartments but less significant in other compartments, due to the different severity of disease and other measurement errors.

$$MD = Mean_{OA} - Mean_{Healthy} \tag{2.4}$$

With the descriptive statistics (Mean, SD, N) of both the healthy and OA groups available, the Welch's t-test (unequal variances t-test) [55] was then performed to test the significance of the MD of the compartments in each of the studies. Here, Welch's t-test was used over Student's t-test for the follow reasons:

(1) The OA and healthy groups were considered independent (unpaired) samples, because we only included case-control studies;

(2) The sample size N in the OA and healthy groups are unequal, in which case a Welch's t-test is more suitable; [55]

(3) Equal variance in the OA and healthy groups can not be assumed. Actually, it can be expected that the variance be very different because the OA group includes all the cases that KLG or WORMS $\geqslant 2$, while the healthy group includes only cases that KLG or WORMS $= 0$ or 1;

(4) Due to (3), a Levene's test [56] must be performed to assess the equality of variances in the two groups if we were to perform a Student's t-test. This is an unnecessary step for Welch's t-test, as it does not assume equal variance in groups [55].

The Welch's t-test was performed in Python by inputting the individual descriptive statistics into the `scipy.stats.ttest_ind_from_stats(Mean1, SD1, N1, Mean2, SD2, N2, equal_var=False)` function [57]. Note that `equal_var=False` is a required parameter, because the default `equal_var=True` performs Student's t-test. The calculated two-tailed P values are listed in Table 2.2 on page 14. Based on the calculated P value and MD we thus use the following decision rules for compartmental data selection.

Decision Rules for Compartmental Data Selection:

I. Remove any compartment with a negative MD;

II. Find the the smallest P in the remaining compartment(s);

(1) If this compartment with the smallest P happens to have the largest positive MD, accept this compartment with smallest P (bold $\boldsymbol{P}$).

(2) If this compartment do not have the largest positive MD, compare its smallest P with the P of the compartment with the largest MD:

a. if the difference between this smallest P and the P of the compartment with the largest MD is within the range of 0–0.001 inclusive, the P difference is considered small and is negligible, accept the compartment with the largest MD (bold **MD**).

b. if the difference between this smallest P and the P of the compartment with the largest MD is larger than 0.001, this P difference is not negligible, still accept the compartment with the smallest P despite a larger MD is available in other compartment (bold $\boldsymbol{P}$).

Data was selected from one compartment per patient in each of the studies. The compartment selected and the P or MD used to make the selection decision is in **bold font**

in Table 2.2 on page 14. Note that the meta-analysis does not consider which specific compartments are selected. The purpose of this compartmental selection step is simply to refine data to the compartment in which OA-related change is observed, removing any presupposition of phenotype or disease pattern and ensuring that a single, independent datum per patient is used. These selected data may or may not be the same as the most significant finding reported in their original studies, due to the difference in the approach to dividing the groups and statistical inference.

The final data, after within-study data pooling and compartmental data selection, is presented in Table 2.3. The final data in Table 2.3 is saved in a 'meta_data.csv' file which will be read by Program Code A.1 on page 253.

Table 2.3: Final Data for Meta-analysis

	OA Pooled			Healthy Pooled		
Study	Mean	SD	N	Mean	SD	N
Knee						
Multanen et al. 2015	42.48	7.06	78(25+53)	38.40	2.40	12
Zuo et al. 2014	52.43	8.07	42(22+20)	47.30	4.50	50
Liebl et al. 2014	30.43	3.12	50	28.91	2.39	80
Prasad et al. 2013	37.46	6.02	27	33.03	2.76	28
Jungmann et al. 2013	33.50	1.80	163(135+6+22)	32.83	1.79	82(72+10)
Mosher et al. 2011	39.32	2.24	32(16+16)	35.18	1.86	18
Zarins et al. 2010	30.56	4.77	40(17+23)	28.14	4.03	23
Bolbos et al. 2008	35.50	2.23	16	31.89	3.09	16
Stahl et al. 2007	44.41	7.27	8	39.62	4.38	10
Li et al. 2007	39.72	1.56	10(2+3+3+2)	34.70	2.50	10
Koff et al. 2007	66.60	9.41	184(89+64+27+4)	65.50	6.10	39
Dunn et al. 2004	40.25	0.70	48(20+28)	34.90	1.00	7
Hip						
Wyatt et al. 2015	28.10	3.50	30	27.30	2.70	54
Subburaj et al. 2013	33.90	3.10	9	31.10	1.70	12
Yamamoto et al. 2011	34.40	3.10	15	30.80	1.20	28
Nishii et al. 2008	37.10	12.00	12	33.40	4.50	14

2.2.1.4 Models for Meta-Analysis

The Fixed Effect Model

There are two common models for carrying out meta-analysis—the fixed effect model and the random effect model [35]. The fixed effect model assumes a single, fixed but unknown underlying population mean, and every study estimates this single value. The meta-analysis combines data from different studies to give a point estimate of effect size (ES) and its confidence interval (CI). Suppose k studies are enter into a fixed effect model meta-analysis, the overall point estimate of ES is calculated as a weighted average of the study means [58]:

$$ES = \frac{\sum_{i=1}^{k} W_i Mean_i}{\sum_{i=1}^{k} W_i} \tag{2.5}$$

where W_i is the weight for study i, calculated as:

$$W_i = \frac{1}{Var_i} = \frac{1}{SE_i^2} = \frac{1}{\left(\frac{SD_i}{\sqrt{N_i}}\right)^2} \tag{2.6}$$

where Var_i is the variance, and SE_i the standard error of study i. The variance of ES is estimated as:

$$Var_{ES} = \frac{1}{\sum_{i=1}^{k} W_i} \tag{2.7}$$

and the 95% CI of the point estimate ES can be constructed to be:

$$\left[ES - 1.96\sqrt{Var_{ES}}, ES + 1.96\sqrt{Var_{ES}}\right] \tag{2.8}$$

As can be expected, large and precise studies are weighted more heavily than small and imprecise studies.

The Random Effect Model

A set of studies is homogeneous if sampling variability can reasonably account for the variation among studies, and heterogeneous if variation among studies is larger than this. When there is heterogeneity that cannot be readily be explained, one analytical

approach is to incorporate it into a random effect model, which assumes that there is a distribution of population parameters, and the effects beings estimated in the different studies are not identical (each denoted ES_i), but follow this distribution. The overall ES^* for the population is then estimated using the individual ES_i values sampled from this distribution [58]. The model represents a lack of knowledge about why the effects differ in different studies as if they were random [35], which is exactly the case for T_2 values of OA and healthy populations reported in the included studies.

The equations for calculating the ES^* and Var_{ES^*}, and the 95% CI are very similar to Eqs. (2.5), (2.7) and (2.8) on the preceding page, except that the weights W_i^* in these equations are calculated in a way different from Eq. (2.6) on the previous page [58]:

$$W_i^* = \frac{1}{Var_i^*} = \frac{1}{Var_i + T^2} = \frac{1}{\left(\frac{SD}{\sqrt{N_i}}\right)^2 + T^2} \tag{2.9}$$

where Var_i is the variability of $Mean_i$ around ES_i, and T^2 (not to be confused with T_2) is the variability of ES_i around the overall population ES. Large T^2 indicates that heterogeneity, or variability between study means is large, and a random effect model is needed. The details of how T^2 is calculated are too complicated to be discussed here, but could be found in [59].

Although the aim of this study is to produce a reference value of T_2 difference between OA and healthy populations, it would be naive to think this value is a fixed constant, and the observed differences among studies are solely due to chance. Instead, this variation of T_2 difference is expected to be affected by many factors other than chance, such as MR scanning, image processing and disease severity—factors that would form a distribution of T_2 difference rather than a fixed constant. That is, the set of studies included in the meta-analysis are heterogeneous and the random effect model should be used.

2.2.1.5 Mean Difference

With the random effect model, the mean difference (MD) is calculated, allowing combination of the MD_i results reported in the individual studies as an overall MD.

The estimation of MD in T_2 through meta-analysis, which normally estimates SMD, is possible because T_2 value, as a measurement of relaxation time, are reported in the same unit of ms in all the included studies.

2.2.1.6 Effect Sizes

To put it simply, the standardised mean difference (SMD, also known as 'effect size', especially when referring to versions of SMD [35]) is a measurement of how many SDs are the two groups separated. The SMD is estimated by dividing the MD by the standard deviation (SD) of the population from which the groups were sampled, i.e. [60]:

$$SMD = \frac{MD}{SD} \tag{2.10}$$

Depending on whether the population SD is known, SMD can be estimated in a number of different ways. Three different methods of estimating the population SD give rise to three of the better-known effect size indexes (versions of SMD) as follows [61; 62]:

$$\text{Cohen's } d = \frac{MD}{SD_{pooled}} \tag{2.11}$$

$$\text{Glass's } \Delta = \frac{MD}{SD_{control}} \tag{2.12}$$

$$\text{Hedge's } g = \frac{MD}{SD^*_{pooled}} \tag{2.13}$$

where

$$MD = Mean_1 - Mean_2 \tag{2.14}$$

(a generalisation of Eq. (2.4) on page 16)

$$SD_{pooled} = \sqrt{\frac{\sum_{i=1}^{N_1}(X_i - Mean_1)^2 + \sum_{i=1}^{N_2}(Y_i - Mean_2)^2}{N_1 + N_2 - 2}} \tag{2.15}$$

$$= \sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2}{N_1 + N_2}} \tag{2.16}$$

$$SD^*_{pooled} = \sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2}{N_1 + N_2 - 2}} \tag{2.17}$$

(same as Eq. (2.2)$|_{k=2}$ on page 13)

Choosing from the three equations requires an examination of the SD of each group. If the SDs are approximately the same it is reasonable to assume a common SD in these two populations, in which case the two SDs can be pooled using Eq. (2.15) to calculated the **Cohen's *d*** index of effect size. However, in practice Eq. (2.16) is often used to calculate the Cohen's d [63].

The **Glass's Δ** [64] index is used when the SDs of the two groups differ, then the homogeneity of variance assumption is violated and pooling the SDs is not appropriate. In this case the SD of the control group is used because it is not affected by the effects of the disease (or more often treatment in a typical meta-analysis) and therefore more closely reflect the population SD. Glass's Δ is used later in Section 2.3.2.6 on page 52 of this chapter when the simulation results are compared with the findings in the meta-analysis.

In the Cochrane Handbook, the SMD is defined as the **Hedge's *g*** [65] index of effect size, which was developed for removing a slight positive bias in the estimation of effect size using Cohen's d.[1] This bias can also be quantified by:

$$g \approx d\left[1 - \frac{3}{4(n_1 + n_2) - 9}\right] \tag{2.18}$$

[1]Terminology is inconsistent when referring to the original studies: (1) g was originally developed by Hedge, but was labeled as d at the time; (2) what was labeled as g in the 1980s is actually what is now Cohen's d. It was originally developed by Cohen but was later refined by Glass (hence g for Glass); (3) Later g became synonymous with 'Hedge's d' in (1), and finally was referred to as Hedge's g, rather than 'Hedge's h' because of (2). [61]

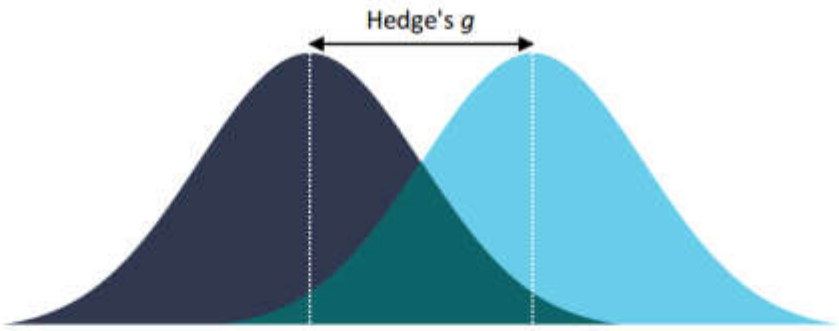


Figure 2.2: Illustration of Hedge's g

An illustration of Hedge's g can be found in Fig. 2.2. In this study, the Hedge's g estimator in Eq. (2.18) was used to calculate the effect size, following the standard procedure in Cochrane Handbook.

2.2.1.7 Meta-Analysis in R

The 'metafor' package in R [66] is utilised to perform the meta-analysis. Random effect model provided in the `rma.uni` function (with alias `rma`) was used to estimate the MD, and the Hedge's g estimator of effect size (`method="HE"`) was applied (discussed in the previous section) to estimate the effect size between OA and healthy population [67]. The R code for performing the meta-analysis of effect size is presented in Program Code A.1 on page 253. The code for meta-analysis of MD is the same except `measure="MD"`.

2.2.1.8 Interpreting Effect Sizes

Effects by themselves are meaningless unless they are contextualised against some frame of reference, such as a well-known scale. The Cochrane Handbook highlights three methods for re-expressing the effect sizes: converting to (log) odds ratio, multiplying effect size by a typical SD (SD of a representative study) to estimate a MD in a familiar scale, or using the Cohen's rules of thumb [68].

However, the effect in this study represents the diagnostic performance of an imaging technique (T_2 measurement), in which context the (log) odds ratio and SD multiplying

methods are rarely used. Authors in this field tend to use statistics such as sensitivity and the area under the receiver operating characteristic curve (AUC, or AUROC; also known as common language effect size (CL) and probability of superiority) to report the diagnostic performance of the imaging techniques. To compare the results in this study with the AUC and sensitivity values in the literature, the methods used to convert effect size to/from AUC and sensitivity are discussed below.

Converting Between Effect Size and AUC

AUC is meant to be more intuitive for people without any training in statistics. It gives the probability that a person picked at random from the case group will have a higher score than a person picked at random from the control group.

Assuming normality and equal-sized, equal-variance groups, AUC can be converted from/to effect size using [69]:

$$AUC = \Phi\left(\frac{ES}{\sqrt{2}}\right) \tag{2.19}$$

$$ES = \sqrt{2}\Phi^{-1}(AUC) \tag{2.20}$$

where Φ is the Cumulative Distribution Function (CDF) of the Standard Normal Distribution (mean $= 0$, SD $= 1$), and Φ^{-1} is the Quantile Function (or Percent Point Function), which is the inverse function of Φ.

Converting Between Effect Size and Sensitivity

As no formula in the literature can be found to convert sensitivity to/from effect size, a formula needs to be derived for this purpose.

Based on the same assumption used for converting effect size to AUC (normality, equal-sized, equal-variance groups), plus if the cut-off bar (threshold of group dividing) is set at where the two normal curves cross (at $\frac{ES}{2}$), and divide the distribution of results into true positive (TP), false positive (FP), true negative (TN), false negative (FN) areas as

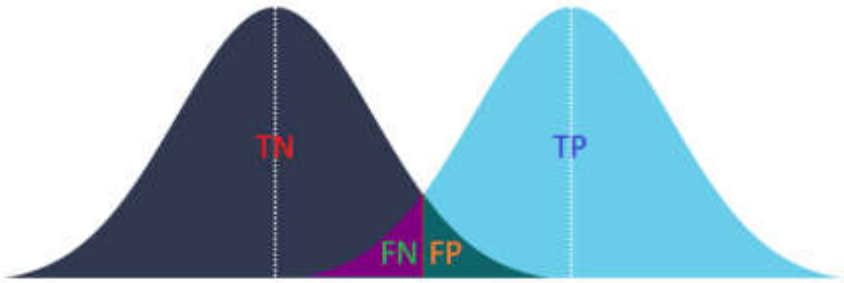


Figure 2.3: Illustration of the TP, FP, TN, and FN Areas Divided by the Distribution of Results and the Cut-off Bar

illustrated in Fig. 2.3, then effect size can be converted to/from sensitivity with:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.21}$$

$$= \frac{1 - 2\Phi\left(-\frac{ES}{2}\right)}{1 - \Phi\left(-\frac{ES}{2}\right)} \tag{2.22}$$

$$ES = -2\Phi^{-1}\left(\frac{1 - \text{sensitivity}}{2 - \text{sensitivity}}\right) \tag{2.23}$$

where Φ is the CDF of the Standard Normal Distribution, and Φ^{-1} is the Quantile Function. The correctness of these formulas will be discussed in the Results and Discussion section.

Cohen's Rules of Thumb

The controversial rules of thumb proposed by Cohen is as follows: 0.2 represents a small effect, 0.5 a medium effect, and 0.8 a large effect [63]. However, many methodologists have argued against such interpretation, because effect size is context-dependent and in some context small effects can be important and meaningful. Nevertheless, these interpretations are easy to grasp and the Cochrane Handbook encourages authors to consider including a rule of thumb in their reviews [68].

Using Eqs. (2.19) and (2.20) on the facing page for converting between effect size and AUC, Eqs. (2.22) and (2.23) for converting between effect size and sensitivity, and Cohen's Rules of Thumb above, it is now possible to convert result from other studies in this field as the benchmark to discuss the findings in effect size of T_2 measurement in OA.

2.2.1.9 Benchmarks for Interpretation

To compare the effect sizes found in this study against findings of others, we extracted effect sizes from some meta-analysis (systematic reviews) of similar purpose, and converted some significant AUC and sensitivity results identified in some original diagnostic performance evaluation studies. The studies used as benchmarks are:

(1) Systematic review and meta-analysis conducted by Menashe et al. in 2012 [70]: evaluated diagnostic performance of all the quantitative and semi-quantitative and compositional (dGEMRIC) MR techniques in OA, which reported AUC = 0.804 (using X-ray, clinical diagnosis, arthroscopy, histology, direct visualisation and CT scans as reference standard), and sensitivity = 60.7% (95% CI 0.497 to 0.708, using X-ray as the reference standard);

(2) Systematic review and meta-analysis conducted by MacKay et al. in 2018 [71]: evaluated diagnostic performance of $T_{1\rho}$ and T_2 for knee OA, which reported effect sizes of $ES = 0.73$ (95% CI 0.40 to 1.06) for discrimination between mild OA and control using $T_{1\rho}$. The study also reported a $ES = 0.49$ (95% CI 0.30 to 0.67) for discrimination between mild knee OA and control, and $ES = 1.24$ (95% CI 0.63 to 1.85) for discrimination between severe knee OA and control using T_2, which averaged to a point estimate of $ES = 0.87$ for discrimination between knee OA and control using T_2 measurement. All results used X-ray as the reference standard in that study;

(3) An original diagnostic value study conducted by Li et al. in 2009 [72]: evaluated diagnostic performance of dGEMRIC in knee OA, which reported an effect size of $ES = 1.34$ (using X-ray, clinical symptoms and MRI as reference standard); and

(4) An original diagnostic and prognostic value study conducted by Palmer et al. in 2017 [73]: evaluated diagnostic performance of dGEMRIC for hip OA, which reported AUC = 0.750 (95% CI 0.541 to 0.959) and sensitivity = 55.6%, if dGEMRIC ratio in a specific region of hip cartilage is used in combination with radiographic alpha angle and positive impingement test on hip, and using JSN as the reference standard.

By re-expressing effect size with AUC and sensitivity, and re-expressing AUC and sensitivity with effect size, we were able to compare our meta-analysis result against these benchmarks, and against Cohen's rules of thumb, in a common context (diagnostic performance) and in common scales (effect size, AUC and sensitivity). The comparison is presented in Table 2.5 on page 32.

2.2.2 Results and Discussion

2.2.2.1 ΔT_2^{META}—the Mean Difference in T_2

One of the most important result of this study, as can be seen in Fig. 2.4 on page 30, is that a point estimate and CI of the Mean Difference (MD) in T_2 (denoted $\boldsymbol{\Delta T_2^{META}}$) between OA and healthy populations was achieved by combining outcomes from studies of knee and hip:

$\boldsymbol{\Delta T_2^{META}}$ **(Knee) = 2.20** (95% CI 1.24 to 3.17) ms for knee,

$\boldsymbol{\Delta T_2^{META}}$ **(Hip) = 3.31** (95% CI 2.48 to 4.13) ms for hip, and

$\boldsymbol{\Delta T_2^{META}}$ **(Overall) = 2.98** (95% CI 2.46 to 3.49) ms for knee and hip combined.

To discuss the meaning of ΔT_2^{META}, the T_2 difference can be classified into two different catalogs—$\Delta T_2^{population}$ and $\Delta T_2^{individual}$ (summarised in Table 2.4 on the following page). $\Delta T_2^{population}$ is the population difference in T_2 values between patients with OA and healthy controls, measured in case-control studies such as those included in this meta-analysis. $\Delta T_2^{individual}$, however, is the difference in T_2 values between an OA ROI and normal ROI, in a single-subject single-scan T_2 mapping, i.e., it is the source of the colour variations in a coloured T_2 map, and thus is T_2 difference used in an individual diagnostic context.

The ΔT_2^{META} values found above are important reference values for describing the average difference in T_2 value between OA and normal population, i.e. $\Delta T_2^{META} = \Delta T_2^{population}$. They are useful to future researchers who use T_2 measurement to separate OA and normal

Table 2.4: Summary of Difference Between $\Delta T_2^{population}$ and $\Delta T_2^{individual}$

	$\Delta T_2^{population}$	$\Delta T_2^{individual}$
Nature of ΔT_2	Between-group effect —population difference in T_2 values between patients with OA and healthy controls	Within-subject biological effect —difference in T_2 values between an OA ROI and healthy ROI
Visualisation	Separation of two distribution curves	Colour variations in a coloured T_2 map
Context	Clinical Trial	Individual Diagnose
Estimation in this Thesis	$\Delta T_2^{META} = \Delta T_2^{population}$	$\Delta T_2^{META} = \Delta T_2^{individual}$

population in clinical trials—ΔT_2^{META} is the value they should expect to replicate before the trials are started, and the value against which their new result can be compared.

Taking one step further, if it can be assumed that ΔT_2^{META} can also describe the expected difference between an OA affected ROI and an normal ROI in the cartilage, then the ΔT_2^{META} values can also serve as the reference value for individual diagnoses in clinic, i.e. $\Delta T_2^{META} = \Delta T_2^{individual}$. Clinicians who perform screening on patient scans will find these values useful—they should typically expect to see a difference of ΔT_2^{META} between the normal and OA ROIs in the cartilage of a patient if their diagnose is solely based on T_2 measurements, and they should typically expect to see a difference of ΔT_2^{META} if they made their diagnoses using other evidence and refer back to T_2 maps to confirm their conclusion.

In this thesis, the ΔT_2^{META} values on the previous page is used in Section 2.3.2.6 on page 52 to define the bottom line criteria of the trueness and precision in resolving $\Delta T_2^{individual}$. ΔT_2^{META} are also used in the calculation of the effect size (Glass's Δ) performance metric, which is then used to compare with the bottomline criteria defined by the ES^{META} (see next section), to evaluate the fitting methods' performance in resolving $\Delta T_2^{population}$.

2.2.2.2 ES^{META}—the Effect Size

We also performed a meta-analysis to estimate the effect sizes (denoted ES^{META}) corresponding the ΔT_2^{META} values found in the previous section. As can be seen in Fig. 2.5 on page 31, the meta-analysis for effect size between the OA and healthy populations showed a Hedge's g effect size of:

$\boldsymbol{ES^{META}}$ **(Knee) = 0.84** (95% CI 0.18 to 1.49) for knee,

$\boldsymbol{ES^{META}}$ **(Hip)= 1.38** (95% CI 0.30 to 2.45) for hip, and

$\boldsymbol{ES^{META}}$ **(Overall) = 1.24** (95% CI 0.42 to 2.05) for knee and hip combined.

In Table 2.5 on page 32, our ES^{META} values in this meta-analysis are compared with the benchmarks we selected. The numbers in bold font represent the original statistics reported in this and the selected studies, and the numbers in normal font are statistics that were converted using Eqs. (2.19), (2.20), (2.22) and (2.23) on pages 24–25. As can be seen in 'Palmer's Hip dGEMRIC' row (0.762 vs 0.750), and in 'Menashe's MRI Overall (Knee and Hip)' row (0.793 vs 0.804), the converted AUC is very close to the reported AUC, which confirmed the correctness of the derivation of Eqs. (2.22) and (2.23) on page 25. The visualisations are produced using [74] and is accurate to one decimal place.

In Table 2.5 on page 32, is very clear that the majority of the results are larger than the 'large' rule of thumb. One could stop here and conclude that all the MRI techniques have a statistically significant diagnostic value larger than the classical 0.8 threshold. However, this also means Cohen's rules of thumb might not be a good criteria in this context, because they fail to distinguish the 'smaller larges' from the 'larger larges' in Table 2.5.

In the rows marked in green colour, the 'ES^{META} Overall T_2' found in this study is very close to the 'Menashe's MRI Overall (Knee and Hip)' benchmark, both approximately at an effect size of 1.2. This not only indicates the correctness of both studies, but also shows that the diagnostic value of T_2 reflects the overall diagnostic value of MRI in OA.

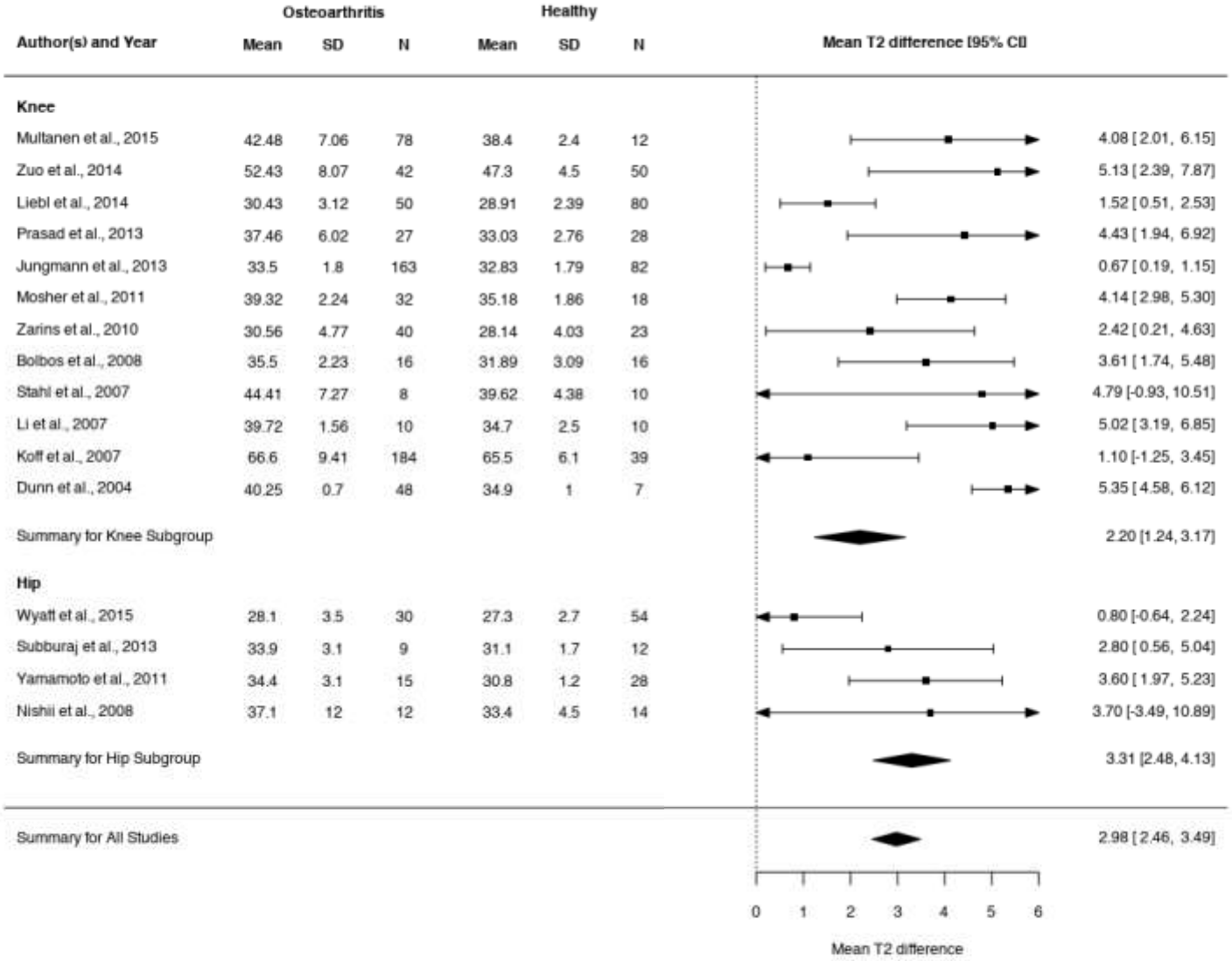


Figure 2.4: Result of Meta-Analysis on Mean Differences (MD) in T_2 Between OA and Healthy Populations

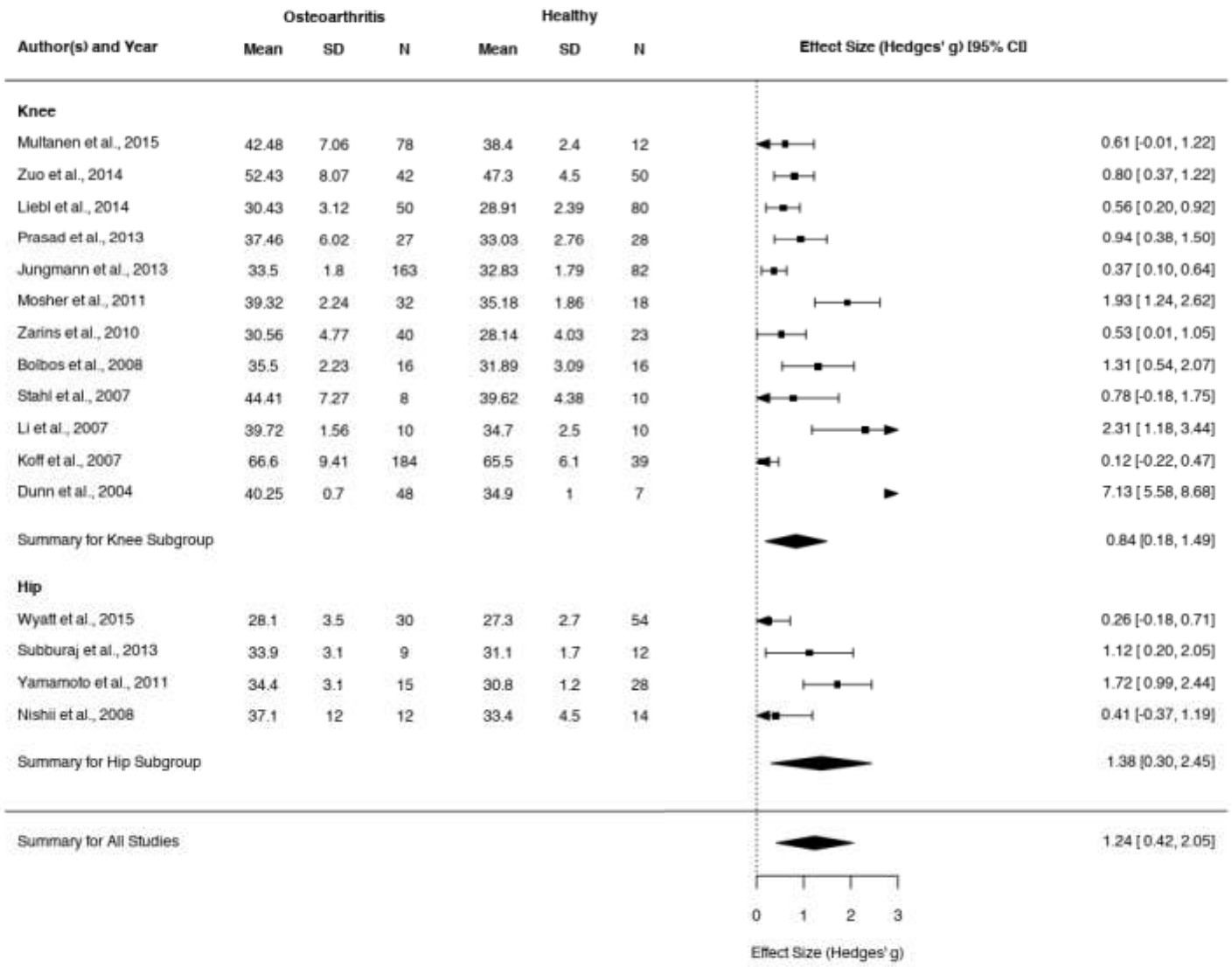
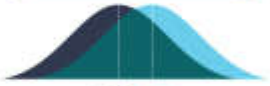


Figure 2.5: Result of Meta-Analysis on Standardised Mean Difference (SMD, Using Hedge's g Effect Size) in T_2 Between OA and Healthy Populations

Table 2.5: Comparison of the ES^{META} of T_2 Measurement with the Rules of Thumbs of Effect Size and Some Significant Results in Other Diagnostic Performance Studies of MR Techniques in Osteoarthritis

Rules of Thumb / Research Findings	Effect Size	AUC	Sensitivity	Visualisation
	0.00	0.500	50.00%	
Cohen's 'Small' [63]	**0.20**	0.556	14.76%	
Cohen's 'Medium' [63]	**0.50**	0.638	32.97%	
MacKay's Knee $T_{1\rho}$ [71]	**0.73**	0.697	44.34%	
Cohen's 'Large' [63]	**0.80**	0.714	47.43%	
ES^{META} Knee T_2	**0.84**	0.724	49.12%	
MacKay's Knee T_2 [71]	**0.87**	0.730	50.14%	
Palmer's Hip dGEMRIC [73]	0.95	**0.750**	53.66%	
	1.01	0.762	**55.60%**	
Menashe's MRI Overall (Knee and Hip) [70]	1.15	0.793	**60.70%**	
	1.21	**0.804**	62.54%	
ES^{META} Overall T_2	**1.24**	0.810	63.46%	
Li's Knee dGEMRIC [72]	**1.34**	0.828	66.41%	
ES^{META} Hip T_2	**1.38**	0.835	67.53%	
	2.00	0.921	81.14%	
	3.00	0.983	92.84%	

Interestingly, using this effect size of 1.2 as the middle point, the effect sizes of dGEMRIC (in purple colour) for knee and hip, and T_2 (in blue colour) for knee and hip approximately mirrored each other's location in the table.

The 'ES^{META} Knee T_2' found in this study is very close to 'MacKay's Knee T_2' (0.84 vs 0.87, in the upper blue coloured rows) , which is a strong indicator of the correctness of both estimations. MacKay also reported a close but smaller effect size (0.73, in the upper yellow coloured row) for $T_{1\rho}$, which is the only effect size that falls into the 'medium' (0.5) to 'large' (0.8) range of Cohen's rules of thumb. These results suggest T_2 measurement performs better in diagnosing knee OA than $T_{1\rho}$. The hip dGEMRIC has almost equal but slightly better performance than knee T_2 (0.95 vs 0.87, in the upper purple coloured row), but its effect size is still smaller than MRI overall (0.95 vs 1.21).

The 'ES^{META} Hip T_2' found in this study has the best performance among all techniques (1.38, in the lower blue coloured row), which is almost twice the effect size of knee $T_{1\rho}$ (1.38 vs 0.73). Li's Knee dGEMRIC' nearly levels the effect size of our 'ES^{META} Hip T_2' (1.34 vs 1.38, in the lower purple coloured row), indicating the superior diagnostic value of both of the techniques. It is reasonable to believe without hip T_2 and knee dGEMRIC, the effect size of 'ES^{META} Overall T_2' and MRI overall, which is now at around 1.2, will lean more towards the Hip dGEMRIC and Knee T_2 side, possibly down to a effect size lower than 1.0.

In summary, the current study shows that the diagnostic value of knee dGEMRIC and hip T_2 are highest, with an effect size of around 1.4, while diagnostic performance of knee T_2 and hip dGEMRIC are lower, at an effect size of around 0.9. The overall diagnostic value of MRI in OA is at an effect size of around 1.2.

These results suggest more work needs to be done to improve the usefulness of knee T_2 and hip dGEMRIC. The Eq. (2.10) on page 21 for calculating effect size has already indicated the direction to improve T_2 measurement—with the MD in the numerator fixed at ΔT_2^{META}, the SD in the denominator needs to be reduced to have better diagnostic

performance.

In the next part of this chapter, I therefore examine, and attempt to resolve, one of the most important factors that contributed to the relatively poor effect sizes in T_2 measurements and calculation—noise.

2.3 Performance Evaluation of Sixteen T_2 Fitting Methods with Monte-Carlo Simulation

2.3.1 Theory

2.3.1.1 The Noise-Free Signal

In the absence of noise, the signal intensity, S, of a magnitude reconstructed spin echo (SE) image decays with echo time (TE) following a monoexponential function [75]:

$$S(TE) = S_0 \exp\left(-\frac{TE}{T_2}\right) \tag{2.24}$$

Equation (2.24) can be implemented in Python as Program Code A.2 on page 253. For multi-echo gradient echo scan, the decay follows the same function except T_2 is replaced by T_2^*.

As the fitting methods discussed in this thesis are suitable for both SE and GE, 'T_2' will be used to represent either T_2 and T_2^* as appropriate.

2.3.1.2 The Rician Distribution of Signal

In the real world, Gaussian (or normal) distributed white noise will always affect the MR signal in both the $\boldsymbol{I}$ ('in phase', 'real') and $\boldsymbol{Q}$ ('quadrature', 'imaginary') channels [76]. Denote the noise-free complex signal as $\mathbf{S}$, and the complex-valued white noise as $\mathbf{N}$, then the measured complex signal given by [34]:

$$\mathbf{S_M} = \mathbf{S} + \mathbf{N} \tag{2.25}$$

is still Gaussian distributed. The final magnitude image is then calculated with a pixel-wise absolute value operation in the complex image:

$$S_M = |\mathbf{S_M}| = \sqrt{[Re(\mathbf{S}) + Re(\mathbf{N})]^2 + [Im(\mathbf{S}) + Im(\mathbf{N})]^2} \tag{2.26}$$

The probability distribution of S_M, after the absolute value operation, is no longer Gaussian but Rician, expressed in terms of the noise-free signal S in Eq. (2.24) on the previous page and the SD of the Gaussian noise σ [77–79]:

$$f_{Rician}(S_M|\theta) = \frac{S_M}{\sigma^2} \exp\left(-\frac{{S_M}^2 + S^2}{2\sigma^2}\right) I_0\left(\frac{S_M S}{\sigma^2}\right) \tag{2.27}$$

where $\theta = (S, \sigma)$, and I_0 is the modified Bessel function of the first kind of 0th order.

2.3.1.3 The Rayleigh Distribution of Signal

For an ROI in the background area, the measured complex signal consists only the Gaussian white noise $\mathbf{N}$, and Eq. (2.25) on the preceding page is reduced to:

$$\mathbf{S_M} = \mathbf{N} \tag{2.28}$$

the magnitude of which is then reduced from Eq. (2.26) to:

$$S_M = |\mathbf{S_M}| = \sqrt{[Re(\mathbf{N})]^2 + [Im(\mathbf{N})]^2} \tag{2.29}$$

The probability distribution of S_M is then reduced from Rician (Eq. (2.27)) to Rayleigh [75; 79]:

$$f_{Rayleigh}(S_M|\sigma) = \frac{S_M}{\sigma^2} \exp\left(-\frac{{S_M}^2}{2\sigma^2}\right) \tag{2.30}$$

Obviously, Eq. (2.30) is a special case (when $\theta = (0, \sigma)$) of the Eq. (2.27). The SD of Rayleigh Distribution can be expressed as [79]:

$$SD_{Rayleigh} = \sqrt{\frac{4-\pi}{2}}\sigma \approx 0.655\sigma \tag{2.31}$$

The factor of 0.655 is the Rayleigh distribution correction factor [75], which is important for estimating SNR required by some fitting methods. The usage of Eq. (2.31) and the Rayleigh factor will be demonstrated in Table B.1 on page 265.

2.3.1.4 The Noncentral Chi Distribution of Signal

Modern T_2 mapping is performed using multiple coils. With an array coil of L units, the magnitude image is then calculated using the multichannel MR signals as [80]:

$$S_M = \sqrt{\sum_{i=1}^{L} |\mathbf{S_M}^i|^2} \tag{2.32}$$

where $\mathbf{S_M}^i$ denotes the measured complex signal of the ith channel, and $||$ the magnitude operation (i.e. Eq. (2.26) on the preceding page). Assuming the SD of the gaussian noise σ are the same in all channels, the probability distribution of S_M, due to the root-sum-square operation in Eq. (2.32), follows a Noncentral Chi distribution expressed as [79; 80]:

$$f_{ncx}(S_M|\theta) = \frac{S_M}{\sigma^2}\left(\frac{S_M}{S}\right)^L \exp\left(-\frac{{S_M}^2 + S^2}{2\sigma^2}\right) I_{L-1}\left(\frac{S_M S}{\sigma^2}\right) \tag{2.33}$$

where $\theta = (S, \sigma)$, and I_{L-1} is the modified Bessel function of of the first kind of $(L-1)$th order. Obviously, Eq. (2.27) on the preceding page is a special case of Eq. (2.33) in which the number of array coil $L = 1$ [80].

2.3.2 Methods

2.3.2.1 Summary of Fitting Methods

To perform a thorough performance analysis for state-of-the-art T_2 fitting methods, this study included not only T_2 fitting methods reported for musculoskeletal MRI, but also included some fitting methods reported in the liver and myocardial MRI communities, where T_2, or transverse relaxation rate R_2 $(= \frac{1}{T_2})$, is used as an important factor for

Table 2.6: Summary of T_2 Fitting Methods Evaluated in This Study

Method Type	Noise Models				Knowledge of SNR
	None	Gaussian		Constant	
Methods of Least Squares	**LLS**	**NLS**	**TRUNC**	**OFFSET**	Not required
	Rician		Non-central Chi		
Methods of Moments	**NCEXP**	**SQEXP**	**M^{1}NCM**	**M^{2}NCM**	
	NCEXP$_\sigma$	**SQEXP$_\sigma$**	**M^{1}NCM$_\sigma$**	**M^{2}NCM$_\sigma$**	Required
Methods of Maximum Likelihood and Bayesian Estimation	**MLE$_{Rice}$**	**BAYES$_{Rice}$**	**MLE$_{nc\chi}$**	**BAYES$_{nc\chi}$**	

liver/myocardial iron quantification, where a lower T_2, or higher R_2 implies higher iron-load [81–83].

In the literature, sometimes T_2 fitting methods of the same principle were named differently by different authors, which makes comparisons confusing. To clearly refer to a specific T_2 fitting method, this thesis uses the name and/or abbreviation used by the author who first (determined by publication year and citation relationships) used a method for medical imaging research, rather than those who derived a method in a purely mathematical context. The fitting methods evaluated in this study are summarised in Table 2.6 in terms of their method type, noise model and requirement for SNR measurement. A detailed mathematical and technical description of the methods is available in the next section.

2.3.2.2 Fitting Methods in Detail

Linear Least Squares (LLS)

The classical linear least squares estimation method assumes the magnitude image is noise-free, i.e., the signal S follows Eq. (2.24) on page 35. The natural logarithm of both sides of Eq. (2.24) gives a linear function of the form $y = ax + b$ (Eq. (2.34)), where slope $a = -\frac{1}{T_2}$ and intercept $b = \ln S_0$. a and b can then be optimised by ordinary linear regression by fitting a series of coordinates (x, y), where $x = TE$ and $y = \ln S_M$ [84].

$$\ln S_M = -\frac{TE}{T_2} + \ln S_0 \tag{2.34}$$

Despite its oversimplified assumption about the noise in MRI, the LLS method is widely used for its simplicity. Another advantage of LLS is that it does not require an initial guess, thus the T_2 and S_0 values calculated using LLS can in turn be used in other methods as a good initial guess. *In this study, LLS is used to calculate such initial guesses for other methods*, and it is implemented in Python using NumPy's least squares polynomial fit function `numpy.polyfit(xdata, ydata, deg=1)` [85], where `deg` represents the degree of the fitting polynomial, which should be set to 1 for linear least squares estimation. The Python code for LLS implementation is presented in Program Code A.3 on page 254.

Non-linear Least Squares (NLS)

It is commonly believed that at a high SNR (for example, when SNR > 5), the distribution of the magnitude signal can be approximated by Gaussian. S_0 and T_2 can then be optimised using a non-linear least squares fitting algorithm [81]:

$$\arg\min_{S_0,T_2} \sum_{i=1}^{N} \left[S_{M_i} - S_0 \exp\left(-\frac{TE}{T_2}\right) \right]^2 \tag{2.35}$$

The Levenberg-Marquardt (LM) algorithm is the most commonly used model to solve non-linear optimisation problems (despite its lack of ability to deal with constrained problems). This study followed this convention and implemented NLS in Python using the LM algorithm provided by SciPy's non-linear least squares fit function [86] `scipy.optimize.cu-`

`rve_fit(T2_func, xdata, ydata, p0, method='lm')`, where `T2_func` is Eq. (2.24) on page 35. The initial guess pair (T_2, S_0) calculated by LLS is passed to the initial guess parameter `p0` to replace the default initial guess $(1, 1)$. The Python code for NLS implementation is presented in Program Code A.4 on page 254.

Truncation Model (TRUNC)

Removing one or more echoes from fitting has been a popular approach in the literature to deal with the noise in T_2 measurement. Which echo/echoes should be removed from the fitting, and which should be kept, remains debatable. Echo-removing methods are also special cases of the family of weighted fitting methods, in which echoes are weighted differently to account for SNR variations, and are removed when weighted zero [81].

Among all the variations of echo-removing methods, Truncation Model (TRUNC) is evaluated in this study because the removal of echoes in TRUNC, unlike the subjective approach in other echo-removing methods, is decided by an algorithm based on Coefficient of Determination, or R-squared value, denoted R^2. R^2 is calculated with [82; 83]:

$$\begin{aligned} R^2 &= \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\ &= 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \end{aligned} \tag{2.36}$$

where y_i denotes the data points, $\hat{y}_i$ the estimated regression line, and $\bar{y}$ the the sample mean of the data points. R^2 is always a number between 0 and 1. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. $R^2 = 1$ indicates that the regression line perfectly fits the data.

TRUNC can be expressed with Eq. (2.37):

$$\underset{S_0, T_2}{\arg\min} \sum_{i=1}^{x} \left[S_{M_i} - S_0 \exp\left(-\frac{TE}{T_2}\right) \right]^2 \tag{2.37}$$

where x is the number of echoes being used for fitting, determined by R^2 in Eq. (2.36) on

the preceding page.

The Python implementation for TRUNC (Program Code A.5 on page 255) is the same as NLS, other than that x is determined by an algorithm: after the initial NLS fit to calculate the R^2, accept the T2 value if $R^2 \geqslant 0.995$, or keep eliminating the last echo and re-calculating R^2, until $R^2 \geqslant 0.995$ or only $x = 2$ echoes are left [82; 83].

Offset Model (OFFSET)

The Offset Model (OFFSET) assumes the signal in magnitude image is produced by adding a constant offset C to the noise-free signal, making a three-parameter model in Eq. (2.38), where the offset C is an unknown parameter to be optimised by the LM algorithm [83; 87]. OFFSET can be expressed as:

$$S_M = S_0 \exp\left(-\frac{TE}{T_2}\right) + C \tag{2.38}$$

Similar to NLS, `scipy.optimize.curve_fit(T2_func_OFFSET, xdata, ydata, p0, method='lm')` [86] is used to implement OFFSET, where `T2_func_OFFSET` is Eq. (2.38). The three parameter initial guess `p0` is set to (T_2, S_0) calculated by LLS, and an additional initial guess of offset $C = 1.0$. The implementation of OFFSET is presented in Program Code A.6 on page 256.

Noise-Corrected Exponential Model (NCEXP) and Noise-Corrected Exponential Model with Known σ (NCEXP$_\sigma$)

The NCEXP model uses the expected value (denoted E$[S_M]$, also known as the first moment) of the Rician random variable S_M (Eq. (2.27) on page 36) to describe the decay of S_M with TE. E$[S_M]$ for Rician is frequently reported as Eq. (2.39) in the literature [34; 77]:

$$\mathrm{E}_{Rician}[S_M] = \sigma\sqrt{\frac{\pi}{2}} \exp(-\alpha)\left[(1+2\alpha) I_0(\alpha) + 2\alpha I_1(\alpha)\right] \tag{2.39}$$

where $\alpha = (\frac{S^2}{2\sigma^2})$, S is the noise-free signal in Eq. (2.24) on page 35, and I_0 the modified Bessel function of the first kind of 0th order and I_1 the 1st order.

Substitution of α into Eq. (2.39) give rise to Eq. (2.40), which relates the $\mathrm{E}[S_M]$ directly with the parameters to be optimised.

$$\mathrm{E}_{Rician}[S_M] = \sigma\sqrt{\frac{\pi}{2}}\exp\left(-\frac{S^2}{4\sigma^2}\right)\left[\left(1+\frac{S^2}{2\sigma^2}\right)I_0\left(\frac{S^2}{4\sigma^2}\right)+\frac{S^2}{2\sigma^2}I_1\left(\frac{S^2}{4\sigma^2}\right)\right] \tag{2.40}$$

$$= \sigma\sqrt{\frac{\pi}{2}}{}_1F_1\left(-\frac{1}{2},1,-\frac{S^2}{2\sigma^2}\right) \tag{2.41}$$

Eq. (2.41) express the same function using the confluent hypergeometric function of the first kind (${}_1F_1$), which *to the best of my knowledge has not been reported in the literature before.* Eq. (2.41) not only is much easier to implement than Eq. (2.40), but also is important for comparing with the Eq. (2.43) on the facing page of the First Moment Noise Corrected Model ($\mathrm{M^1NCM}$) that uses the Noncentral Chi Distribution.

The σ can be treated as a parameter to be optimised by fitting, giving rise to a three-parameter fitting model (NCEXP). It can also be a known value obtained by performing SNR measurement on the image before fitting (or set as a known iterate parameter in a simulation as if it is measured, see Table B.1 on page 265), which gives rise to a two-parameter fitting model (NCEXP with known σ, or NCEXP_σ).

To be comparable with related studies, Eq. (2.40) is used to implement these two models in Python as Program Code A.7 on page 256, although using Eq. (2.41) would produce a more concise code. The three parameter initial guess (`guess_3`) is set to (T_2, S_0) calculated by LLS, plus an additional initial guess of $\sigma = 35.0$. The reason for setting 35.0 as the initial guess for σ will be explained in Section 2.3.2.3 on page 47.

Squared Exponential Model (SQEXP) and Squared Exponential Model with Known σ (SQEXP_σ)

The SQEXP model uses the expected value of the squared random variable (denoted $\mathrm{E}[{S_M}^2]$, also know as the second moment) to describe the decay of the Rician signal S_M

(Eq. (2.27) on page 36) with TE. For Rician Distribution, $\mathrm{E}[{S_M}^2]$ can be expressed as [34; 88]:

$$\mathrm{E}_{Rician}[{S_M}^2] = S^2 + 2\sigma^2 \tag{2.42}$$

where S is the noise-free signal in Eq. (2.24) on page 35. Similar to NCEXP, the σ can be treated as a parameter to be optimised by fitting (SQEXP), or a known parameter (SQEXP$_\sigma$). SQEXP and SQEXP$_\sigma$ are implemented in Python as Program Code A.8 on page 258. The initial guesses are set in the same way as NCEXP.

First Moment Noise Corrected Model (M^{1}NCM) and First Moment Noise Corrected Model with Known σ (M^{1}NCM$_\sigma$)

M^{1}NCM is the Noncentral Chi version of NCEXP. The number of array coils used in imaging is taken into consideration, and the first moment ($\mathrm{E}[S_M]$) of the Noncentral Chi signal S_M (Eq. (2.32) on page 37) is used to describe the decay of S_M with TE. For Noncentral Chi Distribution, $\mathrm{E}[S_M]$ can be expressed as [80; 89]:

$$\mathrm{E}_{ncx}[S_M] = \sigma\sqrt{\frac{\pi}{2}}\frac{(2L-1)!!}{2^{L-1}(L-1)!}{}_1F_1\left(-\frac{1}{2}, L, -\frac{S^2}{2\sigma^2}\right) \tag{2.43}$$

where L is the number of units of array coils, S the noise-free signal in Eq. (2.24) on page 35, ${}_1F_1$ the confluent hypergeometric function of the first kind, and !! the double factorial: $n!! = n(n-2)(n-4)\cdots$.

When $L = 1$, Eq. (2.43) is reduced to Eq. (2.41) on the facing page, so M^{1}NCM can be considered as a generalisation of NCEXP. Again, σ can be treated as a parameter to be optimised by fitting (M^{1}NCM), or a known parameter (M^{1}NCM$_\sigma$). M^{1}NCM and M^{1}NCM$_\sigma$ are implemented in Python as Program Code A.9 on page 259. The initial guesses are set in the same way as NCEXP.

Second Moment Noise-Corrected Model (M^2NCM) and Second Moment Noise-Corrected Model with Known σ (M^2NCM_σ)

M^2NCM is the Noncentral Chi version of SQEXP. In M^2NCM, the second moment ($E[{S_M}^2]$) of the Noncentral Chi Signal (Eq. (2.32) on page 37) is used to describe the decay of S_M with TE. For Noncentral Chi distribution, $E[{S_M}^2]$ can be expressed as [80; 90]:

$$E_{nc\chi}[{S_M}^2] = S^2 + 2L\sigma^2 \tag{2.44}$$

where L is the number of units of array coils, and S is the noise-free signal Eq. (2.24) on page 35. When $L = 1$, Eq. (2.44) is reduced to Eq. (2.42) on the preceding page, so M^2NCM can be considered as a generalisation of SQEXP. Again, σ can be treated as a parameter to be optimised by fitting (M^2NCM), or a known parameter (M^2NCM_σ). M^2NCM and M^2NCM_σ are implemented in Python as Program Code A.10 on page 260. The initial guesses are set in the same way as NCEXP.

When σ is unknown and to be optimised by fitting, the equations for M^2NCM (Eq. (2.44)), OFFSET (Eq. (2.38) on page 41) and SQEXP Eq. (2.42) on the preceding page are of a very similar form. Whether they are actually the same optimisation problem and will produce the same fitting result will be examined in the simulation and be discussed.

Maximum Likelihood Estimation of Rician (MLE_{Rice}) and Noncentral Chi ($MLE_{nc\chi}$) Signal

Estimating parameters from independent measurements of S_M at each TE is classically formulated in a Maximum Likelihood framework. For a given probability distribution function $f(S_M|\theta)$, the likelihood function of the measurements is defined as the product of the S_M at each TE [91; 92]:

$$\mathcal{L}(S_M|\theta) = \prod_{i=1}^{N} f(S_{M_i}|\theta) \tag{2.45}$$

where N is the total number of echoes, and S_{M_i} the measured signal at the ith TE. Substituting $f(S_M|\theta)$ with Eq. (2.27) on page 36 and Eq. (2.33) on page 37 gives rise to

the likelihood function of the Rician (Eq. (2.46)) and Noncentral Chi (Eq. (2.47)) signal respectively [93]:

$$\mathcal{L}_{Rician}(S_M|\theta) = \prod_{i=1}^{N} \frac{S_{M_i}}{\sigma^2} \exp\left(-\sum_{i=1}^{N} \frac{{S_{M_i}}^2 + S^2}{2\sigma^2}\right) \prod_{i=1}^{N} I_0\left(\frac{S_{M_i} S}{\sigma^2}\right) \tag{2.46}$$

$$\mathcal{L}_{nc\chi}(S_M|\theta) = \prod_{i=1}^{N} \left[\frac{S}{\sigma^2}\left(\frac{S_{M_i}}{S}\right)^L\right] \exp\left(-\sum_{i=1}^{N} \frac{{S_{M_i}}^2 + S^2}{2\sigma^2}\right) \prod_{i=1}^{N} I_{L-1}\left(\frac{S_{M_i} S}{\sigma^2}\right) \tag{2.47}$$

The likelihood functions return the probability of the measurement in terms of the parameters $\theta = (S, \sigma)$. The Maximum Likelihood estimator is then found from the global maximum of the likelihood functions:

$$\hat{\theta}_{ML} = \arg\max_{S_0, T_2} \mathcal{L}(S_M|\theta) \tag{2.48}$$

The MLE is implemented in Python by maximising the log-likelihood functions, i.e. natural logarithm of the likelihood functions Eqs. (2.46) and (2.47), which is commonly used as an equivalent to the original likelihood functions. The log-likelihood function converts multiplied terms in the original function into the sum of their logarithms, which makes the maximising (calculating derivative) computationally much easier than if the terms are multiplied together.

Since SciPy only provided a minimisation function `scipy.optimize.minimize(func, guess, method)` [94], the minimums of the negative log-likelihood functions (`-1.0*T2_log_like_Rice` and `-1.0*T2_log_like_Rice`), equivalent to the maximums, are calculated. The quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno (BFGS) (`method='BFGS'`) is used for minimisation for its proven good performance even for non-smooth optimisations [94]. The σ is set as a known parameter that needs to be pre-measured from the image for the two MLE models in this study. The initial guess (`guess`) is set to (T_2, S_0) calculated by LLS. The implementations of MLE_{Rice} and $\text{MLE}_{nc\chi}$ are presented in Program Code A.11 on page 261.

Baysian Estimation of Rician (BAYES$_{Rice}$) and Noncentral Chi (BAYES$_{nc\chi}$) Signal

The Bayesian approach is based on incorporating known information (prior knowledge) into a mathematical framework. The prior knowledge available in MR signal are [95; 96]:

(1) the S and the σ of the noise are non-negative;

(2) S and σ are independent.

Suppose the unknown parameters θ follow the prior distribution $\Pi(\theta)$ with prior density $\pi(S, \sigma)$. The prior information is then translated into mathematical terms results in the following properties [95; 96]:

(1) $\pi(S, \sigma) = 0$ for $A < 0$ or $\sigma < 0$;

(2) $\pi(S, \sigma) = g(S)h(\sigma)$ with g and h positive, real valued functions.

Additionally, using Jaynes' invariance principle [97], it can be shown that:

(3) $\frac{1}{k^2} = \pi(kS, k\sigma)$ with $k > 0$;

The prior knowledge translated in properties (2) and (3) above results in the following family of prior densities [95; 96]:

$$\pi(S, \sigma) = \frac{1}{S^{1-\varepsilon}\sigma^{1+\varepsilon}} \text{ for any } \varepsilon \in \mathbb{R} \tag{2.49}$$

where ε is set to the value that results in minimal uncertainty on the parameter estimates. Following the original study of this method, $\varepsilon = 2$ was used in this study [95]. By applying Bayes' theorem, the posterior density of θ is obtained by multiplying the likelihood function $\mathcal{L}(S_M|\theta)$ (Eq. (2.45) on page 44) with the prior density function $\pi(S, \sigma)$ [98]:

$$f(\theta|S_M) = c\mathcal{L}(S_M|\theta)\pi(\theta) \tag{2.50}$$

In Eq. (2.50), substituting $\mathcal{L}(S_M|\theta)$ with the likelihood functions of Rician (Eq. (2.46) on page 45) and Noncentral Chi (Eq. (2.47) on page 45) signal, and $\pi(S,\sigma)$ with the prior density in Eq. (2.49) give rise to the posterior densities of Rician (Eq. (2.51)) and Noncentral Chi (Eq. (2.52)) signal respectively:

$$f_{Rician}(\theta|S_M) = \frac{c}{S^{1-\varepsilon}\sigma^{1+\varepsilon}} \prod_{i=1}^{N} \frac{S_{M_i}}{\sigma^2} \exp\left(-\sum_{i=1}^{N} \frac{{S_{M_i}}^2 + S^2}{2\sigma^2}\right) \prod_{i=1}^{N} I_0\left(\frac{S_{M_i}S}{\sigma^2}\right) \tag{2.51}$$

$$f_{nc\chi}(\theta|S_M) = \frac{c}{S^{1-\varepsilon}\sigma^{1+\varepsilon}} \prod_{i=1}^{N} \left[\frac{S}{\sigma^2}\left(\frac{S_{M_i}}{S}\right)^L\right] \exp\left(-\sum_{i=1}^{N} \frac{{S_{M_i}}^2 + S^2}{2\sigma^2}\right) \prod_{i=1}^{N} I_{L-1}\left(\frac{S_{M_i}S}{\sigma^2}\right) \tag{2.52}$$

In a Bayesian framework, the maximum a posteriori (MAP) estimator is then found from the global maximum of the posterior densities [99]:

$$\hat{\theta}_{MAP} = \arg\max_{S_0,T_2} f(\theta|S_M) \tag{2.53}$$

Similar to the MLE methods, the log-likelihood functions of the prior density and the posterior densities are used to implement the BAYES methods in Python. The maximisation functions are implemented in a similar way as the MLE methods. The σ is set as a known parameter that needs to be pre-measured for the two BAYES models in this study. The initial guess (`guess`) is set to (T_2, S_0) calculated by LLS. The implementations of BAYES_{Rice} and $\text{BAYES}_{nc\chi}$ are presented in Program Code A.12 Program Code A.12.

2.3.2.3 Monte Carlo Simulation

Simulation Design

The Monte Carlo simulation simulated the imaging protocol used in the Osteoarthritis Initiative (OAI) study, where Sagittal 3 Tesla multi-echo spin-echo (MESE) T_2 images were acquired. TEs were set to be the same as OAI, which are 10, 20, 30, 40, 50, 60, and 70 ms [100]. L is set to 8 to simulate the quadrature transmit, eight-channel PA receive

coil (InVivo Corp) used in the OAI study [101].

In the simulation, True S_0 needs to be kept constant, and SNR_0 varied to obtain the $\sigma = \frac{S_0}{\text{SNR}_0}$ needed by the noise generator. The value of S_0, which was intensionally or unintentionally concealed in previous simulation studies [34; 81], is therefore a critical parameter which will influence the level of noise generated. In Chapter 4 on page 133, a measurement of SNR were performed on the first echo of all the 4,654 sets of eligible T_2 images in the OAI study, and obtained a mean intensity in the ROI at the first echo of $S_1 = 511.78$, and a mean first echo σ of 35.27. Therefore, the S_0 in the simulation is set equal to the S_1 at 512.0, and the initial guess for σ is set to be 35.0 (for methods in which σ is to be optimised rather than pre-measured).

In this study, the True T_2 values are set to iterate from 20 ms to 80 ms inclusive with 1 ms as the step, which is the normal range of cartilage T_2 values. SNR_0 levels are set to iterate from 5 to 105 inclusive with 1 as the step to obtain the σ for the noise generator. This forms $61 \times 101 = 6,061$ pairs of (True T_2, SNR_0), representing 6,061 biochemical and imaging conditions that simulates the conditions in real scans. For each (True T_2, SNR_0) pair, 1,000 sets of noise are generated by the noise generator and are 'added' to the noise-free signal (S, determined by the known S_0 and True T_2), resulting in 1,000 sets of synthetic signals (noisy S) to be fitted by the 16 methods. In total there were $6,161 \times 1,000 \times 16 = 98,576,000$ fittings performed in this simulation study.

The performance metrics are then calculated and compared with the bottom line criterion defined by the meta-analysis in Part I (Section 2.2) of this chapter. The results are finally visualised as three heatmaps for the three performance metrics (discussed in Section 2.3.2.6 on page 52) used in this study.

The main body of the Monte Carlo simulation, showing the parameters used and the structure of the iterations, is presented in Program Code A.13 on page 263. The simulation was run on the workstation 'horla' (see Table 4.1 on page 134) and took 1 day 7 hours 41 minutes and 6 seconds to complete, which approximates to 17.94 seconds per 1,000

fittings at each (T_2, SNR_0) pair.

2.3.2.4 Synthetic T_2 signal

In the simulation, noise and synthetic signals can be generated by using different distributions and different 'noise-adding' approaches. Depending on the assumption made about the noise in MRI, the noise can be Gaussian, Rayleigh, or Rician distributed, and can either be created using the analytical form, or sampled directly from the PDFs of the distributions.

(1) If the noise is assumed to be Gaussian distributed, `np.random.normal(loc, scale, size)` function can be used to sample the noise with `scale` $= \sigma = \frac{S_0}{\mathrm{SNR}_0}$ and `size=(test_num, TE_num)` which is (1,000, 7) for 1,000 sets of noises in the 7 echoes. The synthetic noise is then generated by taking the absolute value of the sum of the noise-free signal and the noise. The Gaussian signal generated using itsanalytical form and the PDF are the same because both the noise-free signal and the noise are real numbers;

(2) If the noise is assumed to be Rayleigh distributed, there are two options to create the synthetic signal. One can either use `np.random.normal(loc, scale, size)` twice to generate the real and imaginary components (both are Gaussian) of the complex noise independently, and create the Rayleigh distributed magnitude noise by using the analytical form by taking the absolute value of the complex noise, or use `scipy.stats.rayleigh.rvs(loc, scale, size)` to sample the Rayleigh distributed magnitude noise directly. The synthetic signal can then be created by taking the absoltue value of the sum of the noise-free signal and the magnitude noise generated in either approach;

(3) If the noise is assumed to be Rician distributed, similar to the Rayleigh assumption, the noise can be generated by using the `np.random.normal(loc, scale, size)` twice independently, but the sum of real part (or imaginary part) and the noise-free

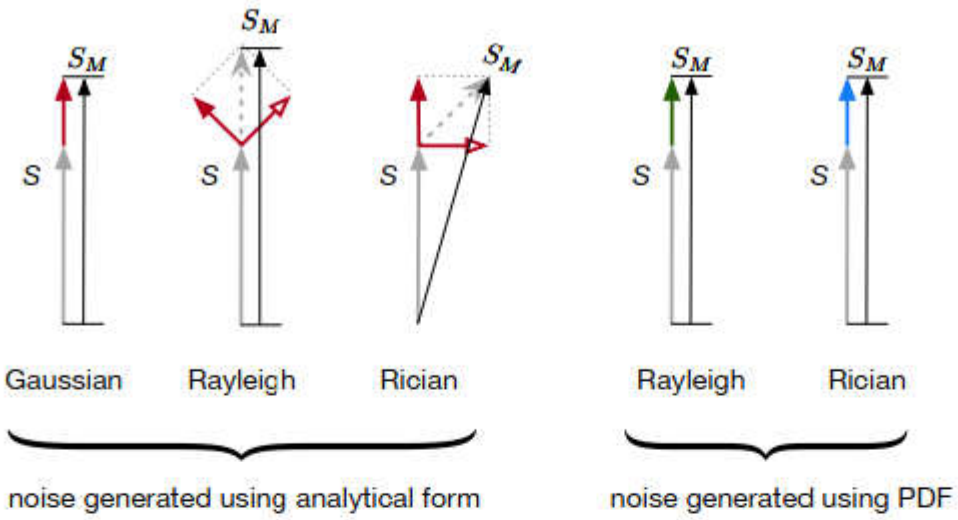


Figure 2.6: Comparison of Methods for Generating Synthetic (Noisy) Signal Affected by Difference Noise Distribution

signal is calculated before performing the absolute operation to get the magnitude signal. The other approach by using the PDF to sample the Rician noise directly is similar to that of the Rayleigh assumption, and is done by using `scipy.stats.rice.rvs(loc, scale, size)`.

The above approaches are illustrated in Fig. 2.6, where S is the noise-free signal and S_M the synthetic magnitude signal. The red arrows represent the Gaussian distributed real and imaginary components in the complex noise if the analytical form is used to generate the noise, while the green and blue arrows represent the real-numbered noise sampled directly from the respective PDFs.

The implementation of the noise generator and the synthetic signal generator are presented in Program Code A.14 on page 264. In this study, the Rician noise generated using the analytical form is used to generator the synthetic signal because it simulates the noise in real scans best.

2.3.2.5 Interpretation of the Synthetic T_2 signal

In order to evaluate the performance of the fitting methods for both individual diagnoses and population separation, the synthetic (noisy) signal created is interpreted from two

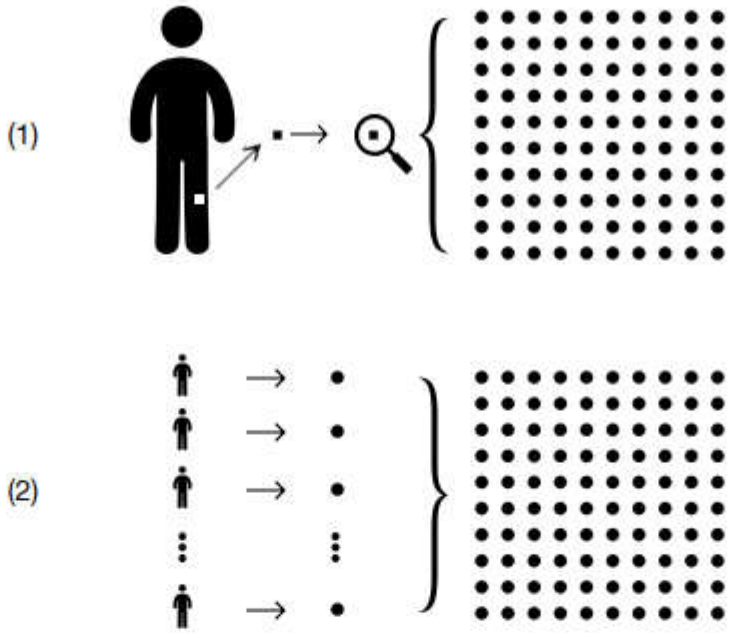


Figure 2.7: Illustration of the Two Perspectives to Interpret the Synthetic T_2 Signal

perspectives (Fig. 2.7):

(1) Assume the 1,000 synthetic signals are voxels in an ROI from the image of a single person, which all have the same True T_2 but are affected by different random noise sampled at certain level of SNR_0 in this scanning section. This enables the calculation of the trueness and precision of the fitting methods at this (True T_2, SNR_0) condition, described in the next section, which evaluate the fitting methods performance in resolving $\Delta T_2^{individual}$ in a single-patient single-scan section for diagnosis in an individual level.

(2) Alternatively it can be assumed that the 1,000 synthetic signals are each contributed by a distinct subject in a group of 1,000 subjects, which all happens to have the same True T_2 value but are affected by different random noise sampled *independently* in their respective scans, all at this SNR_0. This enables us to use the same set of simulation results to further evaluate the effect sizes of the fitting methods at this (True T_2, SNR_0) condition, described in the next section, which deals with the fitting methods performance in resolving the $\Delta T_2^{population}$, for example in a long-term multi-centre multi-scanner clinical trial.

Table 2.7: Summary of Performance Metrics

Performance Metrics	**Trueness**	**Precision**	**Effect Size**
Definition	Closeness of agreement between the mean value of the fitted T_2 and the True T_2 value	Closeness of agreement between independent fitting results	Magnitude of the difference between groups in terms of SD
Numerical Expression	$Mean_{\text{fitted } T_2}$ − True T_2	1.96SD	$\frac{\Delta T_2^{META}}{\text{SD}}$
Nature of the Metric	Systematic fitting error; bias of fitting result	Random fitting error; Spread of fitting result	Re-expressing random fitting error in terms of population difference
Bottom Line Criterion	$\lvert\text{Trueness}\rvert \leqslant \Delta T_2^{META}$	Precision $\leqslant \Delta T_2^{META}$	$S^{SIM} \geqslant ES^{META}$

The definition of the performance metrics of trueness and precision based on Interpretation (1), and effect size based on (2), are described in detail in the next section.

2.3.2.6 Performance Metrics and the Bottom Line Criterion

Three performance metrics—trueness, precision and effect size, are defined to evaluate the simulation results. The main points of these metrics are summarised in Table 2.7. The details are discussed in the following sections.

Trueness

For each (True T_2, SNR_0) pair, trueness of a fitting method at this condition refers to the closeness of agreement between the mean value of the fitted T_2 values and the True T_2 value:

$$\text{Trueness} = Mean_{\text{fitted } T_2} - \text{True } T_2 \tag{2.54}$$

As can be seen in an illustration of trueness in Fig. 2.8 on the next page, trueness refers to the systematic error, the way in which the fitting results may be consistently different

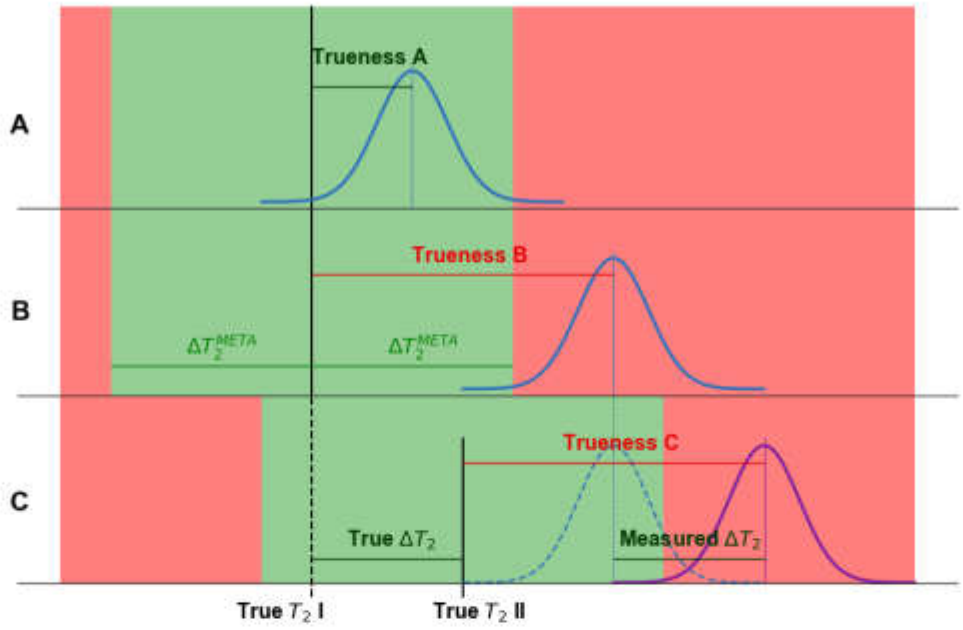


Figure 2.8: Illustrations of Definition, the Bottom Line criterion, and the Importance of Consistency of Trueness

from the true value, or biased.

Trueness A in Fig. 2.8 shows a trueness value smaller than the ΔT_2^{META} ($Mean_{\text{fitted } T_2}$ lies in the green region), which is desirable because the clinical meaningful T_2 difference (ΔT_2^{META}) can still be observed despite the bias in fitting results.

Trueness B in Fig. 2.8, in contrast, shows a trueness value larger than the ΔT_2^{META} ($Mean_{\text{fitted } T_2}$ lies in the red region), which means the clinical meaningful T_2 difference (ΔT_2^{META}) can on longer be observed, and is obscured behind the bias of fitting results caused by systematic fitting error.

To put it simply, trueness, the systematic fitting error, or the bias, must be minimised, and at least be smaller than the clinical meaningful T_2 difference (ΔT_2^{META}), so that the latter could be seen. Thus the bottom line criterion for trueness, defined by the meta-analysis result ΔT_2^{META} is:

$$|\text{Trueness}| \leqslant \Delta T_2^{META} \tag{2.55}$$

It has been argued trueness is irrelevant in clinical MR measurement, because the systematic error is always present, and does not mask the difference in True T_2 values. Consider the Trueness B of measuring True T_2 I, and the Trueness C of measuring True T_2 II in Fig. 2.8 on the preceding page: although they both violated the bottom line criterion and had the $Mean_{\text{fitted } T_2}$ lie in the red region, the difference in True T_2 values can still be correctly measured (Measured ΔT_2 = True ΔT_2) because the values of trueness are equal (Trueness B = Trueness C). The bias in fitting is removed by the subtraction operation when measuring the T_2 difference.

However, the above argument is only true if the systematic fitting error is consistent regardless of the True T_2 being measured, and the SNR_0 under which the measurement is performed. As can be seen in Fig. 2.11 on page 64, this clearly is not the case—trueness values changes at different (True T_2, SNR_0) conditions and how they change varies considerably for different methods. Since absolute consistency of trueness does not exist in any of the methods, a method that has a trueness behaviour more consistent relative to other methods become more valuable because more bias can be removed when the subtraction operation is performed.

The bottomline criterion Eq. (2.55) on the preceding page is still most important because it ensures closeness of the fitted T_2 value to the True T_2—important when an exact measurement of True T_2 is needed. It also ensures the systematic error level is under control as a whole, that is, even inconsistency exists, they all have an upper limit of ΔT_2^{META} and the subtraction of any two trueness values (Measured ΔT_2) will still be very close to the difference in True T_2 (True ΔT_2), although they are not exactly equal.

The bottom line criterion and consistency of the trueness values of the fitting methods will be discussed respectively.

Precision

For each (True T_2, SNR_0) pair, precision of a fitting method is the closeness of agreement between independent fitting results of the same True T_2 at the stipulated SNR_0 condi-

tion. Assuming the fitting results are sampled from a normal distribution with standard deviation SD, precision can be defined as the confidence limit (half of the width of 95% confidence interval, or margin of error) on a single fitted value, which approximately equals to 1.96SD.

$$\text{Precision} = 1.96\text{SD} \tag{2.56}$$

As can be seen in an illustration of precision in Fig. 2.9 on the next page, precision refers to the random error in the fitting results if the fitting of the same True T_2 is repeated many times, shown as the spread of the result, or as the length of the confidence limit (1.96SD); it is also referred to as the reproducibility, or test-retest performance.

Precision A in Fig. 2.9 shows a precision value smaller than the ΔT_2^{META} (1.96SD lies in the green region), which is desirable because the clinical meaningful T_2 difference (ΔT_2^{META}) can still be observed despite the random fitting error, or spread of fitting results.

Trueness B in Fig. 2.9, in contrast, shows a precision value larger than the ΔT_2^{META} (1.96SD lies in the red region), which means the clinical meaningful T_2 difference (ΔT_2^{META}) can on longer be observed, and is obscured behind the spread of fitting results caused by random fitting error.

To put it simply, precision, the random fitting error, or the spread of fitting results, must be minimised, and at least be smaller than the clinical meaningful T_2 difference (ΔT_2^{META}), so that the latter could be seen. Thus the bottom line criterion for precision, defined by the meta-analysis result ΔT_2^{META} is:

$$\text{Precision} \leqslant \Delta T_2^{META} \tag{2.57}$$

It is also necessary to note some terminology issues here: 'trueness' is often referred to as 'accuracy'. However, 'accuracy' is not the correct term to replace 'trueness'. 'Trueness' describe the distance between the mean value of *a large series* of results and the true value,

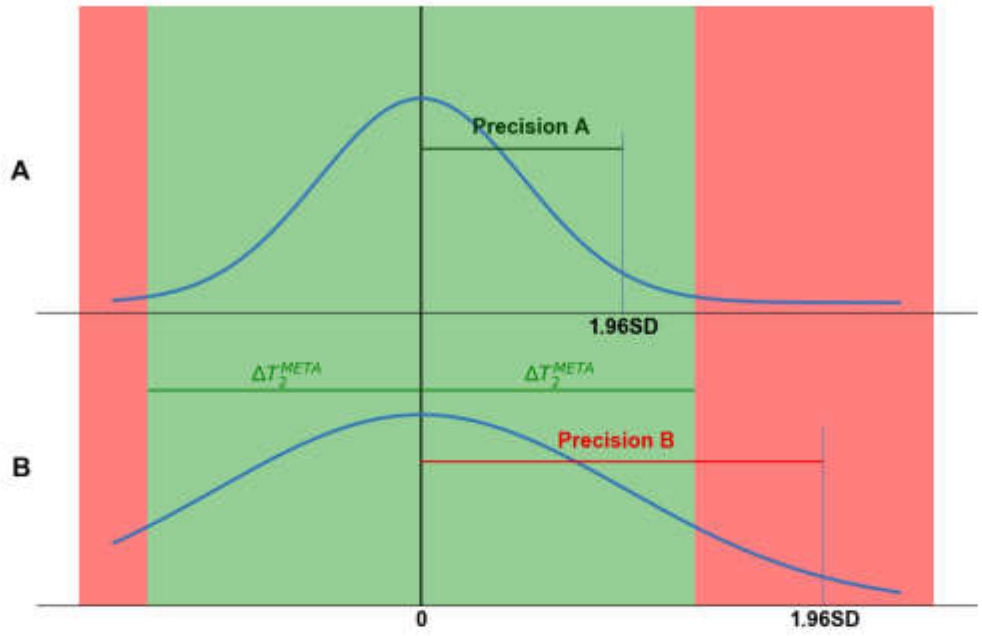


Figure 2.9: Illustration of Definition and the Bottom Line criterion for Precision

while 'accuracy' refers to the closeness of agreement of the result of *one* measurement to the true value, which must be expressed in terms of both bias and SD. That is, *'accuracy' refers to a combination of 'trueness' and 'precision'*. Unfortunately, many literature used the wrong term 'accuracy' when they are really discussing 'trueness'. Nevertheless, in whichever case lower values of trueness and precision are desirable in the context of measurement science, because the lower values represents smaller (systematic or random) errors. It is also very important to bear in mind a 'low value of trueness/precision' or 'low trueness/precision value' should not be confused with 'low trueness/precision' (or 'low accuracy') in colloquial language—they mean exactly the opposite.

Effect Size

Using the Interpretation (2) of the synthetic T_2 signal on page 51, the effect size can also be calculated to measure the fitting methods' performance in resolving the $\Delta T_2^{population}$.

However, the assumption of equal SD in healthy and OA populations for calculating Hedge's g effect size is violated in the simulation, because the SD (and hence 1.96SD—the precision) changes at different (True T_2, SNR_0) conditions, and this change cannot be

ignored otherwise the simulation would be pointless.

As a workaround, the Glass's Δ effect size (see Eq. (2.12) on page 21), which uses the SD of the control rather than the pooled SD as the standardiser, is calculated. Since the mean difference (MD) between OA and healthy population has already been measured in the meta-analysis, which is equal to ΔT_2^{META} (Overall), the effect size at each (True T_2, SNR_0) for a fitting method in the simulation (denoted ES^{SIM}) can then be expressed as:

$$\begin{aligned} ES^{SIM} = \text{Glass's } \Delta &= \frac{\text{MD between OA and Healthy Population}}{\mathrm{SD}_{Healthy}} \\ &= \frac{\Delta T_2^{META}(\text{Overall})}{\mathrm{SD}} \end{aligned} \tag{2.58}$$

As can be seen in Fig. 2.10 on the following page, when the $\Delta T_2^{population}$ is fixed at ΔT_2^{META} (Overall) (shown as the fixed length in red), effect size will decrease with an increasing SD at different (True T_2, SNR_0) pairs. Therefore, ES^{SIM} can be considered as the re-expression of precision (=1.96SD), that better relates the random fitting error to the diagnostic performance (ES^{META}) found in the meta-analysis.

To achieve the diagnostic performance of ES^{META}, the ES^{SIM} should be maximised and be at least not less than the ES^{META}. Thus the bottom line criterion for effect size, defined by the meta-analysis result ES^{META} is:

$$ES^{SIM} \geqslant ES^{META} \tag{2.59}$$

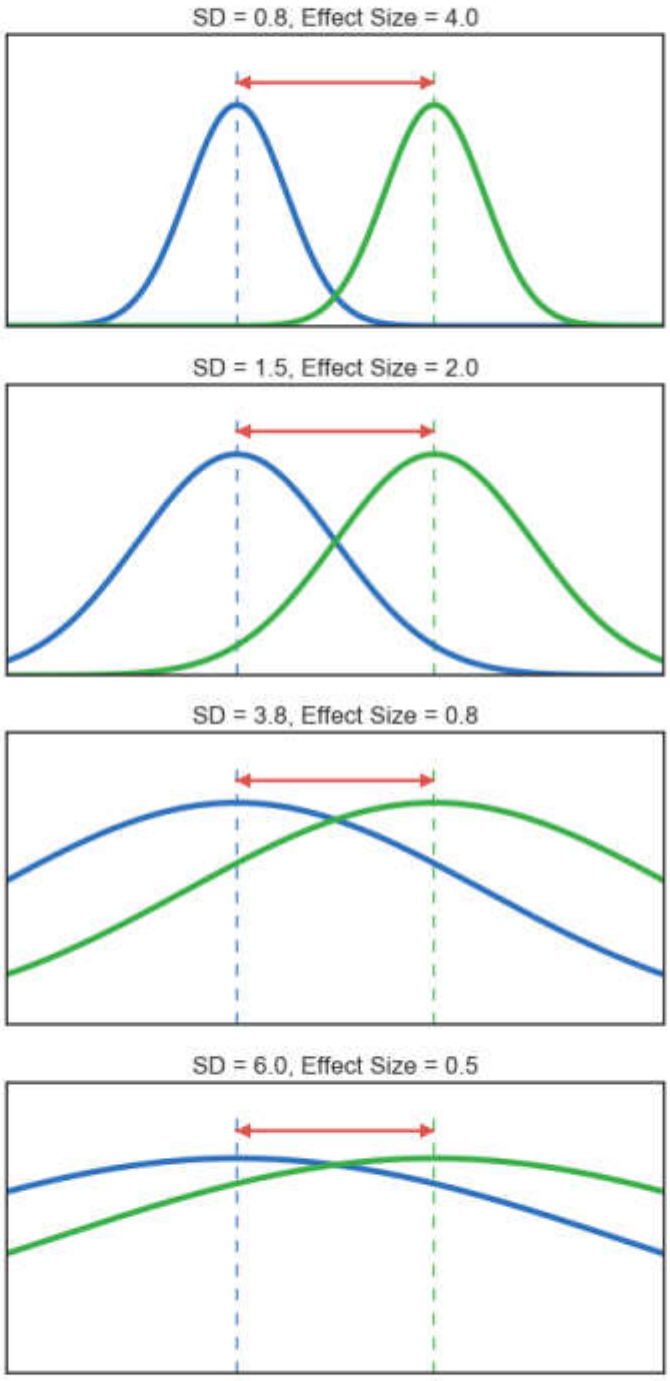


Figure 2.10: Illustration of the Relationship Between Effect Size and SD When $\Delta T_2^{population}$ is Fixed at ΔT_2^{META} (Overall)

2.3.3 Results and Discussion

2.3.3.1 Trueness

Trueness heatmaps of the 16 fitting methods under simulated (True T_2, SNR_0) conditions are plotted in Fig. 2.11 on page 64. There are two key aspects to interpreting trueness: quantity, and consistency. It is desirable for calculations to be as true as possible (values close to zero), and to be consistent with changes in T_2 and SNR_0.

Sign and Quantity of Trueness

The diverged sequential colour palette encodes the sign and quantity of trueness values under different (True T_2, SNR_0) conditions. The warm colour represents positive trueness values and overestimated T_2, while the cold colour represents negative trueness values and underestimated T_2. A light colour, either warm or cold, is desirable because it represents a trueness value (systematic fitting error) close to zero.

Consistency of Trueness

The consistency of trueness can be assessed using the heatmap by inspecting the colour consistency and transition boundaries in the heatmaps. Where colour is consistent in the range of the experiment/measurement(s), calculated T_2 can be directly compared. If the colour transition boundaries are perpendicular to the x-axis (for example, the colour transition at $SNR_0 \approx 25$ in NLS), the trueness is consistent in terms of SNR_0 (consistency in vertical direction). That is, if the SNR_0 is fixed, the trueness values remain the same for different True T_2. If the colour transition boundaries are parallel to the x-axis (for example, the colour transition at True $T_2 \lesssim 30$ in SQEXP), the trueness is consistent in terms of True T_2 (consistency in horizontal direction). That is, if the True T_2 is fixed, the trueness values tend to remain the same for different SNR_0.

It is often assumed SNR_0 is fixed for short-term single-centre studies. Under this assumption, the difference in True T_2 between groups and within an individual is the only changing variable, and using an fitting method that has trueness consistency in vertical

directory will ensure the difference in True T_2 will not be masked by systematic error. As can be seen in Fig. 2.11 on page 64, methods like LLS, NLS, NCEXP$_\sigma$, and MLE$_{Rice}$ all have good vertical consistency that meet this requirement.

However, the assumption of fixed SNR$_0$ is often violated in the real world, because many factors will affect the SNR$_0$. For example, it is often difficult to keep the SNR$_0$ unchanged in long-term, multi-centre studies, due to the different scanner hardware (including hardware upgrade), different software versions (including software upgrades, and not limited to the change in T_2 mapping software), and different operator trainings, etc.

Even the SNR$_0$ in a single scan cannot be assumed fixed. Many factors can affect the SNR$_0$ within an image. These factors include image nonuniformity (B_1 error), parallel imaging (varying g factor by location across the image), and partial volume effects, etc. Thus, a method should have both vertical and horizontal consistency.

It should be noted that for clinical use, in which the separation of normal from OA signals is of primary interest, some inconsistency in trueness may be beneficial. This can be seen in the heatmap of SQEXP$_\sigma$. The trueness of this fitting method is consistent with yet SNR$_0$ and shows a transition with True T_2. Here, the direction of transition is important: it underestimates short True T_2 yet overestimates long True T_2, artificially increasing discriminatory power.

Bottom line Criterion of Trueness

Overlaid on the heatmaps, the bottom line criteria $+\Delta T_2^{META}$ for knee and hip are plotted as solid lines, and $-\Delta T_2^{META}$ as dashed lines. These lines mark the boundary between the 'safe region' (to the right of the lines) where trueness meets the bottom line criterion in Eq. (2.55) on page 53, and the 'failed region' (to the left of the dashed lines) where the criterion was not met. It is therefore desirable for these boundary lines to be as far as possible to the left, leaving a large 'safe region' to the right in which the bottom line criteria are met. The position of this boundary line relates directly to the lowest SNR$_0$ for which fitting can be successfully performed.

As can be seen in the Fig. 2.11 on page 64, LLS, NLS, NCEXP, NCEXP$_\sigma$, SQEXP$_\sigma$ and MLE$_{Rice}$ allowed a large 'safe region'. Among these, NCEXP$_\sigma$ and SQEXP$_\sigma$ performed especially well as the boundary lines in their heatmaps are at SNR$_0 \approx 10$, suggesting a superior low SNR performance of these Rician models of moments.

In low-noise scans (SNR$_0 \gtrsim 35$), LLS can be applied with good trueness and, as the simplest method, may be desirable. The slightly more complicated NLS method can be used at SNR$_0 \gtrsim 25$. For noisy scans, it may be necessary to use complex methods such as NCEXP$_\sigma$, SQEXP$_\sigma$, and MLE$_{Rice}$, however care should be taken to choose a method relative to the noise and T_2 'window' of the scans to avoid regions of inconsistency. For single-patient use, SQEXP$_\sigma$ may be preferred due to its artificial separation of T_2 and good consistency with SNR$_0$.

2.3.3.2 Precision

Precision heatmaps of the 16 fitting methods under simulated (True T_2, SNR$_0$) conditions are plotted in Fig. 2.12 on page 65.

Quantity of Precision

The sequential colour palette encodes the quantity of precision values under different (True T_2, SNR$_0$) conditions. Different to trueness, the precision is always positive as SD cannot be negative. A light colour is desirable because it represents a precision value (random fitting error) close to zero.

Ignoring the lines for now, it can be seen all the methods have a clear dark to light colour transition from top-left to bottom-right, which is different from the heatmaps of trueness. Here, the effect of SNR$_0$ is compounded by the True T_2. Intuitively it may be expected that the short T_2 signal's fast descent into noise would increase random errors in fitting. This effect, however, is dominated by the larger relative noise at high True T_2, requiring better SNR to resolve.

Bottom line Criterion of Precision

As with trueness, the bottom line criteria ΔT_2^{META} for knee and hip overlaid. Again, these boundary lines divide the 'safe region' (on the right) where the criterion in Eq. (2.57) on page 55 is met, and the 'failed region' (on the left) where the criterion is not met.

As can be seen Fig. 2.12 on page 65, the area of the 'safe region' for all methods are limited compared to trueness. This marks one of the biggest challenge in using T_2 for OA. At common levels of SNR_0 (for example, $SNR_0 = 5$–105), the random fitting error due to the noise in the image could become comparable with the clinically meaningful difference (ΔT_2^{META}), or biological effect being measured.

Nevertheless, it can be seen that NLS, $SQEXP_\sigma$, M^1NCM_σ and M^2NCM_σ maintained the largest 'safe region' compared to other methods. In these methods, ΔT_2^{META} for knee requires approximately 15–20 more SNR_0 than hip to resolve, and to make the measurement safe for the whole True T_2 range of 20–80, all these methods would require a $SNR_0 > 105$, outside the simulation range of this study. $NCEXP_\sigma$ have a large safe region for hip, but not for knee.

In summary, NLS, $SQEXP_\sigma$, M^1NCM_σ and M^2NCM_σ are the better fitting methods in terms of precision. Despite this, low precision may make absolute interpretations of T_2 meaningless in a clinical setting.

2.3.3.3 Effect Size

Effect size heatmaps of the 16 fitting methods under simulated (True T_2, SNR_0) conditions are plotted in Fig. 2.13 on page 66. The effect size is a re-expression of SD, and can be interpreted in the same way as precision, though here a deep colour, which represents a higher effect size, is desirable.

The effect size bottom line criteria, relating the random fitting error directly to the population difference that was found in the meta-analysis, are also plotted.

Comparing effect size with precision, one can clearly see that the colour patterns are similar. The bottom line criteria of effect size, however, is less strict than those of precision. This is because precision is for resolving a biological effect of $\Delta T_2^{individual}$, which is used in individual diagnoses, whereas effect size is used to resolve a population difference of $\Delta T_2^{population}$, which permits more overlap and can still differentiate two populations.

Therefore, as can be seen in Fig. 2.13 on page 66, the 'safe region' for knee is very large compared to that of hip, contrary to that in the precision heatmaps. This means the ES^{META} (Knee) is relatively easy to achieve than hip using the current fitting methods. However, this should not be interpreted as a lower SNR_0 being sufficient for a population study of knee. Actually, the low ES^{META} (Knee) we found is a reflection of poor knee scan qualities that need to be improved. Eventually, all the ES^{META} should evolve towards the white dashed line of $ES = 2.0$, and if all the knee and hip scans follow the SNR requirement of that white dotted line, the discriminatory performance (effect size) of knee and hip T_2 in future meta-analyses would all be improved.

It can be seen NLS, $NCEXP_\sigma$, $SQEXP_\sigma$, M^1NCM_σ and M^2NCM_σ can all achieve the current performance of ES^{META} (Knee) at a lower SNR_0 of 15–55, and the ES^{META} (hip) at SNR_0 of 25–95.

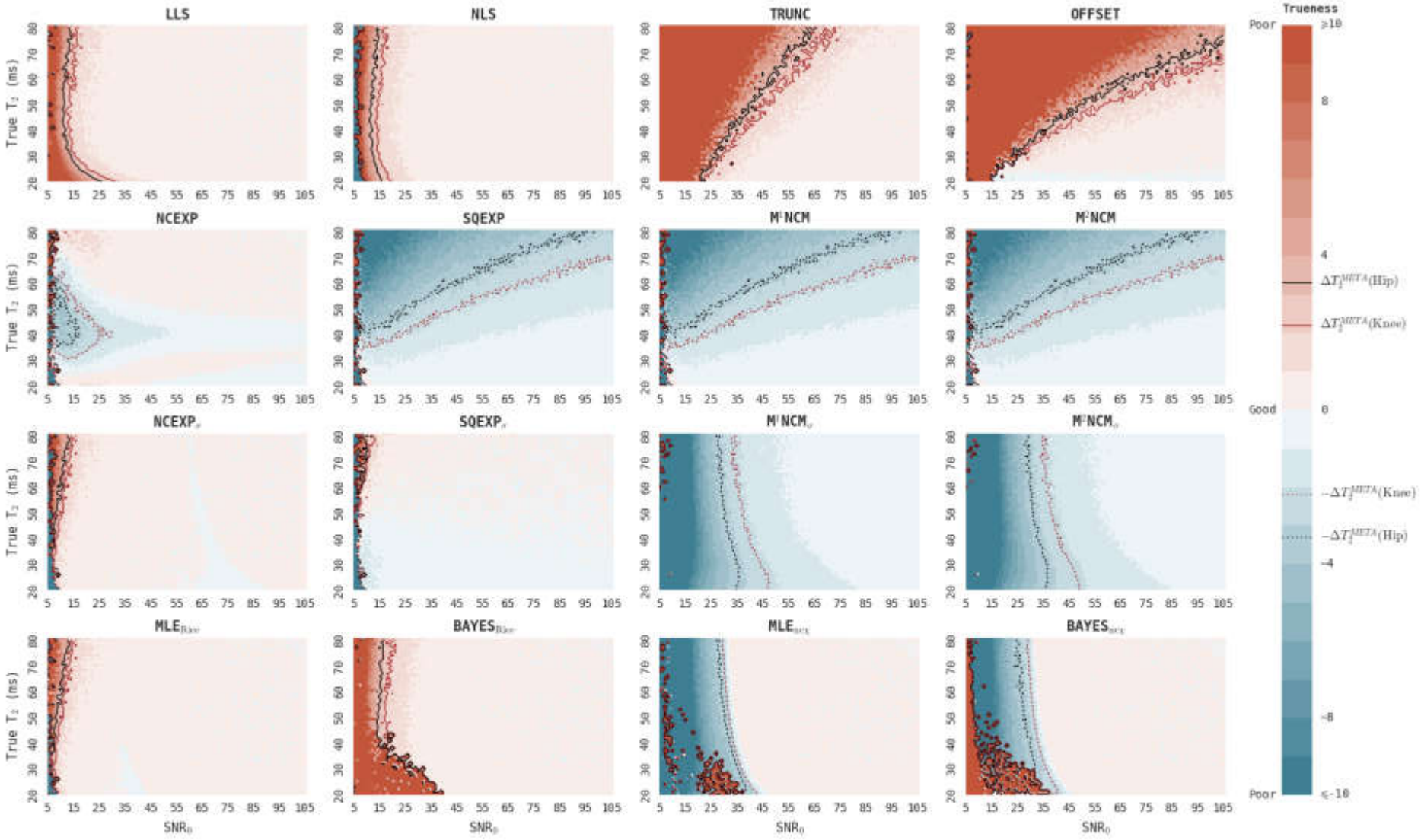


Figure 2.11: Trueness Heatmaps of the 16 Fitting Methods Under Simulated (True T_2, SNR_0) Conditions. The Diverged Sequential Colour Palette Encodes the Sign and Quantity of Trueness. The Solid and Dashed Lines Marks the Bottom Line Criterion.

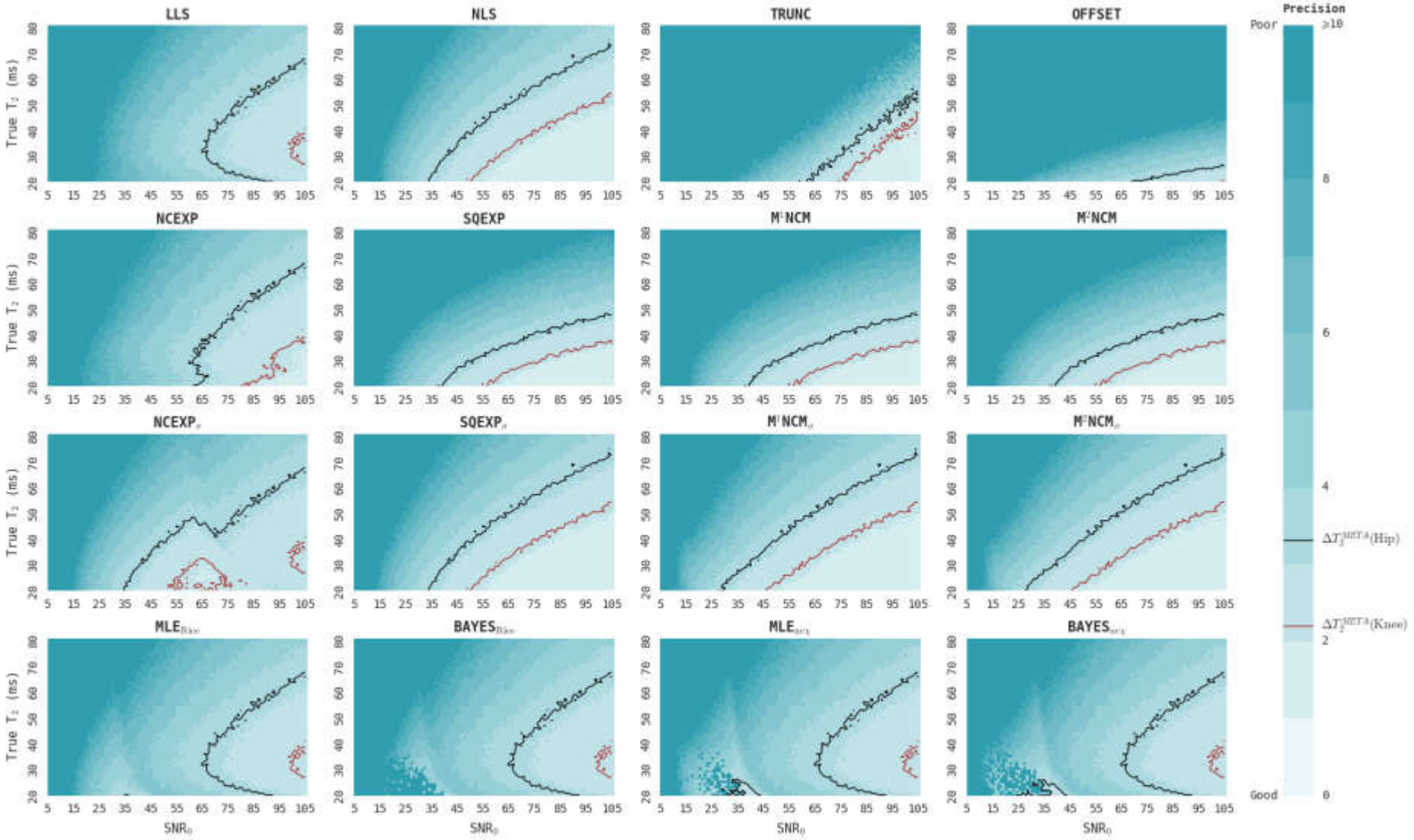


Figure 2.12: Precision Heatmaps of the 16 Fitting Methods Under Simulated (True T_2, SNR_0) Conditions. The Sequential Colour Palette Encodes the Quantity of Precision. The Lines Marks the Bottom Line Criterion.

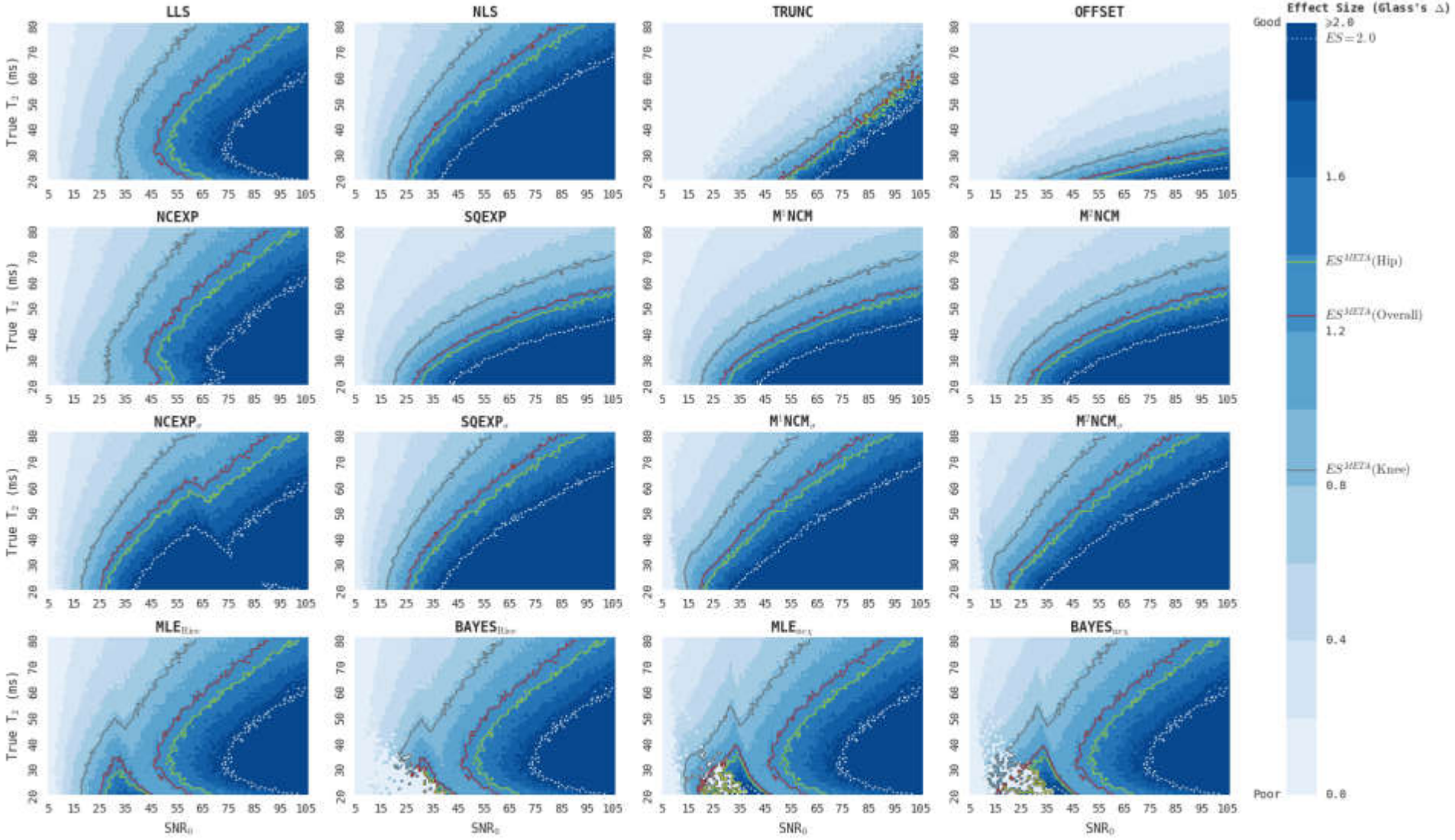


Figure 2.13: Effect Size Heatmaps of the 16 Fitting Methods Under Simulated (True T_2, SNR_0) Conditions. The Sequential Colour Palette Encodes the Quantity of Effect Size. The Solid Lines Marks the Bottom Line Criterion.

2.3.3.4 Overview and Recommendations

The most important and surprising result from this study is that NLS, a simple method, showed a good performance in all the three metrics. The only scenario in which NLS cannot be used is when SNR is low. In this case the more complicated methods NCEXP_σ and SQEXP_σ should be implemented to maintain the consistency of trueness. This improvement in trueness, however, may be meaningless as a very low SNR will surely violate the bottom line criterion for precision in a clinical setting.

The findings in this study indicate that SQEXP_σ should be used in a standard clinical setting, and for trials as the most consistently high-performance method. With the exception of SQEXP_σ we would recommend NLS could also be recommended above all other methods for general use.

Limitations in precision further lead to conclusions about interpretation of T_2 mapping. Our work suggests that absolute measurements are of little use, and that an analysis considering patterns of signal in the tissue will provide stronger diagnostic performance. The following chapters describe the development and implementation of such an approach.

This page intentionally left blank

Chapter 3

Automating the Preprocessing of Musculoskeletal MRI: Development of OxteoFlow

3.1 Introduction

Currently, MR scans, whether anatomical or quantitative, in the musculoskeletal field are analysed mainly in an segmentation-enabled approach. The information (mainly intensity, contrast and some distance/thickness measurements) stored in the MRI, are extracted by performing segmentation before they are ready for analysis. At Oxford, very few clinical trials involving musculoskeletal MRI implemented an automated processing package to process the imaging data. Sadly, the data are often extracted by medical students or junior doctors with months or even years of repeated labour of manual segmentations, which suffers from a great inter- and intra- rater variation and a very long segmentation time per image [102].

The bigger problem with the segmentation-enabled analysis is the quality of the analysis itself. In those analyses, direct comparison of images is not possible because the shape of every subject is different. To work around this and to perform intra- and inter-patient

comparison, cartilage is segmented (manually or automatically) into smaller, more controllable compartments, each approximately corresponding to a particular region of interest (ROI) [73]. In those segmented ROIs, large number of voxels are combined to a single representative number (with confidence interval), and subsequent statistical analyses are based on comparing these numbers. A large amount of information contained in the patterns, interactions, gradients, textures, and local variations of voxels are lost.

In response to this problem, OxteoFlow is developed in this study. This uses an approach based on analytical methods for functional MRI of the brain [103], which are translated to the musculoskeletal field. Specifically, it involves aligning all subjects in a reference space defined by a standard template—a process called spatial registration (or normalisation), then performing voxel-based statistical analysis in that space. The voxel-based statistical analysis of fMRI is now well established in neuroimaging community, and it would be extremely valuable if the same analysis techniques could be used in musculoskeletal MRI.

This chapter discuss the development of OxteoFlow, that is, the translation and optimisation of state-of-the-art image processing techniques from the neuroimaging community to perform the preprocessing for knee and hip images.

3.1.1 Philosophy Behind the OxteoFlow

Self-evolving with State-of-the-art

OxteoFlow is based on packages developed in neuroimaging community. It is a translation of general imaging processing theory and mathematical tools to a specific domain. The life of the OxteoFlow will go beyond the period of a DPhil research project due to the inclusion of these state-of-the-arts and self-evolving packages, most of which are open source and are still under active development. Should any of the package become outdated, they can easily be replaced by an updated version. The following details the goals of OxteoFlow:

1) **Easy to Optimise**

An advantage of OxteoFlow is that it is available as a task oriented software. Although its parameters are only optimised for preprocessing knee T_2 and hip dGEMRIC scans in this study, it can also be optimised for other kind of scans in its Optimisation Mode.

2) **Reproducible**

Reproducibility includes two meanings: the results generated by OxteoFlow can be reproduced, and the OxteoFlow itself can be reproduced as well. Based on open source software, the ultimate goal for OxteoFlow is to be used by musculoskeletal imaging community.

3) **Enabler for Better Analysis**

OxteoFlow is not developed for its own sake. It is developed to enable the advanced voxel-based statistical analysis of musculoskeletal MRI, which can not be done by performing the traditional manual segmentation enabled approach.

3.2 Theory

The goal of preprocessing is to transform the the individual scans so that they are aligned with one another, a process known as registration or spatial normalisation. In this section, the key concepts in registration and some other steps in preprocessing are briefly discussed, as these are necessary for understanding the underlying algorithms and parameters used in OxteoFlow.

3.2.1 Reference Image and Reference Space

In registration, a reference image (also referred to as template, target, or fixed image) need to be used to align the different input images (also referred to as individual, source, object, or moving image) into a common space. In fMRI processing, the coordinate space defined by the reference image is called the reference space. All the input images will

be transformed into, and later analysed within the reference space, so choosing a good template is a very important first step in image preprocessing [104; 105].

A template can be an image from a single individual, or an average of many individuals. In neuroimaging, the first, and best-known template was the brain atlas created by Talairach and Tournoux [106]. The Talairach atlas is not actually an image, but is a set of coordinates and their anatomical landmarks, based on one reference brain they selected. Although the atlas has been widely used as a template and many algorithms have been developed to transform arbitrary neuroimaging scan into the Talairach space, it has a major problem that the original MR scan, on which the atlas is based, is not available, so an accurate image to image registration is not possible using the Talairach atlas.

The templates most commonly used in neuroimaging , are the templates published by Montreal Neurological Institute (MNI), known as the MNI templates, created by averaging together hundreds of individual anatomical scans [107]. The templates are blurry, due to the averaging, but represent the approximate shape of an average human brain. Among all MNI templates, the MNI152 (or ICBM152) template is the most widely used as the standard normalisation template.

As there are no ready reference images like the MNI templates in musculoskeletal MRI, such a reference space needs to be manually defined. In OxteoFlow, one knee and one hip are carefully selected from the imaging database used in this study (OAI and Sibkids) as the reference images. The detailed method of selecting the reference images can be found in Section 3.3.5.3 on page 107.

3.2.2 Linear Transformations

The simplest model used in registration involves the linear, or affine transformations, which can be defined as:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = A \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.1}$$

where A is a 4 × 4 affine transform matrix that is uniquely defined by 12 values (3 for translations, 3 for rotations, 3 for zooms and 3 for shears). The sum of numbers of these values used in a model is called its degrees of freedom (DOF). For example, a 12 DOF means to perform a full 3D registration. A is normally in the form:

$$A = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} & a \\ m_{2,1} & m_{2,2} & m_{2,3} & b \\ m_{3,1} & m_{3,2} & m_{3,3} & c \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

where a, b and c are used for translations and $m_{i,j}$ are used for rotations, zooms and shears. This is sufficient to model overall differences in position and size between different scans of the same tissue. However, a feature of liner transformations is that any set of points that fell on a line prior to the transform will continue to fall on a line after the transform. Thus, it is not possible to model the changes in the shape of an object (such as bending) using linear transformations.

3.2.3 Non-linear Transformations

Non-linear transformation offers much greater flexibility in the registration of images than linear transformations. It can be defined as:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = A \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} + \begin{bmatrix} d_x(x,y,z) \\ d_y(x,y,z) \\ d_z(x,y,z) \\ 1 \end{bmatrix} \tag{3.3}$$

where $d_i(x, y, z)$ are known as displacement fields, or warp fields. They each descibes how far the point at location (x, y, z) should be displaced in the i direction. Different algorithms use different functions in $d_i(x, y, z)$: in the case of FNIRT (FMRIB's Non-linear Image Registration Tool), for example, they are represented by quadratic or cubic B-splines [103].

3.2.4 Cost Function

A standard, and common, way of formulating registration as a mathematical problem is to construct a cost function, which quantifies the dissimilarity between two images, and then search for the transformation which gives the minimum cost. This can be expressed as:

$$T^* = \arg\min_{T \in S_T} C\left[Y, T(X)\right] \tag{3.4}$$

where X is the input images, $T(X)$ is the image after it has been transformed by a transformation T, Y the reference image, $C[I_1, I_2]$ the cost function, S_T the space of allowable transformations, and T^* the transformation that gives the minimum cost [104].

3.2.5 Optimisation of Cost Function

Optimisation methods for image registration attempt to find the particular set of parameter values that minimise the cost function. The simplest method would be to exhaustively search all combinations of possible values for each parameter and choose the combination that minimises the cost function. Unfortunately, this exhaustive search approach is computationally unfeasible for the optimisation of a model parameter. Instead, a common class of methods perform what is called gradient descent, in which the parameters are set to particular starting values, and then the values are modified in the direction that best reduces the cost function, by calculating the first, and possibly second derivatives of the cost function [104; 105].

3.2.6 Regularisation

The gradient descent methods are very powerful in solving cost function optimisation problems, but they are susceptible to local minima—the local valleys in which the optimisation method can get stuck. Local minima can lead to an anatomically unreasonable warps in the transformed image that are complex, sharp and radical in shape. These warps might provide better-intensity matches to the template, but as they are anatomically unreasonable, the output image with these kind of warps cannot be used in subsequent analysis. All non-linear registration is a compromise between minimising the cost function (make the images look similar) and avoiding the local minima (making the warps/displacements reasonable and smooth [103].

This compromise is formalised by regularisation, which are methods that give a penalty for certain values of the parameters. Examples of regularising functions are 'membrane energy' and 'bending energy', which are used to quantify the complexity of the warps themselves. This means that they have a high value for complex, sharp warps and a small value for reasonable, smooth warps. Hence the cost function in Eq. (3.4) on the preceding

page, after including regularisation, becomes [103]:

$$T^* = \underset{T \in S_T}{\arg\min} \{C[Y, T(X)] + \lambda \epsilon(T)\} \tag{3.5}$$

where ϵ is the regularising function and λ the parameter which controls the importance of the regularisation term, which balances between similarity of images and complexity of the warps. In this way, regularisation could prevent unreasonable warps while still allowing relatively fine changes in non-linear transformations.

3.2.7 Sub-sampling

Another way to avoid local minima is sub-sampling, that is, degrading the resolution of the images, by some factor and then register the resulting low-resolution images together. This ensures that the largest structures in the images are aligned first. For example, in OxteoFlow's Non-linear Registration Layer FNIRT is optimised to first perform the registration at a resolution of 8mm. Once this is completed, FNIRT will use warp fields as the initial values to perform a second registration at a finer resolution (in OxteoFlow, 4mm), and so on until finally FNIRT is using the full resolution of the images, i.e. no sub-sampling [103].

3.2.8 Diffeomorphism

A specific kind of transformation, known as a diffeomorphism (for differentiable homeomorphism) has gained great interest recently due to its ability to effectively align structures across individuals in a way that respect anatomical constraints. A diffeomorphism is a differentiable map with a differentiable inverse, which means there can only be one point in the input space mapping into any given point in the reference space, and this mapping can be reversed. Diffeomorphism is a desirable property people try to achieve in non-linear registration. In recent years it has become popular to use algorithms where

diffeomorphism is guaranteed by construction [108].

Diffeomorphism can be understood by considering how the warps are constructed from a series of iterations. In algorithms like FNIRT which are not diffeomorphic in construction, the warp field is constructed as the sum of the updates calculated in each iteration. This means that even if each and every update is diffeomorphic itself, the sum may not necessarily be. To ensure diffeomorphism, FNIRT checks non-diffeomorphic field (by searching for Jacobians within certain range) after each sub-sampling level and if it finds any it will project the non-diffeomorphic field onto the closest diffeomorphic one. That is to say, FNIRT is not constructed diffeomorphic but it does ensure diffeomorphism, although in a less elegant way [103].

In contrast, algorithms like SyN (Symmetric Normalisation) provided in ANTs (Advanced Normalisation Tools) are diffeomorphic in construction, because the warp field is constructed by warping the warps in the last iteration, and that guarantees that the warp field is diffeomorphic as long as each of the iterations are diffeomorphic. This means user can use large regularisation λ when calculating each update to ensure diffeomorphism in each step, perform many updates (iterations) and still obtain a large warp. This is much slower than the FNIRT method but could solve some difficult cases in which FNIRT is not able to provide a satisfactory registration [108].

3.2.9 Interpolation

Once the parameters for transformation have been estimated, they are applied to the original image to create a transformed, or resliced image. This involves filling in the voxel values in the coordinate space using the transformed image coordinates and the original intensities. However, in general the transformations will involve fractions of voxels, making transformed voxels not overlap exactly with the original voxels. Therefore, it is necessary to interpolate the values of the original image intensities in order to obtain the transformed image. Common interpolation methods includes nearest neighbour, (tri)linear, sinc, and

B-spline [103; 108].

3.2.10 Intensity Non-uniformity

Intensity non-uniformity is the smooth intensity variation in MR images caused by inhomogeneity in the static magnetic (polarising) field (B_0), the radiofrequency (excitation) field (B_1) and from regional differences in the magnetic properties of the tissue being imaged [109]. Such intensity variation may have very little effect on the visual inspection of the MR image, but they do significantly reduce the accuracy of automated image processing procedures such as registration [110]. The model commonly used to describe intensity non-uniformity is a multiplicative model with additive noise [111]:

$$v(\mathbf{x}) = u(\mathbf{x})f(\mathbf{x}) + n(\mathbf{x}) \tag{3.6}$$

where at location $\mathbf{x}$, v is the measured signal, u is the true signal emitted by the tissue, f is an unknown smoothly varying bias field, and n is white Gaussian noise assumed to be independent of u. The problem of compensating for intensity non-uniformity is the task of estimating f. The combination of additive and multiplicative interference makes this task difficult [111].

Such intensity non-uniformity can be corrected by performing a bias field removal step before performing the registration, for example, by using the N3 algorithm (see Section 3.3.1.2 on page 82) as we did in OxteoFlow. The registration algorithms often also have the options to model this bias field (for example, the intensity mapping models in FNIRT). However, it is found in this study that performing non-uniformity correction before registration yield significantly better results than only rely on the options provided by the registration algorithm.

Nevertheless, the final cost function, with the regularisation term and the options for intensity non-uniformity correction, can be expressed as:

$$T^* = \arg\min_{T \in S_T} \{C\left[Y, T(\alpha X)\right] + \lambda \epsilon(T)\} \tag{3.7}$$

where α is the bias field correction that are modelled as part of the registration.

3.3 Methods

The OxteoFlow is an image preprocessing workflow consists of layers of different functions. The main image processing tasks are performed by the **Algorithm Layers**, where images are processed by a series of carefully selected MR image processing algorithms from fMRI, translated and optimised for musculoskeletal MRI. The performance of the Algorithm Layers, and that of the whole workflow, are measured by the **Metrics Layers**, where similarity metrics scores for evaluating the goodness of the registration are calculated. Before, after and in-between the Algorithm and the Metrics Layers are the **Utility Layers**, where the inputs (images and the accompanying information) and outputs (processed image and by-products) of each individual layer are collected, modified, and cached before they are sent to the next layer, and eventually to the output folder.

The layers included in the Algorithm, Metrics, and Utility Layers are summarised in Table 3.1 on the next page, and described in detail in the following sections.

3.3.1 Algorithm Layers

3.3.1.1 Hip Extraction Layer

Inspired by the Brain Extraction Tool (BET) in the FSL (FMRIB Software Library) package [112], a Hip Extraction Tool (HET) is developed to remove the non-hip joint tissue from a hip scan, to reduce the size of the MR image and thus boost processing speed.

Table 3.1: Summary of Layers in OxteoFlow

Algorithm Layers		
Layer	Algorithm (Program)	
Hip Extraction	CHT (HET)	
Non-uniformity Correction	N3 (through *mri_nu_correct.mni* in FreeSurfer)	
Linear Registration	FLIRT (FSL)	
Convert Affine	*c3d_affine_tool* (Convert3D)	
Nonlinear Registration	FNIRT (FSL), SyN (ANTs)	
Apply Transforms	*applywarp* (FSL), *antsApplyTransforms* (ANTs)	
Metrics Layers		
Similarity Metric	Range	Priority in Optimisation
Correlation Ratio	$[0, 1]$	High
Normalised Mutual Information	$[0, 1]$	High
Correlation Coefficient	$[-1, +1]$	Medium
Mutual Information	$[0, +\infty)$	Low
Utility Layers		
Layer	Function	
Information Source	Read and pass subject ID to other layers	
Data Source	Get data from hard drive	
Reference Source	Read and pass reference image and related data	
Data Sink	Store the data generated by the workflow	
Assign ID	Prefixing/Suffixing filenames when needed	
Merge Metrics	Merge results of different metrics	
Join Metrics	Collect results of one metric in Optimisation Mode	
Save Metrics	Save results from Merge/Join Metrics Layers	

HET takes advantage of the circle shape of femoral head in the sagittal view, and uses a Circle Hough Transforms (CHT) algorithm [113] to locate the centre of hip joint in the MRI. The program then crops the image to a pre-defined smaller size, while putting the hip joint in the centre of the new image. The process is illustrated in Fig. 3.1. As can be seen, the size of central slice (slice 10) of the hip T_2 image is reduced from the original 880×880 to 400×400 (the red rectangle), after the centre of the femoral head is located and the image cropped. The volume of the new image is approximately 1/4 of the original volume, which would significantly reduce the computation time in the subsequent registration steps. The core program code of HET is translated from the documentation of CHT algorithm in the scikit-image package [114]. Due to the length of the code, the full Python implementation is not presented here.

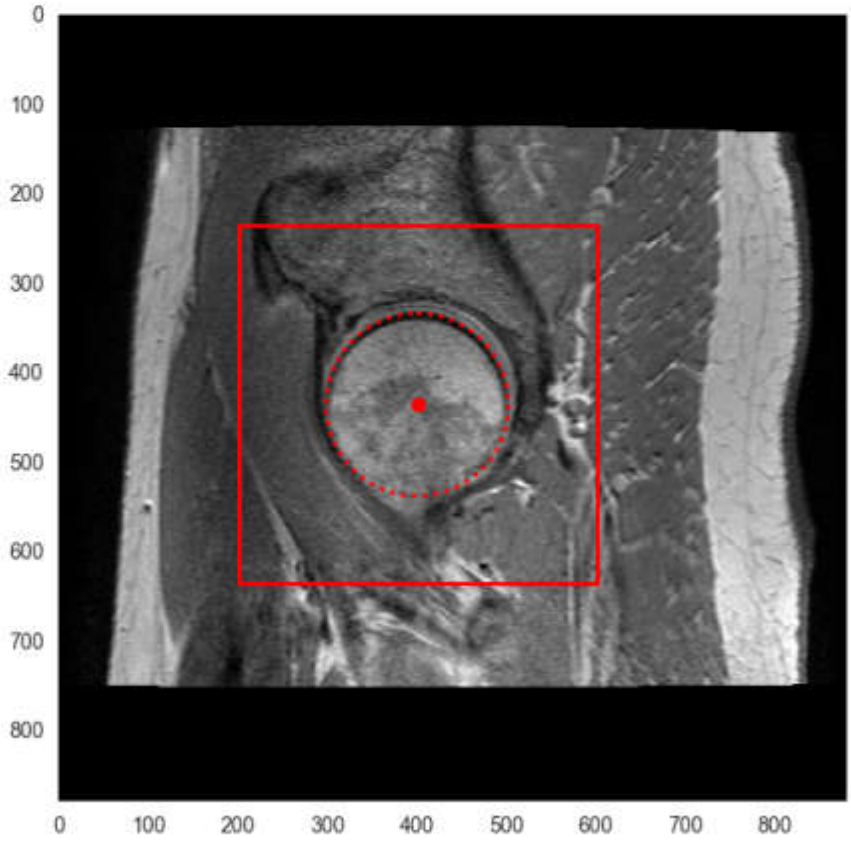


Figure 3.1: Hip Extraction Tool

Similar to HET, an equivalent extraction tool for knee is also developed but is found to decrease the registration performance of knee scans in later steps. A possible reason for

this is that in hip image the gross structure (femoral head) and the ROI (hip cartilage) are overlapped and are both kept after cropping , while in knee image the gross structures (most of the femur and tibia) and the ROI (knee cartilage) are separated. That means cropping the knee image to the ROI also removes most of the femur and tibia from the image, which is not desirable because these gross structures are important to image registration. Therefore, such a tissue extraction layer is only used as the first step for hip image preprocessing, while for knee scans the full images are always used in OxteoFlow. This difference can be seen by comparing OxteoFlow architecture graphs in Figs. C.4 and C.5 on page 271 and on page 272.

3.3.1.2 Non-uniformity Correction Layer

In early versions of the OxteoFlow, an intensity non-uniformity correction layer was not included , and all the original imaging data were input directly (for hip, after HET) into the Linear Registration Layer. This caused many registration failures despite my efforts to optimise the parameters in the registration algorithms.

The inclusion of a Non-uniformity Correction Layer is inspired by FreeSurfer's *recon-all* process, an fully-automated brain MRI processing workflow that performs all, or any part of the cortical reconstruction process [115]. Although most of the processing steps in *recon-all* are irrelevant to musculoskeletal MRI (for example, those regarding removal of skull and neck), the 'NU Intensity Correction' step in *recon-all*, which uses the Nonparametric Nonuniform intensity Normalisation (N3) algorithm [111] for correcting intensity non-uniformity (i.e. bias field) is found to be relevant to musculoskeletal MRI, and is included in OxteoFlow as an essential first step that could significantly improve the preprocessing performance.

N3 algorithm uses the notation $\hat{u} = \log u$ to convert Eq. (3.6) on page 78 into an additive model [111]:

$$\hat{v}(\mathbf{x}) = \hat{u}(\mathbf{x}) + \hat{f}(\mathbf{x}) \tag{3.8}$$

From this model, the following iterative solution is derived for obtaining the corrected image at the nth iteration [111]:

$$\begin{aligned}\hat{u}^n &= \hat{v} - \hat{f}_e^n \\ &= \hat{v} - S\left\{\hat{v} - E\left[\hat{u}|\hat{u}^{n-1}\right]\right\}\end{aligned} \tag{3.9}$$

where $\hat{u}^0 = \hat{v}$, $\hat{f}_e^0$ (the initial bias field estimate) is set to 0, and the smoothing operator S, is a B-spline approximation. The details of how N3 calculates the expected value of the true image given the current estimate of the corrected image, $E\left[\hat{u}|\hat{u}^{n-1}\right]$ can be found in [111].

N3 algorithm does not rely on a parametric model of tissue intensities, nor on segmenting a volume into contiguous regions to correct for intensity non-uniformity. It has the ability to correct for intensity non-uniformity in MR data irrespective of the pulse sequence and without prior training, and the ability to correct for non-uniformity using the structure of a random field, even when the non-uniformity is not clearly visible [111].

In my trials of the N3 algorithm, even the reference image (ID: 9581712) in the OAI, which is chosen because of its best overall image quality (see Section 3.3.5.3 on page 107), is detected by the algorithm to have a significant bias field, which is shown as a spectral image in Fig. 3.2 on the next page (3rd image in the first row). After N3 correction, the bias field is removed but the changes can be difficult to identify in the greyscale images (1st images in both rows). To better visualise this change, spectral colour map of the images are created and presented in Fig. 3.2 (2nd images in both rows), where the intensity of the bone marrow and muscles (in red colour) become clearly more uniform after correction of the bias field.

N3 is developed by McConnell Brain Imaging Centre of the Montreal Neurological Institute (MNI-BIC) and distributed as the MNI_N3 Software Package [116]. The main script *nu_correct* [117] in MNI_N3 package only accepts data in the MINC (Medical Image Network Common Data, .mnc) format. However, after years of development, most

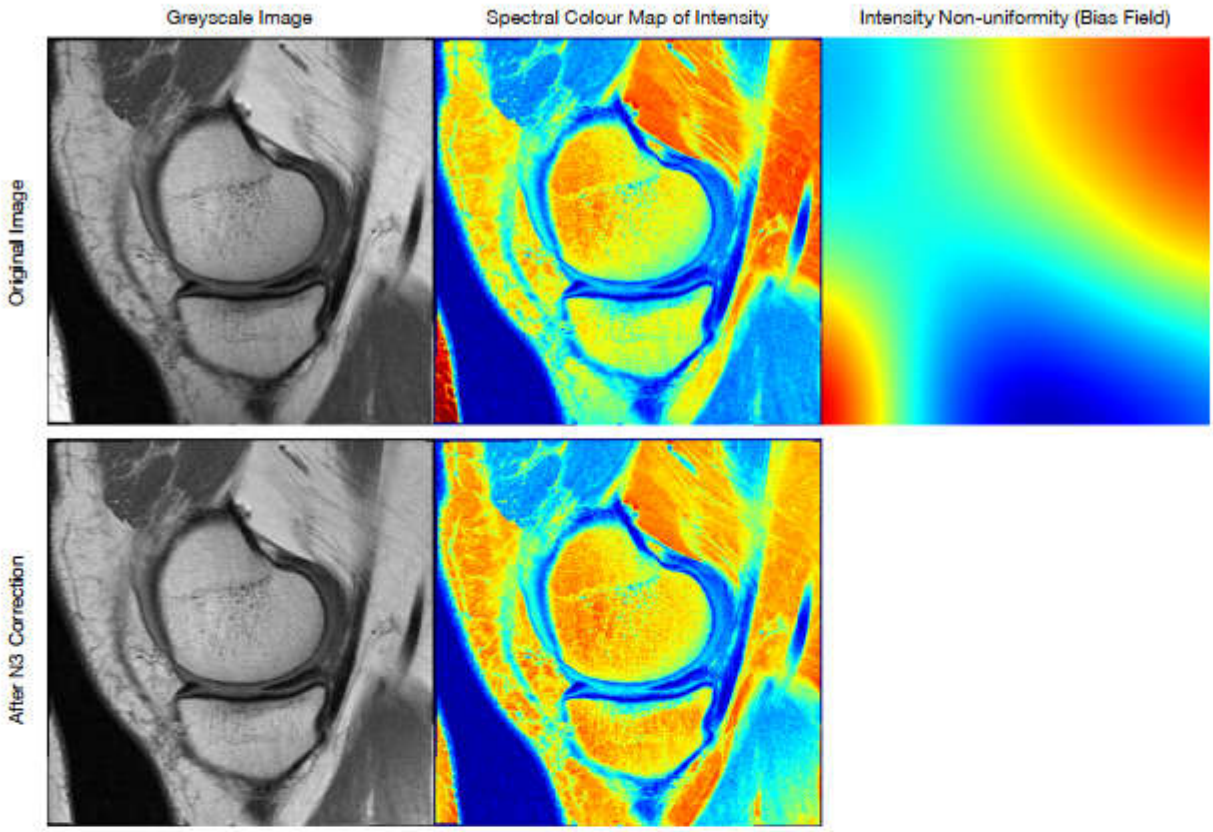


Figure 3.2: Intensity Non-Uniformity (Bias Field) Correction

modern neuroimaging data are now stored in the NIfTI (Neuroimaging Informatics Technology Initiative, .nii or .nii.gz) format, which *nu_correct* seems not support. As all other layers in OxteoFlow uses NIfTI format, it is necessary to use a NIfTI-supporting N3 algorithm.

Fortunately, FreeSurfer provides a wrapper program for *nu_correct*, named *mri_nu_correct.mni* [118], which is the tool that *recon-all* actually calls to run in its 'NU Intensity Correction' step [115]. Compared to *nu_correct*, *mri_nu_correct.mni* not only supports NIfTI format, but also uses float data type internally instead of using uchar (unsigned character), and rescales the output so that the global mean is the same as that of the input, which are useful features in MR image processing.

In OxteoFlow, the *mri_nu_correct.mni* program is optimised and used in the Non-uniformity Correction Layer. The optimised parameters can be found in Table D.2 on page 275.

3.3.1.3 Linear Registration Layer

FMRIB's Linear Image Registration Tool (FLIRT) [104; 105] is used as the algorithm in OxteoFlow's Linear Registration Layer. As part of the FSL package, FLIRT is widely used in the neuroimaging community and has established itself as one of the most important image processing tools in the field. In the well-known large-scale comparative image registration algorithm assessments, FLIRT was the only linear registration algorithm evaluated, and its linear registration result was used as the common input data in the following step where fourteen non-linear registration algorithms were evaluated [119].

FLIRT implements a global–local hybrid optimisation method. The fundamental idea of this method is to combine a fast local optimisation Powell's method with an initial global search phase before it, tuned to be computationally feasible. Here the optimisation stage of FLIRT is briefly summarised below as these are important for understanding the parameters (in parentheses) we need to optimise [104; 105]:

I. *Global search* is performed at a sub-sampling level of 8 mm, which is divided into three stages:

   (1) a coarse search over the rotation parameters (`flirt.inputs.coarse_search`) with a full local optimisation of translation and global scale in each rotation tried;

   (2) a finer search over rotation parameters (`flirt.inputs.fine_search`), but with only a single cost function evaluation (not optimisation) at each rotation (for efficiency);

   (3) a full local optimisation (rotation, translation, and global scale) for each local minimum detected from the previous stage.

II. *Local optimisation with perturbations* at a sub-sampling level of 4 mm are then performed for the best three candidates from the previous stage. For each candidate, an initial optimisation and 10 different perturbations (6 for rotation, and 4 for scale)

are also performed. Overall, a total of 33 optimisations with perturbations are performed at the 4 mm resolution;

III. In the *higher resolution stages* of 2 mm, the single best candidate from the previous stage is optimised three times with DOF = 7 then 9 then 12;

IV. In the *final stage* of 1mm, only one optimisation is performed with DOF = 12. Even so, this stage typically takes 10 minutes to run which is approximately 25% of the entire FLIRT execution time [104; 105].

The correlation ratio (CR) cost function provided by FLIRT is used in both *global search* (`flirt.inputs.cost_func`) and *local optimisation* (`flirt.inputs.cost`) stages in OxteoFlow. The mathematical model of the correlation ratio cost function provided by FLIRT is exactly the same as the correlation ratio metric in OxteoFlow's Metrics Layer, which is described in detail in Section 3.3.2.2 on page 94.

FLIRT also allows usage of cost function weighting in reference space (`flirt.inputs.ref_weight`) to affect the cost function. The voxel values of the weighting image represent how much weighting a particular voxel is given in the cost function. Note that this is different from cropping the original images (for example, using HET), as cropping introduces artificial boundaries whereas weighting does not. In this study, custom-built weighting schemes for knee and hip reference images are developed manually in ITK-SNAP [120].

All the above FLIRT parameters for musculoskeletal MRI are optimised through performing grid search optimisations. The full list of optimised parameters can be found in Table D.3 on page 275.

3.3.1.4 Convert Affine Layer

The Convert Affine Layer is a special algorithm layer placed between the Linear and Nonlinear Registration Layers, which is used to convert the FLIRT output affine matrix into a format that can be read by ANTs.

The neuroimaging convention for linear transformation matrices uses an arbitrary choice of transformation centre, i.e., a point that will be unmoved by that transformation. This arbitrary centre of the transformation for most neuroimaging package is either at the corner voxel $(0, 0, 0)$ or at centre of the FOV. However, FLIRT is an exception—it uses the centre of mass of the volume.

FNIRT has no problem in accepting this transformation matrix produced by FLIRT as they are developed in the same laboratory. However, ANTs uses the ITK toolkit format to handle affine transformation files, which is not compatible with the FSL format.

Currently there is only one tool that can be used to effectively convert between these two types of affine matrices—the *c3d_affine_tool* from the Convert3D package [121]. The program take the FLIRT-generated transformation matrix and the reference image as the input, and convert the FSL matrix into the ANTs-compatible matrix. There is no parameter to optimise in this tool, but the necessary configuration is provided in Table D.4 on page 277.

3.3.1.5 Non-linear Registration Layer (FSL Branch)

Two different algorithms are used in our Non-linear Registration Layer—the FNIRT algorithm provided in FSL, and the SyN algorithm provided in ANTs. The algorithm of FNIRT are described in this section and SyN in the next section.

FNIRT was developed by the FMRIB Centre at Oxford, UK. It implements a sum of squared difference cost function to minimise the difference between the warped input image and the reference image. The cost function can be expressed as [103]:

$$O(\mathbf{w}) = \frac{1}{XYZ} \sum_{z=1}^{Z} \sum_{y=1}^{Y} \sum_{x=1}^{X} [g_{xyz}(\mathbf{w}) - \alpha f_{xyz}]^2 + \lambda \epsilon(\mathbf{w}) \tag{3.10}$$

where g is the input image, f the reference image, and $\mathbf{w}$ the displacement (warp) field that transforms the g from its input space into a warped image $g(\mathbf{w})$ in the reference space. The

α controls intensity non-uniformity mapping and modelling (`fnirt.inputs.intensity_mapping_model`). The $\epsilon(\mathbf{w})$ represents the regularisation function (`fnirt.inputs.regularization_model`) and λ controls the weighting of regularisation (`fnirt.inputs.regularization_lambda`).

FNIRT's non-linear registration is optimised by a multi-scale Levenberg-Marquardt (see Eq. (2.35) on page 39) minimisation of the cost function in Eq. (3.10) on the previous page. The optimisation is performed based on a staged sub-sampling approach, in which the sub-sampling scheme (`fnirt.inputs.subsampling_scheme`) can be controlled by the user. FNIRT is not diffeomorphic by construction (see Section 3.2.8 on page 77). To ensure diffeomorphism, after each resolution stage the warp field is projected back onto the space of a closest diffeomorphic field with Jacobians within a predefined range (`fnirt.inputs.jaco- bian_range`).

A six stage sub-sampling scheme, with a warp resolution of **7mm** in all the three axis (`fnirt.inputs.warp_resolution`, see Section 3.4.2 on page 112) is used in OxteoFlow. The optimisation stages are described in detail below as these are important for understanding the parameters (in parentheses) we optimised [103]:

I. In first level sub-registrations, the reference image is smoothed with a 16 mm FWHM Gaussian filter (`fnirt.inputs.ref_fwhm`), and the input image an 18 mm filter (`fnirt.inputs.in_fwhm`), followed by a sub-sampling (decrease of resolution) of factor 8 (`fnirt.inputs.subsampling_scheme`). The sub-registrations at the first level are regularised using a λ of **100** (`fnirt.inputs.regularization_lambda`), and is repeated for 5 iterations (`fnirt.inputs.max_nonlin_iter`), with an intensity non-uniformity modelling applied (`fnirt.inputs.apply_intensity_mapping`). The optimised warp field is then projected back onto the space of a diffeomorphic field within jacobian range of 0.01–100 (`fnirt.inputs.jacobian_range`), before the result is used as the starting warp field of the next level;

II. In the second level, the reference image is smoothed with a 8 mm FWHM Gaussian

filter, and the input image an 10 mm filter, followed by an sub-sampling of factor 4. The second level sub-registrations are regularised using a λ of **50**, and is repeated for 5 iterations with the intensity non-uniformity modelling applied. The optimised warp field is then projected back onto the space of a diffeomorphic field before the result is used in the next level;

III. In the third level, the smoothing filters are unchanged at 8 mm and 10 mm, and the sub-sampling is unchanged at a factor of 4, but the regularisation λ decreased from 50 to **33**. The sub-registrations are repeated for 5 iterations with the intensity non-uniformity modelling applied. Again, the diffeomorphic projection is performed before passing result to the next level;

IV. In the fourth level, the smoothing filters are 4 mm for reference and 6 mm for input images, the sub-sampling is at a factor of 2, and the regularisation λ decreased from 33 to **16**. The sub-registrations are repeated for 5 iterations with the intensity non-uniformity modelling applied. The diffeomorphic projection is performed before passing result to the next level;

V. In the fifth level, no smoothing is applied to the reference image, but the input images are smoothed with a 2mm filter. Reference masks for hip T_2 scans are applied (`fnirt.inputs.apply_refmask`). No subsampling is applied (= 1) and full resolution image is used. The regularisation λ decreased from 16 to **13**. The sub-registrations are repeated for 5 iterations with the intensity non-uniformity modelling applied. The diffeomorphic projection is performed before passing result to the final level;

VI. In the final level, no smoothing, nor subsampling are applied to both the reference and the input images. Reference masks for hip T_2 scans are applied. The regularisation λ is set at **10**. Sub-registrations are repeated for 10 iterations, instead of 5, to ensure a good convergence in the final level. The intensity non-uniformity modelling is not used in the final level as it should have been determined with sufficient

precision in the previous level, this will save almost 50% of total execution time. No diffeomorphic projection is performed before the resulting image are saved.

All the above FNIRT parameters for musculoskeletal MRI are optimised through grid search optimisations. The full list of optimised parameters can be found in Table D.5 on page 277.

3.3.1.6 Non-linear Registration Layer (ANTs Branch)

ANTs was developed by Avants et al. [108] at the University of Pennsylvania. ANTs provides many registration algorithms, among which the Symmetric Normalisation (SyN) algorithm finished the 1st rank in both the above-mentioned registration algorithm assessments [119] and the EMPIRE10 evaluation of intra-subject thoracic CT registration [122]. The SyN algorithm is included in OxteoFlow as an alternative approach to FNIRT to perform non-linear registration.

Simply speaking, SyN algorithm warps both input and reference image towards their 'mean' images at the same time. On a space of diffeomorphic transformations Ω, define the path of transforming input image into reference space as $\phi(x,t)$, where the optimisation time $t \in [0,1]$. If one deform both image along a path ϕ_1 for the input image, and path ϕ_2 for the reference image towards the time point $t = 0.5$, then the full path ϕ is found by joining the paths ϕ_1 and ϕ_2. The symmetric nature of this problem is due to the interchangeability of the input and reference images: symmetric diffeomorphisms guarantee the path ϕ from input to reference is the same as it is when computed from reference to input (ϕ^{-1}), regardless of similarity metric or optimisation parameters [108].

Now consider a spacial coordinate $\mathbf{x}$. Using the symmetry, it can be seen $\phi(\mathbf{x},0) = \phi^{-1}(\mathbf{x},1) = \mathbf{x}$. Using ordinary differential equation, one can also define a time-dependent, smooth velocity field: $\boldsymbol{\nu}\left[\phi(\mathbf{x},t),t\right] = \dfrac{d\phi(\mathbf{x},t)}{dt}$. Then the corresponding full path can be obtained by integrating the velocity field in time, i.e., $\phi(\mathbf{x},1) = \phi(\mathbf{x},0) + \int_{t=0}^{1} \boldsymbol{\nu}\left[\phi(\mathbf{x},t),t\right] dt$. The length of the path, D, between $\phi(\mathbf{x},0)$ and $\phi(\mathbf{x},1)$, can then be expressed as [108]:

$$D = \int_{t=0}^{1} \| L\boldsymbol{\nu}\left[\phi(\mathbf{x}, t), t\right] \| dt \tag{3.11}$$

where L is a linear regularising operator for the velocity field $\boldsymbol{\nu}$. SyN algorithm performs image registration by minimising the transformation path length term, D, and a similarity metric term, M, within the space of diffeomorphisms Ω. This can be expressed as [108]:

$$\boldsymbol{\nu}^* = \arg\min_{\boldsymbol{\nu}} \left\{ D^2 + \lambda M\left[X, \phi(\mathbf{x}, 1), Y)\right] \right\} \tag{3.12}$$

where X is the input image, Y the reference image, and D is given in Eq. (3.11). M is a similarity metric, and λ is the weight of regularisation controlling the degree of exactness in matching. Now define in $t \in [0, 0.5]$, $\boldsymbol{\nu}\left[\phi(\mathbf{x}, t), t\right] = \boldsymbol{\nu_1}(\mathbf{x}, t)$, and in $t \in [0.5, 1]$, $\boldsymbol{\nu}\left[\phi(\mathbf{x}, t), t\right] = \boldsymbol{\nu_2}(\mathbf{x}, 1 - t)$. Exploiting the fact that the diffeomorphism path, ϕ, can be obtained be join the two paths ϕ_1 and ϕ_2, a symmetric alternative to Eq. (3.12) can be constructed as [108]:

$$\begin{aligned} \{\boldsymbol{\nu_1}^*, \boldsymbol{\nu_2}^*\} = \arg\min_{\boldsymbol{\nu}_{1,2}} \Bigg\{ & \int_{t=0}^{0.5} \left[\| L\boldsymbol{\nu_1}(\mathbf{x}, t)\|^2 dt + \| L\boldsymbol{\nu_2}(\mathbf{x}, t)\|^2 \right] dt \\ & + \lambda \int_{\Omega} M\left[X \circ \phi(\mathbf{x}, 0.5), Y \circ \phi(\mathbf{x}, 0.5)\right] d\Omega \Bigg\} \end{aligned} \tag{3.13}$$

which is the equation of SyN algorithm. Note that the regularisation term (start with λ) is equivalent to that in Eq. (3.12). The only change in Eq. (3.13) compared to Eq. (3.12) is the splitting of the integral into two time intervals reflecting the underlying optimised components of the velocity field. The gradient descent optimisation of the transformation of input and reference images will converge along the diffeomorphism path, at a fixed point midway. In SyN, the similarity metric M is cross-correlation [108].

ANTs SyN has a similar set of parameters as FNIRT that needs to be optimised. A six sub-registration scheme for ANTs is used in OxteoFlow, similar to the strategy used for FNIRT. The optimised parameters for ANTs SyN algorithm can be found in Table D.6 on page 282.

Apply Transforms Layer

The function of the Apply Transforms Layer is to apply the transformation computed by FNIRT and ANTs to an accompanying image of the input image. The accompanying image is typically the functional image in fMRI, but in this thesis the pre-computed quantitative T_2 maps .

FSL and ANTs implemented *applywarp* and *antsApplyTransforms* programs respectively for this task. The only parameter to set in this layer is the interpolation method to use when warping the T_2 maps. We used (tri)linear interpolation in both FSL and ANTs in OxteoFlow, and the configuration can be found in Table D.7 on page 284.

3.3.2 Metrics Layers

Four types of similarity metrics in the Metrics Layers are implemented to measure the registration performance of the OxteoFlow. These similarity metrics form the basis of the grid search optimisation for the (hyper)parameters (see Section 3.3.6 on page 108) in OxteoFlow.

3.3.2.1 Correlation Coefficient

Correlation coefficient (CC) is used to measure how strong a relationship is between two variables. There are several types of CC, but it most commonly refers to the Pearson's correlation coefficient [123] (also known as Pearson product-moment correlation coefficient, Pearson's R, or just R). CC is closely related to the various correlation measures that have been used in image registration. If X and Y represent intensities in two images, then CC is defined as [124; 125]:

$$\mathrm{CC}(X,Y) = \frac{Cov(X,Y)}{SD_X SD_Y} \tag{3.14}$$

$$= \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - (\mathrm{E}[X])^2}\,\sqrt{\mathrm{E}[Y^2] - (\mathrm{E}[Y])^2}} \tag{3.15}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.16}$$

where Eq. (3.14) is the definition by covariance (Cov), Eq. (3.15) the definition by expectation (E), and Eq. (3.16) the defined in terms of samples (x_i and y_i) in X and Y. Remember in Chapter 2 Section 2.3.2.2 we defined the Coefficient of Determination (R^2) (Eq. (2.36) on page 40). Actually, the reason the Coefficient of Determination is denoted R^2 is because it is the square of correlation coefficient (R) in Eq. (3.16). Therefore, CC can be considered the coefficient of the linear relationship between corresponding values in X and Y of two images.

CC varies between -1 and +1. The case CC = +1, the perfect positive correlation, occurs when X and Y perfectly coincide, while the case CC = −1, the perfect negative correlation, occurs when X and negative of Y perfectly coincide. CC is suitable for determining the similarity between images with intensities that are known to be linearly related. It is also a relatively efficient similarity measure as it requires a small number of additions and multiplication at each pixel [124].

Please also note that correlation coefficient is different from the cross-correlation metric used by SyN algorithm in ANTs (see Section 3.3.1.6 on page 90), although they are both frequently referred to as 'CC'.

3.3.2.2 Correlation Ratio

Correlation ratio (CR) is a similarity measure that quantifies the degree to which Y is a single-valued function of X, as was originally proposed by Pearson [126]. It was first used in multimodality image registration by Woods [127] and later improved by Roche [128;

129]. CR is defined as:

$$\mathrm{CR}(Y|X) = \frac{Var(\mathrm{E}[Y|X])}{Var(\mathrm{E}[Y|X]) + \mathrm{E}_X[Var(Y|X=x)]} = \frac{Var(\mathrm{E}[Y|X])}{Var(Y)} \tag{3.17}$$

$$= 1 - \frac{1}{N\sigma^2}\sum_i N_i\sigma^2 \tag{3.18}$$

Eq. (3.17) is CR expressed in terms of the conditional expectations, where $Var(\mathrm{E}[Y|X])$ measures the part of Y which is predicted by X, and $\mathrm{E}_X[Var(Y|X=x)]$ measures the part of Y which is functionally independent of X. In practice, the Eq. (3.18) is used to calculate correlation ratio, where N is the total number of voxels in the overlapping region, N_i the number of voxels that have intensity i in X, and σ^2 the variance of the intensities of the corresponding voxels in Y.

CR lies between 0 and 1, and as the variance of intensities in Y that map to each intensity in X decreases, the correlation ratio between X and Y increases. This property makes correlation ratio suitable for comparing images that have considerable intensity differences when the intensities of one is related to the intensities of the other by some linear or nonlinear function [124].

3.3.2.3 Mutual Information

Collignon and Maes [130], and Viola and Wells [131] introduced the idea that multimodal images could be aligned by maximising their mutual information (MI), which is an information measure that depends on the marginal (or Shannon) entropy (the amount of uncertainty) and joint entropy of the underlying images. Given two images X and Y, one can define MI as [132]:

$$\mathrm{MI}(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.19}$$

$$= \sum_{x,y} P(x,y)\ln\frac{P(x,y)}{P(x)P(y)} \tag{3.20}$$

where $H(X)$ and $H(Y)$ are marginal entropies and $H(X,Y)$ the joint entropy.

Another way to look at MI is through the corresponding marginal probability density functions (PDFs), $P(x)$ and $P(y)$, obtained by simple normalisation of their 2D-histogram, and the joint PDF $P(x, y)$, which gives rise to Eq. (3.20) on the preceding page.

MI represents the amount of information that one random variable contains about a second random variable, and vice versa [133]. $\mathrm{MI}(X, Y)$ yields values from 0 (no mutual information—X and Y are independent) to $+\infty$. The higher the $\mathrm{MI}(X, Y)$, the more information is shared between X and Y. However, high values of MI might be unintuitive and hard to interpret due to its unbounded range of values $\mathrm{MI}(X, Y) \in [0, +\infty)$.

3.3.2.4 Normalised Mutual Information

Studholme [134] argued that any image similarity measure should be invariant to changes in the overlap region through the course of registration. They showed that traditional MI does in fact vary with changing overlap, and they proposed the normalised mutual information (NMI) as an alternative. NMI can be defined as [132]:

$$\mathrm{NMI}(X, Y) = \frac{H(X) + H(Y)}{H(X, Y)} = 1 + \frac{I(X, Y)}{H(X, Y)} \tag{3.21}$$

$$= \frac{\sum_{x,y} P(x, y) \ln \dfrac{P(x, y)}{P(x)P(y)}}{-\sum_{x,y} P(x, y) \ln P(x, y)} \tag{3.22}$$

$$\tag{3.23}$$

where Eq. (3.21) is the definition by entropy and Eq. (3.22) by PDF. NMI measures try to bring the possible values to bounded range $\mathrm{NMI}(X, Y) \in [0, m)$. Specifically, the case of $m = 1$ is useful due to ease of comparison with commonly used CC and CR.

To put it simply, CC and MI are classical models, and CR and NMI are their improved version respectively. The Python implementations of these models can be found in the Nipy package, which are available at [135]. The usage of these metrics will be discussed further in Section 3.3.6 on page 108.

3.3.3 Utility Layers

In OxteoFlow, the Utility Layers controls the data flow in and out of the Algorithm and the Metrics Layers. They are important connector layers to make the workflow functional. Seven types of Utility Layers are used in OxteoFlow, and it would be helpful to refer to Fig. C.2 on page 269 when reading this section.

3.3.3.1 Information Source Layer

In OxteoFlow, the subject ID of an image is an important information used by many layers. For example, the Data Source Layer needs subject ID to search folder names and get data, while the Assign ID Layer needs subject ID to rename an output file by prefixing the ID to its filename. They both need the output of the Information Source Layer, the function of which is to read, save and output the subject ID to other layers.

With the Information Source Layer, a user can specify whether the preprocessing should be run on the whole data set, or only in a subgroup (for example, for optimisation), by providing the Layer a list of subject IDs. The subject IDs are then saved in the Information Source Layer and will be passed to other layers when needed.

3.3.3.2 Data Source Layer

Typical imaging studies often require running preprocessing different types of images for the same subject. For example, the study in Chapter 4 required OxteoFlow to perform registration on the first echo of knee T_2 images, and apply the resulting warp field to the accompanying T_2 maps computed using different fitting methods. One simple approach to capturing those files from the file system would be to perform a search. However, in the context of complex workflows and given that users typically arrange their imaging and other data in a semantically hierarchical folder structure, an alternative mechanism for reading and writing the data arranged in this way is often necessary. Data Source Layer is used to collect data in an organised way from the imaging database.

Data Source Layer allows a user to specify multiple filename templates to match a particular type of data, and output them through different 'fields' when needed. When Data Source Layer receives a list of subject ID from the Information Source Layer, it will search through the image database to locate the folders matching the subject IDs. It will then open the folders and compare the names of the containing files with the templates. If there is a match, it will save the path of the matching files into different fields, and output them to other layers when needed.

For example, when using OxteoFlow in Chapter 4, the Data Source Layer will save filenames ended with '_Sagittal_T2_Map_120mmFOV_0' (the first echo files of T_2 scans) into the 'scan' field, and filenames ended with '_Mapping' (T_2 maps calculated with different fitting methods) into the 'qmap' field. Later, when files are requested through the 'scan' field by, for example, the Linear Registration Layer, the Data Source Layer will output the path of the T_2 scans into the 'input image' parameter of FLIRT. When the registration of T_2 images are finished, the Apply Transforms Layer will request files through the 'qmap' field. The Data Source Layer will this time pass the saved paths of, for example, T_2 maps calculated by SQEXP, NCEXP, NLS and LLS methods into the 'image to apply the warp' field of the Apply Transforms Layer, where the T_2 maps will be transformed into the reference space.

3.3.3.3 Reference Source Layer

The Reference Source Layer is the Data Source Layer for the reference image. It captures the reference image, and all the data related to it, such as the reference weighting image, reference mask image, reference segmentation image, and reference metrics mask image. It is an important layer connecting with most of the layers in OxteoFlow, as the reference image and the related data are needed in all the Algorithm for performing the preprocessing, and in all the Metrics Layers for calculating the metric scores.

As there is only one reference image for a given data set, a user can manually specify

the paths of the reference image and the related images in the Reference Source Layer, for only once and do not need to change it again. It should be noted that the Reference Source do not pass the reference image directly to the Linear Registration Layer, instead, it passes the reference image to the Non-uniformity Correction Layer first. It is the bias field corrected reference image that is output by the Non-uniformity Correction Layer used in the subsequent registration steps. Other reference related data, like the reference weighting and masking schemes, however, are passed to other layers directly from the Reference Source Layer as these files do not need Non-uniformity Correction.

3.3.3.4 Data Sink Layer

In OxteoFlow, each subject's images are preprocessed in a working directory named with its subject ID. A working directory is like a cache. It contains not only the outputs of various processing stages, but also contains various temporary files, execution reports, hash files, and so on. As a result, navigation of useful results from irrelevant file become a difficult task.

This problem can be resolved by using the Data Sink Layer, which is used to extract useful results from the working directories and store them in output folders at a different location. With the Data Sink Layer, a user can also choose to save all the useful results generated by different layers in the same output folder, or to save them in separate output folders depending on whichever is convenient to the subsequent analysis. The resulting output folders will be saved in a hierarchical folder structure, with no irrelevant file included.

Usually a user is only interested in preserving the result that will be used in the subsequent analysis. The user can then safely delete the working folder (cache), which is usually very large. The output (useful results) folders in Chapter 4, for example, are about 100 MB per subject while the working folders (useful results + irrelevant files) are about 300 MB per subject. That means nearly 1 TB hard drive space (200 MB $\times$ 4,654 subjects) could

be saved by storing useful results cleanly and separately in output folders using the Data Sink Layer, and deleting the messy working directories.

3.3.3.5 Assign ID Layer

In the OAI study, the subject ID (for research) is not included in image metadata. As a result, the converted NIfTI images do not include subject IDs in their filenames. This means if we process the images directly using OxteoFlow, we would not be able to distinguish between the results of different subjects. The Assign ID Layer, as the name suggests, is used to rename the image using a simple name, and prefixing the subject ID to it.

For example, before the Data Source Layer passes the image named '20040909112433_RIGHT _8_Sagittal_T2_Map_120mmFOV_0.nii.gz' to the Non-uniformity Correction Layer, it will first pass it to the Assign ID Layer to rename the file. The Assign ID Layer will simplify the filename and prefix its subject ID, and the resulting filename is '9000296_T2.nii.gz'.

The Assign ID Layer is also used to rename the output file of FSL and ANTs, before they are saved in the Data Sink Layer. For example, the FNIRT and ANTs name their output files as '9000296_T2_warped.nii.gz' and '9000296_T2_trans.nii.gz', respectively. For a user with little experience with these packages, it would be difficult to know the source of these warped images. Assign ID Layer can rename the files as '9000296_T2_FSL_warped.nii.gz' and '9000296_T2_ANTs_warped.nii.gz' respectively, before passing them to the Data Sink Layer.

3.3.3.6 Merge Metrics Layer

The Merge Metrics Layer is a simple layer which is used for merging the scores computed by the four similarity metrics (CR, CC, MI, and NMI) into a Python list, before the Python list is saved as a csv file in the Save Metrics Layer.

In a fully activated OxteoFlow architecture, two Merge Metrics Layers are needed—one

for merging the four Metrics Layers connecting the FSL Branch, and another for merging the four Metrics Layers connecting the ANTs Branch. As it is always useful to save the four similarity metrics to quantify the goodness of a registration result, the Merge Metrics Layer is used in both of the Optimisation and the Execution Mode.

3.3.3.7 Join Metrics Layer

The Join Metrics Layers is used to collect metric results of a grid search, or parameter sweep in the Optimisation Mode of OxteoFlow. For example, a user might want to optimise parameters 'a' and 'b' by performing a grid search optimisation (described in detail in Section 3.3.6 on page 108). Due to the parameter sweep, for example, of 10 values for 'a' and 10 values for 'b', a $10 * 10$ parameter matrix will be produced, in which the grid represents 100 different candidate pairs of the two parameters, each producing a different warped images. Performing similarity measurements between these images and the reference image will therefore produce 100 CR, CC, MI, and NMI scores, respectively. The Join Metrics Layer is used to save, for example, the 100 CR results into a Python list, which can later be saved as a csv file by the Save Metrics Layer.

As each Join Metrics Layer can only save one type of metric, in a fully activated OxteoFlow architecture, eight Join Metrics Layers are needed—four for joining the four Metrics Layers (CR, CC, MI, and NMI) connecting the FSL Branch, and four for joining the four Metrics Layers connecting the ANTs Branch. The Join Metrics Layer is only used in the Optimisation Mode, because there is no parameter optimisation needed in the Execution Mode.

3.3.3.8 Save Metrics Layer

The Save Metrics Layer is essentially a Python list to csv writer. It is either used to save the Python list passed from the Merge Metrics Layer, into a csv file containing the 4 metrics scores in a row, or is used to save the Python list passed from the Join Metrics

Layer, into a csv file containing, for example, 100 CR metrics in a column.

For each Merge/Join Metrics Layer, one Save Metrics Layer is needed. Therefore, in a fully activated OxteoFlow architecture, ten Save Metrics Layers are needed—two for the two Merge Metrics Layers, and eight for the eight Join Metrics Layers described above.

3.3.4 Implementation

OxteoFlow is developed based on the framework of Nipype (Neuroimaging in Python—Pipelines and Interfaces), which is an open-source, community-developed Python package under the umbrella of NiPy. Nipype provides a uniform interface to existing neuroimaging software and facilitates interaction between these packages within a single workflow. Nipype provides an environment that allows interactive exploration of algorithms from different packages, eases the design of workflows within and between packages, and reduces the learning curve necessary to use different packages [136].

3.3.4.1 Interfaces and Workflow Engine

Nipype consists of many parts, but the most important ones are interfaces and the workflow engine. OxteoFlow's various layers are based on the interfaces, and the interaction of the layers and the flow of data are controlled by the workflow engine.

Interfaces in the context of Nipype are program wrappers that enable Nipype, which runs in Python, to run a program or function in any other programming language. As a result, Python becomes the common denominator of all neuroimaging software packages and allows Nipype to easily connect them to each other [136].

Workflow engine consists of nodes and workflows, which can be interconnected in various ways. A node provides the information—(parameters, filenames, etc.) needed by an interface to run the program properly for a particular job. A workflow (also called a pipeline), is a Directed Acyclic Graph (DAG) or forest of DAGs consists of vertices (nodes)

and edges (lines connecting the nodes), representing flow of data. Each node or workflow has (at least) one input field and (at least) one output field. Those fields specify the data flow into and out of a node or workflow [136].

Interfaces and the workflow engine can be illustrated by the detailed architecture graph of OxteoFlow in Fig. C.2 on page 269. As can be seen, the whole workflow is interconnected by the rectangle nodes. Each node (layer) consists of IN fields on the left, the interface that run the function in the middle, and OUT fields on the right. Data flow out of the OUT field from one node flows into the IN field of the next node, until all the data finally reaches the Data Sink.

3.3.4.2 Multi-processing Support

Central Processing Units (CPUs) with multiple cores have become the standard in modern computer architectures. Using multiple CPU cores can significantly accelerate 'CPU bound' tasks such as image registration where the majority of the task is spent in the CPU. This could potentially be done by using multi-threading, a technique to divide a *single* process into different 'sub-tasks', each of which are run in parallel in different CPU cores. However, most image processing tools, including FLIRT and FNIRT, do not support multi-threading.

An alternative approach, called multi-processing, is to submit *multiple* processes to separate cores, where every process will run completely independent from each other. This approach can be used to run multiple cases of, for example, FLIRT at the same time. In my 16-core workstation 'horla', 16 FLIRT processes can be run in parallel to perform linear registration for 16 different subjects at the same time when multi-processing is enabled.

OxteoFlow implemented multi-processing through Nipype's multi-processing plugin [136]. A user can specify how many CPU cores (and how much memory) can be used by OxteoFlow. My test shows OxteoFlow performs best when (total number -1) cores are used

for preprocessing, leaving one core for the system processes. That means, in the above example, 15 FLIRT processes, instead of 16, should be used on the 16-core workstation to maximise OxteoFlow's performance.

Multi-processing support is one of the most important features of OxteoFlow. It can significantly reduce the computation time of preprocessing a large database. For example, using 15 CPU cores in 'horla', the preprocessing of 4,654 subjects in Chapter 4 were completed in just about a month. If there was no multiprocessing support and only one CPU core was used (which is how those algorithms would be used normally), it could have taken more than a year to complete the preprocessing in that chapter.

The multiprocessing feature can be best illustrated by the System Resource Profiling Graph in Fig. 3.3 on the following page. The graph is generated using the runtime profiler callback log after an OxteoFlow execution. In Fig. 3.3, the purple bars are the independent processes running in different cores, and the green bars show the system memory usage. As can be seen, the graph showed the system resource usages at the initial stages (Data Source Layer → Non-Uniformity Correction Layer → Linear Registration Layer → Non-linear Registration Layer) of a grid search optimisation task (see Section 3.4.2 on page 112), where all the 15 CPU cores are utilised to perform the preprocessing task right from the beginning.

3.3.4.3 Optimisation and Execution Modes

Two working modes, Optimisation and Execution Modes, are implemented in OxteoFlow for different types of tasks.

The Optimisation Mode is used for fine-tuning the (hyper)parameters in the OxteoFlow for a new image set or image type. Although OxteoFlow has been optimised and validated for preprocessing knee T_2, hip dGEMRIC and hip T_2 scans in this thesis, it is difficult to guarantee the optimised parameters will fit for purpose for images acquired using other musculoskeletal MRI protocols, for example, knee dGEMRIC or sodium imaging. In fact,

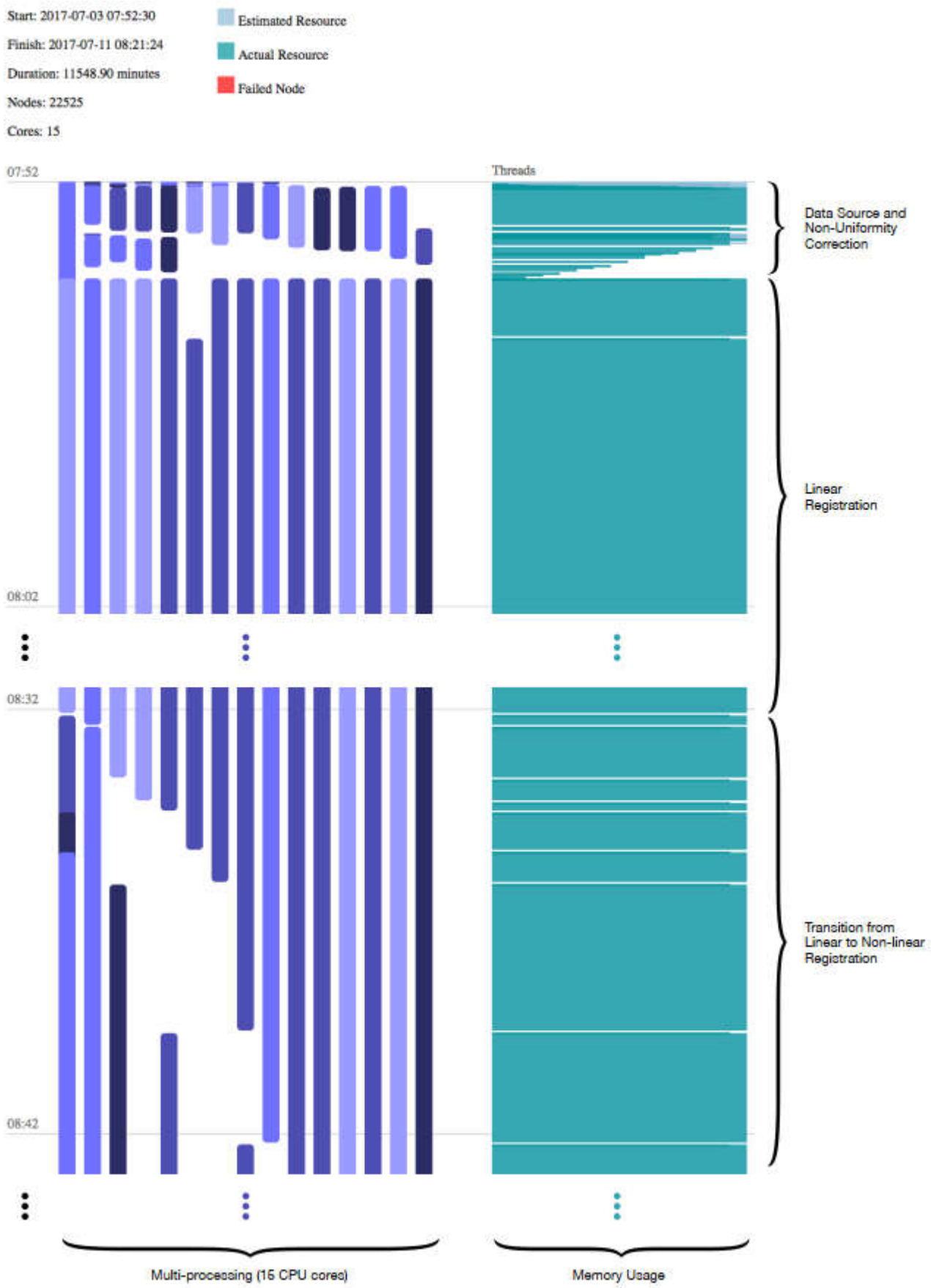


Figure 3.3: System Resource Profiling Graph Showing OxteoFlow's Multi-processing Support

even for a new knee T_2 data set (other than OAI) in which images are acquired using a different, for example, echo time, it would be important to perform an optimisation of key parameters on a small subset first, because even a small refinement in these parameter could make a huge difference in the registration results, and ultimately in the quality of the subsequent analysis.

In the Optimisation Mode, users can set lists of candidate values for different parameters, and OxteoFlow will perform registrations for all the combinations of these candidates. For example, if in an Algorithm Layer the user wants to try two different values a_1, a_2 for parameter a, two different values b_1, b_2 for parameter b, and two different values c_1, c_2 for parameter c, then OxteoFlow will test all the eight combinations of a, b, and c, namely $[a_1,b_1,c_1]$, $[a_1,b_1,c_2]$, $[a_1,b_2,c_1]$, $[a_1,b_2,c_2]$, $[a_2,b_1,c_1]$, $[a_2,b_1,c_2]$, $[a_2,b_2,c_1]$, and $[a_2,b_2,c_2]$. Then the user can decide which set of candidate is the best by comparing the similarity metric scores.

In the architecture graphs of OxteoFlow, the main feature of Optimisation Mode, compared to the Execution Mode, is the use of Join Metrics Layers. The function of this layer is described in Section 3.3.3.7 on page 100.

After all the parameters are optimised for a data type, the OxteoFlow is switched to the Execution Mode to perform an one-off run of a data set and obtain the final result for analysis. In the Execution Mode, no Join Metrics Layers are needed, because there is no parameter to optimise—each similarity metric only produces one final similarity score for a subject.

3.3.5 MR Imaging Data

The MR imaging data we used in this study are knee T_2 images from the Osteoarthritis Initiative (OAI), and hip dGEMRIC and T_2 images from the SibKids study.

3.3.5.1 OAI Knee Image Set

The OAI is a multi-centre, longitudinal, prospective observational study of knee osteoarthritis (OA). The overall aim of the OAI is to develop a public domain research resource to facilitate the scientific evaluation of biomarkers for osteoarthritis as potential surrogate endpoints for disease onset and progression [137]. The imaging protocol used in OAI for acquiring the multi-echo spin echo (MESE) knee T_2 image has been described in Section 2.3.2.3 on page 48.

There are 4,796 subjects enrolled in the OAI study. After data cleaning (see Fig. 4.2 on page 140), 4,666 right knee images are eligible for the studies in this thesis.

3.3.5.2 Sibkids Hip Image Set

The Sibkids study is a prospective longitudinal study of a cohort considered to be at risk for the development of hip osteoarthritis, whose spouses were enrolled as control subjects. The protocol for the sagittal turbo-field-echo (TFE) hip dGEMRIC sequence was TR = 6.0 ms, TE = 2.9 ms, flip angle = 12°, inversion times (TI) = 2100, 1200, 600, 250, and 105 ms, slice thickness/gap = 3 mm/0 mm, and FOV = 180 mm. The imaging protocol for the acquisition of the hip T2Q Sagittal images was TR = 3500 ms, TE = 13.5, 27.0, 40.5, 54.0, 67.5, 81.0, 94.5 ms, flip angle = 90°, slice thickness/gap = 3 mm/0.3 mm, and FOV = 150mm. All hips were scanned with a 3T scanner (Acheiva 3.0T X-series, Philips Healthcare, Eindhoven, Netherlands).

There are 32 right and 28 left hips in the SibKids dGEMRIC image set, and 5 right and 5 left hips in the SibKid T_2 image set. As the preprocessing requires all the images of the same type are of the same side, we flipped all the left hip scans to align their

gross morphological relationships with that of right hips. The resulting SibKid image set consists of 60 dGEMRIC and 10 T_2 scans of the right hips (including those flipped from left hips). We only preprocessed hip dGEMRIC images at TI = 1200 ms, and hip T_2 images at TE = 13.5 ms in this study .

3.3.5.3 Reference Images

Knee T_2 Reference Image

Candidates for reference image were selected by looking for lowest combined MOAKS score, and no individual high scores in the OAI subjects. Eligible candidates then had their image visually inspected by a clinician, who made final recommendations on the reference images to use.

Hip dGEMRIC and Hip T_2 Reference Images

As the number of image included in the SibKids study is small, the candidates for hip dGEMRIC and hip T_2 reference images are selected by visual inspection of all the scans, and selecting the ones with least morphological abnormality. The candidates scans are then read by a clinician, who then made final recommendations on the reference images to use.

3.3.5.4 Optimisation and Validation Cohort

Optimisation Cohort

We randomly selected $n = 20$ knee T_2 images from the 4,666 subjects in the OAI as the Optimisation Cohort for knee, which is used in the parameter optimisation of OxteoFlow.

Due to the small number of hip scans in the SibKids image set, and that hip images are not used in the subsequent studies in this thesis, the optimisation of OxteoFlow for hip images are only limited to parameters that are related to hip preprocessing (HET and reference mask schemes). These are performed on $n = 10$ randomly selected hip

dGEMRIC scans and and $n = 5$ randomly selected hip T_2 scans in the SibKids image set.

Validation Cohort

The optimised parameters are validated by running OxteoFlow in Execution Mode to preprocess all the images in the image sets. Therefore, $n = 4,665$ knee T_2 (= 4,666 in total − 1 reference image) in the OAI image set, $n = 59$ hip dGEMRIC (= 60 in total − 1 reference image), and $n = 9$ hip T_2 (= 10 in total − 1 reference image) in the SibKids image set are used as the Validation Cohorts to evaluate the performance of OxteoFlow on the image sets. The images previously included in the Optimisation Cohort are also included in the Validation Cohort to accurately calculate the mean metrics scores of an image set using all the available image.

3.3.6 Optimisation and Validation

3.3.6.1 Hyperparameters vs Model Parameters

It should be noted there is a difference between the optimisation of cost functions and optimisation of the parameters available as options that can be set by the users. It is easiest to describe this difference using the language from machine learning: the parameters optimised in the next section are the so called 'hyperparameters' which is set before the registration (learning) begins. The values of other parameters in the model (mainly concerning the warp field), which are not set by the user but are derived via optimisation (training) of cost functions (loss function), can be called model parameters. The next section deals with optimising the hyperparameters—the model parameters are optimised by the algorithms, and are not directly controlled by the user.

3.3.6.2 Grid search Optimisation of Hyperparameters

In section Section 3.3.4.3 on page 105, we explained OxteoFlow's Optimisation Mode by listing the eight combinations of $[a_1,\ a_2]$, $[b_1,\ b_2]$, and $[c_1,\ c_2]$. This approach of exhaustively considering all parameter combinations is called grid search, or parameter sweep.

Although arbitrary numbers ($\geqslant 2$) of parameters, and arbitrary numbers ($\geqslant 1$) of candidates for each parameter can be optimised using the grid search method, in this thesis we always perform grid search of parameters in pairs, i.e. optimising two parameters against each other at a time. This is to ensure the grid search result can be plotted into a 2D heatmap to aid visualisation, with the candidates for each parameter forming the x and y axis, and the metric scores of all the candidate pairs as points in the grid.

Assuming the parameters to be optimised in pair are a and b, with three candidates $[a_1, a_2, a_3]$ and $[b_1, b_2, b_3]$ respectively, on which we want to perform a grid search optimisation. Then the approach can be formalised by performing a 'dot product' operation ($\Rightarrow$) of column vector $[a_1, a_2, a_3]^T$, and row vector $[b_1, b_2, b_3]$ to obtain a matrix consists of nine candidate pairs:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \Rightarrow \begin{bmatrix} (a_1, b_1) & (a_1, b_2) & (a_1, b_3) \\ (a_2, b_1) & (a_2, b_2) & (a_2, b_3) \\ (a_3, b_1) & (a_3, b_2) & (a_3, b_3) \end{bmatrix} \tag{3.24}$$

OxteoFlow is then run in the Optimisation Mode to exhaustively test all the nine candidate pairs, keeping all the other parameters in OxteoFlow unchanged. The resulting warped images, one for each candidate pair, are compared with the reference image in the Metrics Layers, where CR, NMI, CC, and MI metrics scores are calculated for each resulting image. This process is repeated for all the 20 subjects in the Optimisation Cohort.

After all the tests are completed, the mean and standard deviation (SD) of the similarity

scores of the 20 subjects are calculated for all the four similarity metrics used. The results are then visualised as interpolated heatmaps overlaid by contours, on top of a parameter grid of the candidate pairs.

3.3.6.3 Decision Rule of Similarity Metrics

Using the contours in the heatmaps, it is easy to locate the highest value in each of the similarity metrics. Based on the descriptions of the similarity metrics in Section 3.3.2 on page 92, and my experience with the reliability of them, I propose a decision rule based on assigning different priority levels to the four similarity metrics, in which I give CR and NMI the highest, CC medium, and MI the lowest priority respectively.

The prioritised decision rule of similarity metrics are as follows:

I. Examine the grid points in mean CR and mean NMI heatmaps, if one, and only one point of highest score can be found by evaluating the intersection of the mean CR and mean NMI heatmaps, and if this point has acceptable SDs (not in the extreme region) in the corresponding SD heatmaps, accept the parameters represented by this point and stop;

II. If no grid point or more than one grid points are found in the intersection of CR and NMI' mean heatmaps, include CC's heatmaps into consideration and see if a decision can be made using the CR, NMI and CC:

    (1) If a decision can be made to accept one grid point, accept it and stop.

    (2) If not, include MI's heatmaps into consideration and see if a decision can be made:

        a. if a decision can be made to accept one grid point, accept it and stop.

        b. if not, accept the grid point of highest mean in CR heatmaps.

An example of applying this prioritised decision rule to find the optimum parameters is illustrated in Section 3.4.2 on page 112.

3.3.6.4 Validation

The validation of OxteoFlow is performed by running it on the Validation Cohorts (see Section 3.3.5.4 on page 108), using the optimised parameters in the Execution Mode. The mean metrics scores of the validations are then compared with the score achieved in the hyperparameter optimisation, to ensure the performance of OxteoFlow is consistent between optimisation using a small subset, and execution on the complete data set.

3.4 Results and Discussion

3.4.1 Architecture of OxteoFlow

3.4.1.1 Optimisation Mode

The full architecture of OxteoFlow is presented in Fig. C.2 on page 269, which is obtained by setting OxteoFlow to run in the Optimisation Mode for hip, so that all the Layers, including the Hip Extraction Layer and the Join Metrics Layers are activated. Fig. C.2 gives details of how the outputs of higher level layers serve as the inputs of the lower level layers, until the final outputs are saved in the Data Sink Layer.

A simplified version of Fig. C.2, which only shows the functions (names) of the layers, is presented in Fig. C.3 on page 270. Further simplification from Fig. C.3 can be obtained by only activating the ANTs branch in the Non-linear Registration Layer, which is shown in Fig. C.4 on page 271.

For preprocessing of knee images, the Hip Extraction Layer is not used and the architecture of OxteoFlow run in the Optimisation Mode for knee images is presented in Fig. C.5 on page 272, where only the FSL branch is activated in the Non-linear Registration Layer. This architecture was used to perform optimisation on knee scans, although we omitted the ANTs branch in Fig. C.5 for simplicity.

3.4.1.2 Execution Mode

The Execution Mode is used to perform a final preprocessing to obtain result, using an already optimised set of parameters for a dateset. As all the parameters have already been optimised in the Optimisation Mode, so the Join Metrics Layers are no longer needed in the Execution Mode.

An example usage of OxteoFlow's Execution Mode is to validate its performance on a random selection of 216 subjects in the OAI image set, shown in Fig. C.6 on page 273. In Fig. C.6, no Join Metrics Layers are used , because no grid search optimisation is performed in the Execution Mode.

3.4.2 Optimisation of Regularisation λ and Warp Resolution

In this section, I present the optimisation procedures of two of the most important parameters in OxteoFlow, using the grid search optimisation approach and the decision rule of similarity metrics.

3.4.2.1 Grid Search Optimisation

One of the key parameters to optimise in OxteoFlow, is the regularisation λ in FNIRT which is used to balances the similarity of the images and complexity of warp field. As I used a six-level sub-sampling scheme in FNIRT, λ has to be a parameter consists of six numbers to be in line with the sub-sampling scheme. Among all the levels, λ in the final level is the most important as it directly controls the regularisation of the warp field in the resulting image.

My experience is that the final level λ needs to be in a range of 5–50, and the first level λ be ten times of it. For example, the default four-level λ in FNIRT is (300, 75, 30, 30). Expending the default parameter into six levels, a possible scheme is (300, $\frac{300}{2}$, $\frac{300}{3}$,

$\frac{300}{2 \times 3}$, 40, 30), that is, (300, 150, 100, 50, 40, 30). Dividing every value in this expended default parameter by 6 a basis parameter with final level $\lambda = 5$, i.e. (50, 25, 16.67, 8.33, 6.67, 5) can be obtained. Multipying this basis parameter by 2, 3, 4, ..., 10, and round all the results down to the nearest integers, one can obtain a list of 10 candidate parameters, with final level $\lambda = \mathbf{5, 10, 15, \ldots, 50}$. The complete list of the 10 candidates is shown as the column vector (the term on the left) in Expression (3.25), where the final level λs are printed in **boldface**.

The second key parameter to optimise in OxteoFlow is the warp resolution in FNIRT, which controls how fast a point in the warp field can change from one position to another position in the x-, y-, and z-direction, in mm. The default value for this parameter in FNIRT is (10, 10, 10). My experience is that it should be in a range between (2, 2, 2) to (11, 11, 11), because the highest resolution (1, 1, 1) takes an unacceptably long time and too much memory to run, while any resolution lower (with valve larger) than (11, 11, 11) will make the registration result too coarse. Therefore, ten possible candidates in this range to evaluate are (2, 2, 2), (3, 3, 3), ..., (11, 11, 11), which are shown as the row vector (the term on the right) in Expression (3.25):

$$\begin{bmatrix} (50, & 25, & 16, & 8, & 6, & \mathbf{5}) \\ (100, & 50, & 33, & 16, & 13, & \mathbf{10}) \\ (150, & 75, & 49, & 24, & 19, & \mathbf{15}) \\ (200, & 100, & 66, & 32, & 26, & \mathbf{20}) \\ (250, & 125, & 82, & 40, & 32, & \mathbf{25}) \\ (300, & 150, & 100, & 50, & 40, & \mathbf{30}) \\ (350, & 175, & 115, & 56, & 45, & \mathbf{35}) \\ (400, & 200, & 132, & 64, & 52, & \mathbf{40}) \\ (450, & 225, & 148, & 72, & 58, & \mathbf{45}) \\ (500, & 250, & 165, & 80, & 65, & \mathbf{50}) \end{bmatrix} \begin{bmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{bmatrix} \Rightarrow \qquad (3.25)$$

where only one number is used to represent this three-dimensional parameter, as the

values in x-, y-, z-direction are the same.

A 'dot product' operation ($\Rightarrow$) of the column vector of the 10 candidates of regularisation λ and the row vector of the 10 candidates of warp resolution can then be performed, as shown in Expression (3.25).

The resulting matrix consists of 100 grids of candidate pairs is shown in Matrix (3.26), where we used the final level λs (in **boldface**) to represent the corresponding six-level regularisation λs.

$$\Rightarrow \begin{bmatrix} (\mathbf{5},2) & (\mathbf{5},3) & (\mathbf{5},4) & (\mathbf{5},5) & (\mathbf{5},6) & (\mathbf{5},7) & (\mathbf{5},8) & (\mathbf{5},9) & (\mathbf{5},10) & (\mathbf{5},11) \\ (\mathbf{10},2) & (\mathbf{10},3) & (\mathbf{10},4) & (\mathbf{10},5) & (\mathbf{10},6) & (\mathbf{10},7) & (\mathbf{10},8) & (\mathbf{10},9) & (\mathbf{10},10) & (\mathbf{10},11) \\ (\mathbf{15},2) & (\mathbf{15},3) & (\mathbf{15},4) & (\mathbf{15},5) & (\mathbf{15},6) & (\mathbf{15},7) & (\mathbf{15},8) & (\mathbf{15},9) & (\mathbf{15},10) & (\mathbf{15},11) \\ (\mathbf{20},2) & (\mathbf{20},3) & (\mathbf{20},4) & (\mathbf{20},5) & (\mathbf{20},6) & (\mathbf{20},7) & (\mathbf{20},8) & (\mathbf{20},9) & (\mathbf{20},10) & (\mathbf{20},11) \\ (\mathbf{25},2) & (\mathbf{25},3) & (\mathbf{25},4) & (\mathbf{25},5) & (\mathbf{25},6) & (\mathbf{25},7) & (\mathbf{25},8) & (\mathbf{25},9) & (\mathbf{25},10) & (\mathbf{25},11) \\ (\mathbf{30},2) & (\mathbf{30},3) & (\mathbf{30},4) & (\mathbf{30},5) & (\mathbf{30},6) & (\mathbf{30},7) & (\mathbf{30},8) & (\mathbf{30},9) & (\mathbf{30},10) & (\mathbf{30},11) \\ (\mathbf{35},2) & (\mathbf{35},3) & (\mathbf{35},4) & (\mathbf{35},5) & (\mathbf{35},6) & (\mathbf{35},7) & (\mathbf{35},8) & (\mathbf{35},9) & (\mathbf{35},10) & (\mathbf{35},11) \\ (\mathbf{40},2) & (\mathbf{40},3) & (\mathbf{40},4) & (\mathbf{40},5) & (\mathbf{40},6) & (\mathbf{40},7) & (\mathbf{40},8) & (\mathbf{40},9) & (\mathbf{40},10) & (\mathbf{40},11) \\ (\mathbf{45},2) & (\mathbf{45},3) & (\mathbf{45},4) & (\mathbf{45},5) & (\mathbf{45},6) & (\mathbf{45},7) & (\mathbf{45},8) & (\mathbf{45},9) & (\mathbf{45},10) & (\mathbf{45},11) \\ (\mathbf{50},2) & (\mathbf{50},3) & (\mathbf{50},4) & (\mathbf{50},5) & (\mathbf{50},6) & (\mathbf{50},7) & (\mathbf{50},8) & (\mathbf{50},9) & (\mathbf{50},10) & (\mathbf{50},11) \end{bmatrix} \tag{3.26}$$

It would be very difficult to manually construct and set the 100 candidate parameter pairs using the Execution Mode of OxteoFlow, as users will need to set the values of parameters for 100 times before each run. However, it can be easily done by using the Optimisation Mode, in which users can input the 10 candidates for regularisation λ and warp resolution respectively, and the 100 candidate pairs shown in Matrix (3.26) will automatically be constructed and tested by OxteoFlow.

For each of the 20 subjects in the optimisation cohort, 100 tests using these candidate pairs need to be run to obtain similarity scores. This means a total of 2,000 preprocessing tests need to be performed by OxteoFlow, which takes a long time. After the 2,000 tests are

completed, the mean and SD of similarity scores of the 20 subjects are plotted in a 10×10 grid, which has the same shape as Matrix (3.26). The value between the grid points are interpolated as a heatmap with contour lines to show trend and encircle the grid points of highest scores. The resulting heatmaps are presented in Figs. 3.4 to 3.7 on pages 118–121, which are then used to select the best parameter pair from the 100 candidates, using the decision rule of similarity metrics in Section 3.3.6.3 on page 110.

3.4.2.2 Decision on Parameters

As can be seen in Fig. 3.4 on page 118, the desirable regions of higher mean metric scores, and lower SDs are encoded with light colour in both the mean and SD heatmaps. The points of highest values in both mean and SD heatmaps can easily be identified using the contour lines. In the mean CR heatmap, two points are shown to have the highest mean CR—$[\mathbf{10}, (7, 7, 7)]$ and $[\mathbf{30}, (9, 9, 9)]$. It can also be seen that these two points both have acceptable SDs in the light coloured area of the SD heatmap.

To make a decision between the two points, the heatmaps of NMI in Fig. 3.5 on page 119 must be examined. As can be seen, $[\mathbf{10}, (7, 7, 7)]$ lies in the highest score region in the mean NMI heatmap, while $[\mathbf{30}, (9, 9, 9)]$ is not in that region. It can also be seen that the SD of $[\mathbf{10}, (7, 7, 7)]$ is acceptable as it avoided the extreme SDs in the dark blue region, in the SD heatmap.

Using the decision rule in Section 3.3.6.3 on page 110, we can stop here and accept the candidate pair $[\mathbf{10}, (7, 7, 7)]$ as it is supported by both of the high priority metrics (CR and NMI). However, it would still be worth checking the CC and MI heatmaps to confirm our decision, although this is not a routine according to the decision rule. As can be seen in Fig. 3.6 on page 120 and Fig. 3.7 on page 121, $[\mathbf{10}, (7, 7, 7)]$ and $[\mathbf{30}, (9, 9, 9)]$ are also supported by the mean and SD heatmaps of CC and MI. That means without the NMI metric, it would not have been possible to make a decision between these two candidates using only CR, CC or MI. That said, NMI is not useful on its own, as the mean NMI

heatmap gives too many candidates (12 eligible grids) in the highest score region.

Overall, using the heatmaps and the decision rule, $[\mathbf{10}, (7, 7, 7)]$, i.e. (100, 50, 33, 16, 13, 10) for regularisation λ, and (7, 7, 7) for warp resolution are selected as the optimum parameters to use in OxteoFlow. By keeping these optimised parameters unchanged, and continue to perform grid search optimisation on other parameters using the same approach, all the parameters in OxteoFlow can be optimised. The complete list of all the optimised parameters are presented in Appendix 5.1 on page 274

3.4.3 Visual Inspection of Preprocessing Outputs

Although it is possible to calculate the similarity metrics scores between the final output image and the reference image, absolute interpretation of the scores is meaningless. One cannot conclude that a higher CR score close to 1.0 is better than 0.7, because the former means the inter-subject difference related to the disease that we want to analyse is removed by preprocessing. Therefore, the similarity scores are only useful when we can linked them to the quality of the preprocessing, and this can only be done by performing visual inspection on the output images. One way to perform visual inspection is to plot the output image and the reference image against each other in a checkerboard plot, where two images are plotted in every other small boxes, and any mismatch can be clearly shown.

As can be seen in Fig. 3.8 on page 122, a representative input knee image in the Optimisation Cohort and the reference knee image are plotted against each other, slice by slice (only 12 slices are shown). Significant mismatch can be observed in all the slices between these two scans, which is especially clear in the cartilage region. Fig. 3.9 on page 123 is the same comparison after the input image is preprocessed by the Linear Registration Layer. As can be seen, the gross structures in the two images are aligned after the linear registration, with only some subtle mismatch in the cartilage region. The final preprocessing output of the OxteoFlow is shown in Fig. 3.10 on page 124. In that

image, nearly all the mismatch between the output and the reference images are removed after the non-linear registration, especially in slices showing knee cartilage.

By performing similar visual inspections on other images in the Optimisation Cohort for knee, it was found OxteoFlow performed equally well in those scans (checkerboard plots not provided here). The overall mean similarity scores between the outputs in the Optimisation Cohort and the reference image are CR = 0.722, NMI = 0.161, CC = 0.712, and MI = 0.838 (shown in Fig. 3.11 on page 125).

The visual inspection process provides a way to link the similarity metrics with the quality of preprocessing outputs. One could expect any new similarity scores comparable to the mean scores of the Optimisation Cohort should also show satisfactory preprocessing quality if they perform visual inspection on their output image.

I have also performed similar visual inspections for hip dGEMRIC and hip T_2 images. The before and after preprocessing checkerboard plots for a representative hip dGEMRIC scan are shown in Figs. 3.14 and 3.15 on page 128 and on page 129. As can be seen, there are clear mismatches between the input image and the reference image before preprocessing, especially in the slices shown in the third row. OxteoFlow successfully removed these mismatches after preprocessing, which can best be shown by the nearly perfect circles formed by the two femoral heads—it is evident even in the third row where the most significant mismatch were observed.

Figs. 3.14 and 3.15 on page 128 and on page 129 are similar comparisons between a representative hip T_2 scan and the reference image. It is clear that the mismatch are also removed after preprocessed by OxteoFlow, despite the very noisy scan and poor image contrast.

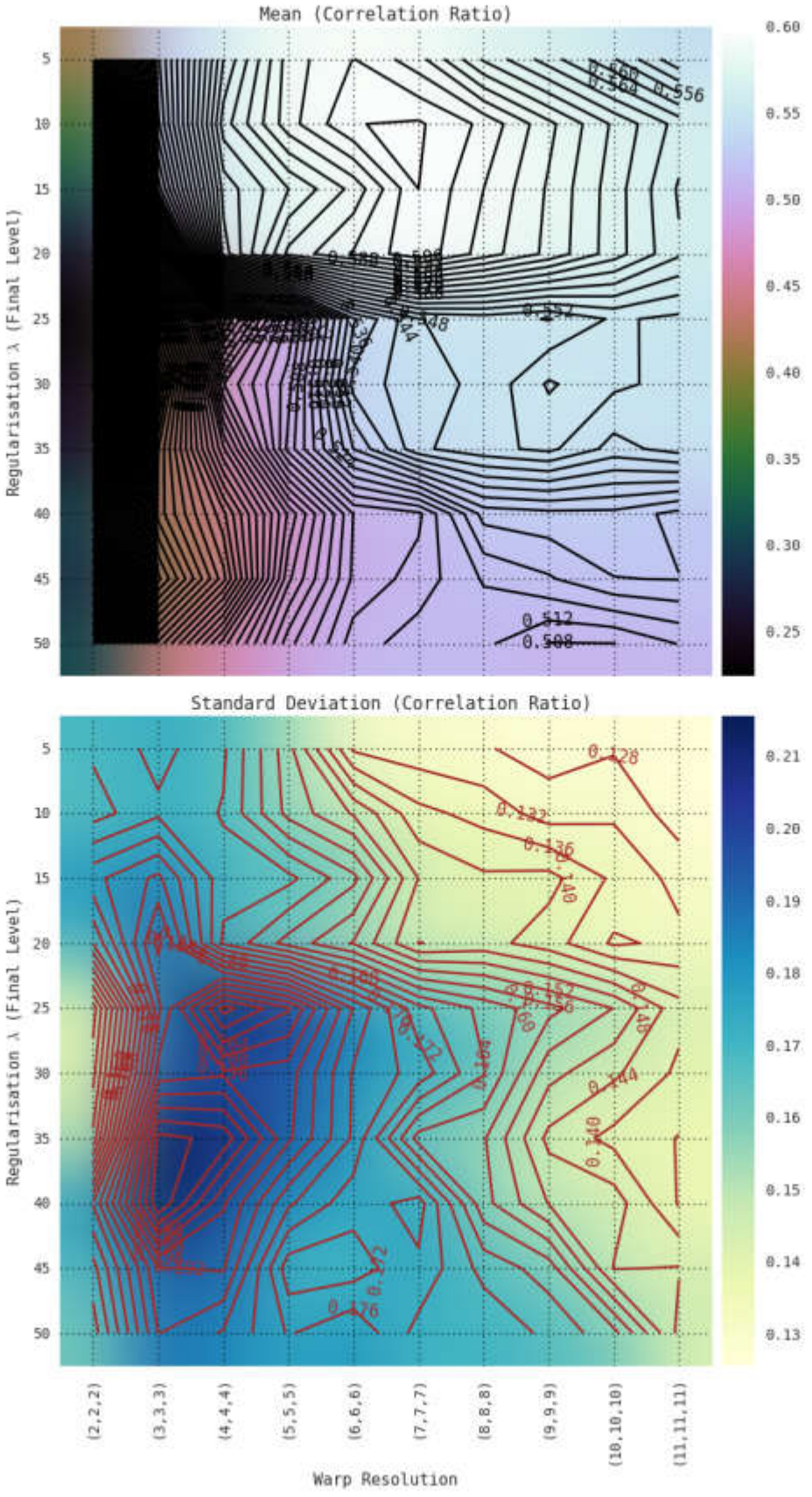


Figure 3.4: Grid Search Optimisation of Regularisation λ and Warp Resolution Guided by Correlation Ratio (CR, High Priority)

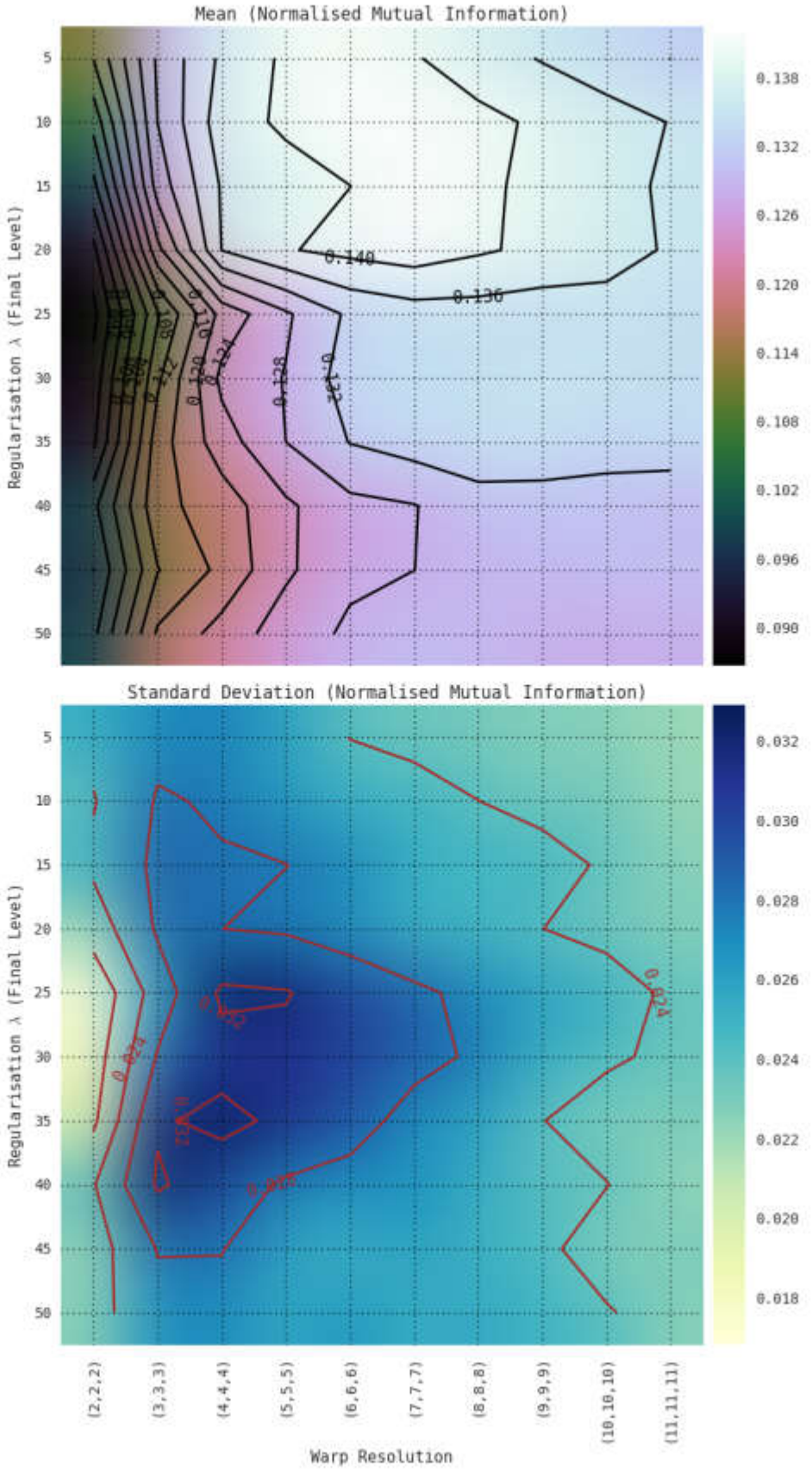


Figure 3.5: Grid Search Optimisation of Regularisation λ and Warp Resolution Guided by Normalised Mutual Information (NMI, High Priority)

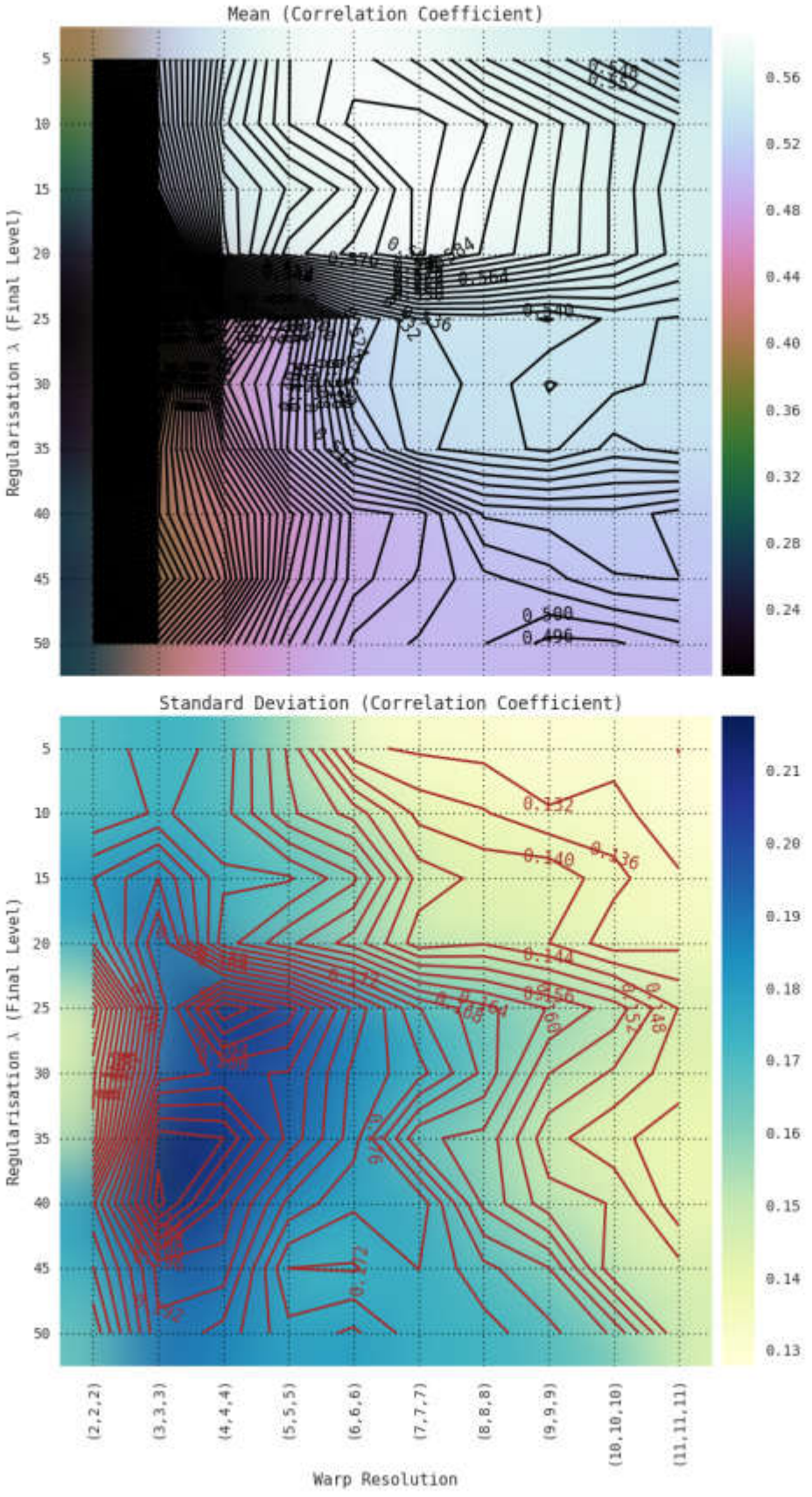


Figure 3.6: Grid Search Optimisation of Regularisation λ and Warp Resolution Guided by Correlation Coefficient (CC, Medium Priority)

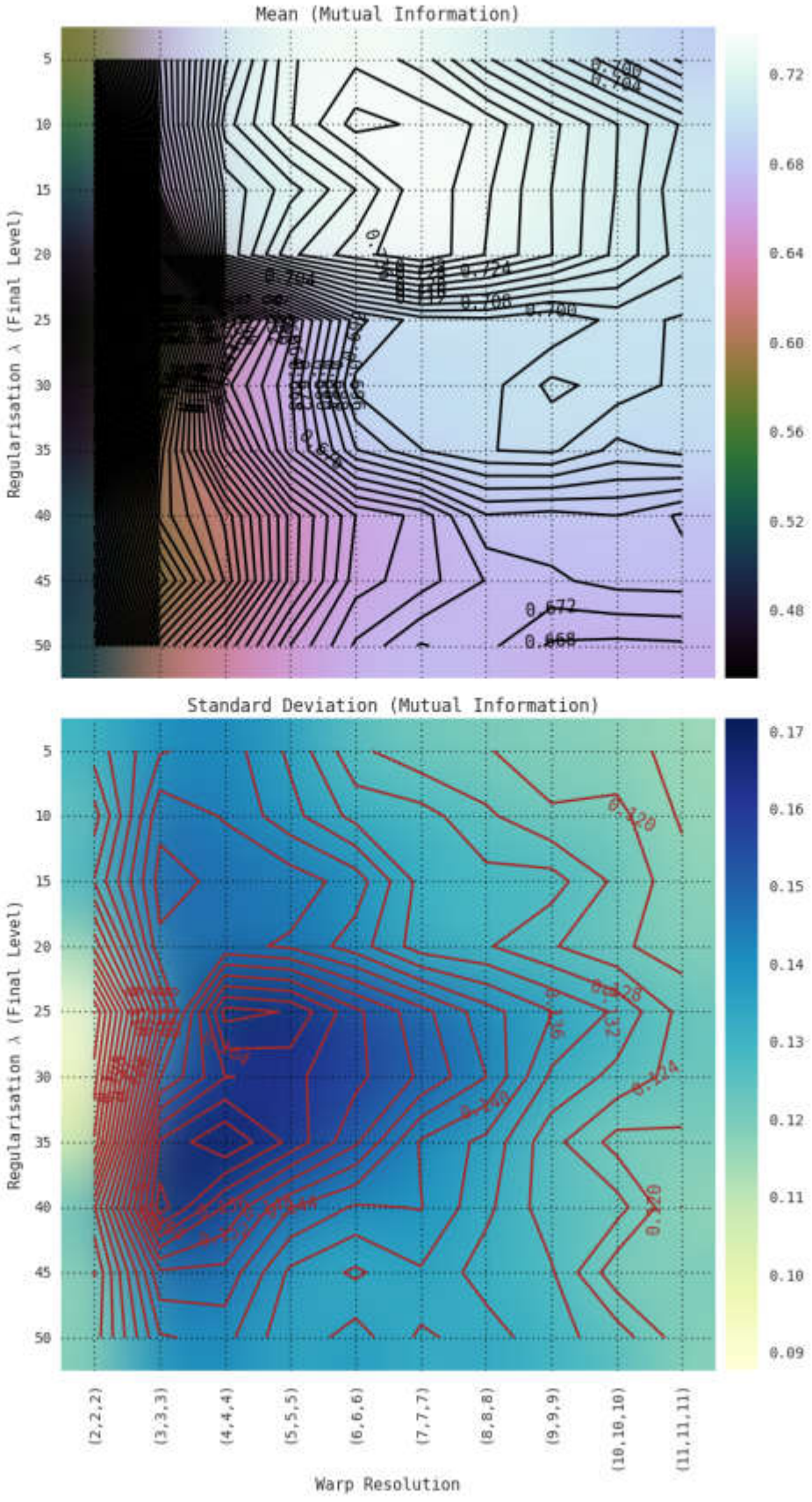


Figure 3.7: Grid Search Optimisation of Regularisation λ and Warp Resolution Guided by Mutual Information (MI, Low Priority)

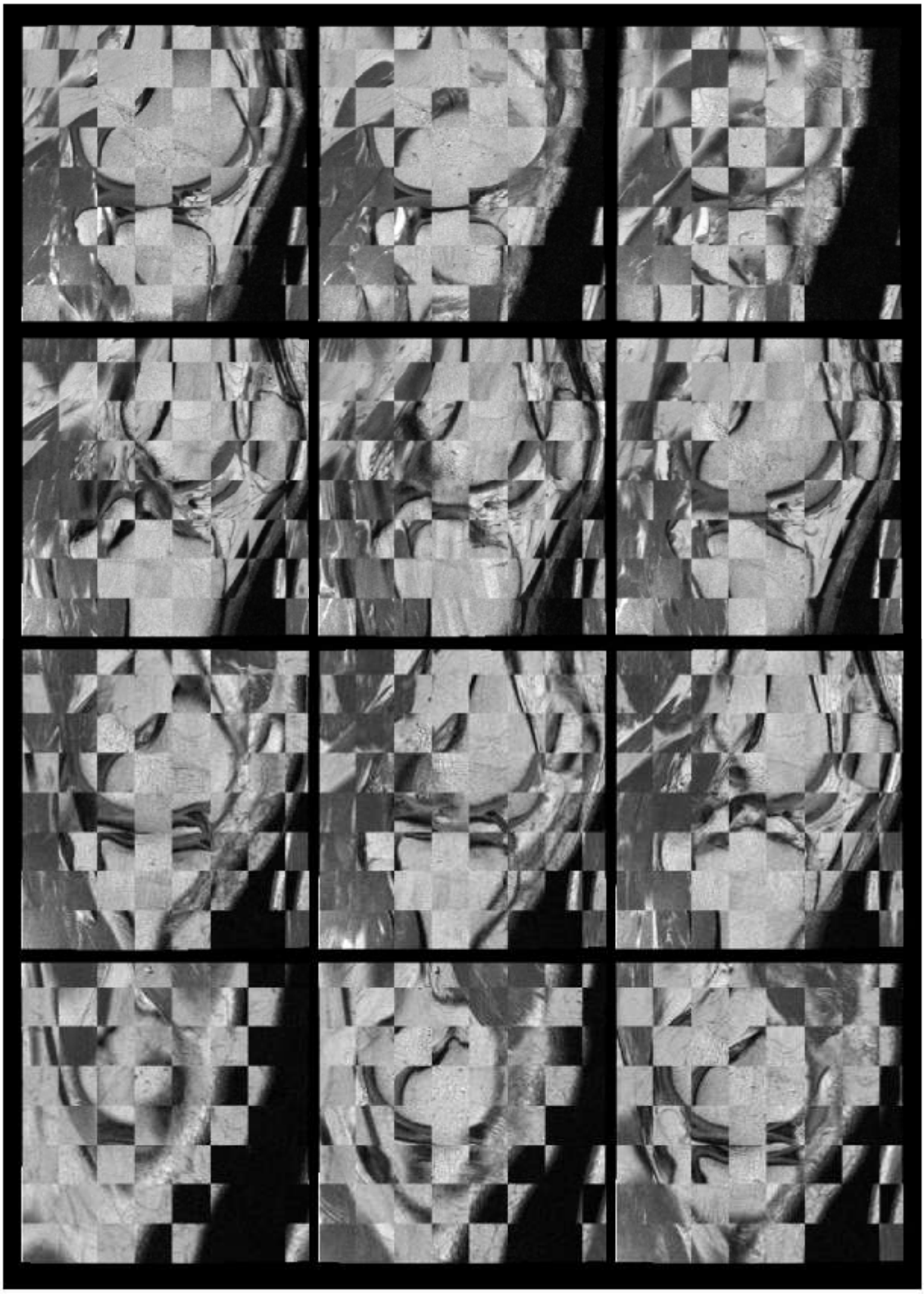


Figure 3.8: Slice-wise Checkerboard Plot of a Knee T_2 Image Against the Reference Image Before Preprocessing

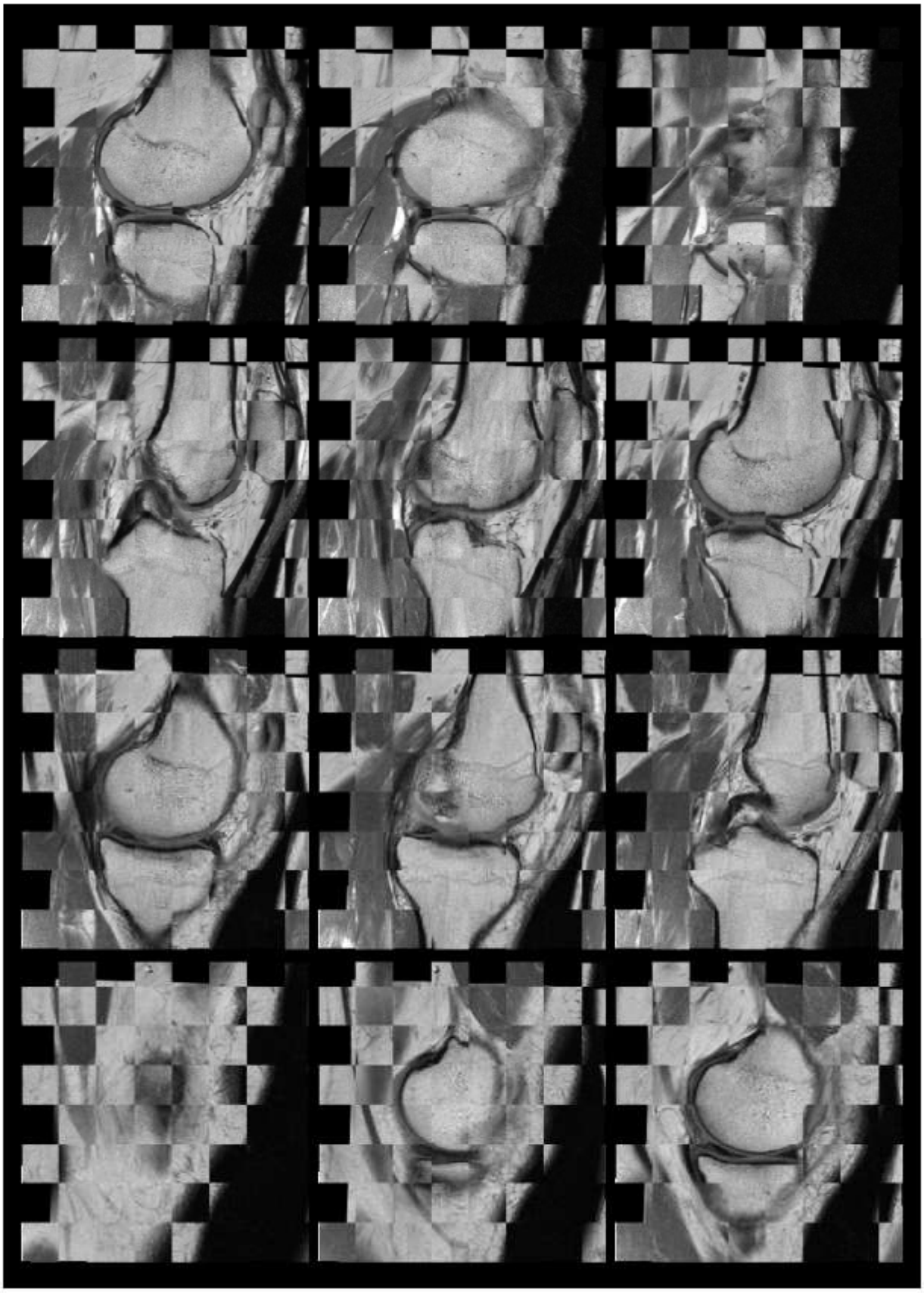


Figure 3.9: Slice-wise Checkerboard Plot of a Knee T_2 Image Against the Reference Image After Being Preprocessed by the Linear Registration Layer in OxteoFlow

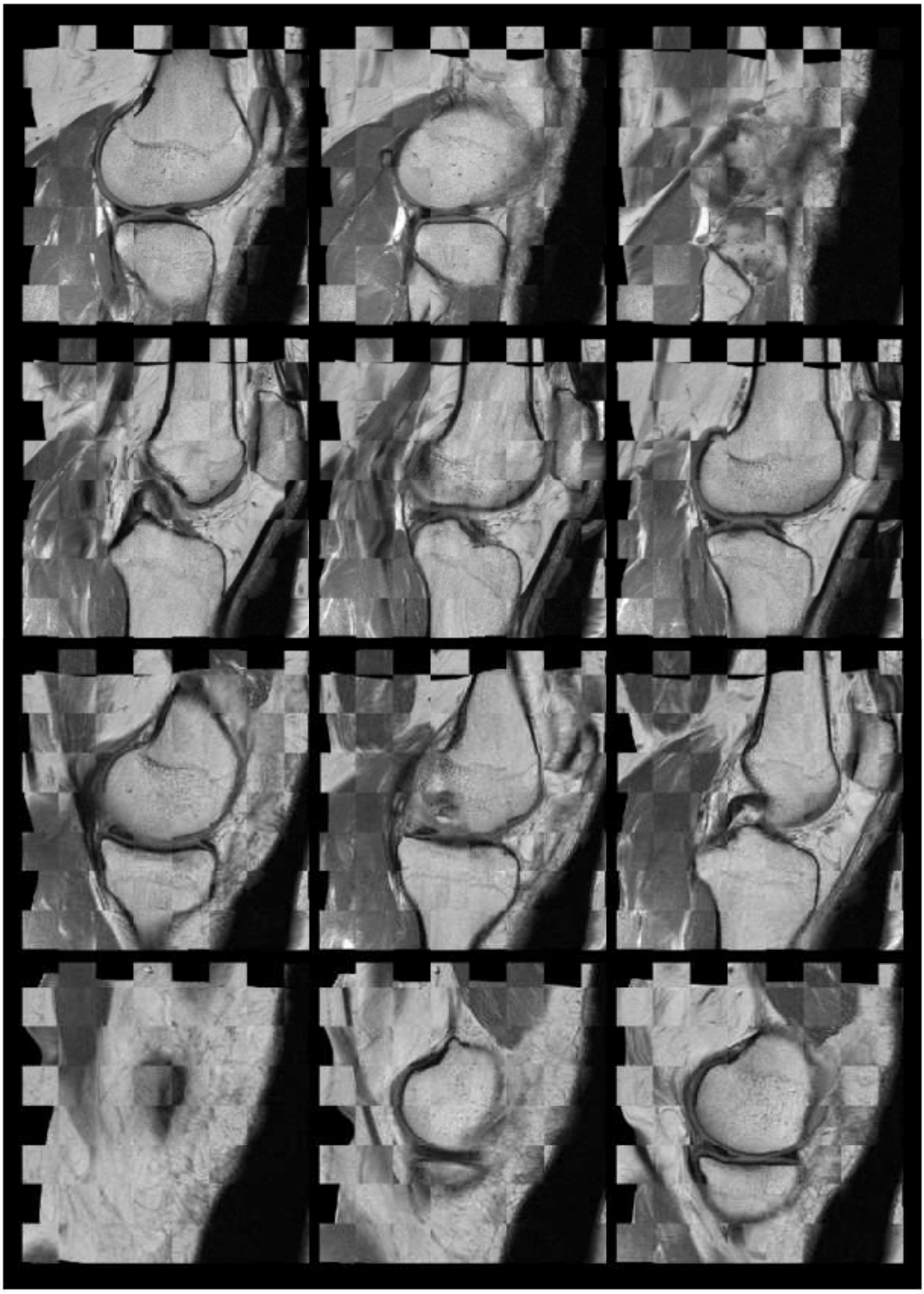
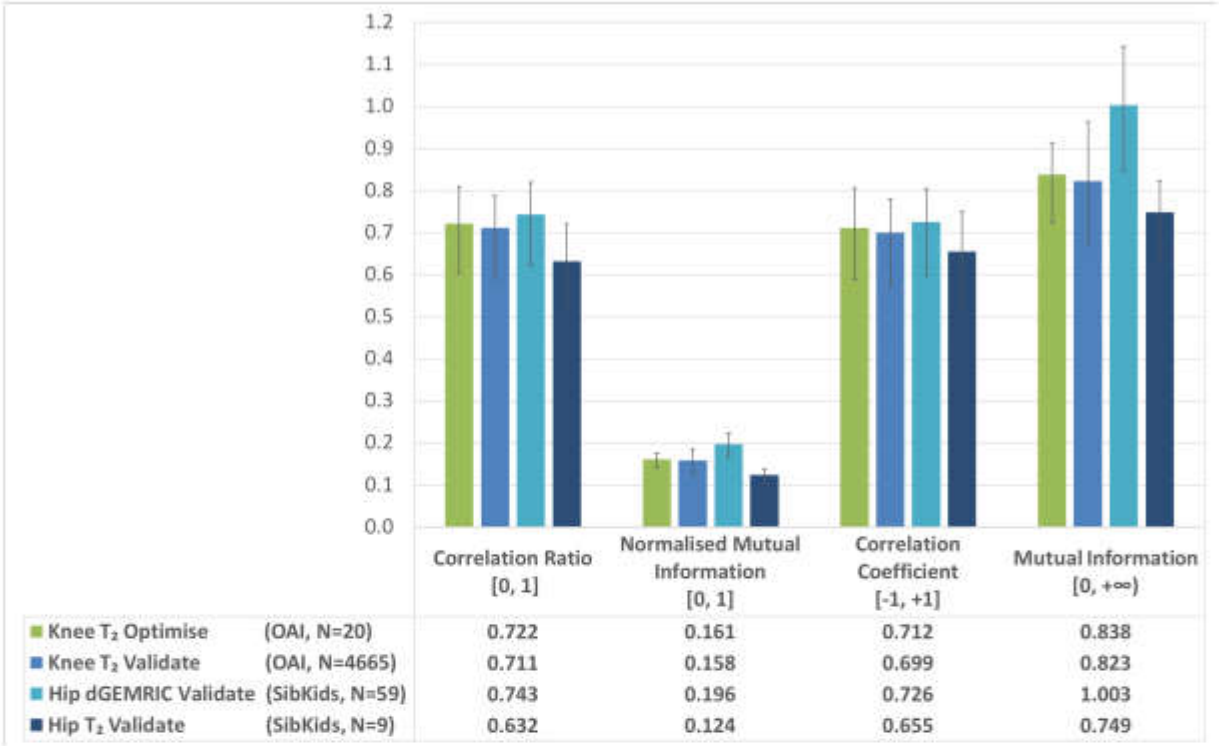


Figure 3.10: Slice-wise Checkerboard Plot of a Knee T_2 Image Against the Reference Image After Being Preprocessed by the Non-linear Registration Layer in OxteoFlow



		Correlation Ratio [0, 1]	Normalised Mutual Information [0, 1]	Correlation Coefficient [-1, +1]	Mutual Information [0, +∞)
Knee T_2 Optimise	(OAI, N=20)	0.722	0.161	0.712	0.838
Knee T_2 Validate	(OAI, N=4665)	0.711	0.158	0.699	0.823
Hip dGEMRIC Validate	(SibKids, N=59)	0.743	0.196	0.726	1.003
Hip T_2 Validate	(SibKids, N=9)	0.632	0.124	0.655	0.749

Figure 3.11: Mean Similarity Metrics Scores of the Optimisation and the Validation Studies

3.4.4 Validation

The mean similarity scores of running OxteoFlow on the complete image sets are shown in Fig. 3.11. As can be seen, OxteoFlow performed consistently well on the complete image set of OAI and dGEMRIC images in SibKids. The mean scores for the complete 'knee T_2 validate' set (N = 4,665) only decreased a very small value (≈ 0.1) in all the four metrics compared to the scores achieved for the 20 subjects in Optimisation Cohort. This means the optimised (hyper)parameters generalised very well in preprocessing unseen images in the OAI image set.

As I only performed a partial parameter optimisation for hip images, and used most of the parameters optimised for knee T_2 scans directly on hip dGEMRIC scans, it is surprising to see that the hip dGEMRIC images (N = 59) showed similar scores as the knee scores. This is a strong evidence that the optimised parameters are suitable for scans in different modality, ensuring OxteoFlow to achieve good preprocessing results on new data type even with very little optimisation.

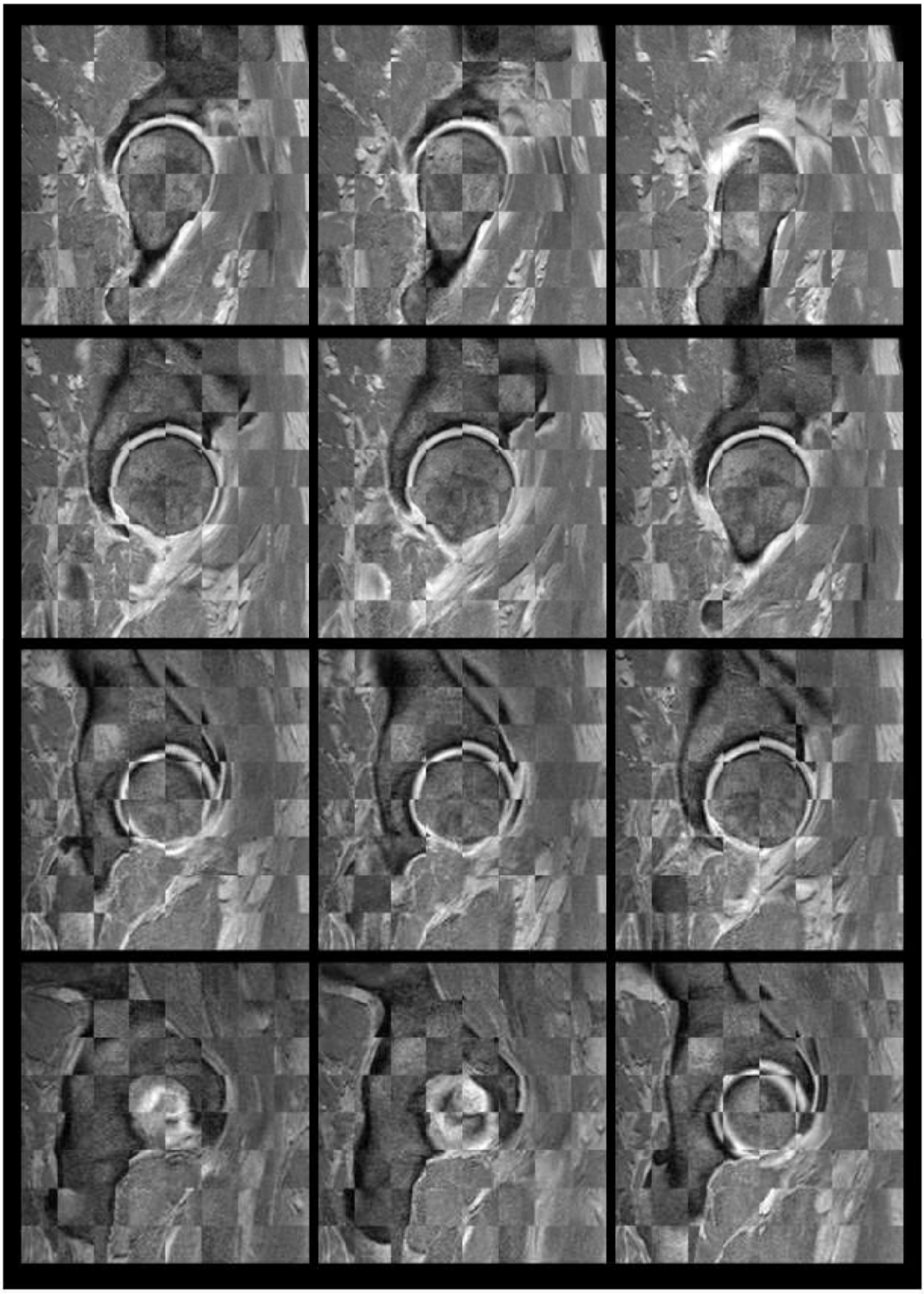


Figure 3.12: Slice-wise Checkerboard Plots of a Hip dGEMRIC Image Against the Reference Image Before Preprocessing

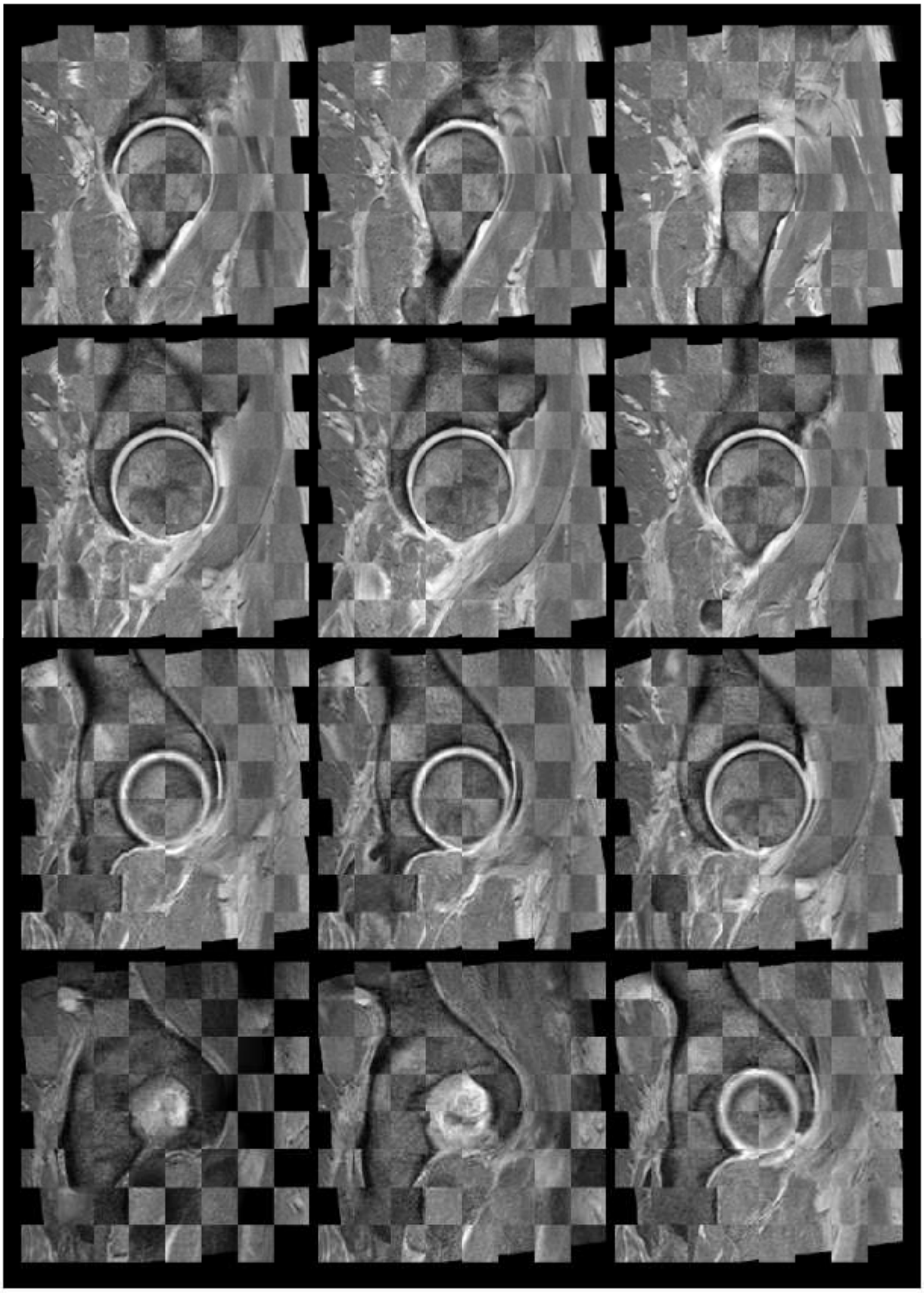


Figure 3.13: Slice-wise Checkerboard Plots of a Hip dGEMRIC Image Against the Reference Image After Being Preprocessed by OxteoFlow

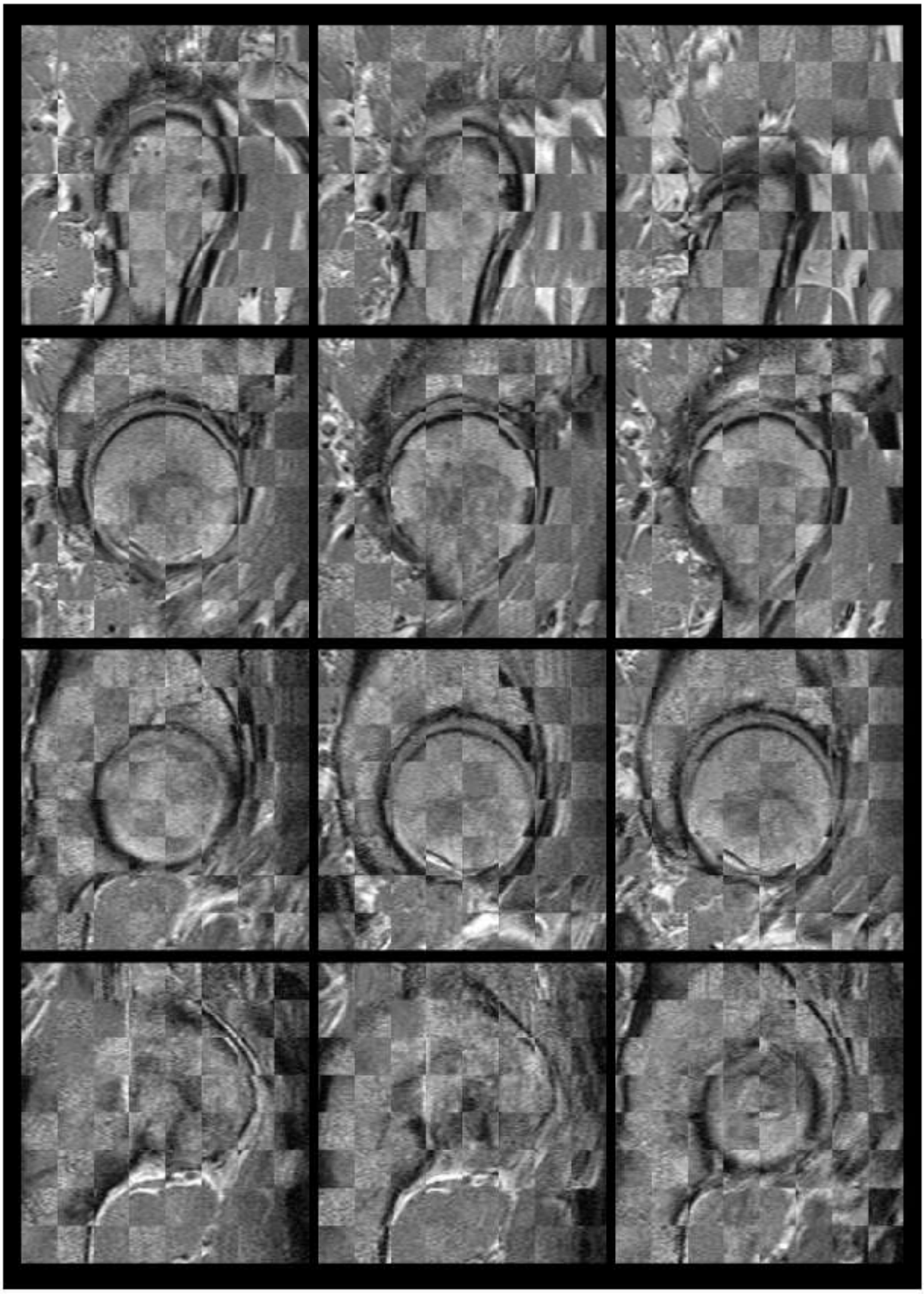


Figure 3.14: Slice-wise Checkerboard Plots of a Hip T_2 Image Against the Reference Image Before Preprocessing

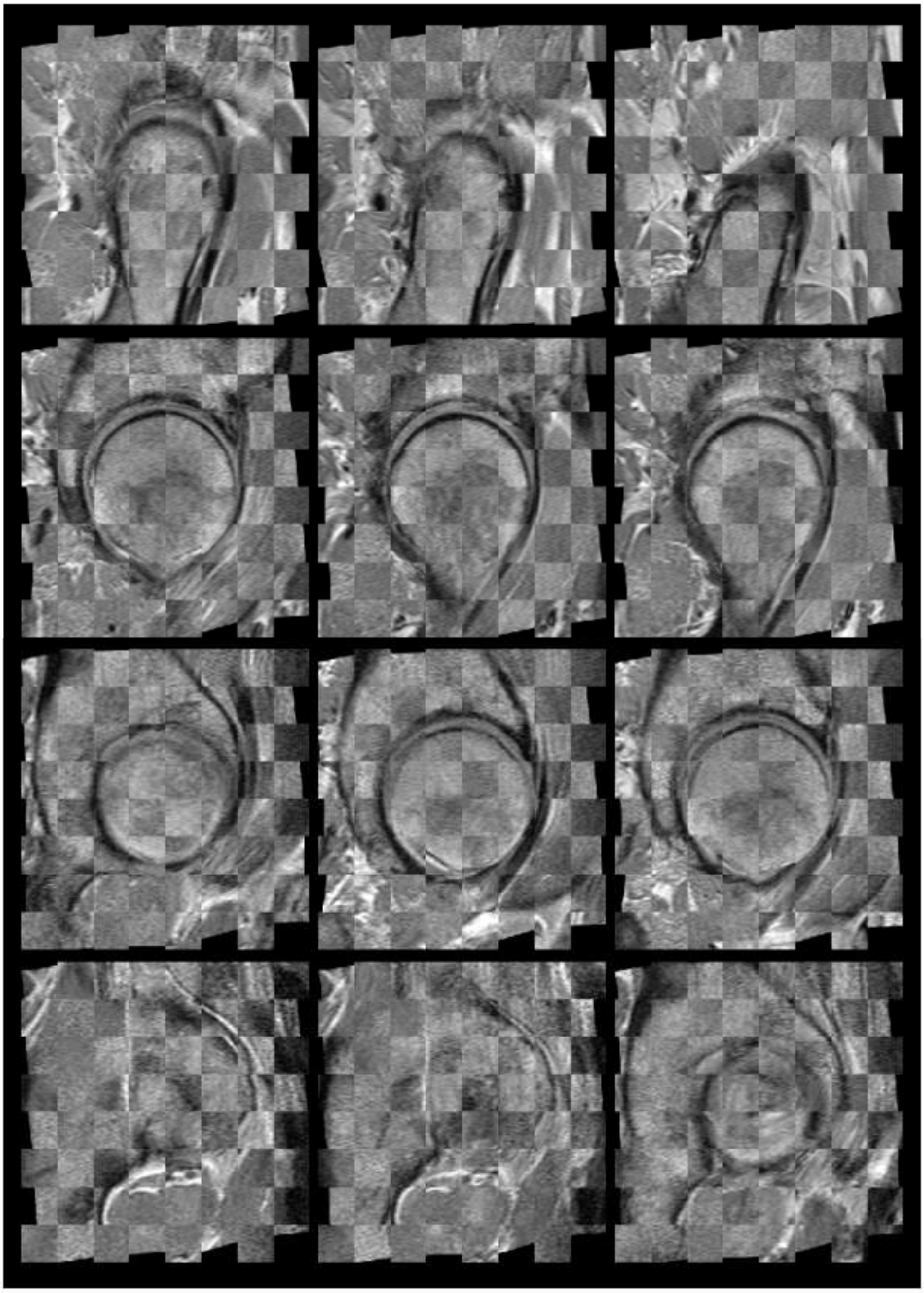


Figure 3.15: Slice-wise Checkerboard Plots of a Hip T_2 Image Against the Reference Image After Being Preprocessed by OxteoFlow

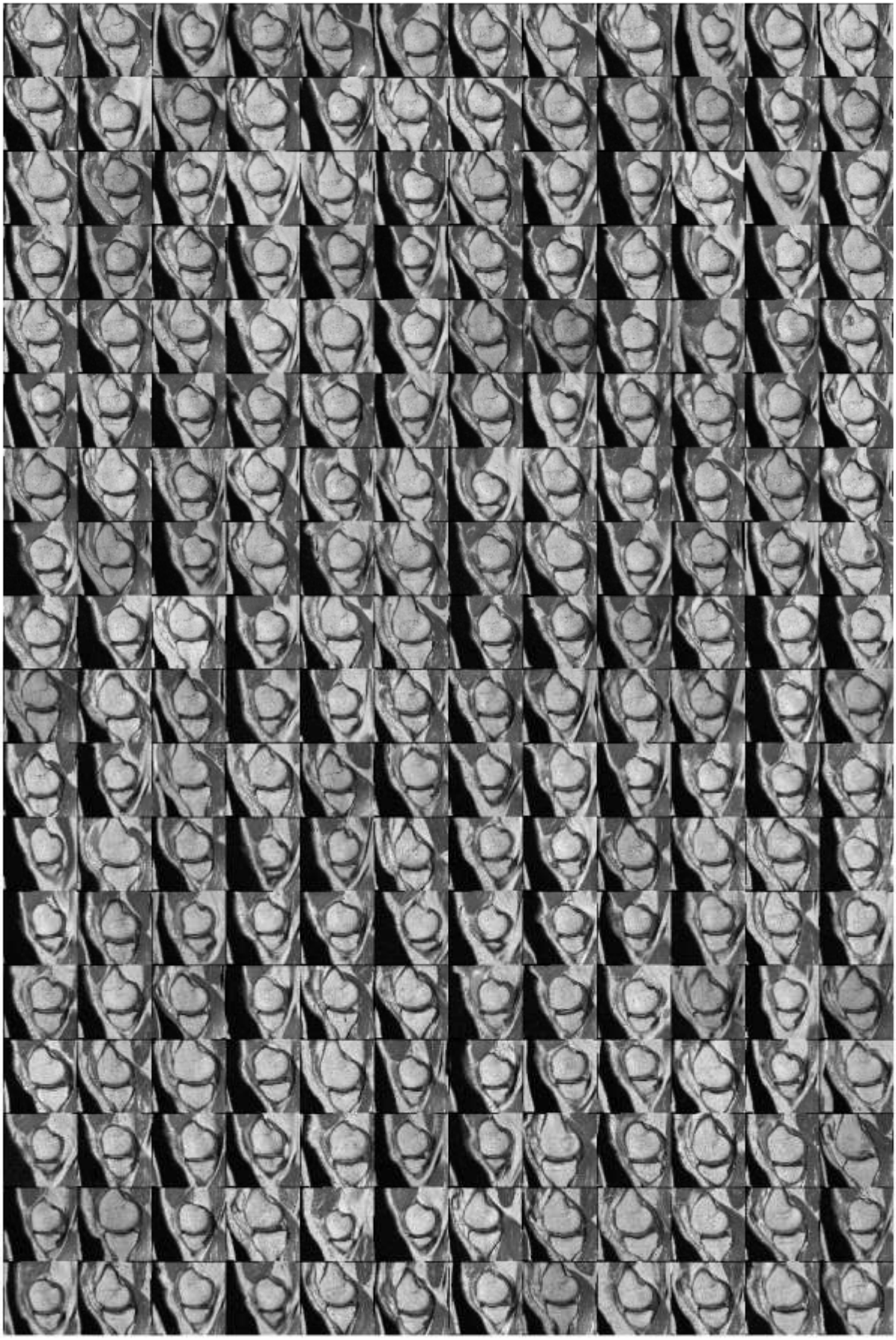


Figure 3.16: Random Selection of 216 (12×18) Knee T_2 images Before Preprocessing



Figure 3.17: Output of the 216 (12×18) Knee T_2 images After Being Processed by OxteoFlow

Although custom-built reference masks were applied on the hip T_2 scan, the similarity scores are significant lower than those of the knee T_2 and hip dGEMRIC. However, visual inspection confirmed that the preprocessing quality of hip T_2 scans were satisfactory, although in some extremely noisy images the OxteoFlow did produce a slightly mismatch result. These lower scores should therefore be interpreted as the result of very low SNR, and very limited sample size in the SibKid hip T_2 set. Again, I must stress that an absolute interpretation of similarity metrics is meaningless, and the preprocessing quality confirmed by visual inspection of output images are always more reliable than the scores.

Finally, as the preprocessed knee T_2 images in the OAI set will be used to perform statistical analysis in the next chapter, I performed a final visual inspection by playing all the preprocessed image as a movie. In this way, 12 unsatisfactory scans, caused by extremely poor original image quality, were removed. To showcase the preprocessing power of OxteoFlow, I randomly selected 216 knees (to fit in a page) from the remaining 4,654 images, and plotted the before and after preprocessing images of the 216 knee in Figs. 3.16 and 3.17 on page 130 and on the previous page respectively. The figures provided a clear overview of the quality of the outputs preprocessed by OxteoFlow. As can be seen, all the knee images originally different in shape (in Fig. 3.16) are all normalised into the same reference space (in Fig. 3.17).

Having developed, optimised, and validated OxteoFlow, and produced satisfactory preprocessing results with it in this chapter, it is now possible to perform the advanced statistical analysis on knee T_2 images using the methods translated from fMRI analysis, as shown in ?? on page ii. In the next chapter, I will discuss some significant applications (case-studies) based on OxteoFlow, including voxel-based analysis, machine learning and deep learning of musculoskeletal MRI.

Chapter 4

T_2 Patterns Associated with OA Manifestation and Progression: Voxel-based Statistical Analysis, Machine Learning, and Development of OxteoNet

4.1 Introduction

Following the development and testing of OxteoFlow in the previous chapter, it is now possible to use it to explore patterns of T_2 associated with the manifestation and progression of disease. The use of OxteoFlow allows the assessment of patterns at the single voxel level, removing bias from manual selection of regions of interest [73] , and providing a substantially richer data set. Only one studies, to our best knowledge, has performed such an analysis for musculoskeletal, but for $T_1\rho$ [138].

Herein we examine the patterns of T_2 with two main goals. First, we attempt to understand the relationship between these patterns and the manifestation and progression of disease. Second, we attempt to use these patterns to predict disease progression. For

both, we utilise the OAI data set—the largest suitable collection of T_2 maps available to the study.

4.2 Voxel-Based Statistical Analysis of the Knee T_2 Maps: Data From the Osteoarthritis Initiative

4.2.1 Methods

4.2.1.1 Hardware and Image Acquisition

Three custom built workstations, named 'xps', 'horla' and 'lab', were used simultaneously for all the computational intensive tasks, including image format conversion, T_2 mapping calculation, and the image preprocessing in this study. The relevant specifications of the three workstations are listed in Table 4.1.

Table 4.1: Hardware Specifications of the Workstations used for Preprocessing

Hardware	Workstations		
	'xps'	'horla'	'lab'
CPU	Intel Core i7-870	Intel Xeon E5-2687W	AMD FX-8320
Number of CPU Cores	4	8	8
Number of Threads	8	16	8
CPU Frequency	2.93 GHz	3.10 GHz	3.50 GHz
Memory Size	16GB	32GB	32GB

Images were obtained from OAI according to standard procedure, and under the OAI Data User Agreement for Research Image Set(s). Original data were acquired and stored on external drives (3 x 4TB, Model STEA4000400, Seagate, California, USA). Additional internal hard drives (2 x 8TB, Model ST8000VN0022, Seagate, California, USA) were used to store the NIfTI images after format conversion, as well as all the intermediate

and final outputs generated by the image processing workflow.

OAI Imaging Data Set

The OAI image set contains imaging data of all 4,796 participants from baseline (0 month) to 96 month. In this study, we use the data at baseline to explore the power of T_2 mapping for future outcomes (incidence and progression). *Unless otherwise specified, the OAI imaging data being processed and analysed in the rest of this thesis are data at the baseline.*

4.2.1.2 Image Format Conversion

To make the OAI image set compatible to the preprocessing workflow, the entire image set was converted from the DICOM format to the NIfTI format. As this task takes extremely long time to complete, a python script was coded to enable parallel conversion (utilising dcm2niiX [139] version 1.0.20170207 [140], and Python API provided by Nipype [136]) to make use of all CPU cores in the three workstations. The dcm2niiX command line options [141] used in this study is listed in Table 4.2. The resulting images have the same voxel values as in the DICOM format, but are much smaller in file size and are compatible to OxteoFlow.

Table 4.2: dcm2niiX Options Used to Convert the OAI Image Set to the NIfTI Format

Nipype Flag	Option Used	Function
out_file_name (-f)	%t_%c_%s_%p_%e	Output filename is formatted as *Time_Comments_Series_Protocol_Echo*
compress (-z)	y	Compress output images as .gz format
has_private (-t)	y	Includes private patient details in the text notes
crop (-x)	False	Do not crop 3D acquisitions
bids_format (-b)	False	Do not create BIDS sidecar file
merge_imgs (-m)	False	Do not merge 2D slices from same series

4.2.1.3 Data Cleaning

Data cleaning was required before preprocessing of the image set due to non-standard scans, mislabelling, repeated scans, and incorrect numbers of echoes.

Sagittal T_2 scans were typically acquired for the right knees, but some exceptions exist where left knees were scanned. As OAI does not specify the participant ID of these exceptions in either release notes, manuals or image assessment data, image metadata was used to identify the right knees from the left. However, due to the large number of images acquired in the OAI study, the metadata of many scans have the side information incorrectly or inconsistently input by the radiologist at the time of the scans. For example, 'RIGHT' was misspelled as 'RIGHRT', 'RIFHT' or even left blank, or its variation 'RT' used across the image set. These metadata were then used by dcm2niiX programme in the filenames (the *_Comments* part, see Table 4.2 on the preceding page) of the converted NIfTI images. These filenames must be corrected for the images to be captured by the T_2 mapping calculation script and OxteoFlow as a valid input, and to correctly count the number of eligible T_2 scans.

All incorrect/inconsistent filenames caused by wrong metadata input were corrected, making it possible to count the number of right and left scans with a simple filename matching. Of all the 4,796 participants, $n = 4,681$ participants with T_2 scans of right knee were identified, $n = 36$ had no T_2 scans on either side, and $n = 79$ had T_2 scans only on left knees. As the number of T_2 scans of left knees ($n = 79$) is small compared to that of right knees ($n = 4,681$), *only T_2 scans of right knees were included for analysis in this study.*

Furthermore, some participants in the OAI have multiple T_2 scans acquired in a single scanning session. To avoid over-counting the number of eligible scans, only one T_2 scan was kept for each participant. As OAI does not explain the reason of repeating the T_2 scans, it was assumed that scans were repeated until satisfactory, and therefore the final scans kept for analysis. Among the $n = 4,681$ participants with right knee scans, $n = 30$ were identified to have repeated scans and had their initial scans removed.

Finally, the OAI T_2 scans were acquired using a seven-echo sequence (every 10ms) with 27 slices in the 120mm FOV. After filename correction and removal of initial scans, the T_2 scan of each participant was checked to contain exactly seven NIfTI files for the seven echoes, named in the form of *Time_Side_Series_Protocol_Echo*. For example, the T_2 images for participant ID 9000296 are:

TE1: '20040909112433_RIGHT_8_Sagittal_T_2_Map_120mmFOV_**0**.nii.gz'

TE2: '20040909112433_RIGHT_8_Sagittal_T_2_Map_120mmFOV_**1**.nii.gz'

TE3: '20040909112433_RIGHT_8_Sagittal_T_2_Map_120mmFOV_**2**.nii.gz'

⋮

TE7: '20040909112433_RIGHT_8_Sagittal_T_2_Map_120mmFOV_**6**.nii.gz'

To be used in this study, each NIfTI image must have 27 slices in the sagittal view, typically with a 384×384 resolution in any slice. T_2 scans of 7 participants with incorrect number of echoes[1] and 8 participants with incorrect number of slices[2] were also identified and removed.

As a result, $n = 4,666$ participants with T_2 scans of right knee are eligible to proceed to the T_2 Mapping step after the data cleaning.

4.2.1.4 T_2 Mapping Calculations and Preprocessing

The predictors in this study are T_2 scans at the baseline, however, OAI only provided ready-to-use T_2 maps for a limited number of subjects in the 72-month follow-up visit only, calculated from the T_2 scans using the Siemens syngo MapIt software. Therefore, I performed T_2 mapping calculation independently for all baseline scans of the 4,666 eligible subjects, using the LLS, NLS, NCECP, SQEXP methods detailed in Chapter 2.

[1]Subject IDs whose number of echoes $\neq$ 7: 9445104, 9733288, 9952664, 9641467, 9545921, 9594551, 9176992

[2]Subject IDs whose number of slices $\neq$ 27: 9283903, 9768219, 9952817, 9560374, 9481413, 9777471, 9006140, 9794763—these information could save days of works for future researchers using the OAI image set.

T_2 maps are then preprocessed (warped into reference space) using OxteoFlow, with only FSL branch activated. ANTs Non-linear Registration Layer are not used in this project due to the large number of images and the considerably slower execution speed of ANTs compared to that of FNIRT. A detailed OxteoFlow architecture used in this project is presented in Fig. 4.1 on the next page.

The workflow from image conversion to preprocessing is summarised in Fig. 4.2 on page 140.

4.2.1.5 Definitions of Incident and Progression of Radiographic OA and Pain

To establish the predictive validity of T_2 mapping to structural (radiographic) and symptomatic (pain) OA outcomes, the study design of FNIH OA Biomarkers Consortium Project [142] is followed, in which the imaging data are stratified into one of the four groups, as shown in Section 4.2.1.6 on page 145. However, one disadvantage of the the FNIH project is that it only followed radiographic OA (ROA) and pain progression of a small subgroup in the OAI image set, without considering the predictive power of T_2 for incident (i.e. onset of) ROA and pain. This study used both indent and progression information of ROA and pain provided by the OAI to stratify the imaging data, which provides a much more complete, and comprehensive review of the predictive validity of T_2 mapping.

This section focuses on discussing the definitions of incident ROA and pain, and ROA and pain progression, which are summarised in Table 4.3.

Table 4.3: Summary of Definitions of Incidence and Progression of ROA and Pain

	Incident	Progression
ROA	Development of KLG $\geqslant 2$ from a knee that was KLG 0 or 1 at baseline regardless of JSN	OARSI JSN score worsened by a partial grade since baseline
Pain	Increase of $\geqslant$ 9 points that was $\leqslant$ 10 at baseline on a 0–100 normalised WOMAC pain subscale score	Persistent increase of $\geqslant 9$ points at two or more times since a visit on a 0–100 normalised WOMAC pain subscale score

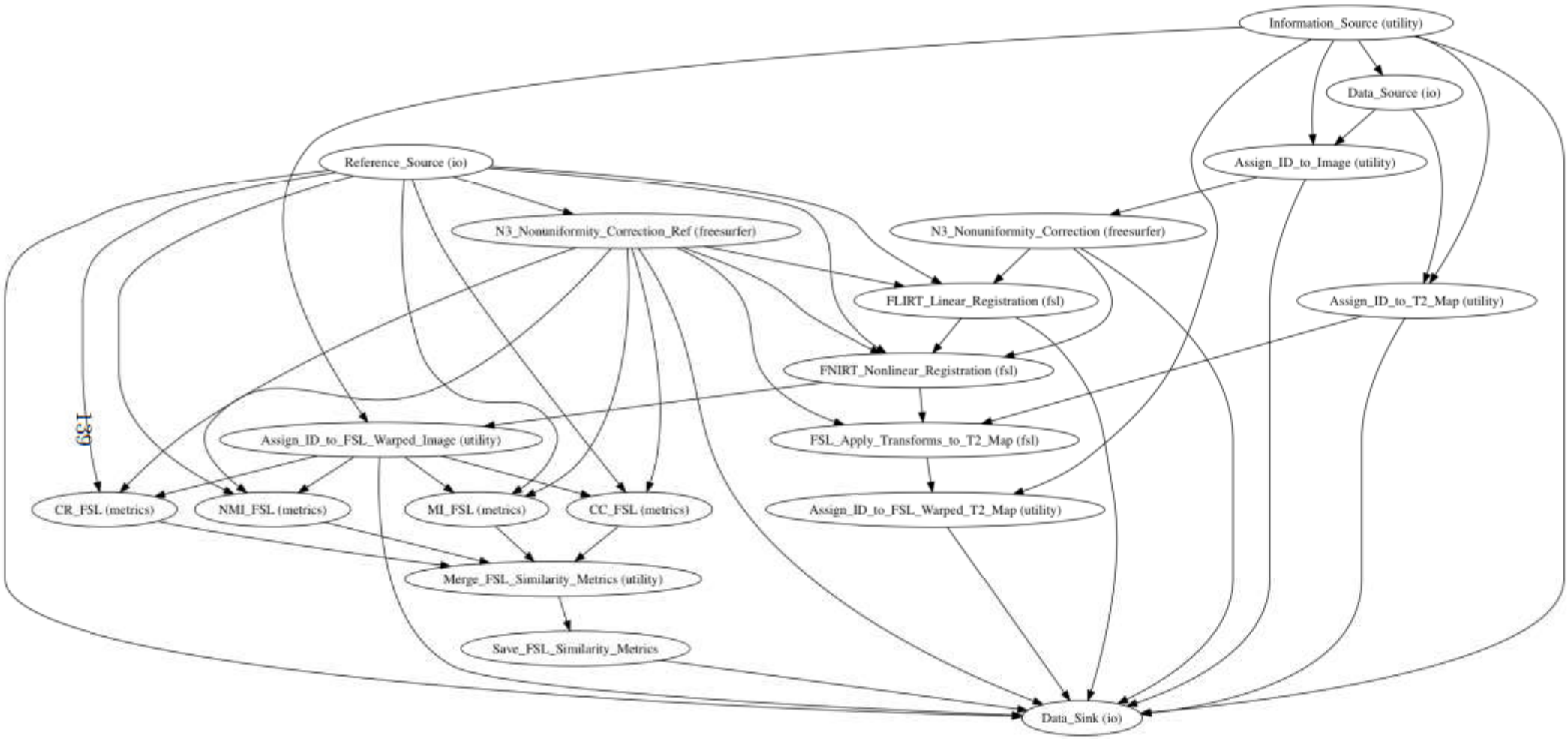


Figure 4.1: Architecture of OxteoFlow Used for Preprocessing the T_2 Maps of OAI Imaging Set (Under Execution Mode, With Only the FSL Branch Activated)

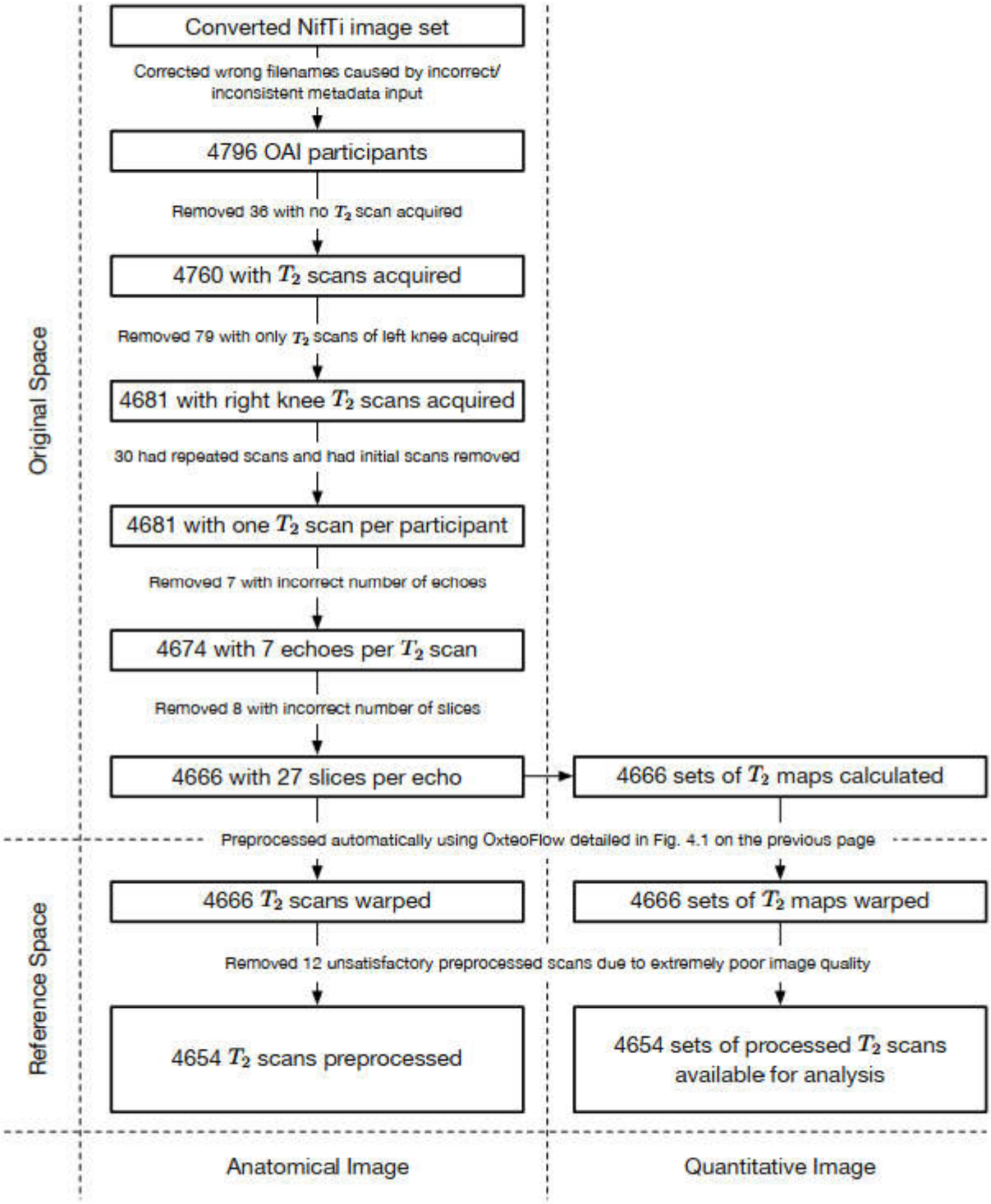


Figure 4.2: Workflow of OAI Imaging Set Before Ready for Analysis

Detailed explanations of how these definitions are used to stratify the OAI data set into the four incidence groups ('Both', 'ROA', 'Pain', and 'None') and the four progression groups (also 'Both', 'ROA', 'Pain', and 'None'), are discussed later in Section 4.2.1.6 on page 145.

Definition of Incident Radiographic OA

In this study, definition [143] recommended in the OAI SAS data set documentation [144] was used to define incident ROA of the knee. The OAI SAS data set OUTCOMES99 [145] has variables that distinguish between two different definitions of incident ROA:

(1) Development of Kellgren & Lawrence Grade (KLG) $\geqslant 2$ from a knee that was KLG 0 or 1 at baseline, regardless of presence/absence of any joint space narrowing (JSN) in either medial or lateral compartments when incidence occurs (OARSI JSN can be 0 at the visit of incidence);

(2) Development of KLG $\geqslant 2$ from a knee that was KLG 0 or 1 at baseline, where the knee has both osteophytes AND an OARSI JSN score $\geqslant 1$ in either medial of lateral compartments at the time point when incidence occurs.

Table 4.4: Frequency Table of Incident Radiographic OA by the KLG

V99ERXIOA	Meaning	Frequency
.M or .A	Missing Data	290
1	KLG $\geqslant 2$ at baseline	2032
2	KLG = 0–1 at baseline, no incident KLG $\geqslant 2$ in follow-up	2103
3	KLG = 0–1 at baseline, incident KLG $\geqslant 2$ at 12 month	97
4	KLG = 0–1 at baseline, incident KLG $\geqslant 2$ at 24 month	45
5	KLG = 0–1 at baseline, incident KLG $\geqslant 2$ at 36 month	57
6	KLG = 0–1 at baseline, incident KLG $\geqslant 2$ at 48 month	40
7	KLG = 0–1 at baseline, incident KLG $\geqslant 2$ at 72 month	84
8	KLG = 0–1 at baseline, incident KLG $\geqslant 2$ at 96 month	48
		4796

As all knees that meet the definition (2) at a specific visit must also meet definition (1) at that visit, or have met definition (1) at an earlier visit if the presence of an osteophyte

preceded development of JSN, the definition (1) is considered a better match of the purpose of this study. That is, to detect/predict disease at the earliest possible time. In addition, using definition (1) will include more subjects into the incidence cohort than the definition (2), which provides additional benefits to balance the incidence and non-incidence cohort. A maximised and balanced data set will be critical for the machine learning study in Part II of this chapter (Section 4.3). Therefore, the definition (1) of incident ROA is used in this study.

In the OAI SAS data set OUTCOMES99, the variable for the traditional definition of incident ROA in the right knee is **V99ERXIOA**, which marks the first visit at which KLG $\geqslant$ 2 regardless of OARSI JSN score (V99ERXIOA = 3–8), or the knee remained KLG 0 or 1 with no incident in the follow-up period (V99ERXIOA = 2), or the knee was already KLG $\geqslant$ 2 at baseline (V99ERXIOA = 1). The detailed meaning and frequency of V99ERXIOA are summarised in Table 4.4 on the preceding page, where subjects with V99ERXIOA = 3–8 (below the dividing line) are those who were radiographically normal at baseline, but developed ROA at some point over the follow-up time of the OAI study.

Definition of Incident Pain

In the OAI study, knee pain is assessed using the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain subscale, one of the three WOMAC subscales of pain, stiffness and disability [146]. The pain subscale ask the participant to describe the pain they experienced with a 0–4 score during five activities: walking, using stairs, in bed, sitting or lie down, and standing, which totals a WOMAC pain subscale possible score range of 0–20.

Incident knee pain is defined as an increase of $\geqslant$ 9 points that was $\leqslant$ 10 at baseline on a 0–100 normalised WOMAC pain subscale score, based on a clinical definition of onset of patient-reported knee pain using WOMAC pain subscale [147].

In the OAI SAS data set, the calculated total WOMAC pain subscale score is stored with

Table 4.5: Frequency Table of Incident Pain by the WOMAC pain subscale

WOMKPR_INCI	Meaning	Frequency
X	Normalised WOMAC pain score $\geqslant 10$ at baseline	1728
0	Normalised WOMAC pain score $\leqslant 10$ at baseline, no increase of $\geqslant 9$ points in follow-up	1318
1	Normalised WOMAC pain score $\leqslant 10$ at baseline, increase of $\geqslant 9$ points in follow-up	1750
		4796

the **VXXWOMKPR** variables in the JointSxXX (Joint Symptoms/Function) SAS Data Sets [148] for each visit, where XX represents the month. For example, V00WOMKPR for 00 month (baseline) can be found in the JointSx00 data set, and V01WOMKPR for 12 month follow-up in the JointSx01 data set, and so on. The VXXWOMKPR data in the separated JointSxXX data sets are collected and processed in a single spreadsheet and the result is saved in a synthetic variable **WOMKPR_INCI**. The number of subjects who had no pain at baseline, but developed knee pain at some point over the follow-up time (below the dividing line) is shown in Table 4.5.

Definition of Progression of Radiographic OA

Instead of using KLG, the recommended definition [143; 149] uses the OARSI JSN score [150; 151] to assess the progression of the ROA [144]. The OAI SAS data set OUTCOMES99 [145] has variables that distinguish between two different definitions of progression of ROA:

(1) OARSI JSN score worsened by a whole grade or more since baseline;

(2) OARSI JSN score worsened by a partial grade or more since baseline.

Definition (2) of the progression of ROA is chosen to be used in this thesis so that an earlier disease of 'within-grade' change could also be considered as progression. This will also maximise the ROA progress cohort and benefit the machine learning study in Part II of this chapter (Section 4.3).

Table 4.6: Frequency Table of Progression Radiographic OA by the OARSI JSN score

V99ERJSFP	Meaning	Frequency
.A	Missing Data	483
0	JSN score remained the same since baseline, no progression	3390
1	JSN score worsened by a partial grade or more at 12 month	351
3	JSN score worsened by a partial grade or more at 24 month	163
5	JSN score worsened by a partial grade or more at 36 month	120
6	JSN score worsened by a partial grade or more at 48 month	117
8	JSN score worsened by a partial grade or more at 72 month	91
10	JSN score worsened by a partial grade or more at 96 month	81
		4796

In the OAI SAS data set OUTCOMES99, the variable for the definition (2) of progression of ROA in the right knee is **V99ERJSFP**, which indicates the first visit at which the OARSI JSN score of a knee was a partial grade or more larger than at baseline (V99ERJSFP = 3–10), or that a knee remained at the same OARSI JSN score at all visits (V99ERJSFP = 0). The detailed meaning and frequency of V99ERJSFP are summarised in Table 4.6, where subjects with V99ERJSFP = 3–10 (below the dividing line) are those who had ROA progression at some point over the follow-up time, compared to baseline.

Definition of Progression of Pain

Pain progression is defined as a *persistent* increase of $\geqslant 9$ points at two or more times since a visit on a 0–100 normalised WOMAC pain subscale score, based on a definition of minimum clinically important difference (MCID) for knee pain worsening [152; 153]. As pain persistence requires a normalised WOMAC score increase at two or more time points since a visit, knees are excluded if there are not enough follow-up time points after the first increase in WOMAC pain subscale score. The same method was also used in the FNIH OA Biomarkers Consortium Project [142] to define progression of knee pain.

The **VXXWOMKPR** data in the separated JointSxXX data sets are collected and processed in the same single spreadsheet used for incident pain, and the result is saved in a

Table 4.7: Frequency Table of Progression of Pain by the WOMAC pain subscale

WOMKPR_PROG	Meaning	Frequency
0	No persistent increase of $\geqslant$ 9 points at two or more times since a visit on a 0–100 normalised WOMAC pain subscale score	4185
1	Persistent increase of $\geqslant$ 9 points at two or more times since a visit on a 0–100 normalised WOMAC pain subscale score	611
		4796

synthetic variable **WOMKPR_PROG**. The number of subjects who had pain progression compared to baseline (below the dividing line) is shown in Table 4.7.

4.2.1.6 Stratifications of the Outcome Groups

Following the approach used in the FNIH OA Biomarkers Consortium Project [142], eligible subjects are classified into one of four groups based on the incidence and progression outcomes of ROA and pain.

Stratifications of the Incidence Outcome Groups

The four incidence outcomes of ROA and pain are:

(1) 'Both': Both incident ROA and pain;

(2) 'ROA': Incident ROA only;

(3) 'Pain': Incident pain only;

(4) 'None': Neither incident ROA nor pain.

Since the predictors are T_2 scans acquired at the baseline, and outcomes are defined by changes over the follow-up time (12-96 month), subjects who do not have T_2 scans at baseline, or who had already met the criteria for incident ROA (V99ERXIOA = 1) or incident pain (WOMKPR_INCI = X) at baseline are excluded. The stratification of the four incidence groups are illustrated in Fig. 4.3 on page 147.

Stratifications of the Progression Outcome Groups

Similarly, the four progression outcomes of ROA and pain are:

(1) 'Both': Both ROA and pain progression;

(2) 'ROA': ROA progression only;

(3) 'Pain': Pain progression only;

(4) 'None': Neither ROA nor pain progression.

The stratification of the four progression groups are illustrated in Fig. 4.3 on the next page.

As can be seen, there are significantly less subjects excluded in the progression outcome groups ($n = 569$) compared to that of the incidence outcome groups ($n = 3,011$). This is because all the subjects, regardless whether or not affected by ROA and/or pain at baseline, are followed up for JSN and pain progression, while a large number of subjects who already had incident ROA and/or pain are excluded from the incidence groups.

4.2.1.7 Voxel-Based T_2 Relaxometry (VBR^{T_2})

FWER and Bonferroni Correction

When running statistical inference on images, a common problem is the so called multiple testing problem. That is, if an image has 100,000 voxels, then at $P < 0.05$ there will be 5,000 false positives. This is a critical problem of voxel-based analysis. Familywise error rate (FWER or FWE), defined for this problem, is the chance of one or more false positives anywhere in the image.

There are many ways that FWER could be controlled. The most famous, and easiest to implement approach, is the so called Bonferroni correction, which involves testing each voxels at an actual significance level α of, take the above image as an example, $\alpha = 0.05/100,000 = 5 \times 10^{-8}$, that is, $P < 5 \times 10^{-8}$. The Bonferroni correction is therefore an extremely conservative procedure to control for FWER, and increase the

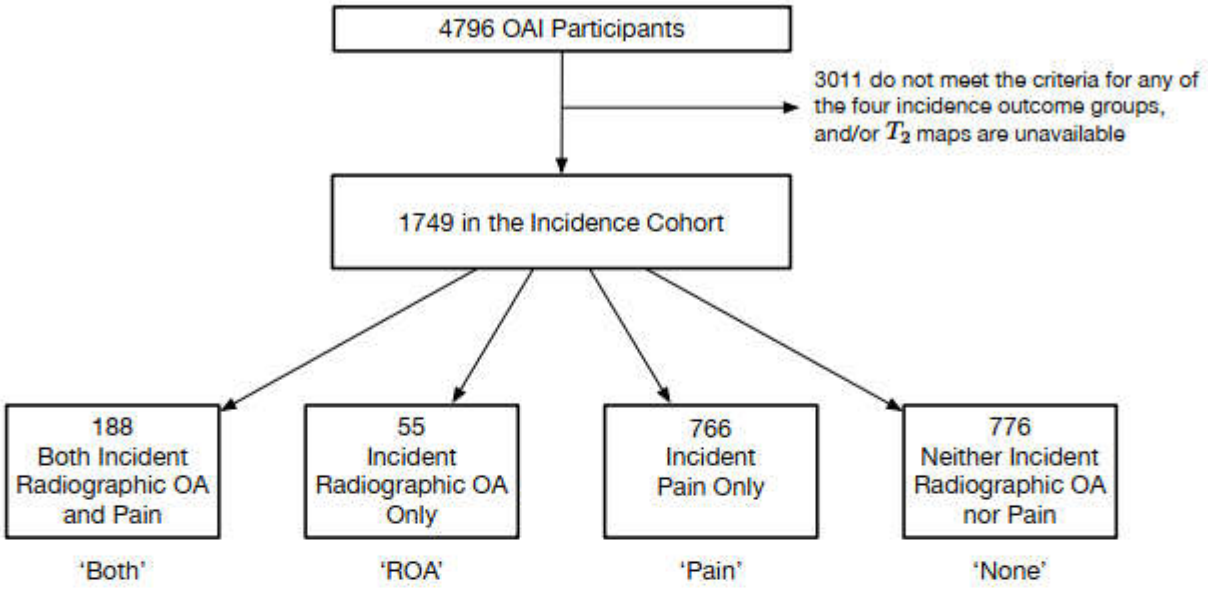


Figure 4.3: Stratification of the Four Incidence Groups

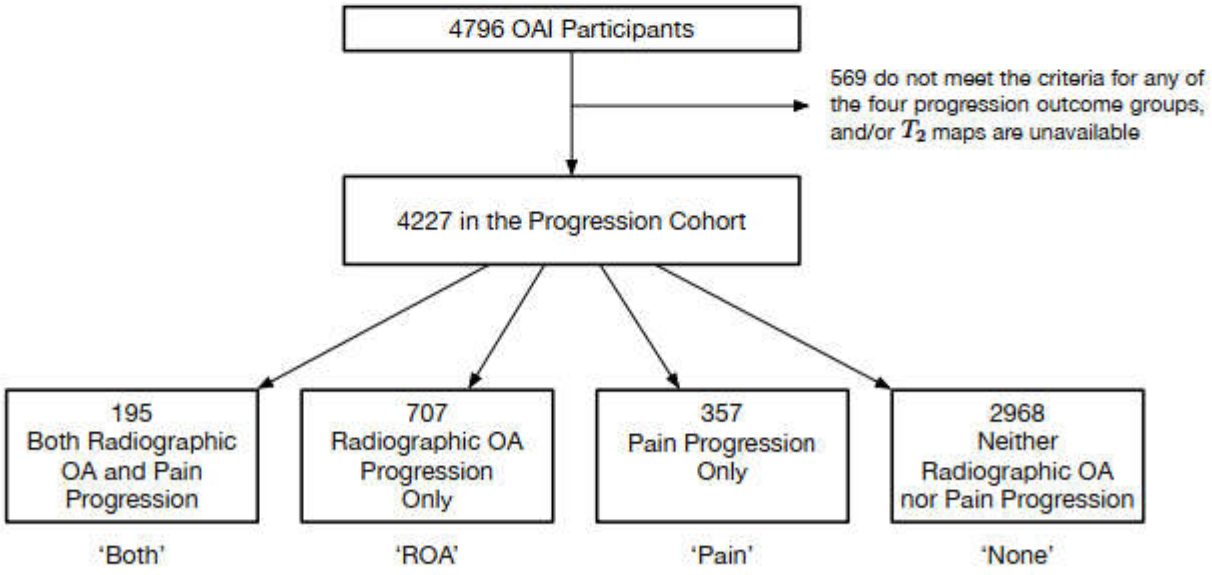


Figure 4.4: Stratification of the Four Progression Outcome Groups

probably of false negative.

In this study, we implemented Bonferroni correction for its simplicity. However, it is only used as an reference to confirm our finding in the non-corrected tests, rather than our main tool to look for pattern. Actually, we will show later most patterns will not survive such a P-value correction procedure, and those survived it also displays significant pattern reduction in the VBR^{T_2} maps.

Number of Voxels Detected Distance Map

To solve the relationship among these outcomes, we created an imaginary two-dimensional space in which the outcomes of 'Both', 'ROA', 'Pain', and 'None' are separated from each other at a distance proportional to the number of voxels detected in the pairwise two sample t-tests between them. These maps are shown in Figs. 4.43 and 4.44 on page 212 and on page 213.

These number of voxels detected distance map provided a clear overview about how easy/difficult these outcome can be differentiated against each other in T_2 maps acquired at the baseline. The maps can also be used as a reference to understand in which way T_2 mapping should be used to a more reliable prognosis, and in which it might not be as effective.

The round node are the groups while the connection line represent the pattern, if exists, between them. Most importantly, the length of the connection line shows the number of voxels detected in the t-test, and therefore is used to show how separated they are.

4.2.1.8 Voxel-based Statistical Analyses: An Overview

This study is designed to support the following group analyses for investigating the prognostic value of T_2 mapping:

1. How do the mean T_2 and standard deviation (SD) patterns of the four outcome groups look like, and different from each other? (Voxel-based descriptive statistics:

mean T_2 and SD patterns of the four outcome groups, 1 analysis);

2. Are those difference in the mean T_2 patterns significant in at least two outcomes? (Voxel-based T_2 relaxometry (VBR^{T_2}): one-way ANOVA of all the four outcome groups, 1 analysis);

3. If so, which specific groups were statistically significantly different from each other? (VBR^{T_2}: 2-sample t-tests between individual groups, $\binom{4}{2} = 6$ analyses, listed below)

   (1) How does T_2 mapping differentiate diseased (ROA and/or pain) knees from normal knees? ('Both' vs 'None', 'ROA' vs 'None', 'Pain' vs 'None', 3 analyses)

   (2) How does T_2 mapping differentiate knees affected by both ROA and pain from those affected by only one of them? ('Both' vs 'ROA', 'Both' vs 'Pain', 2 analyses)

   (3) How does T_2 mapping differentiate knees affected by ROA only from those affected by pain only? ('ROA' vs 'pain', 1 analysis)

4. Would the differentiation be more significant if the groups are combined? (VBR^{T_2}: 2-sample t-tests between combined groups, 3 analyses, listed below)

   (1) How does T_2 mapping differentiate all ROA incidence/progression from no ROA incidence/progression? [('Both' + 'ROA') vs ('Pain' + 'None'), 1 analysis]

   (2) How does T_2 mapping differentiate all pain incidence/progression from no pain incidence/progression? [('Both' + 'Pain') vs ('ROA' + 'None'), 1 analysis]

   (3) How does T_2 mapping differentiate all ROA incidence/progression from normal knees? [('Both' + 'ROA') vs 'None', 1 analysis]

These analyses are performed on both incidence and progression outcome groups, which are discussed in the following sections.

4.2.2 Results and Discussion

4.2.2.1 Voxel-Based Descriptive Statistics for the Incidence Outcome Groups: Visual Assessment

Medial Slices

Figs. 4.5 to 4.9 on pages 151–155 are voxel-based mean T_2 and SD patterns of five slices (slices 4–8 of 27) on the medial side. No visible inter-group mean T_2 difference can be found among the four incidence outcome groups in any slice, although a clear diffused higher SD (more blue colour against yellow) can be observed for the 'ROA' group in the posterior femoral condyle, especially in Figs. 4.5, 4.6, 4.8 and 4.9 (slice 4, 5, 7 and 8 of 27).

The mean T_2 patterns show a laminar appearance in both femur and tibia, which is more obvious in slices where cartilage is thicker, for example in Fig. 4.8 on page 154 (slice 7 of 27). Previous study [40] found superficial cartilage have a higher T_2 value compared to deep zone, which is also observed in the tibial cartilage of all the groups in Fig. 4.8. However, our mean T_2 pattern of the femoral cartilage, for example in Fig. 4.8, shows a different pattern where deep zone cartilage has a higher T_2 value than the superficial layer in all the groups. This pattern is consistent across the central contacting region through the posterior femoral condyle, matching the aforementioned SD patterns.

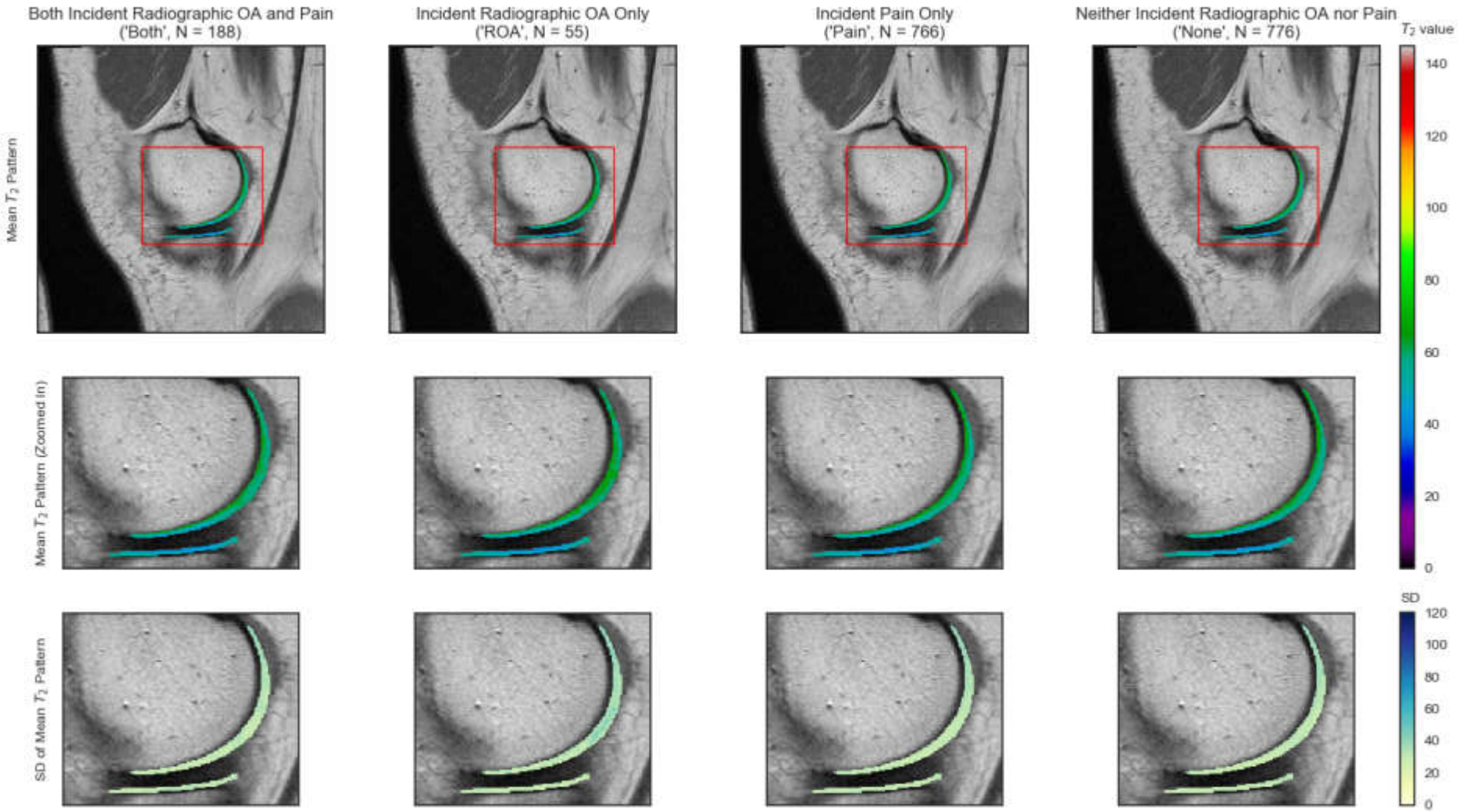


Figure 4.5: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Medial Slice (Slice 4/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.22 on page 177

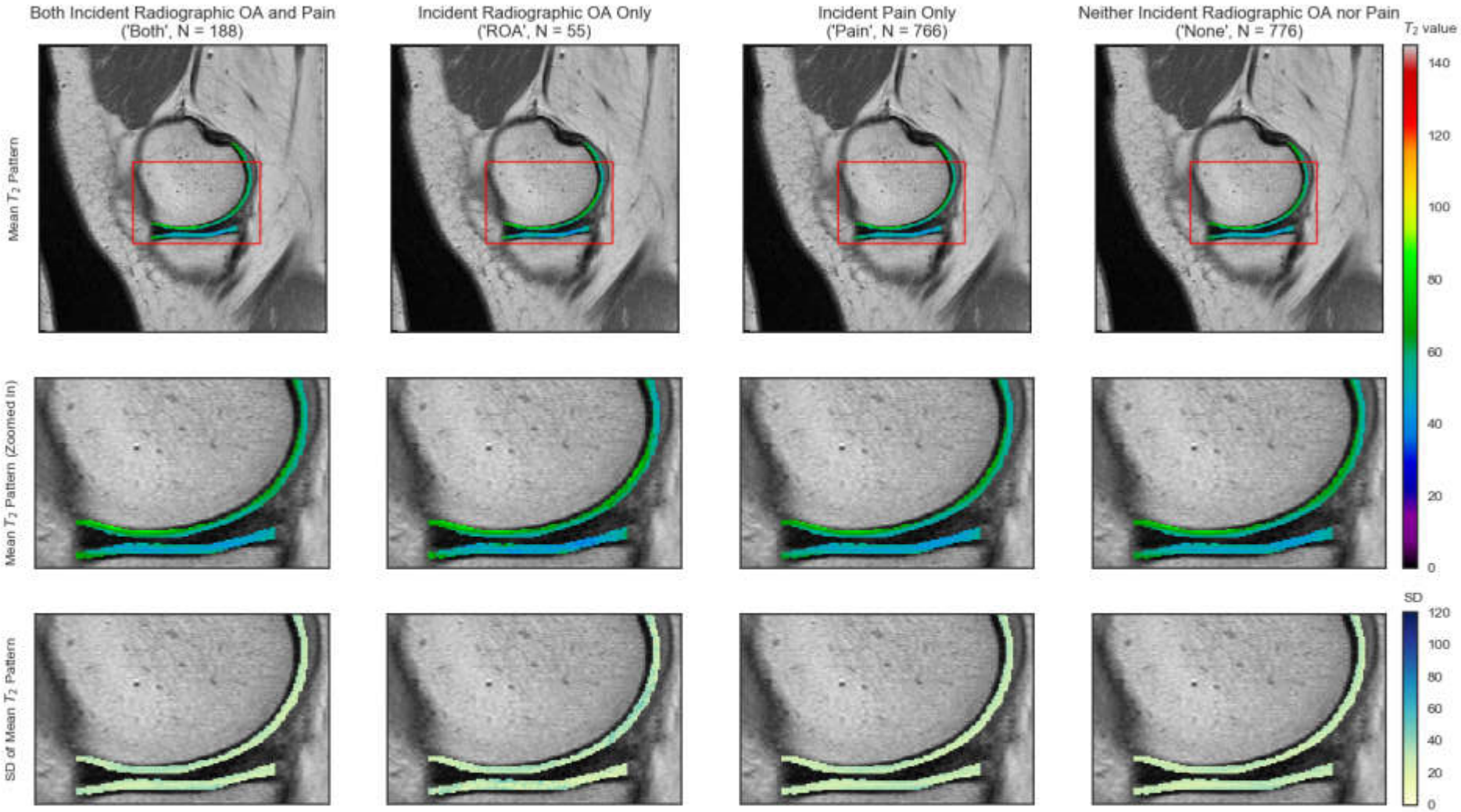


Figure 4.6: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Medial Slice (Slice 5/27[b])

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.23 on page 178

[b]This slice (5/27) is shown in the column '0' of the VBR^{T_2} one-way ANOVA result in Fig. 4.18 on page 167, but is omitted in the VBR^{T_2} two-sample t-tests results in Figs. 4.19 to 4.20 on pages 169–173 due to restricted page size.

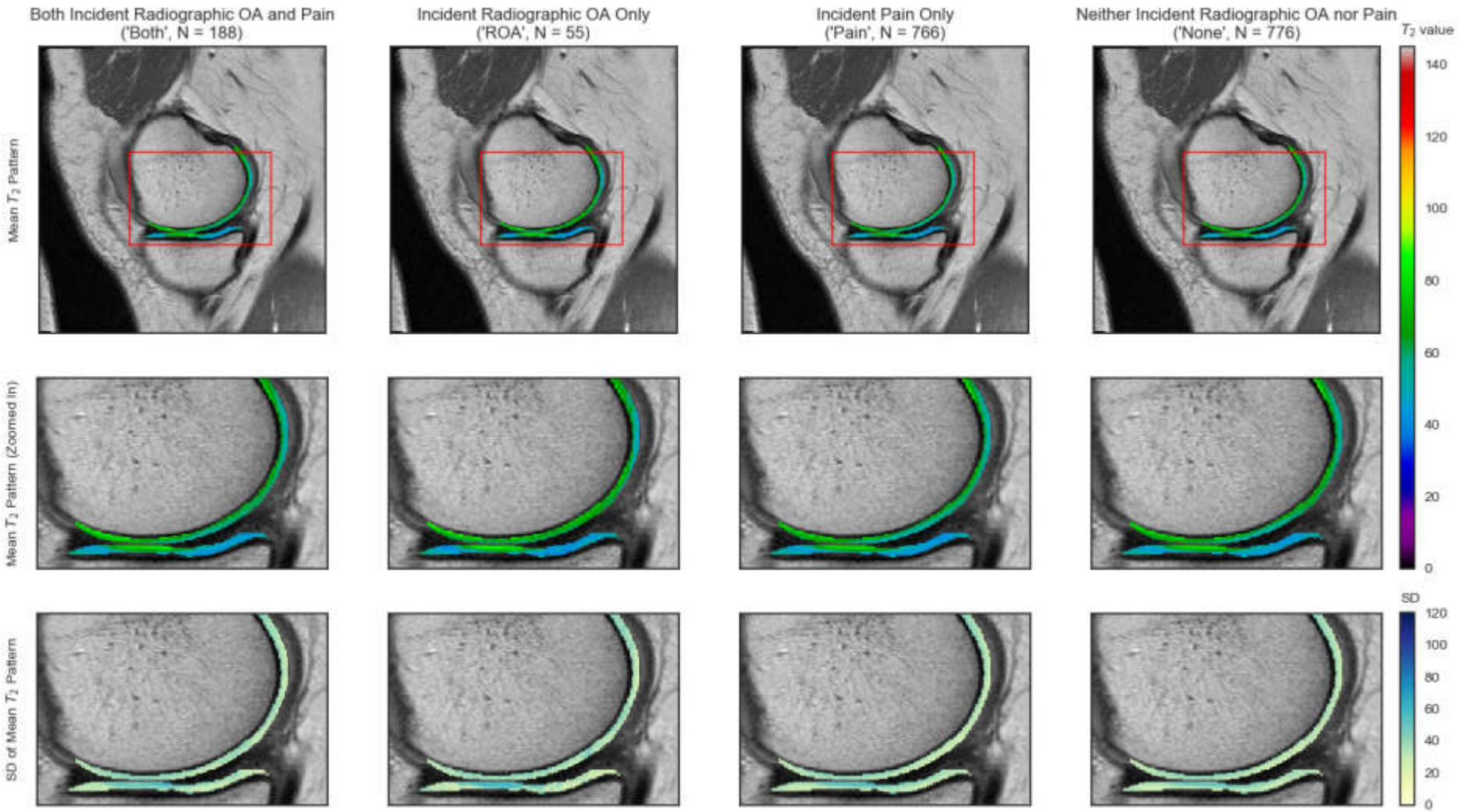


Figure 4.7: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Medial Slice (Slice 6/27[b])

[a] Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.24 on page 179

[b] This slice (6/27) is shown in columns '1's in VBR^{T_2} Figs. 4.18 to 4.20 on pages 167–173

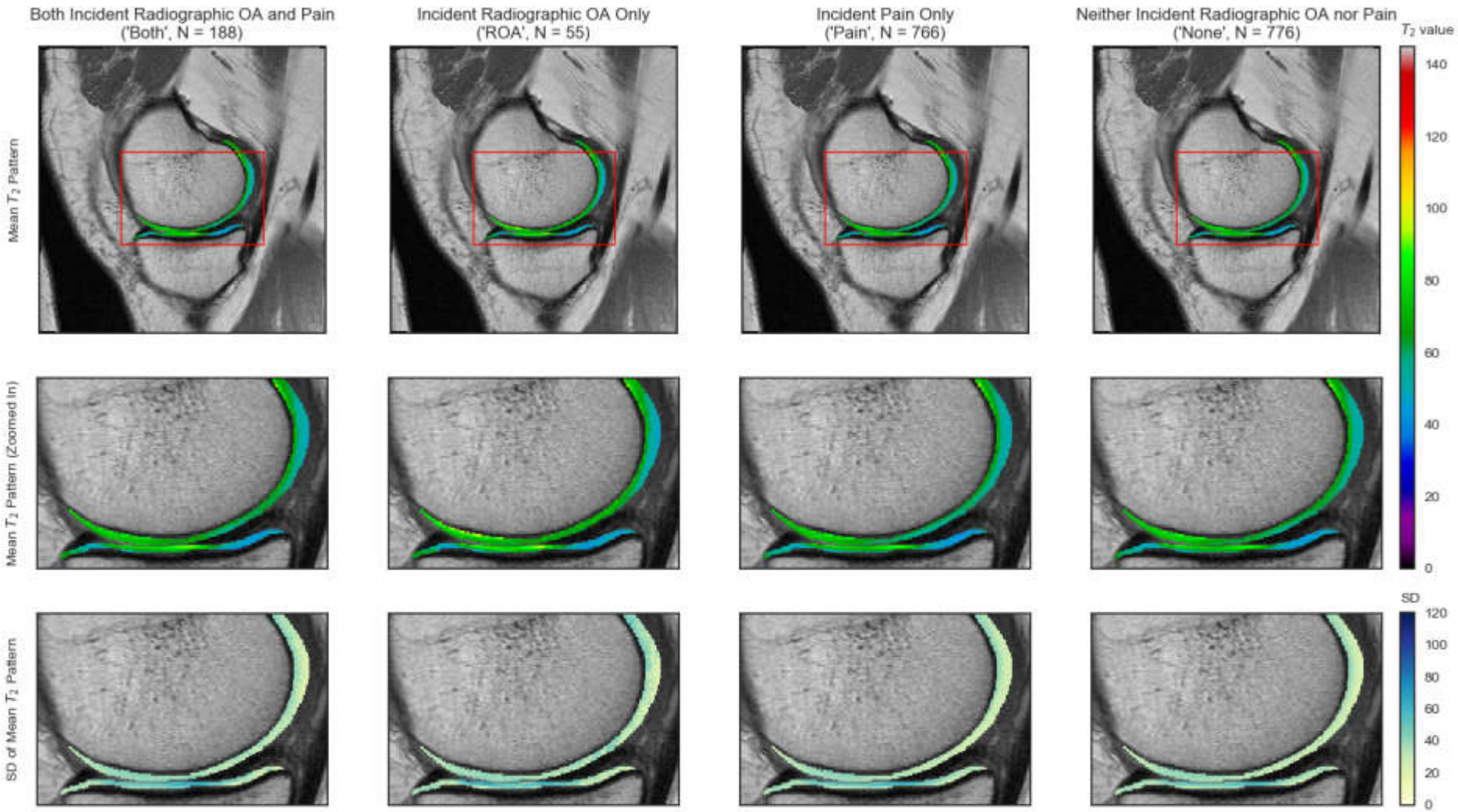


Figure 4.8: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Medial Slice (Slice 7/27[b])

[a] Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.25 on page 180

[b] This slice (7/27) is shown in columns '2's in VBR^{T_2} Figs. 4.18 to 4.20 on pages 167–173

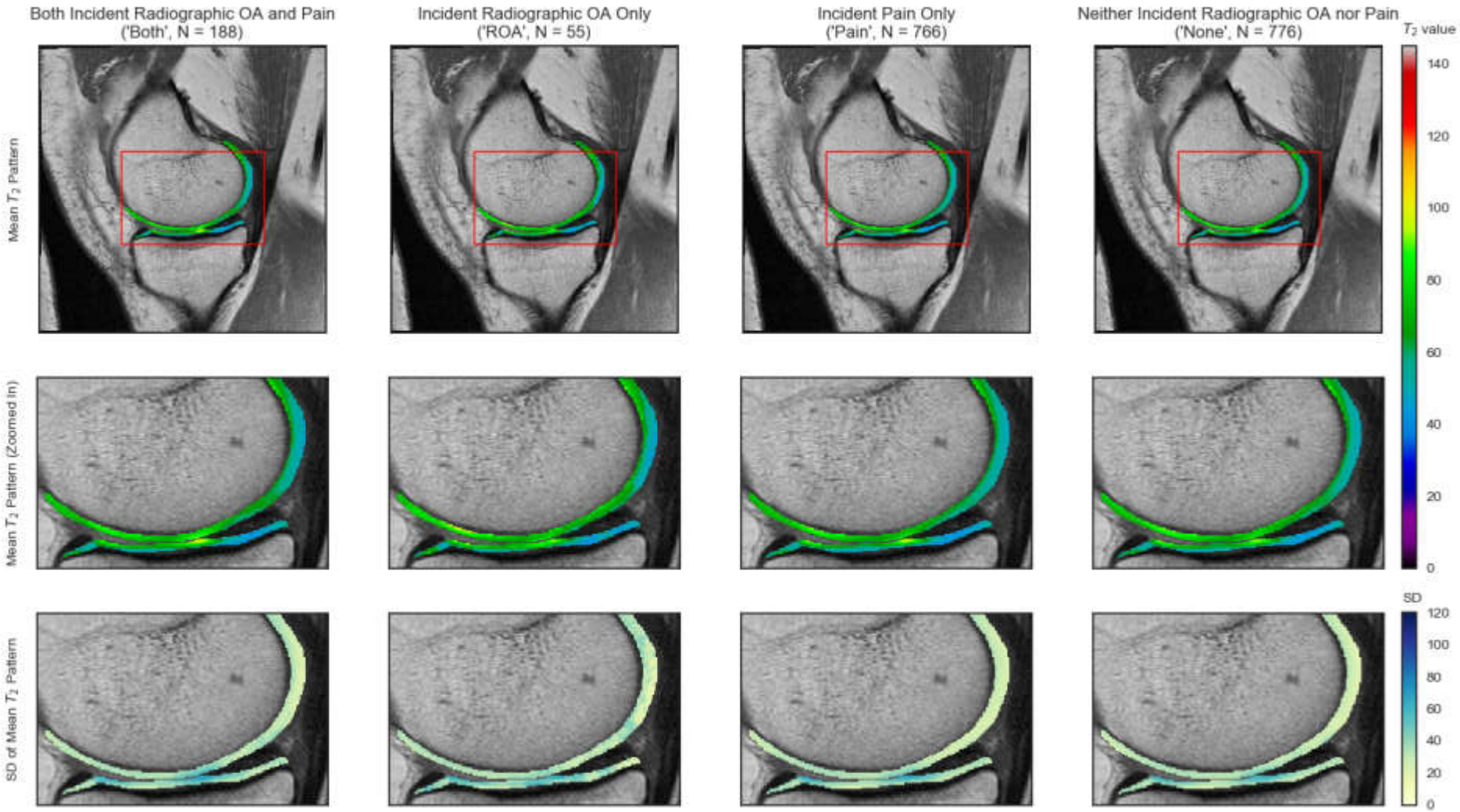


Figure 4.9: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Medial Slice (Slice 8/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.26 on page 181

Patellofemoral Slices

Figs. 4.10 to 4.12 on pages 157–159 are voxel-based mean T_2 and SD patterns of the patella in three patellofemoral slices (slices 13–15 of 27). Fig. 4.10 shows a higher T_2 (in green colour) in the superior patella of 'ROA' group than that of 'Pain' and 'None' groups (in blue colour), with 'Both' slightly lower than the 'ROA' group. In all the three slices in Figs. 4.10 to 4.12, 'ROA' group have the largest and deepest high T_2 region (in yellow and red colours) in the surface of patellar cartilage near the patellofemoral contact region, compared to that of 'Both', 'Pain' and 'None' groups. The anterior femoral cartilage in these slices have visible higher T_2 (in light green colour) in the superficial and deep zones, separated by a lower T_2 zone (in dark green colour) in the middle. No difference could be observed in the SD patterns among groups in these slices.

Lateral Slices

Figs. 4.13 to 4.17 on pages 160–164 are voxel-based mean T_2 and SD patterns of five slices (slices 17–21 of 27) on the lateral side. Apparent in these slices are the higher T_2 regions (in red and yellow colours) in the tibia just below the anterior horn of medial meniscus, where groups with pain ('Both' and 'Pain') show higher T_2 than groups without pain ('ROA' and 'None') in Figs. 4.13 to 4.16. In all the groups, the laminar appearance in the tibia is very clear, where the superficial zone has a higher T_2 than deeper zones.

Interestingly, in Figs. 4.13 and 4.14, groups without ROA ('Pain' and 'None') show a visible higher T_2 layer compared to groups with ROA ('Both' and 'ROA') in the superficial zone of femoral cartilage. However, no substantial difference can be observed among groups in the femoral cartilage in other slices.

Both femoral and tibial cartilage show substantially higher T_2 values in anterior side than posterior side in all the groups and all the slices. However, this pattern is only retained in the SD patterns of tibia, while the SD of femur are higher in the central-posterior region than far anterior/posterior regions.

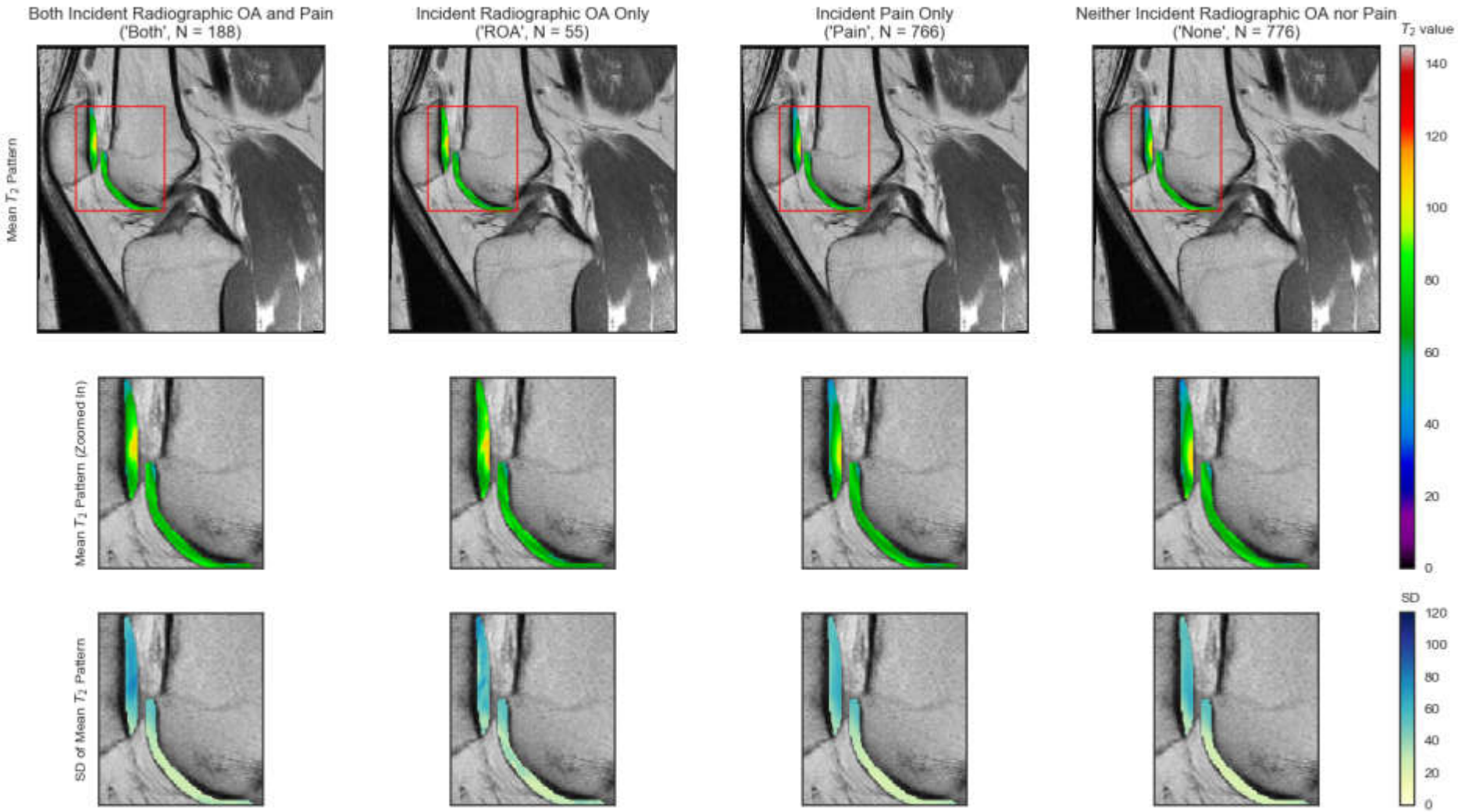


Figure 4.10: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Patellofemoral Slice (Slice 13/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.27 on page 183

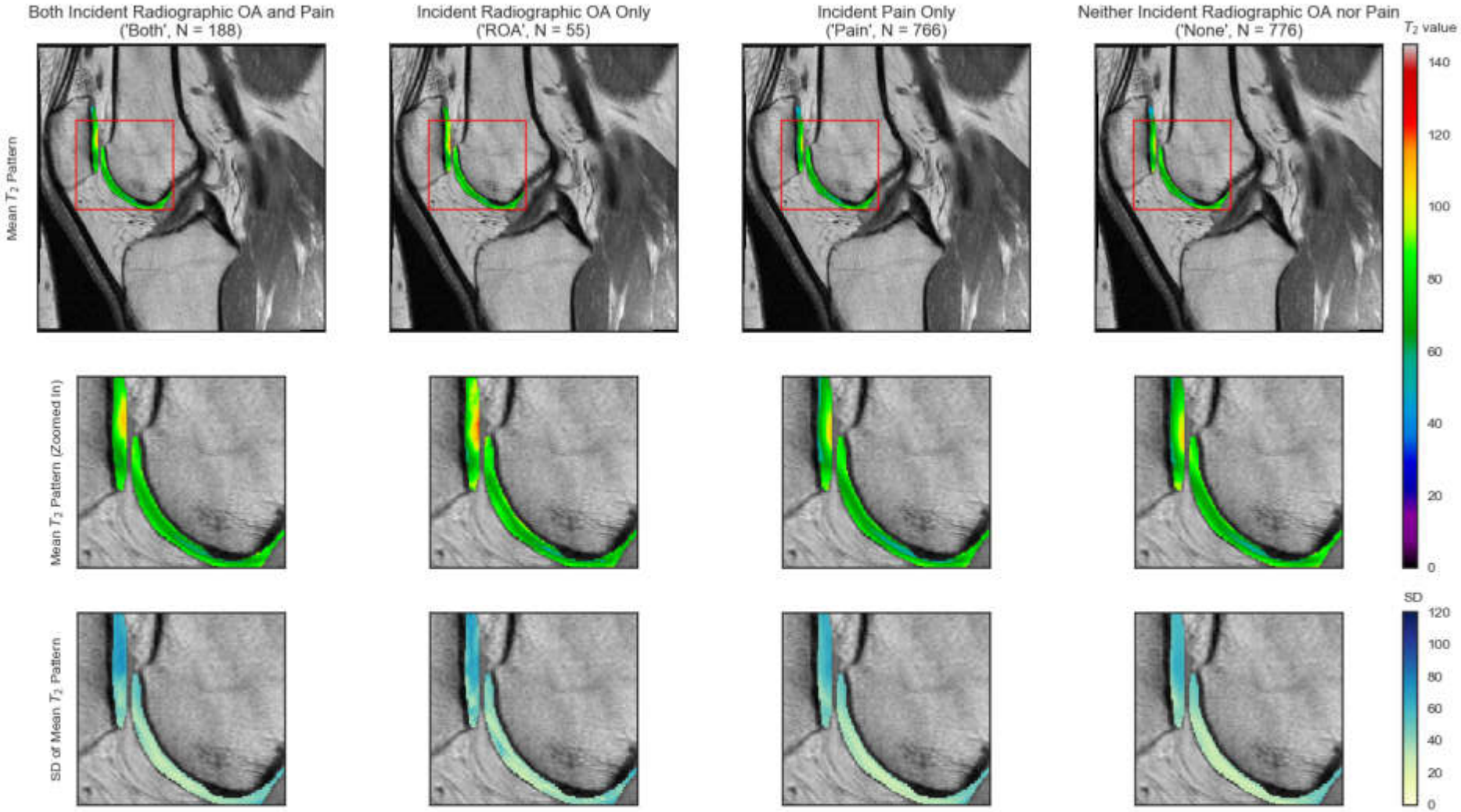


Figure 4.11: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Patellofemoral Slice (Slice 14/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.28 on page 184

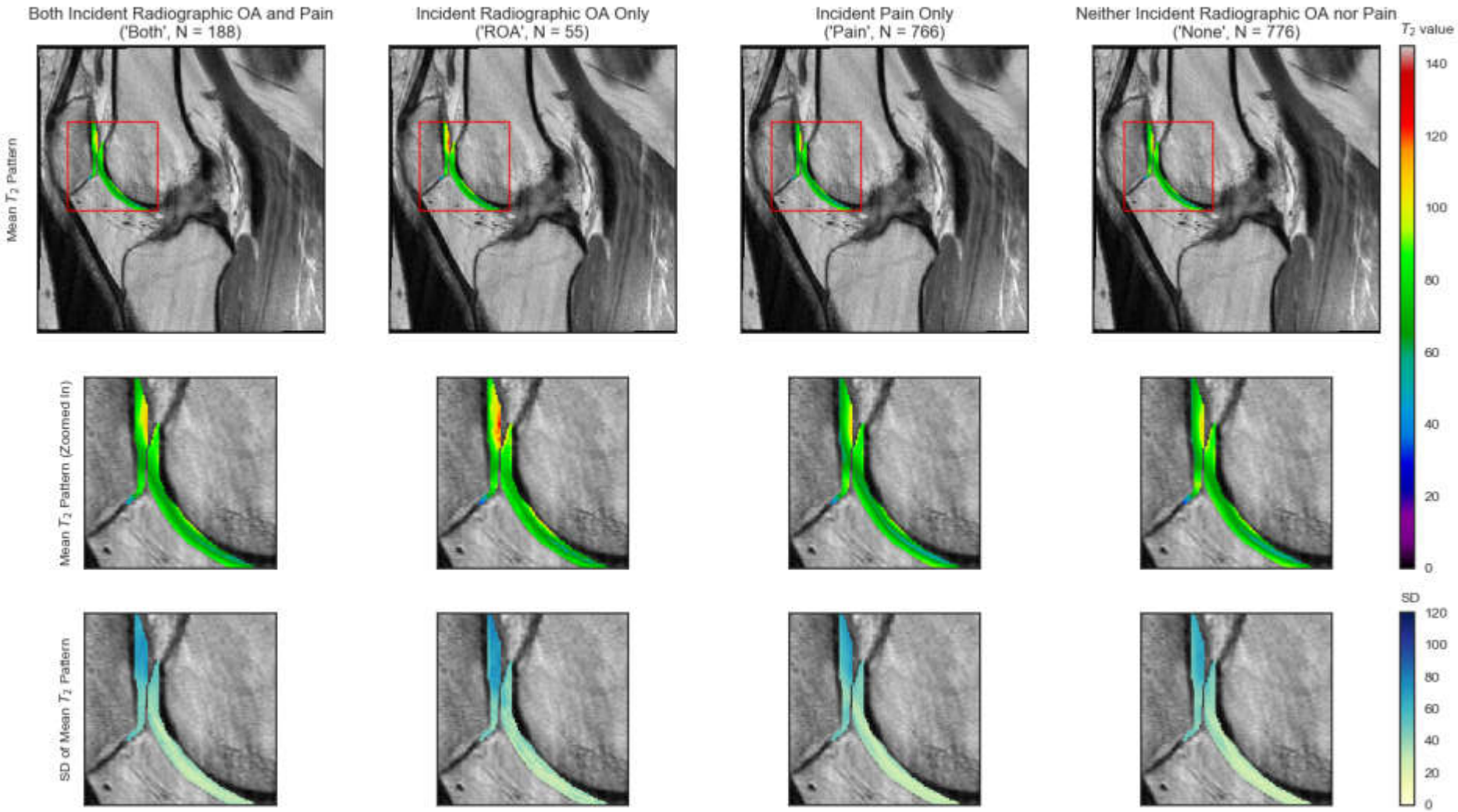


Figure 4.12: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Patellofemoral Slice (Slice 15/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.29 on page 185

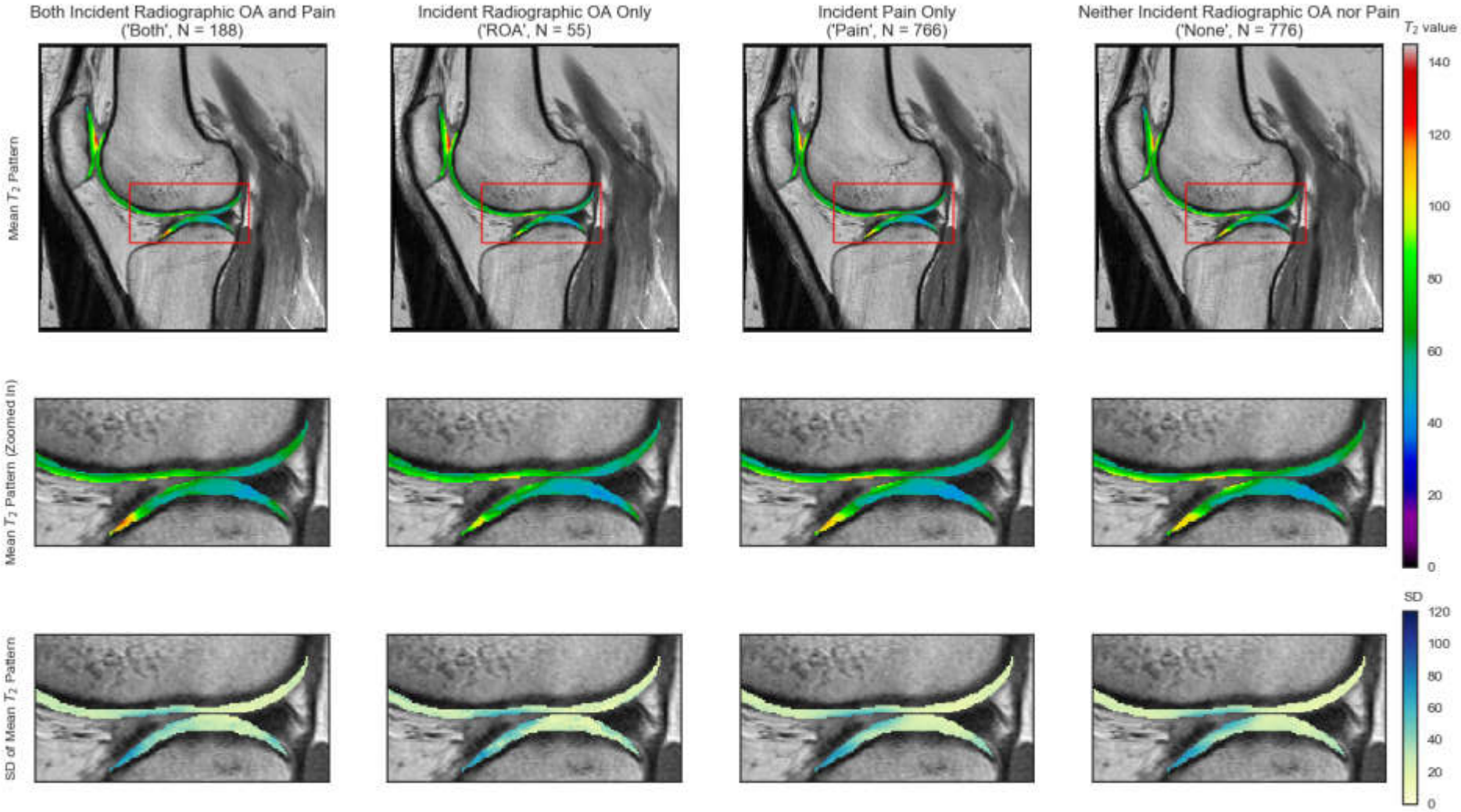


Figure 4.13: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Lateral Slice (Slice 17/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.30 on page 186

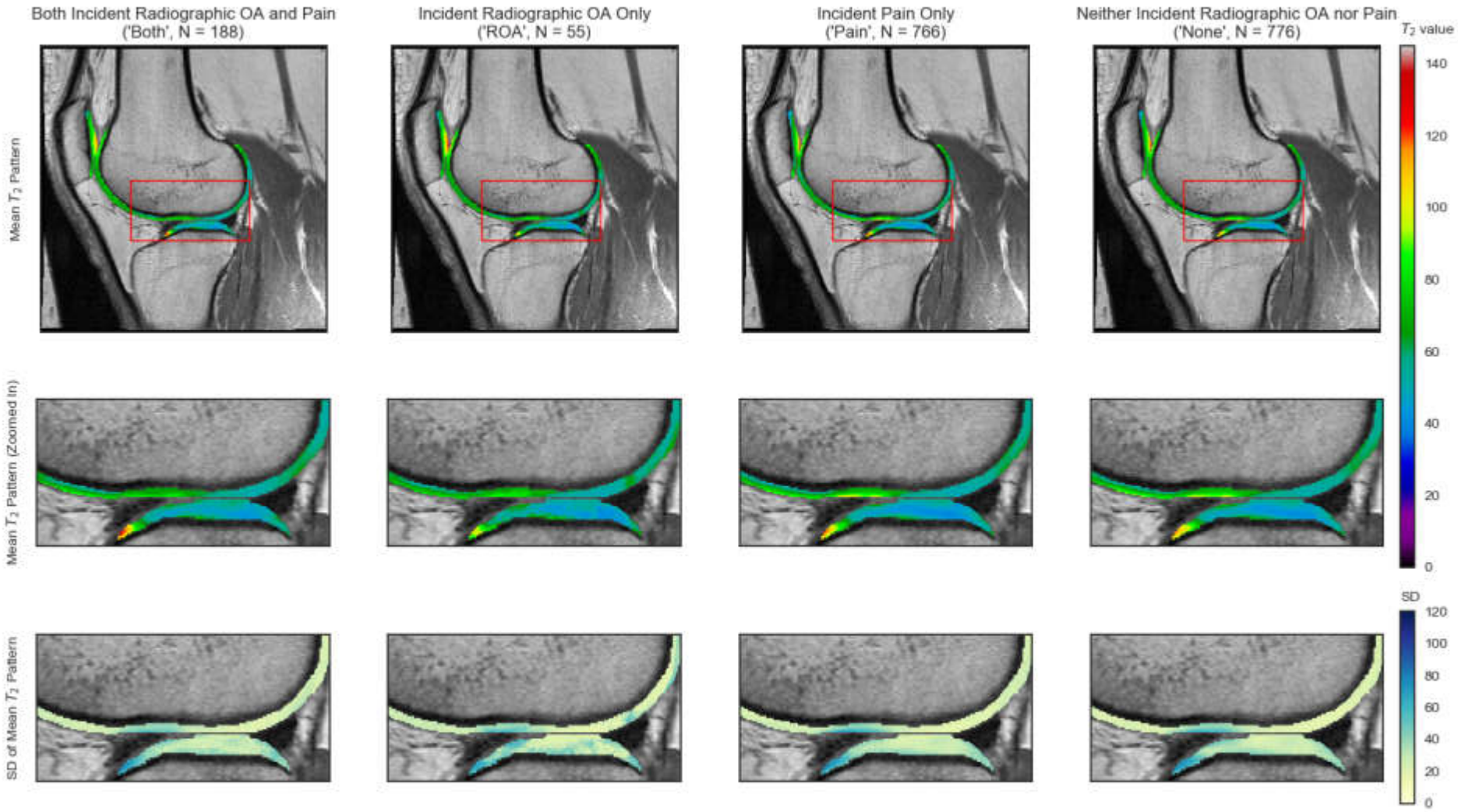


Figure 4.14: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Lateral Slice (Slice 18/27)

[a]Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.31 on page 187

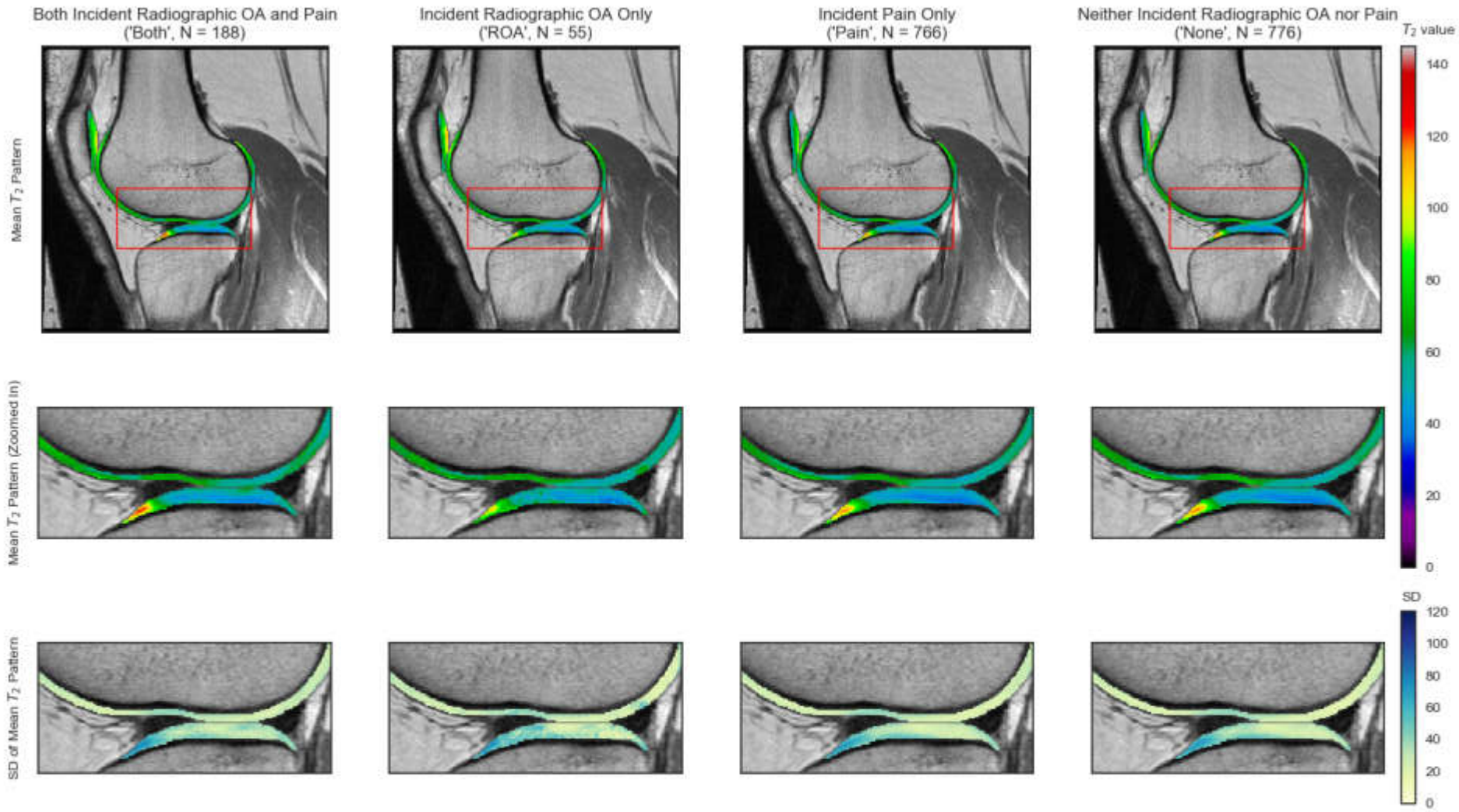


Figure 4.15: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Lateral Slice (Slice 19/27[b])

[a] Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.32 on page 188

[b] This slice (19/27) is shown in columns '3's in VBR^{T_2} Figs. 4.18 to 4.20 on pages 167–173

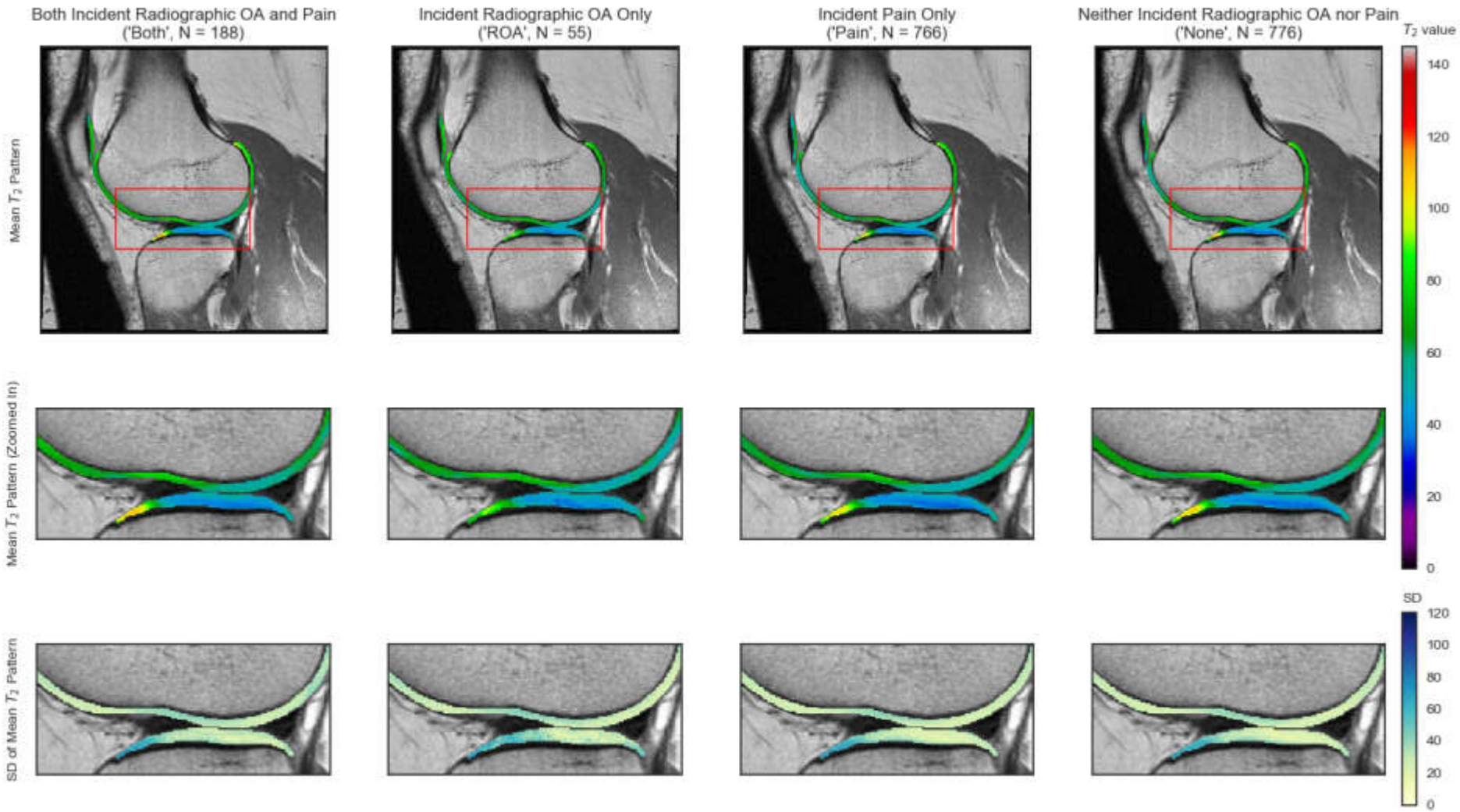


Figure 4.16: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Lateral Slice (Slice 20/27[b])

[a] Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.33 on page 189

[b] This slice (20/27) is shown in columns '4's in VBR^{T_2} Figs. 4.18 to 4.20 on pages 167–173

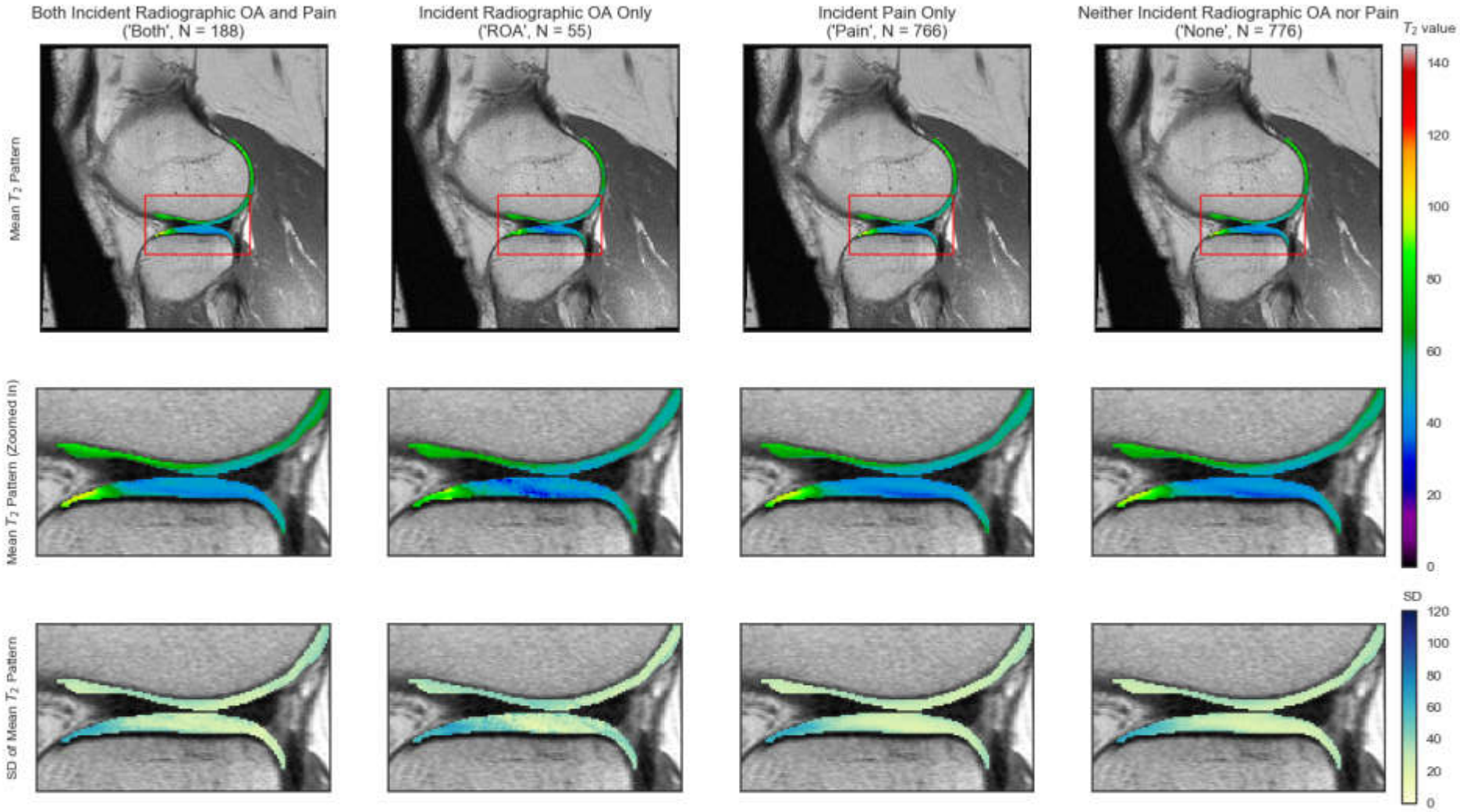


Figure 4.17: Baseline Mean T_2 and Standard Deviation Patterns of the Four Incidence[a] Outcome Groups in a Lateral Slice (Slice 21/27[b])

[a] Cf. the patterns of the four progression outcome groups in the same slice in Fig. 4.34 on page 190

[b] This slice (21/27) is shown in the column '5' of the VBR^{T_2} one-way ANOVA result in Fig. 4.18 on page 167, but is omitted in the VBR^{T_2} two-sample t-tests results in Figs. 4.19 to 4.20 on pages 169–173 due to restricted page size.

4.2.2.2 VBR^{T_2} of the Incidence Outcome Groups: One-way ANOVA

Fig. 4.18 on page 167 shows the one-way ANOVA results of the four incidence outcome groups, where columns 0, 1 and 2 are VBR^{T_2} patterns of three representative medial slices (slices 5–7 of 27), while columns 3, 4 and 5 are the patterns of three representative lateral slices (slices 19–21 of 27), respectively.

Row A in Fig. 4.18 shows the negative $\lg P$-value maps without any P-value correction for FWE controlling, where 7,690 voxels with $P < 0.05$ ($-\lg P > 1.3$) are detected. As can be seen, most activated voxels are in the patella, anterior femur and central–posterior tibial on the lateral side (Fig. 4.18 A3–A5), which means the T_2 differences between the four incidence groups are mainly in the patellofemoral and lateral compartments of the knee. The medial slices in Fig. 4.18 A0–A2, on the other hand, displays minimal and scattered activations in the superficial zone of femoral, and deep zone of tibial cartilage, indicating the T_2 difference among the four incidence groups are not evident on the medial side. This is an interesting finding as it goes against commonly believe initiation in the anteromedial region [154].

After the Bonferroni correction of P-value at different thresholding levels (rows A*, A** and A*** in Fig. 4.18), the vast majority of voxels are deactivated, with an exiguous number of voxels (49 detections at $P < 0.05$ and 18 detections at $P < 0.01$) activated in the tibiofemoral contact region of the lateral tibia (Fig. 4.18 A*3, A**3), and the patellofemoral contact region of the anterolateral femur (Fig. 4.18 A*4, A**4). No voxel is seen to be activated in the medial slices (Fig. 4.18 A*0–A*2, A**0–A**2), indicating there is no significant T_2 difference among in the four incidence group on the medial side of the knee. At $P < 0.001$ (row A*** in Fig. 4.18), there is no voxel detected on either medial and lateral side.

Overall, the Bonferroni corrected negative $\lg P$-value maps in rows A*, A** and A*** of Fig. 4.18 indicate that the T_2 difference among the four incidence groups is not significant

on either medial or lateral side if the FWER is controlled, although a very small number of voxels might be able to reliably differentiate (at least two of) the incidence outcome groups at $P < 0.05$ even after P-value correction. To find out specifically in which (individual or combined) groups can T_2 achieve better differentiation, VBR^{T_2} of two-sample t-tests are performed in the following sections.

4.2.2.3 VBR^{T_2} of the Incidence Outcome Groups: Two-sample t-tests

As has been discussed in Section 4.2.1.8 on page 148, $\binom{4}{2} = 6$ pairwise two-sample t-tests between the individual incidence outcome groups, and 2 t-tests between the combined outcome groups were performed as a decomposition of the one-way ANOVA. In this section, these two-sample t-tests are discussed in an overall ascending order of their numbers of voxels detected. Occasional adjustment is made, however, to this order when some groups display a common pattern and need to be discussed together.

Both Incident ROA and Pain vs Incident ROA Only ('Both' vs 'ROA')

Before Bonferroni correction, the weakest voxel activation among the eight two sample t-tests is observed between the 'Both' and 'ROA' outcome groups, in which only 1,222 voxels are detected as shown in row A of Fig. 4.19 on page 169. Compared with other no-correction t-tests in rows B and C of the same figure and rows A–E of Fig. 4.21 on page 175, the minimal voxel activation pattern in this t-test is not similar, and the number of voxels detected (1,222 voxels) not comparable to any other seven t-tests. After Bonferroni correction, no voxels could be detected in this test at $P < 0.05$ (row A* in Fig. 4.19), which means the 'Both' and 'ROA' groups are not significantly different from each other after FWER controlling.

By definition, the 'Both' group will develop both incident pain and the incident ROA, while the 'ROA' group will develop incident ROA only, over the follow-up period. The difference, reflected in the images as voxel detections (activations), is therefore related

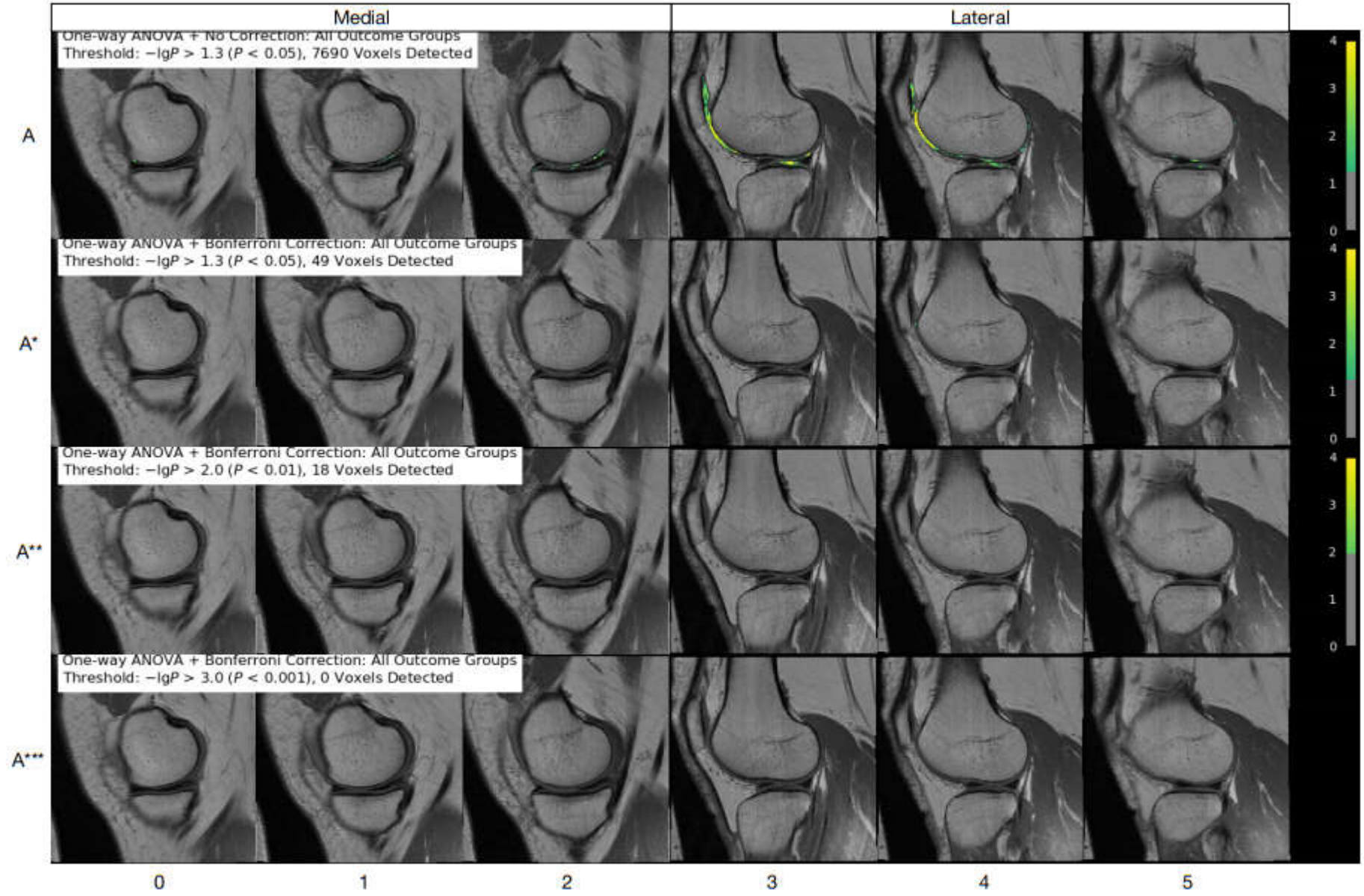


Figure 4.18: Voxel-Based T_2 Relaxometry (VBR^{T_2}): One-way ANOVA of All Incidence[a] Outcome Groups

[a]Cf. VBR^{T_2}: One-way ANOVA of All Progression Outcome Groups in Fig. 4.35 on page 193

to incident pain that comes with (or complicated by) incident ROA, denoted 'Incident Pain | Incident ROA' in the following text. If the activation is significant, then it would be possible to use T_2 mapping to predict an 'Incident Pain | Incident ROA' event in the future, using baseline images at an early time. The purpose of these t-tests is therefore to determine whether we can significantly differentiate these two groups, or in other words, whether we can significantly see, using image at the baseline, the effect of 'Incident Pain | Incident ROA' as a pattern.

As can be seen in rows A and A* of Fig. 4.19, the activations are very weak regardless of whether Bonferroni correction is performed. It is likely, therefore, that T_2 mapping is not capable of differentiating between the 'Both' and 'ROA' incidence outcomes at baseline. This also means, in the context of this chapter, that the 'Both' and 'ROA' outcome groups can be treated as one outcome and are interchangeable. This is a useful finding that will guide us through the analyses in the following two sample t-tests for the incidence outcome groups.

Pattern 1: Effect of Incident Pain

The first voxel activation pattern we identified in the two-sample t-tests is the before-Bonferroni patterns exhibited in the two sample t-tests of 'Pain' vs 'None', and ('Both'+'Pain') vs ('ROA'+'None'), shown in rows B and C of Fig. 4.19 on the facing page. As can be seen in Fig. 4.19 and Fig. 4.21 on page 175, the similar patterns displayed in these two rows, referred to as **Pattern 1** in the following text, are different from patterns of any other six t-tests. On the medial side, Pattern 1 is characterised as fragmented activations in the posteromedial femoral condyle and tibia, but very limited activations in the anterior and central tibiofemoral compartments (Fig. 4.19 B1–B2, C1–C2); on the lateral side, it features congregated detections in the femoral cartilage near the tibiofemoral contact region (Fig. 4.19 B3–B4, C3–C4), but very limited and scattered detections in the tibial cartilage.

By definition, both of the tests represent comparison between a group of incident pain,

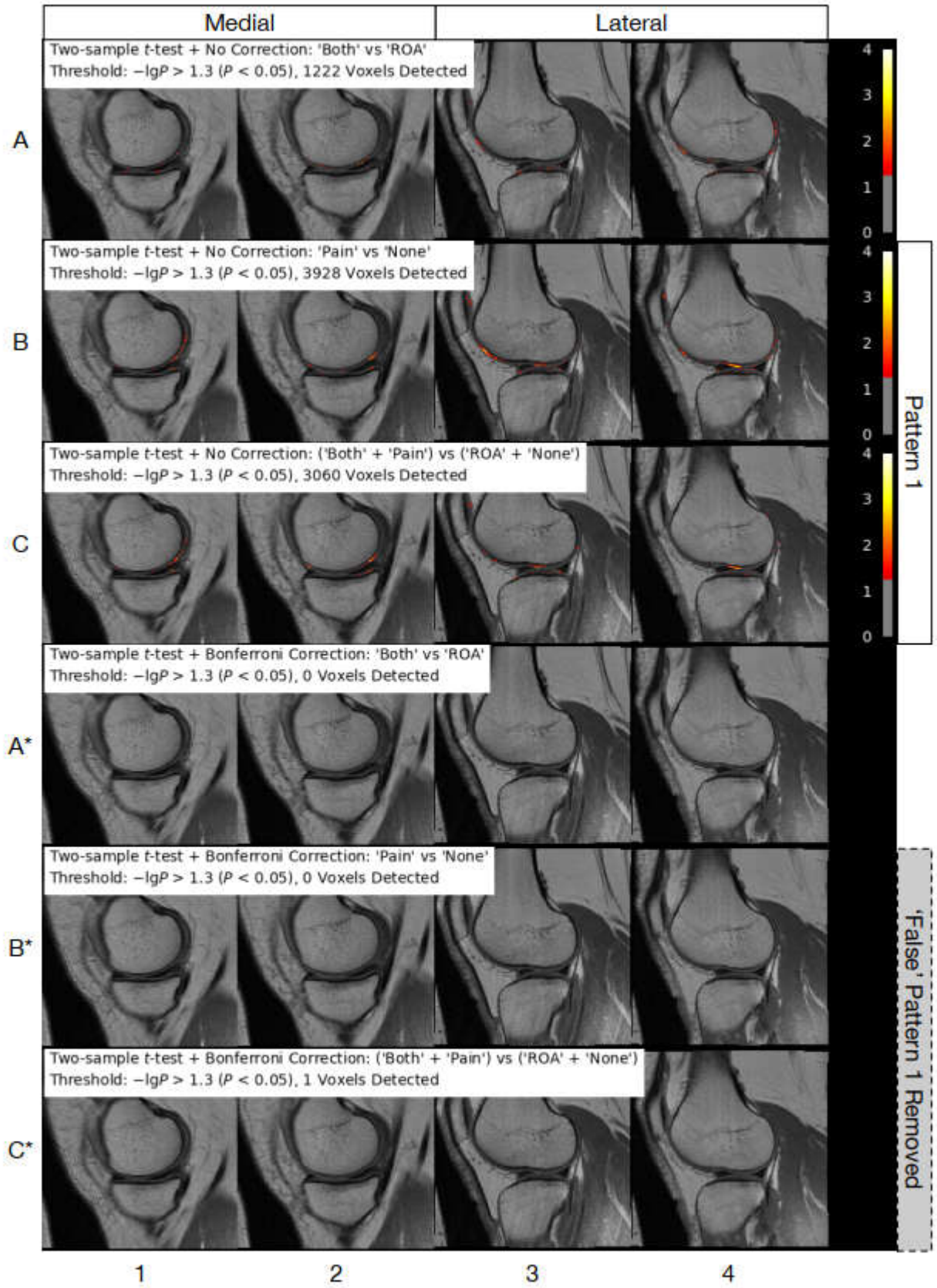


Figure 4.19: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample t-tests Between Incident Pain and No Incident Pain[a] Groups. Rows A–C: No P-value Correction; Rows A*–C*: After Bonferroni Correction.

[a]Cf. VBR^{T_2}: Two-sample t-tests Between Pain Progression and No Pain Progression Groups in Fig. 4.36 on page 195

individually ('Pain') or combined ('Both'+'Pain'), to a individually ('None') or combined ('ROA'+'None') no incident pain group. The similar activated voxel locations displayed in rows B and C of Fig. 4.19, are therefore related to the effect of incident pain relative to normal (no incident pain) knees.

In principle, we expect more activations in t-tests between combined-groups than that of individual groups, which is supported by our VBR^{T_2} results in Fig. 4.21 on page 175, where the row E exhibited more activated voxels than any of its decompositions in row A–D of the same figure, that will be discussed later.

However, the combined ('Both'+'Pain') vs ('ROA'+'None') t-test in row C of Fig. 4.19, have a smaller number of voxels detected compared to that of the 'Pain' vs 'None' t-test in row B (3,060 vs 3,928), which looks abnormal. If we think of row C of Fig. 4.19 as the 'sum' of rows A and B, then by examining the patterns it is easy to see that although the resulting row C is still demonstrating the characteristics of Pattern 1, the pattern is moderated, and tends to converge to the minimal activation pattern shown in row A. This means the dissimilarity between 'Pain' and 'None' in row B of Fig. 4.19 is to some extent neutralised by the similarity between 'Both' and 'ROA' in row A, resulting a weakened Pattern 1 in row C. In this way, the combination of the ('Both'+'Pain') and the ('ROA'+'None') groups in fact weakened, rather than enhanced the Pattern 1 shown in the 'Pain' vs 'None' t-test—involvement of the incident ROA samples in row C effectively prevented some of the voxels from being activated by incident pain. The evidence of this is the weakened activation in the whole lateral femur, particularly in the anterolateral area of Fig. 4.19 C3 compared to the same area in B3. As we will show in the next section, the lateral femur is least activated by incident ROA.

We can therefore achieve a rather remarkable finding, that incident ROA will impair our ability to observe the effect of incident pain, expected to show in locations activated in Pattern 1, using T_2 maps at baseline. We should therefore use subjects with only pain outcome, without ROA, if we want to observe only the effect of pain in their baseline T_2 maps. Failing to eliminate the influence of ROA would mean any effect (shown as the

Pattern 1) relates to a future incident pain would be cancelled out by the effect of incident ROA, in their T_2 maps acquired at the baseline.

After Bonferroni correction, Pattern 1 is completely eliminated in row B*, and only one voxel is detected in row C*, of Fig. 4.19. It means although Pattern 1 shows the effect of incident pain compared to normal knees before P-value correction, the effect is not statistically significant at $P < 0.05$ when the FWER is taken into consideration.

Pattern 2: Effect of Incident ROA

The second voxel activation pattern that was identified, referred to as **Pattern 2** in the following text, is the pattern produced by the two sample t-tests of 'ROA' vs 'None', and 'ROA' vs 'Pain', presented in rows A and B of Fig. 4.21 on page 175. On the medial side, Pattern 2 is characterised as fragmented activations in the anterior–central femur, but very limited activations in the tibia (Fig. 4.21 A1–A2, B1–B2); on the lateral side, strong detections are observed in the patellofemoral and central tibiofemoral contact regions (Fig. 4.21 A3–A4, B3–B4), as well as posterolateral tibia (Fig. 4.21 A4, B4), whereas central–posterior femur and anterior tibia demonstrate weak or no activations.

By definition, both of the t-tests compared the incident ROA group ('ROA') to a no incident ROA group ('None' or 'Pain'). Therefore, Pattern 2 is related to the effect of incident ROA relative to normal (no incident ROA) knees.

Interestingly, compared to row A in Fig. 4.21, row B preserved the Pattern 2 in row A but the number of voxels activated are greatly increased, rather than weakened (3,919 vs 5,065). It means the difference between the 'ROA' group and the 'Pain' group is more significant than that between the 'ROA' and 'None' groups in T_2 mapping at baseline.

If we consider the number of voxels detected as the 'distance' between the two samples in the t-tests, then we can say that the 'Pain' group is further away from 'ROA' than 'None'. This is a surprising result, as we would naturally think the 'distance' (difference) between two diseased status should be smaller than the 'distance' between a diseased status and a healthy status.

The similar patterns in rows A and B of Fig. 4.21 also means that Pattern 1, which was shown when 'Pain' and 'None' were tested against each other in the previous section, become negligible when they are tested against the 'ROA' group individually. Further to the previous discussion that inclusion of incident ROA subjects would weaken the incident pain effect in Pattern 1, the result in this section revealed that Pattern 1 will not only be weakened by incident ROA, but it will be completely masked by Pattern 2 if incident ROA subjects dominate one of the t-test samples.

Pattern 3: Combined Effect of Both Incident Pain and Incident ROA

The strongest, and the easiest pattern to identify among all t-tests, is the **Pattern 3**, which is exhibited in the two sample t-tests of 'Both' vs 'None', 'Both' vs 'Pain', and ('Both'+'ROA') vs ('Pain'+'None'), as shown in rows C, D, and E of Fig. 4.21 on page 175 respectively.

On the medial side, in contrast to Patterns 1 and 2 where no activation can be seen in anteromedial tibia (Fig. 4.19 B1, C1, and Fig. 4.21 A1, B1), Pattern 3 features a clear activation in anteromedial tibia (Fig. 4.21 C1, D1, E1). The fragmented activation shown in the central–posterior femur in Patterns 1 and 2 (Fig. 4.19 B2, C2, and Fig. 4.21 A2, B2) are mostly absent in Pattern 3, in which such activations are restricted in the central tibiofemoral contact region (Fig. 4.21 C2, D2, E2). On the lateral side, strong activations extending from the patellofemoral contact region through anterolateral femur (Fig. 4.21 C3–C4, D3–D4, E3–E4) characterises Pattern 3, which combined features of Pattern 1 and Pattern 2 in those compartments. Strong activations are also presented through the central tibiofemoral contact region (Fig. 4.21 D3–D4, E3–E4) to posterolateral tibia, similar to Pattern 2 in the same locations. These strong, consistent, and distinctive activation features actually made Pattern 3 the easiest pattern to be recognised among Patterns 1–3 in the two sample t-tests between incidence outcome groups.

Combining the effects of incident pain in Pattern 1, and that of incident ROA in Pattern 2, Pattern 3 confirmed that these effects are somehow 'cumulative'—more voxels are activ-

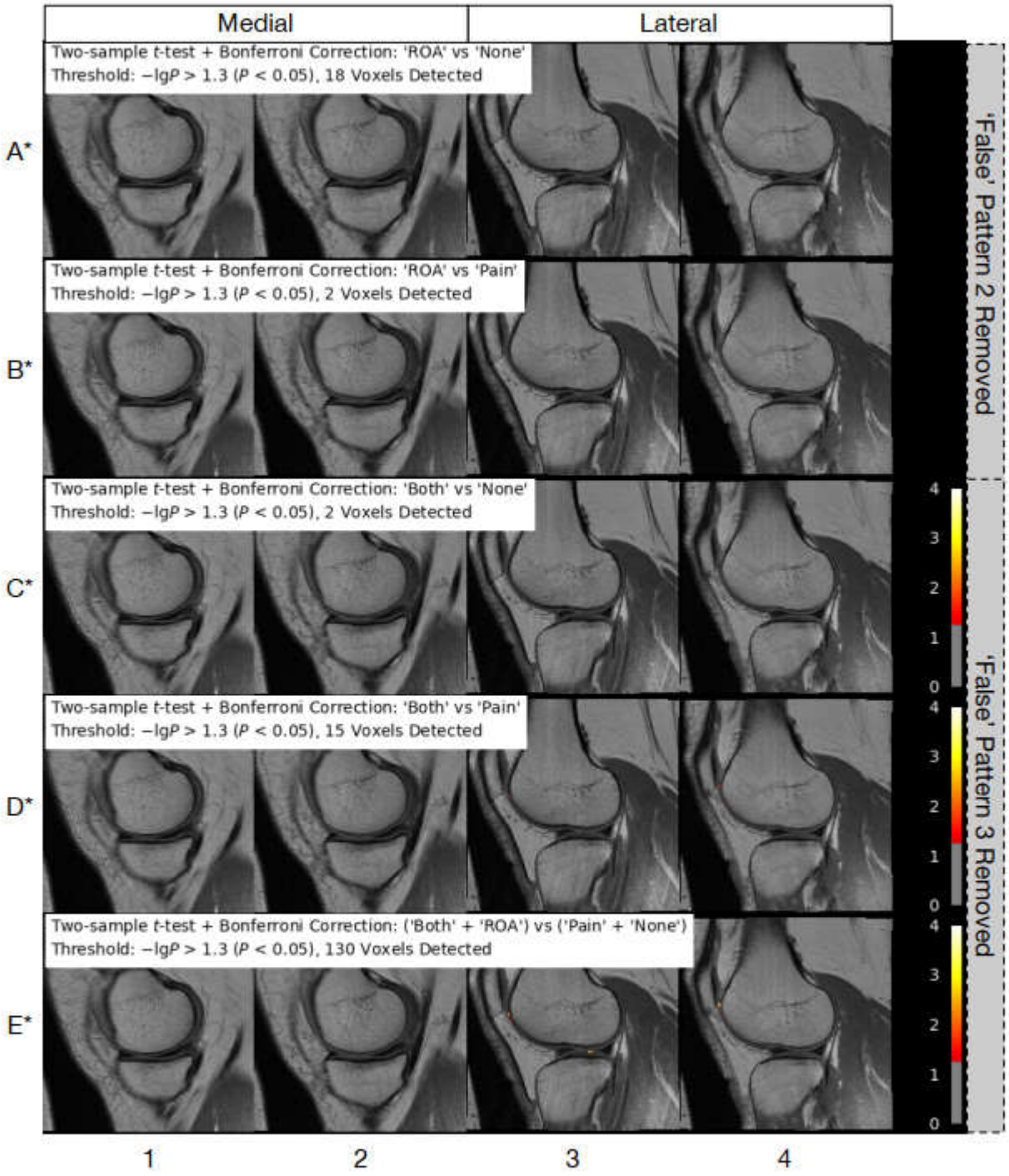


Figure 4.20: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample t-tests Between Incident ROA and No Incident ROA[a] Groups (After Bonferroni Correction)

[a]Cf. VBR^{T_2}: Two-sample t-tests Between ROA Progression and No ROA Progression Groups (After Bonferroni Correction and at $P < 0.05$) in Fig. 4.38 on page 201

ated in the t-tests showing Pattern 3, than the tests showing Patterns 1 and 2. However, such 'cumulative' relationship is much more complex than a simple arithmetic addition, because a single voxel can be related to more than one outcomes and can be activated in different patterns. The relationship can be better understood by inspecting the number of voxels detected map, which will be discussed in Section 4.2.2.7 on page 211.

After Bonferroni correction, as can be seen in Fig. 4.20, the Patterns 2 and 3 are almost totally removed, which again shows the extremity of the correction method. Although it would be unlikely the patterns recognised are all false positives, this indeed shows the effects in these patterns are relatively weak, and the incident outcome groups are challenging to be differentiated using baseline T_2 maps.

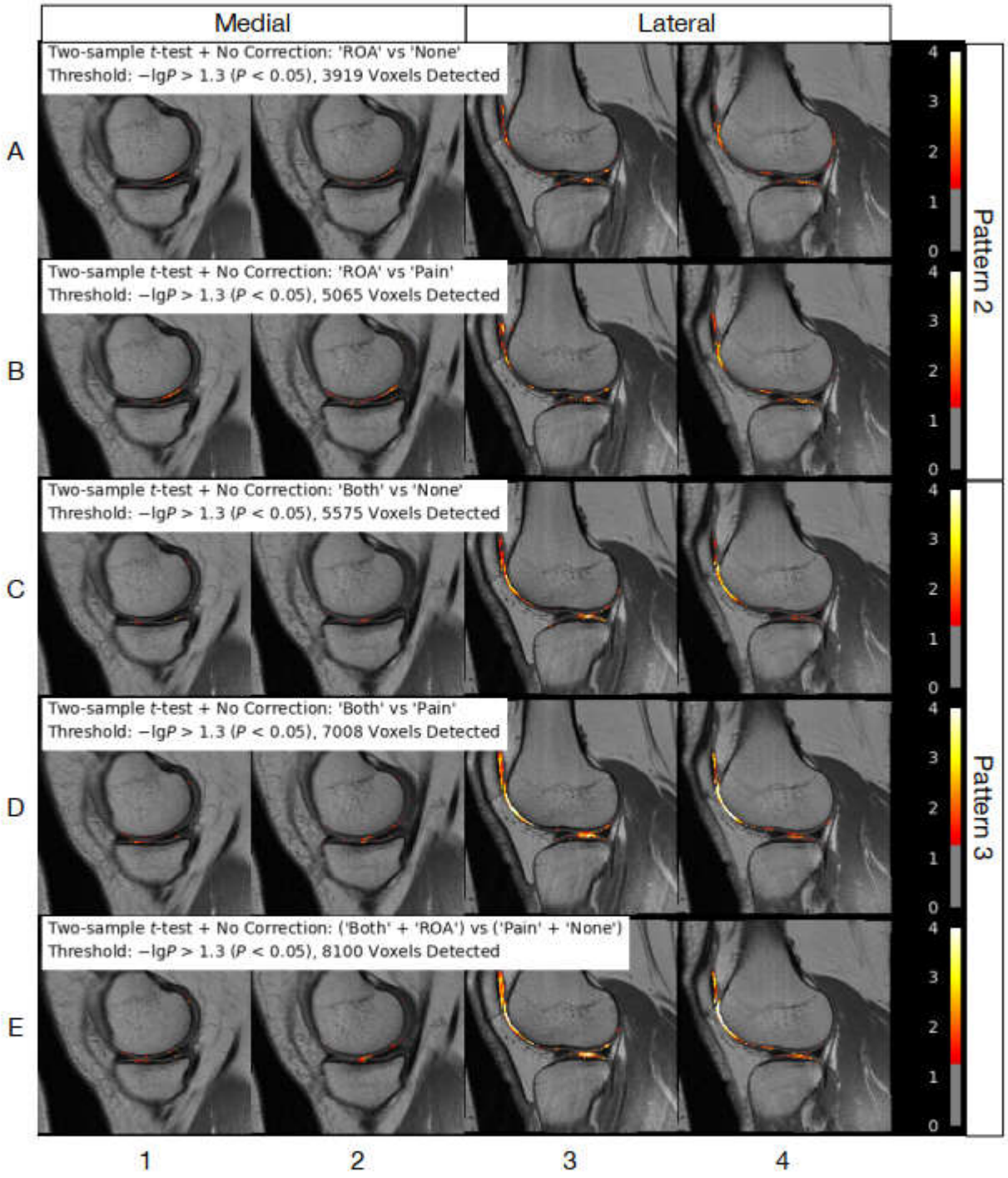


Figure 4.21: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample *t*-tests Between Incident ROA and No Incident ROA[a] Groups (No *P*-value Correction)

[a]Cf. VBR^{T_2}: Two-sample *t*-tests Between ROA Progression and No ROA Progression Groups (No *P*-value Correction) in Fig. 4.37 on page 199

4.2.2.4 Voxel-Based Descriptive Statistics for the Progression Outcome Groups: Visual Assessment

Medial Slices

Figs. 4.22 to 4.26 on pages 177–181 are voxel-based mean T_2 and SD patterns of the same medial slices (slices 4–8 of 27) as the incidence groups discussed in Section 4.2.2.1 on page 150. Unlike their incidence counterparts in Figs. 4.5 to 4.9 on pages 151–155, some clear inter-group mean T_2 differences can be observed among the four progression groups. For example, the 'Both' progression group displays a higher mean T_2 (more green colour against blue) in posterior femoral condyle, compared with the other three groups in Fig. 4.22 on the next page (slice 4 of 27). In Figs. 4.24 to 4.26 on pages 179–181 (slices 6–8 of 27), the 'ROA' progression group have distinct high T_2 spots (more yellow colour against green and blue) on the surface of the tibial cartilage near the tibiofemoral contact region, compared with the other groups.

The SD patterns are consistent with the mean T_2 patterns in Figs. 4.24 to 4.26 on pages 179–181 (slices 6–8 of 27) discussed above, where higher SD regions (more blue colour against yellow) centring around the aforementioned high mean T_2 spots are displayed equally clearly in the 'Both' and 'ROA' groups, compared with the 'Pain' and 'None' groups.

The laminar appearance discussed previously for the incidence groups in Section 4.2.2.1 on page 150 can also be observed in the patterns of their progression counterparts, particularly in Fig. 4.25 on page 180 (slice 7 of 27) where cartilage is thicker. As observed in the incidence group, it can be seen that the superficial layers also have a higher mean T_2 in the tibia, but lower mean T_2 in the femur, than the deep zones in the progression groups.

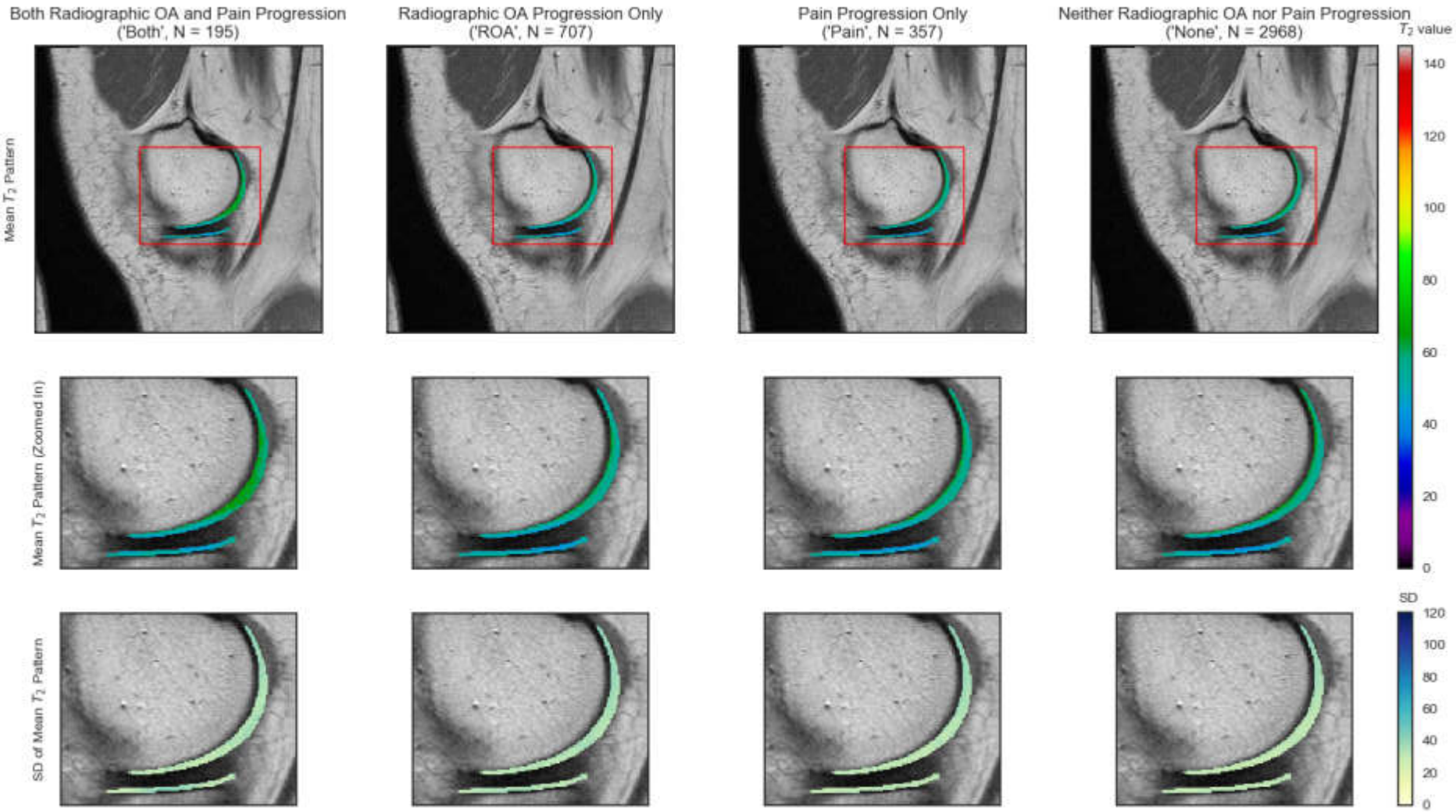


Figure 4.22: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Medial Slice (Slice 4/27)

[a]Cf. the patterns of the four incidence outcome groups in Fig. 4.5 on page 151

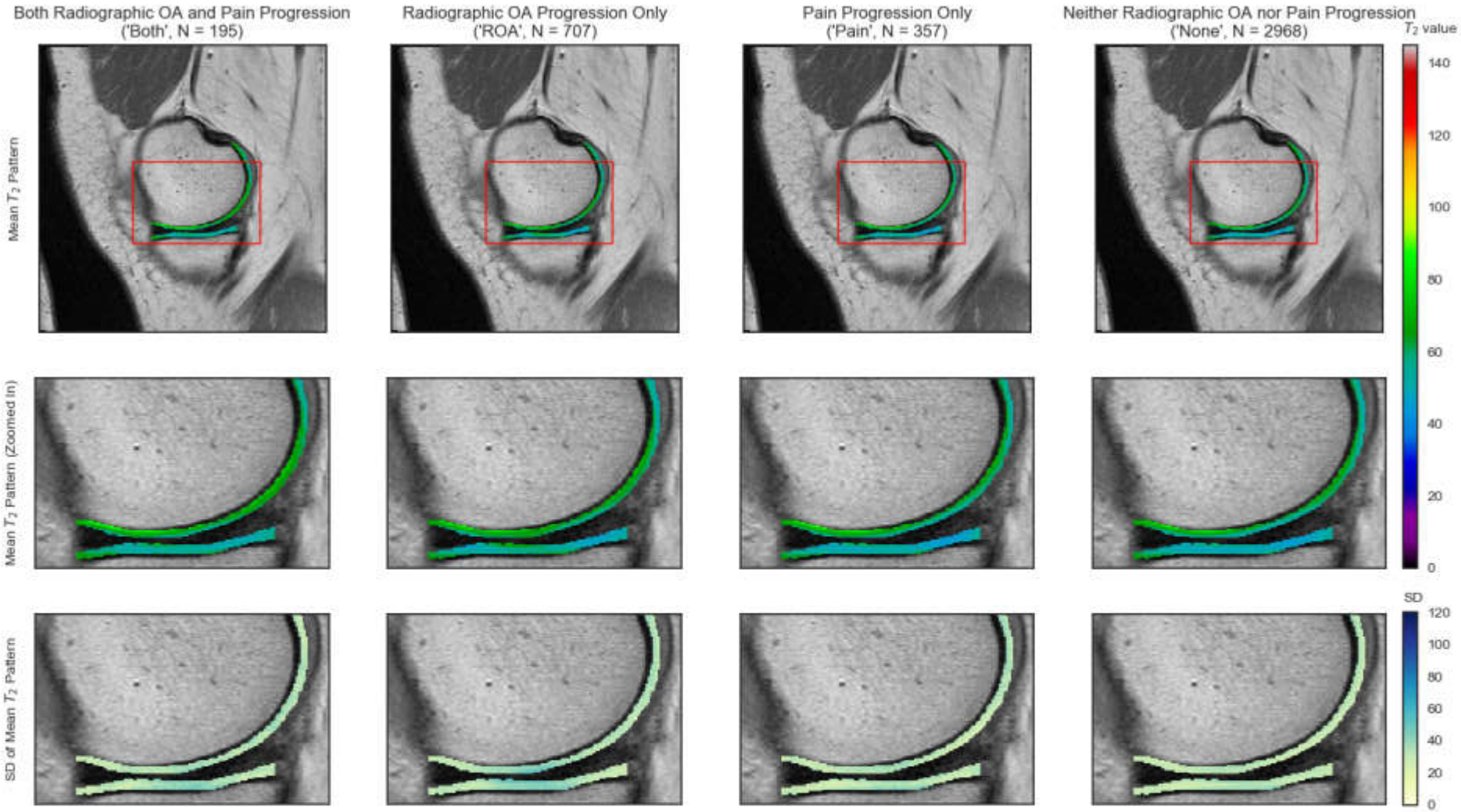


Figure 4.23: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Medial Slice (Slice 5/27[b])

[a] Cf. the patterns of the four incidence outcome groups in Fig. 4.6 on page 152

[b] This slice (5/27) is shown in the column '0' of the VBR^{T_2} one-way ANOVA result in Fig. 4.35 on page 193, but is omitted in the VBR^{T_2} two-sample t-tests results in Figs. 4.36 to 4.40 on pages 195–203 due to restricted page size.

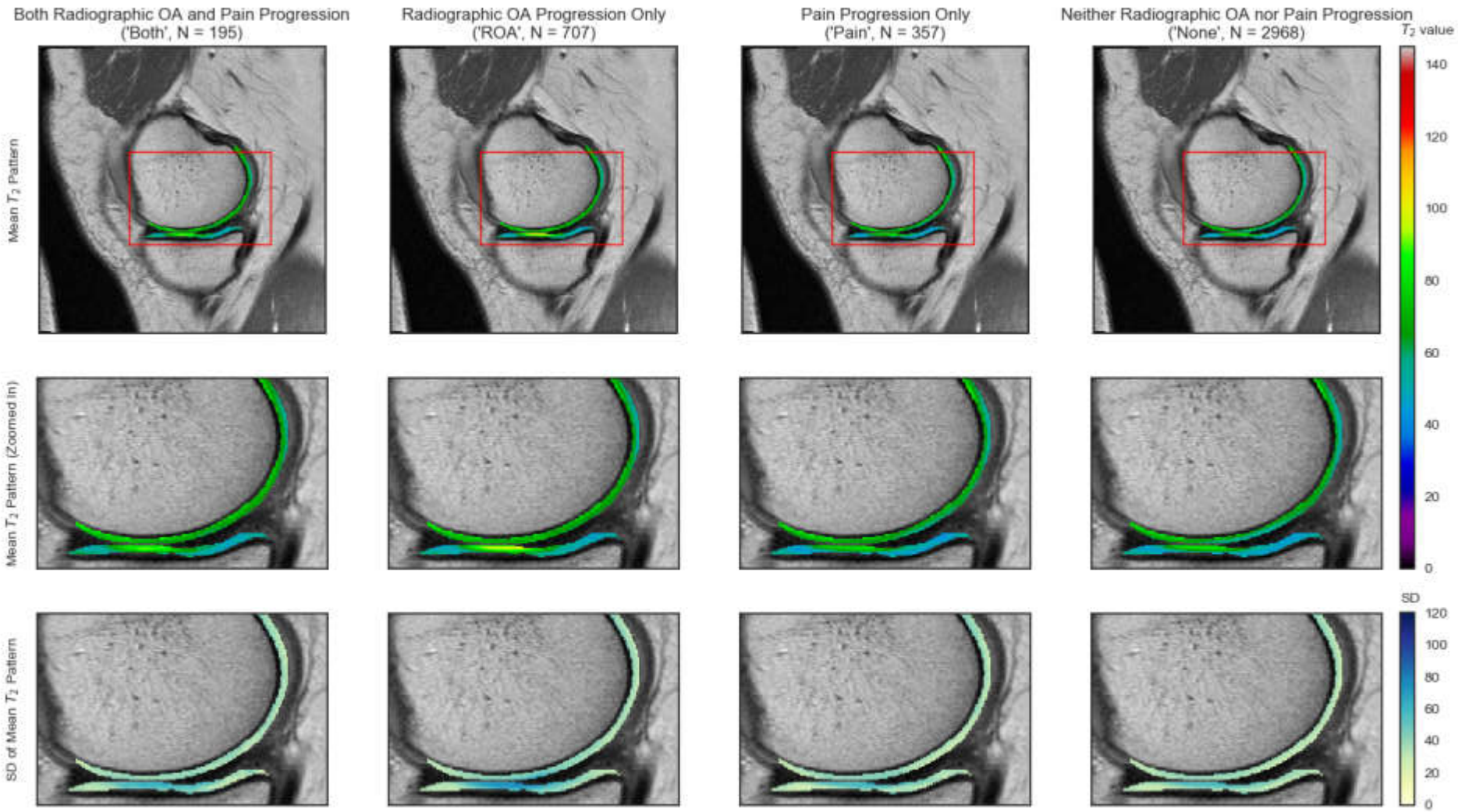


Figure 4.24: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Medial Slice (Slice 6/27[b])

[a] Cf. the patterns of the four incidence outcome groups in Fig. 4.7 on page 153

[b] This slice (6/27) is shown in columns '1's in VBR^{T_2} Figs. 4.35 to 4.40 on pages 193–203

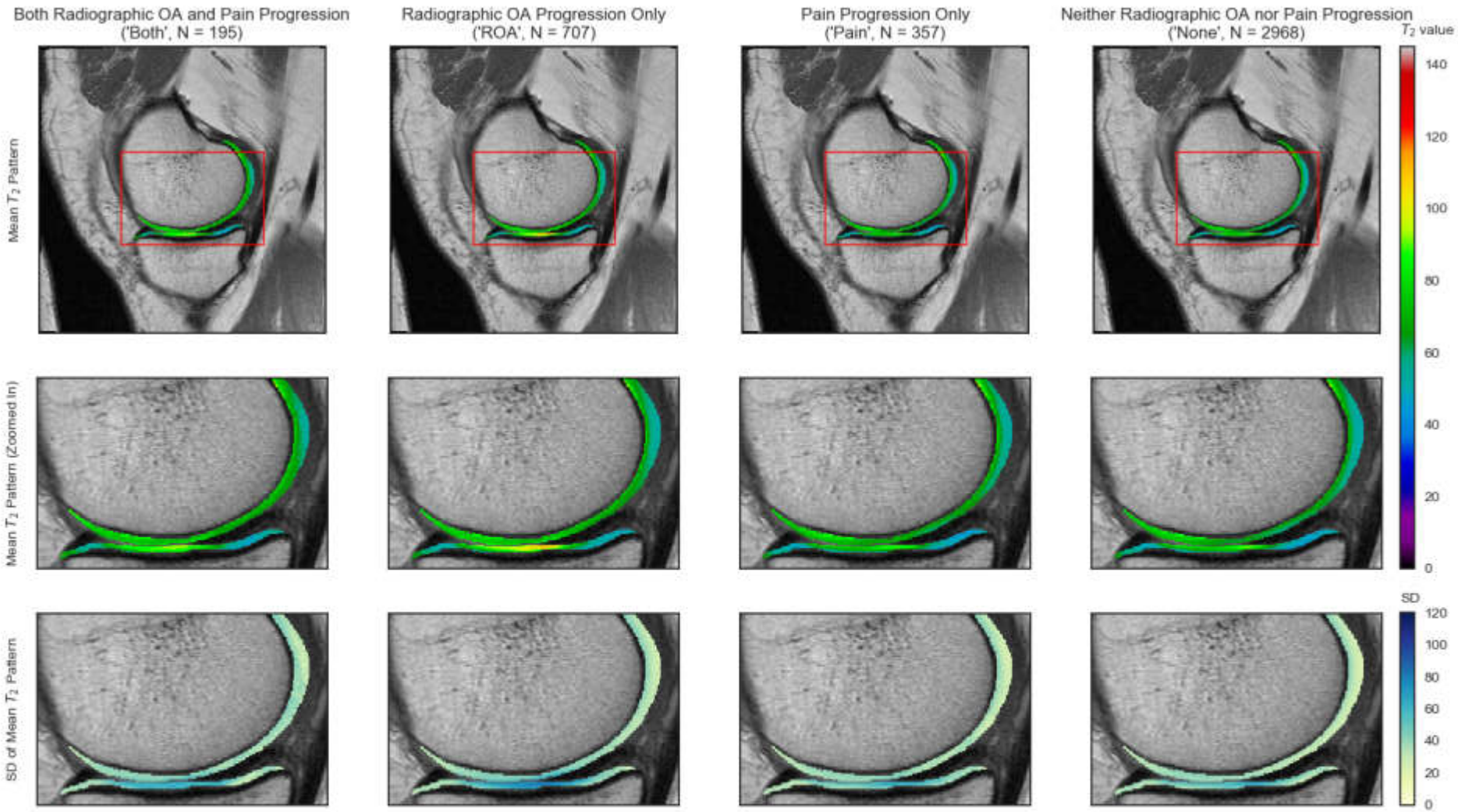


Figure 4.25: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Medial Slice (Slice 7/27[b])

[a] Cf. the patterns of the four incidence outcome groups in Fig. 4.8 on page 154

[b] This slice (7/27) is shown in columns '2's in VBR^{T_2} Figs. 4.35 to 4.40 on pages 193–203

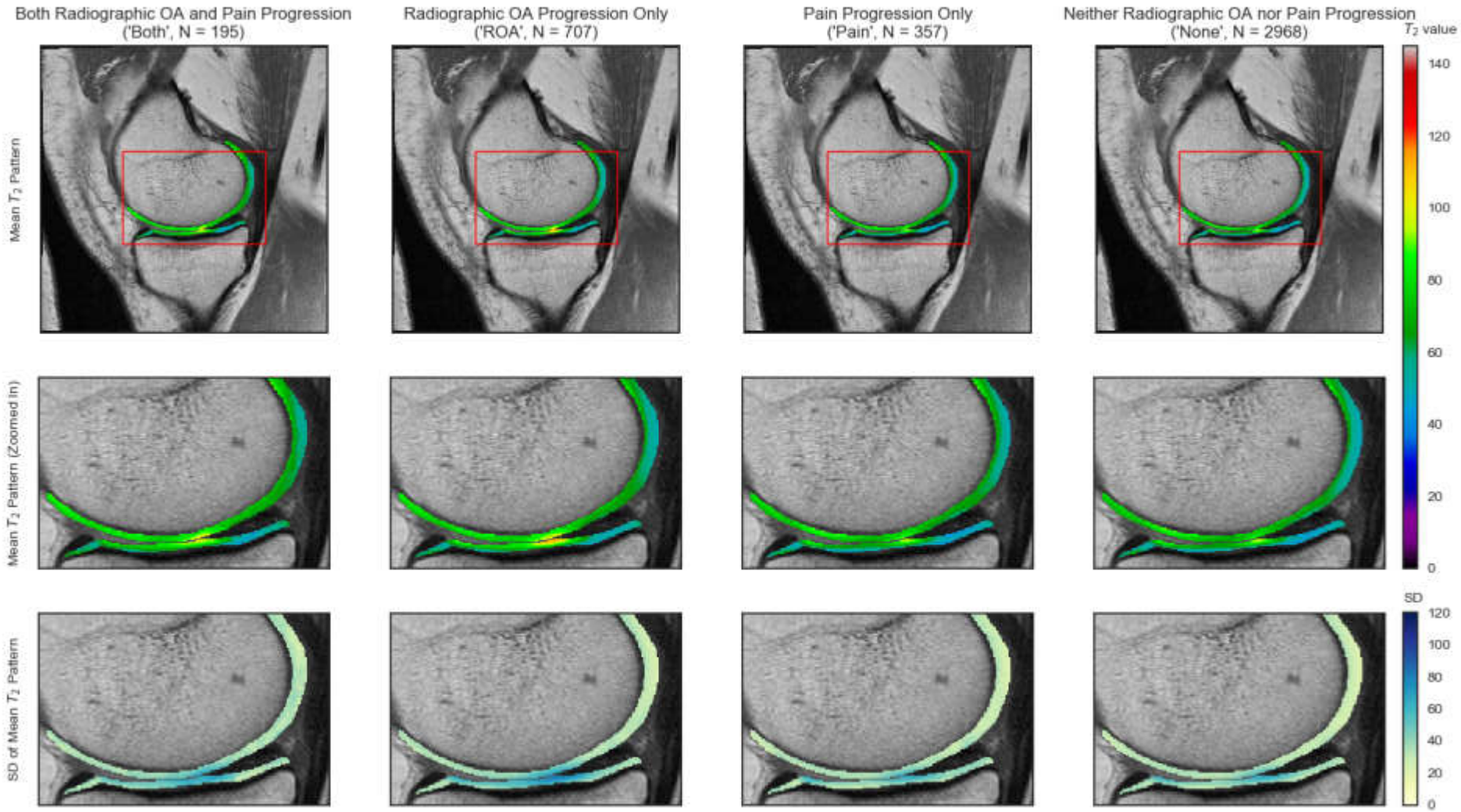


Figure 4.26: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Medial Slice (Slice 8/27)

[a]Cf. the patterns of the four incidence outcome groups in Fig. 4.9 on page 155

Patellofemoral Slices

Figs. 4.27 to 4.29 on pages 183–185 are voxel-based mean T_2 and SD patterns of the same patellofemoral slices (slices 13–15 of 27) as the incidence groups discussed in Fig. 4.10 on page 157. The overall observation of these three slices is that the 'Both' and 'ROA' groups have a larger and deeper high T_2 region (more yellow colour against green) in the surface of patellar cartilage near the patellofemoral region, compared with the 'Pain' and 'None' groups.

The laminar pattern discussed in Fig. 4.10 on page 157 is also shown in the anterior femur of the progression groups, where the superficial and deep zones with noticeable higher T_2 (in light green colour) are separated by a low T_2 zone (in dark green colour) in the middle of the femoral cartilage. Again, no obvious inter-group difference can be observed in the SD patterns in these slices among the progression groups.

Lateral Slices

Figs. 4.30 to 4.34 on pages 186–190 are voxel-based mean T_2 and SD patterns of the same lateral slices (slices 17–21 of 27) as the incidence groups discussed previously on Fig. 4.13 on page 160. Similar to the incidence groups, the higher T_2 regions in the tibia just below the anterior horn (in red and yellow colours) are also shown in these progression groups, although the inter-group difference in these spots are less obvious than their incidence counterparts.

The superficial tibial cartilage near the tibiofemoral contact region, however, displays a conspicuous higher T_2 zone (more green colour against blue) in the 'Both' group (and a smaller one in the 'ROA' group) than the 'Pain' and 'None' groups in all these five slices. The SD patterns in these regions also exhibit substantial differences between groups with ROA ('Both' and 'ROA') and groups without ROA ('Pain' and 'None')—the former clearly have a higher SD zone (more blue colour against yellow) at the superficial tibia matching the higher T_2 zone, while no such higher SD pattern is presented in the latter.

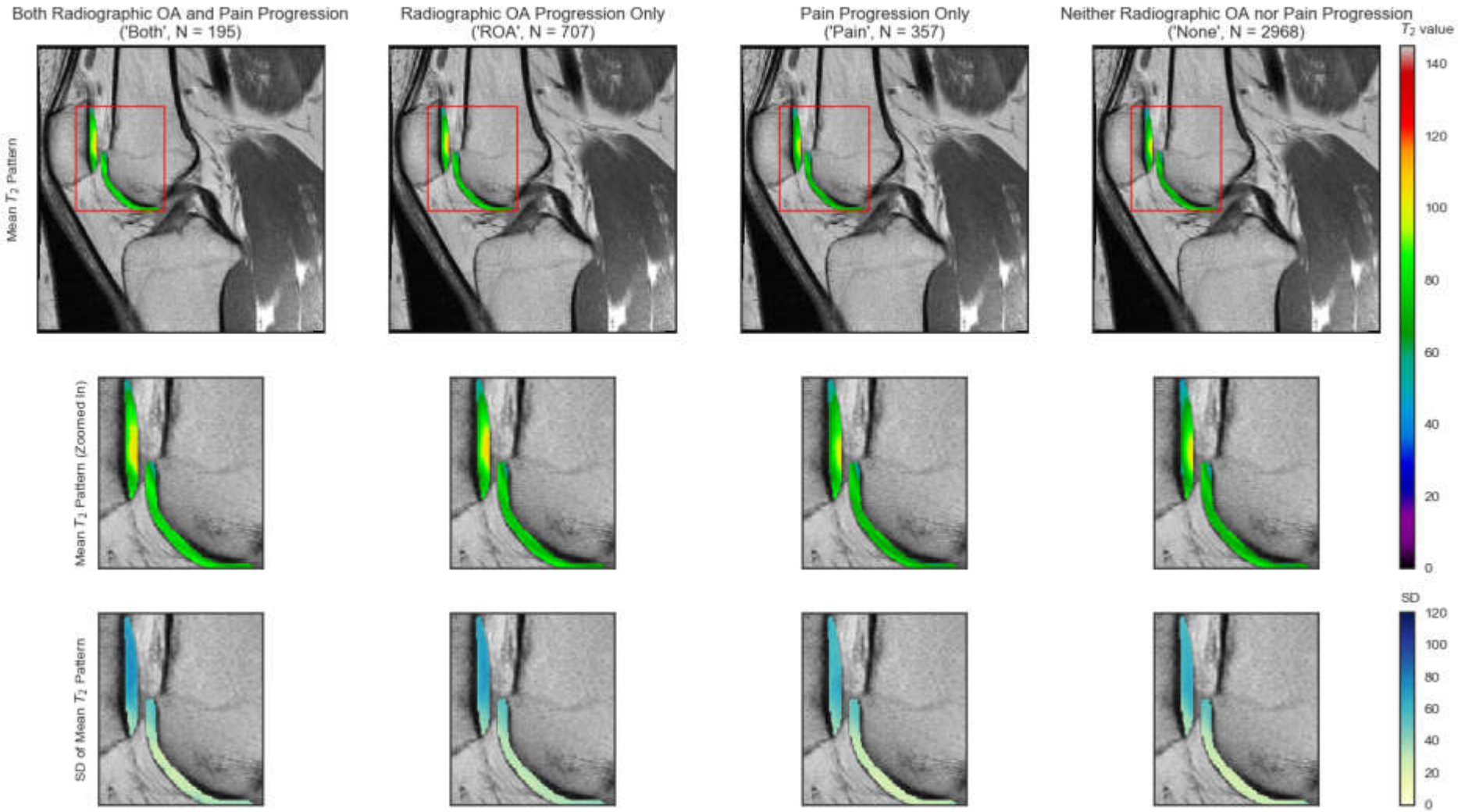


Figure 4.27: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Patellofemoral Slice (Slice 13/27)

[a]Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.10 on page 157

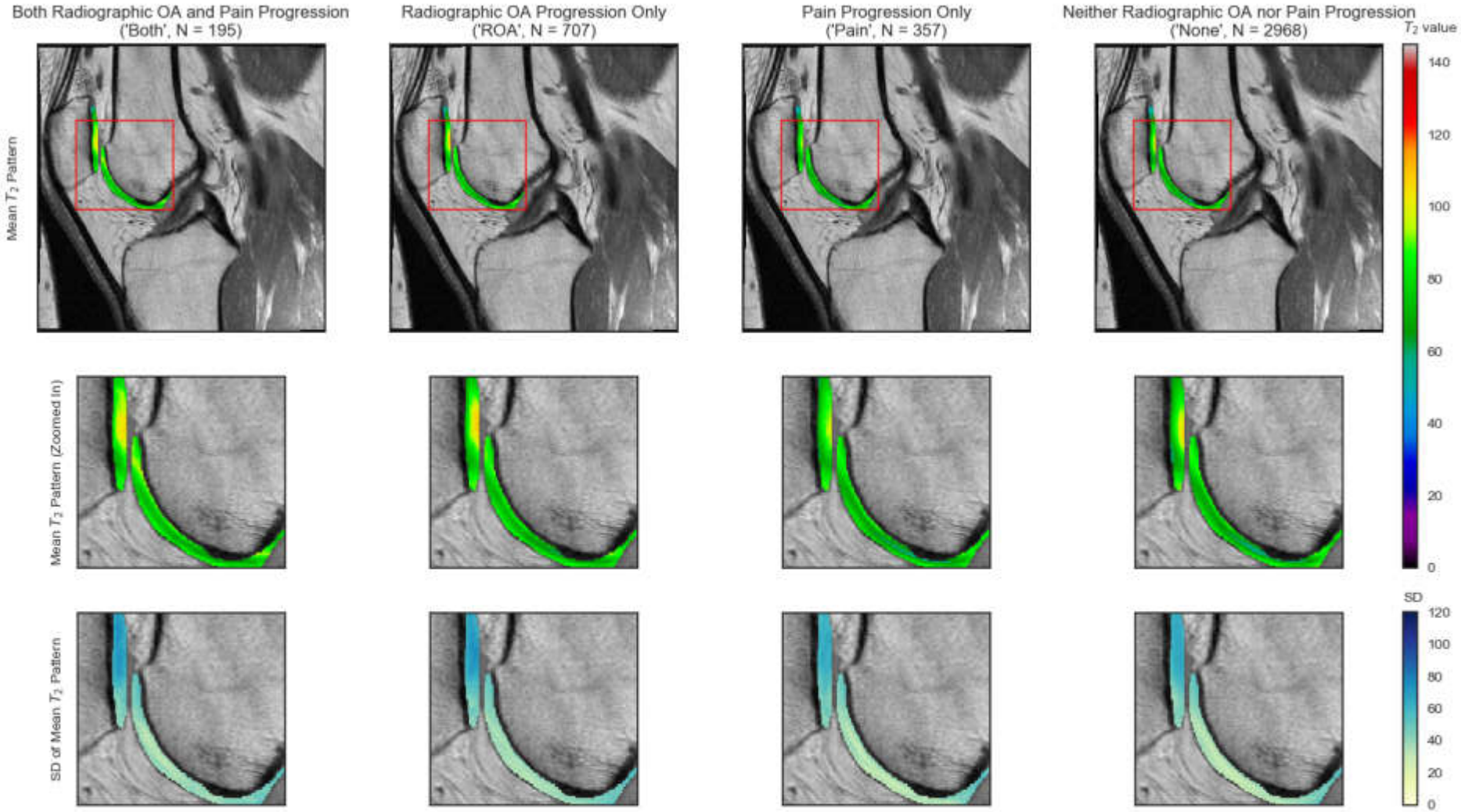


Figure 4.28: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Patellofemoral Slice (Slice 14/27)

[a]Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.11 on page 158

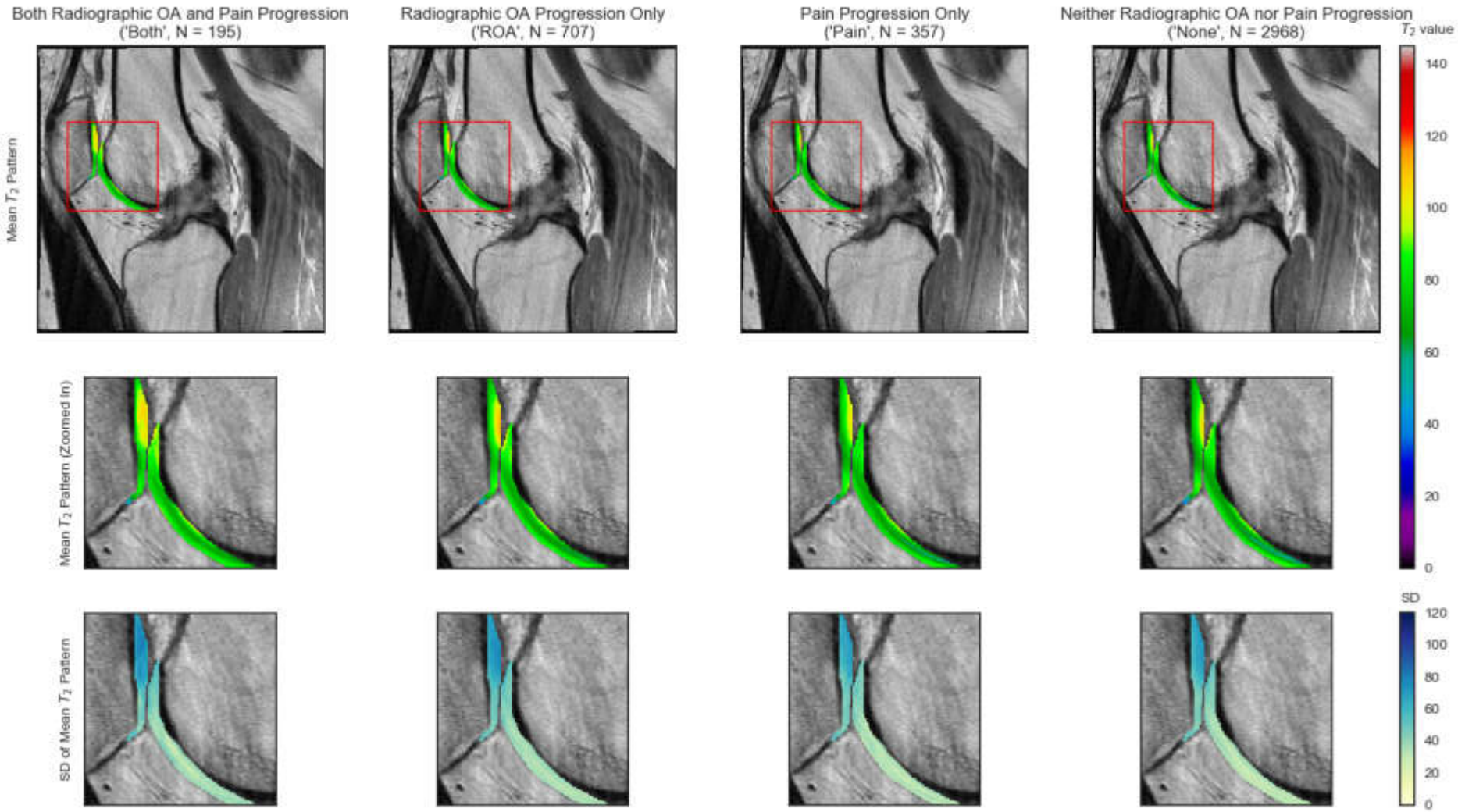


Figure 4.29: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Patellofemoral Slice (Slice 15/27)

[a]Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.12 on page 159

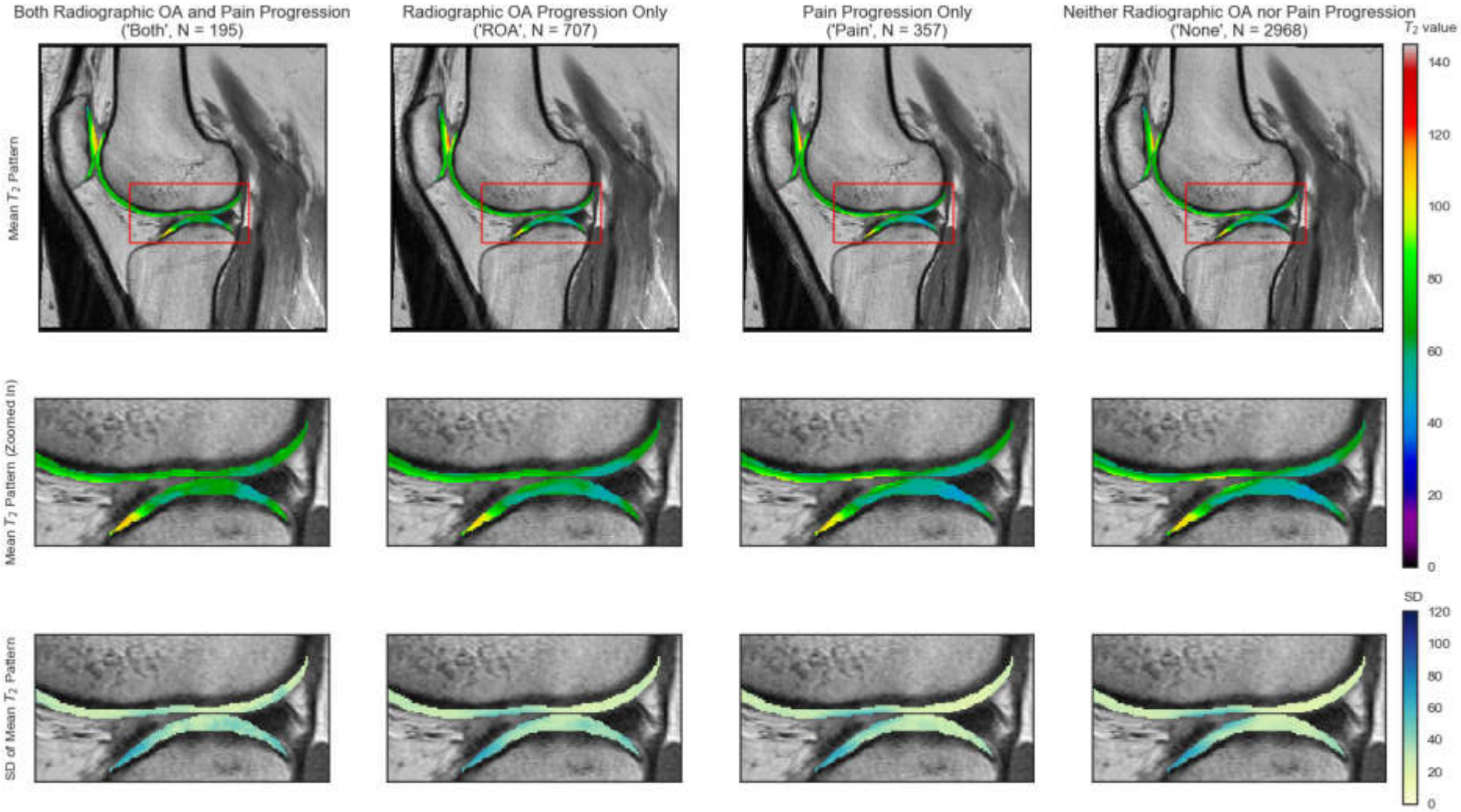


Figure 4.30: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Lateral Slice (Slice 17/27)

[a]Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.13 on page 160

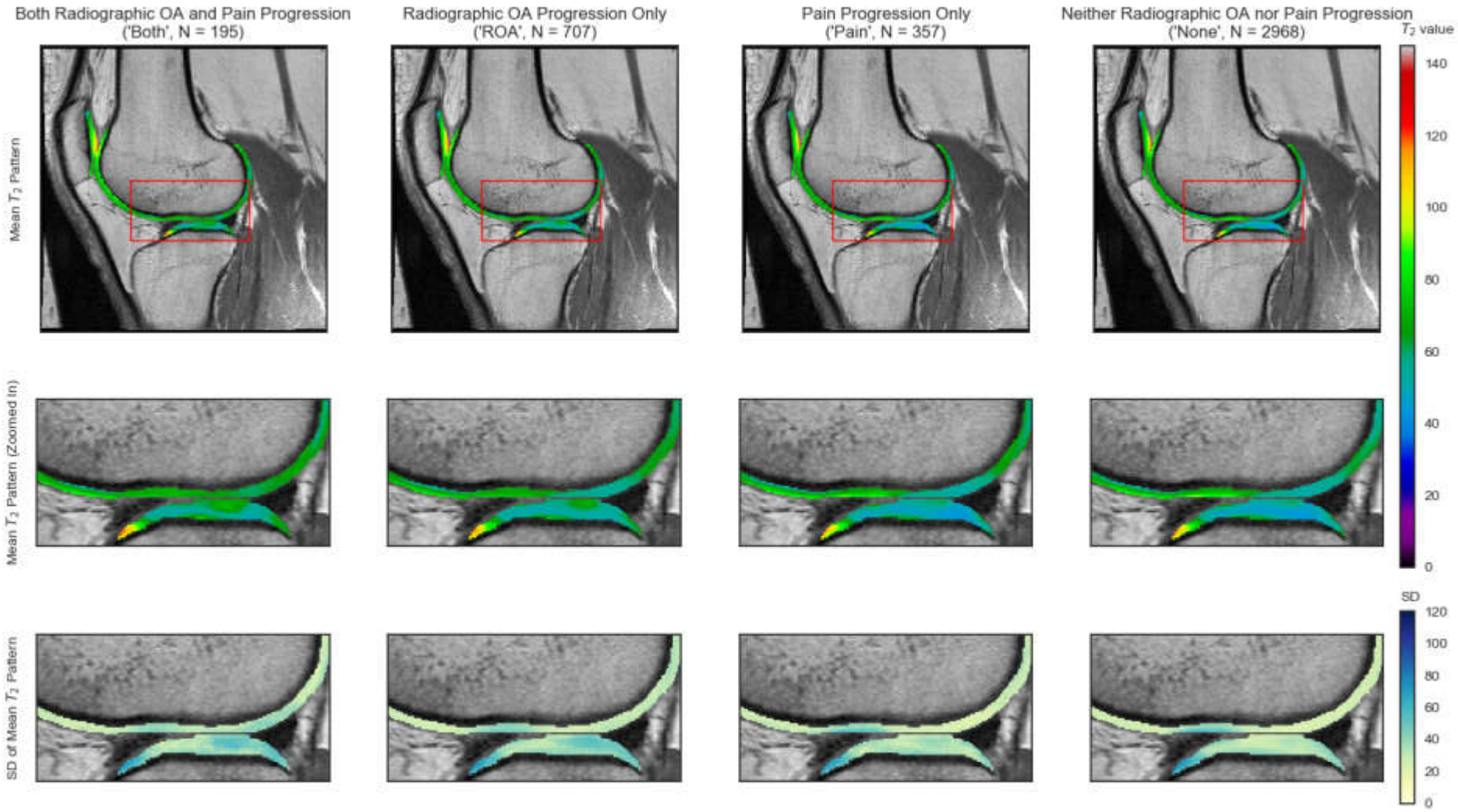


Figure 4.31: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Lateral Slice (Slice 18/27)

[a]Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.14 on page 161

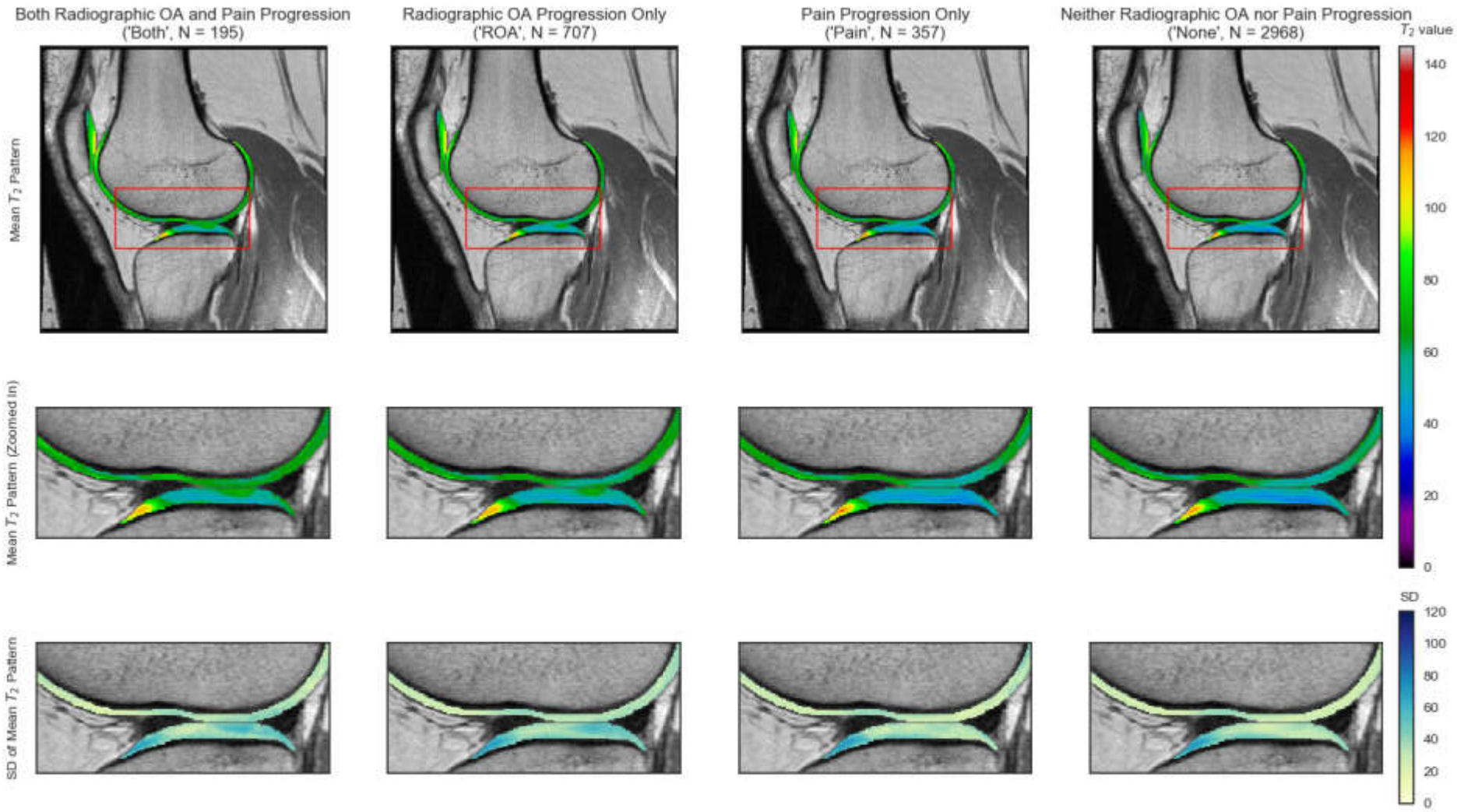


Figure 4.32: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Lateral Slice (Slice 19/27[b])

[a] Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.15 on page 162

[b] This slice (19/27) is shown in columns '3's in VBR^{T_2} Figs. 4.35 to 4.40 on pages 193–203

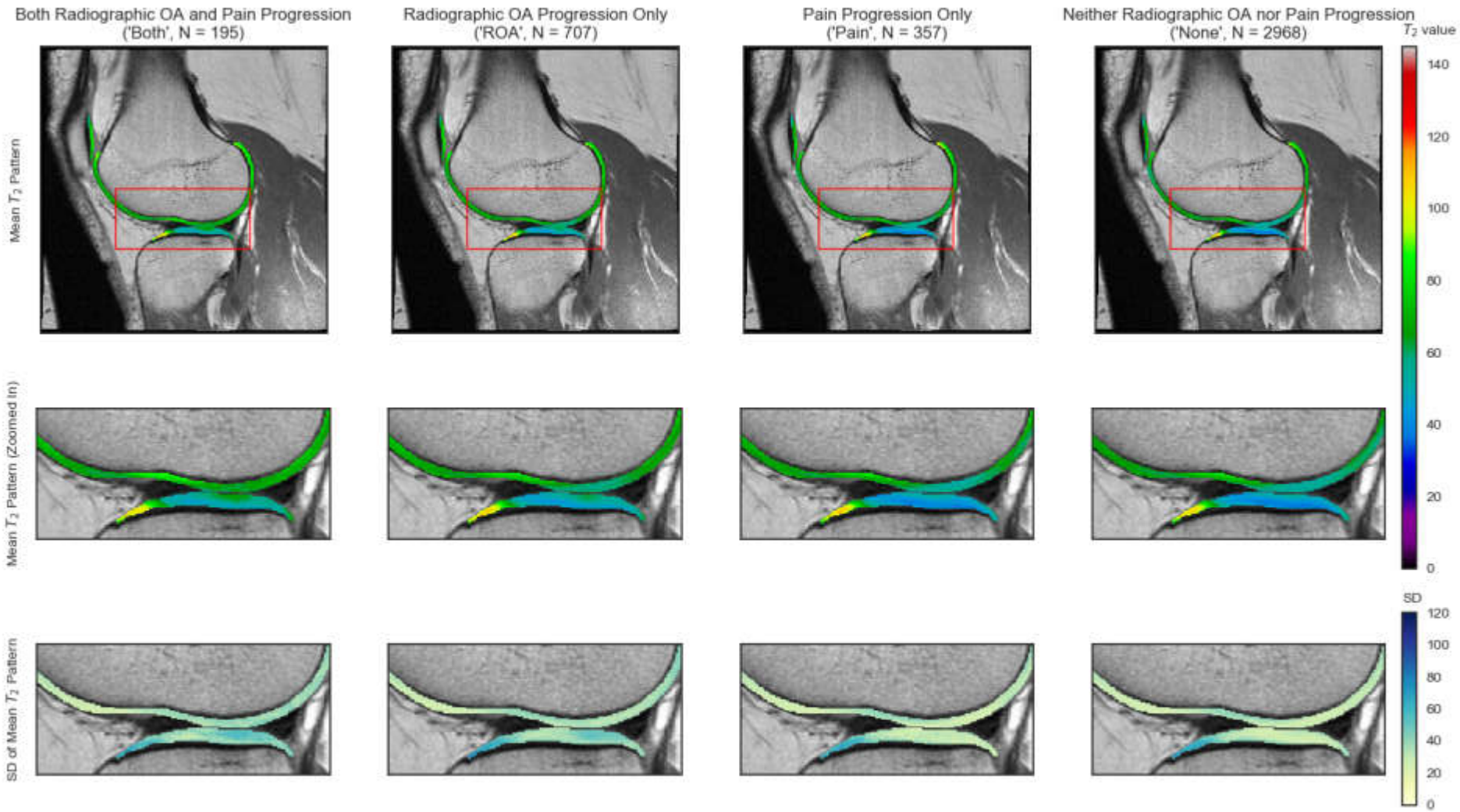


Figure 4.33: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Lateral Slice (Slice 20/27[b])

[a] Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.16 on page 163

[b] This slice (20/27) is shown in columns '4's in VBR^{T_2} Figs. 4.35 to 4.40 on pages 193–203

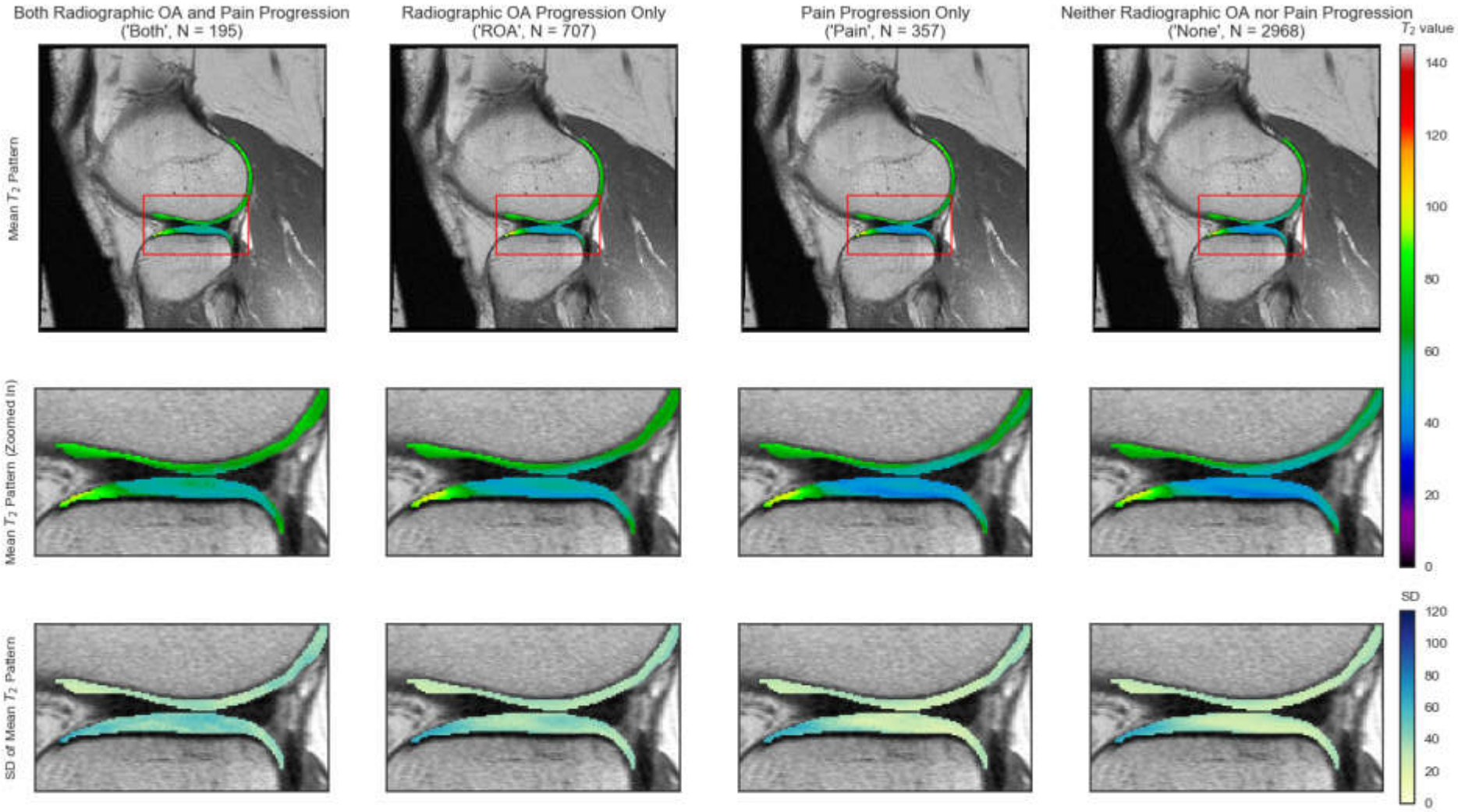


Figure 4.34: Baseline Mean T_2 and Standard Deviation Patterns of the Four Progression[a] Outcome Groups in a Lateral Slice (Slice 21/27[b])

[a] Cf. the patterns of the four incidence outcome groups in the same slice in Fig. 4.17 on page 164

[b] This slice (21/27) is shown in the column '5' of the VBR^{T_2} one-way ANOVA result in Fig. 4.35 on page 193, but is omitted in the VBR^{T_2} two-sample t-tests results in Figs. 4.36 to 4.40 on pages 195–203 due to restricted page size.

4.2.2.5 VBR^{T_2} of the Progression Outcome Groups: One-way ANOVA

Fig. 4.35 on page 193 shows the one-way ANOVA results of the four progression outcome groups in the same medial (slices 5–7 of 27) and lateral (slices 19–21 of 27) slices as the incidence groups discussed in Section 4.2.2.2 on page 165. Compared with their incidence counterparts in Fig. 4.18 on page 167, the ANOVA results of the progression groups clearly exhibit enhanced activation patterns in all slices (columns), and at all P-value correction levels (rows).

In row A of Fig. 4.35 where P-values are not corrected for FWER, 20,331 voxels are detected at $P < 0.05$, which is a massive improvement compared with the 7,690 detections in the same row of Fig. 4.18. Unlike the ANOVA results of the incidence group in which voxel detections are mainly on the lateral side, strong activations are displayed on both medial (Fig. 4.35 A0–A2) and lateral (Fig. 4.35 A3–A5) sides for the progression groups, and covered a large area over the whole femoral and tibial cartilage, rather than limited in the patellofemoral and tibiofemoral contact regions. The massive number of voxels, and the large area they covered in row A of Fig. 4.35, indicates T_2 mapping are much more (almost three times in terms of number of activated voxels) sensitive to the difference of (at least two of) the progression outcome groups, compared with the ANOVA pattern of the incidence outcome groups in the same row of Fig. 4.18 on page 167.

The ANOVA results after the Bonferroni correction of P-value are even more remarkable. As can be seen in rows A*, A** and A*** of Fig. 4.35, thousands of voxels remain activated at rigorous P-values (4,307 detections at $P < 0.05$, 3,561 detections at $P < 0.01$, and 2,832 detections at $P < 0.001$) after corrected for FWER, on both medial and lateral sides. Compared with the incidence results in Fig. 4.18 on page 167 where almost no detections were seen in the three Bonferroni corrected rows, the results here clearly shows that substantial T_2 differences exist between (at least two of) the progression outcome groups. The results are even more notable in terms of the locations of the activations—using only the T_2 maps at the baseline, and knowing nothing more than the category of the

progression outcomes in the future, the VBR^{T_2} clearly marked the tibiofemoral contact regions as the strongest activated compartments among these groups, which matches our expectations about where physiological changes should occur and also where T_2 mapping should show effects.

4.2.2.6 VBR^{T_2} of the Progression Outcome Groups: Two-sample *t*-tests

In the previous section, I performed one-way ANOVA and found that T_2 mapping has remarkable performance in differentiating the progression outcome groups. To understand specifically which groups T_2 mapping differentiates better, two-sample *t*-tests for the progression outcome groups are performed, in the same manner as for the incidence groups in Section 4.2.2.3 on page 166, as a decomposition of the one-way ANOVA results. Again, these two-sample *t*-tests results are discussed in an overall ascending order of their numbers of voxels detected, with some order adjustments to discuss patterns.

Pattern 4: Effect of 'Pain Progression | ROA Progression'

Although the 'Both' vs 'ROA' *t*-test (row A in Fig. 4.36 on page 195) between progression groups does not have the smallest number of voxels detected (4,462 voxels detected, smallest is 3,289) among all the two sample *t*-tests, I discuss this *t*-test first to match the logic, and the order of our discussion for the incidence outcome groups in Section 4.2.2.3 on page 166.

In this *t*-test, a unique activation pattern, referred to as **Pattern 4** in the following text, can be recognised as dissimilar to any of the other *t*-tests. While other *t*-tests in rows B and C of Fig. 4.36 and A–F of Fig. 4.37 on page 199 all exhibited femoral activations to some extent, this *t*-test features a small area of congregated activation only in the centre of tibia, but almost no activation in femur, on the medial side (Fig. 4.36 A1–A2).

By definition, the 'ROA' group had only ROA progression, while the 'Both' group also had a pain progression in addition to ROA progression. The difference between the two

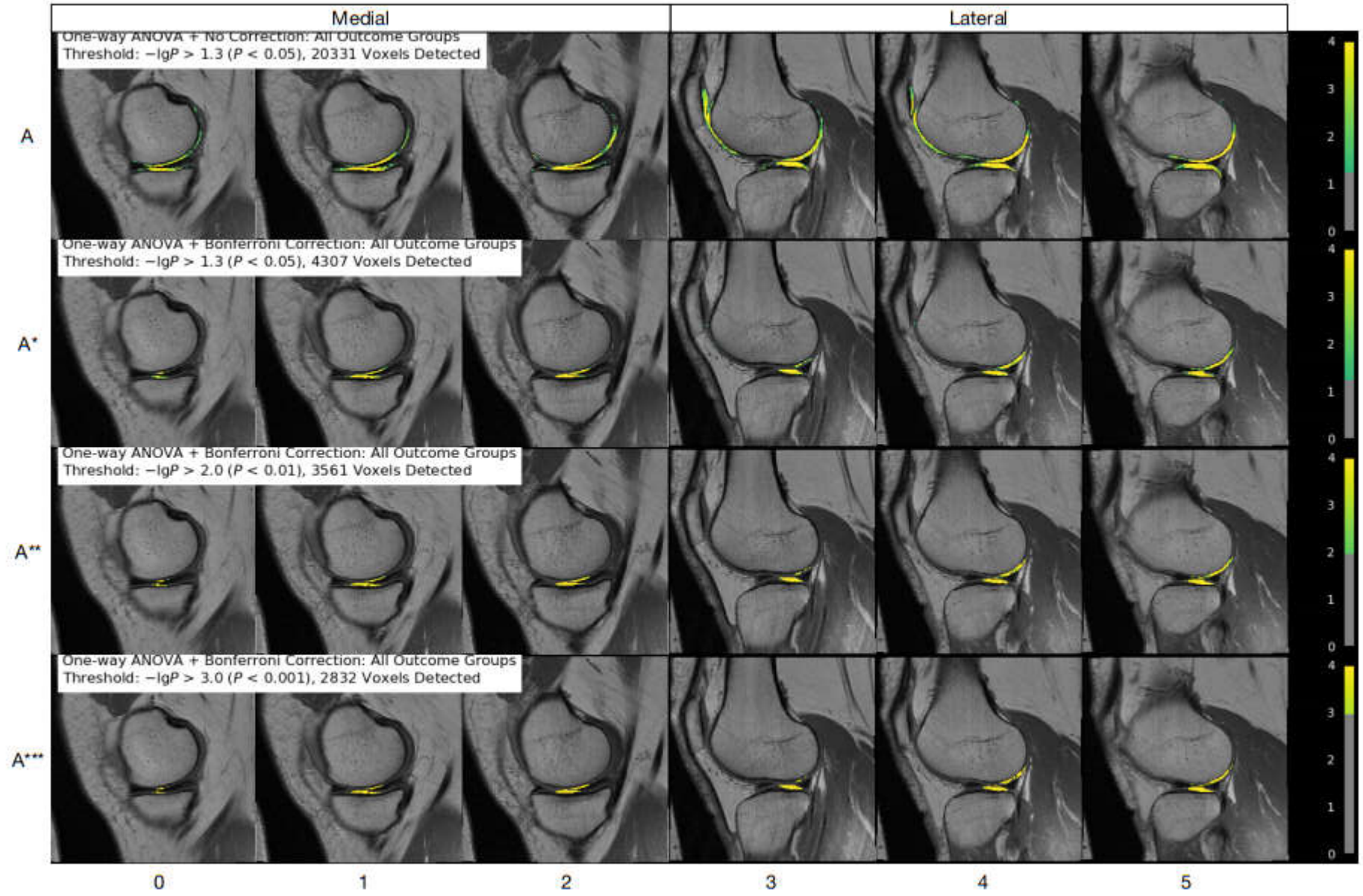


Figure 4.35: Voxel-Based T_2 Relaxometry (VBR^{T_2}): One-way ANOVA of All Progression[a] Outcome Groups

[a]Cf. VBR^{T_2}: One-way ANOVA of All Incidence Outcome Groups in Fig. 4.18 on page 167

group, shown as voxel detections (activations) in the two sample t-test, is therefore related to pain progression that comes with (or complicated by) ROA progression, denoted 'Pain Progression | ROA Progression'.

Unlike its counterpart for the incidence groups, this test displayed a relatively large number of voxels activated (4,462 detections), a level comparable to that of the Pattern 2 (approximately 4,000–5,000 detections, see Fig. 4.21 on page 175). This proved, therefore, that it is possible to use baseline T_2 mapping to differentiate a ROA only progression outcome from a 'Both' progression outcome. It means T_2 mapping can be used to see the effect of 'Pain Progression | ROA Progression'. If we are given the baseline T_2 map of a subject whom we are certain will have only ROA progression (determined with or without T_2 mapping), we will be able to use this T_2 map to further predict if the subject will progress with ROA only, or with (or complicated by) pain progression as well.

After Bonferroni correction, only one voxel is detected in this t-tests at $P < 0.05$ (row A* in Fig. 4.36), which essentially is the same as its incidence counterpart where no voxel is detected (row A* in Fig. 4.19 on page 169).

Pattern 5: Effect of Pain Progression

Similar to their counterparts in the incidence outcome groups in which we identified Pattern 1, the t-tests of 'Pain' vs 'None', and ('Both'+'Pain') vs ('ROA'+'None') for the progression outcome groups, in rows B and C of Fig. 4.36 on the next page, exhibited another pattern that should be discussed separately. This is referred to as **Pattern 5** in the following text.

As presented in Fig. 4.36, rows B and C of Fig. 4.36 on the facing page both displayed congregated activation locations in central–posterior femur, but minimal detections in tibia on the medial side (Fig. 4.36 B1–B2, C1–C2). The lateral side, on the other hand, does not show any obvious unique features when compared against other t-tests. Such a femur-oriented activation pattern is unique to these two t-tests.

From the definition of the groups, the t-tests in these two rows both tested a pain progres-

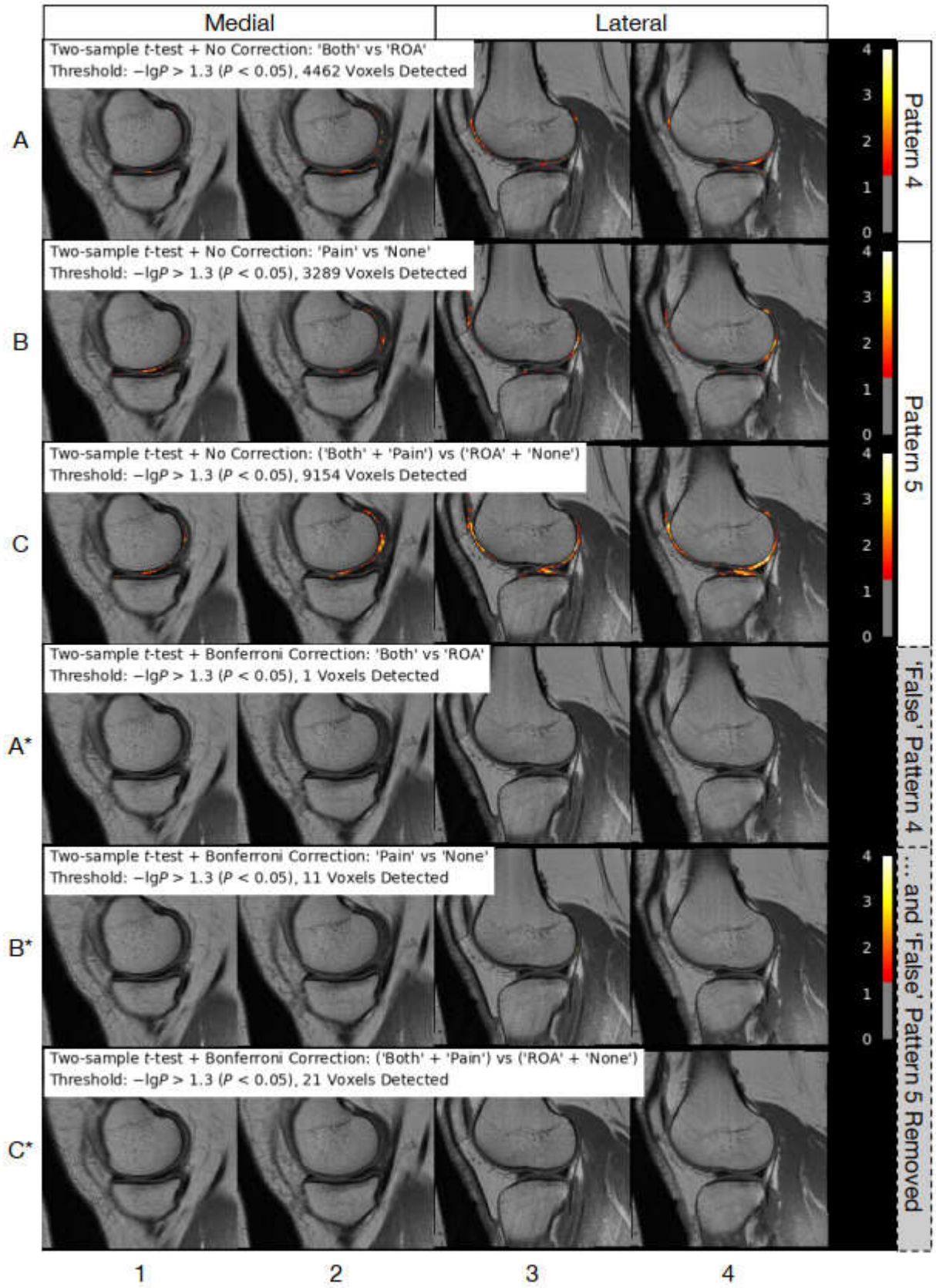


Figure 4.36: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample t-tests Between Pain Progression and No Pain Progression[a] Groups. Rows A–C: No P-value Correction; Rows A*–C*: After Bonferroni Correction.

[a]Cf. VBR^{T_2}: Two-sample t-tests Between Incident Pain and No Incident Pain Groups in Fig. 4.19 on page 169

sion group, individually ('Pain') or combined ('Both'+'Pain'), to a individually ('None') or combined ('ROA'+'None') no pain progression group. Therefore, it is possible that such a pattern is related to the effect of pain progression, relative to a no pain progression outcome.

However, these two t-tests are very different to each other in terms of the numbers of voxels detected—clearly row C (9,154 detections) has an elevated Pattern 5 compared to row B (3,289 detections), despite their similar activation locations. If we think of row C of Fig. 4.36 on the previous page as the 'sum' of rows A and B, then the enhanced Pattern 4 in row C indicates that the dissimilarity between 'Both' and 'ROA' in Pattern 4 (4,462 detections) made a great contribution to the resulting, much stronger dissimilarity between ('Both'+'Pain') and ('ROA'+'None'). This outnumbered rows A and B individually by more than two to one (9,154 vs 4,462, and 9,154 vs 3,289). However, such a contribution only elevated the number of voxels detected, but did not change the general locations of the voxel clusters in Pattern 5.

Recall in Section 4.2.2.3 on page 168 I concluded that incident ROA will prevent the effect of incident pain (Pattern 1) being observed in T_2 maps, as the t-tests between combined groups (row C of Fig. 4.19 on page 169) showed less activated voxels than that between individual groups. The interesting effect of combining the groups have shown again here for the progression groups, but in a different direction—the effect of pain progression (Pattern 5) in row B of Fig. 4.36 on the preceding page can be enhanced when samples with ROA progression in Row A are involved in the t-tests. What this means is that we can use T_2 mapping to predict a pain progression outcome at the baseline, regardless of the ROA status. However, it would be much easier to observe the effect of pain progression (Pattern 5) if the samples have ROA progression as well.

After Bonferroni correction, Pattern 5 is also eliminated in row B* and C*, indicating when FWER is considered, pain progression, like incident pain, is still a challenging target to image and follow up with T_2 mapping. It is unlikely, however, that the patterns we recognised are all false positives—careful interpretation of Bonferroni corrected results is

again needed due to its extremely conservative nature.

Pattern 6: Effect of ROA Progression

The final voxel activation pattern recognised in this study, referred to as **Pattern 6** in the following text, is the pattern produced by six t-tests, presented in rows A–F of Fig. 4.37 on page 199. Although the pattern in rows A–B showed fewer voxels detected, thus smaller clusters than rows C–F, their overall activation locations are very similar and should be classified as one pattern. On the medial side, Pattern 6 can be characterised as very strong activation in the whole tibia, and slightly less strong activation in central–posterior femur (Fig. 4.37 columns 1–2); on the lateral side, strong activations covered almost the whole patellofemoral and tibiofemoral contact regions, through posterolateral femur and tibia, whereas anterior tibia displayed weak activations (Fig. 4.37 columns 3–4).

These t-tests have one thing in common by their definition—they all tested an individually or combined ROA progression group against an individually or combined no ROA progression group. Pattern 6 is therefore related to the effect of ROA progression relative to knees of no ROA progression.

There are some interesting observations though, when the t-tests in this group are compared against each other. Recall in the discussion of Pattern 4 the difference between 'Both' and 'ROA' groups is substantial (4,462 detections), and is expected to relate to the effect of 'Pain Progression | ROA Progression'. Now we have an opportunity to inspect whether such an effect between these two outcomes will remain when they are both tested against a common sample. As can be seen in Fig. 4.37, row A and row B tested the 'Both' and 'ROA' outcome groups against the 'Pain' outcome group individually, and they have shown a clear difference in number of voxels detected (6,952 vs 11,156). This means the said effect indeed remained, or in other words, was still important when tested against the 'Pain' outcome. Otherwise, these two t-tests should show similar numbers of voxels detected. By contrast, if we examine row C and row D in which 'Both' and 'ROA' were tested against the 'None' group individually, it can be seen the number of voxels

detected in these two rows are very similar (18,033 vs 18,206). This in turn indicates that the said effect is no longer important when tested against a 'None' outcome. In common language, the pain progression coming with (or complicated by) ROA progression is different from a ROA only progression outcome when examined using T_2 mapping (this is the conclusion for Pattern 4); however, this difference (Pattern 4) is still small compared to the strong effect of ROA (Pattern 6). Depending on the samples selected (subjects recruited/included), the said effect can be masked by stronger patterns and might not be observed.

The same logic can be used to understand how Pattern 5 can be observed or be masked. As can be seen in Fig. 4.37, row A and row D tested the 'Pain' and 'None' outcome groups against the 'ROA' outcome group individually, and the difference in voxels detected is large (6,952 vs 18,206, difference is 11,254). However, when 'Pain' and 'None' were tested against the 'Both' group individually in rows B and C, the difference narrowed (11,156 vs 18,033, difference is 6,877). This means the difference between 'Pain' and 'None' (Pattern 5) also become less important when they are tested against different outcomes.

The key message here is that if one is interested in the effect of Pattern 4 and 5 and wanted to examine them in T_2 maps, they must carefully select the samples they want to use. Failure to do so would mean these weak effects be masked by the much stronger effect in Pattern 6.

Furthermore, using the logic used before, row E can be regarded as the 'sum' of rows B and D. By combining the groups one essentially are only considering the ROA progression, irrespective of the pain outcome. Row E activated more voxels than either row B or row D (21,638 vs 11,156, and 21,638 vs 18,206). Again this is showing that the dissimilarity between groups are somewhat 'cumulative' when they are combined. The 'cumulative' nature of the effects, however, is not a simple additive, as a single voxel can be related to more than one outcomes and can be activated in different patterns. Therefore, the relationship among these outcome groups are much more complex than one would have thought. For example, Row F, which exhibited the largest number of voxels detected

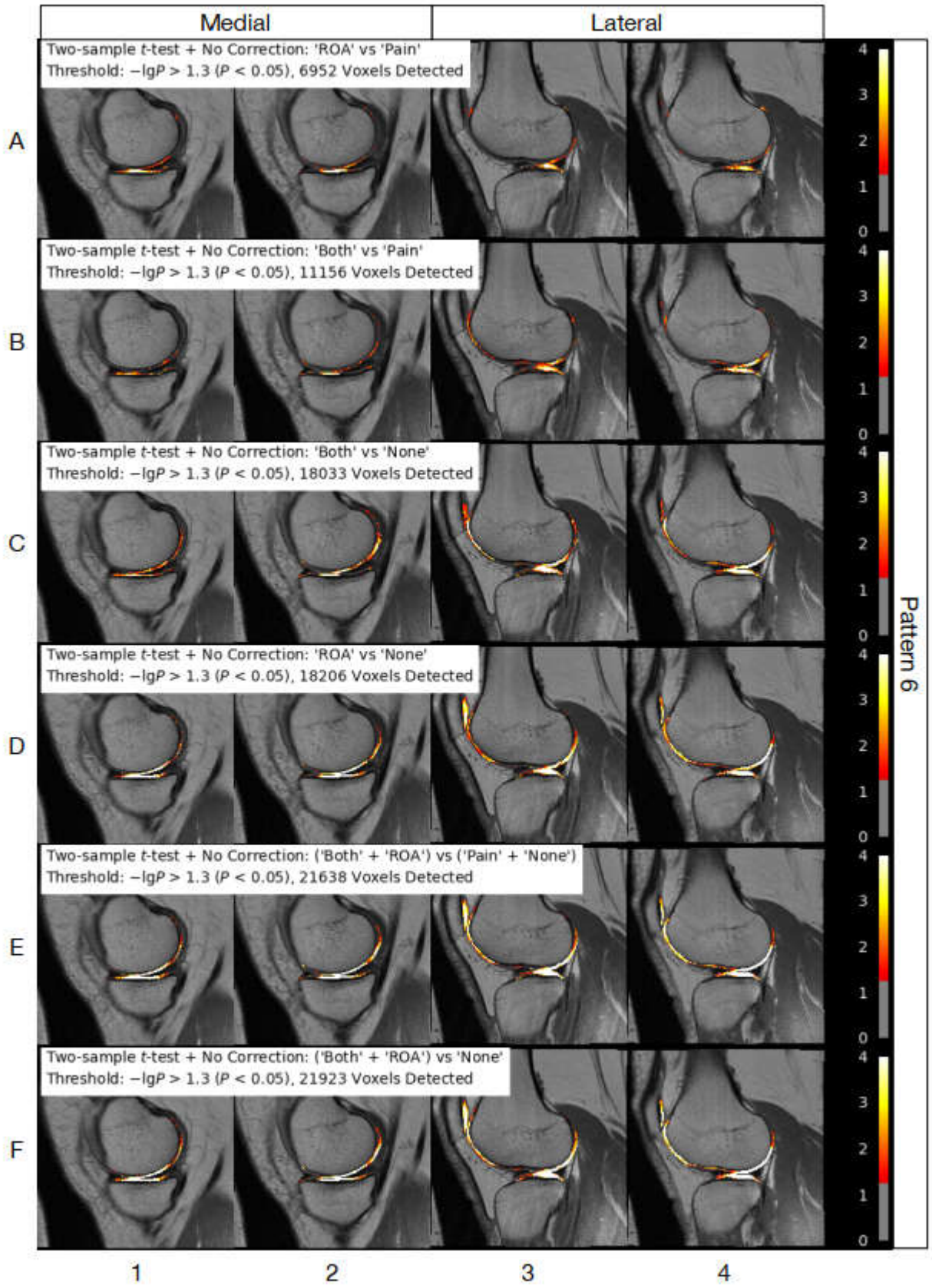


Figure 4.37: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample t-tests Between ROA Progression and No ROA Progression[a] Groups (No P-value Correction)

[a]Cf. VBR^{T_2}: Two-sample t-tests Between Incident ROA and No Incident ROA Groups (No P-value Correction) in Fig. 4.21 on page 175

(21,293) in this chapter, was produced by removing the 'Pain' outcome from the t-tests in row E. Without a big picture of how far away these groups are separated from each other, it would be difficult to understand the result in row F—the highest activation exhibited among all t-tests. This is solved by producing a number of voxels detected distance map, which will be discussed in the next section.

After Bonferroni correction at $P < 0.05$, Pattern 6 shown a remarkable survival in four of its six rows (rows C*–F* of Fig. 4.38 on the next page). As can be seen, although there were dramatic falls in activated voxels in these rows compared to if FWER was not controlled, the Pattern 6 retained in the tibiofemoral contact regions on both medial and lateral sides. More remarkably, these activations are so significant that they even survived the extremely conservative Bonferroni correction at $P < 0.01$ (rows C**–F** of Fig. 4.39 on page 202) and $P < 0.001$ (rows C***–F*** of Fig. 4.40 on page 203), with reasonable fall in numbers but without dramatic change in activation locations.

These results show T_2 mapping is a powerful tool to predict ROA progression outcome, at the earliest possible time (baseline). The flexibility of the definition of progression in this study, which included both diseased and healthy subjects in the baseline, also ensured the ROA progression could be predicted by using any T_2 map as the baseline.

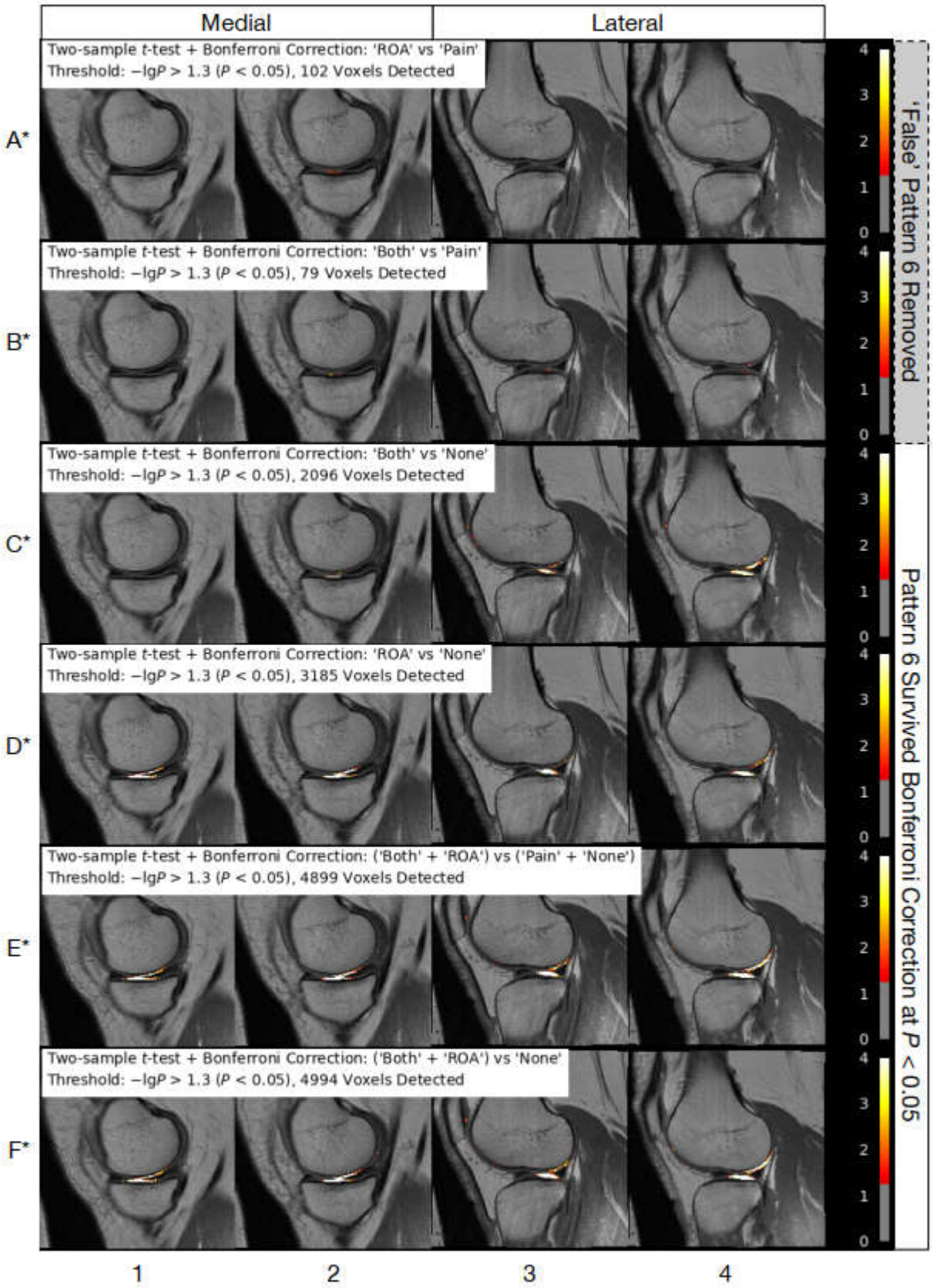


Figure 4.38: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample *t*-tests Between ROA Progression and No ROA Progression[a] Groups (After Bonferroni Correction and at $P < 0.05$)

[a]Cf. VBR^{T_2}: Two-sample *t*-tests Between Incident ROA and No Incident ROA Groups (After Bonferroni Correction) in Fig. 4.20 on page 173

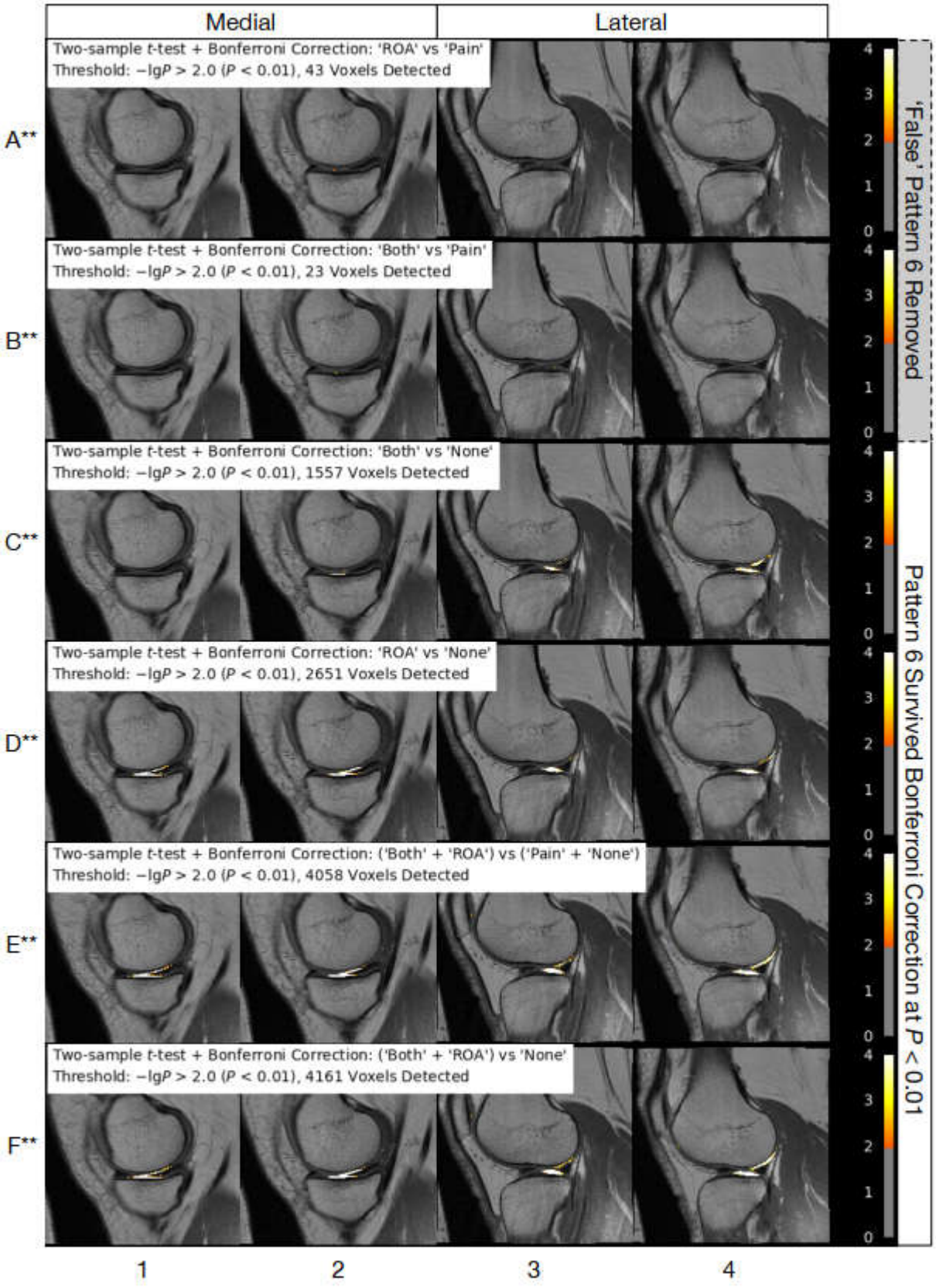


Figure 4.39: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample t-tests Between ROA Progression and No ROA Progression Groups (After Bonferroni Correction and at $P < 0.01$)

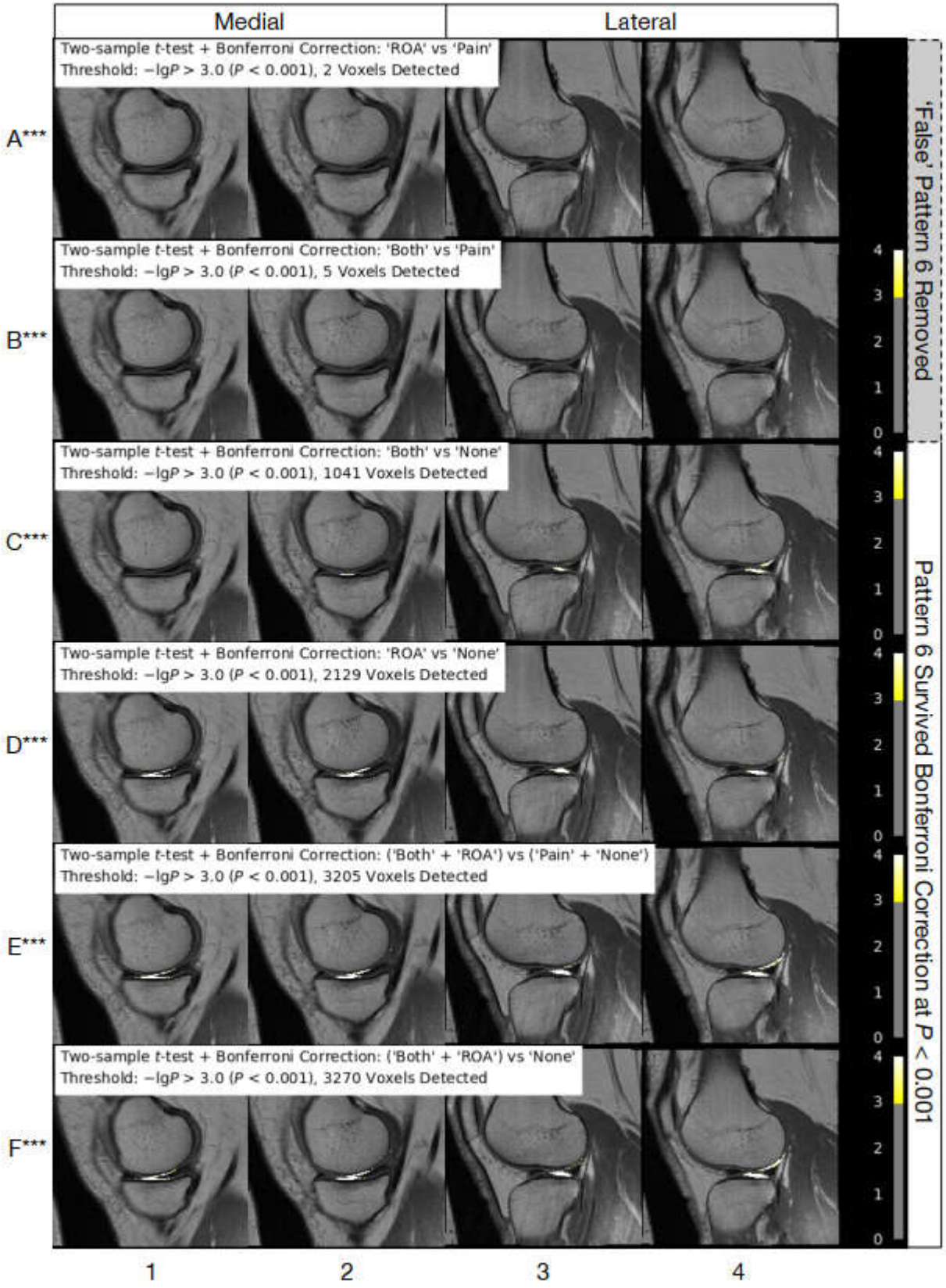


Figure 4.40: Voxel-Based T_2 Relaxometry (VBR^{T_2}): Two-sample t-tests Between ROA Progression and No ROA Progression Groups (After Bonferroni Correction and at $P < 0.001$)

4.2.2.7 Summaries and Interpretations

Summary of Voxel-Based T_2 Relaxometry

The numbers of voxels detected in the one-way ANOVA and two sample t-tests, and the six patterns we recognised from the VBR^{T_2} activations, are summarised in Tables 4.8 and 4.9 on page 206 and on page 207.

Comparing the one-way ANOVA results in the tables, the voxel detections in the progression outcome analysis substantially outnumbered those of the incidence outcomes, which means T_2 mapping is better for detecting/predicting progression outcomes than incidence outcomes.

Breaking the ANOVA result down, the two-sample t-tests further confirmed that T_2 mapping acquired at the baseline are related to the outcome over/after the 96-month follow-up period, which is demonstrated quantitatively by the numbers of voxels detected, and visually by the locations and distributions of these voxels in the activation maps.

By grouping similar activation maps together, six unique voxel activation patterns are recognised in the pairwise two-sample t-tests between outcome groups. The difference in the definitions of these individually or combined outcomes, and the mutual definition they share, were then logically contrasted/associated to explain the origins of the differences, and similarities in their T_2 maps, i.e. the patterns.

Having assigned these patterns to different outcomes, I further performed the same t-tests but controlled for the FWER with a Bonferroni method. Although a large number of the detected voxels and recognised patterns are removed, or even eliminated by the Bonferroni correction, the findings are examined at a much more conservative level. Remarkably, the patterns for progression outcomes, especially Pattern 6 showed considerable significance after the Bonferroni correction, despite the extremity of the correction approach. This again confirmed the exceptional performance of T_2 mapping in differentiating the OA progression outcomes.

Interpretations of the Patterns

There are three perspectives on viewing the patterns:

(1) The most conservative way to view the patterns is to consider them solely as occurred changes at the baseline, that could be used as a predictor of the related outcome. However, these changes are not considered as the origin of the final outcome, in terms of the location and magnitude of changes. These patterns can predict the final outcome only because they appeared in the t-tests between the outcomes, yet we cannot directly say the regions that have shown activation are exactly the sites of change in the final outcome, as the disease could have a mechanism to show early difference at a location other than the final location of lesion. Nevertheless, this interpretation enabled us to look for these patterns in any new T_2 scan (treated as baseline), especially those acquired before any symptom is shown, to predict an outcome in the future.

(2) A slightly less conservative way to view the patterns is to acknowledge these voxels are occurred changes at the baseline, yet also assume these occurred changes themselves are part of the early phases of the disease, and therefore a precursor of the outcome. This is the basis of using the patterns to describe the characteristics of OA at its early phase in a quantitative manner.

(3) The least conservative, but still reasonable, way to view the patterns is to treat these patterns as the early changes of the disease. We also view them as the best estimation of the changes expected to occur in the future and in the eventual outcome. In this way, the T_2 maps captured in a cross-sectional point of time become the closest estimator to a future event longitudinally. These early changes can then be treated as a preview of the outcome, which forms the basis of using these patterns to model the mechanism of OA initiation and progression.

Table 4.8: Summary of VBR^{T_2} for the Incidence Outcome Groups[a]

Voxel-based T_2 Relaxometry (VBR^{T_2}) Tests	Number of Voxels Detected				Voxel Activation Pattern ($P < 0.05$, No Correction)		Related Outcome
	No Correction	After Bonferroni Correction					
	$P < 0.05$	$P < 0.05$	$P < 0.01$	$P < 0.001$	Medial Side	Lateral Side	
One-way ANOVA							
All outcome groups	7,690	49	18	0	-	-	-
Two-sample *t*-tests							
'Both' vs 'ROA'	1,222	0	0	0	-	-	-
'Pain' vs 'None'	3,928	0	0	0			Incident Pain Only
('Both'+'Pain') vs ('ROA'+'None')	3,060	1	0	0			
'ROA' vs 'None'	3,919	18	5	0			Incident ROA Only
'ROA' vs 'Pain'	5,065	2	0	0			
'Both' vs 'None'	5,575	2	1	0			Incident ROA and Pain Combined
'Both' vs 'Pain'	7,008	15	6	0			
('Both'+'ROA') vs ('Pain'+'None')	8,100	130	50	9			

[a]See also Fig. 4.43 on page 212 in which the same results are visualised to better show relationships among the incidence outcome groups.

Table 4.9: Summary of VBR^{T_2} for the Progression Outcome Groups[a]

Voxel-based T_2 Relaxometry (VBR^{T_2}) Tests	Number of Voxels Detected				Voxel Activation Pattern ($P < 0.05$, No Correction)		Related Outcome
	No Correction	After Bonferroni Correction					
	$P < 0.05$	$P < 0.05$	$P < 0.01$	$P < 0.001$	Medial Side	Lateral Side	
One-way ANOVA							
All outcome groups	20,331	4,307	3,561	2,832	-	-	-
Two-sample *t*-tests							
'Both' vs 'ROA'	4,462	1	0	0			Pain Progression | ROA Progression
'Pain' vs 'None'	3,289	11	6	0			Pain progression
('Both'+'Pain') vs ('ROA'+'None')	9,154	21	7	1			
'ROA' vs 'Pain'	6,952	102	43	2			ROA progression
'Both' vs 'Pain'	11,156	79	23	5			
'Both' vs 'None'	18,033	2,096	1,557	1,041			
'ROA' vs 'None'	18,206	3,185	2,651	2,129			
('Both'+'ROA') vs ('Pain'+'None')	21,638	4,899	4,058	3,205			
('Both'+'ROA') vs 'None'	21,923	4,994	4,161	3,270			

[a]See also Fig. 4.44 on page 213 in which the same results are visualised to better show relationships among the progression outcome groups.

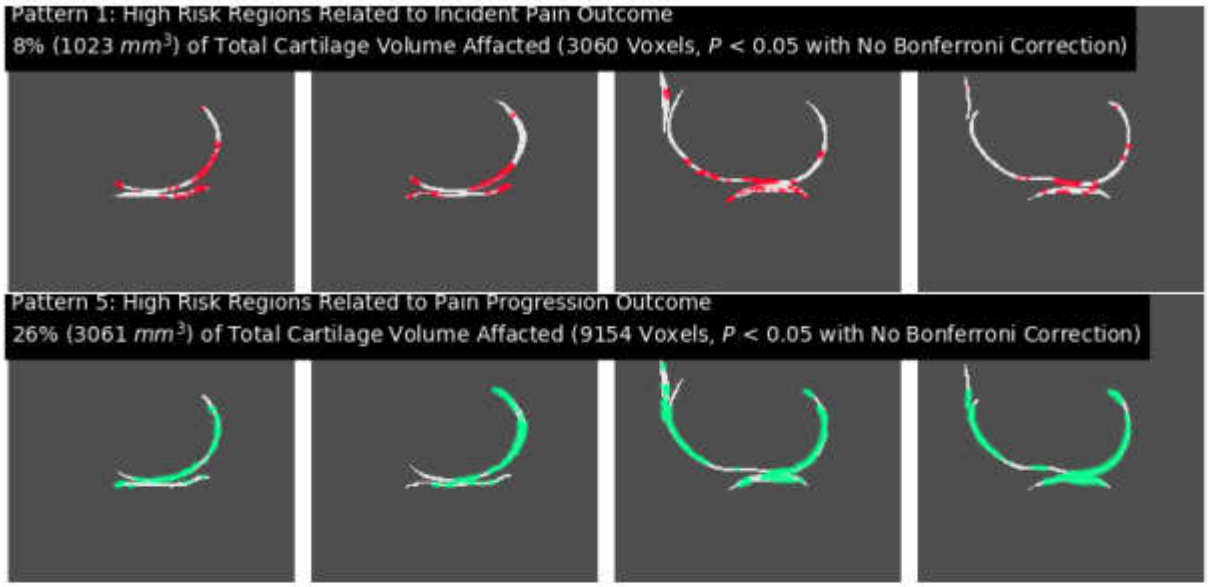


Figure 4.41: High Risk Regions Related to Pain Outcomes

In this thesis, I adopted a perspective that combined (2) and (3), that is, to view the patterns not only as delineation of early disease, but also as an estimation of expected future changes. The advantage of this perspective is that it provides us an opportunity to use the early changes shown in baseline T_2 maps to estimate the later changes yet to develop. Of course, as has been discussed in (1) above, the disease might not develop in a way that is only varying in magnitude but settled in location. However, the voxels in the patterns still have higher risks than other voxels to degenerate as their occurred change are substantial. A combination of perspectives (2) and (3) enables one to view the voxels in the patterns as voxels at risk, which would be difficult to evaluate otherwise. Therefore, an appropriate interpretation of the patterns, in this perspective, is the 'high risk regions' related to the outcomes. This enables a novel, quantitative way to characterise the disease.

First, the patterns provided some quantitative measurements otherwise very difficult to achieve objectively. Quantifying pain, for example, and relating it to specific regions is an extremely difficult task, and modelling pain not yet to come only makes the task more challenging. This is possible, however, by utilising the VBR^{T_2} patterns, as illustrated in the reproduced Patterns 1 and 5 in Fig. 4.41.

The top row of Fig. 4.41 shows Pattern 1 interpreted as high risk regions related to incident pain outcome, while the high risk regions of pain progression (Pattern 5) are presented in the second row. As can be seen, the regions related to pain are quantified in percentage and volume in cubic millimetre (mm^3). It can be seen the incident pain affects approximately 8% of total cartilage volume, while the pain progression affects more than three times of that, at 26%. The pattern for incident pain is more scattered compared to that for pain progression, and there is an overall trend that the sites affected extend from posterior and anterior as the incident pain progresses.

The volume in mm^3, which is calculated by multiplying voxel number with voxel size, provides an intuitive measurement of the regions at risk that can be related to familiar objects. For instance, the high risk region related to incident pain outcome is 1,023 mm^3, which is approximately equal to the volume of a British fifty pence coin[1]. This kind of quantification would not be possible without a voxel-based compare.

Another, more important insight available from these patterns relates to the mechanism of OA initiation and progression: specifically, the role of water, and the related mechanics of the tissue.

T_2 maps are essentially measuring the freedom of water within the tissue, with tightly bound water exhibiting faster relaxation. Control over water is perhaps the single most important factor in cartilage load carriage [19]. With disease, disruption of the collagen network hinders this control, increasing permeability, or 'freedom', and also leading to swelling [19]. Collagen meshwork disruption has been observed in human OA knees at the earliest stages of disease (while tissue remains histologically normal) [155], with wide-spread collagen damage forming a tipping point at which damage becomes energetically irreversible while loading is maintained. The subsequent swelling is indeed considered the best physical indicator of OA [20]. As water is more easily forced out of the tissue, stress

[1]The diameter of a 50p coin is 27.3mm, and the thickness 1.78mm. Assuming it is circular rather than heptagon-shaped, then it becomes a cylinder and the volume is given by $V = \pi r^2 h = \pi \times (\frac{27.3}{2})^2 \times 1.78 \approx 1042$ mm^3 (data from `www.royalmint.com`).

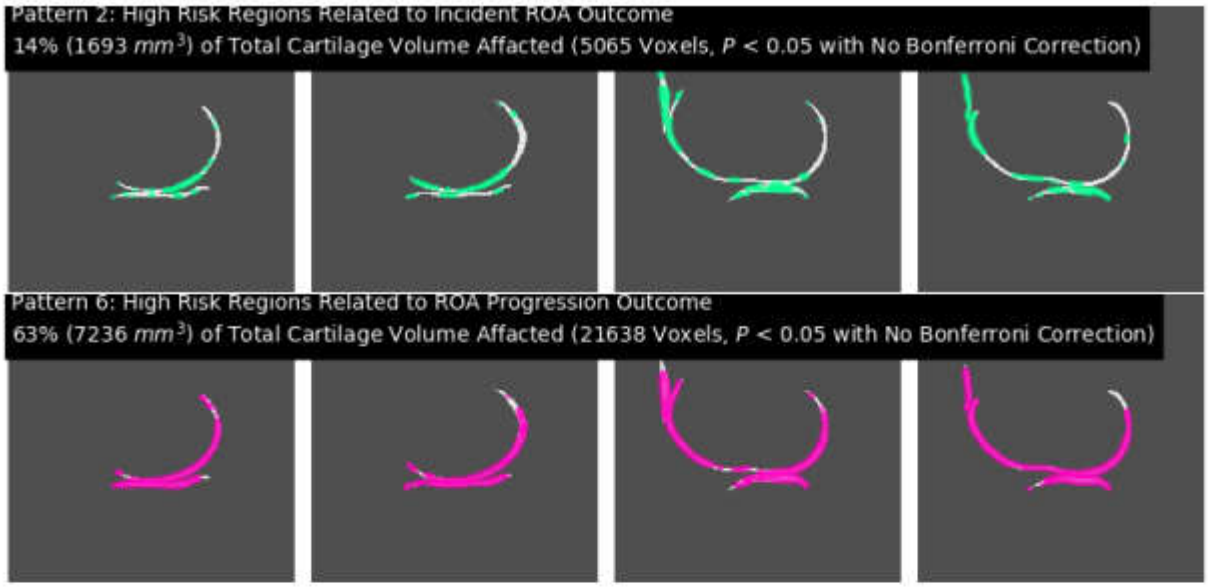


Figure 4.42: High Risk Regions Related to ROA Outcomes

is more quickly transferred to the solid thus continuing the damage process via a positive feedback loop.

Early-stage physical changes of this nature have not, to my knowledge, been observed in intact human knees. Given the findings of Chapter 2 they are unexpected, yet provide an interesting insight into how early the bulk mechanical function of cartilage is compromised in human OA. As can be seen in Fig. 4.42, the regions at risk for incident ROA (defined by KLG) already covered 14% of the cartilage, while that for ROA progression (defined by JSN) increased dramatically to 63%. These patterns means water freedom has already increased in the areas of risk before any macroscopic structural change was observed (worsening of KLG/JSN), highlighting an interesting opportunity for further research into the natural history of OA.

Interpretations of the Number of Voxels Detected Distance Maps

A different way of visualising the results is through a 'distance' interpretation. The number of voxels detected distance between the incidence outcome group is illustrated in Fig. 4.43 on the next page, which is then reproduced, to scale, in Fig. 4.44 on page 213 (on top left) to facilitate comparison with the map of the progression outcome groups.

Unlike the the patterns, which focus on detailed comparisons in voxel locations and distributions, these maps concentrates on illustrating the overall relationship among the groups, which is useful in understanding the transition between different stages/phases of the OA disease in a wider view. The distance between the nodes shows how easy/difficult it would be for every pair of these individual or combined outcomes to be differentiated. It is desirable that two outcomes are connected by a longer line, because this represents there are more voxel detections between the outcome group that made them more separated in the 2D space I created. This is particularly important for a complex problem such as that addressed in this chapter, in which the effects I seek to delineate are masked by others.

As can be seen in Fig. 4.43, the four incidence outcomes formed an overall triangular relationship, in which 'None' and 'Pain' occupy two vertices independently, while 'ROA' and 'Both' can be considered collectively occupying the third vertex. The nodes for the combined groups are drawn on the connection, or extended (dashed) connection lines of the groups being combined, so that the weakened/enhanced pattern exhibited between these combined groups are represented as the shortened/extended distances.

The four progression outcomes in Fig. 4.44, on the other hand, formed a quadrangle in which the two horizontal sides (both showing Pattern 6) are longer but the two vertical sides (showing Patterns 4 and 5 respectively) are shorter. This reflects the much stronger activation effects in Pattern 6 relative to Patterns 4 and 5. Again, the combined groups are shown in connection lines between, or dashed lines extended from the groups being combined, which themselves are also connected to compared against each other.

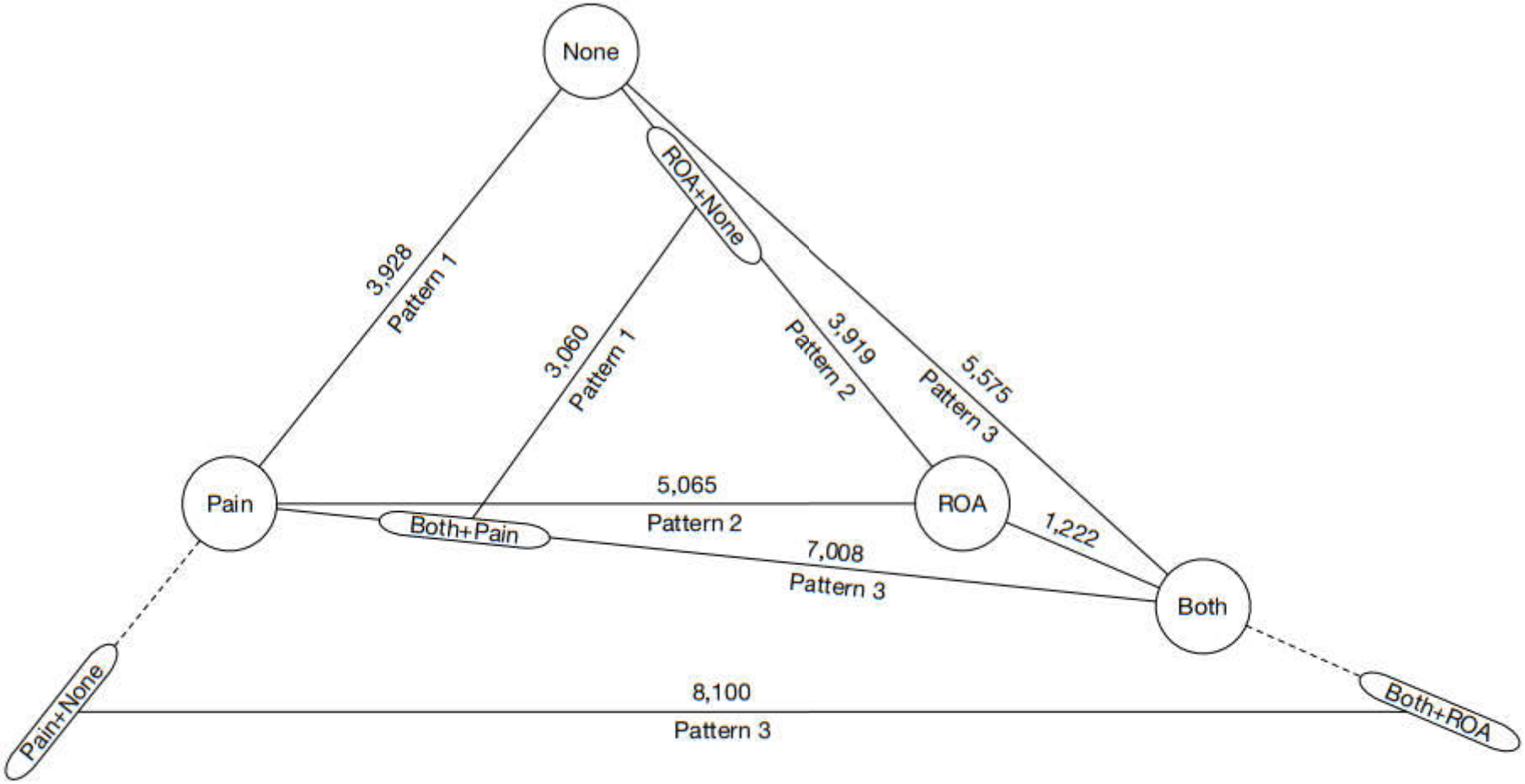


Figure 4.43: Voxel-Based T_2 Relaxometry ($\text{VBR}^{T_2T_2}$): Number of Voxels Detected Distance Between Incidence Outcome Groups

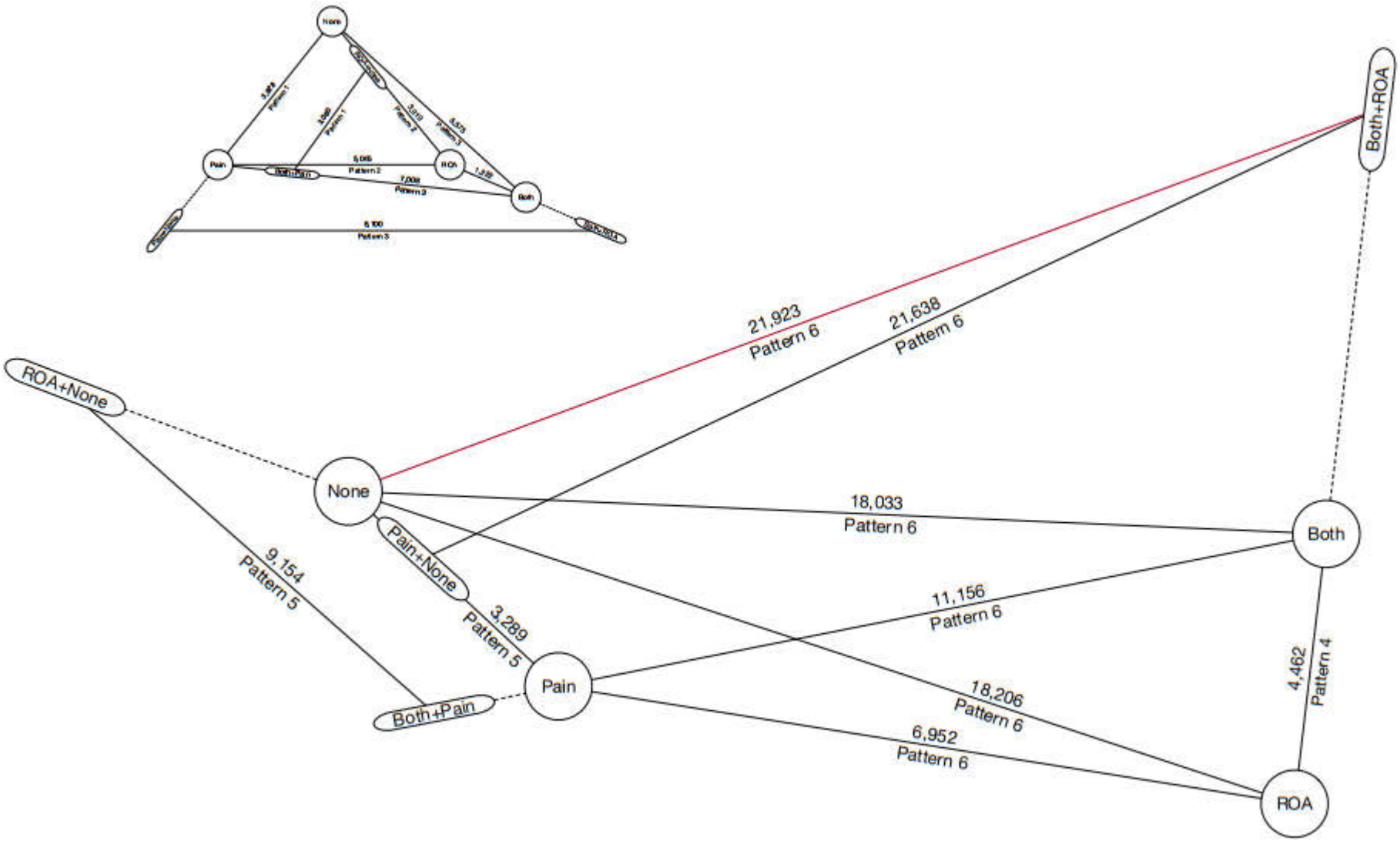


Figure 4.44: Voxel-Based T_2 Relaxometry ($VBR^{T_2T_2}$): Number of Voxels Detected Distance Between Progression Outcome Groups (Fig. 4.43 is reproduced to scale on top left)

There are two main insights that can me drawn from these maps:

First, it demonstrates the best application of T_2 mapping as a prognostic tool—use it to predict ROA progression outcome, regardless of pain. This corresponds to the connection line highlighted in red in Fig. 4.44 on the previous page, which displayed Pattern 6 and is also the strongest activation among all two-sample t-tests in this study. Using T_2 mapping in this setting, as is shown in the map, could achieve best separation between the outcomes of ROA progression and No ROA progression. Clearly, 'Will it get worse?' is a big question people would ask at clinic, and one of the most difficult to answer. As confirmed by the Bonferroni corrected result, baseline T_2 maps between these those radiographically progressed and those not are significantly different from each other even after the the extremely conservative correction at a extremely low P-value ($P < 0.001$), so we can almost be certain to differentiate these outcomes using T_2 maps acquired even when people have no symptom at all (included in the progression cohort), and provide a confident answer to the above question.

Second, it illustrates the strength of T_2 mapping for answering different questions, and highlights the place for this technique in the suite of available tests and assessment strategies. An important problem in image screening by radiologists is the inter-rater and intra-rater reliability. These maps give very clear information about how much room is left between outcome groups to allow for these variations. Similar to Chapter 2, this provides insight into the value of performing the test. For example, according to Fig. 4.44 on the preceding page, it would be easy to tell between 'None' and 'ROA' progression outcomes, but relatively difficult to tell between 'None' and 'Pain', irrespective of training/expertise.

In the design of clinical trials, these maps provide an overview of the significance level they could expect should they decide to use T_2 mapping as part of their research. The delineation of masking effects provides a useful tool for recruitment. According to Fig. 4.44 on the previous page, for example, one may take advantage of the enhanced Pattern 5 (9,154 activations) by ignoring the ROA status if they are imaging for pain progression.

However, according to Fig. 4.43 on page 212, they must be careful to compare pure 'Pain' vs pure 'None' participants, and avoid recruiting anyone who already developed ROA, to be able to see Pattern 1 instead of Pattern 2, if they are imaging for pain incidence. As such they further provide an objective assessment that can serve as an entry point into more complicated pattern analysis techniques such as those employed in the Part II of this chapter (the next section).

This page intentionally left blank

4.3 Prognosis with Machine Learning and Deep Learning: Development of OxteoNet

4.3.1 Philosophy Behind the OxteoNet

Following from the findings in Part I (Section 4.2) of Chapter 4 highlighting the prognostic potential of T_2 mapping, I present the development and investigation of an analysis tool, OxteoNet, to aid clinical decision making with machine learning.

Constrained by a limited timeframe within the scope of this DPhil project, the focus of this work was to produce a working prototype showing feasibility, rather than on producing a fully optimised tool (like OxteoFlow). I also aimed to quickly fill the gap between musculoskeletal MRI analysis and the state-of-the-art in the wider medical imaging community. The work herein is further bound by two main constraints:

No Competition with Human

The OxteoNet is developed as a tool to support clinicians in decision-making. Unlike some of the studies it is based on, OxteoNet is not, in any sense, developed to compete with human clinicians and radiologists. It is not intended to be used without human judgement, and its results should not be interpreted without domain knowledge. Instead, it is intended as a decision aid that can quantify the relationship between an individual (new scan) and a wider group with known outcomes (the OAI).

Data-Driven, Without Cherry-picking

This work is therefore data driven. As a machine (deep) learning project, OxteoNet is based on training the model with a large imaging data set (the OAI). Some significant published (not preliminary) research based on very small subsets of OAI indeed demonstrated good model performance (reference omitted). However, it is difficult to trust machine learning models tested on small fractions of a data set when the complete data set was freely and fully available.

4.3.2 Methods

4.3.2.1 Problem Formulation

In Section 4.2, T_2 mapping demonstrated best prognostic performance for ROA progression outcome, regardless of pain (Pattern 6). This was highlighted in rows F of Fig. 4.37 on page 199, and the red coloured connection line in Fig. 4.44 on page 213, both showing the most significant activation between the combined group of ('Both' + 'ROA'), and the 'None' group.

Based on this finding, and simplifying to a feasibility study, the machine learning problem is limited from predicting the label for all four progression outcome ('Both', 'ROA', 'Pain' and 'None'), to a binary classification problem—where the input is a T_2 map X and the output is a binary label $y \in \{0, 1\}$ representing No ROA progression ($0 =$ 'None') or ROA progression ($1 =$ 'Both' + 'ROA').

4.3.2.2 Study Design

Traditonal machine learning models and a deep learning model (OxteoNet) are investigated simultaneously in this project. To the best of my knowledge, this is the first comprehensive performance test of (most, if not all) popular machine learning (including deep learning) models using the baseline T_2 mapping data from the OAI.

4.3.2.3 Data Preparation

Previous chapters have resolved what is perhaps the most difficult part of using machine learning in medical imaging—controlling 'context' via registration and segmentation (Chapters 3 and 4). These efforts, translated into the language of machine learning, are the so-called feature engineering where only the meaningful data are extracted.

In some medical image machine learning projects, feature engineering ,beyond registration, were achieved through performing ANOVA to extract the most significant (activated) area

from the image and only use data in these area for training the model [156]. However, as can be seen in the ANOVA result in Fig. 4.35 on page 193, nearly the whole cartilage area is activated and there is no added benefit from excluding the smaller regions that are not. Therefore, the same, full cartilage segmentation in the reference space in Section 4.2 is directly reused in this project.

4.3.2.4 Hardware Preparation

The workstation 'horla' (see Table 4.1 on page 134) is upgraded with a NVIDIA GeForce GTX 1080 Ti graphics card, one of the most powerful and commonly used by the deep learning community at the time, in order to train the OxteoNet on the Graphics Processing Unit (GPU). The power supply unit and the liquid cooling system of 'horla' were also upgraded to accommodate the significantly increased power consumption and heat generation from the new configuration.

4.3.2.5 Traditional Machine Learning Models

The following traditional machine learning models are implemented and investigated in this project:

1) LinearSVC: Linear Support Vector Classification;

2) SVC (Linear): Support Vector Classification with a linear kernel function ($\langle x, x' \rangle$);

3) SVC (Poly): Support Vector Classification with a polynomial kernel function ($(\gamma \langle x, x' \rangle + r)^d$, where d (degree) and γ are hyperparamters, and I used $d = 5$);

4) SVC (RBF): Support Vector Classification with a rbf kernel function ($\exp(-\gamma \|x - x'\|^2)$, where γ is a hyperparamter);

5) SVC (Sigmoid): Support Vector Classification with a sigmoid kernel function ($\tanh(\gamma \langle x, x' \rangle + r)$, where γ is a hyperparamter);

6) Logit: Logistic Regression with L2 regularisation, which minimises the cost funtion: $\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\{\exp[-y_i(X_i^T w + c)] + 1\}$;

7) KNN: k-Nearest Neighbors, I used $k = 5$;

8) DTs: Decision Trees Classifier;

9) RFC: Random Forests Classifier;

10) ABC: AdaBoost Classifier;

11) GaussianNB: Gaussian Naive Bayes;

12) MLP: Multi-layer Perceptron Classifier; and

13) Chance: A classifier that makes predictions purely based on guess.

Implementing and optimising these models is itself a challenging task, even without including deep learning. The models are therefore not optimised, instead the default settings in Scikit-learn [157] are adopted to accelerate implementation and limit bias. Technical details of these models are not presented here, but could be found online at [158].

A stratified (preserving the percentage of samples for each class) 10-fold cross-validation strategy is used to train all the traditional machine learning models above. That is, the original data is randomly (but stratified) partitioned into 10 equal sized subsamples. A model is trained using 9 of the 10 folds as training data; the resulting model is validated on the remaining one fold of the data (i.e., it is used as a test set to compute a performance measure such as AUROC score). This procedure is repeated 10 times until all the 10 folds have been used as test set once, before the performance scores of each of the 10 runs are averaged.

4.3.2.6 Deep Learning Model—the OxteoNet

OxteoNet and CheXNet

The OxteoNet is informed by CheXNet [159], a deep learning model trained on ChestX-

ray14 data set which contains over 100,000 chest X-ray scans with 14 diseases. OxteoNet would not have been developed without using many open source codes that reimplemented CheXNet in Python (more specifically, PyTorch [160]). However, OxteoNet is not another reimplementation of CheXNet in Python, as they are a differentiated/related in a number of ways:

(1) OxteoNet is developed for predicting future Osteoarthritis using quantitative MRI (T_2 mapping), while CheXNet is for detecting occurred lung diseases using plain X-ray scans. They are trained using completely different data sets in two medical imaging domains (ChestX-ray14 vs the OAI).

(2) CheXNet intended to compete with human radiologist and claimed success; OxteoNet is developed as a tool to support human clinicians in screening, and also for filling the research gap between musculoskeletal imaging and the state-of-the-art (see Section 4.3.1 on page 217). I currently have no plan to compare OxteoNet's performance with humans, and argue that such a comparison is both inappropriate and uninformative. In this thesis, OxteoNet's performance is only compared with its traditional machine learning counterparts, and with our findings in the voxel-based analysis in Section 4.2.

(3) CheXNet is based on DenseNet [161]; OxteoNet is informed by CheXNet and also DenseNet-based. In this sense CheXNet and OxteoNet are siblings rather than parent and child, although OxteoNet, and many other deep learning models inspired by CheXNet, would not have been developed without this sibling.

Architecture of OxteoNet

OxteoNet is based on a 121-layer Dense Convolutional Network (DenseNet) [161] trained on the OAI data set. DenseNet improved flow of information and gradients through the network, making the optimisation of very deep networks tractable. Following the approach used in CheXNet, the final fully connected layer was replaced with one that has a single output, after which we apply a sigmoid nonlinearity [159].

The weights of the network are initialised with weights from a model pre-trained on ImageNet [162]. The network is trained end-to-end using Adam with standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) [163]. For a single example in the training set, the weighted binary cross entropy (BCE) loss is optimised. The technical details of this loss function can be found online at [164]. 64 mini-batches, 8 data loading workers, and 100 epochs are used. An initial learning rate of 0.0001 is used, and is decayed by a factor of 10 each time the validation loss plateaus after an epoch. The model with the lowest validation loss is then picked [159].

Due to the long training time, a stratified 5-fold, rather than 10-fold cross-validation strategy is used to train the OxteoNet. Due to the restriction of DenseNet, it only accepts RGB images of size 224×224 as input. CheXNet used a downscale strategy to resize the chest X-ray scans into the required size [159]. I consider this could lead to loss of pixel (voxel for MRI) information, so I crop the T_2 maps using a 224×224 window in the reference, which is just enough to cover the cartilage region. Three slices are chosen on medial (slices 5–7 of 27) and lateral (slices 19–21 of 27, see Fig. 4.35 on page 193) sides, which were passed to the three colour channels originally used for inputting the RGB image.

It should be noted, like the analysis of the traditional machine learning methods, OxteoNet is not optimised beyond the typical training regime of DenseNet architecture. There is room for fine-tuning and improvement, but this approach could ensure a fair comparison between OxteoNet and traditional machine learning models, which limits bias in the findings.

4.3.2.7 Performance Scores

Some performance metrics were defined in Chapter 2 for evaluating the performance of the fitting methods, which includes precision. In machine learning and pattern recognition, however, precision is defined differently as the number of true positives (TP) over the

number of true positives (TP) plus the number of false positives (FP). The concepts of specificity and sensitivity are also used in machine learning and have the same definition as in medical diagnosis, only that sensitivity is named 'recall' in the machine learning realm. The equations for these concepts are standardised below [158]:

$$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{4.1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{4.2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.3}$$

The Receiver Operating Characteristic (ROC) curve is created by plotting sensitivity (recall) against (1 − specificity), and the Area Under the ROC curve (AUROC, or AUC) is a measure of how well a parameter can distinguish between two groups. By computing the AUROC, the ROC curve is summarised in one number. Similarly, precision and recall can be plotted against each other to form the Precision-Recall curve. Average Precision ($\text{AP} = \sum_n (\text{Recall}_n - \text{Recall}_{n-1})\text{Precision}_n$) is the single number used the summarise the Precision-Recall curve [158].

In this project, AUROC and AP scores are used to evaluate the prediction performance, as they are widely used in machine learning and deep learning communities. Both of the metrics are between 0 and 1, and higher values are better.

4.3.3 Results and Discussion

4.3.3.1 Performance Comparison

For each model, the 10-fold cross-validation generated 10 groups of AUROC and AP scores, which are presented in Fig. 4.45 on page 225. In addition to the standard box plot (one interquartile range (IRQ) in the box and 1.5 IQRs in each whisker), the mean scores are also plotted as red triangles.

As can be seen in Fig. 4.45, OxteoNet clearly outperforms all the traditional machine learning models with mean AP scores of 0.612 on medial side and 0.604 on lateral side, while the AP scores of all the traditional machine learning models are all below 0.4. However, OxteoNet does not show significant advantage over the RFC, ABC, GaussianNB, and Logit models in the AUROC score, which are all at a level of approximately 0.650. Nevertheless, these two scores should be considered together and OxteoNet clearly demonstrates consistently good performance over its traditional machine learning counterparts.

Other traditional machine learning models, especially the SVC (RBF), SVC (Sigmoid) and the MLP, can be considered as not fit for purpose as those do not show clear improvement over the Chance classifier.

The medial models of OxteoNet showed slightly better, and much more consistent results in both scores than the lateral models. Such consistency in the medial models is a valuable feature, as it shows the learning was effective even if the training and testing data were different in each run of the cross-validation.

4.3.3.2 Model Interpretations

OxteoNet

To interpret OxteoNet's prediction, the class activation maps (CAMs) of the final convolutional layer are generated, following the methods used in CheXNet [159]. The CAM is a heatmap overlaid on the original anatomical image, which highlights regions used by the model to make the prediction (classification). The CAM heatmap has a size of 224×224, matching our cropped input T_2 maps. For better visualisation, I further cropped the CAMs to the size of 134×224, and overlaid them on the original, full sized anatomical images. The resulting CAM heatmaps for different subjects are illustrated in Fig. 4.46 on page 227.

As can be seen, OxteoNet made the prediction based on T_2 patterns in different regions. The continuous and smooth colour change in the heatmap shows that the model used

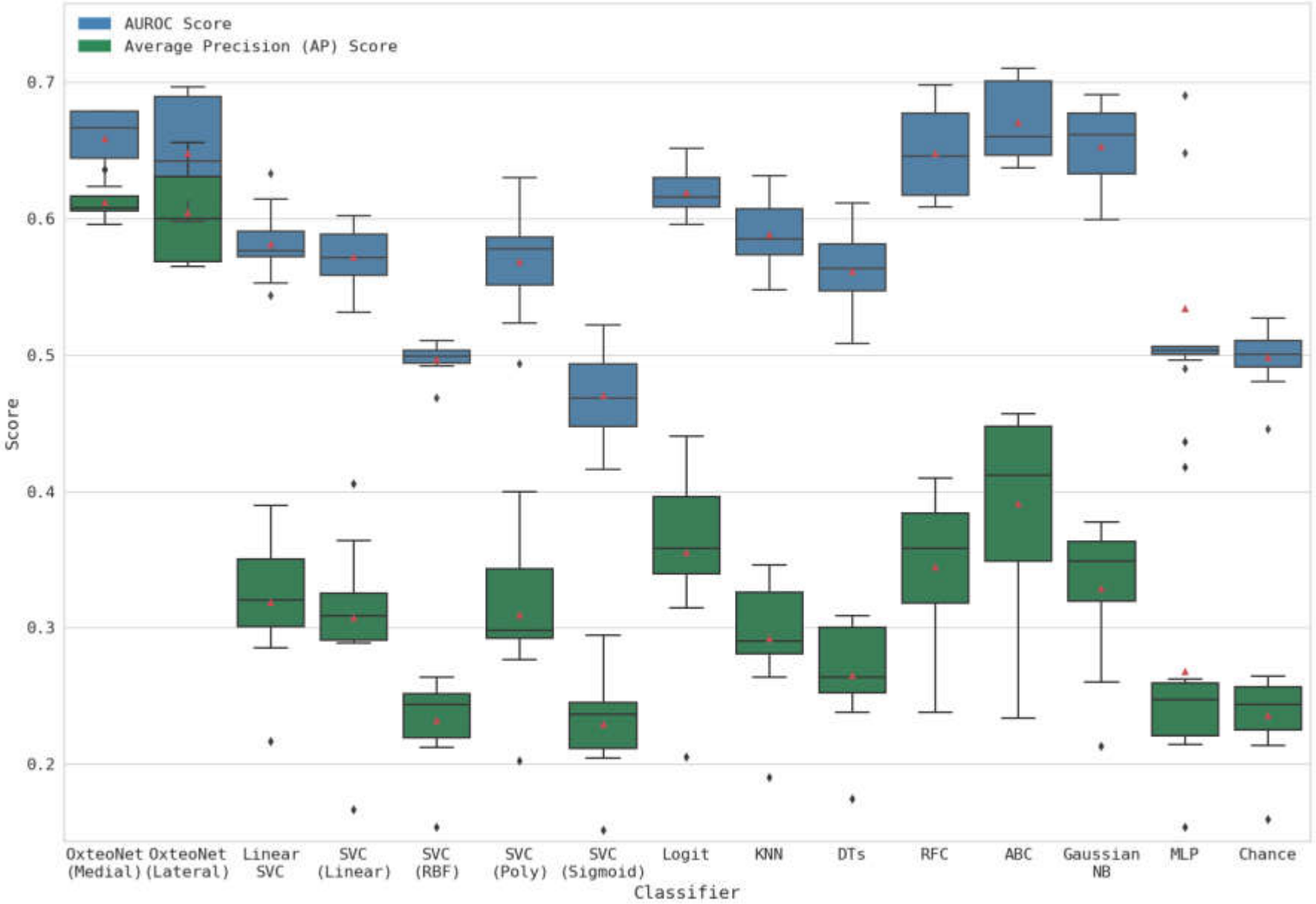


Figure 4.45: Box Plot of Cross-validation Scores of the OxteoNet vs Traditional Machine Learning Algorithms, Red Triangle Represents the Mean Value

the voxels in the cartilage collectively, rather than individually, to make the judgement. Variations in these heatmaps also show that regions contributing to the prediction range throughout the tissue. Coupled with the registration and normalisation in OxteoFlow, this provides a balance between constraint to the cartilage, thus forcing context on the prediction, and flexibility to account for individual variations between patients/subjects. Finally, the illustration of these regions may serve as an indicator for the clinician as site(s) at which pathology is observed, and warrant further inspection.

Random Forests

The Random Forests Classifier (RFC) is a representative traditional model that achieved a relatively good result in this study. The model's Feature Importance Map, similar to the CAM of OxteoNet, could also be visualised by reconstructing the learned features back into the image space. The Feature Importance Map of the RFC model are presented in Fig. 4.47 on page 228.

As can be seen in Fig. 4.47, the Feature Importance Map of RFC model closely matchs the gross appearance of Pattern 6 in the voxel-based analysis, especially those with FWER controlled in Fig. 4.40 on page 203. This could be considered both a strength and a weakness. First, it indicates that the model was successful in determining regions of importance, and showed sensitivity to the spatial patterns associated with disease. However, as the model showed similar result to the VBR^{T_2}, it means RFC still considers voxels individually and rigidly, in a similar manner to the two sample t-tests. The smaller AP score of this model compared to OxteoNet might be a demonstration of this weakness.

Linear Support Vector Classification

Finally, the weights of the Linear Support Vector Classification (LinearSVC) are visualised to illustrate the voxels the model used to make prediction. From Fig. 4.48 on page 229, the LinearSVC weights maps can, in an oversimplified way, be interpreted as the positive voxels (in warm colour) contributing to a decision of progression ($y = 1$), and the negative ones (in cold colour) to no progression ($y = 0$). As can observed in Fig. 4.48,

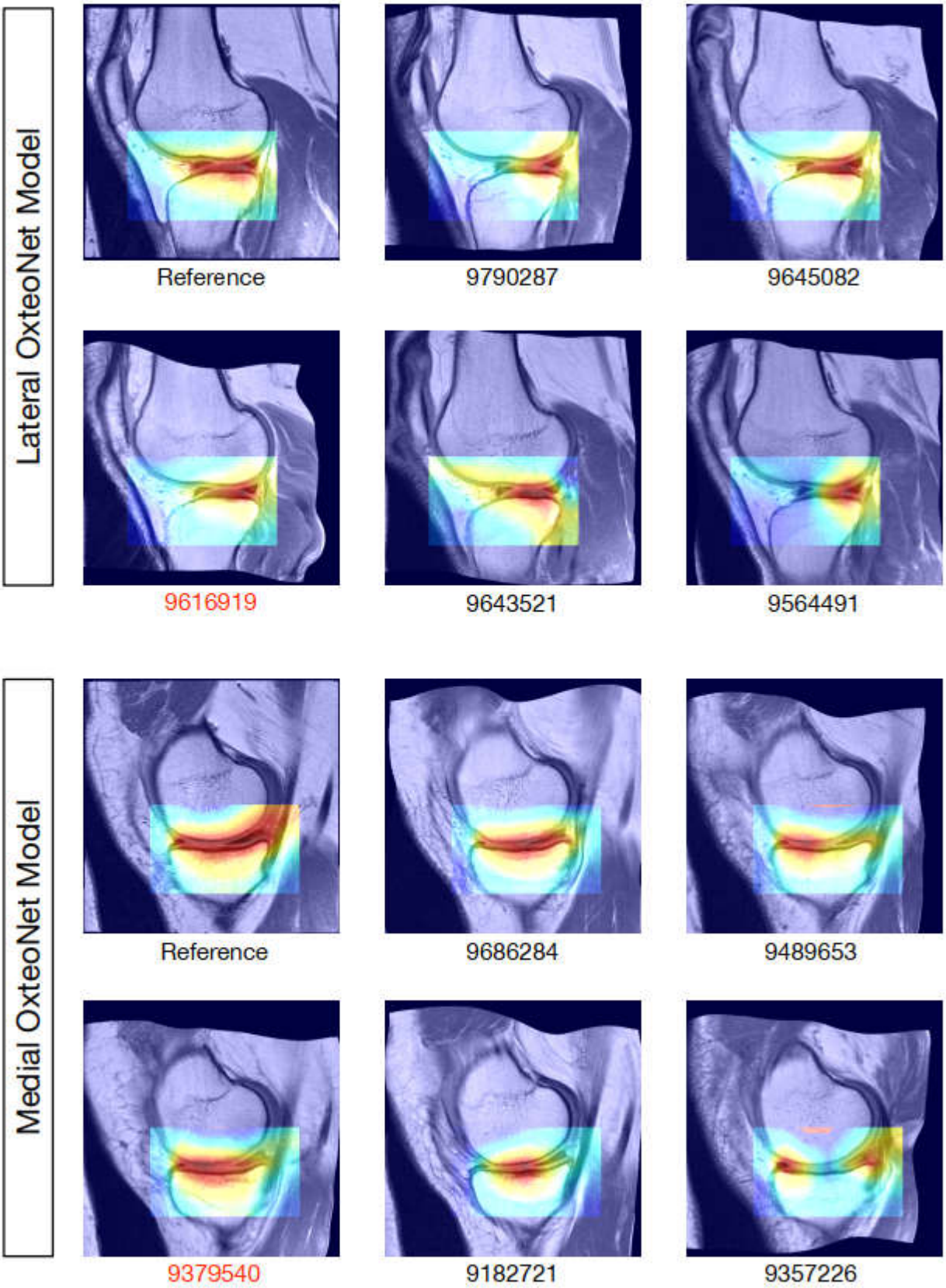


Figure 4.46: Class Activation Maps of OxteoNet for Different Subjects. Subjects with Red ID have an ROA Progression Outcome.

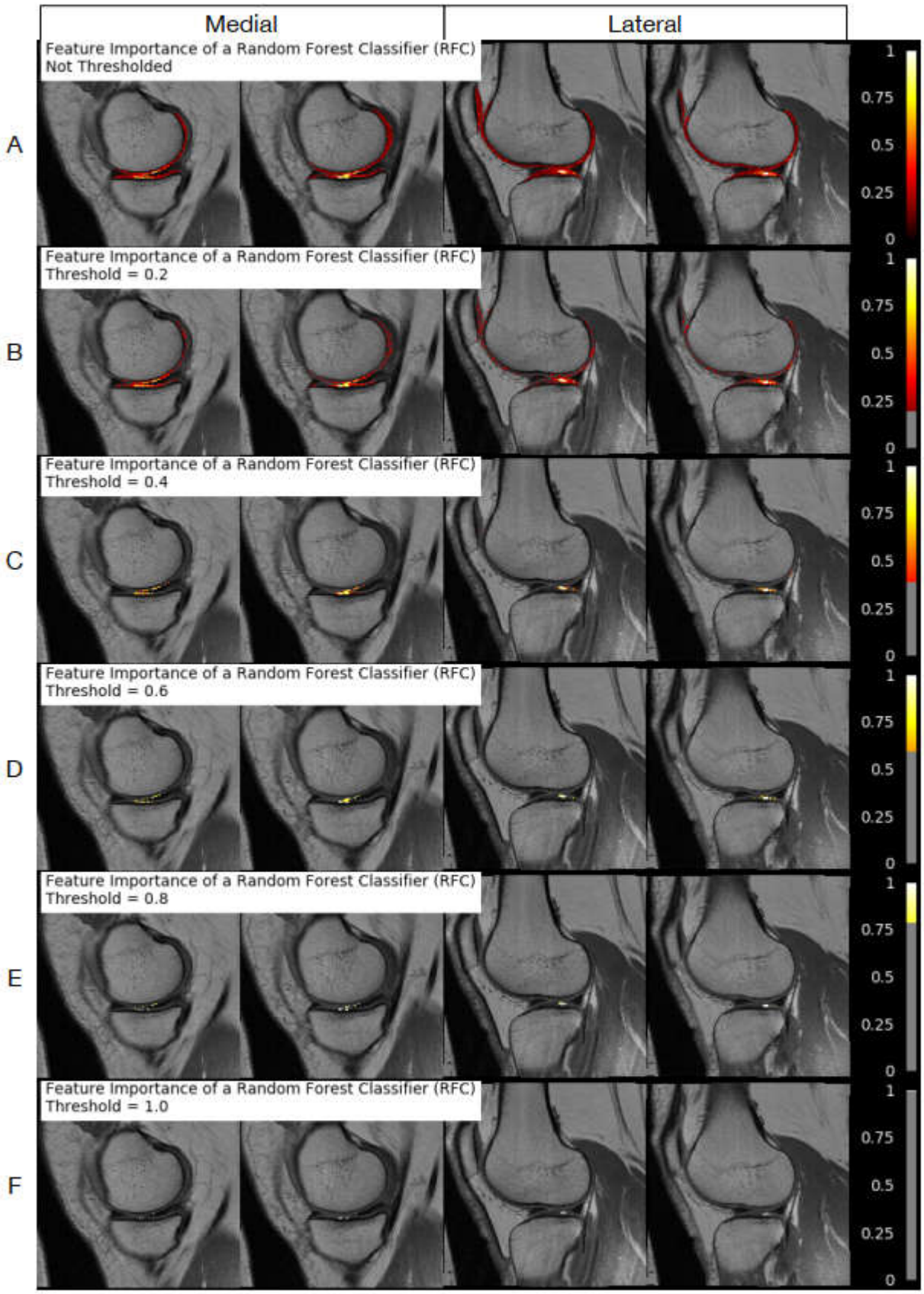


Figure 4.47: Feature Importance of a Random Forest Classifier (RFC)

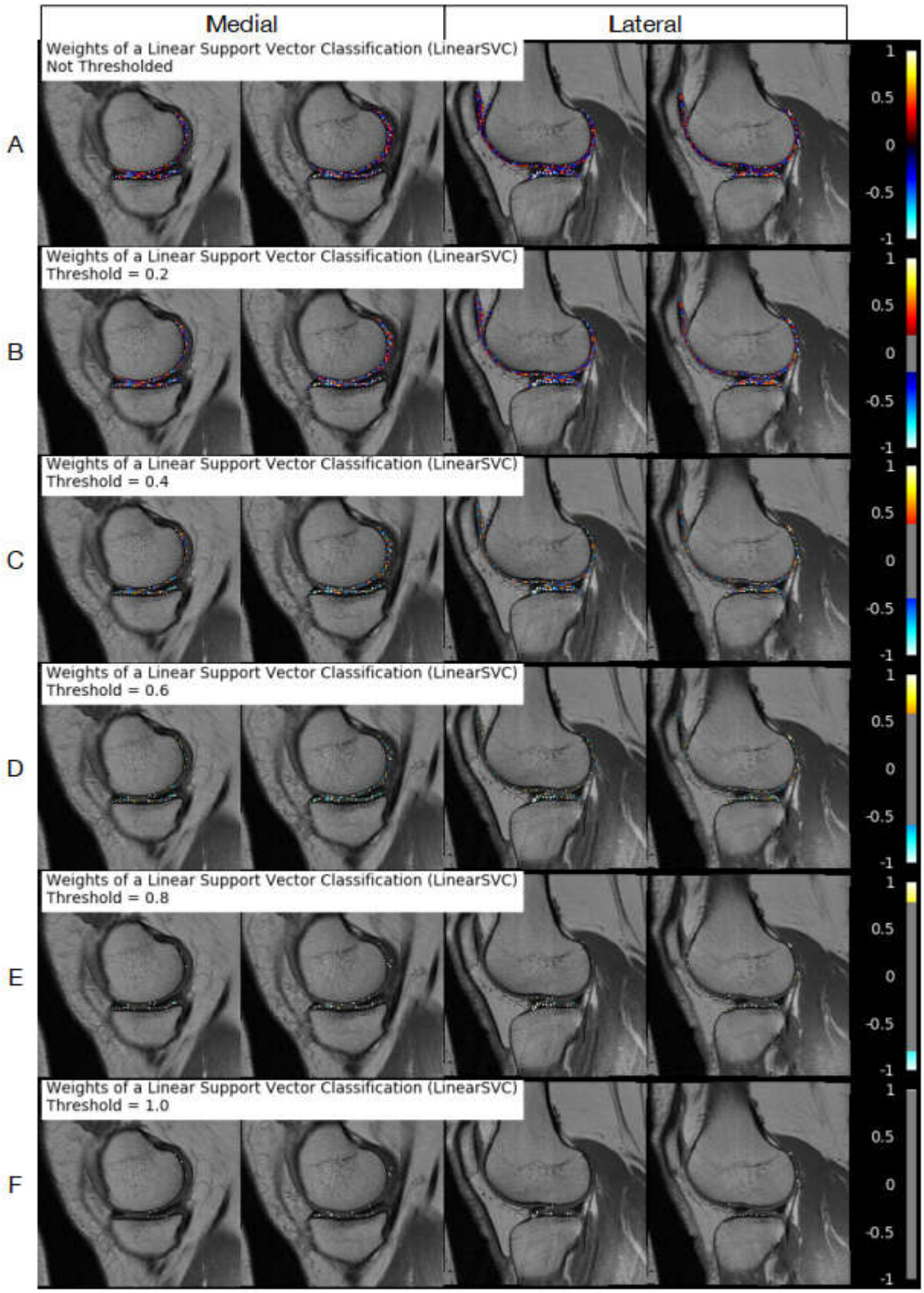


Figure 4.48: Weights of a Linear Support Vector Classification (LinearSVC)

the LinearSVC model uses a highly scattered approach to examine the T_2 map, in which no spatial correlation could be observed.

This further highlights the importance of patterns in T_2 maps: RFC and LinearSVC models considere voxels in T_2 map discretely, while OxteoNet considers T_2 map more collectively and continuously. Indeed, a model that only sees the image as a list of flattened numbers is insufficient to decode the patterns and textures stored in the T_2 maps. In terms of performance and approach, OxteoNet (deep learning) clearly provided better solutions over traditional machine learning models.

4.4 Conclusions

In summary, I developed a deep learning algorithm, OxteoNet, which predicts ROA progression using T_2 maps acquired even if no symptom is shown. The model can also highlight regions that it considers important, which could aid clinicians in routine screening, and provide a 'sense check' for the prediction. This preliminary model is compared with a series of popular traditional machine learning models widely used by the community, and is found to outperform all of them, particularly in specificity. Without extensive optimisation, OxteoNet has already shown promising results in disease prediction. I believe the model shows promise for further development and optimisation, as it has the potential to become an important aid to radiologists and clinicians in making diagnostic and prognostic decisions.

Chapter 5

General Discussion and Conclusions

This thesis has attempted to translate and build a platform technology for the analysis of musculoskeletal MRI. It is presented with a focus on T_2 mapping, essentially a case study based on the most commonly used 'physiological' scan. T_2 measurements probe the freedom of water movement, control over which is central to cartilage function [19] and the best physical indicator for health [20]. T_2 mapping should therefore be an excellent technique for characterisation in diseases such as OA. This has been reflected by encouraging results in tightly controlled in-vitro experiments [22], yet clinical application has been far less successful.

Chapter 2 explores the reasons for this discrepancy. Specifically, noise in T_2 scans plays an important role in the calculation of T_2 values. Although advanced T_2 fitting methods such as NCEXP (where a simplified equation is reported for the first time) and SQEXP can model Rician noise and achieve good performance in controlling systematic fitting errors (trueness), the random fitting error, measured by precision, must be smaller than the biological effect to be resolved. Through a systematic review and meta-analysis, it was found such a biological effect, i.e. difference in T_2 value (ΔT_2^{META}) between OA

and healthy tissue to be extremely small. In a typical T_2 scan, the random fitting error (precision) caused by noise would normally be comparable to ΔT_2^{META}, making the good trueness performance meaningless and absolute measurement of T_2 of little use.

This of course leads to the question of how to harness the theoretical potential of T_2 mapping given the limitations of clinical imaging. The common clinical approach is to assess patterns of T_2, based on both their knowledge of typical patterns of disease manifestation and comparison with patterns in previously viewed images. This approach is inherently subjective, yet nevertheless informative for decision making. My goal was to make it quantitative.

To achieve this goal I had four specific aims:

1) Transform images from different patients to the same space, to allow direct comparison;

2) Compare those images to provide a quantitative view of how early-stage disease manifests;

3) Use these images to predict OA progression; and

4) Develop the above as a platform technology that can be used for a range of musculoskeletal imaging applications.

The problem of transforming individual scans to the same space has been pursued within the neuroscience community, and their methodology is now reaching maturity. In Chapter 3, I harnessed a number of open-source computational tools developed by that community, adapted, and optimised them to the very different morphologies and image qualities in musculoskeletal imaging.

Specifically, I used a two-step approach, performing standard affine transformations followed by a more complex nonlinear transformation (or warping) to register individuals to a reference scan. A grid search optimisation approach was used to optimise the model, and the parameters were decided by examining similarity metrics at each of the grids. I implemented parallel computing to accelerate the process, and the models are further

developed into a automated preprocessing workflow, OxteoFlow, that was tested and optimised for both knee and hip, T_2 and dGEMRIC data. OxteoFlow is the foundation of the analysis platform I have developed in this DPhil. As can be seen by visual inspection (which is arguably more reliable than similarity metrics), OxteoFlow successfully normalised knees (Fig. 3.10 on page 124) and hips (Fig. 3.15 on page 129) of different shapes into their respective reference spaces.

Having established this foundation allowing scans to be aligned in the same space, Chapter 4 demonstrates its application. First, voxel-wise comparison of scans were performed based on structural (ROA) and symptomatic (pain) in the OAI incidence and progression cohorts. Specifically, I stratified these cohorts into four groups—'Both','ROA', 'Pain' and 'None' based on their outcomes. By performing a voxel-based pattern analysis on their baseline T_2 maps, six different patterns, all relating to different incident/progression outcomes were identified. These patterns mean T_2 mapping could image regions at risk for a future outcome, even if the subject has not shown/experienced any symptoms. I found that the relationship among the outcomes are complex, but can be illustrated with a voxels detected distance map, which shows how easily the outcomes can be differentiated using a T_2 map. I found that in order to observe the effect of pain, an investigator needs to carefully rule out the influence of ROA, as the T_2 pattern of former will be masked by the latter.

Encouraged by the findings of distinct anatomical sites associated with disease progression, yet variation between individuals, I have explored the opportunities for predicting disease progression. Specifically targeting the progression cohort (most representative of the patient presenting at clinic), I have tested and compared leading machine learning algorithms with my own model—OxteoNet, based on the Dense Convolutional Network (DenseNet) of [161]. Traditional machine learning approaches, which use combinations of the voxel-wise results illustrated in the two sample t-tests generally produced strong AUROC scores. The OxteoNet provides a small, though significant, improvement in AUROC, yet a dramatic improvement in AP score. I expect this improvement to relate

to the freedom of OxteoNet to detect patterns that are similar, yet not necessarily in the same location for each patient.

The comparison of Figures Fig. 4.47 on page 228 and Fig. 4.46 on page 227, which illustrate important regions for prediction, are also informative. Random forrest predictions, like each of the higher-performing traditional methods, shows distinct regions of importance, closely matching regions highlighted in the ANOVA (Fig. 4.35 on page 193) and two sample t-tests (Figs. 4.37 to 4.40 on page 199 and on pages 201–203) results. OxteoNet in contrast shows some preference for these regions, yet utilises data from a wider region of tissue, perhaps reflecting this variability.

Finally, it is important to note that this comparison does not use optimised parameters, leaving room for improvement in all results. I expect, however, that such an optimisation will be of greater benefit to the more complex, and less established OxteoNet approach.

5.1 Limitations and Future Work

Again, this thesis seeks primarily to develop a platform, and the specific results of Chapters 3 and 4 with respect to T_2 and knee OA should be considered as case studies. As such, there is far more insight to be gained from the results of Chapter 3. Indeed it is expected that separate programmes of work looking deeper into both the exploration of OA initiation and progression (building on Section 4.2 on page 134) and the development of a predictive tool that can be established as part of the radiological pathway (building on Section 4.3 on page 217).

Although considerable effort has been made to reason most of the methodological and analytical choices in this DPhil, I acknowledge a number of limitations and pitfalls in the work. In most of the cases, these limitations were actually identified at early stages of each of the projects, but were recognised as unfeasible to avoid within the time and

resource constraints of the project. On the other hand, the logical flow underlying the chapters in this thesis often meant later projects were dependent upon earlier results to commence, which required compromises and decisions be made to move projects forward. Some of those choices, strictly speaking, could either be oversimplified, too empirical, or far from universal, leaving many questionable and debatable limitations.

First of all, this study concentrates on the more controllable factors in the computational aspects of utilising T_2 mapping, specifically, T_2 calculation and image preprocessing. Some well recognised factors, such as the magic angle and partial volume effects, are not considered or dealt with in this thesis.

The magic angle effect is the focally increased signals observed in curved cartilage surfaces where collagen fibres are oriented at an angle of $\theta = \arccos\left(\sqrt{\frac{1}{3}}\right) \approx 54.74°$ relative to the static magnetic (polarising) field (B_0), such as the acetabular cartilage in hip T_2 scans and femoral condyle in knee T_2 scans. This effect is not considered in this thesis because 1) some T_2 fitting methods that deals with magic angle effect involve excluding the first echo from the the calculation, which in my opinion would cause more harm than benefit because those methods rely on the later, much noisy echoes to perform the fitting, leading to more variation in the resulting T_2 value; and 2) the location and magnitude of the intensity increase involved in the magic angle effect are consistent among subjects, especially among normalised/registered scans in the same image set (such as the OAI). A consistent increase of signals would not affect interpretation and analysis of the T_2 difference among scans, because the bias would be cancelled out by the comparison.

The partial volume effect, which occurs when more than one tissue type are contained in a voxel, is another limiting factor in T_2 mapping where the intensity of the mixed voxels are the weighted sum of signals from different tissues. Such an effect is also not considered in this thesis because 1) the mathematics underlying partial volume effect is essentially similar to the sub-sampling technique in registration and the convolution operation in deep learning, where signals from different voxels are combined to simplify analysis. If sub-sampling and convolution are so important to registration and machine

learning, there is no reason why partial volume effect be harmful to the same processes; and 2) partial volume effect affects boundary voxels of tissues the most (for example, the interface between cartilage and bone), which therefore does not influence interpretation of the patterns within the cartilage. It can, on the contrary, create another layer of gross structure along boundaries of different tissues that could aid registration.

In Chapter 2, I intentionally neglected the fitting methods used by the studies included in the meta-analysis, to avoid the bias of excluding the majority of studies that do not report these information. However, this could introduce another bias that if these included studies all used the simplest NLS fitting (which they probably did), then the good performance of NLS in the fitting method evaluation could just be due to the bottomline criteria favouring it. There were some other limitations in the meta-analysis, for example, the compartmental data selection could be another source of bias as it only chose one most significant compartment from each study. The interpretation of the effect size from the meta-analysis also could have been more insightful if I had more studies to compare against.

In the fitting methods evaluation, I only tested monoexponential fitting models , but did not include the biexponential models also being used by some studies [165]. Also, I did not take the Cramer-Rao lower bound (CRLB) in to consideration, but explained the models solely based on their simulation performance. Such a simulation test could have also benefitted from a phantom study for comparison, which was not performed due to restricted access to scanners. This limitation, however, was offset by the meta-analysis which provided the bottom line criteria of the performance metrics.

The OxteoFlow, which was developed in Python, is still in the form of a command-line tool, and requires a complicated installation and setup process of its dependencies on each of its workstations, which limited its dissemination and testing. A user interface (UI) was planned, but not possible to achieve during the timeframe of this DPhil. In future, I plan to pack this platform, including OxteoFlow and OxteoNet, and all their dependencies into a container image that can be quickly and easily distributed. Also, as

one of its key dependency, FSL, is currently not open-sourced, the OxteoFlow requires further optimisation in the ANTs branch to make itself a fully open source software.

In this study, I had restricted access to suitable imaging sets, forcing me to use the OAI image set. Although OxteoFlow has shown good performance on a small cohort of hip dGEMRIC scans (SibKids), dGEMRIC is now a banned procedure in the UK, so the good performance of OxteoFlow in this modality can only show its robustness but with no practical value. Further testing and optimisation of OxteoFlow and OxteoNet on a larger number of hip, even shoulder and ankle T_2 scans, is required to fully demonstrate its utility for all musculoskeletal scans.

In the voxel-based analysis and machine learning, I only included right knees, and did not include the relatively small numbered left knee scans. Although the left scans could be flipped into right ones, I suspected this could introduce another source of bias to the VBR^{T_2} analysis. This approach, however, might be an effective way to increase the number of samples available for the training of traditional machine learning models and the OxteoNet, and should have been tested and evaluated.

Finally, the machine learning models investigated, including the OxteoNet were not optimised. This was to ensure a fair test of their performance, but also due to the restricted time available. The current OxteoNet model has already shown promising results, and I consider this limitation a subtle one, and one that could easily be resolved by devoting more time in tuning and training.

In summary, this thesis covered a complete research and development (R&D) life cycle of T_2 imaging and its applications (calculation $\longrightarrow$ preprocessing $\longrightarrow$ analysis $\longrightarrow$ diagnosis/prognosis), answered some fundamental and important questions related to this non-invasive OA biomarker, translated some of the latest and best technical advances from a wide range of medical imaging domains into musculoskeletal, and demonstrated its value as an image analysis platform for musculoskeletal studies. A potential implementation enabled by this platform, is illustrated in Fig. 5.1 on the following page. □

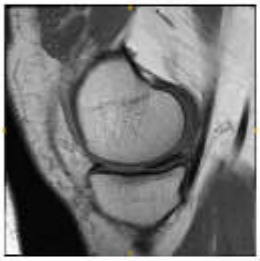
OxteoFlow
OxteoNet | Image Analysis Report

NDORMS
UNIVERSITY OF OXFORD

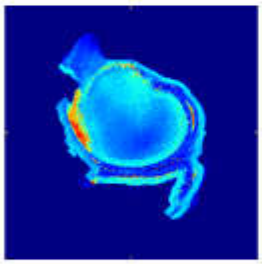
Patient name: XX
Date of Birth: XX-XX-XXXX
Scanner: SIEXXXS
Report Date: XX-XX-XXXX

Radiologist: XX

Centre: Nuffield Orthopaedic Centre



T2



cT2



OxteoNet Class Activation Map
(Important Area)

OxteoNet Favours:
No Osteoarthritis Progression Outcome in 96 months (87%)

WARNING: Further Inspection is ALWAYS Required.

OxteoNet is an AI model based on Densely Connected Convolutional Networks (DenseNets) trained on a subset of Osteoarthritis Initiative (OAI) image set, which contains 3,870 knee T2 scans with known outcomes.

Figure 5.1: A Potential Implementation Based on the Studies in This Thesis

References

[1] Murray, C. J. et al., "UK health performance: findings of the Global Burden of Disease Study 2010", *Lancet*, 381/9871 (2013), 997–1020, https://www.ncbi.nlm.nih.gov/pubmed/23668584.

[2] Arthritis Research UK, "Osteoarthritis in general practice" (2013).

[3] Lin, Elizabeth H. B., "Depression and Osteoarthritis", *The American Journal of Medicine*, 121/11, Supplement 2 (2008), S16–S19, http://www.sciencedirect.com/science/article/pii/S0002934308008644.

[4] Rahman, M. Mushfiqur et al., "The relationship between osteoarthritis and cardiovascular disease in a population health survey: a cross-sectional study", *BMJ Open*, 3/5 (2013), http://bmjopen.bmj.com/content/3/5/e002624.abstract.

[5] Cross, M. et al., "The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study", *Ann Rheum Dis*, 73/7 (2014), 1323–30, http://www.ncbi.nlm.nih.gov/pubmed/24553908.

[6] London, N. J., Miller, L. E., and Block, J. E., "Clinical and economic consequences of the treatment gap in knee osteoarthritis management", *Med Hypotheses*, 76/6 (2011), 887–92, https://www.ncbi.nlm.nih.gov/pubmed/21440373.

[7] Burstein, D. and Hunter, D. J., "'Why aren't we there yet?' Re-examining standard paradigms in imaging of OA: summary of the 2nd annual workshop on imaging based measures of osteoarthritis", *Osteoarthritis Cartilage*, 17/5 (2009), 571–8, https://www.ncbi.nlm.nih.gov/pubmed/19233336.

[8] Matzat, S. J. et al., "Quantitative MRI techniques of cartilage composition", *Quant Imaging Med Surg*, 3/3 (2013), 162–74, http://www.ncbi.nlm.nih.gov/pubmed/23833729.

[9] Chu, C. R., Millis, M. B., and Olson, S. A., "Osteoarthritis: From Palliation to Prevention: AOA Critical Issues", *J Bone Joint Surg Am*, 96/15 (2014), e130, http://www.ncbi.nlm.nih.gov/pubmed/25100783.

[10] Chu, C. R. and Andriacchi, T. P., "Dance between biology, mechanics, and structure: A systems-based approach to developing osteoarthritis prevention strategies", *J Orthop Res*, 33/7 (2015), 939–47, http://www.ncbi.nlm.nih.gov/pubmed/25639920%20https://onlinelibrary.wiley.com/doi/pdf/10.1002/jor.22817.

[11] Braun, H. J. and Gold, G. E., "Diagnosis of osteoarthritis: imaging", *Bone*, 51/2 (2012), 278–88, http://www.ncbi.nlm.nih.gov/pubmed/22155587.

[12] Carr, A. J. et al., "Advances in arthroscopy-indications and therapeutic applications", *Nat Rev Rheumatol*, 11/2 (2015), 77–85, http://www.ncbi.nlm.nih.gov/pubmed/25348038.

[13] Beard, David J. et al., "Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial", *The Lancet*, 391/10118 (2018), 329–38.

[14] Thorlund, J. B. et al., "Arthroscopic surgery for degenerative knee: systematic review and meta-analysis of benefits and harms", *BMJ* 350 (2015), h2747, http://www.ncbi.nlm.nih.gov/pubmed/26080045.

[15] Palmer, A. J. et al., "Non-invasive imaging of cartilage in early osteoarthritis", *Bone Joint J* 95-B/6 (2013), 738–46, http://www.ncbi.nlm.nih.gov/pubmed/23723266.

[16] Bangerter, N. K. et al., "Quantitative techniques for musculoskeletal MRI at 7 Tesla", *Quant Imaging Med Surg*, 6/6 (2016), 715–30, https://www.ncbi.nlm.nih.gov/pubmed/28090448.

[17] Surowiec, R. K., Lucas, E. P., and Ho, C. P., "Quantitative MRI in the evaluation of articular cartilage health: reproducibility and variability with a focus on T2 mapping", *Knee Surg Sports Traumatol Arthrosc*, 22/6 (2014), 1385–95, http://www.ncbi.nlm.nih.gov/pubmed/24170187.

[18] Crema, M. D. et al., "Articular Cartilage in the Knee: Current MR Imaging Techniques and Applications in Clinical Practice and Research", *Radiographics*, 31/1 (2011), 37–U76, %3CGo%20to%20ISI%3E://WOS:000286608900004.

[19] Klika, V. et al., "An overview of multiphase cartilage mechanical modelling and its role in understanding function and pathology", *J Mech Behav Biomed Mater*, 62 (2016), 139–57, https://www.ncbi.nlm.nih.gov/pubmed/27195911.

[20] Broom, Neil D. and Flachsmann, Rene, "Physical indicators of cartilage health: the relevance of compliance, thickness, swelling and fibrillar texture", *Journal of Anatomy*, 202/6 (2003), 481–94, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1571108/.

[21] Oloyede, A. and Broom, N., "The biomechanics of cartilage load-carriage", *Connect Tissue Res*, 34/2 (1996), 119–43, http://www.ncbi.nlm.nih.gov/pubmed/8909876.

[22] Baum, T. et al., "Correlation of magnetic resonance imaging-based knee cartilage T2 measurements and focal knee lesions with body mass index: thirty-six-month followup data from a longitudinal, observational multicenter study", *Arthritis Care Res (Hoboken)*, 65/1 (2013), 23–33, http://www.ncbi.nlm.nih.gov/pubmed/22623435.

[23] Chu, Constance R. et al., "Early diagnosis to enable early treatment of pre-osteoarthritis", *Arthritis Research & Therapy*, 14/3 (2012), 212–2, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3446496/.

[24] Xin, Y, *Novel Imaging and Processing for Early Musculoskeletal Disease: a Transfer Report Submitted for Transferring from Probationer Research Student to DPhil Status*, Report (Trinity Term, 2015).

[25] Bolbos, R. I. et al., "Relationship between trabecular bone structure and articular cartilage morphology and relaxation times in early OA of the knee joint using parallel MRI at 3 T", *Osteoarthritis Cartilage*, 16/10 (2008), 1150–9, http://www.ncbi.nlm.nih.gov/pubmed/18387828.

[26] Li, X. et al., "Quantitative MRI using T1rho and T2 in human osteoarthritic cartilage specimens: correlation with biochemical measurements and histology", *Magn Reson Imaging*, 29/3 (2011), 324–34, http://www.ncbi.nlm.nih.gov/pubmed/21130590.

[27] Rautiainen, J. et al., "Multiparametric MRI assessment of human articular cartilage degeneration: Correlation with quantitative histology and mechanical properties", *Magn Reson Med* (2014), http://www.ncbi.nlm.nih.gov/pubmed/25104181.

[28] Zuo, H. et al., "The biomarkers changes in serum and the correlation with quantitative MRI markers by histopathologic evaluation of the cartilage in surgically-induced osteoarthritis rabbit model", *PLoS One*, 10/4 (2015), e0124717, http://www.ncbi.nlm.nih.gov/pubmed/25884951.

[29] Lee, J. H. et al., "Topographical variations of the strain-dependent zonal properties of tibial articular cartilage by microscopic MRI", *Connect Tissue Res*, 55/3 (2014), 205–16, http://www.ncbi.nlm.nih.gov/pubmed/24559385.

[30] Lammentausta, E. et al., "T_2 relaxation time and delayed gadolinium-enhanced MRI of cartilage (dGEMRIC) of human patellar cartilage at 1.5 T and 9.4 T: Relationships with tissue mechanical properties", *J Orthop Res*, 24/3 (2006), 366–74, http://www.ncbi.nlm.nih.gov/pubmed/16479569.

[31] Miyata, S. et al., "Feasibility of noninvasive evaluation of biophysical properties of tissue-engineered cartilage by using quantitative MRI", *J Biomech*, 40/13 (2007), 2990–8, http://www.ncbi.nlm.nih.gov/pubmed/17442320.

[32] Keenan, K. E. et al., "Prediction of glycosaminoglycan content in human cartilage by age, T1rho and T2 MRI", *Osteoarthritis Cartilage*, 19/2 (2011), 171–9, http://www.ncbi.nlm.nih.gov/pubmed/21112409.

[33] Othman, S. F. et al., "High-resolution/high-contrast MRI of human articular cartilage lesions", *Acta Orthop*, 78/4 (2007), 536–46, http://www.ncbi.nlm.nih.gov/pubmed/17966009.

[34] Raya, J. G. et al., "T_2 measurement in articular cartilage: impact of the fitting method on accuracy and precision at low SNR", *Magn Reson Med*, 63/1 (2010), 181–93, https://www.ncbi.nlm.nih.gov/pubmed/19859960.

[35] Deeks, JJ, Higgins, JPT, and Altman, DG, "Chapter 9: Analysing data and undertaking meta-analyses", in *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]*, ed. Higgins JPT and Green S (The Cochrane Collaboration, 2011), http://handbook.cochrane.org.

[36] Kellgren, J. H. and Lawrence, J. S., "Radiological assessment of osteo-arthrosis", *Ann Rheum Dis*, 16/4 (1957), 494–502, http://www.ncbi.nlm.nih.gov/pubmed/13498604.

[37] Peterfy, C. G. et al., "Whole-Organ Magnetic Resonance Imaging Score (WORMS) of the knee in osteoarthritis", *Osteoarthritis Cartilage*, 12/3 (2004), 177–90, http://www.ncbi.nlm.nih.gov/pubmed/14972335.

[38] Multanen, J. et al., "Bone and cartilage characteristics in postmenopausal women with mild knee radiographic osteoarthritis and those without radiographic osteoarthritis", *Journal of Musculoskeletal & Neuronal Interactions*, 15/1 (2015), 69–77, http://www.ncbi.nlm.nih.gov/pubmed/25730654.

[39] Zuo, H. et al., "Quantitative evaluation in combination with nonquantitative evaluation in early patellar cartilage osteoarthritis at 3.0 T", *Clin Interv Aging*, 9 (2014), 1133–43, http://www.ncbi.nlm.nih.gov/pubmed/25075182.

[40] Liebl, H. et al., "Early T_2 changes predict onset of radiographic knee osteoarthritis: data from the osteoarthritis initiative", *Ann Rheum Dis* (2014), http://www.ncbi.nlm.nih.gov/pubmed/24615539.

[41] Prasad, A. P. et al., "$T_{1\rho}$ and T_2 relaxation times predict progression of knee osteoarthritis", *Osteoarthritis Cartilage*, 21/1 (2013), 69–76, http://www.ncbi.nlm.nih.gov/pubmed/23059757.

[42] Jungmann, P. M. et al., "T_2 relaxation time measurements are limited in monitoring progression, once advanced cartilage defects at the knee occur: longitudinal data from the osteoarthritis initiative", *J Magn Reson Imaging*, 38/6 (2013), 1415–24, http://www.ncbi.nlm.nih.gov/pubmed/24038491.

[43] Mosher, T. J. et al., "Knee articular cartilage damage in osteoarthritis: analysis of MR image biomarker reproducibility in ACRIN-PA 4001 multicenter trial", *Radiology*, 258/3 (2011), 832–42, http://www.ncbi.nlm.nih.gov/pubmed/21212364.

[44] Zarins, Z. A. et al., "Cartilage and meniscus assessment using $T_{1\rho}$ and T_2 measurements in healthy subjects and patients with osteoarthritis", *Osteoarthritis Cartilage*, 18/11 (2010), 1408–16, http://www.ncbi.nlm.nih.gov/pubmed/20696262.

[45] Stahl, R. et al., "MRI-derived T_2 relaxation times and cartilage morphometry of the tibio-femoral joint in subjects with and without osteoarthritis during a 1-year follow-up", *Osteoarthritis Cartilage*, 15/11 (2007), 1225–34, http://www.ncbi.nlm.nih.gov/pubmed/17561417.

[46] Li, X. et al., "In vivo $T_{1\rho}$ and T_2 mapping of articular cartilage in osteoarthritis of the knee using 3 T MRI", *Osteoarthritis Cartilage*, 15/7 (2007), 789–97, http://www.ncbi.nlm.nih.gov/pubmed/17307365.

[47] Koff, M. F., Amrami, K. K., and Kaufman, K. R., "Clinical evaluation of T_2 values of patellar cartilage in patients with osteoarthritis", *Osteoarthritis Cartilage*, 15/2 (2007), 198–204, http://www.ncbi.nlm.nih.gov/pubmed/16949313.

[48] Dunn, T. C. et al., "T_2 relaxation time of cartilage at MR imaging: comparison with severity of knee osteoarthritis", *Radiology*, 232/2 (2004), 592–8, https://www.ncbi.nlm.nih.gov/pubmed/15215540.

[49] Wyatt, C. et al., "Cartilage $T_{1\rho}$ and T_2 Relaxation Times in Patients With Mild-to-Moderate Radiographic Hip Osteoarthritis", *Arthritis Rheumatol*, 67/6 (2015), 1548–56, http://www.ncbi.nlm.nih.gov/pubmed/25779656.

[50] Subburaj, K. et al., "Regional variations in MR relaxation of hip joint cartilage in subjects with and without femoralacetabular impingement", *Magn Reson Imaging*, 31/7 (2013), 1129–36, http://www.ncbi.nlm.nih.gov/pubmed/23684960.

[51] Yamamoto, S. et al., "Quantitative T_2 mapping of femoral head cartilage in systemic lupus erythematosus patients with noncollapsed osteonecrosis of the femoral head associated with corticosteroid therapy", *J Magn Reson Imaging*, 34/5 (2011), 1151–8, http://www.ncbi.nlm.nih.gov/pubmed/21953994.

[52] Nishii, T. et al., "Evaluation of cartilage matrix disorders by T_2 relaxation time in patients with hip dysplasia", *Osteoarthritis Cartilage*, 16/2 (2008), 227–33, http://www.ncbi.nlm.nih.gov/pubmed/17644363.

[53] Fedorov, S., *GetData Graph Digitizer* (version 2.26) [Windows application] (2013), http://getdata-graph-digitizer.com.

[54] Higgins, JPT and Deeks, JJ, "Chapter 7: Selecting studies and collecting data", in *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]*, ed. Higgins JPT and Green S (The Cochrane Collaboration, 2011), http://handbook.cochrane.org.

[55] Welch, B. L., "The generalization of 'Student's' problem when several different population variances are involved", *Biometrika*, 34/1-2 (1947), 28–35, http://dx.doi.org/10.1093/biomet/34.1-2.28.

[56] Brown, MB and Forsythe, AB, "Robust Tests for the Equality of Variances", *Journal of the American Statistical Association*, 69/346 (1974), 364–7, https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1974.10482955.

[57] SciPy Documentation, *scipy.stats.ttest_ind_from_stats*, The SciPy community (2017), https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind_from_stats.html, accessed 6 Mar. 2017.

[58] Cumming, Geoff, *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis* (Multivariate applications series; New York: Routledge, Taylor & Francis Group, 2012), xiv, 519 pages, https://www.routledge.com/Understanding-The-New-Statistics-Effect-Sizes-Confidence-Intervals-and/Cumming/p/book/9780415879682.

[59] Borenstein, Michael, *Introduction to meta-analysis* (Chichester, U.K.: John Wiley & Sons, 2009), xxviii, 421 p.

[60] Coe, R, "It's the effect size, stupid: what effect size is and why it is important.", *Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002. [Available from http://www.leeds.ac.uk/educol/]* (2002).

[61] Ellis, Paul D., *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results* (Cambridge ; New York: Cambridge University Press, 2010), xvii, 173 p. https://effectsizefaq.com/essential-guide-to-effect-sizes/.

[62] "Effect size", Wikipedia (19 Mar. 2018), https://en.wikipedia.org/wiki/Effect_size, accessed 22 Mar. 2018.

[63] Cohen, Jacob, *Statistical Power Analysis for the Behavioral Sciences* (2nd, Hillsdale, N.J. ; Hove: Erlbaum Associates, 1988), xxi, 567 p.

[64] Glass, Gene V., McGaw, Barry, and Smith, Mary Lee, *Meta-analysis in social research* (Beverly Hills: Sage Publications, 1981), 279 p.

[65] Hedges, Larry V., "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators", *Journal of Educational Statistics*, 6/2 (1981), 107.

[66] Wolfgang, Viechtbauer, "Conducting Meta-Analyses in R with the metafor Package", *Journal of Statistical Software, Articles*, 36/3 (2010), 1–48, https://www.jstatsoft.org/v036/i03.

[67] Wolfgang, Viechtbauer, *Reference Manual for Package 'metafor'*, CRAN (22 June 2017), https://cran.r-project.org/web/packages/metafor/metafor.pdf, accessed 21 Nov. 2017.

[68] Schunemann, HJ et al., "Chapter 12: Interpreting results and drawing conclusions", in *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]*, ed. Higgins JPT and Green S (The Cochrane Collaboration, 2011), http://handbook.cochrane.org.

[69] Ruscio, J., "A probability-based measure of effect size: robustness to base rates and other factors", *Psychol Methods*, 13/1 (2008), 19–30, https://www.ncbi.nlm.nih.gov/pubmed/18331151.

[70] Menashe, L. et al., "The diagnostic performance of MRI in osteoarthritis: a systematic review and meta-analysis", *Osteoarthritis Cartilage*, 20/1 (2012), 13–21, http://www.ncbi.nlm.nih.gov/pubmed/22044841.

[71] MacKay, J. W. et al., "Systematic review and meta-analysis of the reliability and discriminative validity of cartilage compositional MRI in knee osteoarthritis", *Osteoarthritis Cartilage* (2018), https://www.ncbi.nlm.nih.gov/pubmed/29550400.

[72] Li, W. et al., "Value of precontrast T_1 for dGEMRIC of native articular cartilage", *J Magn Reson Imaging*, 29/2 (2009), 494–7, https://www.ncbi.nlm.nih.gov/pubmed/19161210.

[73] Palmer, A. et al., "Diagnostic and prognostic value of delayed Gadolinium Enhanced Magnetic Resonance Imaging of Cartilage (dGEMRIC) in early osteoarthritis of the hip", *Osteoarthritis Cartilage*, 25/9 (2017), 1468–77, https://www.ncbi.nlm.nih.gov/pubmed/28506842.

[74] Magnusson, Kristoffer, *Interpreting Cohen's d effect size: an interactive visualization* [web application] (2014), http://rpsychologist.com/d3/cohend/.

[75] McRobbie, Donald W., Moore, Elizabeth A., and Graves, Martin J., *MRI from picture to proton* (Third edition., Cambridge: Cambridge University Press, 2016), 1 volume.

[76] Elster, Allen D. and Burdette, Jonathan H., *Questions & answers in magnetic resonance imaging* (2nd, St. Louis, MA ; London: Mosby, 2001), xiv, 333 p.

[77] Gudbjartsson, H. and Patz, S., "The Rician distribution of noisy MRI data", *Magn Reson Med*, 34/6 (1995), 910–4, http://www.ncbi.nlm.nih.gov/pubmed/8598820.

[78] Andersen, Anders H., "On the Rician distribution of noisy MRI data", *Magnetic Resonance in Medicine*, 36/2 (1996), 331–2.

[79] Aja-Fernandez, S. and Vegas-Sanchez-Ferrero, G., *Statistical Analysis of Noise in MRI: Modeling, Filtering and Estimation* (Springer International Publishing, 2016), https://books.google.co.uk/books?id=5FjHjwEACAAJ.

[80] Feng, Y. et al., "Improved MRI R_2^* relaxometry of iron-loaded liver with noise correction", *Magn Reson Med*, 70/6 (2013), 1765–74, http://www.ncbi.nlm.nih.gov/pubmed/23359410.

[81] Yokoo, T. et al., "Quantitative R_2^* MRI of the liver with rician noise models for evaluation of hepatic iron overload: Simulation, phantom, and early clinical experience", *J Magn Reson Imaging*, 42/6 (2015), 1544–59, http://www.ncbi.nlm.nih.gov/pubmed/25996989.

[82] He, Taigang et al., "An non subjective method for myocardial T_2^* curve fitting in thalassemia", *Journal of Cardiovascular Magnetic Resonance*, 10/Suppl 1 (2008), A107.

[83] He, T. et al., "Myocardial T_2^* measurement in iron-overloaded thalassemia: an ex vivo study to investigate optimal methods of quantification", *Magn Reson Med*, 60/2 (2008), 350–6, http://www.ncbi.nlm.nih.gov/pubmed/18666131.

[84] Otto, R. et al., "Limitations of using logarithmic transformation and linear fitting to estimate relaxation rates in iron-loaded liver", *Pediatric Radiology*, 41/10 (2011), 1259–65, %3CGo%20to%20ISI%3E://WOS:000294690500006.

[85] NumPy Documentation, *numpy.polyfit*, The SciPy community (2017), https://docs.scipy.org/doc/numpy/reference/generated/numpy.polyfit.html, accessed 6 Mar. 2017.

[86] SciPy Documentation, *scipy.optimize.curve_fit*, The SciPy community (2017), https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html, accessed 6 Mar. 2017.

[87] Akcakaya, M. et al., "Improved quantitative myocardial T_2 mapping: Impact of the fitting model", *Magn Reson Med* (2014), http://www.ncbi.nlm.nih.gov/pubmed/25103908.

[88] Miller, A. J. and Joseph, P. M., "The Use of Power Images to Perform Quantitative-Analysis on Low SNR MR Images", *Magnetic Resonance Imaging*, 11/7 (1993), 1051–6, %3CGo%20to%20ISI%3E://WOS:A1993MB03400017.

[89] Feng, Y. et al., "A novel semiautomatic parenchyma extraction method for improved MRI R2* relaxometry of iron loaded liver", *J Magn Reson Imaging*, 40/1 (2014), 67–78, http://www.ncbi.nlm.nih.gov/pubmed/24677406.

[90] Bidhult, S. et al., "Validation of a new T_2^* algorithm and its uncertainty value for cardiac and liver iron load determination from MRI magnitude images", *Magn Reson Med* (2015), http://www.ncbi.nlm.nih.gov/pubmed/26010550.

[91] Sijbers, J. et al., "Maximum-likelihood estimation of Rician distribution parameters", *IEEE Trans Med Imaging*, 17/3 (1998), 357–61, http://www.ncbi.nlm.nih.gov/pubmed/9735899.

[92] Sijbers, J. and Dekker, A. J. den, "Maximum likelihood estimation of signal amplitude and noise variance from MR data", *Magn Reson Med*, 51/3 (2004), 586–94, http://www.ncbi.nlm.nih.gov/pubmed/15004801.

[93] Golub, F. et al., "Estimation of spin-echo relaxation time", *J Magn Reson*, 237 (2013), 17–22, http://www.ncbi.nlm.nih.gov/pubmed/24125956.

[94] SciPy Documentation, *scipy.optimize.minimize*, The SciPy community (2017), https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html, accessed 10 Mar. 2017.

[95] Lauwers, L. et al., "Estimating the parameters of a Rice distribution: a Bayesian approach", *I2mtc: 2009 Ieee Instrumentation & Measurement Technology Conference, Vols 1-3* (2009), 110–3.

[96] Lauwers, L. et al., "Analyzing Rice distributed functional magnetic resonance imaging data: a Bayesian approach", *Measurement Science and Technology*, 21/11 (2010), 115804.

[97] Jaynes, E. T., "Prior Probabilities", *IEEE Transactions on Systems Science and Cybernetics*, 4/3 (1968), 227–41.

[98] Walker-Samuel, S. et al., "Bayesian estimation of changes in transverse relaxation rates", *Magn Reson Med*, 64/3 (2010), 914–21, http://www.ncbi.nlm.nih.gov/pubmed/20806382.

[99] Bouhrara, M., Reiter, D. A., and Spencer, R. G., "Bayesian analysis of transverse signal decay with application to human brain", *Magn Reson Med*, 74/3 (2015), 785–802, https://www.ncbi.nlm.nih.gov/pubmed/25242062.

[100] Wirth, W. et al., "Layer-specific femorotibial cartilage T2 relaxation time in knees with and without early knee osteoarthritis: Data from the Osteoarthritis Initiative (OAI)", *Sci Rep*, 6 (2016), 34202, https://www.ncbi.nlm.nih.gov/pubmed/27670272.

[101] Eckstein, F. et al., "Impact of coil design on the contrast-to-noise ratio, precision, and consistency of quantitative cartilage morphometry at 3 Tesla: a pilot study for the osteoarthritis initiative", *Magn Reson Med*, 57/2 (2007), 448–54, https://www.ncbi.nlm.nih.gov/pubmed/17260363.

[102] Stehling, C. et al., "A novel fast knee cartilage segmentation technique for T_2 measurements at MR imaging–data from the Osteoarthritis Initiative", *Osteoarthritis Cartilage*, 19/8 (2011), 984–9, http://www.ncbi.nlm.nih.gov/pubmed/21515391.

[103] Andersson, Jesper LR, Jenkinson, Mark, and Smith, Stephen, *Non-linear registration, aka spatial normalisation. FMRIB technical report TR07JA1*, Report (2007).

[104] Jenkinson, Mark et al., "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images", *NeuroImage*, 17/2 (2002), 825–41.

[105] Jenkinson, M. and Smith, S., "A global optimisation method for robust affine registration of brain images", *Med Image Anal*, 5/2 (2001), 143–56, https://www.ncbi.nlm.nih.gov/pubmed/11516708.

[106] Talairach, J and Tournoux, P, "Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging" (1988).

[107] Fonov, V. et al., "Unbiased average age-appropriate atlases for pediatric studies", *Neuroimage*, 54/1 (2011), 313–27, https://www.ncbi.nlm.nih.gov/pubmed/20656036.

[108] Tustison, N. J. and Avants, B. B., "Explicit B-spline regularization in diffeomorphic image registration", *Front Neuroinform*, 7 (2013), 39, http://www.ncbi.nlm.nih.gov/pubmed/24409140.

[109] Sled, John G. and Pike, G. Bruce, "Understanding intensity non-uniformity in MRI", in (Medical Image Computing and Computer-Assisted Intervention - MICCAI'98; Springer Berlin Heidelberg), 614–22.

[110] Arnold, J. B. et al., "Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects", *Neuroimage*, 13/5 (2001), 931–43, https://www.ncbi.nlm.nih.gov/pubmed/11304088.

[111] Sled, J. G., Zijdenbos, A. P., and Evans, A. C., "A nonparametric method for automatic correction of intensity nonuniformity in MRI data", *IEEE Trans Med Imaging*, 17/1 (1998), 87–97, https//www.ncbi.nlm.nih.gov/pubmed/9617910.

[112] FSL Documentation, *BET*, Oxford Centre for Functional MRI of the Brain (FMRIB) (2017), https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET, accessed 10 Mar. 2017.

[113] Xie, Y. H. and Ji, Q., "A new efficient ellipse detection method", *16th International Conference on Pattern Recognition, Vol Ii, Proceedings* (2002), 957–60, %3CGo%20to%20ISI%3E://WOS:000177886900234.

[114] scikit-image Documentation, *Circular and Elliptical Hough Transforms*, scikit-image development team (2017), http://scikit-image.org/docs/dev/auto_examples/edges/plot_circular_elliptical_hough_transform.html, accessed 10 Mar. 2017.

[115] FreeSurfer Documentation, *recon-all*, Laboratory for Computational Neuroimaging (2017), http://freesurfer.net/fswiki/recon-all, accessed 10 Mar. 2017.

[116] Sled, J. G., *The MNI_N3 Software Package* (version 1.08) (2004), http://www.bic.mni.mcgill.ca/software/N3/.

[117] MNI_N3 Documentation, *MINC/Tools/N3*, McConnell Brain Imaging Centre of the Montreal Neurological Institute (2017), https://en.wikibooks.org/wiki/MINC/Tools/N3, accessed 10 Mar. 2017.

[118] FreeSurfer Documentation, *mri_nu_correct.mni*, Laboratory for Computational Neuroimaging (2013), http://freesurfer.net/fswiki/mri_nu_correct.mni, accessed 10 Mar. 2017.

[119] Klein, A. et al., "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration", *Neuroimage*, 46/3 (2009), 786–802, http://www.ncbi.nlm.nih.gov/pubmed/19195496.

[120] Yushkevich, Paul A. et al., "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability", *Neuroimage*, 31/3 (2006), 1116–28, http://www.itksnap.org.

[121] Yushkevich, Paul, *Convert3D Medical Image Processing Tool* (version 1.0.0) (2016), http://www.itksnap.org/c3d/.

[122] Murphy, K. et al., "Evaluation of Registration Methods on Thoracic CT: The EMPIRE10 Challenge", *IEEE Transactions on Medical Imaging*, 30/11 (2011), 1901–20.

[123] "Mathematical contributions to the theory of evolution.III. Regression, heredity, and panmixia", *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187 (1896), 253–318, http://rsta.royalsocietypublishing.org/content/roypta/187/253.full.pdf.

[124] Goshtasby, A. Ardeshir, "Similarity and Dissimilarity Measures", in *Image Registration: Principles, Tools and Methods* (London: Springer London, 2012), 7–66, https://doi.org/10.1007/978-1-4471-2458-0_2.

[125] "Pearson correlation coefficient", Wikipedia (31 Mar. 2018), https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, accessed 31 Mar. 2018.

[126] Pearson, Karl, *On the general theory of skew correlation and non-linear regression* (London, Dulau and co., 1905), 54 p.

[127] Woods, R. P., Mazziotta, J. C., and Cherry, S. R., "MRI-PET registration with automated algorithm", *J Comput Assist Tomogr*, 17/4 (1993), 536–46, https://www.ncbi.nlm.nih.gov/pubmed/8331222.

[128] Roche, A. et al., *Multimodal Image Registration by Maximization of the Correlation Ratio*, Report (INRIA, 1998-08 1998), https://hal.inria.fr/inria-00073311.

[129] Roche, A. et al., "The correlation ratio as a new similarity measure for multimodal image registration", *Medical Image Computing and Computer-Assisted Interven-*

tion - Miccai'98, 1496 (1998), 1115–24, https://pdfs.semanticscholar.org/ae72/6fcc91664d3e710ba32429d2f3befb15cf17.pdf.

[130] Maes, F. et al., "Multimodality image registration by maximization of mutual information", *IEEE Trans Med Imaging*, 16/2 (1997), 187–98, https://www.ncbi.nlm.nih.gov/pubmed/9101328.

[131] Viola, Paul and Wells III, William M., "Alignment by Maximization of Mutual Information", *International Journal of Computer Vision*, 24/2 (1997), 137–54, https://doi.org/10.1023/A:1007958904918.

[132] Bouma, Gerlof, "Normalized (Pointwise) Mutual Information in Collocation Extraction", in *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, 31–40, https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf.

[133] Bardera, Anton, Feixas, Miquel, and Boada, Imma, "Normalized similarity measures for medical image registration", in *Medical Imaging 2004*, mmmmmccclxx (SPIE), 11.

[134] Studholme, C., Hill, D. L. G., and Hawkes, D. J., "An overlap invariant entropy measure of 3D medical image alignment", *Pattern Recognition*, 32/1 (1999), 71–86, https://www.sciencedirect.com/science/article/pii/S0031320398000910.

[135] Roche, A. and Brett, M., *nipy/nipy/algorithms/registration/similarity_measures.py* [Python code] (Oct. 2015), https://github.com/nipy/nipy/blob/master/nipy/algorithms/registration/similarity_measures.py.

[136] Gorgolewski, K. et al., "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python", *Front Neuroinform*, 5 (2011), 13, https://www.ncbi.nlm.nih.gov/pubmed/21897815.

[137] Schneider, E. and Nessaiver, M., "The Osteoarthritis Initiative (OAI) magnetic resonance imaging quality assurance update", *Osteoarthritis Cartilage*, 21/1 (2013), 110–6, http://www.ncbi.nlm.nih.gov/pubmed/23092792.

[138] Pedoia, V. et al., "Fully automatic analysis of the knee articular cartilage T1rho relaxation time using voxel-based relaxometry", *J Magn Reson Imaging*, 43/4 (2016), 970–80, https://www.ncbi.nlm.nih.gov/pubmed/26443990.

[139] Li, X. et al., "The first step for neuroimaging data analysis: DICOM to NIfTI conversion", *J Neurosci Methods*, 264 (2016), 47–56, https://www.ncbi.nlm.nih.gov/pubmed/26945974.

[140] Chris Rorden, *dcm2niiX* (version 1.0.20170207) [macOS and Linux application] (7 Feb. 2017), https://github.com/rordenlab/dcm2niix.

[141] Nipype Documentation, *interfaces.dcm2nii*, Neuroimaging in Python Team (2017), http://nipype.readthedocs.io/en/latest/interfaces/generated/nipype.interfaces.dcm2nii.html, accessed 6 Mar. 2017.

[142] OA Biomarkers Consortium FNIH Project Documentation, *Osteoarthritis Biomarkers Consortium FNIH Project: Study Design*, The Osteoarthritis Initiative

(2015), https://oai.epi-ucsf.org/datarelease/docs/FNIH/OaBioFnihDataOverview.pdf, accessed 21 Nov. 2017.

[143] Felson, D. T. et al., "Defining radiographic incidence and progression of knee osteoarthritis: suggested modifications of the Kellgren and Lawrence scale", *Ann Rheum Dis*, 70/11 (2011), 1884–6, https://www.ncbi.nlm.nih.gov/pubmed/21908453.

[144] OAI SAS Dataset Documentation, *Central Reading of Knee X-rays for Kellgren & Lawrence Grade and Individual Radiographic Features of Tibiofemoral Knee OA*, The Osteoarthritis Initiative (2016), https://oai.epi-ucsf.org/datarelease/SASDocs/kXR_SQ_BU_descrip.pdf, accessed 21 Nov. 2017.

[145] OAI SAS Dataset, *OUTCOMES99: Outcomes* (The Osteoarthritis Initiative, 2017), https://oai.epi-ucsf.org.

[146] Bellamy, N. et al., "Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee", *J Rheumatol*, 15/12 (1988), 1833–40, http://www.ncbi.nlm.nih.gov/pubmed/3068365.

[147] Hensor, E. M. et al., "Toward a clinical definition of early osteoarthritis: onset of patient-reported knee pain begins on stairs. Data from the osteoarthritis initiative", *Arthritis Care Res (Hoboken)*, 67/1 (2015), 40–7, http://www.ncbi.nlm.nih.gov/pubmed/25074673.

[148] OAI SAS Dataset, *JointSx: Joint Symptoms/Function* (The Osteoarthritis Initiative, 2017), https://oai.epi-ucsf.org.

[149] Felson, D. T. et al., "A new approach yields high rates of radiographic progression in knee osteoarthritis", *J Rheumatol*, 35/10 (2008), 2047–54, https://www.ncbi.nlm.nih.gov/pubmed/18793000.

[150] Altman, R. D. et al., "Atlas of individual radiographic features in osteoarthritis", *Osteoarthritis Cartilage*, 3 Suppl A (1995), 3–70, http://www.ncbi.nlm.nih.gov/pubmed/8581752.

[151] Altman, R. D. and Gold, G. E., "Atlas of individual radiographic features in osteoarthritis, revised", *Osteoarthritis Cartilage*, 15 Suppl A (2007), A1–56, http://www.ncbi.nlm.nih.gov/pubmed/17320422.

[152] Angst, F., Aeschlimann, A., and Stucki, G., "Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities", *Arthritis Rheum*, 45/4 (2001), 384–91, http://www.ncbi.nlm.nih.gov/pubmed/11501727.

[153] Angst, F. et al., "Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities", *J Rheumatol*, 29/1 (2002), 131–8, http://www.ncbi.nlm.nih.gov/pubmed/11824949.

[154] Gulati, Aashish et al., "Localization of the full-thickness cartilage lesions in medial and lateral unicompartmental knee osteoarthritis", *Journal of Orthopaedic Research*, 27/10 (2009), 1339–46, https://doi.org/10.1002/jor.20880.

[155] Brown, Cameron P. et al., "Imaging and modeling collagen architecture from the nano to micro scale", *Biomedical Optics Express*, 5/1 (2014), 233–43, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891335/.

[156] Abraham, Alexandre et al., "Region segmentation for sparse decompositions: better brain parcellations from rest fMRI" (Sept. 2014), 8, https://hal.inria.fr/hal-01093944.

[157] Pedregosa, F. et al., "Scikit-learn: Machine Learning in Python ", *Journal of Machine Learning Research*, 12 (2011), 2825–30.

[158] "Scikit-learn User Guide", Scikit-learn developers (29 June 2018), https://scikit-learn.org/stable/user_guide.html, accessed 1 Sept. 2018.

[159] Rajpurkar, Pranav et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning", *arXiv preprint arXiv:1711.05225* (2017).

[160] Paszke, Adam et al., "Automatic differentiation in PyTorch", in *NIPS-W* (2017).

[161] Huang, Gao et al., "Densely connected convolutional networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[162] Deng, J. et al., "ImageNet: A large-scale hierarchical image database", in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (June 2009), 248–55.

[163] Kingma, Diederik P and Ba, Jimmy, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980* (2014).

[164] "PyTorch Documentations: BCELoss", Torch Contributors (1 June 2018), https://pytorch.org/docs/stable/nn.html, accessed 1 Sept. 2018.

[165] Bouhrara, M. et al., "Incorporation of Rician Noise in the Analysis of Biexponential Transverse Relaxation in Cartilage Using a Multiple Gradient Echo Sequence at 3 and 7 Tesla", *Magnetic Resonance in Medicine*, 73/1 (2015), 352–66, %3CGo%20to%20ISI%3E://WOS:000346908800035.

[166] Moreland, Kenneth, "Diverging Color Maps for Scientific Visualization", in *Advances in Visual Computing*, ed. George Bebis et al. (Berlin, Heidelberg: Springer Berlin Heidelberg, 2009), 92–103.