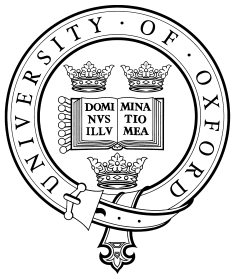
Development and Application of Image Analysis Techniques to Study Structural and Metabolic Neurodegeneration in the Human Hippocampus using MRI and PET

Courtney A. Bishop
Merton College



FMRIB Centre
University of Oxford
Department of Clinical Neurosciences
John Radcliffe Hospital
Oxford

Hilary Term 2012

This thesis is submitted to the FMRIB Centre, Department of Clinical Neurosciences, University of Oxford, for the degree of Doctor of Philosophy. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

To my beloved parents...

Acknowledgements

Firstly, I would like to thank my three supervisors, Dr Mark Jenkinson (FMRIB), Dr Jerome Declerck (Siemens Molecular Imaging), and Dr Dorit Merhof (Konstanz University), for their collective assistance throughout the course of my D.Phil. We predicted that it would be a challenge to juggle a project between three research sites, with so many differing research interests and ideas, but I believe that it has been a huge success. I thank you all for handling this collaboration with immense skill and consideration, to culminate in a mutually-productive, well-rounded body of work that I am proud to present here. Particular thanks go to Mark for providing overall direction and guidance to the project, and for giving me the confidence and opportunity to broaden my academic experiences. To Jerome, I am grateful for your wealth of enthusiasm and ideas, and for the opportunity to work amongst the vibrant and high-achieving employees of Siemens Molecular Imaging. I thank Dorit for continuing to support my D.Phil project after returning to Germany, as well as her hospitality and support on academic visits, and for her continued friendship.

I also owe a great deal to several members of FMRIB: to Gwenaelle Douaud — for supervising my DTC Summer Project; to Professors Irene Tracey and Steve Smith — for their support with grant applications, and for their contributions to FMRIB; to Dr Jesper Andersson — for his fantastic lectures and support with FNIRT; to Dr Giovanna Zamboni — for generously contributing her time and knowledge to this work; to Andri Tziortzi — for her advice regarding the processing of PET data, and for her valued friendship; to Sue Field and Marilyn Goulding — for their impeccable organization of FMRIB and FMRIB events, and for their generosity shown during my D.Phil; and to all remaining members of the FMRIB Analysis Group and the FMRIB Physics Group — for their valued company at conferences and in the research laboratory.

Additionally, I would like to thank all of the Science and Technology Team at Siemens Molecular Imaging, for welcoming me into the office and for providing such a supportive environment to work in. Particular thanks go to Dr Matt Kelly for his constant friendship and support (as well as his sportsmanship at the pool table). I am also indebted to Professor Sir Mike Brady for originally introducing me to the field of image analysis, and for his brilliantly conducted and informative Journal Review Classes.

I wish to acknowledge and thank the EPSRC for funding this research through the Life Sciences Interface Doctoral Training Centre (LSI DTC) Programme, as well as all of the teaching and support staff at the LSI DTC for their efforts in running this fantastic programme. For additional funding, I am grateful to Siemens Molecular Imaging, the Guarantors of Brain, Merton College, and the University of Konstanz. Most crucially, I thank the collaborating research sites that provided clinical data and expertise for this project: (i) the Center for Morphometric Analysis, Massachusetts General Hospital, USA; (ii) Harvard Medical School, Massachusetts, USA; (iii) the Alzheimer's Disease Neuroimaging Initiative; (iv) and the Max-Planck-Institute, University of Leipzig, Germany.

Finally, I would like to thank my family and friends for supporting me throughout my D.Phil. In particular, I would like to thank my parents, Robert and Davina Birch — for always believing in me, for selflessly providing me with the best opportunities in life, and for being the most incredible parents — and my husband, Martin, for his encouragement, his willingness to listen and to assist with my research, but above all, for helping me to smile when times got tough.

Courtney A. Bishop
Merton College

Doctor of Philosophy
Hilary Term 2012

Development and Application of Image Analysis Techniques to Study Structural and Metabolic Neurodegeneration in the Human Hippocampus using MRI and PET

Abstract

Despite the association between hippocampal atrophy and a vast array of highly debilitating neurological diseases, such as Alzheimer's disease and frontotemporal lobar degeneration, tools to accurately and robustly quantify the degeneration of this structure still largely elude us. In this thesis, we firstly evaluate previously-developed hippocampal segmentation methods (FMRIB's Integrated Registration and Segmentation Tool (FIRST), Freesurfer (FS), and three versions of a Classifier Fusion (CF) technique) on two clinical MR datasets, to gain a better understanding of the modes of success and failure of these techniques, and to use this acquired knowledge for subsequent method improvement (e.g., FIRSTv3). Secondly, a fully-automated, novel hippocampal segmentation method is developed, termed Fast Marching for Automated Segmentation of the Hippocampus (FMASH). This combined region-growing and atlas-based approach uses a 3D Sethian Fast Marching (FM) technique to propagate a hippocampal region from an automatically-defined seed point in the MR image. Region growth is dictated by both subject-specific intensity features and a probabilistic shape prior (or atlas). Following method development, FMASH is thoroughly validated on an independent clinical dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI), with an investigation of the dependency of such atlas-based approaches on their prior information. In response to our findings, we subsequently present a novel label-warping approach to effectively account for the detrimental effects of using cross-dataset priors in atlas-based segmentation. Finally, a clinical application of MR hippocampal segmentation is presented, with a combined MR-PET analysis of wholefield and subfield hippocampal changes in Alzheimer's disease and frontotemporal lobar degeneration. This thesis therefore contributes both novel computational tools and valuable knowledge for further neurological investigations in both the academic and the clinical field.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Imaging and analysis techniques . . . 3
- 1.3 Thesis goals . . . 5
- 1.4 Thesis overview . . . 6

2 Background Information **8**
- 2.1 Neuroanatomy . . . 8
- 2.2 Neurodegenerative diseases and affective disorders . . . 14
- 2.3 Neuroimaging: the fundamentals of MRI and PET . . . 18
- 2.4 Image processing techniques . . . 30
- 2.5 Segmentation evaluation: performance measures . . . 40
- 2.6 Clinical data . . . 44

3 Method Comparison and Evaluation **49**
- 3.1 Introduction . . . 50
- 3.2 Materials and methods . . . 54
- 3.3 Results . . . 62
- 3.4 Discussion and conclusions . . . 74

4 Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH) **78**
- 4.1 Introduction . . . 79
- 4.2 Materials and methods . . . 81
- 4.3 Results . . . 92
- 4.4 Discussion and conclusions . . . 101

5 FMASH Validation and Improvement: Accounting for Changes in Data and Labelling Protocol **106**
- 5.1 Introduction . . . 108
- 5.2 Materials and methods . . . 110
- 5.3 Results . . . 114
- 5.4 Discussion and conclusions . . . 123

6 Clinical Application: a Combined MR-PET Analysis of Wholefield and Subfield Hippocampal Changes in AD and FTLD **130**
6.1 Introduction . . . 131
6.2 Materials and methods . . . 136
6.3 Results . . . 145
6.4 Discussion and conclusions . . . 160

7 Conclusions and Future Work **170**
7.1 Thesis summary and key contributions . . . 171
7.2 Future work . . . 177
7.3 Concluding remarks . . . 179

A Computation of the FM Arrival Time Surface **181**
A.1 Methods . . . 181

B Leipzig Data: Scanner and Sequence Variability **183**
B.1 Materials and methods . . . 183
B.2 Results . . . 184
B.3 Conclusions . . . 184

C Hippocampal Metabolism from Uncorrected PET Data **185**
C.1 Statistical analysis . . . 185
C.2 Results . . . 186
C.3 Conclusions . . . 187

List of Figures

2.1 Illustration of a typical neuron. . . . 9
2.2 Illustration of brain lobes, directions and sections. . . . 11
2.3 The limbic system, hippocampal dimensions and subfields. . . . 13
2.4 Illustration of a typical MRI scan. . . . 20
2.5 A T1-weighted MR image of the brain. . . . 21
2.6 Schematic illustration of a typical PET scan. . . . 23
2.7 An FDG-PET image of the brain. . . . 25
2.8 The spatial resolution limit in 1D. . . . 27
2.9 Spillover effects in 2D. . . . 29
2.10 A typical segmentation workflow for the hippocampus. . . . 32
2.11 The boundary problem. . . . 33
2.12 Segmentation performance measures. . . . 41
2.13 Structural MR images. . . . 45

3.1 FMRIB's Integrated Registration and Segmentation Tool (FIRST) modelling. . . . 55
3.2 Freesurfer's Aseg (FS) segmentation estimate. . . . 57
3.3 Classifier Fusion (CF) methods. . . . 58
3.4 Method comparison on the CMA data. . . . 63
3.5 Pairwise comparison of methods on the CMA data. . . . 65
3.6 CMA segmentation estimates. . . . 66
3.7 False positive (FP) and false negative (FN) maps for the CMA data. . . . 67
3.8 Method comparison on the BPSA data. . . . 69
3.9 Pairwise comparison of methods on the BPSA data. . . . 70
3.10 BPSA segmentation estimates. . . . 71
3.11 FP and FN maps for the BPSA data. . . . 73

4.1 FMASH seed point selection. . . . 84
4.2 The Fast Marching (FM) arrival time surface. . . . 85
4.3 FMASH potential maps. . . . 93
4.4 FMASH performance: left hippocampus (LHipp). . . . 95
4.5 FMASH performance: right hippocampus (RHipp). . . . 96
4.6 Method comparison. . . . 98
4.7 False positive (FP) maps. . . . 99

4.8 False negative (FN) maps. 100

5.1 FMASH ADNI validation: comparison of method performance. 116
5.2 FMASH ADNI validation: comparison of false positive (FP) and false negative (FN) maps. 117
5.3 Data and label investigation: dependency of atlas-based methods on hippocampal priors. 119
5.4 Data and label investigation: FMASH false positive (FP) maps. 121
5.5 Data and label investigation: FMASH false negative (FN) maps. 122
5.6 FMASH ADNI validation: comparison of segmentation estimates. 124
5.7 Data and label investigation: thresholded priors and FMASH segmentations. . . 129

6.1 Hippocampal subfields. 140
6.2 Wholefield hippocampal measures. 147
6.3 Vertex analysis with false discovery rate (FDR) threshold. 148
6.4 Vertex analysis without false discovery rate (FDR) threshold. 150
6.5 Linear discriminant analysis. 154
6.6 Hippocampal subfield metabolism. 157
6.7 Pairwise comparisons of subfield metabolism. 158
6.8 Support vector machine (SVM) analysis. 165

C.1 Measures of wholefield hippocampal metabolism from uncorrected PET data. . 186

List of Acronyms

AD	Alzheimer's Disease
ADNI	Alzheimers Disease Neuroimaging Initiative
ANOVA	Analysis of Variance
BET	Brain Extraction Tool
BP	Bipolar Disorder
BPSA	Bipolar Data from San Antonio
CA	Cornu Ammonis
CDR	Clinical Dementia Rating
CF	Classifier Fusion
CFNS	Classifier Fusion with Native Space Selection
CFRS	Classifier Fusion with Reference Space Selection
CMA	Center for Morphometric Analysis
CMRglc	Cerebral Metabolic Rate of Glucose Consumption
CON	Controls
CPM	Combined Performance Measure
CSF	Cerebrospinal Fluid
DG	Dentate Gyrus
DOF	Degrees of Freedom
FDG	18-F Fluorodeoxyglucose
FIRST	FMRIBs Integrated Registration and Segmentation Tool
FIRSTv3	FIRST Variation with a Three-Step Registration
FLIRT	FMRIB's Linear Integrated Registration Tool
FM	Fast Marching
FMASH	Fast Marching for Automated Segmentation of the Hippocampus
FMRIB	Oxfords Centre for Functional MRI of the Brain
FN	False Negative
FNIRT	FMRIB's Nonlinear Integrated Registration Tool
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FS	Freesurfer
FSL	FMRIB's Software Library
FDR	False Discovery Rate
FTD	Frontotemporal Dementia
FTLD	Frontotemporal Lobar Degeneration
FWHM	Full Width at Half Maximum

GM	Grey Matter
LDA	Linear Discriminant Analysis
LHipp	Left Hippocampus
mMG	modified Müller-Gärtner
MMSE	Mini Mental State Exam
MNI152	The Montreal Neurological Institute Standard Brain Atlas
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
NC	Normal Controls
NMI	Normalized Mutual Information
PA or PNFA	Progressive Non-Fluent Aphasia
PET	Positron Emission Tomography
PSF	Point Spread Function
PVE	Partial Volume Effect
PVEC	Partial Volume Effect Correction
RHipp	Right Hippocampus
ROI	Region of Interest
SD	Semantic Dementia
SVM	Support Vector Machines
T	Tesla
TIV	Total Intracranial Volume
WM	White Matter

Nomenclature

Γ_n	Amplitude of local signal variations at voxel n
(i, j, k)	Each of 6 neighbouring grid points of $(i_{min}, j_{min}, k_{min})$
$(i_{min}, j_{min}, k_{min})$	Trial point with smallest U value
I_j	Signal at local neighbouring voxel j
I_n	Signal at any voxel n
j	Local neighbouring voxel
λ	Weighting parameter for prior term of the input potential $\tilde{P}$
Λ	Set of all local spatial minima of the Γ map
m	Most central local minimum and hippocampal seed point p_0
μ_m	Mean signal within V_m
μ_n	Mean signal within V_n
n	Voxels in the hippocampal ROI defined by the spatial prior
p	Points in the image grid
p_0	Hippocampal seed point m
$\tilde{P}$	Input potential
$p(Hipp(n))$	Probability of voxel n in the hippocampal prior
U	Arrival time surface or minimal action
V_m	Local neighbourhood of seed point m
V_n	Local neighbourhood of voxel n

1

Introduction

1.1 Motivation

The brain is the most complex, fascinating and fundamental part of our central nervous system; the function of which faciliates our every existence, from the most basic everyday processes, to our most ingenious thoughts and actions. Like any other organ, however, the brain is susceptible to the degenerative and dysfunctional processes of disease, and when these arise, there can be catastrophic consequences. Aside from the high mortality rates, as well as the physical, mental and emotional turmoil (for both patients and carers), neurological diseases and disorders impose an extraordinarily large social and economic burden worldwide. Consequently, there exists a large and important area of neuroscience focused on understanding the workings of the human brain, and most crucially, how and why it fails when it does. This desire to understand brain degeneration and dysfunction is driven by a need for more reliable patient diagnosis and monitoring, and ultimately, more effective treatment strategies.

To halt or reverse the impacts of neurological disease, and ultimately maintain the function of our precious brain, requires an in-depth understanding of all aspects of the disease(s) over multiple spatial scales — wholebrain changes, affected neural circuits, anatomical structures and sub-structures, neurons, proteins and genes — as well as the associated clinical features. However, interdependence of these scales means that knowledge acquired at any one of them

may provide key information for a comprehensive understanding of the whole disease process.

Here, we focus on a subcortical structure called the hippocampus, because it is known to be involved in many highly-debilitating neurological diseases: Alzheimer's disease (AD), frontotemporal lobar degeneration (FTLD), and medial temporal lobe epilepsy (TLE), to name just a few [Geuze et al., 2005]. In order to establish to what extent the hippocampus (and indeed any other target structure) is involved in such diseases, it must be accurately identified/delineated in a subject with that disease, and subsequently compared to that of a non-diseased subject (or monitored over time). Whilst this might sound invasive, advances in imaging techniques have enabled high-resolution imaging and analysis of the brain *in vivo*, allowing the morphology and function of the hippocampus to be studied non-invasively. Furthermore, such investigations enable neuroscientists to search for characteristic, disease-related hippocampal changes, which (as discussed above) improves our understanding of the whole disease process, for more reliable patient diagnosis, monitoring, and effective therapies.

This thesis is therefore motivated by a clinical need to robustly and accurately analyze the hippocampus in the imaged brains of pathological (and healthy) subjects, and to better understand how this information may be used in patient diagnosis and monitoring. In a sense, there is always a symbiotic relationship between the clinicians and the analysts that study such problems: the clinical manifestations and impacts of the neurological disease (i.e., the clinical problem) provides an analytical problem — that of identifying characteristic imaging markers of disease — whilst a developed image analysis tool (i.e., an analytical solution) may facilitate more reliable patient diagnosis and monitoring in the clinic, thus providing the foundations on which a clinical solution (or treatment) may be found.

1.2 Imaging and analysis techniques

Two particularly effective tomographical imaging techniques, providing images of the structure and function of the human brain, respectively, are conventional magnetic resonance imaging (MRI) and positron emission tomography (PET). Conventional MRI uses the magnetic properties of protons to image the internal structure of the brain *in vivo*, providing high-resolution, anatomical images with good soft-tissue contrast. PET, on the other hand, uses the decay properties of radiolabelled biomolecules to generate images of brain function and physiology in living subjects. A particularly desirable attribute of PET is that the observed signal is intrinsically quantitative — the signal intensity in the image is proportional to the amount of radioisotope present, measured in units of Becquerels per millilitre (Bq/ml). The spatial resolution of PET is, however, somewhat poorer than structural MRI, so the precise location of specific brain structures in PET data is more problematic. For example, our structure of interest (the hippocampus), is at the spatial-resolution limit of current PET scanners, so it is particularly susceptible to associated imaging artifacts (known as partial volume effects), and is often poorly defined in the reconstructed PET image. It is therefore preferrable to identify the hippocampus in structural magnetic resonance (MR) images, owing to their superior spatial resolution and good soft-tissue contrast. Collectively then, conventional MRI and PET provide complementary structural and functional information for more detailed understanding of disease-related hippocampal changes — highlighting the importance of multi-modal image analysis. It is worth noting that functional MRI techniques also faciliate the study of brain activity, but PET retains the significant advantage of being able to image particular neurotransmitter systems (using radiolabelled receptor ligands that bind to specific brain receptors) as well as brain metabolism.

The identification and classification of geometrical structures in imaging data is known as *segmentation*. Traditionally, segmentation would be performed manually on the structural MR images, by expert raters trained in the task of delineating the hippocampus. However, this manual intervention requires knowledge of a particular segmentation protocol (of which there are

many [Konrad et al., 2009]) and it is extremely labour-intensive, thus impractical for large-scale studies and clinical trials. Consequently, the last two decades have seen the emergence of many semi-automated and automated segmentation methods for this purpose (although expert manual labels are still frequently used as a gold standard (reference) for initial training and validation of these methods).

As well as the obvious time- and labour-saving benefits of automated segmentation methods, the results of any given method are often highly reproducible, providing a means of consistently defining the hippocampus in a large clinical dataset, without the intra- and inter-rater variability introduced by human intervention. The main problem, therefore, is the accuracy of these automated segmentation methods, which is controlled by several limiting factors. Firstly, image artifacts (such as signal inhomogeneities, noise, and partial volume effects) make it a challenging task to automatically segment any medical image. Secondly, the automated segmentation of the hippocampus is particularly difficult because of the complex shape of this structure, high inter-subject and disease variability, as well as borders of relatively poor contrast in the MR image. In light of these image artifacts and computational challenges, hippocampus segmentation algorithms for MR data generally cannot rely on intensity information alone. Instead, prior information about the target anatomical structure is used to guide the segmentation process. This prior information (also known simply as a *prior*, an *atlas*, or a *template*) captures common attributes of the structure in a population, such as the spatial distribution and/or intensity profiles. The prior can then be used, for example, to reduce the space of possible segmentations for the hippocampus. The dependency of the hippocampal segmentation methods on this prior information does, however, mean that inaccuracies or bias in the prior could be passed on to the resultant segmentations. As a possible consequence of this, automated hippocampal segmentation tools tend to perform well on healthy subject data (from which the prior information is typically established), but they often fail to capture the substantial and specific atrophy present in clinical datasets.

In addition to the difficulties associated with accurate segmentation of the hippocampus, most hippocampal segmentation methods are validated on a specific, *single* dataset, so their cross-dataset performance (as well as their relative performance compared to other methods) is generally unknown. There is also uncertainty over the clinical utility of the computed hippocampal measures. Collectively, these shortcomings may be limiting the use of hippocampal measures in clinical practice.

1.3 Thesis goals

First and foremost, the main goal of this thesis is to develop novel computational tools for accurately and robustly segmenting the hippocampus in structural MR images. A prerequisite of these methods is that they should be robust to the changes in image quality, disease-status and demographics that typically characterize differing clinical datasets. A combination of subject-specific intensity features and prior information (in the form of spatial distributions and/or intensity profiles) will be used throughout. To achieve this main goal, we aim to initially perform a thorough evaluation of existing hippocampal segmentation methods — to understand their cross-dataset performance, and their modes of success/failure — so that we may use the acquired information to both improve upon these methods, and to guide the development of a more accurate, novel segmentation method. Subsequent goals include deriving and analyzing measurements of hippocampal morphology (i.e., size and shape) and functionality (i.e., tissue glucose utilization) from these MR hippocampal segmentations and corresponding PET data, respectively, to better inform our understanding of certain neurodegenerative dementias and to assess the possible clinical utility of these measures. Achieving such goals will reduce the uncertainty surrounding the relative computational accuracy of hippocampal segmentation methods and the clinical use of the derived hippocampal measures, with the hope that they will eventually be incorporated into routine clinical practice — for patient diagnosis, monitoring, and assessment of treatment efficacy.

1.4 Thesis overview

The outline of this thesis, and how the above-mentioned goals will be addressed, is now described. Following this introduction, Chapter 2 presents the background information required for a comprehensive understanding of the remainder of the thesis, including some basic brain anatomy and a description of the studied neurological diseases/disorders, as well as the fundamentals of neuroimaging and image processing techniques. Particular emphasis is placed on the computational tools required for the work in this thesis, including those used to assess the performance accuracy of the hippocampal segmentation methods. To close the chapter, we present details of the particular clinical data utilized in this thesis.

In Chapter 3, a review of previous work in the field of hippocampal segmentation is initially given, before several existing hippocampal segmentation methods are presented, evaluated, and improved upon. The review highlights the need for a thorough evaluation of developed hippocampal segmentation methods, across multiple clinical datasets, and using a unified set of measures for assessment of segmentation accuracy/performance. This need is largely met in the subsequent method comparison and evaluation study, which reveals the relative performance, strengths and weaknesses, of several current hippocampal segmentation methods. Furthermore, from an understanding of the modes of failure of these segmentation methods, mechanisms for improvement are proposed, which are then put into practice.

Building on the knowledge acquired in the previous chapter, Chapter 4 presents the development and initial validation of a novel, fully-automated hippocampal segmentation method for MR data. This approach, termed FMASH (Fast Marching for Automated Segmentation of the Hippocampus, [Bishop et al., 2011]), uses a 3D Sethian Fast Marching (FM) technique [Deschamps and Cohen, 2000] to grow a hippocampal region from an automatically-defined seed point in the MR image. The growth of the hippocampal region is dictated by both image-derived intensity features and atlas-derived prior information. In the closing section of this chapter, the

performance of FMASH is compared to similar, recently-developed hippocampal segmentation methods, and possible extensions to the FMASH algorithm are proposed.

Chapter 5 provides a thorough investigation of the dependency of atlas-based hippocampal segmentation methods on their training data (i.e., the prior information). Firstly, our previously-developed automated segmentation method, FMASH, is further validated on a clinical dataset that is completely independent from its training data. Secondly, FMASH and a simple atlas-based segmentation method are implemented using prior information established from one or both of two datasets, to better appreciate how prior information can introduce bias into the final segmentations. By understanding this bias, we are also able to propose a novel approach to account for it.

Broadly speaking, Chapter 6 investigates the potential clinical utility of hippocampal segmentation methods. From MR-derived hippocampal segmentations, morphological (size and shape) properties of the hippocampus are computed, and functional (metabolic) information is extracted from corresponding PET data. Further computational analysis of these hippocampal measures allows us to explore: (i) whether derived measures of hippocampal morphology and metabolism can better inform our understanding of certain neurodegenerative dementias; and (ii) whether one or both of these measures can be used as a complementary biomarker for more reliable clinical diagnosis of subjects to a particular dementia syndrome.

Conclusions are drawn at the end of each chapter, but Chapter 7 draws the various threads together to summarize the main findings of the thesis and its contribution to the field, before extensions and longer-term directions of the work are proposed. As a final close to Chapter 7, and the thesis as a whole, we provide some concluding remarks.

2

Background Information

This Chapter provides some background information which will be useful for a comprehensive understanding of the remainder of the thesis. It includes an introduction to both wholebrain and hippocampal anatomy, some key neurodegenerative diseases, the fundamentals of imaging and image processing techniques, as well as details of the clinical data used in subsequent chapters.

2.1 Neuroanatomy

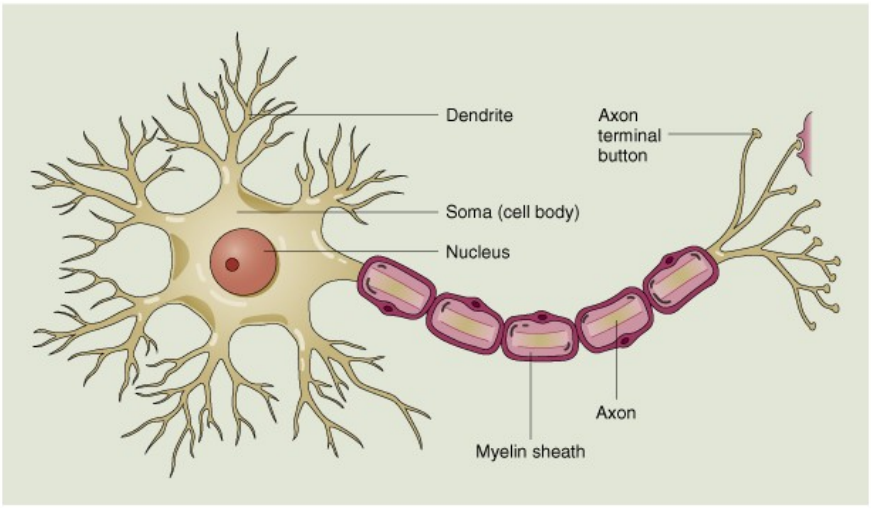
This section focuses on providing a background in the neuroanatomy that underlies the images that we are aiming to segment. We will initially review the basic structure of the central nervous system (CNS), followed by the main subdivisions of the brain, before finally focusing on the hippocampus itself.

2.1.1 The brain

The brain (cerebrum, cerebellum, brainstem) and the spinal cord constitute the CNS — a complex network of nerve cells (neurons) and supporting neuroglial cells that ultimately forms the basis for all sensory processes, perception, cognition, and behaviour. Neurons are the information processing units of the CNS, receiving and transmitting electrochemical signals throughout

the brain via synaptic terminals. Figure 2.1 shows an illustration of a typical neuron. Incoming information is received at the neuronal dendrites (highly-branched fibres projecting from the cell body) and directed along the length of the neuron to the origin of the neuronal axon (a single, slender fibre projecting from the cell body, contralateral to the dendrites), where it is transmitted to target cells (typically other neurons) in the cellular ensemble (or circuit). Axons bundled together into long cable-like structures make-up the body's nerves. Many of these axons are covered by a lipid-rich wrapping called a myelin sheath, which electrically insulates the axons so that they can conduct nerve impulses much more rapidly. This myelin sheath also gives axons a fatty white appearance. Consequently, two gross histological terms are used to describe regions that are either rich in cell bodies or rich in axons: these are grey matter (GM) and white matter (WM), respectively.

Figure 2.1 *Illustration of a typical neuron.* (Image courtesy of fourthingsabout.blogspot.com)



On a global scale, the brain consists of a highly folded outer layer of cortical GM, which overlays an abundance of WM tracts and deep (subcortical) GM nuclei buried further within. The brain is divided into two main hemispheres (left and right) which are connected by a thick central band of nerve tracts called the corpus callosum. Each of these hemispheres is then conventionally subdivided into four main regions or lobes (frontal, parietal, occipital, and tem-

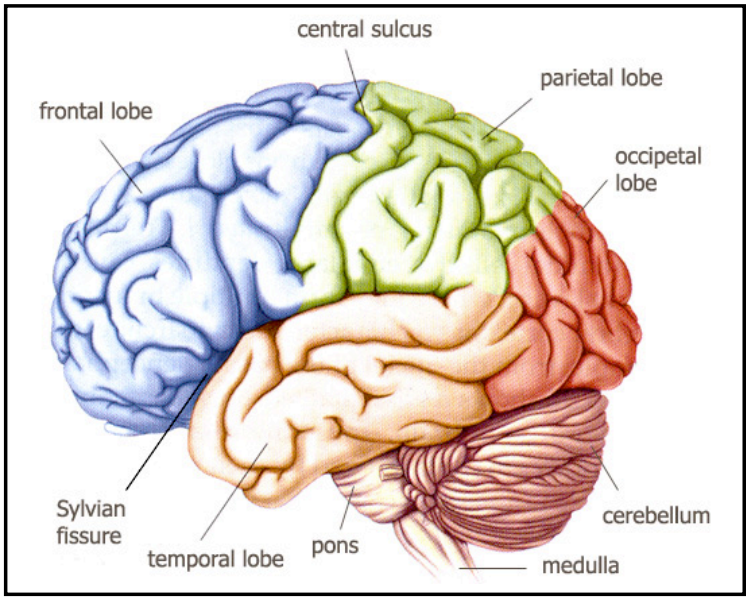
poral) as depicted in Figure 2.2a, and so-called because of the skull bones above them. In order to fully describe the spatial organization of these four brain lobes, and indeed the spatial organization and orientation of all substructures in the brain (such as the later-mentioned hippocampus), some gross anatomical terminology is required. Thus, anterior and posterior indicate the front and back; dorsal (or superior) and ventral (or inferior), the top and bottom; and medial and lateral, the midline or the side. Each of these spatial notations is depicted in Figure 2.2b. Cross-sections through the brain can then be defined relative to these spatial notations (also shown in Figure 2.2b). An axial (horizontal or transverse) section separates dorsal and ventral brain regions; a sagittal (mid-sagittal or paramedian) section divides the two hemispheres; and a coronal (or frontal) section is in the plane of the face, separating anterior from posterior. As the name suggests, the frontal lobe is situated at the front of the brain. It is separated from the more posterior parietal lobe by a fold called the central sulcus. The temporal lobe extends almost as far anterior as the frontal lobe, but is ventral to it. The divide between the temporal and the frontal lobes is demarcated by a deep fissure called the lateral (or Sylvian) fissure. Finally, the occipital lobe is found at the back of the brain, separated from the parietal lobe by the parieto-occipital sulcus, and concealing the cerebellum beneath.

2.1.2 The hippocampus

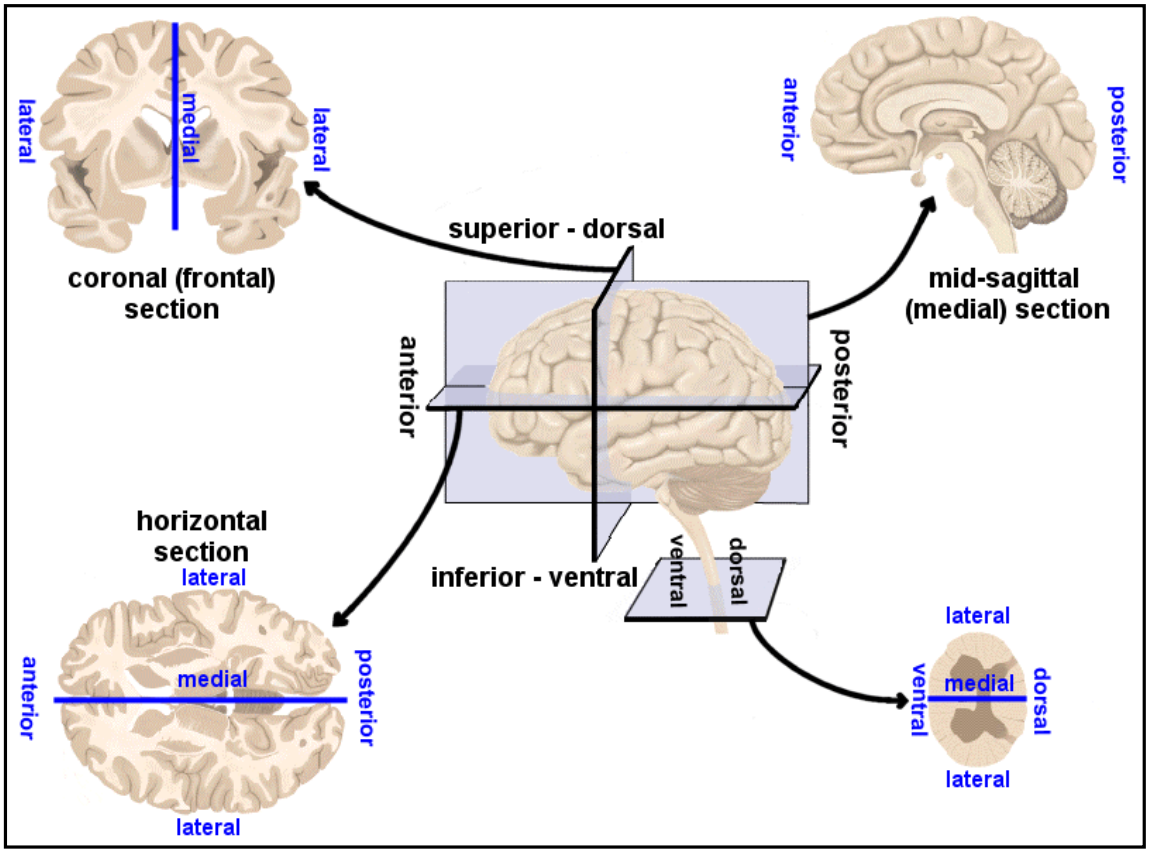
The hippocampus belongs to the limbic system of the brain (illustrated in Figure 2.3a) and is located inside the medial temporal lobe, bordering the lateral ventricles, the thalamus and the amygdala. The limbic system controls emotional behaviour and short-term declarative memory (i.e., that which is available to consciousness and can be expressed/declared by language). The major role of the hippocampus in the short-term storage and establishment of new declarative memories was first brought to light by the study of amnestic patients, where a difficulty storing and retrieving declarative information was found to depend on the integrity of a subset of limbic circuits, and in particular, the hippocampus.

Figure 2.2 *Illustration of brain lobes, directions and sections.* (a) The brain has four main lobes (frontal, parietal, occipital, and temporal), and the boundaries between these lobes are demarcated by a fold or fissure in the cortex. (b) The spatial notations used to describe the human brain, including the three main sections or slices through it. (Images courtesy of neuro.sofiatopia.org and homepage.smc.edu)

(a)



(b)

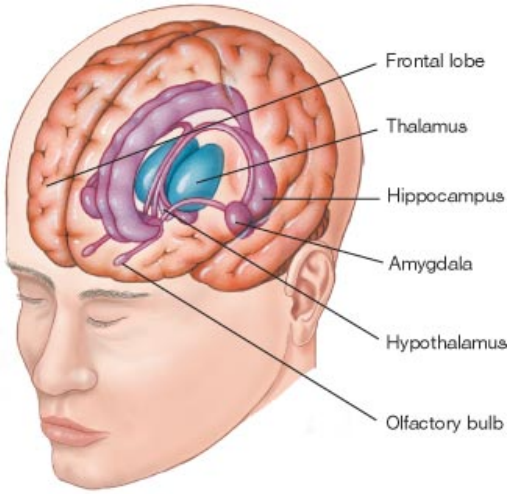


Crucially involved in episodic and spatial memory processes, the hippocampus structure has therefore been implicated in the pathophysiology of many debilitating neurological and psychiatric diseases [Delacourte et al., 1999; Geuze et al., 2005; Konrad et al., 2009]; hippocampal volume loss is a characteristic feature of Alzheimer's Disease (AD), frontotemporal lobar degeneration (FTLD) and temporal lobe epilepsy (TLE), with more ambiguous findings of volume decreases in schizophrenia (SCZ) and in affective disorders such as major depression (MD) and bipolar disorder (BP). The name "hippocampus" stems from the Greek words for horse (hippos) and sea monster (kampos), so-called because the curvature of the structure, typically seen in post-mortem studies and high-resolution magnetic resonance (MR) images, resembles the tail of a seahorse. An illustration of the hippocampus proper — its typical shape and dimensions — is given in Figure 2.3b. Viewed in the sagittal plane, the hippocampus is broadly divided into three sections (along its anteroposterior axis): the hippocampal *head* and the hippocampal *tail* constitute the most anterior and posterior portions of the hippocampus, respectively, whilst the hippocampal *body* refers to the middle section of this subcortical structure.

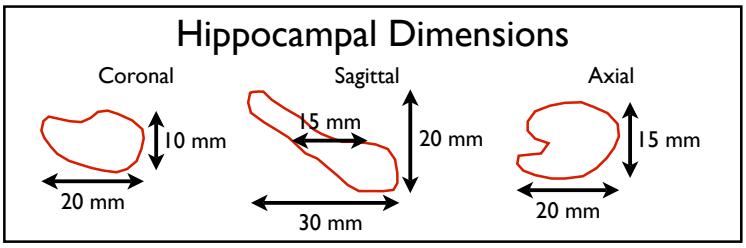
Considering the finer-scale details of the hippocampus, several functionally and molecularly distinct subregions can be identified *in vitro* and with high-resolution focal imaging techniques, revealing a laminar organization of discrete hippocampal subfields. Here, we define the hippocampus as consisting of five main subfields, each shown in Figure 2.3c — the four cornu ammonis fields (CA1, CA2, CA3, CA4) and the dentate gyrus (DG) — with a smallest dimension of typically 2mm for any given subfield. As for all brain structures, the hippocampus does not function in isolation; it forms part of a broader network of structures called the "hippocampus formation", which includes the pre-defined hippocampus, the subiculum, presubiculum, parasubiculum, and the entorhinal cortex (EC). Predominant unidirectional circuitry is somewhat unique to the hippocampus formation, and can be viewed most easily in the axial (or transverse) plane, as illustrated in Figure 2.3c. In the "trisynaptic pathway", the EC passes input to the DG, which is turn passes to the CA3, to the CA1, and then to the subiculum. The subiculum

Figure 2.3 *The limbic system, hippocampal dimensions and subfields.* (a) The hippocampus forms part of the middle arc of the limbic system, which collectively controls emotional behaviour and memory. Patients with hippocampal damage typically experience amnesia — a difficulty establishing and retrieving memories. (b) Typical dimensions for the hippocampus (in mm) are illustrated in the coronal (left), sagittal (middle) and axial (right) plane. (c) In the axial plane, the subdivisions of the hippocampus can be clearly identified, including the cornu ammonis (CA) fields 1-4 and the dentate gyrus (DG). The trisynaptic circuit — connecting the DG, CA3, CA1, and the subiculum (Sub) — is illustrated by the bold arrows. (Image (a) courtesy of cwx.prenhall.com)

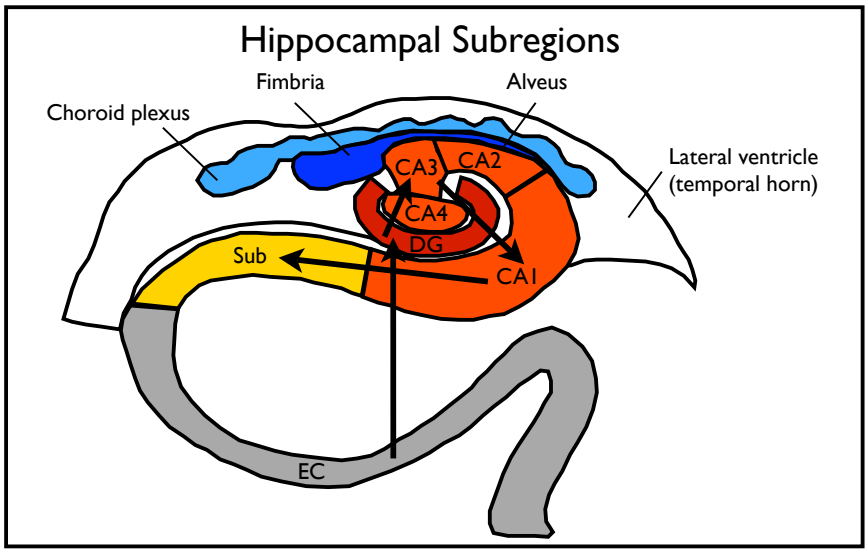
(a)



(b)



(c)



therefore provides the main hippocampal outflow.

2.2 Neurodegenerative diseases and affective disorders

From an understanding of the neuroanatomy and vocabulary presented above, we can describe the clinical features and global patterns of brain change typically found in each of the diseases and disorders studied in this thesis: the neurodegenerative dementias known as Alzheimer's disease (AD) and frontotemporal lobar degeneration (FTLD), and the affective disorder that is bipolar disorder (BP).

2.2.1 Alzheimer's disease (AD)

AD is the most common cause of dementia, causing debilitating neurodegeneration of individuals worldwide. It is typically a disease of the elderly, with the likelihood of developing dementia roughly doubling every five years after the age of 65 (www.alz.co.uk/research/world-report). Of the aforementioned diseases associated with hippocampal atrophy, AD therefore presents one of the largest socio-economic impacts: Alzheimer's Disease International estimate that there are 35.6 million people currently living with dementia worldwide (www.alz.co.uk/research/world-report), increasing to 65.7 million by 2030, and costing the global economy US$604 billion per year. Consequently, considerable effort is being made to understand the pathophysiological mechanisms of AD in order to further the development of effective treatment strategies for the disease.

AD is a progressive disease, characterized by initial episodic memory impairment (i.e., difficulty remembering specific events, situations and experiences) and loss of attention, followed by a failure of language skills, visual-spatial orientation, abstract thinking and judgement. Eventually, it is inevitable that some aspects of personality may also be affected. An initial diagnosis of AD is typically made based on the identification of these characteristic clinical features,

but solid confirmation requires evidence of the following three features (either from specific neuroimaging/sampling techniques, or a post-mortem examination of the brain):

1. **neurofibrillary tangles** — intraneuronal lesions composed of hyperphosphorylated (i.e., abnormal) forms of the microtubule-stabilizing tau protein;

2. **amyloid plaques** — extracellular deposits of abnormal protein (mainly amyloid beta, Aβ) processed from the amyloid precursor protein (APP);

3. **diffuse loss of neurons** — particularly in the neocortex, limbic structures (hippocampus, amygdala, and associated cortices), and selected brain-stem nuclei (such as the basal forebrain nuclei).

In particular, the loss of neurons gives rise to a well-described pattern of global GM atrophy; initially affecting the medial and then the lateral temporal lobes. There is, however, a discrepancy between this global pattern of GM atrophy and the (functional) metabolic profile observed in AD patients: dysfunction of the glucose metabolic system, on the contrary, is characterized by early involvement of the posterior cingulate cortex (PCC), subsequently spreading to the neighbouring precuneus and temporoparietal neocortical areas [Chételat et al., 2003; Meguro et al., 2001; Mevel et al., 2007; Nestor et al., 2003; Schroeter et al., 2009; Villain et al., 2010]. The most widely accepted hypothesis for this discrepancy is a disconnection or "diaschisis" hypothesis, whereby the specific alteration of hippocampal and parahippocampal neurons (involved in connecting the hippocampal formation with other brain structures) leads to reduced blood flow, and reduced metabolic activity, in the distant posterior areas such as the PCC [Jobst et al., 1992]. AD post-mortem studies have shown that the hippocampus is affected early in the course of the disease, with approximately 20-50% loss of neurons by the time that individuals are moderately affected [Bobinski et al., 1997].

2.2.2 Frontotemporal lobar degeneration (FTLD)

The typical age of onset of FTLD is lower than that of AD, with the greatest prevalence in those younger than 65 years. However, elderly subjects may also be affected by FTLD. On the whole, FTLD is therefore one of the most common causes of dementia worldwide.

As the name suggests, FTLD is broadly associated with atrophy of the frontal and temporal lobes of the brain, but it encompasses a heterogeneous group of patients with differing clinical, pathological and genetic characteristics. According to the clinical diagnostic criteria established by Neary et al. [1998], FTLD can be divided into three subtypes — progressive non-fluent aphasia (PA, or PNFA), semantic dementia (SD), and frontotemporal dementia (FTD) — each identified by a specific pathological and neuropsychological profile: symmetric, bilateral frontal lobe atrophy with behavioural changes and disproportionate executive dysfunction in FTD; asymmetric (left>right), bilateral atrophy in SD, most marked in the anterior temporal neocortex, coupled with disproportionate semantic memory impairment (i.e., word comprehension, name finding, face and/or object recognition); and asymmetric atrophy in PA, mainly affecting the left frontotemporal lobe, together with disproportionate language breakdown that gives rise to nonfluent aphasia (expression by speech or writing) in the absence of severe amnesia. Additionally, imaging studies have shown reduced glucose utilization (i.e., reduced brain activity) in the frontal and prefrontal cortices, anterior cingulate regions and anterior temporal lobes in FTD [Diehl et al., 2004; Garraux et al., 1999; Ishii et al., 1998; Jeong et al., 2005; Schroeter et al., 2008]. By comparison, similar studies in SD have found reduced brain activity in more temporal regions [Diehl et al., 2004; Drzezga et al., 2008; Mummery et al., 1999; Schroeter et al., 2007], with some spatial overlap of the pattern of temporal hypometabolic findings in AD and SD [Drzezga et al., 2008].

In a similar manner to AD, clinical diagnosis of FTLD (and its constituent subtypes) is typically made based on a positive presentation of the above clinical features and characteristic

brain pathology. In some cases, complementary genetic information may also be considered. Confirmation of this clinical diagnosis can then be made at post-mortem with the identification of three main histological subtypes:

1. **FTLD-tau** — presenting with intraneuronal lesions composed of abnormal tau protein (similar to AD);
2. **FTLD-TDP43** — broadly characterized by the presence of ubiquitin (a regulatory protein signalling protein recycling) and a DNA-binding protein (TDP-43), with an absence of tau pathology;
3. **FTLD-FUS** — having intranuclear and cytoplasmic inclusions of the RNA-binding protein FUS, as well as neuritic threads, present in the hippocampus, medulla and cortex.

2.2.3 Bipolar disorder (BP)

BP is an affective (i.e., emotional) disorder with a typical age of onset between 15-25 years. The worldwide prevalence of this mental disorder is thought to be in the region of 0.3-1.5%, although values vary depending on nationality and diagnostic criteria [Weissman et al., 1996].

BP is characterized by persistent alterations in mood, varying from episodes of severe depression to periods of extreme irritability and hyperactivity (or mania), which affects thoughts, emotions, and behaviours. Periods of mania typically alternate with those of depression, but both can occur simultaneously, and the severity of this mania is used to distinguish between two types of BP (I and II). The mania phase can last from days to months, whilst depressive moods and low self-esteem are a daily occurrence. Consequently, there is an elevated risk of suicide in those suffering from BP.

With the aforementioned clinical features, it is perhaps no surprise to learn that there is increased engagement of the limbic system of the brain in BP. There are reports of GM changes in the prefrontal and premotor cortex, limbic structures (such as the hippocampus and the amygdala),

the thalamus, striatum, and the cerebellum [Strakowski et al., 2002; James et al., 2011; Lisy et al., 2011; Watson et al., 2012], but the findings are mixed (suggesting an increase, a decrease, or no change in GM volume). Additionally, WM alterations are observed in the anterior corpus callosum (connected with the prefrontal cortices) and the limbic structures [James et al., 2011; Watson et al., 2012], indicating disturbed limbic-cortical connectivity in BP.

Finally, it is worth noting that all of the above-mentioned brain changes, associated with the diseases and disorders herein, are in addition to those typially observed in normal ageing (which is itself the most common example of neurodegeneration and hippocampal atrophy). There are also other common types of dementia, including vascular dementia (VD) and dementia with lewy bodies (DLB).

In the next section of this chapter, we present two advanced imaging techniques that have enabled such a comprehensive understanding of the global structural and functional brain changes associated with the diseases and disorders herein.

2.3 Neuroimaging: the fundamentals of MRI and PET

There are two main tomographic imaging modalities used to assess the internal structure and function of the human brain: magnetic resonance imaging (MRI) and positron emission tomography (PET). Both modalities are invaluable in the study of imaging biomarkers for the detection and diagnosis of neurological and psychiatric disorders. In this section, we initially provide an introduction to MRI and PET, describing the fundamental processes involved and the typical images produced, before focusing on their clinical utility, resolution limits and associated image artifacts.

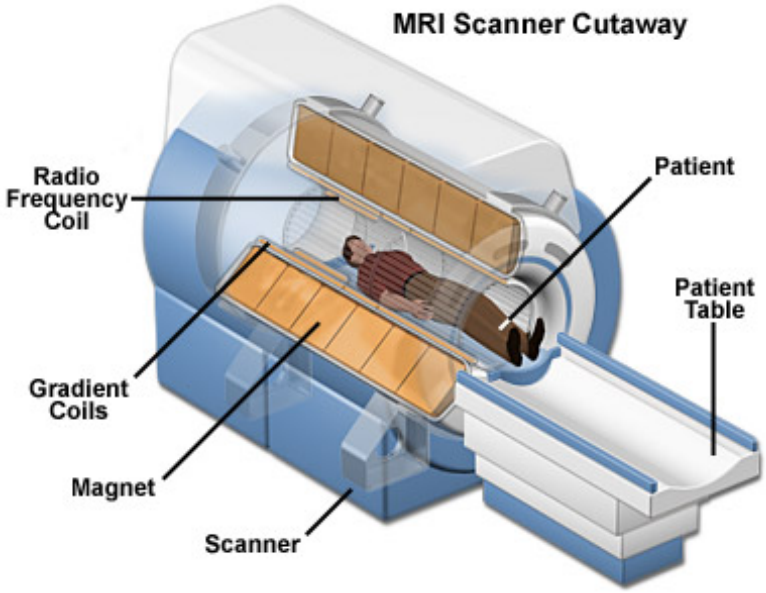
2.3.1 Magnetic resonance imaging (MRI)

MRI is a completely non-invasive tomographic imaging technique that uses the magnetic properties of protons (typically hydrogen nuclei) and externally measured nuclear magnetic resonance (NMR) signals to image the internal structure and workings of the brain *in vivo*. A nucleus with nonzero spin (i.e., intrinsic angular momentum) has a nuclear magnetic moment, which in the presence of an applied magnetic field, precesses about this field such that the net magnetization is aligned with it. The precessing nucleus can be tilted or "excited" out of alignment by a brief radiofrequency (RF) pulse perpendicular to the main magnetic field. The effect of the RF pulse is to generate an oscillating field perpendicular to the main field, which can be detected by receiver coils surrounding the target nucleus. The excited nucleus will then emit energy ($\sim$$10^6$-$10^8$ Hz) as it returns to the alignment imposed by the external magnetic field — a process known as *relaxation*. The frequency of the oscillating field (Larmor frequency) detected by the external receiver is proportional to the magnetic field strength. Therefore, in the presence of a magnetic field gradient (i.e., spatially varying magnetic field strength), the detected frequency of the oscillating nucleus can be used to infer its spatial location — this is the basis of MRI.

In a typical MRI scan (Figure 2.4), the subject is placed inside a large cylindrical magnet that produces a strong and uniform, static magnetic field (termed B_0) along the length of the scanner, with a typical field strength of 1.5-7 Tesla. Three sets of gradient magnets oriented along orthogonal axes are used to generate spatially-varying magnetic field gradients. These static magnetic fields align the atomic nuclei within the tissue of the subject to produce a net nuclear magnetization (M). A radiofrequency transmitter coil then generates a rotating magnetic field (termed B_1) that perturbs the static magnetic field, provides the required excitation and transverse magnetization (M_{xy}), resulting in oscillatory radiofrequency signals. These signals are detected by the receiver coil and subsequently converted into an electrical signal for encoding of the final magnetic resonance (MR) image. Specifically, sampling of the time-varying electri-

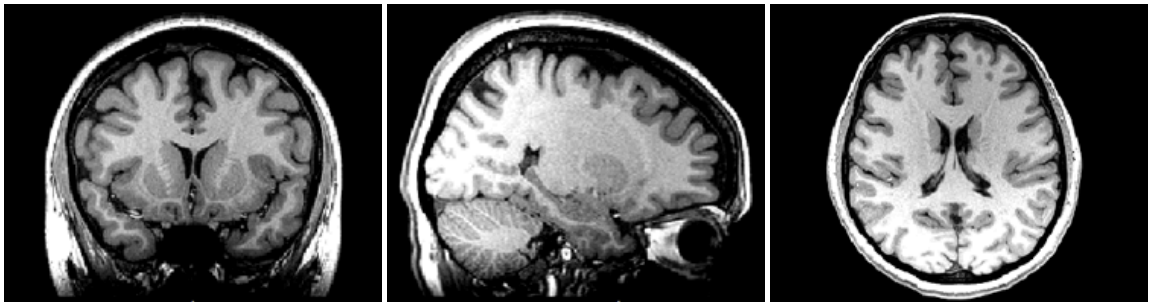
cal signal $S(t)$ is equivalent to sampling the k-space signal $S(k)$, where k is a spatial frequency variable, and a reconstruction method (an inverse Fourier transform) is required to generate the desired image-space data (i.e., the MR image, $I(x)$).

Figure 2.4 *Illustration of a typical MRI scan.* (Image courtesy of Wikimedia Commons)



The measured MRI signal depends on the proton density, two relaxation constants, T1 and T2, as well as the relative timing of the external RF pulses and the magnetic field gradients. Proton density, T1 and T2 are all tissue-specific properties. In particular, T1 is a measure of the time required for the net magnetization to return to equilibrium after excitation. MRI pulse sequences that are designed to be sensitive to this tissue property result in structural T1-weighted images. These T1-weighted MR images are especially good at distinguishing between GM and WM tissue, as well as the cerebrospinal fluid (CSF) that fills the neighbouring ventricles. On a normal T1-weighted MR image (such as that shown in Figure 2.5), the brightest, highest intensity features are WM, the dark grey regions are GM, and the very dark areas at the centre of the image are CSF. Typically, a three-dimensional (3D) image of the brain is acquired, which is then processed with the appropriate image analysis tools. In this thesis, T1-weighted images (generated with conventional MRI) are the only MR images used.

Figure 2.5 *A T1-weighted MR image of the brain.* The bright white matter (WM) in the image is surrounded by the dark grey cortical grey matter (GM) along the outside edge of the brain. Subcortical structures, such as the hippocampus, are located towards the centre of the image and vary in intensity from dark grey to almost as bright as the WM. The very dark area at the centre of the image is cerebrospinal fluid (CSF). The intensity bias along the anterior-posterior line of the brain is an imaging artifact (discussed in Section 2.3.4).



The typical resolution of a structural MR image, acquired on a 3T scanner, is 1 mm isotropic. Considered together with the typical dimensions of the human hippocampus (illustrated in Figure 2.3b), it is apparent that conventional MRI provides us with sufficient resolution to image the hippocampus *in vivo*. Furthermore, the hippocampal atrophy that typically characterizes the aforementioned neurological and psychiatric disorders is thought to occur on a spatial scale large enough to be detected with structural MRI [Bobinski et al., 1996], providing us with a 'window of opportunity' for early diagnosis and preservation of neurological function at the highest possible level. Early diagnosis of such diseases may therefore be made through morphological (size and shape) change in the hippocampus, requiring accurate, reproducible delineation of this region-of-interest (ROI) in structural MR images; a task complicated by its complex shape, large inter-subject variability and fairly poor contrast at the hippocampus-amygdala border.

Finally, it is worth noting that other MRI sequences exist. Whilst conventional MRI samples spin density and proton relaxation times to distinguish between different tissue types (GM, WM, CSF), several variants exist: diffusion-weighted imaging (DWI) measures diffusion coefficients to quantify the diffusion of water molecules in the brain, and functional MRI (fMRI) uses the paramagnetic properties of haemoglobin to measure localized blood oxygen-level dependent (BOLD) changes in response to external stimuli. As such, MRI is a highly powerful imaging

technique for studying both the structure and function of the human brain.

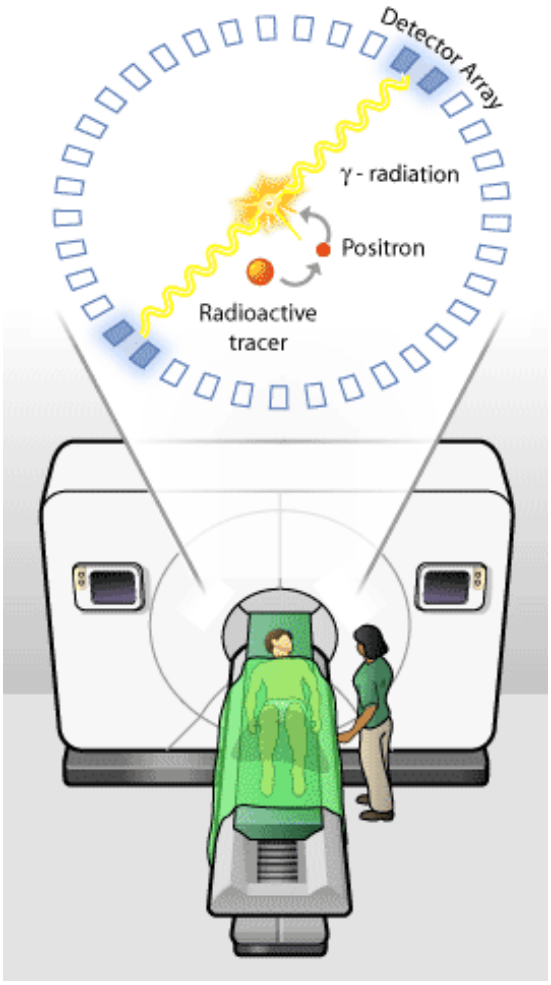
2.3.2 Positron emission tomography (PET)

PET is another valuable *in vivo* tomographic imaging technique, using radioisotopes to measure local biochemical and pharmacological parameters relatively noninvasively in both animals and humans. Radioisotopes, which decay by positron emission, are typically produced in a cyclotron and used to label compounds of biological interest. Each labelled compound, termed a PET tracer, typically contains $\sim 10^{13}$-10^{15} radioactive molecules. When placed inside the subject of interest, this radioactive PET tracer is analogous to a beacon, with the location and brightness of the observed radioactive decay signal providing information about both the location and concentration (i.e., utilization) of the labelled biological molecule in the body. Naturally, the choice of PET tracer determines the types of intrinsic molecular processes that can be observed.

As the radioisotope undergoes positron emission decay (also known as positive beta decay), it emits a positron, the antimatter counterpart of an electron. Each positron has a very short lifetime in electron-rich material such as tissue, travelling only a short distance ($\sim 10^{-1}$-10^{-2}cm) before combining with an electron to form a hydrogen-like state known as a *positronium*. This state lasts for $\sim 10^{-10}$s before a process known as *annihilation* occurs, where the mass of the electron and the positron is converted into electromagnetic energy in the form of gamma photons. The positron and the electron are almost at rest when they collide, with a net momentum close to zero. Since momentum, as well as energy, must be conserved during the annihilation process, two photons are emitted simultaneously in opposite directions ($\sim 180^{\circ}$ apart) as illustrated schematically in Figure 2.6. Simultaneous or coincident detection of this pair of photons, contralateral to one another, can be used to infer the approximate location of the radioactive source within the subject, and forms the basis of PET imaging. Furthermore, the number of photon pairs (or events) arising from a given spatial location correlates with the amount of PET

tracer, providing information about the relative utilization of the metabolically-active, labelled molecule in that particular tissue.

Figure 2.6 *Schematic illustration of a typical PET scan.* (Image courtesy of www.scq.ubc.ca)



A typical scan involves intravenous injection of the PET tracer, followed by a waiting period, where the metabolically-active tracer becomes concentrated in the tissue(s) of interest. The subject is then placed in the imaging scanner (also shown in Figure 2.6) and data is collected over a series of time points, for a total duration of around 1 hour (depending on the tracer used and the number of time frames collected). The scanner itself consists of a ring (or several rings) of detectors coated in a scintillator material, which creates a burst of light (i.e., optical photons) upon detection of a gamma photon. This light is then detected by photomultiplier tubes or

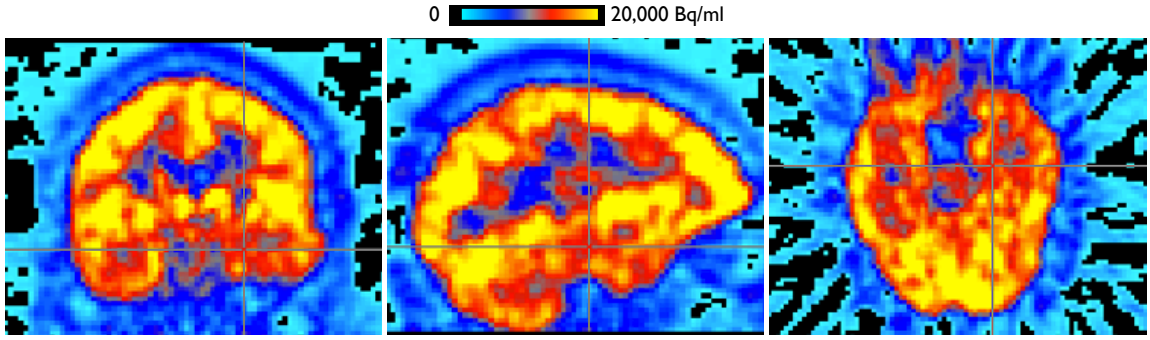
silicon avalanche photodiodes (Si APD), and converted to an electrical signal, before being fed to the subsequent electronics. Since PET imaging depends on the simultaneous or coincident detection of a pair of annihilation photons (generated by the annihilation process), those which do not arrive in pairs (i.e., within a few nanoseconds of each other) are ignored.

The most common PET tracer is 18-F fluorodeoxyglucose (FDG), an analogue of glucose, for which the pre-scan waiting period is approximately 1 hour and the scan period is typically 40-60 minutes. During this time, 10^6-10^9 events (decays) will typically be detected. These events are corrected for a number of factors and then reconstructed into a 3D tomographic image using mathematical algorithms, such as Filtered Back Projection (FBP). The spatial distribution of 18-F FDG (and indeed any other radiolabelled tracer) is mapped quantitatively in the subject of interest, since the signal intensity in any particular voxel (3D pixel) is proportional to the amount of radioisotope (i.e., activity) in that voxel, measured in units of Becquerels per millilitre (Bq/ml). This value will vary depending on the amount of the injected tracer and the dynamics of that tracer (including tissue uptake/washout, and the radioactive half-life). By collecting data over a series of time points during the scan period, the specific uptake curve (or time activity curve, TAC) of the tracer is sampled, providing information on tissue type and functionality. This dynamic data can either be analyzed directly, or more commonly, a mean sum image is generated by averaging the signal intensity over a chosen time frame (e.g., 40-60 mins). An example of such a mean sum FDG-PET image of the human brain is presented in Figure 2.7, showing the spatially-varying glucose utilization of different structures and tissue types. GM and WM regions of high metabolic activity consume large quantities of glucose, which translates into bright, high intensity features on the FDG-PET image. On the contrary, areas of reduced glucose consumption (such as CSF) appear as darker areas on the FDG-PET image.

The typical resolution of an FDG-PET image is 3-4mm isotropic, which is much poorer than the aforementioned T1-weighted MR images. As such, the hippocampus is much more difficult to

identify in FDG-PET images. However, the functional information provided by PET is highly complementary to the anatomical information obtained from other modalities such as structural MRI. Compared to fMRI, PET is a much more established imaging technique for the assessment of brain dysfunction in neurological diseases, and in particular, dementia syndromes. Lower regional brain metabolism was observed in dementia patients almost 30 years ago, and has been used in routine clinical practice for many years since. Functional MRI, on the other hand, is a relatively modern imaging technique (first developed in the 1990's), that has yet to establish itself in the routine assessment of neurodegenerative diseases. One major disadvantage of fMRI (compared to PET) is that the subject has to remain extremely still during the scan, because tiny movements can obscure and ruin the fMRI data. Additionally, fMRI data is relatively difficult to interpret. PET data, on the other hand, is less affected by subject movement in the scanner, and PET retains the significant advantage of being able to image particular neurotransmitter systems (using radiolabelled receptor ligands that bind to specific brain receptors) as well as brain metabolism. The complementarity of conventional MRI and PET, as well as the maturity of these two imaging techniques, thus motivates their inclusion in this thesis. Furthermore, the current development of hybrid MR-PET scanners promises new opportunities (and analytical requirements) in the field of MR-PET.

Figure 2.7 *An FDG-PET image of the brain.* Areas of high brain activity and high glucose utilization appear bright red (i.e., hot), whilst regions of reduced brain activity are dark blue (i.e., cold). The streaking around the outside of the brain is an imaging artifact (discussed in Section 2.3.4).



2.3.3 The changing utility of MRI and PET

Stemming from the differing and complementary information provided, MRI and PET have traditionally served different, but complementary, clinical purposes in neuroimaging. For example, conventional MRI has traditionally been used for treatment monitoring, since the high level of anatomical detail (provided by structural MR images) facilitates the precise monitoring of tissue growth or shrinkage. On the contrary, the majority of PET brain scans have previously been aimed at diagnosis. More recently, however, data mining and machine learning techniques (used to discover and learn patterns of change in large datasets) have enabled the quantification of disease-specific changes in a population, thereby making great inroads towards using MRI for individual subject classification and clinical diagnosis. Likewise, there is an increased demand to use the quantitative properties of PET imaging for patient monitoring as well. This shift in the clinical utility of MRI and PET is clearly dictated by advances in image processing and data analysis techniques for both imaging modalities, but also by advances in image quality (amounting to improved spatial and temporal resolution as well as better signal-to-noise (SNR) ratio). We now talk a little more about the inherent resolution limits of imaging systems, as well as a common and problematic imaging artifact that arises from this.

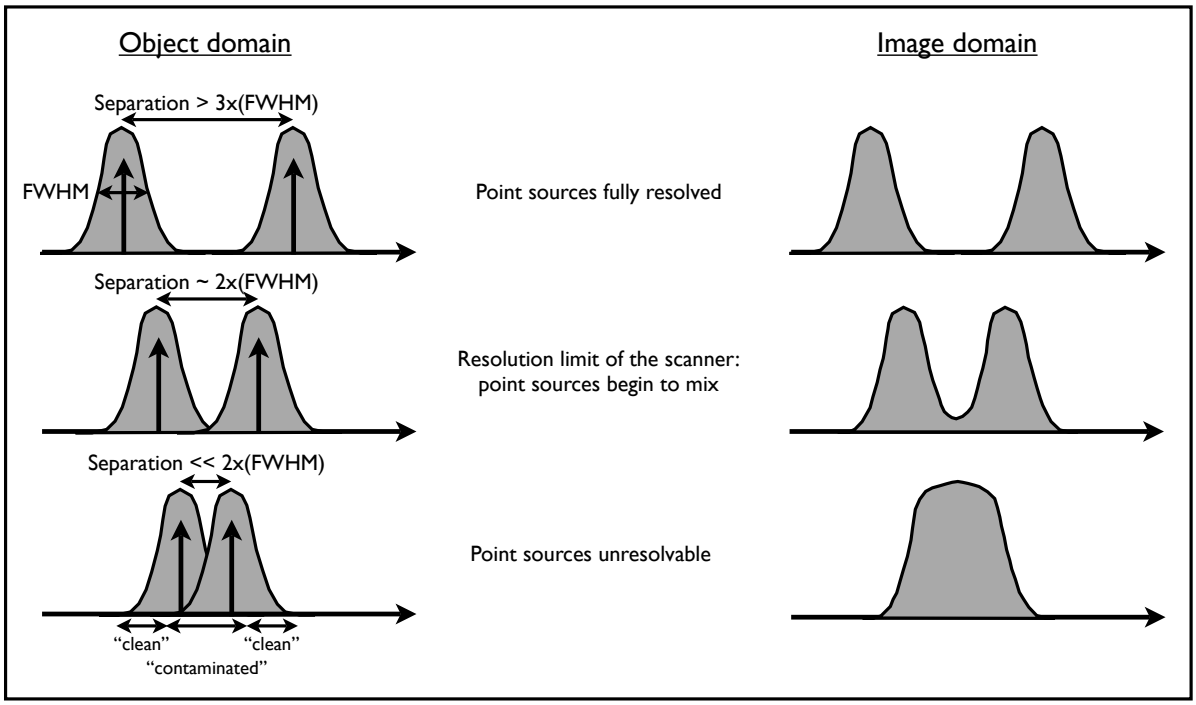
2.3.4 Image resolution and imaging artifacts

When imaging an object with any imaging system/scanner, the observed image is never an exact replica of the original object, because there are always sources of error. Such errors can arise from the limited resolution of the imaging system, scanner and physiological noise, as well as image artifacts. One resolution-limiting factor is the point spread function (PSF) of the scanner, and the partial volume effect (PVE) is a common image artifact associated with the image resolution, both of which will be described in more detail here.

The point spread function (PSF)

The spatial resolution of an imaging system is formally described as the smallest separation δx of two point sources that is necessary for them to be resolvable (i.e., separately identified) in the resultant image, and this is largely determined by the PSF of the scanner. If we initially consider just one point source, the PSF of the scanner has the effect of generating a blurred feature in the observed image. Mathematically, the PSF is essentially a smoothing kernel $h(x)$, which convolves with the object function $I(x)$ to generate the blurred output image $\hat{I}(x)$.

Figure 2.8 *The spatial resolution limit in 1D.* Due to the smoothing effects of the scanner's point spread function (PSF), point sources in the object domain (left side) are seen as smoothly distributed signals in the image domain (right side). When the point sources are separated by less than 2-3 times the full-width-at-half-maximum (FWHM) of the PSF, the signal from these point sources begins to mix, until they are no longer resolvable in the image domain — here, we say that the original "clean" signal has been "contaminated".



The amount of blurring/smoothing in the image is proportional to the effective width (or more specifically, the full-width-at-half-maximum, FWHM) of $h(x)$, and only point sources separated by more than 2-3 times the FWHM of $h(x)$ can be successfully resolved in the output

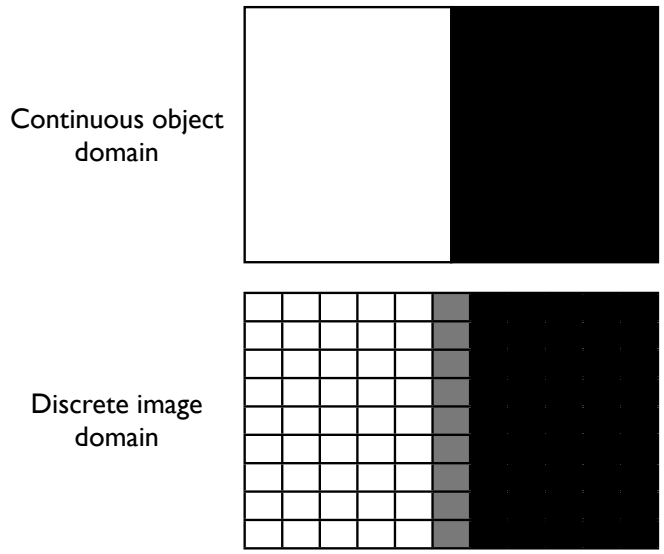
image. When this criterion is not met, the observed image is a mixture of the signal from the two point sources (as illustrated in Figure 2.8) and there is an overall loss of information. For example, the FWHM of a PET scanner is in the region of 5-8mm, and the smallest dimension of the hippocampus in any given plane is typically 10-20mm, so this subcortical structure is at the spatial resolution limit of current PET scanners.

The partial volume effect (PVE)

We have just seen that loss of signal is a direct consequence of the spatial resolution (i.e., PSF) of an imaging system in the object domain — the imaging system can blur information from sources that partially occupy the PSF of the scanner. Additionally, as a consequence of forming a discrete image from the continuous object space, there is mixing/contamination of signals from neighbouring features in the image domain (otherwise known as *spillover* effects in PET imaging). The signal in each discrete image compartment (i.e., voxel) is a weighted average of all of the signals contributing to that compartment, with the relative weighting determined by the fraction of the compartment occupied by that tissue. Both of these effects (i.e., information loss as a result of partial occupancy of the PSF in the object domain and signal mixing/spillover effects in the image domain) are a consequence of limited spatial resolution, and are collectively known as partial volume effects (PVEs).

We can illustrate signal mixing in the image domain by considering the boundary between two adjoining objects of highly-differing signal intensity (Figure 2.9). In the continuous object domain, the white and black objects are perfectly defined, but in the discrete image domain, half of the central column is occupied by the white object and the other half is occupied by the black object, so that the output image is a greyish mixture of the signals from the two objects. In a T1-weighted MR image of the human brain, such signal mixing/contamination effects can make it extremely difficult to determine the difference between a pure GM compartment and one that is a mixture of bright WM and dark CSF.

Figure 2.9 *Signal mixing effects in 2D.* In the continuous object domain (top) the boundary between the white and black objects is perfectly defined, but in the discrete image domain (bottom) the boundary becomes a greyish mixture of the two objects.



Similarly, in FDG-PET images, where there is poorer spatial resolution and correspondingly greater PVEs, the glucose utilization observed in a given image compartment can either be greater or less than the true activity of the region being imaged, owing to spill-in effects (from surrounding tissue of higher activity) or spill-out effects (into surrounding tissue of lower activity), respectively. It is therefore extremely difficult to distinguish between true variations in glucose utilization (i.e., hyper- or hypo-metabolic findings resulting from true physiological brain changes), and those that are merely artifacts of differing morphology and associated spillover effects.

Other imaging artifacts

As well as the above-mentioned PVEs, image distortions or artifacts can arise from noisy, inaccurate or insufficient data, and/or errors in the image reconstruction. For example, variations or fluctuations in the RF signal during MR imaging causes variation in the proton spin tilts and in the sensitivity of the scanner to detect the emitted RF signal, resulting in gradual intensity drifts in the reconstructed MR image. These *bias field* effects can be clearly seen in Figure 2.5,

where the absolute intensity/brightness of the MR image decreases from the posterior to the anterior portion of the brain. Additionally, in MRI, *ringing* artifacts appear as a series of lines in the MR image, parallel to sharp boundaries. There are several possible causes of these ringing artifacts, including: (i) an overshoot or ringing of the Fourier series occurring at discontinuities (i.e., the Gibbs phenomenon); or (ii) finite sampling (and missing of high-frequency data) in k-space (although this is relatively unusual because the reconstruction algorithms deal with the finite sampling issues quite well). *Aliasing* and *motion* are other imaging artifacts, evident in both MR and PET images. There are two common types of aliasing: firstly, undersampling of k-space in MR imaging leads to wrap-around effects, whereby the object appears to warp around the edges of the image (e.g., with the subject's forehead truncated on one side of the image and re-appearing again on the other); secondly, angular undersampling in both MR and PET image reconstruction results in a streaking artifact in the output image (e.g., Figure 2.7). Motion artifacts arise from respiratory and cardiac motion, as well as gross movement of the subject in the scanner, which manifest as image blurring and ghosting (misalignment) in both MR and PET images. PET imaging also suffers from signal *attenuation* artifacts — a reduction in the number of true coincidence events due to absorption and scatter of the annihilation photons in the body.

2.4 Image processing techniques

Having described the fundamentals of MRI and PET, we now turn to the image analysis techniques used to process and extract quantitative information from these diagnostic brain images. Whilst careful and methodical qualitative evaluation of data (by visual assessment) is always recommended, it is quantitative information that provides solid evidence of characteristic brain abnormalities in the studied neurological and psychiatric disorders. As such, it is this quantitative information that proves invaluable in the search for disease biomarkers, and ultimately, in the clinical diagnosis and monitoring of patients suffering from these neurological and psychi-

atric disorders.

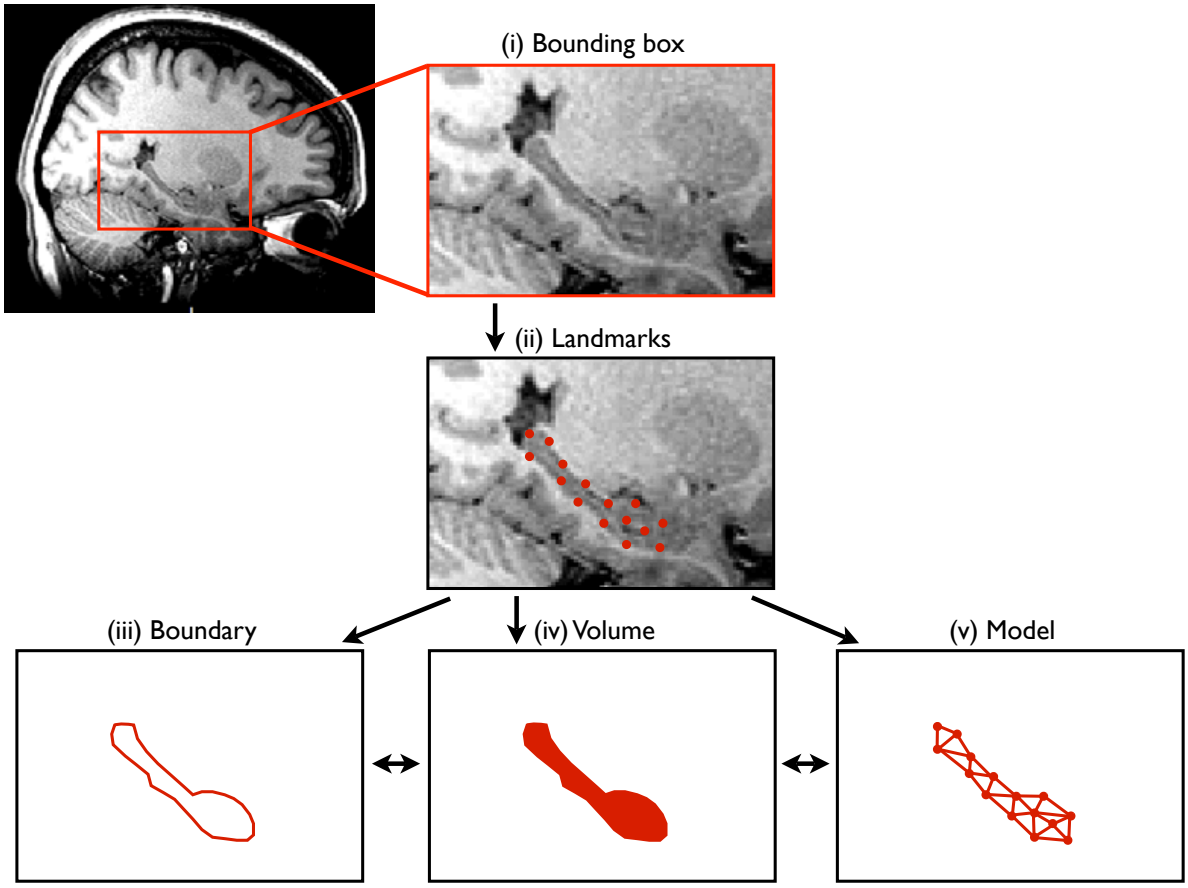
2.4.1 Segmentation

In image analysis, segmentation refers to the identification and classification of geometrical structures in imaging data; typically the selection of landmarks or the delineation of a region of interest (ROI) which has some defining characteristic (e.g., anatomy (brain, heart, liver), MR intensity, radiotracer activity). Segmentation of clinical data is crucial for the quantification of ROIs, facilitating the search for disease biomarkers, diagnosis of patients and treatment planning. For example, if we wish to segment, quantify, and model the hippocampus, starting from a wholebrain T1-weighted MR image, we may implement the following steps, illustrated in Figure 2.10: (i) identification of a bounding box encompassing the hippocampus and surrounding structures (amygdala, ventricles, cortical GM and WM); (ii) generation of hippocampal-specific landmarks for basic quantification (width, length etc.) and subsequent isolation of the hippocampus proper; either as (iii) a boundary contour, (iv) a region-based, volumetric output, or even (v) a parameterized model of the hippocampal surface. For region-based segmentation, the output may either be a 'hard' segmentation (with boundary voxels either completely included or completely excluded), or a 'soft' segmentation (with boundary voxels shared between neighbouring classes). To demonstrate this, we present the well-known boundary problem below.

The boundary problem

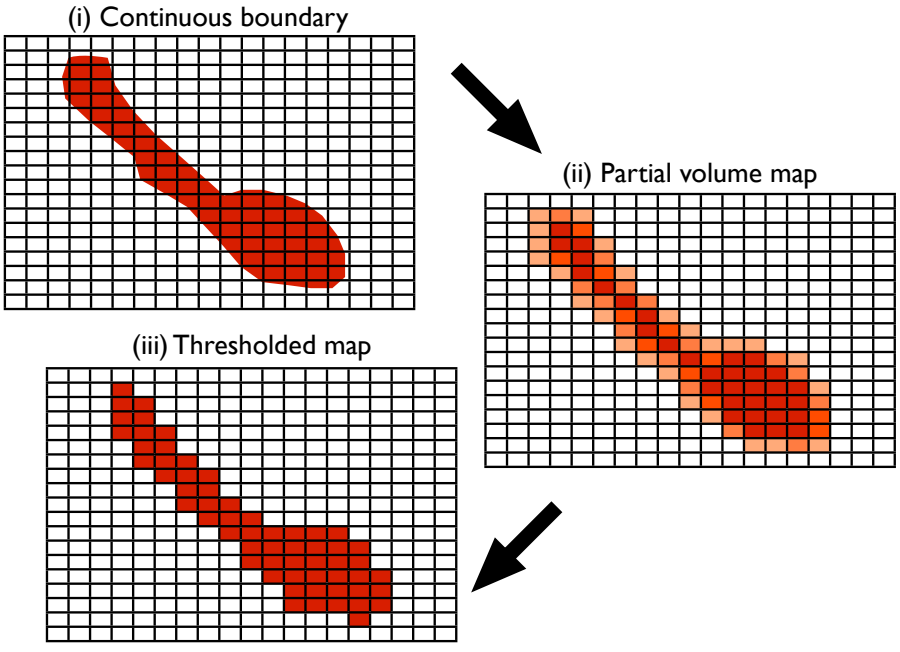
With the definition of a continuous boundary contour in a discrete image space, we are faced with the well-known boundary problem illustrated in Figure 2.11. In a sense, this is the reciprocal of the PVE problem described in Section 2.3.4. We have two types of voxels: interior voxels (which are completely contained within the hippocampal contour), and boundary voxels (which are partially occupied by the contour). There is a certain fractional value associated with each

Figure 2.10 *A typical segmentation workflow for the hippocampus.*



voxel, arising from the partial volume of that voxel contained within the hippocampal contour, which can be used to generate a hippocampal partial volume map (i.e., a spatial map showing the partial volume of each voxel belonging to the hippocampus). With a fraction of 1 assigned to all of the interior voxels, and a fraction <1 assigned to each of the boundary voxels, the hippocampal partial volume map provides a soft segmentation of the hippocampus. One method of generating a hard, binary segmentation of the hippocampus is the application of a fraction threshold (i.e., an inclusion/exclusion criteria) to the partial volume map, retaining only those voxels that have a partial volume fraction greater than the applied threshold. For example, a fraction threshold of 0.5 retains only those voxels with more than a 50% partial occupancy (i.e., with 50% or more of their volume within the hippocampal contour).

Figure 2.11 *The boundary problem.* In region-based segmentation, the representation of (i) a continuous contour in discrete image space results in two types of voxels (interior and exterior). (ii) A partial volume map can be generated according to the fraction of each voxel contained within the contour, and subsequent thresholding of this partial volume map gives (iii) a hard, binary segmentation.



Conventionally used manual segmentation can only provide hard segmentations, with the expert rater choosing to include or exclude discrete image voxels in the final segmentation. Furthermore, as mentioned in the thesis introduction (Chapter 1), manual segmentation requires expert

knowledge (both of the ROI and the chosen labelling protocol) and is extremely time consuming, thus impractical for large-scale clinical studies. For example, it can take an expert rater up to two hours to manually segment a single hippocampus on a high-resolution T1-weighted MR image [Morey et al., 2009; Hogan et al., 2000]. Consequently, the last decade has seen a surge in the development of semi-automated and automated segmentation methods for this purpose. The task of automatically segmenting the hippocampus, is, however, complicated by image artifacts, the complex shape of this structure, high inter-subject and disease variability, as well as borders of relatively poor contrast in the structural MR image. Therefore, intensity features alone generally cannot provide sufficient information for robust and accurate, automated segmentation of the hippocampus. Rather, prior (shape and intensity) information is established from a training dataset, generating a computational atlas/template of common attributes that can be used to guide the subsequent segmentation process. For example, shape distributions in a population may be represented by a probabilistic spatial prior, where each voxel value can be considered as either (i) the average percentage of that voxel contained within the hippocampal contours, or (ii) the percentage of the hippocampal segmentations that classify that voxel as hippocampus (rather than non-hippocampus).

To establish the common attributes of the hippocampus in a population requires the alignment of corresponding image features in multiple (training) images — a process known as *registration*, which will be described in more detail in the following section. Variability in the training images arises from natural anatomical variations between subjects, as well as differences in image quality. However, if the image registration is poor, the variability of common attributes within a population can be misrepresented, so that the prior information is misinformed (i.e., inaccurate and biased). Furthermore, if prior spatial information is established (with or without intensity profiles), registration is typically required to align the target (to-be-segmented) image with the training data. Inaccuracies in this alignment process will usually result in misguided segmentation and poorer segmentation results. Therefore, without accurate image registration,

segmentation methods generally cannot reap the benefits offered by prior information.

On the one hand, then, prior information is extremely valuable for accurate, automated segmentation of the hippocampus, but the dependency of the segmentation on the validity of this prior information can also be the downfall of a segmentation method. Considering the latter further, we first note that prior information is established from a training dataset, where the "true" segmentation of the hippocampus is already "known". By this, we mean that the common attributes of our target anatomical structure are established from a set of MR images that have already been manually segmented by one or more expert raters, under the assumption that these expert manual segmentations (or labels) provide the most truthful representation of the underlying image anatomies. However, there are two issues with this: (i) the wealth of different manual labelling protocols currently adopted for segmentation of the hippocamus on MR images [Konrad et al., 2009] means that there is uncertainty over what constitutes a "truthful" representation of the hippocampus; (ii) even if a single manual labelling protocol is adopted, intra- and inter-rater variability means that, at best, the segmentations deviate from the chosen protocol by 10-20%. Secondly, we note that only attributes present in the training data can be captured/learnt and subsequently used as prior information for segmentation of other MR images. Therefore, if this prior information (e.g., mean hippocampal size and shape) is not representative of the images to be segmented, the hippocampal segmentation estimates may be poor.

Since we typically do not want to make strong assumptions about the target (to-be-segmented) image, the prior information is often designed to be as general as possible — establishing common attributes from many images, acquired using the same scan protocol (e.g., T1-weighted, or T2-weighted), but with differing brain anatomy, intensity, and/or image contrast. The impact of the manual labelling protocol on the prior information and the final hippocampal segmentations is, however, not yet fully understood. Clearly, the more dependent a hippocampal segmentation method is on the prior information, the more the final segmentation estimates are likely to be

affected by the above issues associated with these priors. In general, automated hippocampal segmentation tools tend to perform well on healthy subject data, obtaining comparable results to manual labels, but they often fail to capture the substantial and specific atrophy displayed in clinical data. Reasons for this may include the limited availability of suitable clinical datasets (for method training), as well as the difficulty in capturing the complex morphological (size and shape) changes and inter-subject variability typically present in clinical populations.

2.4.2 Registration

Registration is the process of aligning two or more images in order to achieve the best possible anatomical, functional, or other feature correspondence. A mathematical function (or spatial transform, T) is estimated that maps points from one image to homologous points in another image. If we wish to monitor a given subject's brain over time, or make between-subject comparisons, we need the images in question to be aligned (i.e., registered) to one another. We say that the resultant images are in the same image space. Typically, one image is chosen as the fixed target (or reference) image to which all other (moving) images will be registered. When an image is registered to another, some degree of image degradation will occur, owing to the fact that the moving image will need to be rearranged/transformed (often translated, rotated, scaled and skewed) and resampled/interpolated in order to conform to the (position, orientation, size and shape of the) target, reference image.

Degrees of freedom

Broadly speaking, there are three different registration algorithms commonly employed, each categorized by the types of allowable transformations and therefore the number of degrees of freedom (DOF) for the spatial transform: rigid-body, affine (or linear), and non-linear. A rigid body moving in 3D has three possible translations (along the x, y and z dimension) as well as three possible rotations about the set of orthogonal (x, y, z) axes, giving a total of 6 DOF for

rigid body registration in 3D. Affine registration allows up to 6 more transformations, giving a maximum of 12 DOF: the aforementioned 3 translations and 3 rotations, plus both a scaling and a shear (or skew) in each of the three orthogonal dimensions. Finally, non-linear registration introduces up to three fields of displacement (or warp fields), one for each dimension (x, y, z).

With more DOF comes increased flexibilty and potentially greater registration accuracy, but this is at the expense of both computational complexity and time. It is therefore always important to select the most approriate registration protocol for the task at hand, which is largely dependent on the types of images that we wish to align. For example, if we consider two T1-weighted MR images of the brain: a rigid-body 6 DOF alignment will only correct for differences in spatial position and orientation, and not differences in brain anatomy, so this type of registration should only be used for intra-subject registration (i.e., for two images of the same subject, either from the same or differing image modality); an affine registration protocol will also correct for global brain differences between individuals, and is therefore suitable for inter- (i.e., between) subject registration tasks; finally, non-linear registration accounts for finer-scale, local differences in brain anatomy (e.g., ventricular enlargement of dementia patients), and is therefore recommended when highly-accurate inter-subject alignment is required.

Often, the number of degrees of freedom is progressively increased to achieve correspondingly better alignment of images. For example, affine registration of two T1-weighted MR images from different subjects produces a transformation matrix (providing the instructions for 3D transformation of the moving image to the reference image space) that can be used to initialize a non-linear alignment of these images. This progressive alignment of images has two key attributes: firstly, it reduces the computational complexity and duration (e.g., by reducing the search range) of the later steps in this hierarchical process; and secondly, it increases the likelihood of finding the globally-optimal alignment of the two images.

Similarity metrics

Data correspondence is achieved by maximizing any similarities (or minimizing any dissimilarities) between the two images. To quantify this similarity, several similarity measures are commonly used, such as *cross-correlation*, *least squares*, *correlation ratio*, *normalized cross-correlation*, *mutual information*, and *normalized mutual information* [Rizzo et al., 1991; Arun et al., 1987; Barnea and Silverman, 1972; Collignon et al., 1995; Viola and Wells III, 1997; Studholme et al., 1998]. Some similarity measures are better than others for certain registration tasks. For example, cross-correlation, normalized cross-correlation, and least squares are typically used to register images from (within) the same modality (e.g., two T1-weighted MR images, two FDG-PET images), whereas correlation ratio, mutual information, and normalized mutual information are used for cross- (i.e., between) modality registrations.

Interpolation methods

In order to compute these similarity metrics, the information contained within the moving image must be interpolated (i.e., re-sampled) to match the size and resolution of the underlying reference image. Interpolation itself is the process of using known (neighbouring) data to estimate values at unknown data points. Typical interpolation algorithms in 3D include *nearest neighbour*, *trilinear*, and *spline*, each of which differ in complexity (i.e., the number and weighting of adjacent voxels considered in estimating an unknown data point) and the quality of the resultant interpolated image. Generally speaking, image quality can be improved at the expense of increasing computational complexity and time, but the choice of interpolation method largely depends on the specific registration task being performed, because each method has its own advantages and disadvantages. For example, nearest neighbour is the simplest interpolation method, but the interpolated image is shifted with respect to the original image. Linear interpolation is also relatively simple, and each segment (bounded by two points) can be interpolated independently, but the method results in discontinuities at each point, so the function is gener-

ally not differentiable. Spline interpolation, on the other hand, offers a smooth transition and true continuity between data points, and the interpolant is differentiable at all points, but the method is relatively complex because it requires the fitting of a polynomial of degree n within each segment. Additionally, spline interpolation requires knowledge of the two endpoints of the segment plus the two points on either side of them, which raises the issue of how to interpolate between the first and last segments.

2.4.3 FMRIB's software library (FSL)

The FMRIB Software Library (FSL, www.fmrib.ox.ac.uk/fsl) [Smith et al., 2004; Woolrich et al., 2009] is a comprehensive suite of image analysis tools for conventional MRI, fMRI, and DWI brain data, provided by Oxford's Centre for Functional MRI of the Brain (FMRIB). This library includes software for image pre-processing (cropping, bias-field correction, brain extraction, timing and distortion effects), registration (rigid-body, affine, non-linear) and segmentation (using manual definition or parameterized-modelling algorithms), as well as a whole host of tools for quantifying information in the processed images (i.e., label volumes, vertex analysis, voxel-based morphometry). Throughout this thesis, tools from the FSL will be utilized for analysis of all clinical data, unless otherwise stated. Most frequently, we use the brain extraction tool (BET), which strips the skull, neck, and other non-brain features from the T1-weighted MR images, FMRIB's linear image registration tool (FLIRT), it's non-linear counterpart (FNIRT), and the later-mentioned integrated registration and segmentation tool (termed FIRST).

2.4.4 PVElab software

The PVElab software [Quarantelli et al., 2004] is a comprehensive software package for correction of PET and single PET (SPET) data from the PVEs introduced by the PSF of the scanner. Through the integrated analysis of brain PET/SPET and MRI data, the PVElab software is able to correct for the contaminating contribution of surrounding tissue to each specific voxel. As

illustrated above, structures (in the object domain) that are smaller than two times the FWHM of the scanner's PSF are especially affected by PVEs (e.g., the hippocampus). The observed PET signal (also known as the *apparent tracer concentration*) at each voxel in the vicinity of the hippocampus (and indeed elsewhere in the brain) therefore reflects the average tracer concentration from all contributing tissues (GM, WM, CSF). Since WM and CSF typically have lower tracer uptake than GM, the tracer concentration can be underestimated in subcortical GM structures such as the hippocampus. To address this issue, the PVElab software essentially removes any contributions of the surrounding tissue to the apparent GM tracer concentration.

The PVElab software provides a choice of several different methods for PVE-correction (PVEC) of PET and SPET data. In this work, the modified Müller-Gärtner (mMG) method [Müller-Gärtner et al., 1992; Rousset et al., 1998] is implemented to correct FDG-PET data. This mMG PVEC method is based on the theory that the observed PET image (I_{obs}) is a linear sum of the actual (i.e., "true") radioactivity distribution of each tissue compartment (I_{grey}, I_{white}, I_{CSF}) convolved with the PSF of the imaging system. Assuming that the spatial distribution of each tissue compartment is known from MRI, and the actual radioactivity within WM and CSF is constant and known (from computation of the mean radioactivity within these tissue compartments), a series of modified MR and PET images are generated, leaving a simple ratio for estimating the actual GM concentration.

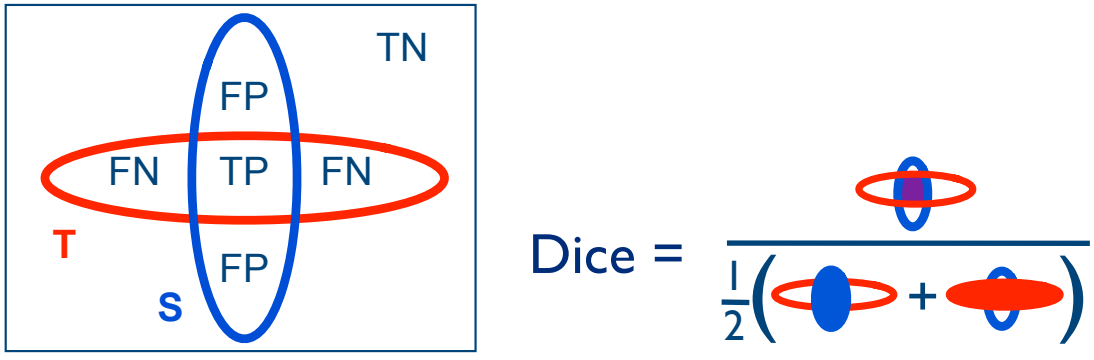
2.5 Segmentation evaluation: performance measures

Once delineation of a target ROI has been made, it is important to evaluate the accuracy of this segmentation to ensure that we have a realistic representation of the underlying imaged object. If this segmentation estimate is not sufficiently accurate, then subsequently derived properties of the ROI (e.g., length, volume, texture, etc.) may themselves be inaccurate and misleading, potentially impacting on the diagnosis and future treatment of the subject under investigation. Again, whilst an initial qualitative evaluation by visual assessment is always encouraged, thor-

ough quantitative evaluation of segmentation estimates is crucial for a comprehensive understanding of segmentation accuracy, and for comparison of differing segmentation methods. To this end, a set of easily-interpretable and informative performance measures are required, together with a set of expert manual labels/segmentations to be used as a gold standard (i.e., a reference against which all other segmentation estimates are compared).

There is a vast array of published performance measures for quantitative evaluation of segmentation methods. However, with such a variety of possible measures, and with no universal standard agreed upon, it is often difficult to formulate direct comparisons of segmentation methods. In this thesis, we report a combination of metrics used in a recent online resource for validation of brain segmentation methods [Shattuck et al., 2009] and the MICCAI 2008 competition workshop on MS lesion segmentation: Dice coefficient, false positive rate (FPR) and false negative rate (FNR). Using the hippocampus as our example ROI, if X is the set of all voxels in the image, we define the gold standard (i.e., truth set) $T \subset X$ as the set of voxels labelled as hippocampus by the expert manual rater. Similarly, we define the set $S \subset X$ as the set of voxels labelled as hippocampus by the segmentation algorithm or method being tested, as illustrated in Figure 2.12 (a) for a 2D section.

Figure 2.12 *Segmentation performance measures.* Illustrations of (a) the success and error measures for a segmentation estimate S compared to the gold standard truth set T: true positive (TP), true negative (TN), false postive (FP) and false negative (FN) count; and (b) the Dice coefficient similarity metric (the formula for which is given in Equation 2.3).



(a) (b)

2.5.1 Success and error rates

The true positive set is the set of voxels common to both T and S, defined as $TP = T \cap S$. The true negative set is the set of voxels that are labelled as non-hippocampus in both sets, defined as $TN = \overline{T} \cap \overline{S}$. Similarly, the false positive set is defined as $FP = \overline{T} \cap S$ and the false negative set is $FN = T \cap \overline{S}$, also illustrated in Figure 2.12 (a). From these four sets, we can compute various success and error rates for an image segmentation:

$$\text{FPR} = \frac{\mid FP \mid}{\mid FP \mid + \mid TN \mid} \tag{2.1}$$

$$\text{FNR} = \frac{\mid FN \mid}{\mid FN \mid + \mid TP \mid} \tag{2.2}$$

It is worth noting that the size of the image field of view (FOV) affects the FPR but not the FNR, such that the FPR are very low when considered over the whole image as opposed to a smaller ROI. To maintain a constant FOV in the presence of differing image dimensions, we always consider a ROI of volume 91125 mm^3 (equivalent to a 45x45x45 mm cuboid) in our calculations of the TN set.

2.5.2 Standard-space FP and FN maps

Both FPR and FNR are useful quantitative measures for the assessment of segmentation method performance, but they tell us nothing about the spatial distribution of the FP and FN set. Knowledge of these spatial distributions provides valuable information about the modes of failure of the segmentation method and therefore clues to how it might be improved — of utmost importance for image analysts striving to develop the most accurate and robust segmentation tool possible. To this means, we introduce the concept of standard-space FP and FN maps for additional asessment of segmentation method performance.

As illustrated in Figure 2.12 (a) for a 2D segmentation estimate, a spatial map of the FP or FN set can be found for a given segmentation method applied to a given image of an object (e.g., a subject's T1-weighted MR image of the brain). If this same segmentation method is applied to a large cohort of subjects, with the corresponding FP and FN maps computed for each, it is impossible to directly compare these subject-specific maps because they are all in a different image space, unique to each particular subject. In order to directly compare the FP and FN maps of all subjects, they first need to be registered to a common image space, such as the standard MNI space. In this standard reference space, it is then possible to compute an average FP and FN map for the given segmentation method — providing a visual summary of the relative spatial distribution and frequency of segmentation errors on the entire data cohort. In 3D, voxel values in these average FP and FN maps correspond to the fraction of subjects showing a FP or FN result, respectively, at that voxel. It is perhaps worth noting that, whilst small localization errors may be introduced by registration of each subject's FP and FN map to the standard MNI space, the false voxels themselves do not rely on the registration — they are computed before the registration and just transformed into the common reference space.

2.5.3 Similarity metrics: Dice coefficient

The Dice coefficient (illustrated in Figure 2.12: b) provides information about the degree of overlap between two segmentations, and is defined as the size of the intersection of two sets divided by their average size:

$$D(T,S) \;=\; \frac{2 \mid T \cap S \mid}{(\mid T \mid + \mid S \mid)} \tag{2.3}$$

It is also worth noting that whilst the aforementioned performance measures are used here to evaluate the accuracy of automated segmentation methods with respect to a set of gold standard expert manual labels, the Dice coefficient is also frequently used to assess the agreement between two sets of manual labels, provided either by the same rater at different time points

(i.e., an intra-rater reliability measure) or by different expert rates (i.e., an inter-rater reliability measure).

2.6 Clinical data

Four clinical datasets (detailed below) are utilized in this work. For our development and evaluation of hippocampal segmentation methods, we require structural (T1-weighted) MR data with corresponding expert manual labels for the hippocampus (i.e., the CMA, BPSA, and ADNI datasets), whilst multi-modal (MR and FDG-PET) data is necessary for investigating the combined structural and metabolic changes that occur in neurodegenerative dementia syndromes (i.e., the Leipzig data). Dataset selection is therefore restricted by current availability of both structural MR images and the required complementary data (i.e., expert manual labels or FDG-PET data), but where possible, the patients within these datasets have diseases/disorders associated with hippocampal atrophy.

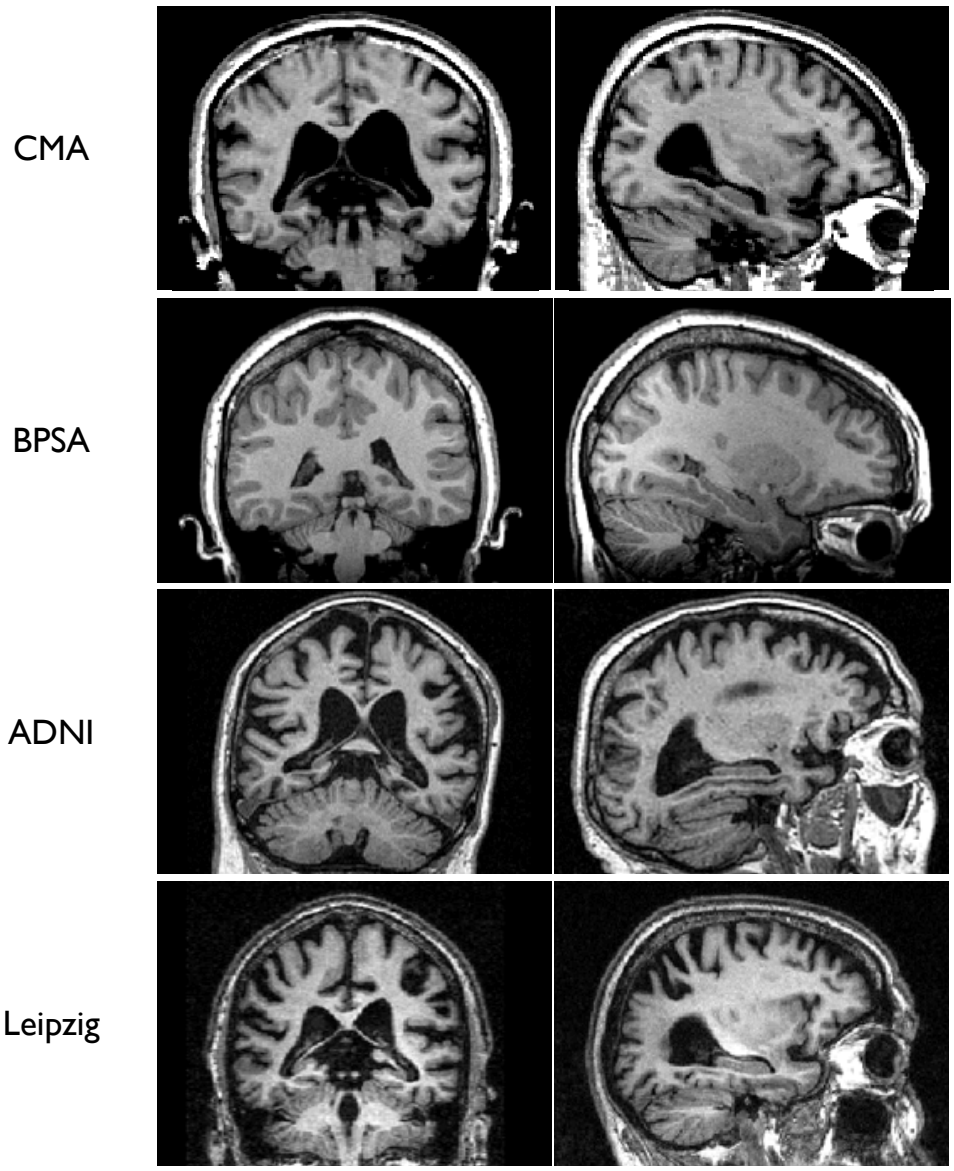
2.6.1 Data from the Center for Morphometric Analysis (CMA)

The first clinical dataset is provided by collaborators at the Center for Morphometric Analysis (CMA). It consists of T1-weighted MR images and corresponding expert manual labels for 9 normal control (NC) subjects and 8 atrophied AD patients (Figure 2.13: first row), the demographics of which are given in Table 2.1. The resolution of the T1-weighted MR images is 0.94 x 1.50 x 0.94 mm.

Table 2.1 *Demographics of the CMA data.* Group size (N), mean age (in years), and gender (Female (F) or Male (M)) for the normal controls (NC) and Alzheimer's disease (AD) patients in the CMA dataset.

	NC	**AD**
N	9	8
Age (years±SD)	73.7±5.0	72.3±3.9
Gender (F, M)	5, 4	2, 6

Figure 2.13 *Structural MR images.* Coronal- (left) and sagittal- (right) view images of an Alzheimer's Disease (AD) patient from the CMA (first row), ADNI (third row) and Leipzig dataset (fourth row), showing ventricular enlargement and hippocampal atrophy, as well as an exemplar bipolar (BP) patient from the BPSA dataset (second row).



2.6.2 Bipolar dataset from San Antonio (BPSA)

The second clinical dataset is a collection of T1-weighted MR images (Figure 2.13: second row) and corresponding expert manual labels for 16 NC and 16 bipolar (BP) patients from collaborators at the University of Texas Health Science Center in San Antonio (denoted BPSA). Image resolution is 0.8 x 0.8 x 0.8 mm, with further details provided in Table 2.2.

Table 2.2 *Demographics of the BPSA data.* Group size (N), mean age (in years), and gender (Female (F) or Male (M)) for the normal controls (NC) and bipolar patients (BP) in the BPSA dataset.

	NC	**BP**
N	16	16
Age (years±SD)	37.1±10.5	38.2±12.6
Gender (F, M)	7, 9	10, 6

Since hippocampal volume decreases in BP patients are fairly ambiguous, we appreciate that BP is not necessarily the best choice of psychiatric disorder for looking at hippocampal abnormalities. However, inclusion of this dataset in our evaluation of hippocampal segmentation methods will test their accuracy and robustness with respect to different disorders and imaging parameters.

The research sites that provided the CMA and BPSA data used their own expert raters and semi-automated contouring tools to perform a similar manual definition of the hippocampus in sequential coronal slices, based on intensity boundaries and well-established geometrical rules of neuroanatomy [Kennedy et al., 1989]. All experts underwent a period of training (of up to three months) until they had reached a defined reproducibility. As a measure of agreement between the manual labels of different raters (0 being completely no agreement, 1 being total agreement), the inter-rater reliability for the BPSA and CMA data is 0.90 and 0.80, respectively.

2.6.3 Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Similar to the aforementioned datasets, the third clinical dataset provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI, publicly available at http://adni.loni.ucla.edu) consists of T1-weighted MR images and corresponding expert manual labels. Image resolution varies, but is typically 1.20 x 0.94 x 0.94 mm, and the data demographics are given in Table 2.3. In addition to the information on age and gender, measures of dementia severity (Mini Mental State Examination (MMSE) scores [Folstein et al., 1975], and Clinical Dementia Rating (CDR) [Morris, 1993]) were available for all three clinical groups (NC, MCI, and AD). Figure 2.13 (third row) presents the T1-weighted MR image of an exemplar AD patient from this ADNI dataset.

Table 2.3 *Demographics of the ADNI data.* Group size (N), mean age (in years), gender (Female (F) or Male (M)), Mini Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) scores for the ADNI data used in this study, including normal controls (NC), mild cognitively impaired (MCI) and Alzheimer's Disease (AD) patients.

	NC	**MCI**	**AD**
N	20	15	20
Age (years±SD)	76.0±5.7	75.2±7.2	77.4±7.2
Gender (F, M)	7, 13	5, 10	6, 14
MMSE (score±SD)	29.4±0.5	22.5±0.5	17.7±2.6
CDR (score±SD)	0.23±0.26	0.80±0.25	0.85±0.24

2.6.4 Leipzig data

For our fourth and final dataset, collaborators at the Max-Planck-Institute in Leipzig provided both structural MR and (functional) FDG-PET data of 13 controls (CON), 21 AD, and 10 frontotemporal lobar degeneration (FTLD) patients. The T1-weighted MR image of one AD patient is shown in Figure 2.13 (fourth row). The FTLD group is further categorized into two FTLD subtypes, with 6 semantic dementia (SD) and 4 frontotemporal dementia (FTD) patients. The demographics of this data are provided in Table 2.4 below. Image resolution for the MR data is 1.5 x 1.0 x 1.0 mm, whilst each transaxial slice of the reconstructed FDG-PET data has a matrix

of 128 x 128 voxels with an edge length of 2.45 mm.

Table 2.4 *Demographics of the Leipzig data.* Group size (N), mean age (in years), gender (Female (F) or Male (M)), Mini Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) scores for the Leipzig data used in this study, including controls (CON), Alzheimer's Disease (AD) and Frontotemporal Lobar Degeneration (FTLD) patients. Details of the two FTLD subtypes are also given: Semantic Dementia (SD) and Frontotemporal Dementia (FTD). MMSE scores for controls were not available (n.a.) because other tests had already identified these subjects as being cognitively normal elderly.

	CON	**AD**	**FTLD**	**SD**	**FTD**
N	13	21	10	6	4
Age (years±SD)	53.9±6.0	61.1±6.7	61.8±5.9	64.2±3.4	58.3±7.6
Gender (F, M)	6, 7	12, 9	4, 6	3, 3	1, 3
MMSE (score±SD)	n.a.	23.2±3.9	24.5±4.9	24.5±4.7	24.5±5.8
CDR (score±SD)	0.23±0.26	0.71±0.25	0.85±0.47	0.83±0.61	0.88±0.25

3

Method Comparison and Evaluation

Here, several MR hippocampal segmentation methods are presented and evaluated: two versions of FMRIB's Integrated Registration and Segmentation Tool (FIRST and FIRSTv3), Freesurfer's Aseg (FS), and three versions of an atlas-based label-fusion approach termed Classifier Fusion (CFNS, CFRS and CFRS-NMI). Segmentation performance is assessed on two separate clinical datasets, utilizing expert manual labels as gold standard to quantify Dice coefficients, false positive rates (FPR) and false negative rates (FNR). The first clinical dataset, from the Center for Morphometric Analysis (CMA), contains 9 normal controls (NC) and 8 highly-atrophied AD patients, whilst the second is a collection of 16 NC and 16 bipolar (BP) patients from San Antonio (denoted BPSA). Results show that CFNS, CFRS and CFRS-NMI outperform all other methods on the BPSA data, whilst FIRST and FIRSTv3 perform best on the CMA data, with average bilateral Dice coefficients of 0.80±0.05, 0.82±0.05, 0.82±0.05, 0.85±0.02 and 0.85±0.02, respectively. This work brings to light several strengths and weaknesses of the evaluated hippocampal segmentation methods, of utmost importance for robust and accurate segmentation in the presence of specific and substantial pathology.

3.1 Introduction

Conventionally used manual segmentation of neuroanatomical structures requires expert knowledge and is extremely time consuming, thus impractical for clinical routine or large-scale clinical studies that typically require several hundred images. Nevertheless, it is still the accepted gold standard for hippocampus segmentation, with the performance of emerging semi-automated and automated segmentation methods assessed by comparison to this standard. This section provides a more technical review of the methodology and performance of published hippocampal segmentation methods, broadly classified into three groups: region-growing, atlas-based, and parameterized-modelling approaches. Whilst some methods may still be identified as belonging to one particular classification [Taylor and Barrett, 1994; Carmichael et al., 2005; Heckemann et al., 2006; Aljabar et al., 2007; Patenaude, 2007], there is a progressive trend towards 'multi-approach' segmentation methods that combine aspects of two or more groups to achieve increased robustness and segmentation accuracy [Fischl et al., 2002; Chupin et al., 2007; Colliot et al., 2008; Morra et al., 2008; van der Lijn et al., 2008; Chupin et al., 2009b; Lotjonen et al., 2010; Wolz et al., 2010].

As the name suggests, region-growing methods typically grow a region-of-interest (ROI) from an initial start/seed point manually defined on the target image. The ROI grows and deforms according to various constraints (e.g., anatomical landmarks, surface curvature, energy constraints) imposed by the respective algorithms. In 2007 Chupin et al. [2007] presented a dual region-growing approach for segmentation of both the hippocampus and the amygdala, based on competitive growth of homotopically deforming regions. The deformation constraint was based on prior knowledge of anatomical features in the vicinity of 11 sets of automatically-detected landmarks from the MRI data, combining both aspects of region-growing and model-based methods. The relative intensities and positions of structures in the vicinity of the hippocampus and amygdala were modelled using a series of 2D 3x3-voxels pattern masks (voxel size: 0.94-0.98mm), with these intensity and positional rules subsequently used to detect the anatomical

landmarks during region growing. The study consisted of 16 young normal controls (NC) and 8 AD patients, finding mean relative volume error, overlap and maximum boundary distance for the hippocampus of 7%, 84% and 4.5 mm for NC and 9%, 84% and 6.5 mm for AD patients, respectively. Likewise, in a clinical application of this technique [Colliot et al., 2008], significant hippocampal volume reductions were detected for all patient groups (mild cognitively impaired (MCI) and AD), with the normalized hippocampal volume giving correct group classification rates of 73% for MCI and 84% for AD with respect to NC, as well as 69% for AD with respect to MCI. Whilst these region-growing methods can provide relatively efficient segmentation accuracy and group classification compared to human raters, they often require time-consuming and knowledge-based manual intervention (for definition of a specific bounding box and seed points), with the resultant segmentation dependent on good initialization. This issue will be discussed further in the following chapter.

Atlas-based methods traditionally align a single atlas, consisting of an anatomical image plus corresponding manual labels, to a target query image (or vice versa) to provide a segmentation estimate for this query image [Barnes et al., 2008; Carmichael et al., 2005; Haller et al., 1997; Qiu et al., 2009; Xu et al., 2008]. Registration aligns the atlas anatomies to a target image, with the atlas labels propagated using the warp field derived from this registration. When the atlas anatomies are aligned with a target image, the term "classifier" may be used to describe the atlas labels. In 2005, Carmichael et al. [2005] compared the relative performance of the standard atlas-based approach (using Harvard and MNI atlases) with that of randomly selected, manually labelled subject images ("cohort atlases"), as well as the influence of different registration methods (AIR, SPM, FLIRT, Chen [Woods et al., 1998; Friston et al., 1995; Jenkinson et al., 2002; Chen, 1999]). They concluded that cohort atlases and fully-deformable registration provide best results, with an overlap ratio (OR) of 0.6 (equivalent to a Dice coefficient of 0.75). Heckemann et al. [2006] found that segmentation performance could be greatly improved by fusing multiple, propagated atlas labels in a multi-classifier decision fusion approach, with

a reported hippocampus OR of 0.70±0.05 (Dice coefficient of 0.82±0.03). Understandably, Heckemann's label-fusion approach suffers from problems of scale when the number of classifiers is large, fuelling various extensions to the method [Aljabar et al., 2007 2009; van der Lijn et al., 2008; Lotjonen et al., 2010].

Aljabar et al. [2007] introduced the concept of classifier selection (using a subset of best-matched atlases, ranked according to segmentation similarity, image similarity and demographics) and found an asymptotic level of segmentation accuracy with increasing numbers of atlases. Mapping all repository images to a single reference space, top-ranked classifiers were selected and fused using a majority vote rule. Overlap values were highest using normalized mutual information (NMI), closely followed by cross-correlation (CC), for image similarity selection and fusion of 10-20 classifiers (ordered according to segmentation similarity). Mean (SD) overlap values after fusion of 1000 sets of 15 random classifiers for all 20 query subjects were 0.79 (0.03) for the hippocampus, 0.90 (0.02) for the lateral ventricle and 0.88 (0.02) for the thalamus, with NMI image similarity selection resulting in 14 of the 20 subjects having overlap values in the upper quartile of this distribution for the hippocampus. A later study [Aljabar et al., 2009] evaluated image similarity in MNI reference space, selected 20 top-ranked atlases and registered these directly to the query image to obtain a segmentation estimate in native space, giving a higher average, combined left-right Dice accuracy for the hippocampus of approximately 0.83. One possible downside of this approach is that the set of top-ranked atlases in standard space may not necessarily be the same as the set of top-ranked atlases in native space. Indeed, during development of a combined multi-atlas and intensity-modelling segmentation method, Lotjonen et al. [2010] found that label-fusion generally performs best in the native space, with intensity-modelling (using either graph-cuts or an expectation maximization (EM) algorithm) as a post-processing step. Interestingly, this study reported clear differences in Dice coefficient between datasets (0.88 for the Alzheimer's Disease Neuroimaging Initiative (ADNI), 0.82 for the Internet Brain Segmentation Repository (IBSR) data), suggesting differences in image

quality, manual segmentation protocol, clinical status and demographics as possible reasons for this.

Parameterized-modelling approaches generally learn characteristics of a given structure from a training dataset, generate an appropriate model, and fit this to a new query image to provide a segmentation estimate [Fischl et al., 2002; Kelemen et al., 1999; Morey et al., 2009; Patenaude, 2007; Patenaude et al., 2011; Pitiot et al., 2002]. These characteristics may include intensity profiles, shape, or spatial distributions, represented by a probability map, a neighbourhood function or a statistical model of shape and appearance. Both the later-mentioned FIRST and Freesurfer (FS) tools use parameterized-modelling to segment the hippocampus. During development and validation of FIRST, Patenaude [2007] reported an average hippocampal overlap of approximately 78% on test data with highly-varied subject pathology and demographics. In a comparative study, Morey et al. [2009] evaluated the performance of FIRST and FS on healthy subject data, using percentage volume overlap, percentage volume difference, across-sample correlation and 3D group-level shape analysis. The reported average left and right overlaps for the hippocampus were 79%$\pm$3.6 and 80%$\pm$2.9 with FIRST, 82%$\pm$1.5 and 82%$\pm$2.8 with FS. In other work, Pitiot et al. [2002] segmented the hippocampus using deformable templates constrained by statistical shape models, calculating a root mean square error (RMSE) of 1.3$\pm$0.6 mm between the computed boundary and that of the manually-defined hippocampus.

As is evident from the literature, most studies focus on the development and evaluation of an individual segmentation method on a single dataset, reporting any number of performance measures, such that direct comparison of hippocampal segmentation methods is compromised. Consequently, it is often difficult to decide upon an appropriate segmentation method for any given study. Comparative evaluation of several methods, on multiple clinical datasets, may highlight the relative strengths and weaknesses of these different approaches, facilitating informed selection of appropriate segmentation methods and guiding their future development. To this end, we evaluate here the aforementioned model-based methods, FIRST and FS, and three adaptations

of the atlas-based label-fusion approach of Aljabar et al. [2007]. In the subsequent chapter, we combine aspects of region-growing and atlas-based methods to develop a novel hippocampal segmentation method, termed FMASH (Fast Marching for Automated Segmentation of the Hippocampus).

3.2 Materials and methods

3.2.1 Clinical MR data

The CMA and BPSA datasets (detailed in Section 2.6) were used in this study: the first consisting of 9 normal controls (NC) and 8 highly-atrophied AD patients, and the second a collection of 16 NC and 16 bipolar (BP) patients.

3.2.2 Image processing and segmentation methods

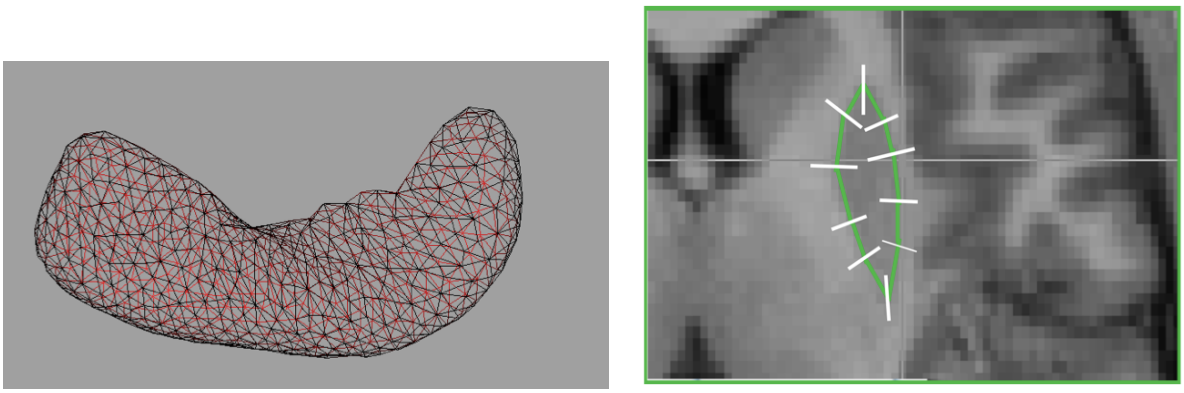
Unless otherwise stated, all MR data was processed using tools from the FMRIB Software Library (FSL [Smith et al., 2004; Woolrich et al., 2009], www.fmrib.ox.ac.uk/fsl). The MR images were initially cropped to remove excess neck, and BET [Smith, 2002] was used to remove most of the non-brain structure. The brain-extracted data was used for affine registration tasks, whilst the whole-brain data was used for non-linear transformations, as described below. The following segmentation methods were used to delineate the hippocampi of every subject, with the accuracy of the resultant segmentation estimates assessed by comparison to expert manual labels, using the performance measures described in Section 3.2.3.

FMRIB's Integrated Registration and Segmentation Tool (FIRST)

FIRST is a Bayesian statistical shape and appearance (intensity) model [Patenaude, 2007; Patenaude et al., 2011]. From a training dataset, shape is modelled as a multivariate Gaussian point distribution model (PDM), parameterized by a linear combination of the mean (Figure 3.1a)

and eigenvectors. Intensity is also modelled by its mean and eigenvectors, with intensity profiles sampled along the surface normal at each vertex location (Figure 3.1b). Following intensity normalization, each new image is spatially normalized to the MNI152 reference space using a two-stage linear registration process; whole-brain followed by a subcortical-weighted affine transformation, maximizing the correlation ratio similarity function. Specifically, the image similarity function is initially computed over the whole brain to estimate the global alignment of each new image with the MNI152 reference, before a binary subcortical mask restricts computation of the similarity function to the subcortical regions, refining the alignment of these brain structures. In the reference space, the average mesh is initialized and iteratively deformed (for a fixed number of iterations) to give the resultant structure mesh. This structure mesh is itself a parameterized model of the hippocampal surface, with numbered vertices and information about the spatial location of these vertices, which can be used for future vertex analysis — more on this in Chapter 6 (Clinical Application). Finally, 3D surface rendering and filling produces a volumetric FIRST output.

Figure 3.1 *FMRIB's Integrated Registration and Segmentation Tool (FIRST) modelling.* (a) The mean surface mesh for the point distribution model (PDM) of the left hippocampus and (b) a 2D illustration of the intensity profiles for the left putamen. (Images courtesy of users.fmrib.ox.ac.uk/∼brian/brianp_thesis.pdf)



(a) (b)

Here, FIRST was run simultaneously for both the hippocampus and the amygdala, giving a resultant segmentation with no voxel overlap at the hippocampus-amygdala border. Individual

hippocampus (and amygdala) segmentations for the left and right hemispheres were extracted from the FIRST output volume.

FMRIB's Integrated Registration and Segmentation Tool v3 (FIRSTv3)

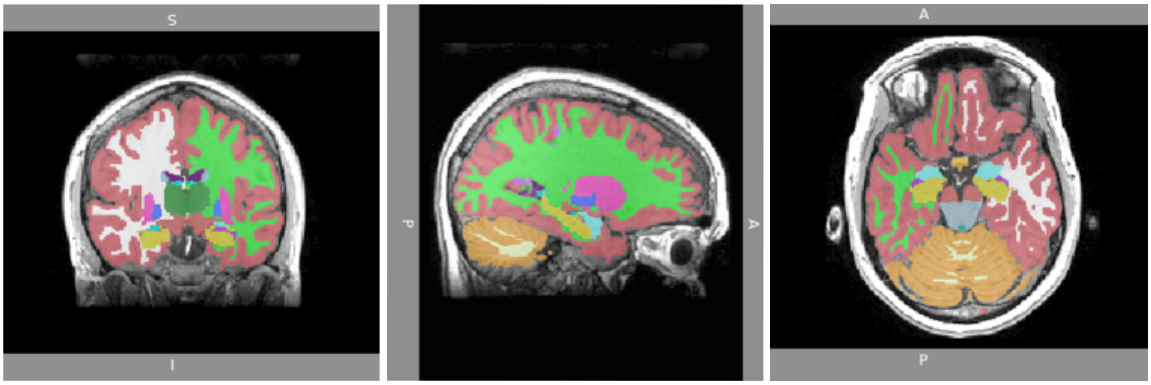
FIRSTv3 is an adaptation of the aforementioned FIRST tool, using a novel structure-specific spatial normalization before model fitting. Following intensity normalization, each new image is spatially normalized to the MNI152 reference space using a three-stage affine registration; whole-brain followed by subcortical- then hippocampus-weighted alignment, using three-times and eight-times dilated masks, respectively. As decribed above for FIRST, initial computation of the similarity function over the entire brain region finds an estimate of the gobal alignment of each new image with the MNI152 reference, before the binary subcortical mask restricts and refines this alignment to the subcortical regions. The difference here is an additional, third registration step, using a binary hippocampal mask to further refine the alignment of the hippocampal regions. The original (non-dilated) subcortical mask is provided with the standard MNI templates in FSL, and the Harvard-Oxford atlas (part of FSL) provides the original hippocampal mask, before standard morphological (binary) dilation to generate the final weighting masks. The computed three-step affine transform is then used as input for FIRST model fitting, segmenting each subject's hippocampi.

Here, the three-stage affine registration was computed with FLIRT [Jenkinson and Smith, 2001], using 12 degrees of freedom (DOF) and maximization of normalized cross-correlation. Like FIRST, FIRSTv3 was run simultaneously for both the hippocampus and the amygdala, giving a resultant segmentation with no voxel overlap at the hippocampus-amygdala border. Individual hippocampus (and amygdala) segmentations for the left and right hemispheres were extracted from the FIRSTv3 output volume.

Freesurfer's Aseg (FS)

FS uses Bayesian parameter estimation theory [Fischl et al., 2002], with voxel classification determined by the segmentation that maximizes the (posterior) probability of input data given prior probabilities from the training set. Each new image is initially segmented, assigning voxels to the class of maximum probability determined by an initial probability map. Class probabilities are re-computed using a local neighbourhood function and the image is then re-segmented, repeating these two steps until the segmentation does not change (typically 5-6 iterations). An example of the final FS segmentation estimate is presented in Figure 3.2; providing a volumetric output for each segmented subcortical structure.

Figure 3.2 *Freesurfer's Aseg (FS) segmentation estimate.* Coronal- (left), sagittal- (middle) and axial- (right) view images of the FS volume-based (subcortical) labelling, with the hippocampus shown in yellow.

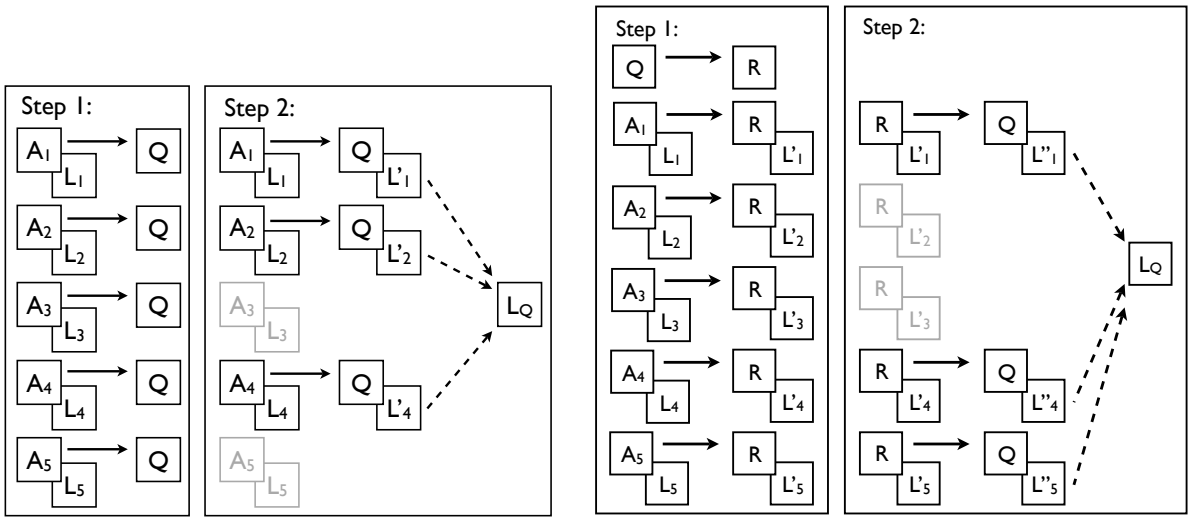


In this work, the first step of the program 'recon-all' was run for all subjects to transform images from their native space to FS Talairach space, with the 'subcortseg' option used to segment all subcortical structures. The affine transformation between Talairach space and each image's native space was computed using the 'tkregister2' function. In FS Talairach space, hippocampus labels were extracted from the FS output volume, binarized and transformed back into their native space (using the aforementioned affine transform). Here, the final FS hippocampal labels were obtained by threshold (probability = 0.5).

Classifier Fusion (CF)

The CF methods presented here (Figure 3.3) are extensions of the label-fusion atlas-based approach of Aljabar et al. [2007], with classifier selection either performed in the native space (NS) of the target query image, or a standard reference space (RS), giving rise to two versions of the CF method: CFNS and CFRS. In this work, the term 'atlas' refers to a paired T1-weighted anatomical image and its corresponding manual labels. When the atlas anatomies are aligned (i.e., registered) to a new query image, the atlas labels can be considered as a 'classifier', providing a segmentation estimate for this query image.

Figure 3.3 *Classifier Fusion (CF) methods.* A flow chart of the CF methods, with (a) classifier selection in the native subject space (CFNS), and (b) classifier selection in the standard reference space (CFRS). (a) — Step 1: affine register all A_i to Q. Step 2: non-linearly register selected A_i to Q, transform associated labels L_i, and fuse. (b) — Step 1: non-linearly register all A_i and Q to R, and transform associated labels L_i. Step 2: non-linearly transform selected labels to Q, and fuse. Please see main text for more details.



For the CFNS method, all atlas anatomies A_i are registered to the query image Q using a three-step affine transformation process; whole-head followed by subcortical- and hippocampus-weighted transformations (using three-times and eight-times dilated masks, respectively). Top-ranked atlas anatomies are selected and non-linearly registered to the query image, using the previously-computed three-step affine transformation for initialization. The resultant non-linear

transforms propagate the corresponding atlas labels L_i to the query image, providing a series of segmentation estimates L_i' for this image. Finally, in the query image space, the labels L_i' are fused (using majority voting) to give a single segmentation estimate for the query image. This output provides volumetric information (similar to FS), but simple fusion of labels (without application of a voxel-occupancy threshold) can generate a structure probability map, where voxel values correspond to the fraction of subjects classifying that voxel as the structure of interest — in this case, the hippocampus.

For the CFRS method, all atlas anatomies A_i and the query image Q are registered to the standard reference image R using an initial three-step affine transformation (whole-head, subcortical- and hippocampus-weighted transformations) followed by a non-linear warp. The corresponding atlas labels L_i are transformed to the reference space using these computed warps. The inverted (reference-to-query) non-linear transform is also computed. Top-ranked atlas anatomies are selected in the standard reference space, then the corresponding atlas labels L_i' are propagated to the native query image using the pre-computed inverted warp. This provides a series of segmentation estimates L_i'' for the query image, which are fused (using majority voting) to give a single segmentation estimate.

The CFRS method is similar to the approach of Aljabar et al. [2007] which employs a reference space to avoid the computational burden of non-linearly registering all atlas anatomies to the query image space. The CFNS method, on the other hand, performs CF in the native subject space to avoid degradation of the query image. This is in line with a recent study by Lotjonen et al. [2010] suggesting that CF performs best in the query image space. To avoid the computational burden associated with non-linear registration of every atlas to each query image, and to maximize classifier separation, the CFNS approach selects top-ranked classifiers after a three-step affine transformation.

Here, all affine transformations were computed with FLIRT [Jenkinson and Smith, 2001], using 12 DOF and maximization of the normalized cross-correlation similarity function, with the

same function evaluated over a restricted hippocampal region (i.e., a four-times dilated hippocampal mask) for selection of top-ranked atlas anatomies. Specifically, higher/better ranked atlas anatomies were those with higher/larger values of the computed image similarity measure. FNIRT was used for all non-linear registrations [Andersson et al., 2008]; specifying a control point spacing of 10mm, a sub-sampling scheme of 4,4,2,2,1,1, a bending energy model for regularization of the warp-field, and weightings for the membrane energy regularization of 300,150,100,50,40,30. Additionally, the CFRS method was implemented using the normalized mutual information (NMI) similarity function for both the affine registrations and selection of top-ranked atlas anatomies, termed CFRS-NMI. The purpose of this was to investigate whether the image similarity function significantly affected the resultant CF segmentation estimates. For the CMA and BPSA data, a leave-one-out approach was implemented to give $n - 1$ out of n atlases as potential classifiers within each dataset. Owing to the relative size of the two clinical datasets, 8 and 15 top-ranked classifiers were selected for the CMA and BPSA query images, respectively — equating to $\sim$50% of the total number of classifiers.

3.2.3 Performance measures

The metrics described in Chapter 2 (Section 2.5) were used to directly compare the segmentation performance of the methods herein. Dice coefficient, false positive rate (FPR) and false negative rate (FNR) were computed for all segmentation estimates and compared using the statistical analysis techniques described below.

Standard-space FP and FN maps

Additionally, standard-space maps were generated to assess the spatial distribution of FP and FN voxels for each segmentation method. We previously described the implementation of a three-step affine transform mapping each image to the standard MNI152 space, using 12 DOF and maximization of normalized cross-correlation. For each segmentation method, these affine

transforms were used to map all subjects' FP and FN voxels from their native space to MNI152 space. Here, the average FP and FN maps were computed (summing trilinearly interpolated values), with voxel values corresponding to the fraction of subjects showing a FP or FN result at that voxel.

3.2.4 Statistical analysis

SPSS statistics software (Version 19.0, www.spss.com/statistics/) was used for all statistical analysis. Initially, a chi-square test for independent samples was performed to evaluate group-wise gender differences for each dataset, and group comparisons for age were tested with an independent samples t-test (confidence interval = 0.95). For each segmentation method, a repeated measures analysis of variance (ANOVA) was initially employed to test for main effects of side and group for each performance measure (Dice, FPR and FNR), with side (left, right) as the within-subjects repeated measure and clinical group as the between-subjects factor (i.e., a 2*2 (side*group) ANOVA design). Paired-sample t-tests with Bonferroni correction for multiple comparisons were used to test for differences in side and group, with F_d, F_{fp}, F_{fn} and p_d, p_{fp}, p_{fn} denoting the F-statistic and corrected p-values for Dice, FPR and FNR, respectively. Following this, the left- and right-side performance measures were combined to give a bilateral average, which was used for all subsequent statistical analysis.

To test for overall difference across the segmentation methods and clinical group-by-method interactions, a 6*2 (method*group) mixed ANOVA design was subsequently employed for each performance measure, with method as the within-subjects factor and clinical group as the between-subjects factor. Where Mauchly's test indicated that the assumption of sphericity for method had been violated ($p<0.05$), degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. This time, paired-sample t-tests with Bonferroni correction for multiple comparisons were used to identify between-method differences, with F_d, F_{fp}, F_{fn} and p_d, p_{fp}, p_{fn} similarly denoting the F-statistic and corrected p-values for Dice, FPR and FNR,

respectively.

3.3 Results

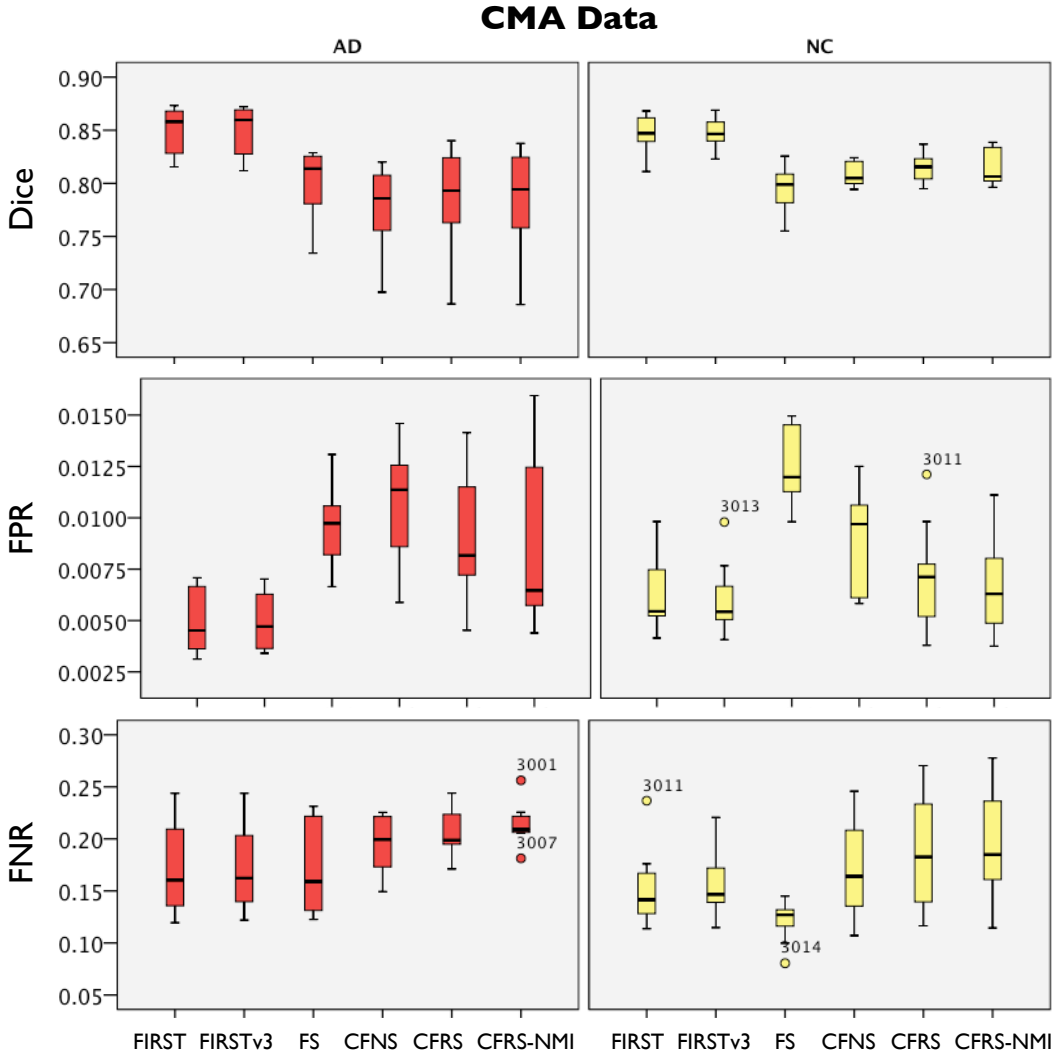
For both the CMA and BPSA data, Pearson's chi-square test for independent samples did not reveal any statistical difference in gender between the groups (CMA: NC=AD, χ^2(1, N=17) =1.633, p=0.201; BPSA: NC=BP, χ^2(1, N=32) =1.129, p=0.288), and the independent samples t-test identified no groupwise difference in age (CMA: p=0.531; BPSA: p=0.786). With regards to the hippocampal segmentation estimates, we generally did not observe any left-right asymmetries in method performance. Consequently, results presented here are the bilateral average of the performance measures for each subject's hippocampi. Box plots showing method performance on the CMA and BPSA data are given in Figures 3.4 and 3.8, respectively, together with the results of the pairwise comparisons (Figures 3.5 and 3.9) and some sample segmentation estimates in Figures 3.6 and 3.10. Additionally, Figures 3.7 and 3.11 present the standard-space FP and FN maps for each method's left hippocampal estimates on the CMA and BPSA data, respectively. For each ANOVA, Mauchly's test indicated that the assumption of sphericity for method had been violated ($p<0.0005$) so degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity.

3.3.1 CMA data

Testing for left-right asymmetries in performance for each segmentation method, the initial repeated measure ANOVA designs generally revealed no significant effect of side, with the exception of the FNR for FS (where left<right, F_{fn}(1,15)= 6.33, p_{fn}= 0.012), and the FNR for CFRS-NMI (where left>right, F_{fn}(1,15)= 7.90, p_{fn}= 0.007). The subsequent 6*2 (method*group) mixed ANOVA designs, using the bilateral average performance measures, revealed a significant effect of method for each performance measure (F_d(1.63,24.41)= 50.75, p_d <0.0005;

F_{fp}(1.34,20.14)= 13.76, p_{fp} <0.0005; F_{fn}(1.29,19.30)= 6.62, p_{fn}= 0.013), a significant method*group interaction for Dice (F_d(1.63,24.41)= 6.53, p_d= 0.008) and a trend towards a method*group interaction for FPR (F_{fp}(1.34,20.14)= 3.31, p_{fp}= 0.073), together with a strong trend towards a main effect of group for FNR (F_{fn}(1,15)= 4.39, p_{fn}= 0.053).

Figure 3.4 *Method comparison on the CMA data.* Box plots showing Dice coefficients (top), false positive rate (FPR: middle) and false negative rate (FNR: bottom) for each segmentation method on the CMA data, with separate plots for AD patients (left) and NC (right). Boxes have lines at the lower quartile, median, and upper quartile values, with whiskers extending to 1.5 times the inter-quartile range. Outliers are indicated by a circle and their corresponding subject identification data (ID).



The significant method*group interaction for Dice was a result of the CF methods having an opposite trend in groupwise Dice coefficients compared to the parameterized models: all of

the CF methods (CFNS, CFRS, and CFRS-NMI) had higher Dice for NC compared to AD patients, with this difference reaching significance for CFNS (F_d(1,15)= 5.23, p_d= 0.037 two-tailed); on the other hand, FIRST, FIRSTv3 and FS showed slightly increased Dice for AD patients compared to NC, although these groupwise differences were not significant. Similarly, the trend towards a method*group interaction for FPR was driven by an opposite trend in the groupwise FPR for the CF methods compared to the parameterized models: whilst FIRST, FIRSTv3 and FS displayed reduced FPR for AD patients compared to NC, all of the CF methods had higher FPR for the AD patients. With the exception of FS (F_{fp}(1,15)= 10.29, p_{fp}= 0.006 two-tailed), however, none of these groupwise differences in FPR were significant. Finally, the trend towards a main effect of group for FNR reflects the fact that all methods had higher FNR for AD patients compared to NC, but this groupwise difference was only significant for the FS method (F_{fn}(1,15)= 6.04, p_{fn}= 0.027 two-tailed).

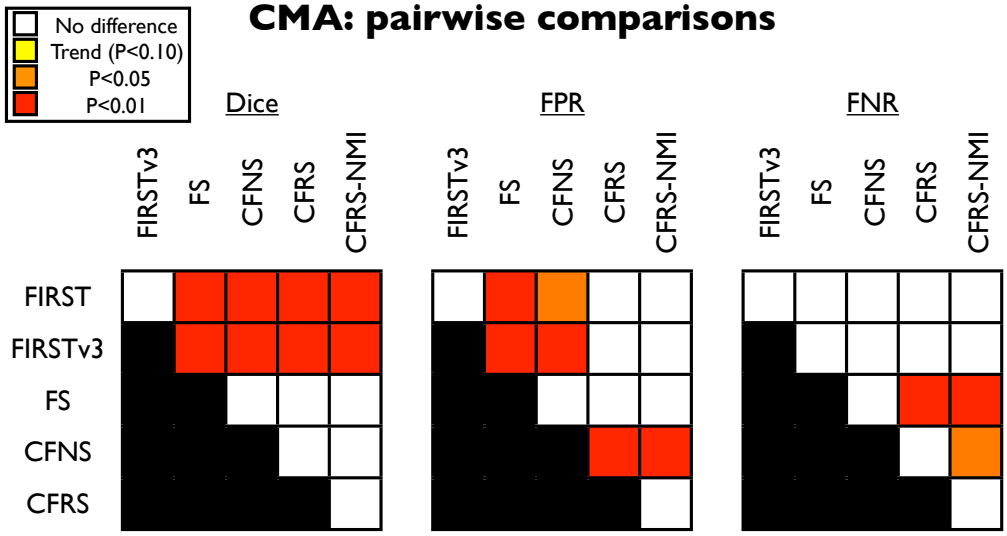
Thus, with the exception of the Dice coefficients for the CFNS method, the FPR and FNR for FS, there was no effect of group on segmentation performance for each method. The segmentation methods were therefore compared and ranked in terms of their performance on the entire CMA data (i.e., AD patients and NC combined). For each performance measure, the pairwise comparisons revealed a strict relative ordering (1-6) of the segmentation methods (with a value of 1 assigned to the best performing / first ranked method, a value of 2 to the second ranked method, and so on). Based on a simple sum of these rankings, the overall ordering of the segmentation methods was as follows

(i) FIRSTv3; (ii) FIRST; (iii) FS; (iv) CFRS; (v) CFRS-NMI; (vi) CFNS.

The differences in performance between FIRSTv3 and FIRST, as well as between FS and the CF methods, were minimal, with both FIRST and FIRSTv3 clearly outperforming all other segmentation methods on the CMA data. The results of the pairwise comparisons are presented in Figure 3.5. It is worth noting that for every performance measure, the pairwise comparisons revealed no significant difference between FIRST and FIRSTv3, and also no significant differ-

ence between CFRS and CFRS-NMI, so the relative ordering of these methods is based on their relative performance compared to the others.

Figure 3.5 *Pairwise comparison of methods on the CMA data.* Schematic representation of the two-tailed, paired-sample t-test results for the CMA data, comparing Dice coefficients (left), FPR (middle) and FNR (right) across the hippocampal segmentation methods.



The dominance of FIRST and FIRSTv3 on the CMA data was cemented by significantly higher overall Dice coefficients (0.848±0.020 and 0.848±0.019, respectively) compared to all other methods (p_d <0.0005 for all), and significantly lower FPR (both 0.006±0.002) than FS and CFNS (FS: p_{fp} <0.0005 for both FIRST and FIRSTv3; CFNS: p_{fp}= 0.007 and 0.005 for FIRST and FIRSTv3, respectively), combined with relatively low FNR (0.162±0.042 and 0.163±0.039, respectively). Typical segmentation estimates are given in Figure 3.6 (first and second row, respectively). For both FIRST and FIRSTv3, FP findings occured mainly in the ventromedial region of the hippocampus head (Figure 3.7 (left panel): first and second row, respectively), whilst medial boundaries and posterior regions of the hippocampus head tended to be underestimated (Figure 3.7 (right panel): first and second row, respectively).

FS had a respectable average Dice coefficient of 0.797±0.028 and significantly reduced FNR (0.145±0.043) compared to CFRS and CFRS-NMI (p_{fn}= 0.005 and 0.003, respectively), although visual assessment of results (e.g., Figure 3.6: third row) indicated a 'greedy' labelling

Figure 3.6 *CMA segmentation estimates.* Coronal- (left), sagittal- (middle) and axial- (right) view images of the left hippocampal segmentation estimates for subject 1 of the CMA dataset, showing good performance for FIRST and FIRSTv3 (first and second row, respectively), a 'greedy' labelling tendency for FS at the dorsoanterior and lateral hippocampal boundaries (third row), and slight mis-alignment of the CF results (fourth, fifth and sixth row) compared to the manual labels (seventh row). For direct qualitative comparison of the segmentation estimates, the cursor position is the same for each image.

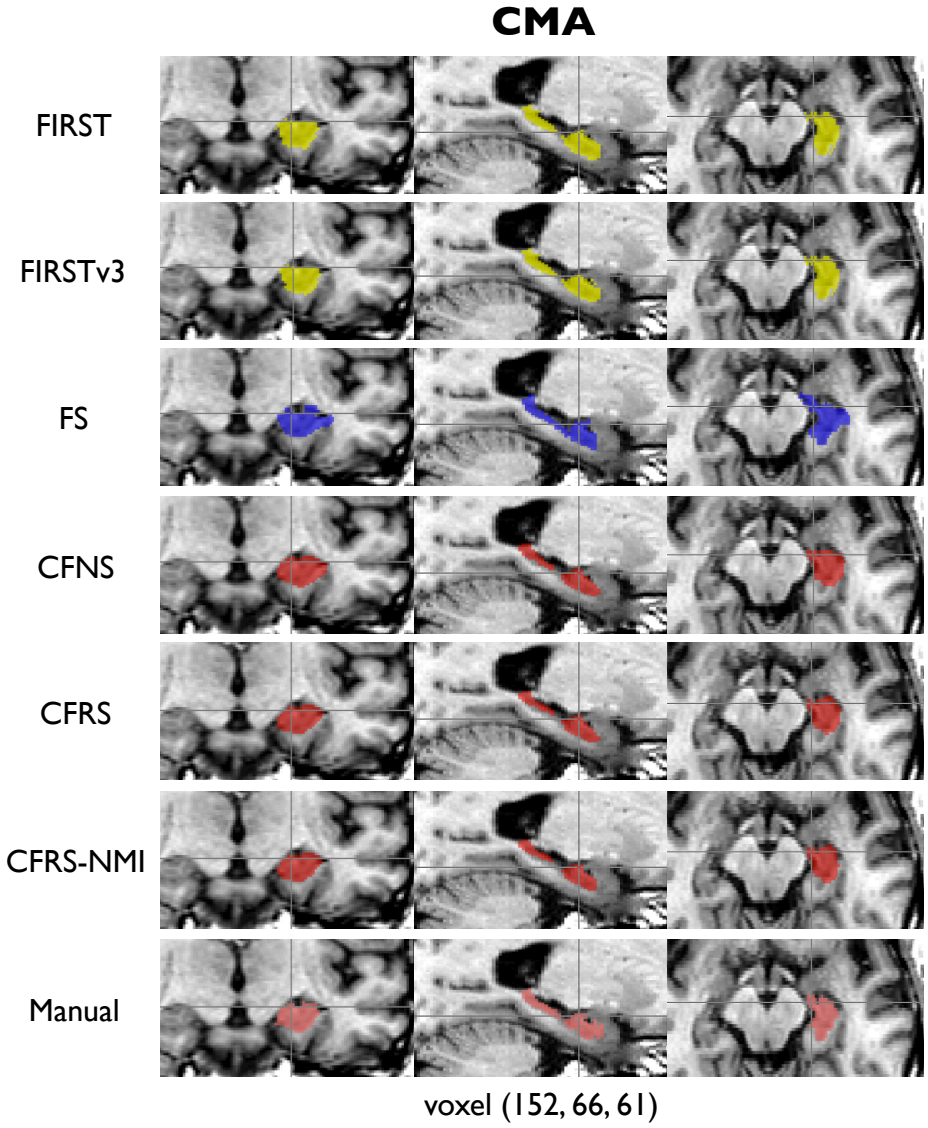
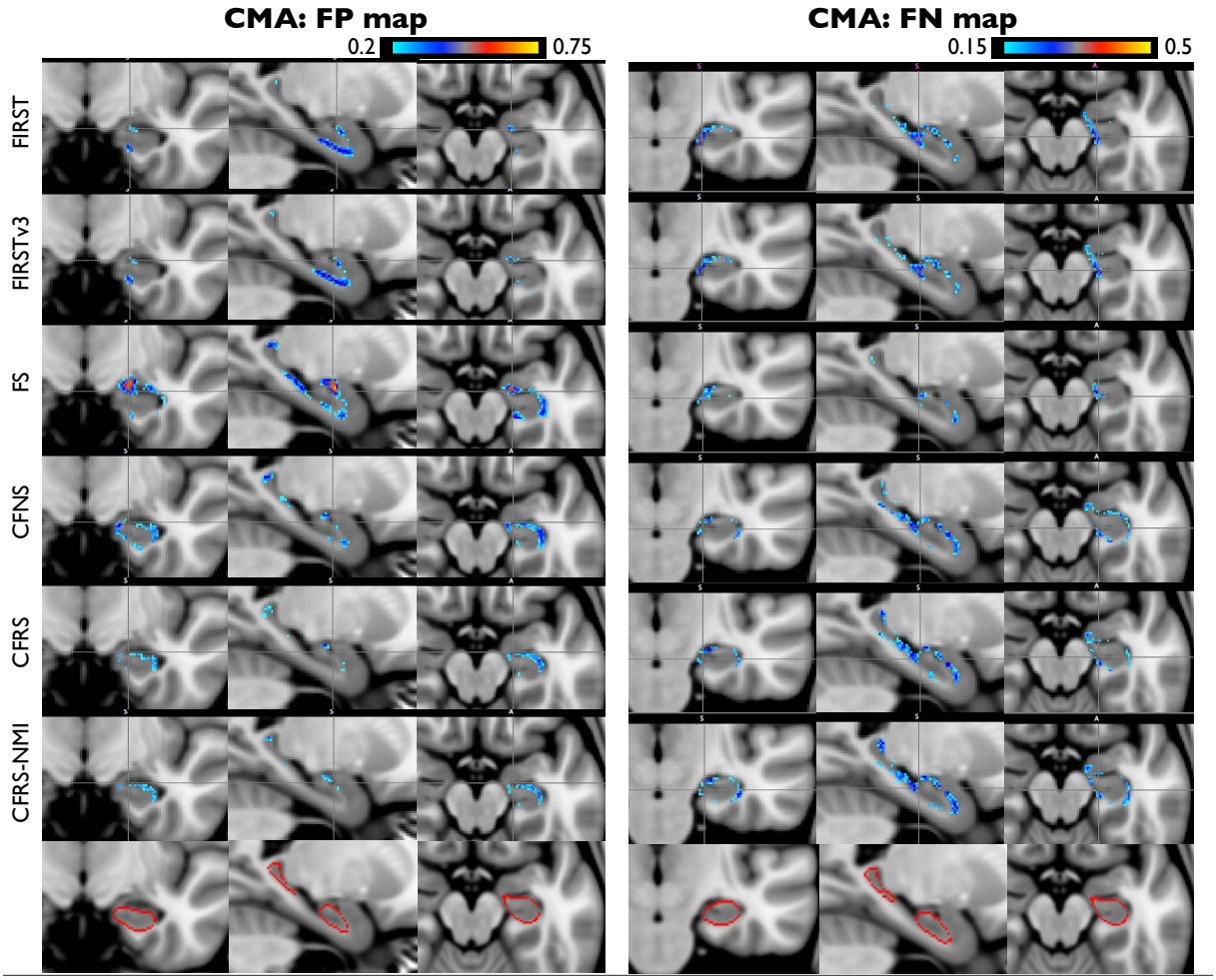


Figure 3.7 *False positive (FP) and false negative (FN) maps for the CMA data.* Left and right panels show coronal- (left), sagittal- (middle) and axial- (right) view images of the left hippocampus standard-space FP and FN maps, respectively, for each segmentation method on the CMA data. A voxel-wise threshold of 20% FP finding across the dataset was applied to all FP maps, whilst the FN maps were thresholded at 15%. To aid visual assessment, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (seventh row) and the cursor position is the same for all images in each panel.



tendency (especially in NC), leading to significantly higher FPR compared to both FIRST and FIRSTv3 (p_{fp} <0.0005 for both). The standard-space maps in Figure 3.7 (third row) showed a general overestimation at the hippocampus-amygdala border (i.e., most dorsoanterior direction), the body and dorsal tail sections (left panel), whilst FN findings (right panel) were relatively infrequent and limited to medial boundaries of the hippocampus head.

There was no significant difference in overall Dice coefficients for the three CF methods (CFNS: 0.794±0.032; CFRS: 0.801±0.040; CFRS-NMI: 0.801±0.039, p_d=1.00 for all), but the CFNS method had significantly higher FPR (0.010±0.003) than both of the reference-space CF selection methods (CFRS: 0.008±0.003; CFRS-NMI: 0.008±0.004, p_{fp}= 0.0005 for both). This was, however, somewhat counter-balanced by a significantly reduced FNR for CFNS (0.181 ±0.042) compared to CFRS-NMI (0.203±0.043, p_{fn}= 0.007). For both CFRS and CFRS-NMI, the standard-space maps (Figure 3.7: fifth and sixth row, respectively) showed a relatively low frequency of FP findings (left panel), restricted to dorsolateral boundaries of the hippocampus, but underestimation of the hippocampal body and dorsoanterior aspects of the hippocampus head and tail (Figure 3.7 (right panel): fifth and sixth row, respectively). The pattern of FP and FN findings for CFNS (Figure 3.7: fourth row) was similar to that of the CFRS methods, except the spatial extent and frequency of FP findings generally increased (left panel) whilst the frequency of FN findings decreased (right panel).

3.3.2 BPSA data

Similar to the CMA data, the performance of each method on the BPSA data was generally not affected by side. The only exception to this trend was the FPR for FIRST, where there was a left<right asymmetry (F_{fp}(1,30)= 9.42, p_{fp}= 0.003). The initial repeated measure ANOVA designs also revealed no groupwise differences in segmentation performance for each method. As for the CMA data, the paired t-test results (presented in Figure 3.9) gave a strict relative ordering (1-6) of the segmentation methods for each performance measure. Based on a simple

sum of these rankings, the overall ordering of the segmentation methods on the BPSA data was as follows

(i) CFRS; (ii) CFRS-NMI; (iii) CFNS; (iv) FIRSTv3; (v) FS; (vi) FIRST.

The three CF methods clearly outperformed all other segmentation methods on the BPSA data, whilst the overall performance of FIRST was relatively poor.

Figure 3.8 *Method comparison on the BPSA data.* Box plots showing Dice coefficients (top), FPR (middle) and FNR (bottom) for each segmentation method on the BPSA data, with separate plots for BP patients (left) and NC (right). Outliers are indicated by a circle, extreme outliers by an asterisk. Please refer to Figure 3.4 for more details.

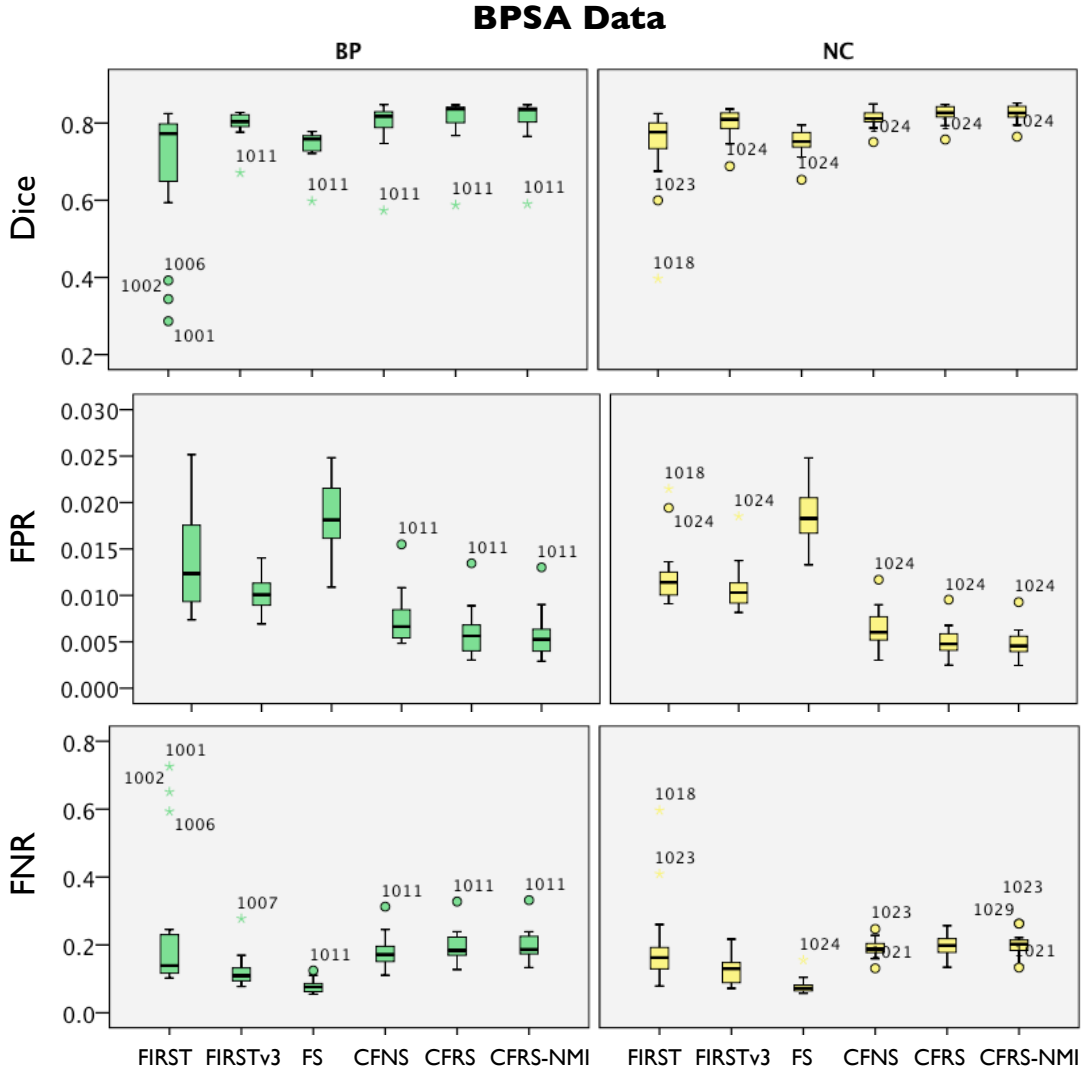
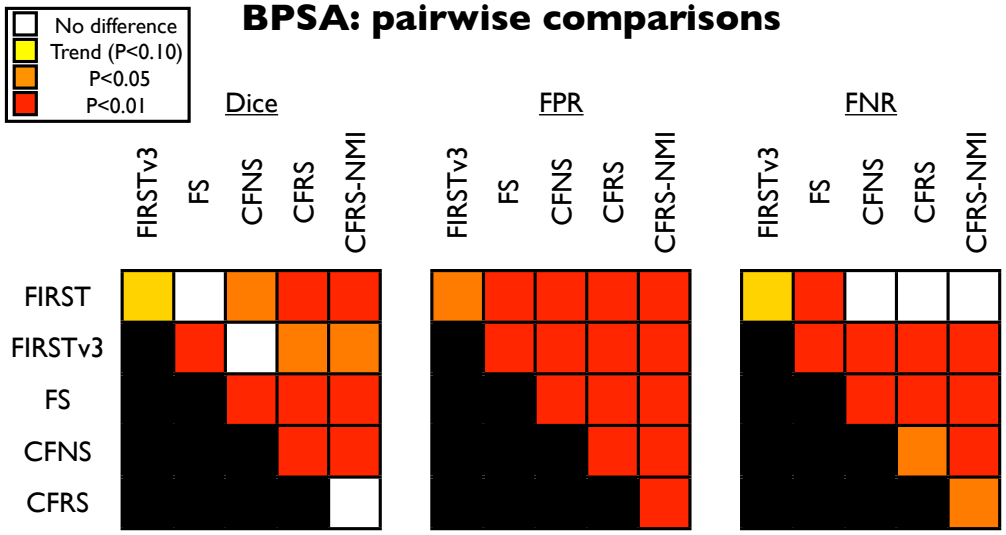


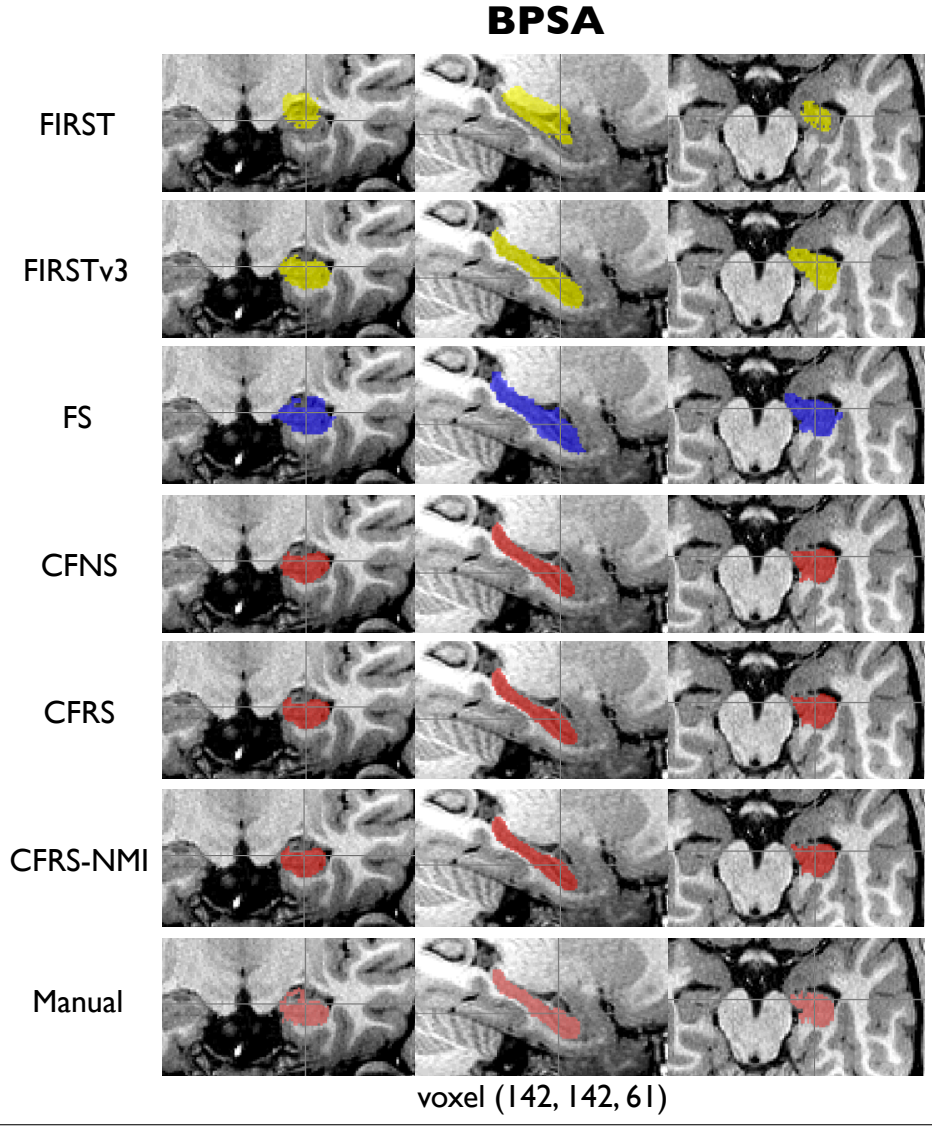
Figure 3.9 *Pairwise comparison of methods on the BPSA data.* Schematic representation of the two-tailed, paired-sample t-test results for the BPSA data, comparing Dice coefficients (left), FPR (middle) and FNR (right) across the hippocampal segmentation methods.



In fact, FIRST performed relatively well for the majority of subjects, but its overall segmentation performance was hindered by that of a few subjects, where poor initial registration with the standard (MNI152) space resulted in poor segmentation estimates (e.g., Figure 3.10: first row). Consequently, whilst FIRST displayed a median Dice coefficient comparable to other methods, and a lower median FNR than all of the CF methods (Figure 3.8: top and bottom), the method had significantly reduced Dice coefficients, higher FPR and FNR compared to most other methods (FIRSTv3: p_d= 0.029, p_{fp}= 0.018, p_{fn}= 0.035; FS: p_{fn}= 0.0005; CFNS: p_d= 0.016, $p_{fp} < 0.0005$; CFRS: p_d= 0.005, $p_{fp} < 0.0005$; CFRS-NMI: p_d= 0.005, $p_{fp} < 0.0005$). Interestingly, the standard-space maps in Figure 3.11 show overestimation of all anterior and dorsal hippocampal borders by all model-based methods (FIRST: top row, FIRSTv3: second row, FS: third row) and a high prevalence of FN counts at ventromedial borders for FIRST, attributed to a dorsoanterior shift of the entire hippocampus as shown in Figure 3.10 (first row).

The registration errors experienced with FIRST on the BPSA data were largely addressed by FIRSTv3, incorporating a hippocampus-specific spatial normalization step prior to model-fitting. Consequently, FIRSTv3 displayed significantly higher Dice coefficients and lower FNR

Figure 3.10 *BPSA segmentation estimates.* Multi-view images of the left hippocampal segmentation estimates for subject 1 of the BPSA dataset, showing the effect of poor registration of the query image to the standard (MNI152) space for FIRST (first row), the much improved FIRSTv3 result (second row), a 'greedy' labelling tendency for FS (third row), and relatively good volume estimation for all CF methods (fourth, fifth and sixth row) compared to the manual labels (seventh row). As for the CMA data, the cursor position is the same for each image to facilitate direct qualitative comparison of the segmentations.



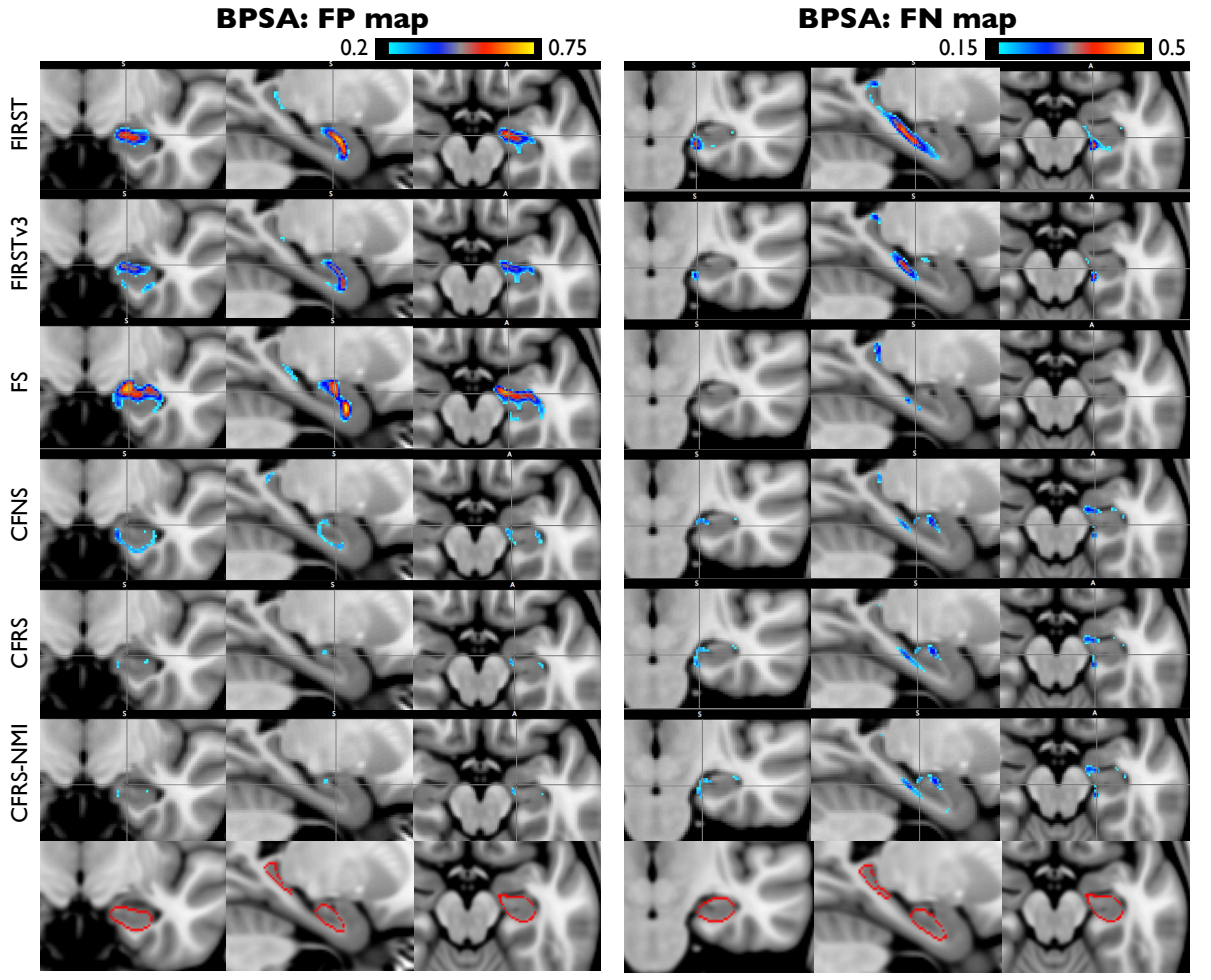
(0.798±0.037 and 0.125±0.045, respectively) compared to most other methods (FIRST: p_d= 0.029, p_{fn}= 0.035; FS: p_d <0.0005; CFNS, CFRS and CFRS-NMI: p_{fn} <0.0005 for all), although FN findings were still evident at ventromedial boundaries (Figure 3.11 (right panel): second row).

The FS tool displayed Dice coefficients comparable to that of human raters (0.746±0.039) and a significantly reduced FNR compared to all other methods (p_{fn} <0.0005 for all), only underestimating the most dorsoanterior sections of the hippocampus tail (Figure 3.11 (right panel): third row). However, visual assessment of segmentation estimates and the standard-space FP map (Figures 3.10 and 3.11 (left panel): third row) indicated a 'greedy' labelling tendency, giving significantly higher FPR compared to all other methods (p_{fp} <0.0005 for all).

Both reference-space CF methods (CFRS and CFRS-NMI) had significantly higher Dice (both 0.817±0.048) and lower FPR (both 0.005±0.002) compared to all other segmentation methods (FIRST: p_d= 0.005 for both, p_{fp} <0.0005 for both; FIRSTv3: p_d= 0.015 and 0.013, respectively, p_{fp} <0.0005 for both; FS: p_d <0.0005 for both, p_{fp} <0.0005 for both; CFNS: p_d <0.0005 for both, p_{fp} <0.0005 for both CFRS and CFRS-NMI). The segmentations for subject 1 of the BPSA dataset are given in Figure 3.10 (fifth and sixth row, respectively). Whilst there was no significant difference in Dice between CFRS and CFRS-NMI, the balance between FP and FN findings was somewhat different, with the CFRS-NMI method giving more conservative segmentation estimates. Consequently, the CFRS-NMI method had significantly increased FNR (0.200±0.041) compared to CFRS (0.196±0.041, p_{fn}= 0.006) and most other segmentation methods (FIRSTv3: 0.125±0.045, p_{fn} <0.0005; FS: 0.079±0.022, p_{fn} <0.0005; CFNS: 0.185±0.039, p_{fn}= 0.0005), but FPR was undoubtedly the lowest of all methods, even bettering the FPR of the CFRS method (p_{fp}= 0.002). The standard-space maps for CFRS and CFRS-NMI (Figure 3.11: fifth and sixth row, respectively) show an exceptionally low frequency of FP findings (left panel), with FN counts (right panel) in the most dorsoanterior and posterior parts of the hippocampus head. There was a similar pattern of FN findings for the CFNS method (Fig-

ure 3.11 (right panel): fourth row), but this native-space selection CF method overestimates medial and ventral borders of the hippocampus, and in particular, the posterior boundary of the hippocampal head.

Figure 3.11 *FP and FN maps for the BPSA data.* Left and right panels show multi-view images of the left hippocampus standard-space FP and FN maps, respectively, for each segmentation method on the BPSA data. A voxel-wise threshold of 20% FP finding across the dataset is applied to all FP maps, whilst the FN maps are thresholded at 15%. To aid visual assessment, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (seventh row) and the cursor position is the same for all images in each panel.



Of final note here is the repeated identification of the same subject outliers for the performance measures shown in Figure 3.8. Consistently poorer segmentation estimates were computed for subjects 01011 (BP) and 01024 (NC), giving rise to reduced segmentation performance. For

subject 01011, it appears that an unusually large amount of choroid plexus in the left lateral ventricle is causing overestimation of the left hippocampal tail by FS and the CF methods, whilst noisy data and particularly poor contrast seems responsible for overestimation of the left hippocampal-amygdala border and the ventral boundary (bilaterally) by all methods. For subject 01024, however, all methods "overestimate" the anterior hippocampal boundary, but the manual labels appear too conservative relative to the underlying neuroanatomy (as viewed on the T1-weighted MR image).

3.4 Discussion and conclusions

When comparing the performance of segmentation methods on different datasets, it is important to appreciate that a number of factors can affect segmentation accuracy, including image quality, manual labelling protocol, clinical status and demographics. In addition, model-based methods may be biased towards data used in their training. Indeed, decreased performance of FIRST, and to a lesser extent FIRSTv3 and FS, on the BPSA data suggests that these model-based methods are biased towards the CMA data previously used in their training. Such detrimental effects (of training bias) may stem from an over-dependency of model-based methods on their prior information. One prime example of this is the suprising, and somewhat concerning, output produced by FIRST when the query image was clearly misregistered with the standard (MNI152) image and the prior information therein. In several instances, the default two-step registration protocol placed the target hippocampus ventral and posterior to the FIRST model, such that the prior information did not correspond with the underlying subject anatomy. Even so, FIRST proceeded to produce a seemingly reasonable segmentation of the hippocampus, but one which was shifted dorsal and anterior to the true location of the hippocampus in the target MR image. On occasion, the hippocampal segmentation produced by FIRST was even in the ventricular space above the target hippocampus. We raise our concerns here, firstly because we wish to highlight the strong dependency of model-based methods on their prior information, but also

because we wish to reiterate the importance of visualizing the output segmentations produced by any given automated segmentation method — if only summary statistics were observed for these erroneous FIRST segmentations, such as the hippocampal surface area or volume, it may be wrongly assumed that all is well with the FIRST segmentations, when in reality, there may be misalignments such as that observed herein. Thankfully, the registration errors experienced with FIRST on the BPSA data were largely overcome with the hippocampus-specific spatial normalization of FIRSTv3, giving significantly improved segmentation estimates compared to FIRST. We predict similar improvements in registration accuracy and segmentation performance with other subcortical structures, using an appropriate structure-specific normalization step prior to model fitting with FIRST.

For the BPSA data, previously unseen by any of the segmentation methods, the three CF methods clearly outperformed all other methods, with Dice coefficients comparable to those previously reported [Aljabar et al., 2007; Heckemann et al., 2006; Lotjonen et al., 2010]. In particular, the improved segmentation result over FIRST and FS suggests a potential advantage of using flexible subject-specific selection methods like CF over more constrained model-based methods trained from more diverse data. Care should be taken, however, as previous results for the CMA data suggest that CF performance is restricted by reduced image quality and registration error. Previous work, investigating the performance of many of these segmentation methods on the CMA data [Bishop et al., 2010], found that bias field effects and excess neck in the T1-weighted MR images gave poor brain extraction and subsequent registration errors. These registration errors contributed to both FP and FN findings at most hippocampal boundaries for the CFNS method, although the relatively low frequency and even distribution of FP and FN counts suggested that the registration errors were uncorrelated. In this study, use of cropped CMA data rectified most of the previously-encountered inter-subject registration errors, but the CFNS method (and to a lesser extent the CFRS methods) still suffered from FP and FN findings at most of the hippocampal boundaries. The fairly ubiquitous distribution of

FP and FN findings suggests that registration errors are once again uncorrelated, but we must be mindful of the dependency of these atlas-based methods on robust and accurate registration. Finally, it is possible that a sub-optimal number of classifiers were chosen for CF on the CMA dataset, reducing overall segmentation accuracy.

Comparing all of the CF methods, our results suggest that similar segmentation estimates, and comparable overall segmentation performance, can be achieved with CFNS and the CFRS methods (CFRS and CFRS-NMI). However, the CFNS method is relatively more computationally expensive than the CFRS methods. This is because the CFNS method requires on-the-fly registration of every atlas anatomy A_i (in the data cohort) to each new query image Q, which in this case means i affine registrations plus s non-linear registrations (for s selected classifiers), and a computational cost that increases linearly with the number of atlases. On the other hand, each CFRS method only requires on-the-fly registration of the single query image Q to the reference R (i.e., 1 non-linear registration), since all of the atlas anatomies A_i can be processed 'offline' beforehand. As a result, the CFRS methods are somewhat more practical for large-scale clinical investigations, where typically several hundred images are processed. Finally, there was no overall difference in segmentation performance between CFRS and CFRS-NMI, so the choice of image similarity function (for the affine registrations and subsequent selection of top-ranked atlas anatomies) does not significantly impact on the final hippocampal segmentations.

In summary, the work presented here is an evaluation of both existing and improved MR hippocampal segmentation methods, highlighting general issues hampering the current clinical utility of these methods (potential model-bias, image preprocessing and registration errors) and therefore possible mechanisms for making them more robust and accurate in the presence of specific and substantial pathology. This work is an investigatory stepping-stone towards developing a dedicated hippocampal segmentation method for pathological subjects. As practical guidance towards the design and construction of this method, we suggest exploration of combinatorial methods that incorporate a spatially-varying weighting of subject-specific intensities

and model-based shape features. Where possible (i.e., at good contrasting boundaries) subject-specific image information should drive the segmentation, with models used to constrain the segmentation at poorer contrasting borders such as the hippocampus-amygdala border.

4

Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH)

Here we present Fast Marching for Automated Segmentation of the Hippocampus (FMASH): a novel approach using the Sethian Fast Marching (FM) technique to grow a hippocampal region of interest (ROI) from an automatically-defined seed point. Segmentation performance is assessed on two separate clinical datasets, utilizing expert manual labels as gold standard to quantify Dice coefficients, false positive rates (FPR) and false negative rates (FNR). The first clinical dataset (denoted CMA) contains 9 normal controls (NC) and 8 atrophied AD patients, whilst the second is a collection of 16 NC and 16 bipolar (BP) patients (denoted BPSA). An optimal and robust stopping criterion is established for the propagating FM front, and the final FMASH segmentation estimates are compared to two commonly-used methods: FIRST/FSL and Freesurfer (FS). Results show that FMASH outperforms both FIRST and FS on the BPSA data, with significantly higher Dice coefficients (0.80 ± 0.01) and lower FPR. Despite some intrinsic bias for FIRST and FS on the CMA data, due to their training, FMASH performs comparably well on the CMA data, with an average bilateral Dice coefficient of 0.82 ± 0.01. Furthermore, FMASH most accurately captures the hippocampal volume difference between NC

and AD, and provides a more accurate estimation of the problematic hippocampus-amygdala border on both clinical datasets. The consistency in performance across the two datasets suggests that FMASH is applicable to a range of clinical data with differing image quality and demographics.

4.1 Introduction

As mentioned in Chapter 3, previous hippocampal segmentation methods can be broadly classified into three groups: region-growing, atlas-based, and parameterized-modelling approaches. Whilst some methods may still be identified as belonging to one particular classification [Taylor and Barrett, 1994; Carmichael et al., 2005; Heckemann et al., 2006; Aljabar et al., 2007; Patenaude, 2007], there is a progressive trend towards 'multi-approach' segmentation methods that combine aspects of two or more groups to achieve increased robustness and segmentation accuracy [Fischl et al., 2002; Chupin et al., 2007; Colliot et al., 2008; Morra et al., 2008; van der Lijn et al., 2008; Chupin et al., 2009b; Lotjonen et al., 2010; Wolz et al., 2010]. This work will focus on region-growing methods and parameterized-models, but will utilize aspects of atlas-based approaches. In turn, the resultant segmentation performance will be compared to two commonly-used methods: the parameterized model FIRST/FSL [Patenaude, 2007; Woolrich et al., 2009], and the combined atlas-based and parameterized-model Freesurfer (FS) [Fischl et al., 2002].

Region-growing methods were first introduced in Chapter 3. Here we provide more examples and focus on the manual intervention often required for this category of segmentation methods. In previous work [Taylor and Barrett, 1994], competitive region-growing from one or more chosen seeds was performed by comparison of border voxels to the first-order statistics of voxels that had already been absorbed (or added) to the ROI. Combining this concept with a deformable model, Ashton et al. [1997] elastically deformed a region from a line of seeds in the presence of constraining forces (elastic surface tension, deviation from the expected sur-

face normal and resistance from surrounding tissue). Both the initial line of seeds and the deformation constraint were derived from manually-segmented contours in three orthogonal slices. In turn, a statistical model of hippocampal geometry was combined with elastic deformations, utilizing displacement forces derived from matching local intensity profiles [Kelemen et al., 1999]. More recently, several methods have focused on the simultaneous extraction of both the hippocampus and the amygdala [Yang and Duncan, 2004; Yang et al., 2004; Chupin et al., 2007; Colliot et al., 2008; Chupin et al., 2009b]. The work of Chupin et al. [2009b] has proved particularly effective, so it is useful to track the progressive development of this method over time. The competitive region-growing method previously mentioned in Chapter 3 [Chupin et al., 2007] was the initial development of the algorithm, requiring manual definition of two seed points and a bounding box, with the deformation constraint based on prior knowledge of anatomical features in the vicinity of automatically-detected landmarks. Applied to clinical data [Colliot et al., 2008], the algorithm detected significant hippocampal volume reductions for both mild cognitively impaired (MCI) and AD patients with respect to normal controls, facilitating accurate group classification. Subsequent development gave rise to the final automated region-growing method [Chupin et al., 2009b], imposing additional energy constraints and automated detection/correction of atlas mismatch for improved segmentation performance, and similar group classification accuracy [Chupin et al., 2009a]. Instead of initializing the algorithm with a manually-defined seed point, Chupin et al. [2009b] utilized probabilistic atlas priors for automated initialization, defining much larger hippocampal and amygdala objects from the maximal probability zone of the respective atlases. As such, the method was essentially a voxel re-classification approach on these initial objects, and not strictly a region-growing method per se.

Most previous and existing region-growing methods therefore still require time-consuming and knowledge-based manual intervention (for definition of a specific bounding box and seed points), with the resultant segmentation dependent on good initialization. Furthermore, the ma-

jority of studies focus on development and evaluation of an individual segmentation method on a single dataset, reporting a non-standardized selection of performance measures, such that direct comparison of hippocampal segmentation methods is compromised.

Here we present *Fast Marching for Automated Segmentation of the Hippocampus* (FMASH): a novel region-growing method for automated hippocampal segmentation on structural MR data. The algorithm uses a 3D Sethian Fast Marching (FM) technique (Section 4.2.5) to propagate a hippocampal ROI from an automatically-defined seed point p_0 (Section 4.2.4). The propagating ROI is driven by a potential function (Section 4.2.7) comprised of both subject-specific intensity features and a model-based shape prior. In contrast to Chupin et al. [2009b], the FMASH algorithm extracts a subject-specific, single-voxel hippocampal seed from the structural MR data (based on local means analysis) and does not rely on competitive growth of the neighbouring amygdala structure. The FMASH algorithm is applied to two clinical datasets (Section 4.2.1) with segmentation performance assessed according to three common measures (Section 4.2.9), providing both valuable assessment of method performance on multiple datasets and ease of comparison with previous reported methods. An optimal and robust stopping criterion is established for the propagating FM front (Section 4.3.3) and the resultant FMASH segmentation compared to FIRST and FS (Section 4.3.4). Additionally, Section 4.4 provides context for the FMASH algorithm based on the most recent literature.

4.2 Materials and methods

4.2.1 Clinical MR data

This work utilized the CMA and BPSA clinical data detailed in Section 2.6, consisting of T1-weighted MR images and their corresponding expert manual labels: 9 normal controls (NC) and 8 highly-atrophied AD patients for CMA; 16 NC and 16 bipolar (BP) patients for BPSA. Evaluation of the FMASH algorithm on both of these datasets was motivated by the desire to

construct a general hippocampal segmentation method that worked well for a range of diseases and different image acquisition parameters.

4.2.2 Software and image preprocessing

Tools from the FMRIB Software Library (FSL [Smith et al., 2004; Woolrich et al., 2009], www.fmrib.ox.ac.uk/fsl) were used to generate the hippocampal spatial priors (Section 4.2.3) and standard-space performance maps (Section 4.2.9), with the remainder of the algorithm currently implemented in MATLAB (www.mathworks.com).

The MR images were initially cropped to remove excess neck, and BET [Smith, 2002] was used to remove most of the non-brain structure. The brain-extracted data was used for affine registration tasks, whilst the whole-brain data was used for non-linear transformations, as described below.

4.2.3 Generation of hippocampal spatial priors

Atlas-based, label-propagation and decision-fusion approaches (such as Classifier Fusion, CF) have previously provided segmentation estimates for the hippocampus, with label-fusion performed either in the native (subject) space [Heckemann et al., 2006; Aljabar et al., 2009; Bishop et al., 2010] or a standard reference space [Aljabar et al., 2007]. Here, a similar approach was used to construct the hippocampal spatial priors, propagating all subject data (MR images and corresponding manual labels) to the standard MNI152 space via a multi-step registration scheme: an initial three-step affine registration with FLIRT [Jenkinson and Smith, 2001], followed by a multi-resolution non-linear warp with FNIRT [Andersson et al., 2008]. From this repository of spatially-normalized data, the target (to-be-segmented) image was excluded, manual labels were fused (summing trilinearly-interpolated values) and averaged to generate a standard-space hippocampal prior for each individual subject. Each inverted non-linear warp subsequently transferred the standard-space hippocampal prior to the corresponding subject

(native) space for use in the FMASH algorithm. For all affine registrations, FLIRT was executed with 12 DOF and maximization of the normalized cross-correlation similarity function. For all non-linear registrations with FNIRT, a control point spacing of 10mm was specified, with a sub-sampling scheme of 4,4,2,2,1,1, a bending energy model for regularization of the warp-field, and weightings for the membrane energy regularization of 300,150,100,50,40,30.

4.2.4 Image processing and seed extraction

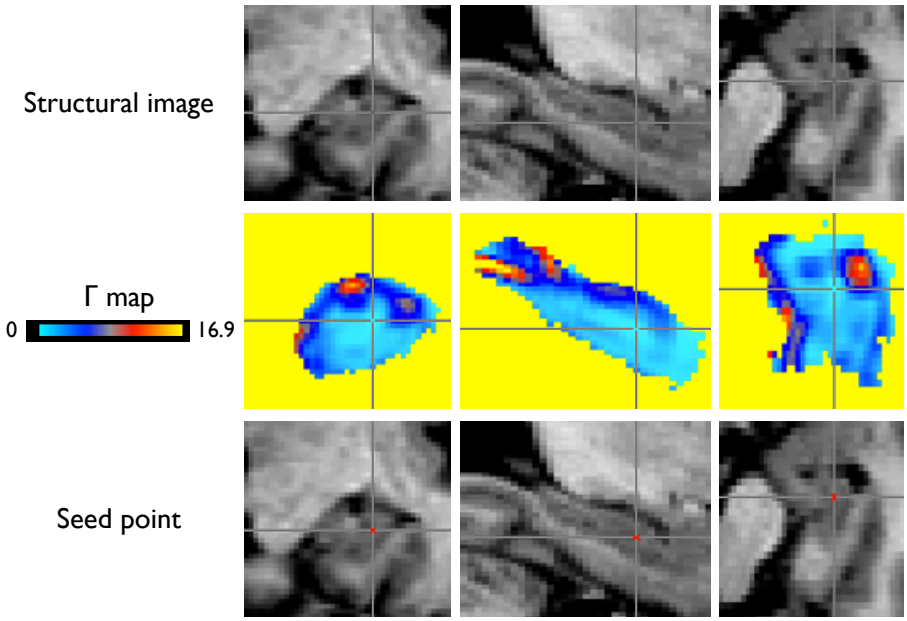
For each subject, the minimum and maximum index of non-zero voxels along each dimension of the hippocampal spatial prior defined a rectangular box (i.e., cuboid) surrounding the hippocampus. Maroy's Local Means Analysis (LMA) method for dynamic PET images [Maroy et al., 2008] was adapted to extract a series of points with minimal local intensity variance within this ROI. For all voxels n in the hippocampal ROI, the amplitude of local signal variations was estimated by

$$\Gamma_n = \frac{1}{\#V_n - 1} \times \sum_{j \in V_n} \frac{(I_j - \mu_n)^2}{\mu_n} \tag{4.1}$$

where $\#V_n$ is the number of voxels contained within the local neighbourhood V_n of voxel n, μ_n is the mean signal within V_n, and I_j is the signal at a neighbouring voxel j. Spatial neighbourhoods of 5x5x5 and 5x3x5 voxels were considered for the BPSA and CMA data, respectively, maintaining a similar spatial extent (in mm) of V_n for differing voxel dimensions.

We defined $\Lambda = \{n \,|\, \forall j \in V_n, \Gamma_n < \Gamma_j\}$ the set of locations of all local minima of the resulting "Γ map" (i.e., the set of voxels that are local minima with respect to the local intensity variance measure). These local minima are points of stability in the hippocampal ROI, usually the least affected by noise and image artifacts. Defining the most central local minimum as the single hippocampal seed point $m = p_0$ thus provided stable and robust initialization of the FM algorithm. Figure 4.1 shows the original T1-weighted MR image (top row), the Γ map (middle row), and the single hippocampal seed point (bottom row) for subject 1 of the CMA dataset.

Figure 4.1 *FMASH seed point selection.* Coronal- (left), sagittal- (middle) and axial- (right) view images of the structural T1-weighted MR data (top row), the Γ map (middle row), and the selected seed point (bottom row) for subject 1 of the CMA dataset. For ease of comparison, the cursor shows the position of the seed point in each image.



4.2.5 Active contours: a global minimum

Classical deformable models are typically driven by an energy term composed of both internal and external constraints, requiring precise initialization for extracting the path of minimum energy between two fixed extremities. Applied to aerial road images, angiographic images of the eye and cardiac MR images, Cohen and Kimmel [1997] simplified this model to external forces only and solved the minimal energy path problem in 2D, using the Sethian FM algorithm for propagation of the initial front [Sethian, 1996]. Later, the FM method was extended to 3D images [Deschamps and Cohen, 2000] with application to path tracking in virtual endoscopy, computed tomography (CT) colonoscopy, and extraction of brain vessels in a magnetic resonance angiography (MRA) scan. These minimal energy path active contour approaches aim to find the curve of minimum energy amongst all curves joining a seed point p_0 to any other point p. A path C(s) in the image is found that minimizes the energy

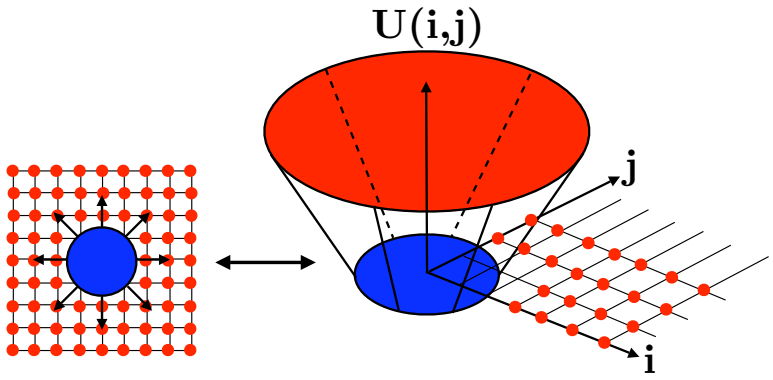
$$E(C) = \int_{\Omega} \tilde{P}(C(s))ds \tag{4.2}$$

where Ω is the curve domain $[0, L]$, L is the length of the curve and $\tilde{P}$ is the input potential. The arrival time surface or minimal action U, is then defined as the minimal energy integrated along the path

$$U(p) = \inf_{\mathcal{A}_{p_0,p}} E(C) = \inf_{\mathcal{A}_{p_0,p}} \left\{ \int_{\Omega} \tilde{P}(C(s))ds \right\} \tag{4.3}$$

with $\mathcal{A}_{p_0,p}$ denoting the set of all possible paths between p_0 and p. The arrival time surface is so-called because it gives the arrival time of the propagating front at any given point in space, illustrated in Figure 4.2 for a 2D problem. The front progresses along the path of least resistance, adding voxels which constitute the lowest potential $\tilde{P}$, thus finding the minimum energy curve for the structure of interest (in this case, the hippocampus).

Figure 4.2 *The Fast Marching (FM) arrival time surface.* Schematic illustration of the minimal energy path active contour approach in 2D, with the initial curve shown in blue and pixels represented by red dots. On the left, we present a hypothetical isotropic propagation of the FM front across a homogeneous image field, with a later arrival time surface $U(i, j)$ shown on the right. The arrival time surface is so-called because it gives the arrival time of the propagating front at any given point in space. Conceptually similar to level-set methods, a slice through the arrival time surface at any given height (or time) gives the spatial extent of the propagating front at that time.



4.2.6 FM resolution in 3D

Noticing that the arrival time surface U satisfied the Eikonal equation

$$\|\nabla U\| = \tilde{P} \tag{4.4}$$

Sethian [1996] presented a fast and efficient method for its construction in 2D, based on the fact that information was propagating *outwards* from the initial seed point (p_0), needing only

one pass on the image. Extended to 3D [Deschamps and Cohen, 2000], the details of the FM algorithm are as follows

1. *Initialization*

   - For each grid point p, let $U(p) = \infty$ (large positive value). Label all points as *far*.
   - Set the start point p_0 to be zero: $U(p_0) = 0$ and label it *trial*.

2. *Marching Loop*

   - Let $(i_{min}, j_{min}, k_{min})$ be the trial point with the smallest U value.
   - Label the point $(i_{min}, j_{min}, k_{min})$ as *alive*, and remove it from the trial list.
   - For each of the 6 neighbouring grid points (i, j, k) of $(i_{min}, j_{min}, k_{min})$
     - if (i, j, k) is labelled as far, then label it trial and compute $U_{i,j,k}$ according to the following equation

$$\begin{aligned} \left(max\left\{u - min\left\{U_{i-1,j,k}, U_{i+1,j,k}\right\}, 0\right\}\right)^2 &+ \\ \left(max\left\{u - min\left\{U_{i,j-1,k}, U_{i,j+1,k}\right\}, 0\right\}\right)^2 &+ \\ \left(max\left\{u - min\left\{U_{i,j,k-1}, U_{i,j,k+1}\right\}, 0\right\}\right)^2 &= \tilde{P}_{i,j,k} \end{aligned} \tag{4.5}$$

       where $u = U_{i,j,k}$;

     - if (i, j, k) is *trial*, then recompute $U_{i,j,k}$ according to equation 4.5 and update if the new value computed is smaller.

With each iteration through the *Marching Loop*, the seed point p_0 propagates outwards along the path of least resistance, adding a voxel which constitutes the lowest potential $\tilde{P}$, thus finding the (minimal action) arrival time surface U. A series of quadratic equations need to be sequentially solved in order to compute U, as detailed in Appendix A. Here, U was computed for all voxels in the hippocampal ROI, before thresholding to give the resultant segmentation estimate — a volumetric output. Computation of an accurate and robust stopping criterion is detailed in

Section 4.2.10. The novel aspect of this work is the automated initialization of the propagating front using a single-voxel hippocampal seed derived from the structural MR data (Section 4.2.4), and the potential function presented below.

4.2.7 A novel potential function

We defined a novel input potential $\tilde{P}_m$ at voxel n, associated with the curve joining the seed point m to any other voxel n:

$$\tilde{P}_m(n) = \frac{(I_n - \mu_m)^2}{{\mu_m}^2} - \lambda \log p\left(Hipp(n)\right) \tag{4.6}$$

Here, μ_m is the mean signal within the local neighbourhood V_m of the seed point m, I_n is the signal at any other voxel n, $p(Hipp(n))$ is the probability of voxel n being labelled as hippocampus in the training data, and λ is the weighting parameter for the prior term. This potential utilizes both image intensity features (first term) and prior probabilities established from training data (second term) to compute the path of lowest potential and least resistance for the propagating FM front.

Our initial work, utilizing only an intensity-based potential function, demonstrated the need for the hippocampal spatial prior; providing further constraint of the propagating FM front at poor-contrast intensity boundaries, such as the hippocampus-amygdala border. The form of the potential function is based on the observation that the first term takes the form of a log Gaussian likelihood probability, and thus the natural Bayesian formulation for a log posterior probability involves the addition of a log prior probability term. The weighting parameter λ provides a relative normalization of the likelihood and prior terms, as is common in such posterior probabilities. In addition, as each term is dimensionless, λ is independent of the scale of the image intensities and should therefore remain constant over a large range of datasets.

4.2.8 Determining the weighting parameter λ

Where possible, we wished the propagating FM front to be driven by subject-specific intensity features, only incorporating model-based shape features when intensity contrast was lacking. Consequently, we searched for the minimum value of the parameter λ that constrained the FM front at these poor-contrast boundaries. For a given λ and seed point m, we computed the potential at every voxel n in the hippocampal ROI (Equation 4.6) and used this information to generate a 3D potential map. By empirically tuning the value of λ (searching over the range 0 to 100 using increments of 10, then 1, then 0.01), a series of 3D potential maps were generated for a small subset of test subjects (2 NC and 2 patients) in each dataset. Section 4.3.1 describes the impact of λ on these 3D potential maps and its determined value. The value of λ was subsequently fixed for all subjects.

4.2.9 Performance measures

Segmentation performance was assessed using the same combination of performance metrics as the previous chapter: Dice coefficient, false positive rate (FPR) and false negative rate (FNR). Additionally, these three metrics were combined to give an indication of overall segmentation performance, as described below. The performance measures served two purposes: firstly, during development of the FMASH algorithm, they were used to compute the FM stopping criterion (Section 4.2.10), and secondly, they evaluated the resultant FMASH segmentation (Section 4.2.11).

Success and error rates

As noted in Chapter 2 (Section 2.5.1), the size of the image field of view (FOV) affects the FPR but not the FNR, such that the FPR are very low when considered over the whole image as opposed to a hippocampal ROI. However, since we define the hippocampal ROI as the minimum

and maximum index of non-zero voxels along each dimension of the hippocampal spatial prior, the size of the ROI (and FOV) varies across individuals. Thus, to maintain a constant FOV, we again consider a ROI of volume 91125 mm^3 (equivalent to a 45x45x45 mm cuboid) in our calculations of the TN set.

Standard-space FP and FN maps

Standard-space maps were similarly generated to assess the spatial distribution of FP and FN voxels for each segmentation method. The three-step affine transformation, computed during generation of the hippocampal spatial priors, propagated all subjects' FP and FN voxels from their native space to MNI152 space. Here, the average FP and FN maps were computed (summing trilinearly-interpolated values), with voxel values corresponding to the fraction of subjects showing a FP or FN result at that voxel.

Combined Performance Measure (CPM)

Segmentation estimates of the left and right hippocampi were obtained for each segmentation method, and the corresponding performance measures computed. These measures were normalized to 0-1 (inverting FPR and FNR such that a high normalized measure corresponded to better segmentation performance). For each side s, the automated segmentation method a, and each performance metric M, the normalized performance measure $M_{norm}(i, a, s)$ for subject i was computed using the following expression

$$M_{norm}\left(i,a,s\right) = \frac{M\left(i,a,s\right) - \min_{\{i,a\}}\left(M\left(i,a,s\right)\right)}{\max_{\{i,a\}}\left(M\left(i,a,s\right)\right) - \min_{\{i,a\}}\left(M\left(i,a,s\right)\right)} \tag{4.7}$$

For each subject, the normalized performance measures were then summed to give the left- and right-side *combined performance measure* (CPM) for any given segmentation method. Since there are three composite performance measures (Dice, FPR, FNR), each normalized to 0-1, the CPM can take any value between 0 and 3, corresponding to the minimum and maximum obtainable segmentation performance, respectively.

4.2.10 FM stopping criterion

During development of the FMASH algorithm, we computed the stopping time, U, for the propagating FM front using the following approach. Firstly, we thresholded each subject's left and right hippocampal arrival time surface at a range of different arrival times ($0.12 \leq U \leq 0.30$ in increments of 0.02), generating a series of left and right segmentation estimates for each subject, from which the corresponding performance measures were computed. The arrival time corresponding to the segmentation with the *maximum* combined performance measure (CPM: Section 4.2.9) defined the optimal stopping criterion for the propagating FM front. In this way, we computed the optimal stopping time for the left and right hippocampus of each subject.

Within each clinical dataset, we then adopted a leave-one-out approach to compute an independent (i.e., unbiased) stopping time for each subject. This involved excluding the target (to-be-segmented) subject and averaging the remaining optimal stopping times for the other subjects in that dataset. These independent stopping times were extremely stable across all subjects and both clinical datasets (Section 4.3.3). Consequently, we computed the dataset-average of these independent stopping times, which was identical for both clinical datasets. This single stopping time was subsequently used to threshold every subject's arrival time surface, with the resulting FMASH segmentations used for validation of the algorithm, as described below.

4.2.11 Segmentation evaluation

Following development of the FMASH algorithm and computation of the FM stopping criterion, FMASH segmentation performance (on the CMA and BPSA datasets) was compared to FIRST and FS using the aforementioned performance measures and the ANOVA statistical analysis technique (SPSS statistics software, version 19.0, www.spss.com/statistics/). The left and right hippocampal segmentation estimates and corresponding performance measures (Dice, FPR, FNR and CPM) were computed separately for each method on both clinical datasets.

For each dataset and each of the three common measures (Dice, FPR and FNR), we initially employed a 2*2 (side*group) ANOVA to test for main effects of side (left, right) and group (patients, controls) for each method, with side as the within-subjects repeated measure and clinical group as the between-subjects factor. Paired-sample t-tests with Bonferroni correction for multiple comparisons were used to identify differences in side and group, with F_d, F_{fp}, F_{fn} and p_d, p_{fp}, p_{fn} denoting the F-statistic and corrected p-values for Dice, FPR and FNR, respectively. It is worth noting that we did not perform this initial ANOVA for the CPM. Since separate computation of the left- and right-side normalized performance measures (Equation 4.7) results in slightly different normalization (owing to slightly varying min/max range of the composite values), the left- and right-side normalized performance measures, and the corresponding CPM, should not be directly compared.

For each side, we subsequently tested for overall difference across the segmentation methods and clinical group-by-method interactions using a 3*2 (method*group) mixed ANOVA design, with method as the within-subjects factor and clinical group as the between-subjects factor. Where Mauchly's test indicated that the assumption of sphericity for method had been violated ($p<0.05$), degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. Paired-sample t-tests with Bonferroni correction for multiple comparisons were used to identify between-method differences; Lp_d, Rp_d, Lp_{fp}, Rp_{fp}, Lp_{fn}, Rp_{fn}, Lp_{cpm}, Rp_{cpm} denote bilateral corrected p-values for Dice coefficient, FPR, FNR and CPM, respectively.

Finally, to investigate the dependence of the FMASH algorithm on the hippocampal spatial prior, FMASH segmentation estimates were obtained for both clinical datasets using a series of different priors, and the corresponding performance measures were computed. For the results presented herein (Figures 4.4 - 4.8), a leave-one-out approach was adopted for the generation of a general, mixed-dataset prior (representing differing labelling protocols and expert raters) with no bias towards the target (to-be-segmented) image anatomies. In addition, we explored the cross-dataset performance; segmenting the CMA data using a single-dataset prior gener-

ated from the BPSA labels only, and vice versa. For both clinical datasets, the left and right hippocampal performance measures were computed, and the aforementioned ANOVA designs were repeated to evaluate FMASH performance using the single-dataset priors.

4.3 Results

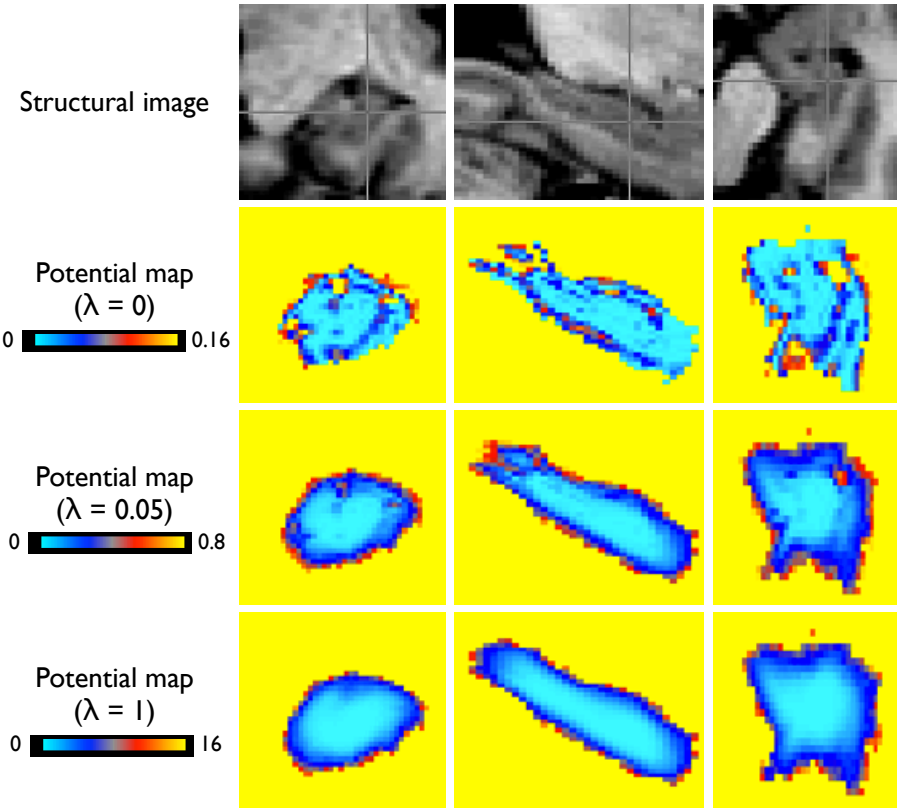
The results may be considered in two parts: Sections 4.3.1 - 4.3.3 refer to the initial development of the FMASH algorithm, whilst Sections 4.3.4 - 4.3.5 detail the segmentation evaluation results of the final FMASH algorithm. For both datasets, we present box plots of FMASH segmentation performance at the ten arrival time thresholds and 3D scatter plots of the normalized performance measures for the left and right hippocampus (Figures 4.4 and 4.5, respectively), computing the FM stopping criterion. In turn, Figures 4.6 - 4.8 compare the resultant FMASH segmentation with those obtained from FIRST and FS.

4.3.1 The weighting parameter λ

The weighting parameter λ controls the contribution of the hippocampal spatial prior towards the FM input potential. At $\lambda = 0$, the potential function depends only on image intensities (first term), with features on the potential map (Section 4.2.8 and Figure 4.3) corresponding only to intensity features in the MR image. Consequently, at good contrast boundaries in the MR image, the potential map was well-defined, showing the intricate details and complexity of the hippocampus boundary and providing sufficient resistance for the propagating FM front. However, at poor-contrast intensity boundaries, such as the hippocampus-amygdala border, there was inconsistent contrast on the potential map and therefore insufficient resistance to stop region growth; regions of lower contrast can be thought of as *holes* in the hypothetical barrier provided by high-contrast regions on the potential map, such that the FM front spills-out across the true hippocampal boundary. As the value of λ increased from zero, previously well-defined features on the potential map were smoothed by incorporation of probabilistic shape features

(second term), whilst poor contrast boundaries became more defined. We therefore searched for a value of λ that provided sufficient resistance (i.e., contrast) on the potential map at ill-defined intensity boundaries in the MR image, whilst maintaining subject-specific features at good-contrast hippocampal borders. Visual assessment of the 3D potential maps for a small subset of test subjects (Section 4.2.8), and the contrast at the hippocampal borders, estimated the value of λ at 0.05 for both datasets.

Figure 4.3 *FMASH potential maps.* Coronal- (left), sagittal- (middle) and axial- (right) view images of the T1-weighted MR data (top row) and a selection of potential maps for subject 1 of the CMA dataset, showing the impact of varying the weighting parameter λ from its minimum value (λ=0: second row) to an arbitrarily large value (λ=1: fourth row), before finding the optimized value of λ=0.05 (third row).



4.3.2 FMASH performance at varying arrival times

As shown in Figures 4.4 and 4.5, FMASH segmentation performance depended on the arrival time threshold applied to the corresponding arrival time surface. In the lower range, Dice co-

efficients increased with U, peaking for values of U in the range 0.16-0.18 on both datasets (Figures 4.4 and 4.5 (A): first row) before subsequently decreasing. As U increased, the FM front propagated further and obtained progressively larger estimates of hippocampal volume, with a resultant increase in FPR (Figures 4.4 and 4.5 (A): second row). On the other hand, a progressively larger hippocampal volume reduced the tendency for FN findings, with FNR showing an inverse dependency on U (Figures 4.4 and 4.5 (A): third row). Clearly, no single arrival time is optimal for all three performance measures; there is always a trade-off or compromise to be made. In Section 4.3.3 below, we present the results of combining these performance measures and computing the FM stopping criterion.

4.3.3 An accurate and robust stopping criterion

3D plots of the left and right hippocampal normalized performance measures are given in Figures 4.4 and 4.5 (B) for the CMA and BPSA data (left and right columns, respectively). A good arrival time threshold corresponds to high Dice coefficients, low FPR and low FNR, and will therefore cluster in the back top corner of the plots. As previously described (Section 4.2.10), we defined the arrival time corresponding to the *maximum* combined performance measure (CPM) as the optimal stopping criterion for the propagating FM front, before utilizing a leave-one-out approach to compute an independent (i.e., unbiased) stopping time for each subject. The (left, right) mean$\pm$SD of these independent stopping times was 0.18$\pm$0.00, 0.18$\pm$0.00 and 0.18$\pm$0.00, 0.17$\pm$0.00 for the CMA and BPSA data, respectively, giving the same bilateral-average stopping time of 0.18 for each dataset. Consequently, the stability of the stopping criterion, across all subjects and both clinical datasets, justified thresholding every subject's arrival time surface at this single stopping time of 0.18. We compare the resultant segmentation performance of FMASH with that of FIRST and FS below.

Figure 4.4 *FMASH performance: left hippocampus (LHipp).* (A) Plots of Dice coefficient (top row), FPR (second row) and FNR (third row) for the left hippocampus showing FMASH performance on the CMA data (left column) and BPSA data (right column) at varying arrival time thresholds. Boxes have lines at the lower quartile, median, and upper quartile values, with whiskers extending to 1.5 times the inter-quartile range. Outliers are indicated by a plus sign. (B) 3D scatter plot of the normalized performance measures, with the optimal arrival time clustering in the back top corner of the plot.

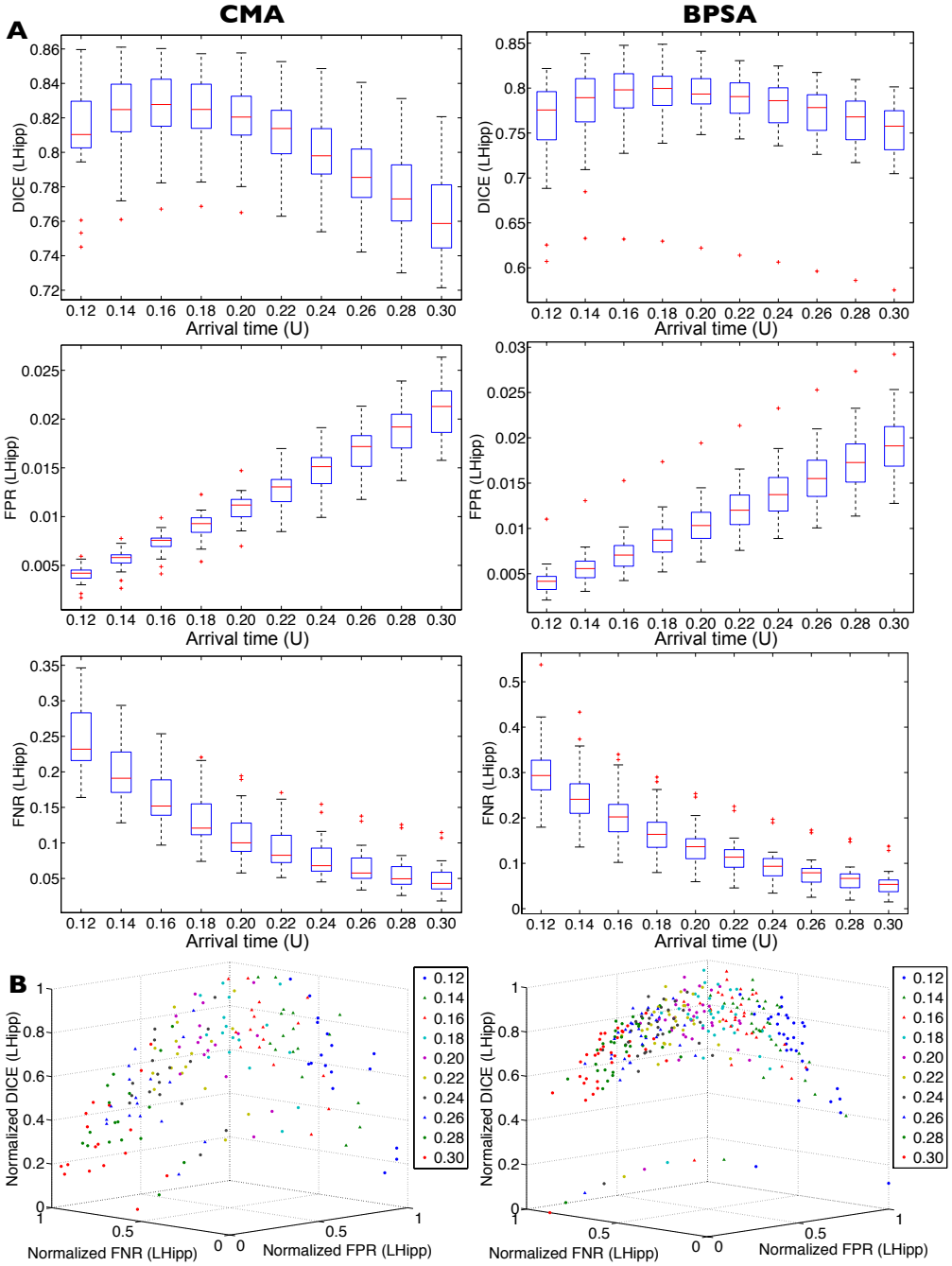
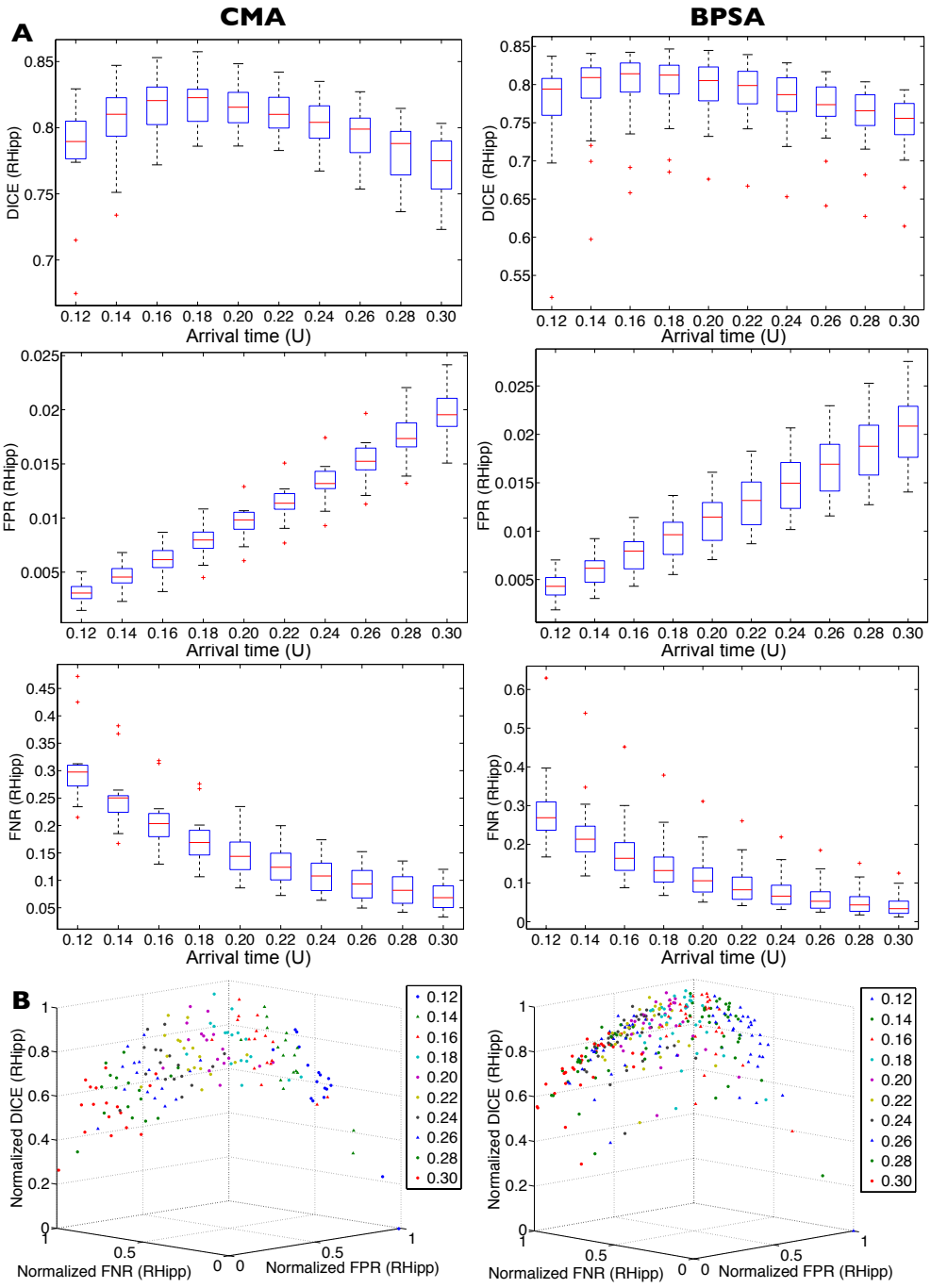


Figure 4.5 *FMASH performance: right hippocampus (RHipp).* Plots of right hippocampal FMASH performance on the CMA data (left column) and BPSA data (right column) at varying arrival time thresholds. Please refer to Figure 4.4 for more details.



4.3.4 Comparison to FIRST and FS

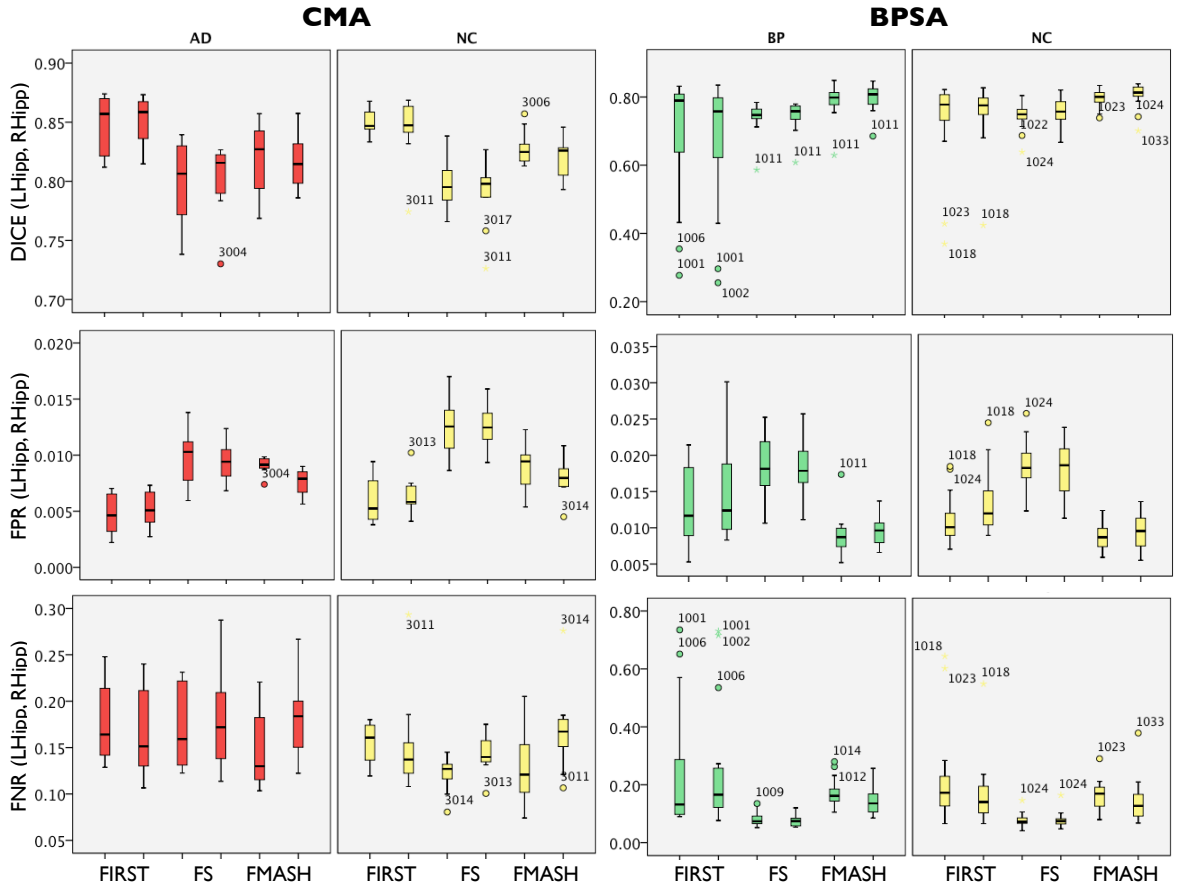
As mentioned in Chapter 3, we generally did not observe any left-right asymmetries in method performance for FIRST and FS using the initial 2*2 (side*group) ANOVA designs; the only exceptions being the FS FNR on the CMA data (where left<right, F_{fn}(1,15)= 6.33, p_{fn}= 0.012), and the FIRST FPR on the BPSA data (where left<right, F_{fp}(1,30)= 9.42, p_{fp}= 0.003). For FMASH, there were similarly some exceptions to this trend of left-right symmetry in segmentation performance: on the CMA data there was a left>right asymmetry in FPR (F_{fp}(1,15)= 5.27, p_{fp}= 0.019) and a correspondingly opposite trend in FNR (left<right, F_{fn}(1,15)= 6.42, p_{fn}= 0.012), whilst on the BPSA data, FMASH FNR was greater on the left side than the right (F_{fn}(1,30)= 4.38, p_{fn}= 0.023).

In general, results of the 3*2 (method*group) mixed ANOVA designs revealed a significant effect of method for each performance measure, but no clinical group-by-method interaction or group effect for either dataset. The only exceptions to this trend arose from a subset of the CMA data performance measures, where there was a group effect for right hippocampal FPR (AD<NC, Rp_{fp}= 0.018), a slightly significant bilateral group-by-method interaction for FPR (Lp_{fp}= 0.046, Rp_{fp}= 0.023), and no significant method effect for the right hippocampal FNR. The bilateral group-by-method interaction for the CMA data FPR was a result of both FIRST and FS having considerably higher FPR for NC compared to AD patients, whilst FMASH FPR showed negligible bilateral group effect: a bilateral increase of 23.1% was observed for FIRST (left: 23.3%; right: 22.9%), but the effect was even more pronounced for FS, with a bilateral increase of 31.7% (left: 28.2%; right: 35.2%).

On the CMA data (Figure 4.6: left panel), FMASH obtained significantly higher Dice coefficients (left: 0.824±0.024; right: 0.818±0.020) and lower FPR (left: 0.009±0.002; right: 0.008±0.001) compared to FS (Lp_d= 0.005, Rp_d= 0.009, Lp_{fp}= 0.003, Rp_{fp} <0.005), with FNR comparable to FIRST. Furthermore, FMASH outperformed both FIRST and FS on the

BPSA data (Figure 4.6: right panel), with significantly higher Dice coefficients (left: 0.793±0.039; right: 0.801±0.038) and lower FPR (0.009±0.002 for both left and right side; FIRST: Lp_d= 0.012, Rp_d= 0.004, Lp_{fp} and Rp_{fp} <0.005; FS: Lp_d, Rp_d, Lp_{fp} and Rp_{fp} <0.005).

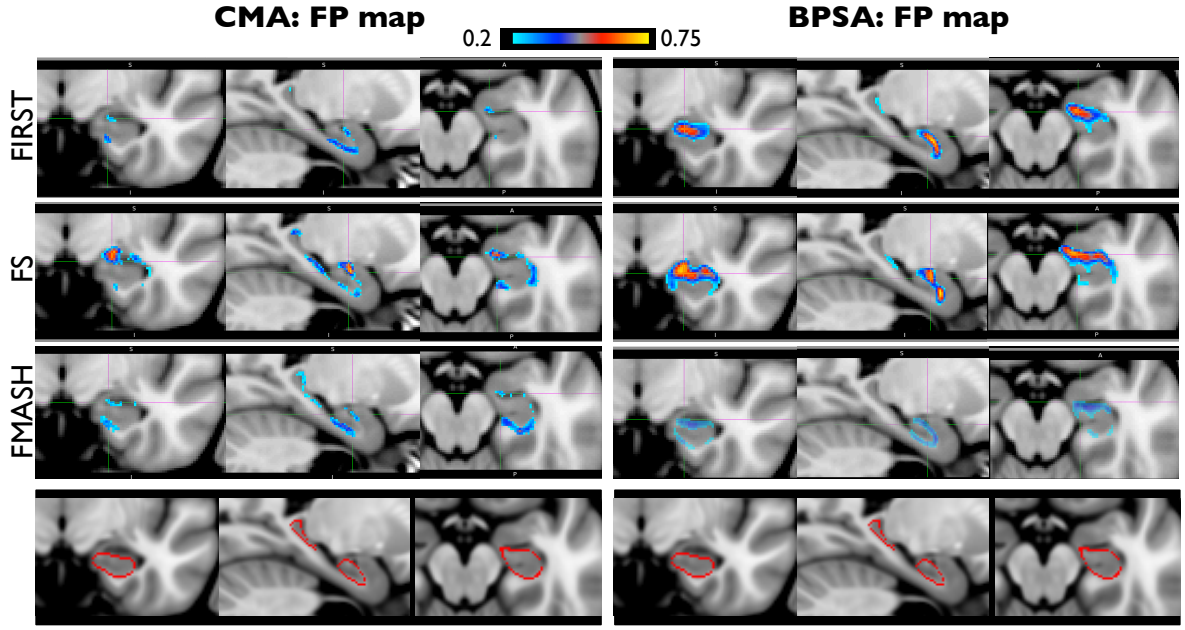
Figure 4.6 *Method comparison.* Box plots comparing FMASH segmentation performance with that of FIRST and FS on the CMA data (left panel) and BPSA data (right panel), showing (within each panel) Dice coefficients (top), FPR (middle) and FNR (bottom) for patients (left) and controls (right). Each duplet presents the left (left) and right (right) hippocampal measures. Boxes have lines at the lower quartile, median, and upper quartile values, with whiskers extending to 1.5 times the inter-quartile range. Outliers and extreme outliers are indicated by a circle and an asterisk, respectively, in addition to their corresponding subject ID.



Just as we used the CPM at different arrival times to find the FM stopping criterion, the CPM for each segmentation method revealed a strict overall ranking of the methods on each dataset. Considering both the left and right hippocampus of all subjects, we had 34 segmentation estimates with corresponding CPM for each method on the CMA data, and an additional 64 measurements for the BPSA data. FMASH had significantly higher CPM on the BPSA data compared

to both FIRST and FS (FIRST: Lp_{cpm}=0.006, Rp_{cpm} <0.005; FS: Lp_{cpm}, Rp_{cpm} both <0.005), obtaining the highest CPM for 86% of segmentations (left: 2.54±0.17; right: 2.62±0.14), with FIRST performing best for the remaining 14%. The ordering of FIRST and FMASH was reversed on the CMA data: FMASH performed best for 15% of segmentations, whilst FIRST obtained the highest CPM for 85% (left: 2.08±0.28; right: 2.28±0.36). Consequently, FIRST had significantly higher CPM than both FMASH and FS on the CMA data (FMASH: Lp_{cpm}= 0.002, Rp_{cpm}= 0.001; FS: Lp_{cpm}, Rp_{cpm} both <0.005). We discuss this heightened performance of FIRST on the CMA data in Section 4.4. Of final note here was the failure of FS to obtain a maximum CPM for any subject in either dataset, with a reduced average CPM of 32.4% (left: 31.1%; right: 33.8%) and 15.3% (left: 17.2%; right: 13.4%) relative to the best performing method on the CMA and BPSA data, respectively.

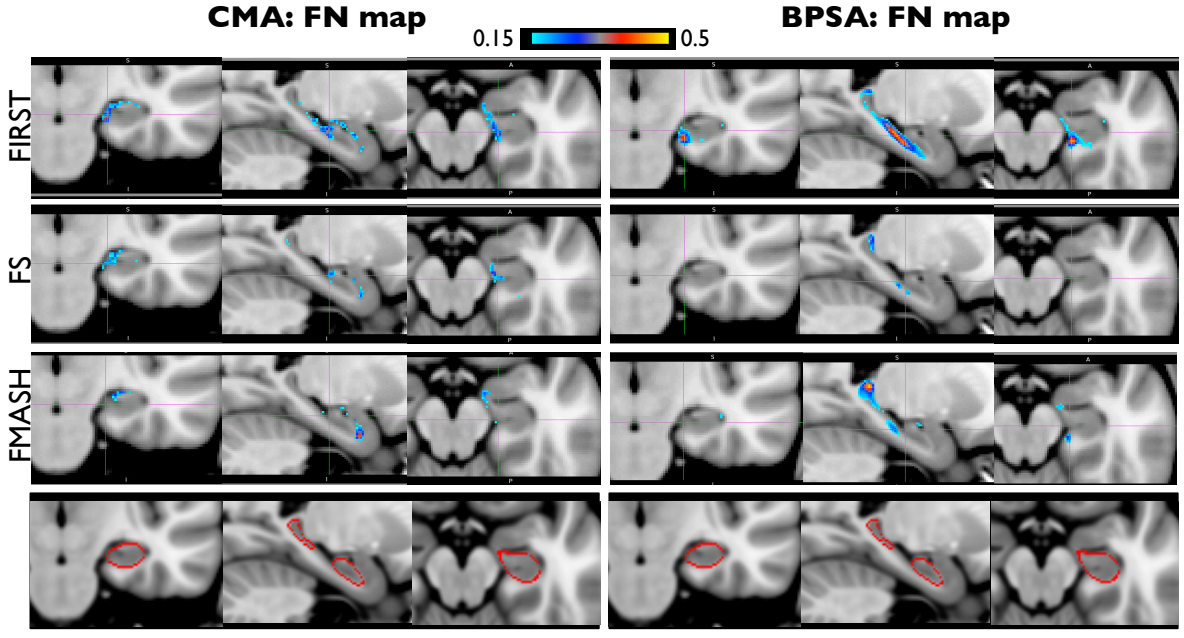
Figure 4.7 *False positive (FP) maps.* Each panel shows coronal- (left), sagittal- (middle) and axial- (right) view images of the left hippocampus standard-space FP maps for FIRST (top row), FS (second row) and FMASH (third row) on the CMA and BPSA data (left and right panel, respectively). A voxel-wise threshold of 20% FP finding across the dataset is applied to all FP maps. To aid visual comparison, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (fourth row) and the cursor position is the same for all images. Similar patterns of FP findings were observed for the right hippocampus.



The standard-space FP maps for the CMA data (Figure 4.7: left panel) showed a low frequency

of FP voxels for FIRST, with only slight overestimation (or spillover) at the most ventromedial boundary of the hippocampus head — a pattern similarly displayed by the FMASH algorithm. FS, on the other hand, displayed FP voxels in the body of the hippocampus and more severe spillover in the most dorsoanterior regions of the hippocampus head (i.e., hippocampus-amygdala border). On the BPSA data (Figure 4.7: right panel), both FIRST and FS showed a marked increase in FP findings, with severe spillover at the hippocampus-amygdala border. In contrast, FMASH displayed a low frequency of FP voxels at this boundary, and reduced spillover at the most ventromedial boundary compared to both itself and FIRST on the CMA data.

Figure 4.8 *False negative (FN) maps.* In a similar layout to Figure 4.7, each panel shows coronal- (left), sagittal- (middle) and axial- (right) view images of the left hippocampus standard-space FN maps for FIRST (top row), FS (second row) and FMASH (third row) on the CMA and BPSA data (left and right panel, respectively). A voxel-wise threshold of 15% FN finding across the dataset is applied to all FN maps. The hippocampal boundary defined by FIRST on the MNI152 standard image is shown (fourth row) and the cursor position is the same for all images. Similar results were observed for the right hippocampus.



Additionally, the standard-space FN maps for the CMA data (Figure 4.8: left panel) showed that medial boundaries and posterior regions of the hippocampus head were underestimated by FIRST (top row) and to a lesser extent FS (second row), whilst FMASH displayed FN voxels

at the most anterior tip of the hippocampus head (third row). FS maintained a low frequency of FN voxels on the BPSA data (Figure 4.8: right panel, second row). However, FIRST severely underestimated ventromedial borders (top row) and FMASH had a propensity for FN voxels in the most dorsoposterior regions of the hippocampus tail (third row).

4.3.5 Evaluation of the hippocampal spatial prior

Investigating the dependence of the FMASH algorithm on the hippocampal spatial prior revealed that, on average, the single/cross-dataset prior (using labels from the other dataset) resulted in a 3.1% drop in Dice coefficients for the CMA data (left: 0.9%; right: 5.3%) and a 5.1% drop for the BPSA data (left: 5.1%; right: 5.2%), compared to the results presented in Figure 4.6 (top row) for the mixed-data prior. Despite this reduction in Dice coefficients using the single-dataset prior, FMASH still obtained significantly lower FPR and higher CPM than FS on the CMA data (Lp_{fp}= 0.001, Rp_{fp} <0.005, Lp_{cpm}= 0.001, Rp_{cpm} <0.005), with FNR comparable to FIRST. For the BPSA data, FMASH retained the highest mean Dice coefficient and lowest FPR of all methods (0.756±0.006 and 0.013±0.000, respectively), with significance achieved against the FPR of FS (Lp_{fp}, Rp_{fp} both <0.005). Furthermore, FMASH still had the highest CPM of all methods on the BPSA data (left: 2.62±0.15; right: 2.33±0.19), with significantly improved segmentation performance compared to FS (Lp_{cpm}, Rp_{cpm} both <0.005).

4.4 Discussion and conclusions

We found that FMASH segmentation performance depends on the arrival time threshold applied to the corresponding arrival time surface — no single arrival time is optimal for all three performance measures; there is always a trade-off or compromise to be made. We presented the results of combining these performance measures and computing the stopping time for the propagating FM front (Section 4.3.3). Given that the same value of λ (0.05) and the same bilateral-average stopping time (0.18) were computed for both datasets, this suggests that we

have a robust stopping criterion that is applicable to a range of clinical data with differing image quality, disease-status and demographics.

As we previously reported [Bishop et al., 2010], both model-based methods, FIRST and FS, have an inherent bias towards their training data, of which this CMA data forms a small subset (5.1%). Contrary to the our initial findings, however, subsequent analysis showed that exclusion of this CMA subset from a re-trained FIRST model *did* alter its segmentation performance: there was a highly significant reduction in bilateral Dice coefficients for the re-trained FIRST model on our 17 CMA subjects ($p<0.0005$; 1.9% decrease for NC, 2.0% decrease for AD patients), suggesting that FIRST results are both dataset dependent and biased explicitly by inclusion of this particular CMA subset. We therefore expected FIRST and FS to dominate on this CMA data, but results presented herein demonstrate that this was not the case. FMASH still obtained significantly higher Dice coefficients and lower FPR compared to FS, lower FNR than FIRST, and a maximum CPM for 15% of segmentations on the CMA dataset. In contrast, FS failed to obtain a maximum CPM for any subject in either dataset, most likely due to a slight "greedy labelling" tendency yielding a higher FPR.

Comparing FMASH performance with that of FIRST and FS also revealed a surprising group-wise discrepancy in FPR for the two parameterized models: both FIRST and FS had considerably higher FPR for NC compared to AD patients, with this effect even more pronounced for FS, and in particular, the right hippocampus. This finding suggests that these commonly-used methods exaggerate the true hippocampal volume difference between NC and AD patients, by as much as two-fold, and that FS has a tendency to exaggerate these differences more for the right hippocampus; providing a strong argument for using the proposed FMASH method instead.

Previous studies comparing FIRST and FS [Morey et al., 2009; Pardoe et al., 2009], reported Dice coefficients in the range of 0.71-0.80 for FIRST and 0.73-0.82 for FS on NC data. The variability in Dice coefficients for FIRST and FS on our NC data (0.73-0.85 and 0.74-0.80,

respectively) are therefore in-line with previous studies, with the exception of the high Dice coefficients for FIRST on the CMA data (discussed above).

Here we provide context for the FMASH algorithm based on the most current literature, comparing FMASH Dice coefficients with those of recently-published methods. In the literature, mean Dice coefficients are either separately reported for NC and diseased patients, or as reported herein, for mixed cohorts (containing both NC and diseased patients). Recent studies using atlas-based approaches have reported Dice coefficients in the range 0.75-0.83 for NC [Carmichael et al., 2005; Heckemann et al., 2006; Aljabar et al., 2007 2009], 0.74-0.76 for temporal lobe epilepsy (TLE) patients (diseased side) [Hammers et al., 2007; Avants et al., 2010] and 0.80 for mixed cohorts [Leung et al., 2010]. However, values as high as 0.86 for NC [van der Lijn et al., 2008] and 0.82-0.88 for mixed cohorts [Lotjonen et al., 2010; Wolz et al., 2010] have been achieved with graph-cuts, embedded learning, and intensity modelling, respectively. Lotjonen et al. [2010] suggested differences in image quality, manual protocol, clinical status and demographics as possible causes of discrepancy between their two clinical datasets, which may also be contributing factors herein. The aforementioned semi-automated region-growing method of Chupin et al. [2007] found an average Dice coefficient of 0.84 for NC data, whilst their more recent automated method utilized a probabilistic prior, with automatic detection and correction of atlas mismatch, to achieve Dice coefficients of 0.87 ± 0.03 for NC and 0.84 ± 0.05 for a mixed cohort of NC and TLE patients [Chupin et al., 2009b]. The improved segmentation performance of Chupin et al. [2009b] on the NC data is not surprising given that the probabilistic prior was generated from NC subjects, segmented by the same investigator and using the same manual labelling protocol. The FMASH bilateral-average Dice coefficients of 0.82 ± 0.01 and 0.80 ± 0.01 for the CMA and BPSA dataset, respectively, are nevertheless comparable with the current literature for mixed cohorts. This said, there are possible improvements and extensions to be made to the FMASH method. For example, automated detection and correction of atlas mismatch, as implemented in Chupin et al. [2009b], and

generation of hippocampal priors using alternative registration algorithms, could provide improved segmentation performance. FMASH group classification accuracy could be investigated with application of the algorithm to publicly available data, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI, www.loni.ucla.edu/ADNI). Additionally, with an appropriate spatial prior, the FMASH algorithm has the ability to segment other brain structures, such as the neighbouring amygdala.

Finally, even with a small value of the weighting parameter λ, we appreciate that there is a dependence of the FMASH algorithm on the hippocampal spatial priors, generated from a set of manually-labelled images. The leave-one-out approach, used to generate the mixed-dataset priors and the results in Figures 4.4 - 4.8, ensured no bias towards the target (to-be-segmented) image anatomies, whilst incorporation of labels from both datasets aimed to generate more general priors that were representative of differing labelling protocols and expert raters. We did, however, also explore the cross-dataset performance; segmenting the CMA data using a spatial prior generated from the BPSA labels only, and vice versa. These single-dataset priors favour a specific labelling protocol and expert raters, so we expected differences in segmentation performance compared to using the mixed-dataset hippocampal priors. Dice coefficients were reduced using the single-dataset priors, but FMASH cross-dataset performance was still superior to FS on the CMA data and had the highest overall performance of all methods on the BPSA data. Furthermore, the observed difference in performance (between the mixed- and single-dataset priors) was comfortably within the range of inter-rater variability, differences in labelling protocol and registration error, although it does not rule-out bias of the mixed-dataset priors towards the two clinical datasets. The following chapter will look to explore this effect further. Nevertheless, for the aforementioned reasons, we recommend the use of a mixed-data hippocampal prior for the FMASH algorithm and we are confident that this will result in a more robust and accurate segmentation method.

Overall, this novel hippocampal segmentation method shows extremely good, consistent per-

formance on the two clinical datasets compared to the most widely-used alternatives, with more accurate estimation of the problematic hippocampus-amygdala border. This fully-automated approach also performs comparably well against the most recently-published methods, as described above. Primarily driven by subject-specific intensities in the MR image, this method is capable of capturing both subject- and disease-specific features, with results suggesting no inherent bias towards either dataset. We envisage accurate and robust hippocampal segmentation estimates using this fully-automated method on a range of clinical datasets, with differing image quality, disease-status and demographics.

5

FMASH Validation and Improvement: Accounting for Changes in Data and Labelling Protocol

In the previous chapter, we presented the development and initial evaluation of FMASH; a combined region-growing and atlas-based approach for automated hippocampal segmentation of structural MR data. The method was developed and evaluated on two clinical datasets (CMA and BPSA), but there were unresolved concerns over data and labelling bias. Here, these issues are addressed using a third, independent clinical dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI), together with the aforementioned CMA and BPSA data. Two main investigations are performed: (i) to ensure complete independence of the FMASH training and test data, and to better appreciate the accuracy and robustness of the FMASH algorithm for external users, the segmentation performance of FMASH (using the mixed CMA+BPSA data prior) is compared to that of FIRST and FS on a test set of 55 ADNI subjects (20 AD, 15 mild cognitively impaired (MCI), and 20 NC); and (ii) to investigate the dependency of atlas-based segmentation methods on data and labelling protocol, without the confounding effect of demographic differences, the performance of both FMASH and the simple atlas-based approach (thresholded prior) is assessed on an ADNI test set of 12 AD and 12 NC,

using a series of different priors (ADNI, CMA, Warped-CMA, Mixed) generated from the same demographic of 8 AD and 8 NC. Utilizing expert manual labels as a gold standard, we quantify Dice coefficients, false positive rates (FPR) and false negative rates (FNR). Results from Part (i) (FMASH ADNI Validation) highlight the current utility and great future potential of FMASH, as the method outperforms both FIRST and FS with significantly higher Dice coefficients (0.717±0.004), lower FPR (0.013±0.000), and the lowest left-side FNR (0.062±0.068). Interestingly, all three methods (FIRST, FS, FMASH) have a similar pattern and high propensity of FP findings, most likely due to differences between the CMA (training) and ADNI (test) data. From this initial investigation, however, we cannot disentangle the relative contribution (to the method training and resultant segmentation estimates) from each of the data variables (i.e., scanner, acquisition sequence, image resolution, subject demographics, manual segmentation protocol). To shed some light on this conundrum, we turn to the Data and Label Investigation in Part (ii): firstly, results show that FMASH performance with each of the hippocampal priors is far superior to the thresholded prior itself, strongly supporting the current finding that combined, multi-approach segmentation methods can vastly improve the accuracy of automated hippocampal delineation; secondly, use of cross-dataset priors identifies a strong dependency of atlas-based approaches on data and labelling protocol, with typically a 6-8% decrease in Dice coefficients. Consequently, a method is devised to account for the detrimental effects of using such cross-dataset priors — by non-linearly warping a cross-dataset prior onto a representative sample of labels from the target, reference dataset, major differences in data and labelling protocol can be overcome, and significantly improved segmentation performance achieved. In fact, the cross-dataset, Warped-CMA prior achieves higher Dice coefficients (0.764±0.007) than the same-dataset ADNI prior (0.740±0.014) on the independent ADNI test data. Differences in labelling protocol are, therefore, most likely re-

sponsible for the relatively diminished Dice coefficients of all methods on the ADNI test data, compared to those previously observed on the CMA and BPSA datasets. This finding strongly supports the use of the proposed label-warping method; facilitating both the "re-training" of current atlas-based approaches (for conformation to a uniform standard (i.e., consensus) protocol and more appropriate segmentation evaluation) as well as providing a means of generating this consensus labelling in the first instance.

5.1 Introduction

In Chapter 4, we presented the development and initial evaluation of FMASH; a combined region-growing and atlas-based approach for automated hippocampal segmentation on structural MR data. The method was developed and evaluated on two clinical datasets, CMA and BPSA: the first consisting of AD patients and normal controls (NC) from the Center for Morphometric Analysis (CMA), and the second containing both NC and bipolar (BP) patients from a collaborator in San Antonio (BPSA). As with the development of any image analysis tool, we believe that it is important to perform further validation on a dataset that is completely independent from the training data (i.e., different scanner, acquisition parameters) in order to fully assess the accuracy and robustness of the tool for external users. Additionally, for any model- and atlas-based approaches (i.e., methods that require training), it is useful to know how dependent these methods are on the specific training data. However, with so many factors affecting both the training and testing data (and the final segmentation estimates), it is often extremely difficult to disentangle the relative contribution from each. For example, in developing the FMASH approach (Chapter 4) we recommended using a mixed-dataset prior for implemention of FMASH — incorporating labels from different datasets to generate more general priors that were representative of differing demographics (disease status, age), labelling protocols and expert raters. Whilst we used a leave-one-out approach to generate the mixed- (CMA+BPSA)

dataset priors, with no bias towards the target (to-be-segmented) image anatomies, we could not rule-out other sources of bias (of these mixed-dataset priors and final FMASH segmentation estimates) towards the two (CMA and BPSA) clinical datasets (e.g., scanner, acquisition sequence, labelling protocol). In an attempt to address this issue, we explored the cross-dataset performance (segmenting the CMA data using a single-dataset prior generated from the BPSA labels only, and vice versa), where we typically observed a 3-5% decrease in FMASH Dice coefficients. With these single/cross-dataset priors, however, demographic differences between the two clinical datasets could have also contributed to the drop in segmentation performance.

This chapter is therefore dedicated to exploring the aforementioned issues of unknown data and labelling bias in atlas-based segmentation methods. As well as the CMA and BPSA datasets, we utilize a third, independent clinical dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.ucla.edu/). Firstly, to assess the performance of FMASH on a completely independent test cohort (i.e., with different demographics, labelling protocol, scan parameters, etc.), our previously-developed FMASH approach (using the mixed (CMA+BPSA) dataset prior) is validated on a test set of ADNI subjects. For this *FMASH ADNI validation*, FMASH segmentation performance is assessed by comparison to that of FIRST and FS. Secondly, to assess the cross-dataset performance of FMASH without the confounding effects of demographic differences, we implement the approach using a series of different hippocampal priors, each with the same demographic makeup of NC and AD patients. With the demographic variable removed, only differences in data (i.e., scan parameters) and labelling protocol remain. For this *data and label investigation*, the resultant FMASH segmentation performance (using each of the hippocampal priors) is compared to that of the simple, atlas-based approach (i.e., the thresholded prior). The work herein can therefore be separated into two parts: (i) *FMASH ADNI validation*; and (ii) *data and label investigation*.

5.2 Materials and methods

5.2.1 Clinical MR data

All data was described in detail in Section 2.6. For the *FMASH ADNI validation*, the complete ADNI dataset (consisting of T1-weighted MR data for 20 normal controls (NC), 15 mild cognitively impaired (MCI) and 20 AD patients, together with their corresponding expert manual labels) was utilized as the test set, evaluating the segmentation performance of FMASH (and other methods) from the hippocampal estimates of these subjects. Expert manual labels from the CMA and BPSA cohorts (25 NC, 8 AD and 16 BP patients in total) were used to generate the FMASH mixed-dataset prior (i.e., used as training data), in keeping with Chapter 4.

For the *data and label investigation* in Part (ii) of this chapter, several alternative hippocampal priors were generated using labels from differing datasets and labelling protocols (ADNI and CMA) but the same demographic of 8 NC and 8 AD patients (Section 5.2.3). The test data was an independent collection of 24 ADNI subjects (12 NC, 12 AD).

5.2.2 Software and image preprocessing

As per Chapter 4, tools from the FMRIB Software Library (FSL [Smith et al., 2004; Woolrich et al., 2009], www.fmrib.ox.ac.uk/fsl) were used to generate the hippocampal spatial priors (Section 5.2.3) and standard-space performance maps (Section 5.2.5), with the remainder of the FMASH algorithm implemented in MATLAB (www.mathworks.com).

The MR images were initially cropped to remove excess neck, and BET [Smith, 2002] was used to remove most of the non-brain structure. The brain-extracted data was used for affine registration tasks, whilst the whole-brain data was used for non-linear transformations, as described below.

5.2.3 Generation of hippocampal spatial priors

All T1-weighted MR data (training and testing) was aligned to the standard MNI152 space via a multi-step registration scheme: an initial three-step affine registration with FLIRT [Jenkinson and Smith, 2001], followed by a multi-resolution non-linear warp with FNIRT [Andersson et al., 2008]. The training labels were propagated to the standard space using these computed warps. Additionally, the inverted (standard-to-subject) non-linear warp was computed for each test subject.

For all affine registrations, FLIRT was executed with 12 DOF and maximization of the normalized mutual information similarity function (with 256 histogram bins). For all non-linear registrations (of the T1-weighted MR data) with FNIRT, a control point spacing of 10mm was specified, with a sub-sampling scheme of 4,4,2,2,1,1, a bending energy model for regularization of the warp-field, and weightings for the membrane energy regularization of 300,150,100,50,40,30.

For the *FMASH ADNI validation*, all spatially-normalized CMA and BPSA manual labels were fused (summing trilinearly-interpolated values) and averaged to generate a standard-space, mixed-dataset hippocampal prior, similar to that utilized in Chapter 4. The inverted non-linear warp of each ADNI test subject subsequently transformed this mixed- (CMA+BPSA) dataset hippocampal prior to the corresponding subject (native) space for use in the FMASH algorithm.

For the *data and label investigation*, similar fusion of spatially-normalized labels generated several hippocampal priors, each with the same demographic makeup of 8 NC and 8 AD patients: a separate prior each from ADNI and CMA, and an ADNI-warped CMA prior. The ADNI-warped CMA prior was created by non-linearly warping the CMA prior to the ADNI prior with FNIRT, using the identity matrix for initialization, and no intensity mapping. Additionally, a combined, mixed- (ADNI+CMA) dataset prior was generated by fusion of the CMA and ADNI priors. In total, this constituted four hippocampal spatial priors for the *data and label investigation*: *ADNI*, *CMA*, *Warped-CMA*, and *Mixed* (ADNI+CMA).

5.2.4 Segmentation methods

For the *FMASH ADNI validation* in Part (i) of this chapter, FMASH segmentation estimates were compared to FIRST and FS using the performance measures detailed below. For this validation of the previously-developed FMASH approach, it is worth remembering that the original mixed- (CMA+BPSA) dataset prior was used, in keeping with Chapter 4.

For the later *data and label investigation*, investigating the dependency of atlas-based approaches on data and labelling protocol, eight different segmentation methods were evaluated: firstly, the four aforementioned hippocampal priors (*ADNI*, *CMA*, *Warped-CMA*, *Mixed*) were thresholded (at a probability threshold of 0.5) to generate four simple, atlas-based segmentation estimates for each ADNI test subject, and secondly, FMASH was performed using each of the (non-thresholded) priors to generate a further four segmentation estimates for each subject.

5.2.5 Performances measures

Here, we assessed the same combination of performance measures as the previous chapters: Dice coefficient, false positive rate (FPR) and false negative rate (FNR). Once again, we considered a ROI of volume 91125 mm^3 (equivalent to a 45x45x45 mm cuboid) to maintain a constant FOV in our calculations of the TN set (and FPR). Standard-space maps were generated as per Chapter 4, to assess the spatial distribution of FP and FN voxels for each segmentation method on the ADNI test data.

In the *data and label investigation*, the ADNI prior was generated using a (training) ADNI dataset that was independent from the ADNI test subjects, but both sets of T1-weighted MR images are manually segmented using the same ADNI labelling protocol, so we expect the ADNI prior to provide the most accurate segmentation estimates (i.e., segmentation estimates that most closely resemble the gold standard ADNI labels). We therefore also refer to the ADNI prior as the 'reference' prior, and the remaining three priors are collectively referred to as 'non-

reference' priors.

Prior difference maps were generated in the *data and label investigation* to assess the difference between the 'reference' *ADNI* prior and each of the other ('non-reference') hippocampal atlas priors (*CMA*, *Warped-CMA*, and *Mixed*): in standard MNI152 space, each non-reference prior was initially subtracted from the reference *ADNI* prior, retaining only positive voxel values, to visualize regions where the prior probability of observing hippocampus was greatest in the *ADNI* atlas; whilst the opposite computation was performed (subtracting the reference ADNI prior from all non-reference atlases and retaining only positive voxel values) to view regions where the prior probability of observing hippocampus was larger than the reference ADNI atlas.

5.2.6 Statistical analysis

All statistical analysis was performed using the SPSS statistics software (Version 19.0). For the larger ADNI dataset utilized in the *FMASH ADNI validation*, a chi-square test for independent samples was initially performed to evaluate groupwise gender differences, and group comparisons for age, MMSE and CDR were assessed using three-way, between-subjects analysis of variance (ANOVA) designs. For both the *FMASH ADNI validation* and the *data and label investigation*, the left- and right-side hippocampal segmentation estimates and corresponding performance measures (Dice, FPR, FNR) were computed separately for each method on the respective ADNI test subjects. Similar to the statistical analysis performed in Chapter 4, we initially tested for asymmetric and groupwise differences in performance for each method using mixed ANOVA designs, with side as the two-way, within-subjects repeated measure, and clinical group as the between-subjects factor. Paired-sample t-tests with Bonferroni correction for multiple comparisons were used to identify differences in side and group, with F_d, F_{fp}, F_{fn} and p_d, p_{fp}, p_{fn} denoting the F-statistic and corrected p-values for Dice, FPR and FNR, respectively.

For each side, we subsequently tested for overall difference across the segmentation meth-

ods and clinical group-by-method interactions using a mixed ANOVA design, with method as the within-subjects factor and clinical group as the between-subjects factor. For the *FMASH ADNI validation*, this involved a 3*3 (method*group) mixed ANOVA design for each performance measure. In the *data and label investigation*, FMASH performance (with each of the aforementioned priors) was initially compared to that of the simple thresholded prior, using a 2*2 (method*group) mixed ANOVA design for each performance measure. Following this, all of the FMASH methods were compared using 4*2 (method*group) mixed ANOVA designs.

Where Mauchly's test indicated that the assumption of sphericity for method had been violated ($p<0.05$), degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. Paired-sample t-tests with Bonferroni correction for multiple comparisons were used to identify between-method differences; Lp_d, Rp_d, Lp_{fp}, Rp_{fp}, Lp_{fn}, Rp_{fn} denote left (L) and right (R) corrected p-values for Dice coefficient, FPR, and FNR, respectively.

5.3 Results

5.3.1 FMASH ADNI validation

There was no statistical difference in gender ($\chi^2(2, N=55)=0.117$, $p=0.943$) or age ($F(2,52)=0.458$, $p=0.635$) between the three clinical groups (NC, MCI, and AD). However, the three-way ANOVA designs revealed a significant main effect of group for MMSE ($F(2,52)=258.3$, $p<0.0005$) and CDR scores ($F(2,52)=37.99$, $p<0.0005$). For the MMSE scores, all pairwise comparisons were significant (NC>MCI>AD: $p<0.0005$ for all), whilst similar comparisons for CDR scores showed significant differences between patients and controls only (i.e., significant differences for MCI versus (v.) NC and AD v. NC: both $p<0.0005$; but not for MCI v. AD: $p=1.000$). Overall, these results confirm the expected increase in dementia severity for MCI and AD patients compared to controls.

Figure 5.1 presents box plots of the segmentation performance for FIRST, FS and FMASH on

the ADNI test data. For this validation of the previously-developed FMASH approach, it is worth remembering that the original mixed- (CMA+BPSA) dataset prior was used, in keeping with Chapter 4. The initial 2*3 (side*group) ANOVA designs revealed no side-by-group interactions for each method, but about half of the analyses revealed a significant main effect of side (FIRST: F_{fp}(1,52)= 8.49, p_{fp}=0.005; FS: F_d(1,52)= 18.01, p_d <0.0005, F_{fp}(1,52)= 20.16, p_{fp} <0.0005; FMASH: F_{fp}(1,52)= 22.06, p_{fp} <0.0005; F_{fn}(1,52)= 5.85, p_{fn}= 0.019). Pairwise comparisons showed that there was a left<right asymmetry for FS Dice coefficients (p_d <0.0005), the FPR of all methods displayed a left>right asymmetry (FIRST: p_{fp}= 0.003; FS: p_{fp} <0.0005; FMASH: p_{fp} <0.0005), and there was a correspondingly opposite left<right asymmetry for FMASH FNR (p_{fn}= 0.010). In addition, there was generally a significant main effect of group, with the following groupwise trends observed for each method

- Dice: AD < MCI < NC;
- FPR: AD < MCI < NC;
- FNR: AD > MCI > NC.

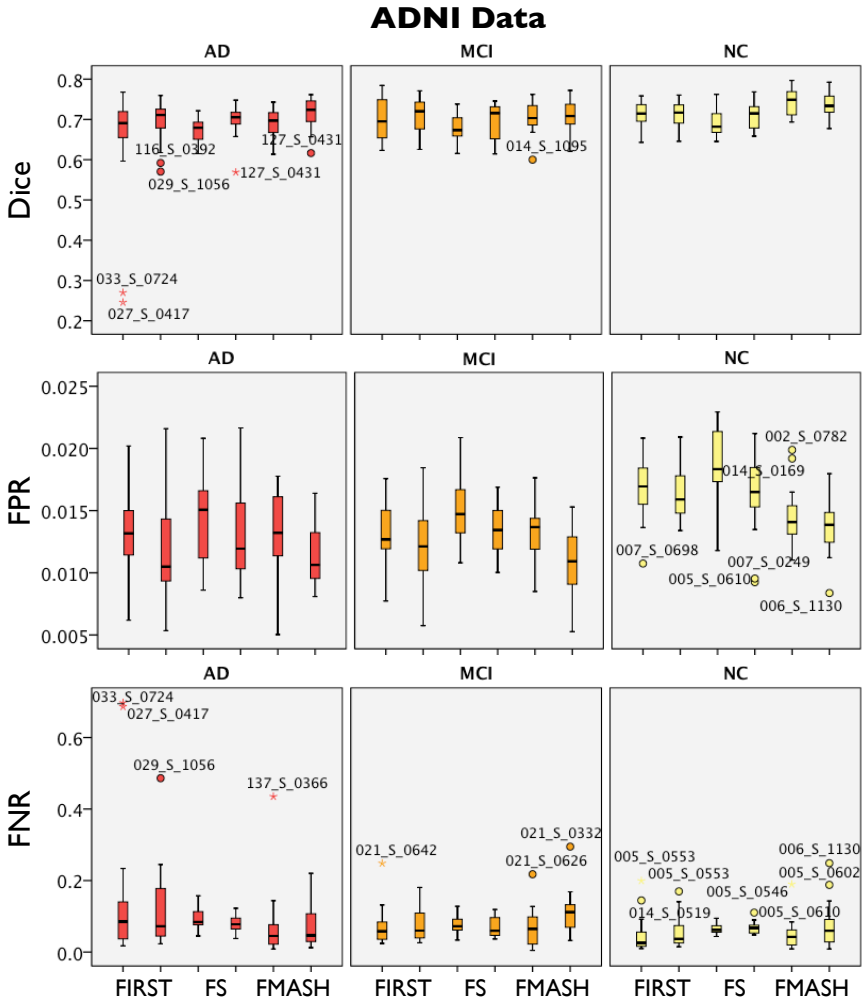
The only exceptions were the FS Dice coefficients and the FMASH FNR, where no main effect of group was observed.

Expectedly, the subsequent 3*3 (method*group) mixed ANOVA designs identified the same groupwise trends for each performance measure, which generally gave rise to a significant main effect of group (Lp_d= 0.012, Lp_{fp} <0.0005, Rp_{fp} <0.0005, Lp_{fn}= 0.008, Rp_{fn}= 0.020). For both Dice and FPR, highly-significant main effects of method were also observed (Lp_d, Rp_d, Lp_{fp}, Rp_{fp} all <0.0005), with subsequent pairwise comparisons ordering the methods as follows

- FMASH > FIRST > FS.

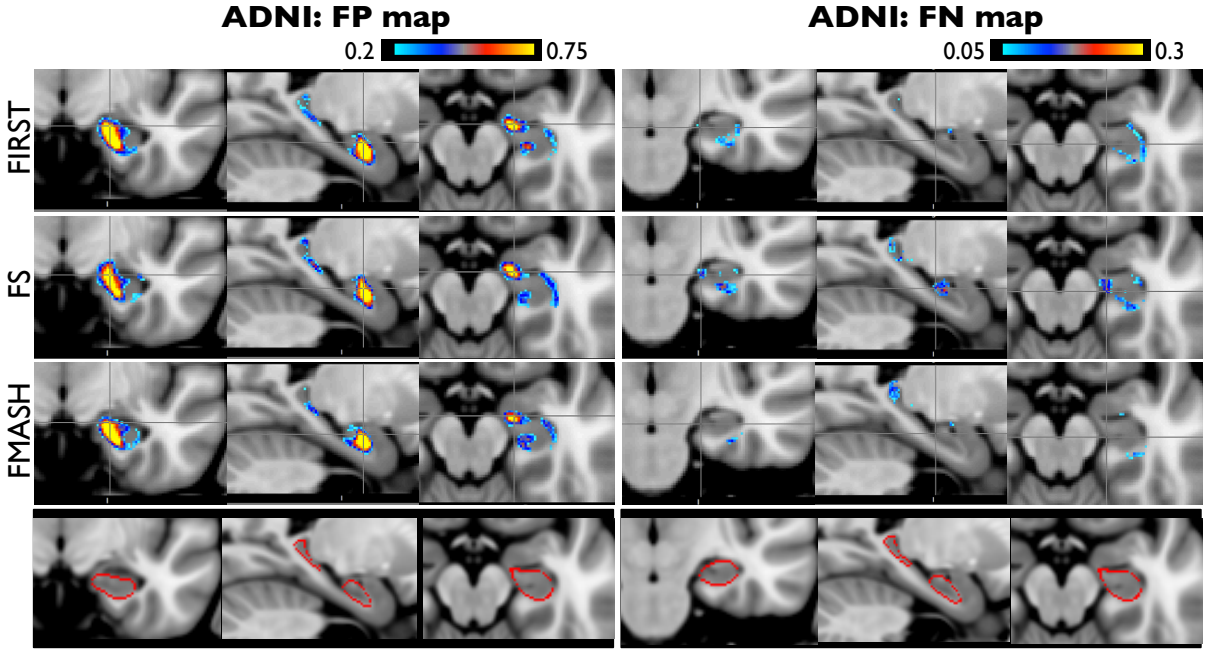
In accord with the results presented previously for the BPSA data (Chapter 3), FIRST performed well for the majority of ADNI subjects, but its overall segmentation performance on the

Figure 5.1 *FMASH ADNI validation: comparison of method performance.* Box plots comparing FMASH segmentation performance to that of FIRST and FS on the ADNI data, showing Dice coefficients (top), FPR (middle) and FNR (bottom) for AD (left plot), MCI (middle plot) and NC (right plot). Within each plot, each duplet presents the left (left) and right (right) hippocampal measures for FIRST (left duplet), FS (middle duplet) and FMASH (right duplet). As per previous chapters, boxes have lines at the lower quartile, median, and upper quartile values, with whiskers extending to 1.5 times the interquartile range. Outliers and extreme outliers are indicated by a circle and an asterisk, respectively, in addition to their corresponding subject ID.



ADNI data was hampered by that of a few subjects (Figure 5.1) — where poor initial alignment with the standard MNI152 space resulted in poor segmentation estimates. Interestingly, FMASH achieved the highest Dice coeffcients (left: 0.713±0.042; right: 0.723±0.036), the lowest FPR (left: 0.014±0.003; right: 0.012±0.003), and the lowest left-side FNR of all methods (0.062±0.068), such that it was the best performing method overall on the ADNI data. Pairwise comparisons revealed that FMASH outperformed both FIRST and FS with significantly higher Dice coefficients (FIRST: Lp_d= 0.039, Rp_d= 0.021; FS: Lp_d, Rp_d both <0.0005) and lower FPR (FIRST: Rp_d= 0.027; FS: Lp_d, Rp_d both <0.0005).

Figure 5.2 *FMASH ADNI validation: comparison of false positive (FP) and false negative (FN) maps.* Left and right panels show multi-view images of the standard-space FP and FN maps, respectively, for FIRST (first row), FS (second row) and FMASH (third row) on the ADNI test data. A voxel-wise threshold of 20% FP finding across the dataset is applied to all FP maps, whilst the FN maps are thresholded at 5%. To aid visual assessment, the hippocampal boundary defined by FIRST on the MNI152 standard image is shown (fourth row) and the cursor position is the same for all images in each panel.



The standard-space maps in Figure 5.2 are particularly interesting because they show a similar pattern of FP findings (left panel) for all three methods, with pronounced overestimation of the hippocampal head compared to our gold standard ADNI labels. The FN maps (right panel) are

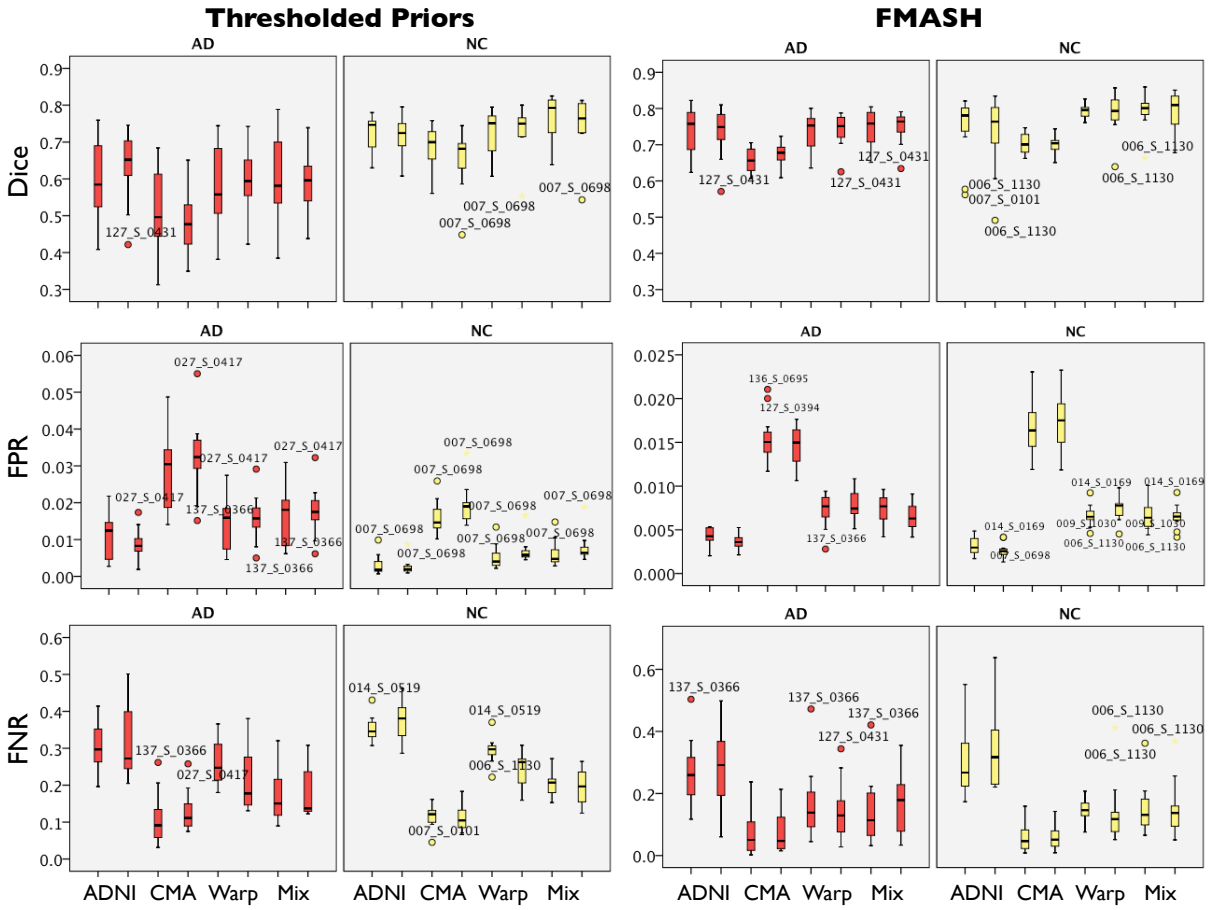
much less similar across methods: whilst FIRST (first row) displays a low frequency of FN findings at the lateral hippocampal boundary, FS (second row) has a tendency to underestimate posterior aspects of the hippocampus head, and FMASH (third row) underestimates the tip of the tail.

5.3.2 Data and label investigation

Box plots showing the relative performance of the four thresholded priors and the four FMASH methods (using each of the hippocampal priors: *ADNI*, *CMA*, *Warped-CMA*, *Mixed*) are given in Figure 5.3 (left and right panel, respectively), whilst the prior difference maps and the FMASH FP and FN maps are displayed in Figures 5.4 and 5.5, respectively.

Interestingly, whilst the initial 2*2 (side*group) ANOVA designs generally found no main effects of side for the Dice coefficients and FNR of each method, all but one of the analyses for FPR revealed either a significant main effect of side (thresholded *ADNI* prior: F_{fp}(1,22)= 5.05, p_{fp}= 0.035; thresholed *CMA* prior: F_{fp}(1,22)= 17.37, $p_{fp} < 0.0005$; thresholded *Warped-CMA* prior: F_{fp}(1,22)= 4.76, p_{fp}= 0.040; FMASH with *ADNI* prior: F_{fp}(1,22)= 11.98, p_{fp}= 0.002; FMASH with *Warped-CMA* prior: F_{fp}(1,22)= 5.88, p_{fp}= 0.024) or a trend towards asymmetry (thresholded *Mixed* prior: F_{fp}(1,22)= 3.92, p_{fp}= 0.060; FMASH with *Mixed* prior: F_{fp}(1,22)= 3.13, p_{fp}= 0.091). Exploring these findings further, the pairwise comparisons revealed that the *ADNI* prior gave a significant left>right asymmetry in FPR (p_{fp}= 0.018), whilst the *CMA* prior had the opposite asymmetry in FPR (left<right, p_{fp} <0.0005), such that these asymmetric tendencies were somewhat counter-balanced in the *Warped-CMA* and the *Mixed* hippocampal priors. Additionally, a main effect of side was observed for the thresholded *CMA* prior Dice coefficients (F_d(1,22)= 6.96, p_{fp}= 0.015) and the thresholded *Warped-CMA* prior FNR (F_{fn}(1,22)= 11.12, p_{fn}= 0.003), with pairwise comparisons identifying a left>right asymmetry in each case (p_d <0.0005 and p_{fn}= 0.002, respectively). However, no side-by-group interactions were observed.

Figure 5.3 *Data and label investigation: dependency of atlas-based methods on hippocampal priors.* Box plots comparing the thresholded prior segmentation performance (left panel) with that of FMASH (right panel) on the ADNI test data, showing (within each panel) Dice coefficients (top), FPR (middle) and FNR (bottom) for AD patients (left plot) and NC (right plot). From left to right in each plot, the duplets give the left (left) and right (right) hippocampal measures using the ADNI- (first duplet), CMA- (second duplet), Warped-CMA (third duplet) and Mixed (fourth duplet) hippocampal prior. Please see Figure 5.1 for details of the boxes.



For each prior, the 2*2 (method*group) mixed ANOVA designs showed that FMASH performance was superior to the thresholded prior for atlas-based segmentation, with significant main effects of method found for all Dice (*ADNI* prior: Lp_d, Lp_d both 0.002; *CMA* prior: Lp_d, Lp_d both <0.0005; *Warped-CMA* prior: Lp_d, Lp_d both <0.0005; *Mixed* prior: Lp_d, Lp_d both <0.0005), all FPR (*ADNI* prior: Lp_{fp}= 0.001, Lp_{fp} <0.0005; *CMA* prior: Lp_{fp}= 0.001, Lp_{fp} <0.0005; *Warped-CMA* prior: Lp_{fp}= 0.011, Lp_{fp}= 0.001; *Mixed* prior: Lp_{fp}= 0.001, Lp_{fp} <0.0005), and most FNR (*CMA* prior: Lp_{fn}, Lp_{fn} both <0.0005; *Warped-CMA* prior: Lp_{fn}, Lp_{fn} both <0.0005; *Mixed* prior: Lp_{fn}= 0.011).

Subsequent comparison of the four FMASH approaches, using the 4*2 (method*group) ANOVA designs, showed some interesting results

- There was a highly-significant main effect of method (i.e., hippocampal prior) for each performance measure (Lp_d, Rp_d, Lp_{fp}, Rp_{fp}, Lp_{fn}, Rp_{fn} all <0.0005), but generally no method-by-group interactions and no group effects, with the exception of a left-side group effect for Dice (Dice(NC)>Dice(AD): Lp_d= 0.016), a method-by-group interaction for the right-side FNR (Rp_{fn}= 0.031, with FNR(AD)>FNR(NC) for all except the *ADNI* prior), and a method-by-group interaction for the right-side FPR (Rp_{fp}= 0.001, with FPR(AD)<FPR(NC) for the *CMA* and *Mixed* priors only).

- The *ADNI* prior had expectedly higher Dice (left, right: 0.744±0.075, 0.735±0.084) and lower FPR (left, right: 0.004±0.001, 0.003±0.001) compared to the (cross-dataset) *CMA* prior on the ADNI test data (Lp_d= 0.0005, Rp_d= 0.028, Lp_{fp} and Rp_{fp} both <0.0005), but a surprisingly higher FNR (left, right: 0.286±0.119, 0.314±0.130; Lp_{fn} and Rp_{fn} both <0.0005).

- The *Warped-CMA* prior was able to recover most of the "ADNI-labelling characteristics" mentioned above, but FPR and FNR were still significantly different to the *ADNI* prior (FPR(*Warped-CMA* prior) > FPR(*ADNI* prior): Lp_{fp} and Rp_{fp} both <0.0005;

FNR(*Warped-CMA* prior) < FNR(*ADNI* prior): Lp_{fn} and Rp_{fn} both <0.0005).

- The *Mixed* prior achieved a similar overall performance to the *Warped-CMA* prior.
- Surprisingly, both the *Warped-CMA* prior and the *Mixed* prior achieved higher Dice coefficients than the *ADNI* prior itself (*Warped-CMA* prior (left, right): 0.765±0.049, 0.764±0.055; *Mixed* prior (left, right): 0.769±0.053, 0.771±0.053), but with significance achieved for the *Mixed* prior on the right side only (Rp_d= 0.009).

Figure 5.4 *Data and label investigation: FMASH false positive (FP) maps.* Multi-view images of the prior difference maps (left panel) and the FMASH FP maps (right panel) for the left hippocampus (LHipp). Here, the difference maps correspond to regions where the specified prior probability is greater than that of the ADNI prior.

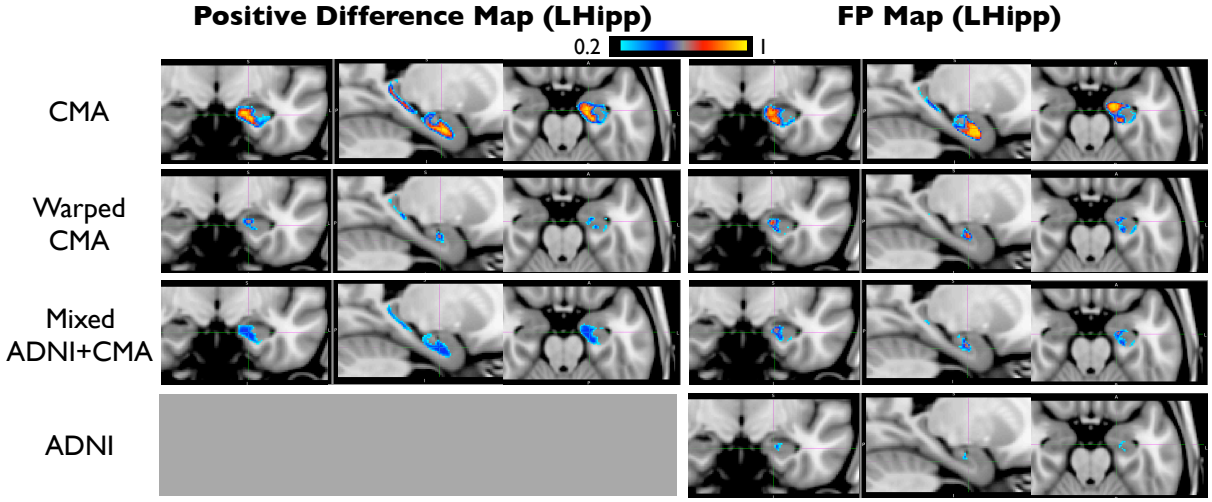
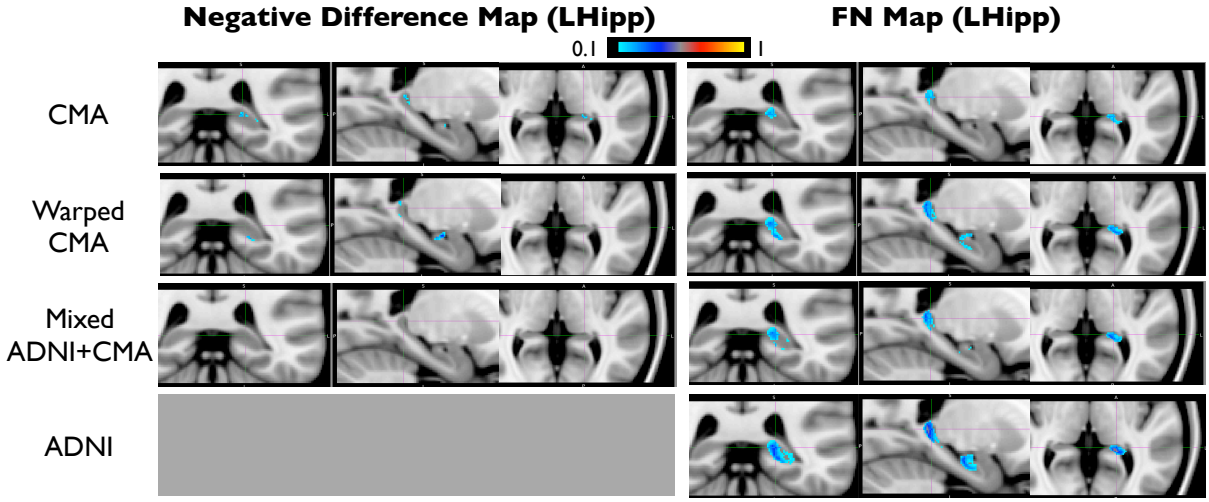


Figure 5.4 shows that the pattern and frequency of FP findings for each of the four FMASH methods (right panel) closely resembles the prior difference maps (left panel) which depict regions where the current prior probability (i.e., *CMA*, *Warped-CMA*, and *Mixed* prior probability) is greater than that of the *ADNI* prior. Compared to the *ADNI* prior, the *CMA* prior (top row) provides a more generous ventral and medial hippocampal boundary, and extends the hippocampal head much further towards the neighbouring amygdala (i.e., dorsoanterior direction). These differences between the *ADNI* and *CMA* priors are largely overcome in the *Warped-CMA* prior (second row), with differences now limited to a small region at the dorsoanterior boundary

of the hippocampal head. Expectedly, the regions previously described as being more generously delineated in the *CMA* prior (i.e., ventral and medial boundaries, hippocampal head) are also observed in the prior difference map of the *Mixed* (ADNI+CMA) prior (third row), albeit with a diluted, reduced frequency of occurence. Surprisingly, whilst the difference maps for the *Warped-CMA* prior and the *Mixed* prior are quite different, the FP maps for the final FMASH segmentations are remarkably similar.

Figure 5.5 *Data and label investigation: FMASH false negative (FN) maps.* Multi-view images of the prior difference maps (left panel) and the FMASH FN maps (right panel) for the left hippocampus (LHipp). Here, the difference maps correspond to regions where the current prior probability is less than that of the ADNI prior.



In contrast to the aforementioned difference maps and FP findings, the difference maps in Figure 5.5 (left panel) show that there are extremely few regions where the prior probability (of the *CMA*, *Warped-CMA*, and *Mixed* prior) is less than that of the *ADNI* prior. Despite this, the final FMASH segmentations underestimate the tip of the hippocampal tail (for all priors) and the posterior boundary of the hippocampal head (for the *Warped-CMA* and *ADNI* priors only).

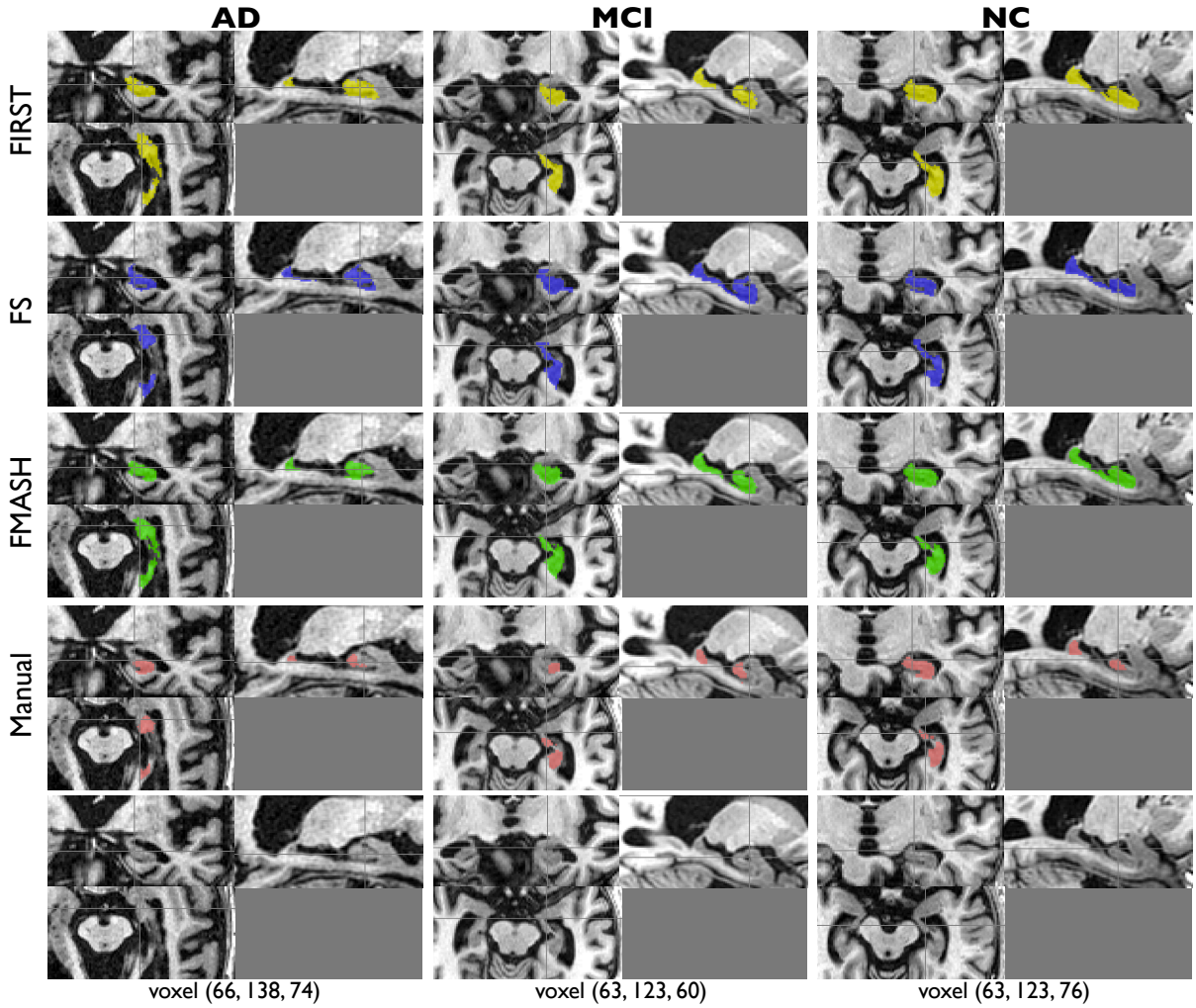
5.4 Discussion and conclusions

In Part (i) of this chapter (*FMASH ADNI validation*) we validated our previously-developed FMASH approach (using the mixed CMA+BPSA data prior) on an independent dataset of 55 ADNI subjects (20 AD, 15 MCI, 20 NC), comparing segmentation performance to that of FIRST and FS. Interestingly, FMASH outperformed the other methods with significantly higher Dice coefficients and lower FPR, coupled with the lowest left-side FNR. We believe that these results showcase the current utility of FMASH and suggest that it has great potential compared to other commonly-used approaches. The FN maps also showed that there was a general infrequent and mixed distribution of FN findings across the methods. However, the extremely similar pattern and high propensity of FP findings for all methods was rather intriguing. Considered together with the relatively low Dice and high FPR compared to that obtained on other datasets (e.g., CMA and BPSA), these results suggest pronounced and systematic differences between the gold standard, manually defined ADNI labels and the hippocampal boundary defined by the three automated segmentation methods (FIRST, FS, and FMASH). If we now take a closer look at both the manual ADNI labels and the automated hippocampal segmentations for some representative subjects (one from each group: AD, MCI, NC), we gain a better understanding of the differences between the defined hippocampal boundaries (Figure 5.6). Compared to the underlying neuroanatomy (as viewed on the T1-weighted MR image), it appears that the manual ADNI labels provide a very conservative definition of the anterior and medial aspects of the hippocampal head, whilst all of the automated methods are much more generous in this region.

The big question here is whether the gold standard ADNI labels or the automatically-defined hippocampal boundaries are closer to the ground truth of the underlying neuroanatomy? Of course, by definition, the gold standard labels are the reference by which we compare other methods, so for the purpose of segmentation evaluation and method validation we must accept these as our best representation of the ground truth. One consideration, which is quite controversial and rarely mentioned when validating segmentation methods, is that the gold standard labels

are in fact at odds with the underlying neuroanatomy, such that the automated segmentation estimates are better identifying the true hippocampal boundary, but the computed performance measures (and false error maps) do not reflect this.

Figure 5.6 *FMASH ADNI validation: comparison of segmentation estimates.* Multi-view images of the automated segmentation estimates (FIRST: first row; FS: second row; FMASH: third row) and the manual ADNI labels (fourth row) for the left hippocampus of three representative subjects (AD: left; MCI: middle; normal control (NC): right). For a clear view of the underlying neuroanatomy, the fifth row shows the T1-weighted MR images with no overlay.



This is not to say that the ADNI labels shown here are in any way inferior to the hippocampal estimates provided by the three automated methods, as the same could be said for any other test data and corresponding manual labels, but it raises the interesting point that the *perceived*

performance of evaluated segmentation methods is completely dependent on the gold standard to which the resultant segmentation estimates are compared. If we now produced our own set of manual labels for the ADNI test subjects, which defined a more generous hippocampal head, for example, the FPR of all three automated methods would be vastly reduced, and overall segmentation performance would be increased, but the automated segmentation estimates themselves would not have changed at all! With such a plethora of different protocols for delineation of the hippocampus [Konrad et al., 2009], differences in opinion as to what does and does not constitute the hippocampus is a common and problematic occurence. In this case, we know that FMASH uses a subset of CMA data as part of its training (i.e., during initial method development, and in the mixed data prior), whilst both FIRST and FS are entirely trained from CMA data, so differences between the CMA and ADNI data are most likely responsible for the aforementioned effects. However, since the CMA and ADNI datasets differ in their initial acquisition (i.e., scanner, acquisition parameters), image resolution, subject demographics, and final manual segmentation protocol, we cannot say for certain what the relative contribution from each of these differences is. Consequently, for the *data and label investigation* in Part (ii), we removed the demographic variable to test the dependency of FMASH and a simple atlas-based approach (thresholded prior) on data and labelling protocol alone.

FMASH performance with each of the hippocampal priors was far superior to the thresholded prior itself, strongly supporting the current finding that combined, multi-approach segmentation methods can vastly improve the accuracy of hippocampal delineation. With regard to the FMASH results, the *ADNI* prior gave expectedly higher Dice and lower FPR compared to the cross-dataset *CMA* prior on the ADNI test data, but a surprisingly higher FNR. Visually assessing the expert manual labels (and the prior difference maps), we found that the *ADNI* prior generally had a more conservative labelling than the *CMA* prior, which likely explains the relatively high FPR (and correspondingly low FNR) for the latter. Quite unexpectedly, however, both the *Warped-CMA* prior and the *Mixed* prior gave higher FMASH Dice coefficients than

the *ADNI* prior itself. As such, it appears that both the *Warped-CMA* prior and the *Mixed* (ADNI+CMA) prior combine the most desirable features from ADNI (i.e., low FPR) and CMA labels (i.e., low FNR) to achieve even greater overall segmentation accuracy.

The pattern and frequency of FMASH FP findings closely followed that of the high-probability zones in the corresponding prior difference maps (Figure 5.4), suggesting two things: (i) the final FMASH segmentation estimates are strongly dependent on the high-probability zones of the input prior; and (ii) the variable that predominantly controls FP errors in the FMASH segmentation estimates is differences in labelling protocol (between the input prior and the gold standard to which the hippocampal output volumes are assessed).

With regard to the first point (i) raised here, it is no surprise that the FMASH segmentation estimates are strongly dependent on the high-probabilty zones of the hippocampal prior — since these high-probability zones, by definition, correspond to regions where hippocampal voxels are most likely to be found, we have developed the method to use this information as such. Reassuringly, the observation that the FMASH FP maps were very similar for the *Warped-CMA* and *Mixed* priors, despite the clear discrepancies in their corresponding prior difference maps, suggests that the final FMASH segmentations are not strongly influenced by erroneous low-probability zones (in the *Mixed* prior), and the image-based intensity features dominate in these regions, which is good news. FMASH did, however, have a tendency to increase the propensity of FN findings relative to the corresponding prior difference maps (Figure 5.5), and this effect was slightly enhanced by the use of more conservative priors (i.e., *ADNI* and the *Warped-CMA* prior). It appears that there is an inherent tendency of FMASH to underestimate the tip of the hippocampal tail — the main limitation of the method, which should be addressed in future work.

What is more surprising here, and somewhat concerning, is the marked difference between the *CMA* and *ADNI* priors: the *CMA* prior "grossly overestimates" the hippocampus relative to the *ADNI* prior (and the reference labelling protocol). These stark differences in labelling protocol

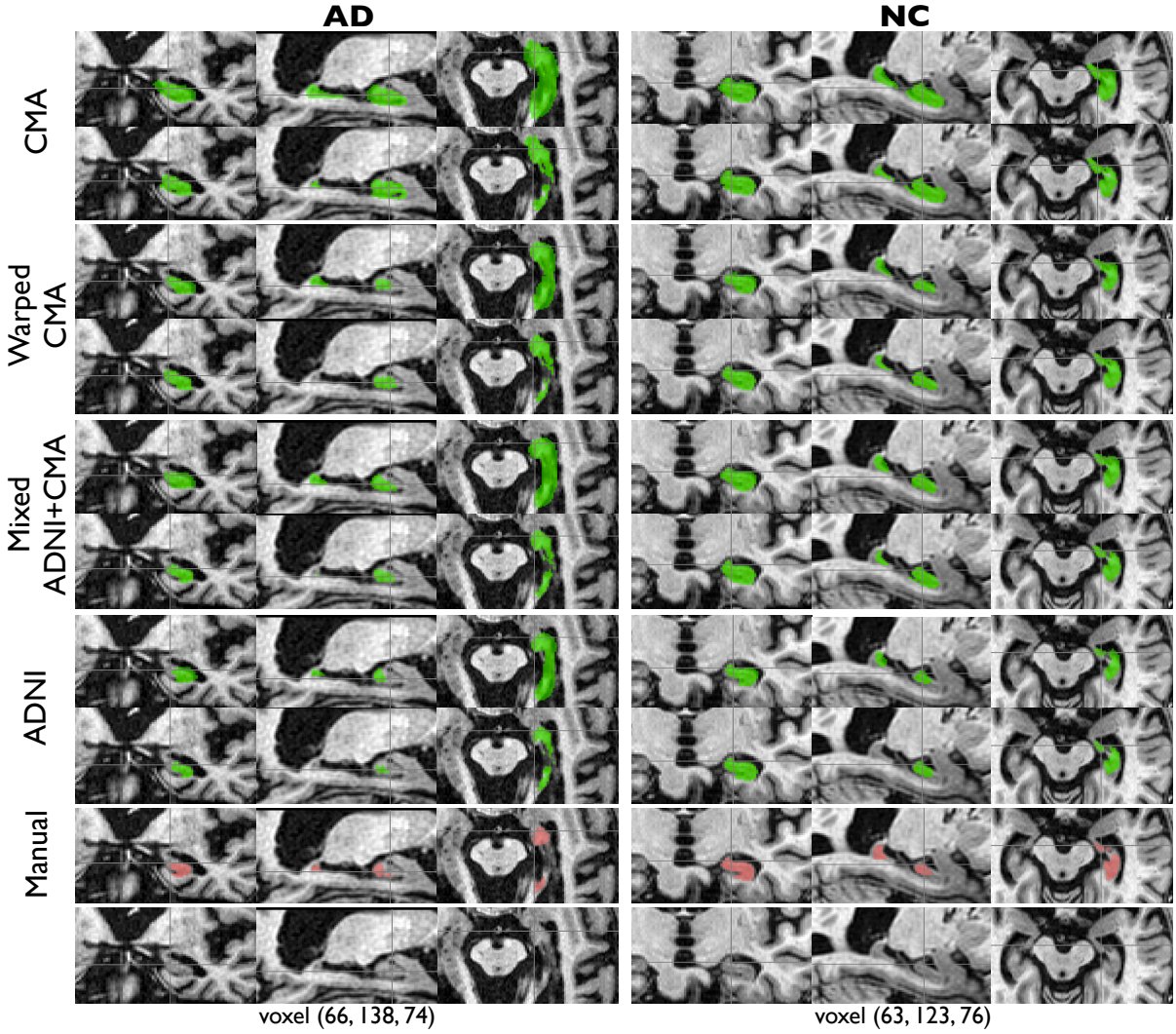
give rise to "erroneous" high-probability zones in the *CMA* atlas prior (with respect to the reference *ADNI* prior), which as we have just discussed, is translated into the final FMASH segmentation estimates. However, if we account for differences in labelling protocol, by non-linearly warping a cross-dataset prior onto a representative sample of labels from the reference (i.e., gold standard) protocol, such as demonstrated here with the *Warped-CMA* prior, the labelling bias can be largely overcome and FP errors significantly reduced. This is a novel approach to account for differences in data and labelling procotol, facilitating the "re-training" of atlas-based approaches so that they conform to a standard protocol, without the tedious and laborious task of manually labelling all of the training data again. As such, this technique would provide a means of evaluating all previous and current segmentation methods against a uniform standard.

We mentioned previously that differences between the *CMA* and *ADNI* data were most likely responsible for the high frequency and similar pattern of FP findings observed for FIRST, FS and FMASH on the ADNI test data, but we could not disentangle the root cause of these FP errors. By excluding demographic differences from the FMASH priors, we can now clearly see what a profound effect differences in data and labelling protocol have on method training, final segmentation estimates, and computed segmentation performance. To more clearly visualize these differences on an individual subject level, we now take a look at the thresholded priors and FMASH segmentation estimates (Figure 5.7) for the same representative AD patient and NC as Figure 5.6. Again, we see that the manual ADNI labels (ninth row) and the automated segmentations from the *ADNI* prior (rows 7-8) provide a much more conservative definition of the anterior and medial aspects of the hippocampal head compared to the other thresholded priors and FMASH segmentation estimates. The biggest discrepancy between the final segmentation estimates is observed for the *CMA* and *ADNI* priors, with these differences largely reduced in the *Warped-CMA* and *Mixed* priors.

Once again, we raise the question of whether the CMA or the ADNI labelling protocol is closer to the ground truth of the underlying neuroanatomy? At present, we believe that there is no

easy answer to this, because a universal standard for hippocampus delineation has not yet been agreed upon (although an initiative is underway at www.hippocampal-protocol.net). As demonstrated in a recent review by Konrad et al. [2009], there are at least 71 different anatomical procotols for manual delineation of the hippocampus on T1-weighted MR images, with qualitatively large variations between them. We must therefore select a gold standard against which to compare other segmentation methods, and if we maintain this standard across all studies, we can make fair comparisons and evaluations of relative segmentation performance, even if this gold standard is not the best estimate of the underlying neuroanatomy. However, if consistent and reproducible errors are found with an array of evaluated segmentation methods, perhaps it is time to take a closer look at the reference labels and consider whether or not they are the best representation of the underlying ground truth neuroanatomy and local image features. Perhaps it is time to establish a universal standard for hippocampus delineation — a consensus labelling protocol. Again, the proposed label-warping approach may not only provide a means of evaluating all previous and current segmentation methods against a uniform standard, but may also provide a means of generating this consensus labelling in the first instance. This label-warping approach therefore makes great inroads towards reliable assessment and comparison of atlas-based segmentation methods, without the confounding effects of differences in data and labelling protocol.

Figure 5.7 *Data and label investigation: thresholded priors and FMASH segmentations.* From top to bottom, each pair of rows (1-2, 3-4, 5-6, 7-8) shows multi-view images of the thresholded prior (upper row of pair) and FMASH segmentation estimates (lower row of pair) for each of the hippocampal priors (CMA, Warped-CMA, Mixed, and ADNI, respectively) compared to the manual ADNI labels (ninth row) for the left hippocampus of two representative subjects (AD: left; normal control (NC): right). For a clear view of the underlying neuroanatomy, the tenth row shows the T1-weighted MR images with no overlay.



6

Clinical Application: a Combined MR-PET Analysis of Wholefield and Subfield Hippocampal Changes in AD and FTLD

The pattern of wholefield hippocampal atrophy in Alzheimer's disease (AD) is relatively well established from MRI, but that of frontotemporal lobar degeneration (FTLD) is less known. FDG-PET findings for these neurodegenerative dementias are heterogeneous, and limited to the whole hippocampus, so the mechanism of local hippocampal dysfunction is unclear. It is also uncertain whether single- or multi-modality measures are more informative for maximal group discrimination and individual clinical diagnosis of dementia. We aim to address these queries and study limitations using 3 Tesla high-resolution T1-weighted MRI combined with FDG-PET data. Using a newly-improved, fully-automated hippocampal segmentation tool for MRI data (termed FIRSTv3), we performed an individual region-of-interest analysis of wholefield hippocampal atrophy and FDG metabolism in AD (N=21) and FTLD (N=10) compared to controls (N=13), including: (i) analysis of variance on wholefield measures; (ii) vertex analysis of variations in hippocampal size/shape; and (iii) linear discriminant analysis to investigate classification accuracies of single- and multi-modality hippocampal measures. Additionally, novel analysis of hippocampal subfield metabolism was performed using state-of-the-art

image processing (including partial volume correction of high-resolution PET data, and subfield atlas mapping to the anatomically precise MR images). FTLD was either considered as one group, or split into semantic dementia and frontotemporal dementia subtypes. Both AD and FTLD displayed significant hippocampal atrophy (p=0.010 and 0.005, respectively), but atrophy on the right side showed greater similarity than on the left. AD-associated atrophy mapped to lateral CA1 areas of the body and tail (bilaterally), lateral and dorsolateral CA1 areas of the hippocampal head (left side only), and the CA23 subfield, whilst the FTLD group displayed additional left-sided medial and posterior atrophy of the hippocampal body and tail. Significant wholefield hypometabolism was only found in FTLD, driven primarily by left-side dysfunction in semantic dementia (p=0.001). Interestingly, results suggest that combined MR+PET measures give better dementia versus non-dementia diagnoses (79.4% - 100% overall classification accuracy), but single-modality measures perform better for subsequent discrimination between dementias. Furthermore, findings support the hypothesis that in AD a compensatory mechanism maintains neuronal activity despite structural atrophy, and the left CA23 subfield might be the first place where severe atrophy overwhelms this mechanism. These results support multi-modal image analysis for improved understanding and diagnosis of dementia, as well as providing novel insight into regional vulnerability and compensatory mechanisms of the human hippocampus. Furthermore, the left CA23 subfield is identified as an important site for drug action to preserve metabolic function in AD.

6.1 Introduction

This chapter explores a clinical application of hippocampal segmentation methods. Once accurate and robust hippocampal segmentation estimates are obtained, the normalized hippocampal

volume can be used to investigate hippocampal atrophy, and measures of hippocampal glucose metabolism (derived from corresponding FDG-PET data) facilitate the investigation of (functional) metabolic changes in target diseased groups. As previously reported in Chapter 2 (Section 2.2), Alzheimer's disease (AD) and frontotemporal lobar degeneration (FTLD) are two of the most common causes of dementia, causing debilitating neurodegeneration of individuals. Generally speaking, these dementias are characterized by a well-described pathological and neuropsychological profile, facilitating diagnosis based on a combination of imaging markers and neuropsychological tests [Dubois et al., 2007; Neary et al., 1998]. We have seen that both MRI and FDG-PET are invaluable in this respect; detecting structural abnormalities and changes in the cerebral metabolic rate of glucose consumption (CMRglc), respectively. On a more individual basis, dementia is a broad spectrum of disease and there is often significant overlap in these profiles for any given subject, making individual diagnosis extremely difficult. Academic research therefore continues to focus its efforts on identifying unique differences in the pathological and neuropsychological profiles of AD and FTLD [Rascovsky et al., 2011], to facilitate better differentiation and individual diagnosis. Since the hippocampus is known to be affected early in the course of disease for both AD and FTLD patients [Delacourte et al., 1999], it is our structure of interest here. We have already described the general findings of global/wholefield hippocampal atrophy in these neurodegenerative dementias (Chapter 2, Section 2.2), so we now consider more detailed/local hippocampal changes; describing observed laterality differences, as well as local patterns of atrophy and hypometabolism.

Previous MRI studies typically report bilateral, wholefield hippocampal atrophy in both semantic dementia (SD) and frontotemporal dementia (FTD), but not in progressive non-fluent aphasia (PA, or PNFA) [Van de Pol, 2006]. Atrophy in FTD and SD patients is thought to be localized predominantly in the anterior hippocampus [Laakso et al., 2000; Chan et al., 2001], with additional reports of an asymmetric (left>right) pattern of hippocampal atrophy in SD [Galton et al., 2001; Van de Pol, 2006]. In contrast, AD patients typically exhibit more dif-

fuse, bilateral atrophy of the hippocampus [Csernansky et al., 2000; Laakso et al., 2000; Chan et al., 2001; Galton et al., 2001; Van de Pol, 2006], with less severe left-sided atrophy than SD patients [Galton et al., 2001; Van de Pol, 2006]. Additionally, there are both reports of symmetric and asymmetric (but non-lateralized) hippocampal atrophy in AD [Jack et al., 1998; Chan et al., 2001; Jonathan and Frosta, 2005; Shi et al., 2009]. In recent years, several groups [Apostolova et al., 2006; Becker et al., 2006; Frisoni et al., 2008] have used a radial mapping technique [Thompson et al., 1996 2004; Narr et al., 2004] to assess local hippocampal changes in mild cognitively impaired (MCI) and AD patients relative to controls (CON). By mapping cytoarchitectural subfields onto 3D models of the hippocampal formation, Frisoni et al. [2008] were able to identify subfield-specific atrophy in both normal aging and AD. AD-associated atrophy mapped to the cornu ammonis 1 (CA1) areas of the body and tail, and to dorsolateral CA1 areas of the hippocampal head, whilst regions corresponding to the CA2-3 subfields were relatively (although not altogether) spared. Other studies [Csernansky et al., 2000; Wang et al., 2003; Chételat et al., 2008b] also reported atrophy on the lateral surface of the hippocampal body and the head. The pattern of hippocampal atrophy in AD patients is therefore relatively well established.

In contrast, FDG-PET studies have yet to agree on the extent of hippocampal hypometabolism in AD patients; some studies find significant hippocampal hypometabolism early in the course of the disease, at the stage of mild cognitive impairment (MCI) and mild AD [De Santi et al., 2001; Nestor et al., 2003; Mosconi, 2005; Mevel et al., 2007; Mosconi et al., 2008], whilst others observe no hippocampal dysfunction at all [Minoshima et al., 1997; Ishii et al., 1998; De Leon et al., 2001; Herholz et al., 2002]. In 2008, one particular report [Chételat et al., 2008a] touched on the idea that, in the hippocampus, where the aforementioned disconnection processes are likely to occur, relative synaptic compensatory mechanisms may be taking place, maintaining neuronal activity in the face of structural alterations. Although less widely documented, studies of CMRglc in FTD also report heterogeneous findings for the hippocam-

pus [Franceschi et al., 2005; Diehl-Schmid et al., 2007; Mosconi et al., 2008]. The extent of hippocampal hypometabolism in both AD and FTLD patients is therefore still fairly unknown.

Such FDG-PET studies either adopt an individual region of interest (ROI) approach (where a hippocampal ROI is manually defined on individual MRI scans), a standard template-based ROI approach (where a single hippocampal ROI is manually drawn on the mean MRI template of the whole data cohort), or a whole-brain voxel-wise analysis as implemented in the Statistical Parametric Mapping (SPM) software package. Thus far, reports of significant hippocampal hypometabolism in MCI and mild AD have primarily been found using rigorous individual ROI analyses, but not with wholebrain voxel-wise analysis. Several studies investigating both methods on the same PET data have reported lower sensitivity of voxel-wise analysis compared to the ROI approach [De Leon et al., 2001; Nestor et al., 2003; Mosconi, 2005]; thought to be due to image degradation during preprocessing (registration to a common space followed by a smoothing at 8-14mm FWHM). One recent study [Mevel et al., 2007] found that significant hippocampal hypometabolism in amnestic MCI patients could be found using voxel-wise analysis with *a priori* hypothesis upon the hippocampus (corresponding to a more liberal statistical threshold), suggesting that the more stringent statistical threshold used in previous voxel-wise approaches (without *a priori* hypotheses) most likely accounts for the lack of signficant findings using this method. However, the authors also state that the reliability of the voxel-based approach in their study may be due, at least in part, to minimal atrophy in the population of patients studied, so the same sensitivity may not be achieved using voxel-wise analysis on other, more atrophied data. It therefore appears that individual ROI analysis, with the PET data preserved in the native subject space, is currently the most sensitive method for detecting hippocampal hypometabolism in dementia patients. Furthermore, if automated hippocampal segmentation methods can be used for individual delineation of each subject's hippocampi, we no longer suffer from the burden of observer-dependent, time-consuming manual delineation of these ROIs.

Here, we initially present an individual ROI analysis of wholefield hippocampal volume and FDG metabolism in both AD and FTLD patients (including FTD and SD subtypes) compared to a control group. Working in the native subject space, with no spatial registration or smoothing of the data, we take advantage of the high sensitivity of individual ROI approaches to detect volumetric and metabolic changes in dementia patients. Standard vertex analysis [Patenaude et al., 2011] is also performed to assess local hippocampal changes in size and shape between groups. Discriminant analysis is employed to investigate whether single- or multi-modality hippocampal measures best discriminate dementia patients from controls and other patients. By examining both structural and functional information in the same clinical dataset, this study builds upon previous single-modality analysis of wholefield hippocampal atrophy [Thompson et al., 2004; Frisoni et al., 2007; Laakso et al., 2000; Chan et al., 2001; Galton et al., 2001; Van de Pol, 2006] and wholefield hippocampal metabolism [Mevel et al., 2007; Nestor et al., 2003; Ishii et al., 1998] in one or both of these diseases. The few studies exploring relationships between hippocampal atrophy and hypometabolism typically focus on AD patients [Villain et al., 2008; Chételat, 2003; Chételat et al., 2008a; De Santi et al., 2001], so this comparison with FTLD provides additional insight into the possible mechanistic differences between these diseases.

Secondly, we investigate metabolic changes in three hippocampal subfields, the cornu ammonis (CA) 1, CA23, and the dentate gyrus hilus (DGH), identified via shape-based alignment of a high-resolution postmortem atlas [Yushkevich et al., 2009] to each subject's hippocampus labels (themselves derived from the anatomically precise MR data). Thus far, studies of hippocampal hypometabolism have been limited to the whole hippocampus, primarily due to the poor spatial resolution of PET data and spillover of intensity information due to partial volume effects, coupled with an inability to manually identify the hippocampal subfields on the corresponding MRI scans. In this study, however, a combination of native-space analysis and state-of-the-art image processing enables us to detect clear differences in the local metabolic profile of AD and FTLD patients. We believe that this is the first time that such measures of hippocampal subfield

metabolism have been reported. This latter investigation is somewhat exploratory because it is yet unclear whether the above methodological steps provide us with sufficient resolution to extract metabolic information from the relatively small hippocampal subfields. At the end of the chapter, we provide a discussion of this issue, which we believe justifies our analysis of hippocampal subfield metabolism.

6.2 Materials and methods

6.2.1 Subjects

This study analyzed both structural MR and (functional) FDG-PET data of 13 controls (CON), 21 AD and 10 FTLD patients provided by collaborators at the Max-Planck-Institute in Leipzig. These subject numbers are in accord with previous studies exploring relationships between hippocampal atrophy and hypometabolism in AD patients [Villain et al., 2008; Chételat, 2003; Chételat et al., 2008a; De Santi et al., 2001]. The FTLD group was further categorized into two FTLD subtypes, with 6 semantic dementia (SD) and 4 frontotemporal dementia (FTD) patients. The full demographics of this data were provided in Section 2.6. Probable AD was diagnosed according to the original and revised NINCDS-ADRDA criteria [McKhann et al., 1984; Dubois et al., 2007]. Diagnosis of FTLD was based on the criteria suggested by Neary et al. [1998]. The CON group included people who visited the Day Clinic of Cognitive Neurology at the University of Leipzig with subjective cognitive complaints, but these could not objectively be verified by a subsequently performed comprehensive neuropsychological and clinical evaluation. In fact, the comprehensive evaluation identified these subjects as being cognitively normal elderly. FDG-PET and MRI for these subjects was conducted for diagnostic reasons within the clinical assessment. This control group was chosen because, of the cognitively normal elderly subjects, only those with subjective cognitive complaints are referred to the clinic for further assessment, and so it is these subjects that provide an accurate representation of clinical, elderly controls.

As such, in clinical practice, it is crucial to discriminate between these elderly control subjects (i.e., subjects showing a normal age-related decrease in cognitive performance) and patients with an early stage of dementia. Patients were excluded if structural imaging revealed lesions due to stroke, traumatic head injury, brain tumor or inflammatory diseases. Informed consent was obtained from all subjects. The research protocol was approved by the ethics committee of the University of Leipzig, and was in accordance with the latest version of the Declaration of Helsinki.

6.2.2 MR imaging

For each subject, a high-resolution T1-weighted MRI scan was obtained, consisting of 128 sagittal slices adjusted to the AC-PC line, a slice thickness of 1.5 mm, and a pixel size of 1x1 mm^2. MRI was performed on two different 3T scanners (MedSpec 30/100, Bruker Biospin, Ettlingen, Germany, and Magnetom Trio, Siemens, Erlangen, Germany) using two different T1-weighted sequences (MDEFT or MP-RAGE with TR =1300 ms, TI = 650 ms, TE = 3.93 ms or TE = 10 ms; FOV 25x25 cm^2; matrix = 256x256 voxels). On the MedSpec scanner, only the MDEFT-sequence was used, whilst on the Magnetom Trio scanner, either MDEFT or MP-RAGE sequences were used. The distribution of scanner types and sequences used to obtain the MRI data was balanced across subjects and did not differ significantly in its distribution between the groups, or for scanner type, or for sequence (as detailed in Appendix B).

6.2.3 FDG-PET imaging

All PET data were acquired on a Siemens ECAT EXACT HR+ scanner (CTI/Siemens, Knoxville, TN, USA) under a standard resting condition in 2-dimensional (2D) mode. In keeping with the Society of Nuclear Medicine procedure guidelines (interactive.snm.org, [Waxman et al., 2009]), all subjects were injected with $\sim$370 MBq (10 mCi) of the FDG tracer 30 min after the start of a 60 min period in which the subjects were placed in a quiet, dimly-lit room with eyes open.

The PET data were acquired dynamically up to 60 min post injection of the tracer. For further analysis, the averaged 40-60 min post-injection PET data were used. Sixty-three slices were simultaneously collected with an axial resolution of 5 mm full width at half maximum (FWHM) and in-plane resolution of 4.6 mm. After correction for attenuation, scatter, decay and scanner-specific dead time, images were reconstructed by filtered back-projection using a Hann-filter of 4.9 mm FWHM. The 63 transaxial slices obtained had a resolution of 128x128 voxels with an edge length of 2.45 mm.

6.2.4 MR image processing

Unless otherwise stated, both MR and FDG-PET data were processed using tools from the FMRIB Software Library (FSL [Smith et al., 2004; Woolrich et al., 2009], www.fmrib.ox.ac.uk/fsl). The MR images were initially interpolated to an isotropic resolution of 1x1x1 mm^3, cropped to remove excess neck, and BET [Smith, 2002] was used to remove most of the non-brain structure. The subsequent brain masks were manually edited using the FSLview software to provide an accurate measure of total intracranial volume (TIV) for each subject. For later processing of PET data, the cerebellum of each subject was identified on the corresponding MR image using FIRST [Patenaude et al., 2011].

Whole-hippocampus segmentation

Chapter 3 introduced a wholefield hippocampal segmentation method termed FIRSTv3. By utilizing a novel structure-specific spatial normalization prior to model-fitting, this adaptation of the aforementioned FIRST method provided more robust alignment of hippocampal regions and resulted in improved segmentation performance. Because FIRSTv3 (and FIRST) output a tessellated hippocampal surface mesh for each individual subject, storing the 3D spatial coordinates of each vertex and establishing vertex correspondence across all subjects, it is possible to perform vertex analysis on these resultant surface meshes. We wished to perform such anal-

ysis of surface meshes in the current work (Section 6.2.6), so FIRSTv3 was the segmentation method of choice, rather than any of the previously evaluated segmentation methods.

For implementation of FIRSTv3, the aforementioned three-step affine transform to the standard MNI space was computed with FLIRT [Jenkinson and Smith, 2001], using 12 degrees of freedom (DOF) and maximization of normalized mutual information (with 256 histogram bins). The computed three-step affine transform was then used as input for FIRST model fitting, segmenting each subject's hippocampi. As per Chapter 3, FIRSTv3 was actually run simultaneously for both the hippocampus and the amygdala, giving a resultant segmentation with no voxel overlap at the hippocampus-amygdala border. Individual hippocampus (and amygdala) segmentations for the left and right hemispheres were extracted from the FIRSTv3 output volume. The resultant whole-hippocampus volumes were normalized by TIV prior to statistical analysis.

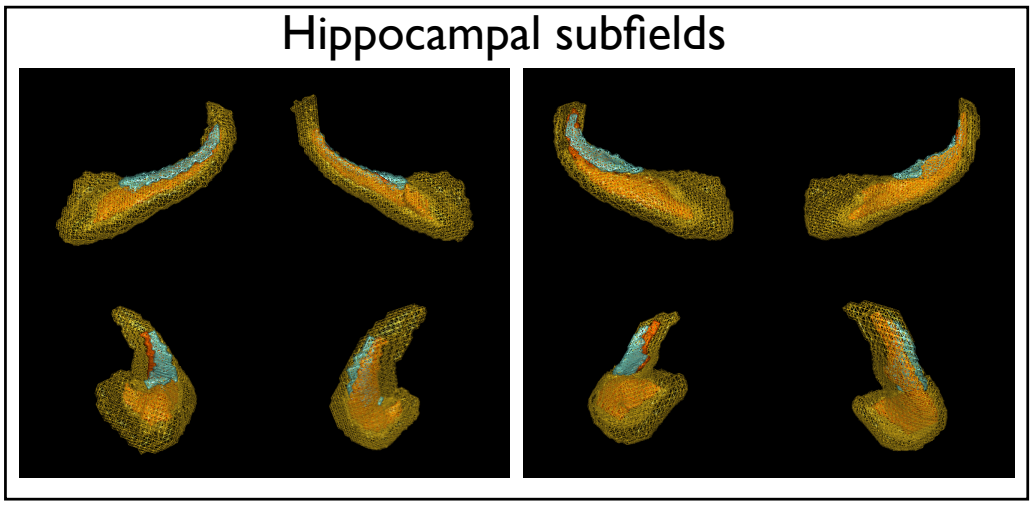
Hippocampal subfield segmentation

Shape-based alignment of binary masks was used to map a high-resolution computational atlas [Yushkevich et al., 2009] to each subject's native space, from the standard MNI brain. The computational Penn Hippocampus Atlas is derived from high-resolution 9.4 Tesla MRI of post-mortem samples, and includes the main subfields of the hippocampus (cornu ammonis fields CA1, CA2/3; the dentate gyrus; and the vestigial hippocampal sulcus). In this work, only the three largest subfields were investigated — CA1, CA23, and the hilus of the dentate gyrus (DGH) — so the smallest dimension of any given subfield was typically 2mm.

The Harvard-Oxford atlas (part of FSL) provided binary hippocampal masks of the standard MNI brain, whilst the previously-computed FIRSTv3 labels guided the shape-based registration to native subject space. Specifically, a binary mask of the Penn Hippocampus Atlas was mapped, in turn, to the left- and right- hippocampal masks of the standard MNI brain, with trilinear interpolation of values. This transformation of the original binary mask generated

a hippocampal partial volume map in standard space, which was thresholded at 0.5 and binarized, before being mapped from the standard MNI space to each subject's FIRSTv3 label. In each case, shape-based registration proceeded via an initial 12 DOF alignment with FLIRT (maximizing normalized correlation) followed by a non-linear warp with FNIRT [Andersson et al., 2008]. The computed non-linear warps were subsequently applied to binary masks of the hippocampal subfields, thresholded at 0.5 and binarized, to provide conservative subfield labels for the left and right hippocampus of each subject. Figure 6.1 shows multiple views of the left (left panel) and right (right panel) hippocampal subfields (CA1: gold; CA23: turquoise; DGH: copper) in standard space, prior to registration to each subject's native space. The wireframe surface meshes were constructed with the marching cubes algorithm in BriView (http://web.me.com/sebastianpatenaude/BriView), and the opacity of the CA1 subfield is reduced to aid visibility of the DGH within the hippocampus core.

Figure 6.1 *Hippocampal subfields.* Lateral- (top left), medial- (top right), anterior- (bottom left) and posterior-view (bottom right) wireframe rendering of the left (left panel) and right (right panel) hippocampal subfields (CA1: gold; CA23: turquoise; DGH: copper) in standard space, constructed with BriView (http://web.me.com/sebastianpatenaude/BriView).



Yushkevich *et al.* [Yushkevich et al., 2009] state that shape-based normalization (registration) is appropriate for mapping anatomical data between structures in situations where image intensity does not provide sufficient cues for establishing a mapping. The boundaries between subfields

are very well defined in the postmortem atlas, but these same boundaries are blurred or invisible on the *in vivo* T1-weighted images. When registering the postmortem atlas to MNI space, and to each subject's native space, we therefore cannot rely on the boundaries between the subfields to guide the registration. Instead, we must compute the mapping between wholefield hippocampal labels and then apply this transform to the subfield labels, making the assumption that the position of the subfields relative to the whole hippocampus remains roughly the same in the *in vivo* hippocampus and the postmortem atlas. Like Yushkevich et al. [2009], we appreciate that this is a strong assumption that leads to an inherent level of error in the subfield boundary estimation, so this technique is not appropriate for investigating subfield volumes and we do not report such measures here. However, since errors in the subfield boundary estimation become less of an issue when integrating voxel-wise quantities over the entire subfield, the labels may be used to investigate subfield metabolism by averaging FDG uptake over the entire subfield. By thresholding the transformed subfield labels, we also find a conservative estimation of each subfield boundary, reducing the susceptibility of our measures to small errors in the subfield boundary estimation.

6.2.5 FDG-PET data processing

Each subject's FDG-PET data was initially corrected for partial volume effects (PVEs) using the modified Müller-Gärtner (mMG) method [Müller-Gärtner et al., 1992; Rousset et al., 1998] implemented in the PVELab software [Quarantelli et al., 2004], and co-registered to the corresponding MR image using a 6 DOF rigid-body registration with FLIRT, maximizing normalized mutual information (with 256 histogram bins). The PVE-corrected, co-registered PET data (cPET) was intensity-normalized using the mean cerebellar uptake, derived from the aforementioned FIRST segmentations. Finally, the average FDG uptake within the whole-hippocampus and each hippocampal subfield was assessed.

PVE-correction (PVEC) of FDG-PET data accounts for the initial point spread function (PSF)

of the PET scanner, thereby correcting for the contribution of surrounding tissue to each specific voxel. Using the substantially more precise anatomical information provided by the subject's MRI data, PVEC methods are able to increase the effective anatomical resolution of the PET signal [Müller-Gärtner et al., 1992; Rousset et al., 1998; Quarantelli et al., 2004]. A phantom study of four different PVEC methods (termed M-PVEc, MG-PVEc, R-PVEc, and mMG-PVEc) revealed higher accuracy of all PVEC PET measures compared to non-corrected measures [Quarantelli et al., 2004], and higher sensitivity of PVEC measures for detecting hippocampal hypometabolism in amnestic MCI patients has been shown *in vivo* [Mevel et al., 2007]. Furthermore, the detrimental effects of partial-voluming are heightened in small ROIs such as the hippocampus, so it is both appropriate and important to perform such PVEC herein.

The superiority of the mMG PVEC method has been shown in a comparison of currently available PVEC methods [Quarantelli et al., 2004]. With voxel-based correction for both the loss of GM activity (as a result of spill-out onto non-GM tissues) and the gain in GM activity (as a result of spill-in from adjacent GM), Quarantelli et al. [2004] report superior precision and accuracy of GM concentration estimates (of more than 96%) using the mMG-PVEC method. Results therefore indicate that the mMG-PVEC method is the best currently available voxel-based approach for PVEC in FDG-PET data, hence our decision to use this particular method here.

Additionally, for intensity normalization of PET data, a metabolically stable reference region is required. According to Yakushev et al. [2008], this reference area should be minimally susceptible to external physiological stimuli, unaffected by the disease of interest, and relatively easy to determine. A whole array of different reference regions have been used in the past, including the cerebral metabolic rate of glucose consumption (CMRglc) [Chételat et al., 2003; Mevel et al., 2007; Mosconi et al., 2008; Yakushev et al., 2008], the cerebellum [Minoshima et al., 1995; Santens et al., 2001; Jeong et al., 2005; Mevel et al., 2007; Yakushev et al., 2008], the entire pontine [Minoshima et al., 1997] or the center of the midpontine slice [Minoshima et al.,

1995; Mosconi, 2005], the sensorimotor cortex [Minoshima et al., 1995; Yakushev et al., 2008], the visual cortex [Minoshima et al., 1995; Santens et al., 2001], and even fairly arbitrary slabs of tissue encompassing the entire anterior-posterior span of the temporal lobes [De Leon et al., 2001]. However, the cerebellum has been shown to be least affected in FTD and mild to moderate stages of AD [Jeong et al., 2005; Ishii et al., 1997], and normalization to this region improves the statistical discrimination between dementia patients and controls [Dukart et al., 2010], so we chose the cerebellum as our reference for intensity normalization here. Furthermore, we felt that the globally differing metabolic profiles of AD, FTLD and NC, would far outweigh any cerebellar differences, making it a more suitable reference compared to the CMRglc.

6.2.6 Statistical analysis

Chi-squared tests, t-tests, and analysis of variance

SPSS statistics software (Version 19.0) was used throughout. Initially, group comparisons for age and Clinical Dementia Rating (CDR) [Morris, 1993] were performed using a univariate analysis of variance (ANOVA) design with group as a three-level fixed factor (AD, FTLD, CON). Independent t-tests with Bonferroni correction for multiple comparisons were also performed. Mini Mental State Examination (MMSE) scores [Folstein et al., 1975] were available for 20 AD and 8 FTLD patients (4 SD, 4 FTD), but not for the CON subjects (due to expected ceiling effects), so the two patient groups were compared using an independent samples t-test (confidence interval = 0.95). Group comparisons for gender were evaluated using a chi-square test for independent samples.

Subsequently, ANOVA was performed on the measures of wholefield hippocampal volume, wholefield hippocampal metabolism, and hippocampal subfield metabolism. For both the wholefield and subfield analysis, side was considered as a within-subjects factor with two levels (left and right). Depending on whether the FTLD group was classified as one entity or two separate subtypes (SD and FTD), analysis was performed with group/disease as a three-level (CON,

AD, FTLD) or four-level (CON, AD, SD, FTD) between-subjects factor. Age and sex were added as covariates throughout. Paired-sample t-tests with Bonferroni correction for multiple comparisons were also performed.

For the whole-hippocampus analysis, 2*3 and 2*4 (side*group) mixed ANOVA designs were initially employed to test for main effects of side or group and any side-by-group interactions. Following this, two-way ANOVA designs investigated the effect of side for each individual group, whilst three- and four-way ANOVA designs tested for a between-subjects effect of group for each side. For the hippocampal subfield analysis, we initially tested for main effects and interactions using 3*2*3 and 3*2*4 (subfield*side*group) mixed ANOVA designs. In both cases, Mauchly's test indicated that the assumption of sphericity for subfield had been violated ($p<0.005$) so degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. The metabolism of each hippocampal subfield was subsequently analyzed using the ANOVA designs described above for the whole-hippocampus.

Owing to the relatively low number of subjects in the SD and FTD subgroups compared to the number of variables, we appreciate that it will be difficult to achieve significance for some of the four-level group analyses. Therefore, in these instances, we report trends in the data that should be further investigated on larger clinical datasets.

Vertex analysis

Vertex analysis was performed to assess group differences in size and shape (i.e., atrophy) of the whole-hippocampus. Once again, FTLD patients were either categorized as one group (FTLD) or as two separate subtypes (FTD and SD). For all group pairs (e.g., CON v. AD, CON v. FTLD, CON v. SD etc.), vertex analysis was performed on the previously-generated FIRSTv3 hippocampal segmentations, with reconstruction of surface meshes in the standard MNI space, effectively normalizing for head size. Statistical analysis was performed using a multi-variate general linear model (GLM) with Pillai's Trace as the test statistic. Subsequently,

results were corrected for multiple comparisons using false discovery rate (FDR). Again, due to the relatively low number of SD and FTD patients, shape differences for these FTLD subtypes may not survive the FDR threshold. In these cases, we view the vertex analysis results prior to correction for multiple comparisons, and report relevant findings.

Discriminant analysis

To investigate whether single- (MR, PET) or multi-modality (MR+PET) hippocampal measures best discriminate patients from controls and other patients, we used the SPSS statistics software (Version 19.0) to perform two-way linear discriminant analysis (LDA), with group as the dependent variable and image measures (wholefield hippocampal metabolism and volume, both left and right side) as predictor variables. For classification, groups were considered to have equal prior probabilities, and appropriate cross-validation was done — treating n-1 out of n observations as the training dataset to determine the discriminant function/rule that best separates the two groups, and using this rule to classify the one observation left out. Initially, we determined the ability of the measures (PET, MR, MR+PET) to discriminate between controls and all remaining patients (either AD+FTLD, or AD+FTD+SD), before investigating the classification accuracy for each group pair (e.g., AD v. FTLD, AD v. SD, etc.). Classification accuracy was defined as the percentage of cases correctly classified, both for each group and overall.

6.3 Results

6.3.1 Demographics

Pearson's chi-square test for independent samples did not reveal any difference in gender between the groups (χ^2(2, N=44) =0.905, p=0.636). The three-way ANOVA designs revealed a significant main effect of group for both age (F(2,41)=6.36, p=0.004) and CDR (F(2,41)=13.30, p$<$0.0005). Post-hoc t-tests for independent samples showed that there was no difference in age

between the patient groups (p=1.000), so the main effect of group was driven by younger control subjects compared to AD and FTLD patients (p=0.007 and p=0.015, respectively) — accounted for by the presence of the age covariate in all remaining statistical analyses. Likewise, the patient groups did not differ in their CDR scores (p=0.815), but as expected, both AD and FTLD patients had significantly higher CDR scores compared to controls ($p<0.0005$ for both). Finally, for the two patient groups, the independent samples t-test identified no difference in MMSE scores (p=0.448), indicating a similar severity of dementia in each patient group.

6.3.2 Wholefield hippocampal atrophy

Three-level group analysis

Testing for significant differences in hippocampal volume, the 2*3 (side*group) RM-ANOVA design revealed a significant effect of group (p=0.006), but no main effect of side and no interactions. Wholefield hippocampal atrophy was evident in both the AD and FTLD groups compared to controls (p=0.010 and 0.005, respectively), as shown in Figure 6.2 (left panel, top row).

In general, vertex analysis results for AD patients showed a similar magnitude and pattern of local hippocampal atrophy bilaterally (Figure 6.3: top panel, left column), with the exception of the dorsolateral surface of the hippocampal head, where highly-significant atrophy ($F(3,30)\sim5$) was observed for the left side only. Regions of highly-significant bilateral hippocampal atrophy were largely restricted to the lateral and dorsal surfaces, (in particular, lateral and dorsal regions of the body (CA1 and CA23), and posterior aspects of the body and tail (CA1)), whilst there was very little atrophy on the medial surface (limited to a very small medioanterior region of the head).

In contrast, FTLD patients displayed a very different pattern and magnitude of local hippocampal atrophy bilaterally (Figure 6.3: top panel, middle column). Additionally, AD and FTLD

Figure 6.2 *Wholefield hippocampal measures.* Estimated marginal means of normalized hippocampal volume (top) and normalized hippocampal metabolism (bottom) across group, for the three- and four-level group analysis (left and right panel, respectively). Each duplet of error bars show the standard errors for the left (left) and right (right) hippocampal measures. Volume is normalized by total intracranial volume (TIV), and metabolism is normalized by cerebellar FDG uptake.

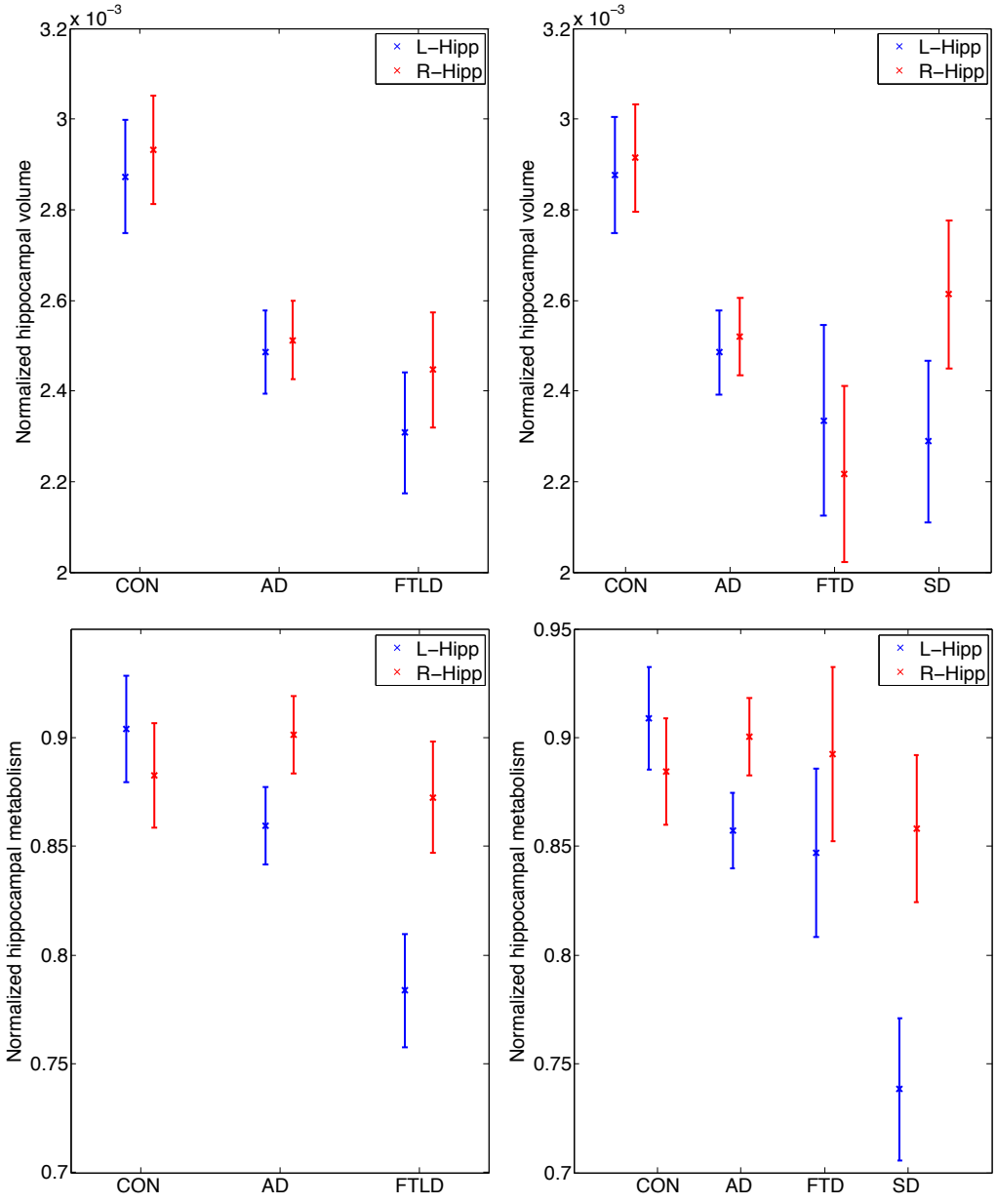
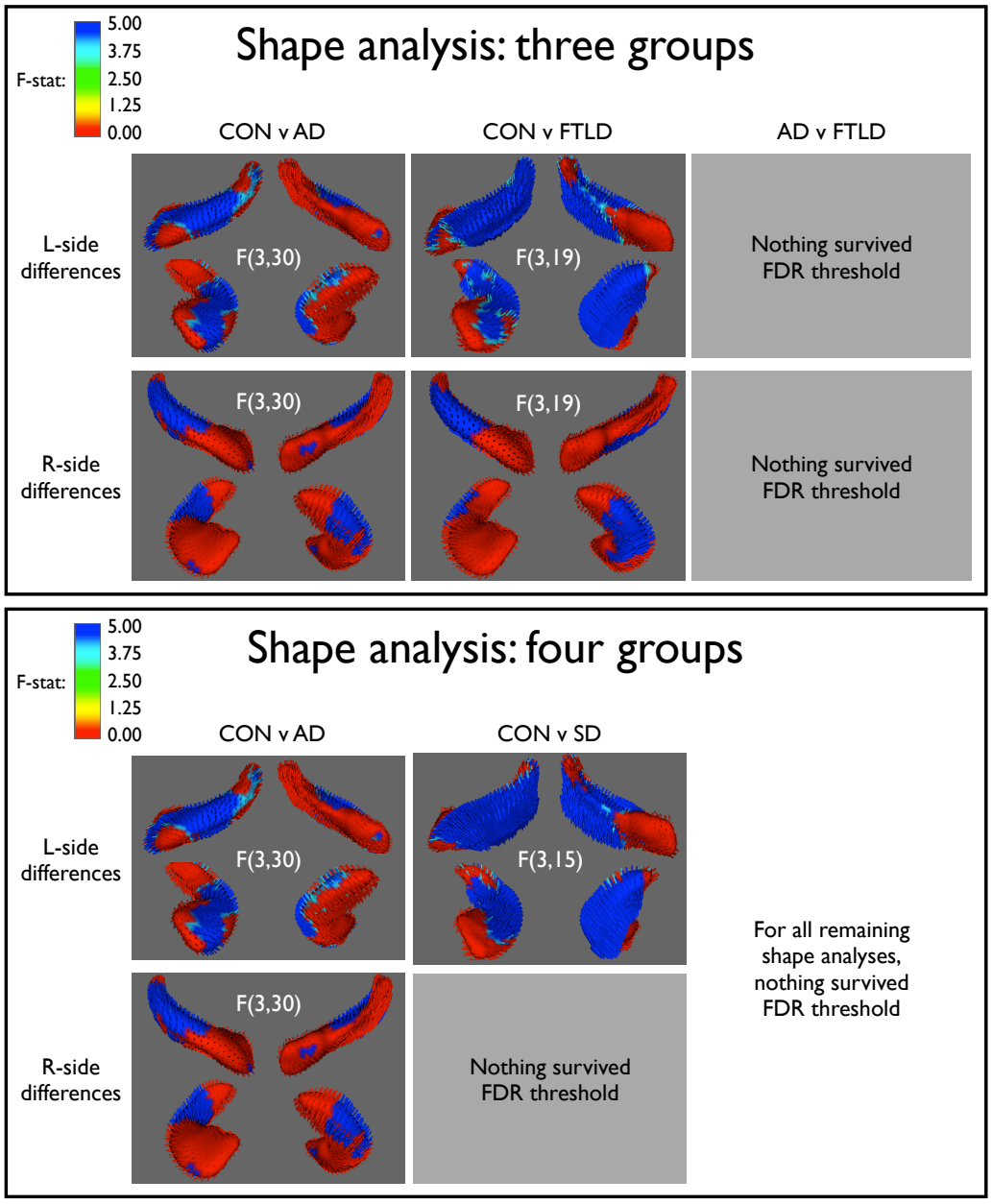


Figure 6.3 *Vertex analysis with false discovery rate (FDR) threshold.* Vertex analysis results for group comparisons in the three- (top panel) and four- (bottom panel) group scheme, with correction for multiple comparisons (i.e., false discovery rate (FDR) threshold). Each 2x2 grid within a panel displays the lateral- (top left), medial- (top right), anterior- (bottom left) and posterior-view (bottom right). Surface colour gives the F-statistic for each group comparison, with blue indicating highly-significant group differences. Degrees of freedom (DOF) for each F-test are also given.



patients displayed more similar atrophy on the right side than the left. As such, with the exception of some additional atrophy along the ventral and posterior surfaces of the body and tail (CA1), the pattern of right-sided atrophy in FTLD was similar to that described above for AD patients. On the left side, however, and in contrast to AD, FTLD patients showed additional highly-significant atrophy ($F(3,19)\sim5$) on medial and posterior surfaces of the hippocampal body and tail (CA1). Consequently, on the left side, only the tip of the tail and medial aspects of the hippocampal head and head-to-body transition zone were spared in FTLD.

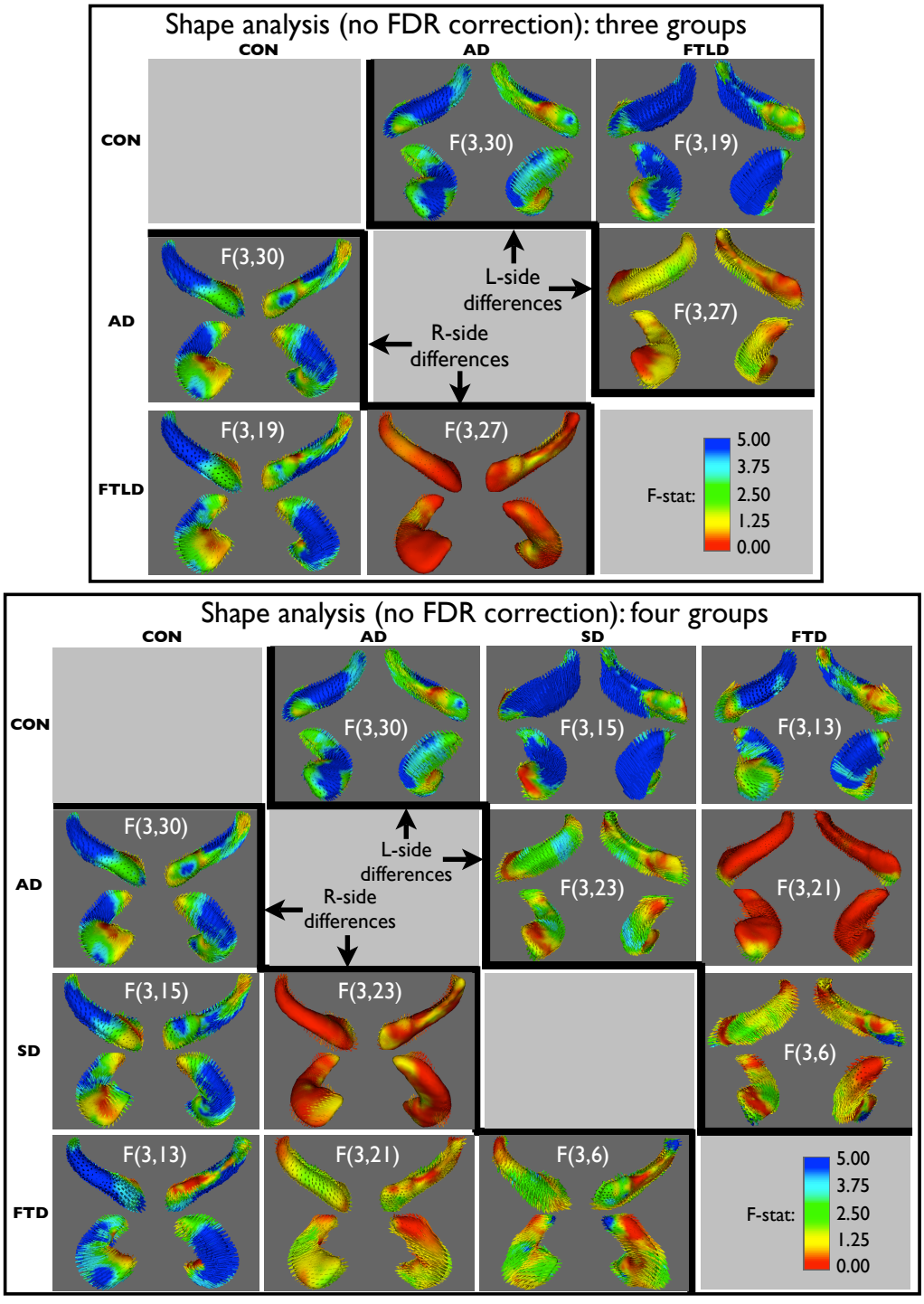
As expected from the similar pattern and magnitude of right-sided atrophy, no significant surface differences were found between AD and FTLD on the right side. For the left side, however, significant surface differences ($F(3,27)\sim3$) were detected for medial and posterior aspects of the hippocampal body prior to correction for multiple comparisons (Figure 6.4: top panel, middle row, right column).

For both AD and FTLD, comparison of the average surfaces with those of the CON group generally showed an inwards and downwards contraction of the lateral and dorsal surfaces, towards the medial and ventral boundaries. Viewing the hippocampus along its long axis, we found that medial surfaces of the hippocampal head and tail act as stable anchor points, whilst the rest of the hippocampus shifts medially and ventrally. Additionally, the aforementioned atrophic regions of the hippocampal head displayed an inward and upwards contraction.

Four-level group analysis

As for the three-level group analysis of hippocampal volume, a main effect of group was found for the 2*4 (side*group) RM-ANOVA design (p=0.015), but no main effect of side and no interactions. For this four-level group analysis, the significance of hippocampal atrophy was slightly reduced for the AD group compared to CON (p=0.027), but the analysis allowed us to differentiate between the FTD and SD subtypes: considering the left and right side together, there was significant bilateral wholefield atrophy for FTD compared to CON (p=0.016), but

Figure 6.4 *Vertex analysis without false discovery rate (FDR) threshold.* Vertex analysis results before correction for multiple comparisons, for the three- (top panel) and four- (bottom panel) group scheme. Within each panel, left hippocampal differences are displayed above the main diagonal, right hippocampal differences below. Further details are per Figure 6.3 caption.



no such bilateral atrophy for SD patients (p=0.193). However, when considering the left and right hippocampus separately in subsequent ANOVA designs, the SD patients had significant left hippocampal atrophy compared to CON (p=0.047). This dominant left-sided atrophy in SD patients can be clearly seen in Figure 6.2 (top right).

Vertex analysis results for AD patients were as described above. As expected, the vertex analyses for SD patients revealed a similar pattern of atrophy as that described above for FTLD patients (combining FTD and SD subtypes), except that the magnitude of the change in SD (i.e., the length of the displacement vectors) was generally enhanced on the left side and reduced on the right. Consequently, whilst the SD versus (v.) CON left-side surface differences survived the correction for multiple comparisons (Figure 6.3: bottom panel, top right), the effect size for the right-side differences was now insufficient to survive the FDR threshold. Results of right-side surface atrophy in SD are therefore shown in Figure 6.4 (bottom panel, third row, first column), before correction for multiple comparisons. Here, in addition to the pattern of atrophy described above for FTLD patients, a region of highly-significant atrophy ($F(3,15)\sim5$) can be seen on the dorsomedial surface of the hippocampal head.

For the FTD group, effect sizes for the vertex analyses were simply too small relative to the current subject numbers, so surface differences did not survive the FDR threshold, and are therefore shown before correction for multiple comparisons (Figure 6.4: bottom panel). Similarly, surface differences for AD v. SD, AD v. FTD, and SD v. FTD are shown before correction for multiple comparisons (Figure 6.4: bottom panel). Of particular note here were the different vertex analysis results for FTD and SD compared to CON, and the results of comparing these two FTLD subtypes. Compared to the atrophy described in SD patients, the vertex analysis results for FTD patients (v. CON) showed that the surface coverage and magnitude of left-hippocampal atrophy generally decreased, whilst that of the right hippocampus increased. Consequently, hippocampal atrophy in FTD patients was much more evenly distributed bilaterally (and of similar pattern on each side). As an additional contrast to SD, highly-significant atrophy ($F(3,13)\sim5$)

was found for anterior aspects of the hippocampal head bilaterally in FTD. In many respects, the pattern of local hippocampal atrophy found in FTD resembles both that of AD and SD patients.

6.3.3 Wholefield hippocampal metabolism

Three-level group analysis

A side-by-group interaction was found for the 2*3 (side*group) RM-ANOVA design of wholefield hippocampal metabolism (p=0.020), but no main effects. Pairwise comparisons revealed a left $<$ right asymmetry (p$<$0.005), which was driven mainly by the FTLD group (p=0.001), but also AD patients (p=0.026) (Figure 6.2: bottom left). Consequently, there was a significant main effect of group for the left hippocampus (p=0.008), with pairwise comparisons revealing left-sided hypometabolism for FTLD patients compared to CON (p$<$0.005) as well as a trend for different uptake between AD and FLTD on the left side (p=0.058, two-tailed).

Four-level group analysis

The 2*4 (side*group) RM-ANOVA found a side-by-group interaction similar to the aforementioned three-level group analysis (p=0.021), but also a trend towards a main effect of group (p=0.064), with subsequent pairwise comparions showing a left $<$ right asymmetry (p$<$0.001), hypometabolism in SD patients compared to CON (p=0.037) and a trend for differential uptake in AD and SD bilaterally (p=0.087, two-tailed). The observed left $<$ right asymmetry results in a main effect of group for the left hippocampus (p=0.003), driven by hypometabolism in SD patients compared to both CON and AD patients (p=0.001 and 0.006, respectively).

6.3.4 Classification accuracy of wholefield hippocampal measures

Three-group scheme

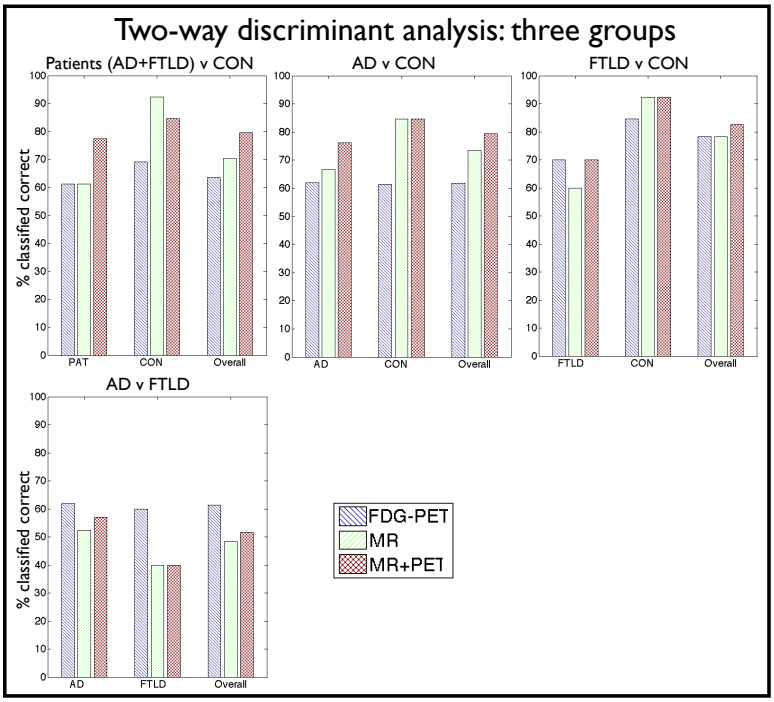
From a first glance at Figure 6.5 (top panel), it is immediately clear that the LDA was better at discriminating patient groups from controls (top row) than it was at discriminating between patient groups (bottom row). For discriminant analysis of patient groups v. controls (top row), the relative importance of the single modality measures for classification of cases in each group was MR $>$ FDG-PET for AD and CON, FDG-PET $>$ MR for FTLD, such that hippocampal volume was an equal or better overall predictor of group than FDG metabolism (third triplet). Combining all predictor variables, the joint MR+PET analysis generally equalled or bettered the single-modality classification accuracy for each group (Patients v. CON: 77.4% for patients; AD v. CON: 76.2% for AD, 84.6% for CON; FTLD v. CON: 70.0% for FTLD, 92.3% for CON) with the exception of CON classification in the patient (AD+FTLD) v. CON analysis (where MR measures alone outperformed the multi-modality approach, correctly classifying 92.3% of CON cases). Consequently, the overall classification accuracy (third triplet) was highest when both MR and PET measures were used in combination (79.5% for patients v. CON, 79.4% for AD v. CON, 82.6% for FTLD v. CON).

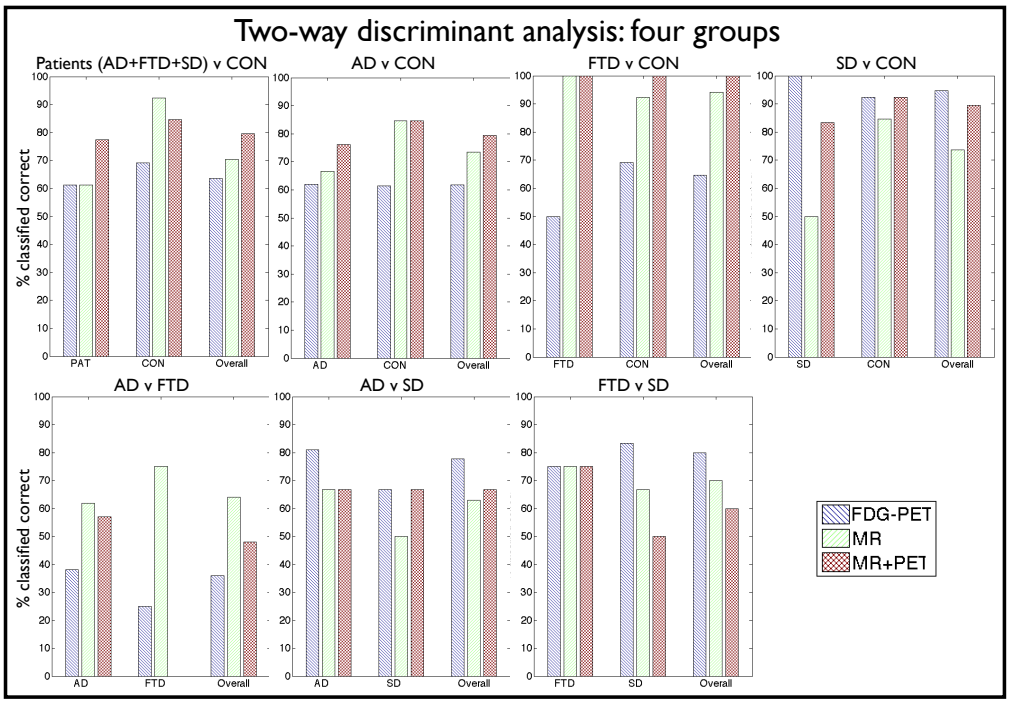
The results of the discriminant analysis were somewhat different for AD v. FTLD patients (bottom row). Interestingly, FDG metabolism alone was the best predictor of group (61.9% of cases correctly classified for AD, 60.0% for FTLD, 61.3% overall), and addition of the MR predictor variables for the multi-modality analysis resulted in a dramatic drop in classification accuracy (57.1%, 40.0%, and 51.6%, respectively).

Four-group scheme

Similar to the discriminant analysis results for the three-group scheme (described above), the first point to note from Figure 6.5 (bottom panel) is that better classification accuracy was

Figure 6.5 *Linear discriminant analysis.* Percentage of cases (subjects) correctly classified using hippocampal FDG-PET metabolism (blue left-slanted lines, left bar of each triplet), MR volume (green right-slanted lines, middle bar of each triplet), and combined MR+PET measures (red criss-crossing lines, right bar of each triplet) as predictor variables in two-way linear discriminant analysis, for the three- (top panel) and four- (bottom panel) group scheme. For each plot, classification accuracy for each group (first and second triplet) is followed by the overall classification accuracy (third triplet).





achieved when discriminating patient groups from controls (top row) than when discriminating between patient groups (bottom row). For discriminant analysis of patient groups v. controls (top row), the relative performance of the single modality measures for classification of cases in each patient group was MR > FDG-PET for AD and FTD, FDG-PET > MR for SD, and no difference for the combined patient group (AD+FTD+SD), whilst for classification of control cases the relative performance of the single modality predictors was dependent on the contrasting patient group: for both AD v. CON and FTD v. CON, hippocampal volume was a better predictor of controls than FDG metabolism, whilst for SD v. CON the opposite was true (i.e., FDG-PET was the better predictor of controls). Consequently, MR-derived measures of hippocampal volume outperformed FDG metabolism in the overall percentage of correctly classified cases for patients v. CON, AD v. CON, and FTD v. CON, but FDG metabolism was the superior single modality predictor overall for SD v. CON (94.7% of cases correctly classified). In fact, whilst joint MR+PET analysis generally provided equal or better classification of cases to each group in the patients v. CON analyses (Patients v. CON: 77.4% for patients, 84.6% for CON; AD v. CON: 76.2% for AD, 84.6% for CON; FTD v. CON: 100% for FTD, 100% for CON; SD v. CON: 83.3% for SD, 92.3% for CON), and the highest overall classification accuracy (79.5% for patients v. CON, 79.4% for AD v. CON, 100% for FTD v. CON), FDG metabolism alone was always the best predictor in the SD v. CON analysis.

For discriminant analysis of patient groups (Figure 6.5: bottom panel, bottom row), the ability of the single modality measures to correctly classify cases within each patient group again depended on the particular groups compared: for classification of AD patients, MR > PET for the AD v. FTD analysis, but PET > MR for AD v. SD (61.9% and 81% of AD cases correctly classified, respectively); for classification of FTD patients, MR > PET for AD v. FTD (75% correct classification of FTD cases), but there was no difference between the single modality predictors for FTD v. SD (again, 75% correct classification of FTD cases); whilst for classification of SD patients, PET > MR for both discriminant analyses (66.7% and 83.3% of SD cases

correctly classified in the AD v. SD and FTD v. SD analyses, respectively). In summary, it therefore seems that hippocampal volume is generally the best predictor for AD and FTD cases, but when either of these patient groups is compared to SD patients, hippocampal metabolism becomes the dominant predictor for all patient groups. As a result, FDG metabolism provides superior overall classification accuracy (third triplet) for AD v. SD (77.8%) and FTD v. SD (80%), whilst the opposite is true for AD v. FTD (64% of cases correctly classified with MR). Of particular surprise here is the relatively poor performance of the multi-modality approach at discriminating between patients groups. For classification of SD cases in the AD v. SD analysis and for classification of FTD cases in FTD v. SD, the joint MR+PET approach matched the classification performance of the best single-modality predictor (66.7% for SD and 75% for FTD, respectively), but for all remaining analyses classification accuracy was worse when MR and PET predictors were combined. This was certainly a surprising result, and will be discussed in Section 6.4.

6.3.5 Hippocampal subfield metabolism

Three-level group analysis

A side-by-group interaction was observed when testing for significant differences in subfield metabolism using the 3*2*3 (subfield*side*group) ANOVA design (p=0.049), but no main effects of subfield, side or group (Figure 6.6: top row). For the CA1 subfield, the 2*3 (side*group) ANOVA design similarly revealed a side-by-group interaction (p=0.015), no main effect of side, but a trend towards a main effect of group (p=0.056). Pairwise comparisons (depicted in Figure 6.7: top panel) also revealed a trend for overall asymmetry (p=0.085, two-tailed), driven by significant left $<$ right differences in FTLD patients (p=0.004). This asymmetry gave rise to a main effect of group for the left CA1 subfield (p=0.004), owing to significant left-side CA1 hypometabolism in FTLD patients compared to both CON and AD (p=0.002 and 0.012, respectively). There was no main effect of side or group, and no interactions, for the CA23 subfield,

Figure 6.6 *Hippocampal subfield metabolism.* Estimated marginal means of left- (left panel) and right- (right panel) hippocampal subfield metabolism across group, for the three- and four-level group analysis (top and bottom panel, respectively). Each triplet of error bars show the standard errors for the CA1 (left), CA23 (middle) and DGH (right) subfield. Metabolism is normalized by cerebellar FDG uptake.

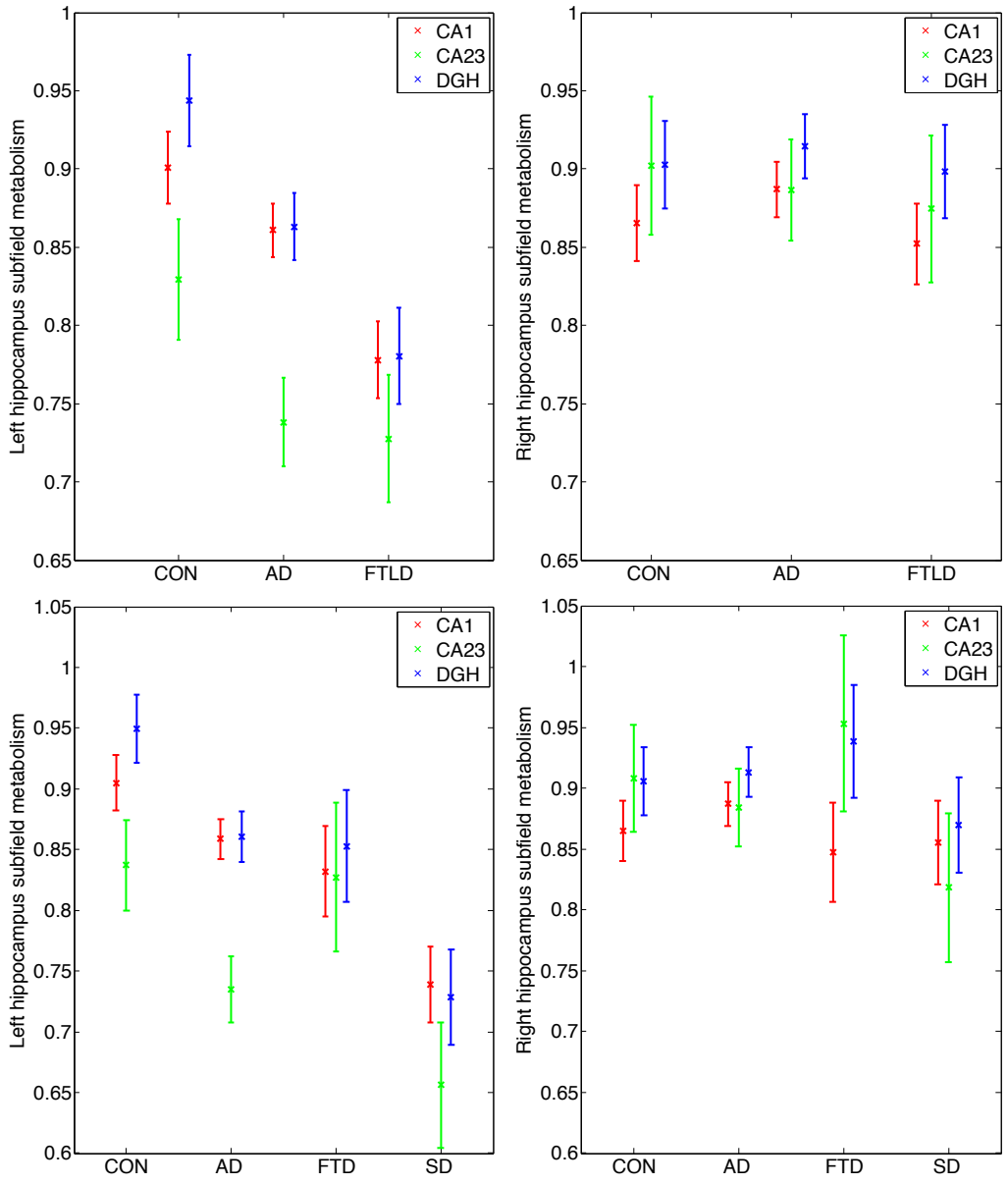
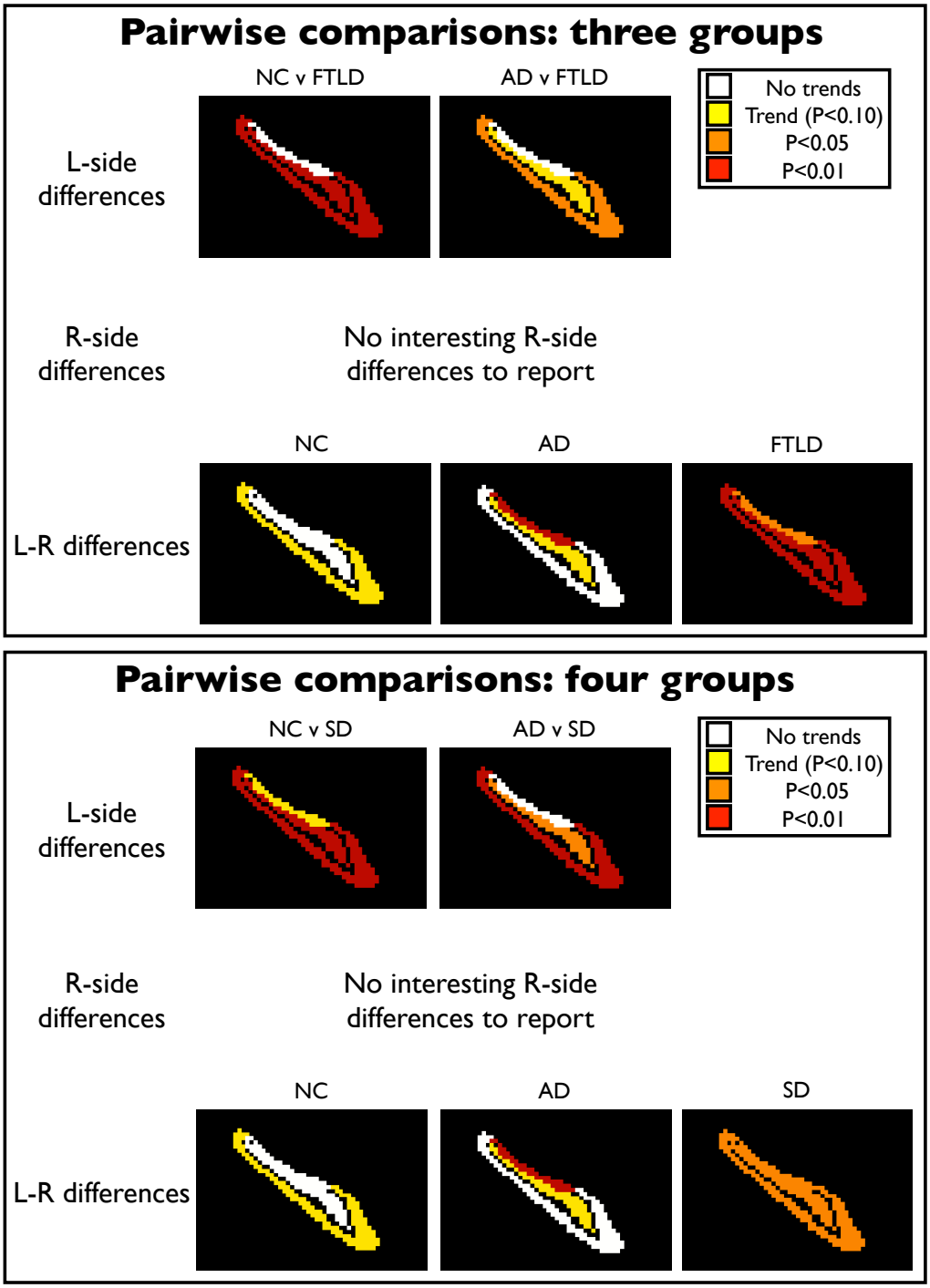


Figure 6.7 *Pairwise comparisons of subfield metabolism.* Schematic representation of the two-tailed, paired-sample t-test results for the three- (top panel) and four- (bottom panel) group comparisons of hippocampal subfield metabolism.



but a significant left < right asymmetry revealed by pairwise interactions ($p<0.001$). In contrast to the CA1 subfield, asymmetry in the CA23 subfield was found in both the AD and FTLD patients (p=0.002 and 0.009, respectively). Results for the DGH subfield were similar to those for CA1: the 2*3 (side*group) ANOVA design found a side-by-group interaction (p=0.004), no main effect of side, but a trend towards a main effect of group (p=0.086). Pairwise comparisons showed significant left < right asymmetry (p=0.004), mainly as a result of asymmetry in FTLD patients (p=0.001), but also AD patients (p=0.059, two-tailed; Figure 6.7: both panels, bottom row), and a trend for different metabolism in FTLD compared to CON bilaterally (p=0.089, two-tailed). Consequently, there was a main effect of group for the left DGH subfield (p=0.003), with left-side DGH hypometabolism in FTLD compared to CON (p=0.001).

Four-level group analysis

The 3*2*4 (subfield*side*group) ANOVA design showed a significant main effect of group (p=0.034) and a trend for side-by-group interaction (p=0.075). The 2*4 (side*group) ANOVA design for the CA1 subfield, revealed a trend for a main effect of group (p=0.087), and a side*group interaction (p=0.006). Pairwise comparisons (depicted in Figure 6.7: bottom panel) also identified an overall trend for left-right asymmetry (p=0.087, two-tailed), attributed to left < right asymmetry in SD patients (p=0.012). The subsequent four-way ANOVA design revealed a main effect of group for the left CA1 subfield, with significant left-side CA1 hypometabolism in SD compared to both CON and AD patients ($p<0.001$ and p=0.004, respectively). For the CA23 subfield, there was no main effect of side and no interactions, but a trend towards a main effect of group (p=0.058) and a left < right asymmetry found by pairwise comparisons ($p<0.001$). Two-way ANOVA designs revealed that this asymmetry was significant in both the AD and SD patients (p=0.002 and 0.019, respectively). For the left CA23 subfield only, there was a significant main effect of group (p=0.044), with significant left-side CA23 hypometabolism in SD compared to CON (p=0.032). There was a significant main ef-

fect of group (p=0.041) and a side-by-group interaction (p=0.010) for the DGH subfield, with pairwise comparions finding a left $<$ right asymmetry (p=0.001) and significant bilateral DGH hypometabolism in SD patients compared to CON (p=0.015). This asymmetry was mainly driven by SD patients (p=0.016), giving rise to a main effect of group for the left DGH subfield (p=0.001) and left-side DGH hypometabolism in SD compared to both CON and AD patients (p=0.001 and 0.012, respectively).

6.4 Discussion and conclusions

This chapter initially presented a detailed individual ROI analysis of wholefield hippocampal atrophy and FDG metabolism in both AD and FTLD patients compared to controls, including comparison of the classification accuracies of both single- and multi-modality hippocampal measures across all of these groups. In addition, novel analysis of hippocampal subfield metabolism was presented. FTLD patients were either considered as one group, or further categorized into SD and FTD subtypes, giving rise to three-level (AD, FTLD, CON) or four-level (AD, FTD, SD, CON) group analyses.

As described above, several FDG-PET studies have investigated both individual ROI and whole-brain voxel-wise analysis on the same clinical data, reporting lower sensitivity of the voxel-wise analysis for detecting hippocampal hypometabolism in MCI and mild AD patients [De Leon et al., 2001; Nestor et al., 2003; Mosconi, 2005]; thought to be due to image degradation during preprocessing (registration to a common space followed by a smoothing at 8-14mm FWHM). We therefore adopted an individual ROI analysis here; preserving data in the subjects' native space, with no image degradation, to take advantage of the higher sensititvity of such approaches to detect metabolic changes in dementia patients. Unlike previous individual ROI studies, however, we did not suffer from the burden of manual delineation of the hippocampal ROIs; instead using FIRSTv3 to automatically and accurately delineate each subject's hippocampi on the corresponding MR image. Furthermore, whilst previous FDG-PET studies have

been limited to investigations of wholefield hippocampal metabolism (primarily due to the poor spatial resolution of PET data and partial volume effects, coupled with an inability to manually identify the hippocampal subfields on the corresponding MRI scans), we provide a novel analysis of hippocampal subfield metabolism — with the higher sensitivity of the individual ROI analysis, and state-of-the-art image processing (including partial volume correction of high-resolution PET data, and shape-based mapping of a subfield atlas to the anatomically precise MR images), we were able to detect clear differences in the local metabolic profile of AD and FTLD patients. We believe that this is the first time that such measures of hippocampal subfield metabolism have been reported.

6.4.1 Hippocampal atrophy

In summary, results of wholefield hippocampal atrophy were in agreement with previous reports [Csernansky et al., 2000; Laakso et al., 2000; Chan et al., 2001; Galton et al., 2001; Van de Pol, 2006]: bilateral atrophy in all of AD, FTD, and SD patients, with an additional asymmetric (left>right) pattern of hippocampal atrophy in SD. Furthermore, this cohort of AD patients experienced symmetric (left=right) degeneration of the hippocampus compared to controls. Vertex analysis results showed subtle differences between left and right hippocampal atrophy for each patient group, and in general, the pattern of right hippocampal atrophy in AD and FTLD patients was more similar than that of the left side. Compared to the AD-associated atrophy, the FTLD group showed additional atrophy on the medial and posterior surface of the hippocampal body and tail, suggesting a differing mechanism of degeneration in these dementias.

The differences between the AD and FTLD groups were observed in the vertex analysis results prior to correction for multiple comparisons (Figure 6.4: top panel), but were unfortunately not sufficiently strong to withstand the FDR threshold. Furthermore, since our FTLD group contains both SD and FTD patients, each of which share some similarities with the pattern of hippocampal atrophy found in AD, it was not surprising to find no overriding significant dif-

ferences between the FTLD and AD groups in the vertex analysis. For the four-group scheme, effect sizes were too small to achieve significance given the size of the FTD and SD groups, and so differences are shown before correction for multiple comparisons to highlight trends that may lead to statistically significant findings in larger group studies.

Although other groups have also investigated the pattern of AD-associated hippocampal atrophy [Csernansky et al., 2000; Wang et al., 2003; Apostolova et al., 2006; Becker et al., 2006; Chételat et al., 2008b; Frisoni et al., 2008], it is difficult to make direct comparisons because of the different shape analysis techniques used, the different subfields investigated, as well as changes in scanner, imaging protocol and data cohorts. This considered, our findings are largely in accord with these previous studies; the only exception being our observation of significant atrophy in areas corresponding to the CA23 subfield in AD patients, which is in contrast to Frisoni et al. [2008].

6.4.2 Wholefield hippocampal metabolism

Despite the marked hippocampal atrophy, no significant wholefield hypometabolism was detected for the AD patients; supporting the aforementioned hypothesis that relative synaptic compensatory mechanisms may be taking place, maintaining neuronal activity in the face of structural alterations [Chételat et al., 2008a] — a very interesting prospect. Furthermore, our results suggest that this compensatory mechanism may be more active on the right side. The plots of wholefield FDG metabolism in Figure 6.2 (bottom row) showed relatively *increased* FDG uptake in the right hippocampus of AD patients compared to CON, and the previously mentioned ANOVA designs found a significant (left<right) metabolic asymmetry in AD ($p=0.026$). In recent work, Tondelli et al. [2011] observed greater and earlier atrophy of the right hippocampus in the preclinical phase of AD (i.e., 10 years before AD manifests clinically). If neuronal function is indeed maintained during this earlier right hippocampal atrophy, these results also suggest that the right hippocampus has a greater "functional reserve" in the presence of at-

rophy. Consequently, the mechanism leading to later clinical manifestation of AD may only initiate when the right hippocampal compensatory function is overwhelmed by severe atrophic damage.

In FTLD patients, significant wholefield hypometabolism was found, but this was largely driven by hippocampal dysfunction in the SD subtype, and in particular, on the left side. Consequently, measures of left hippocampal metabolism were sufficient to differentiate between the SD and AD groups, with the former having significantly reduced wholefield FDG metabolism on the left side.

6.4.3 Discriminant analyses

Results from the discriminant analyses revealed that wholefield hippocampal metabolism was the best predictor of group for AD v. SD and FTD v. SD patients, with overall classification accuracies of 77.8% and 80%, respectively, using both left and right hippocampal metabolism as predictor variables. Interestingly, in subsequent discriminant analysis, a higher overall classification accuracy of 85.2% was achieved with just left hippocampal metabolism as the predictor variable for AD v. SD patients.

In general, the discriminant analysis results showed that group classification accuracies using combined MR+PET measures were equal or better than the top single-modality measure for discriminating patients from controls, with overall accuracies ranging from 79.4% to 100% of cross-validated cases correctly classified. On the other hand, the best discrimination between different dementias was typically found using single-modality predictor variables: if SD patients were involved in the analysis, FDG metabolism always provided the best classification accuracy (e.g., AD v. FTLD, AD v. SD, FTD v. SD), whilst hippocampal volume was the best overall predictor of group for AD v. FTD patients. Although currently established from a relatively small data cohort, these results could have important future clinical implications and suggest that, when in doubt if a patient has dementia or not (i.e., to discriminate between con-

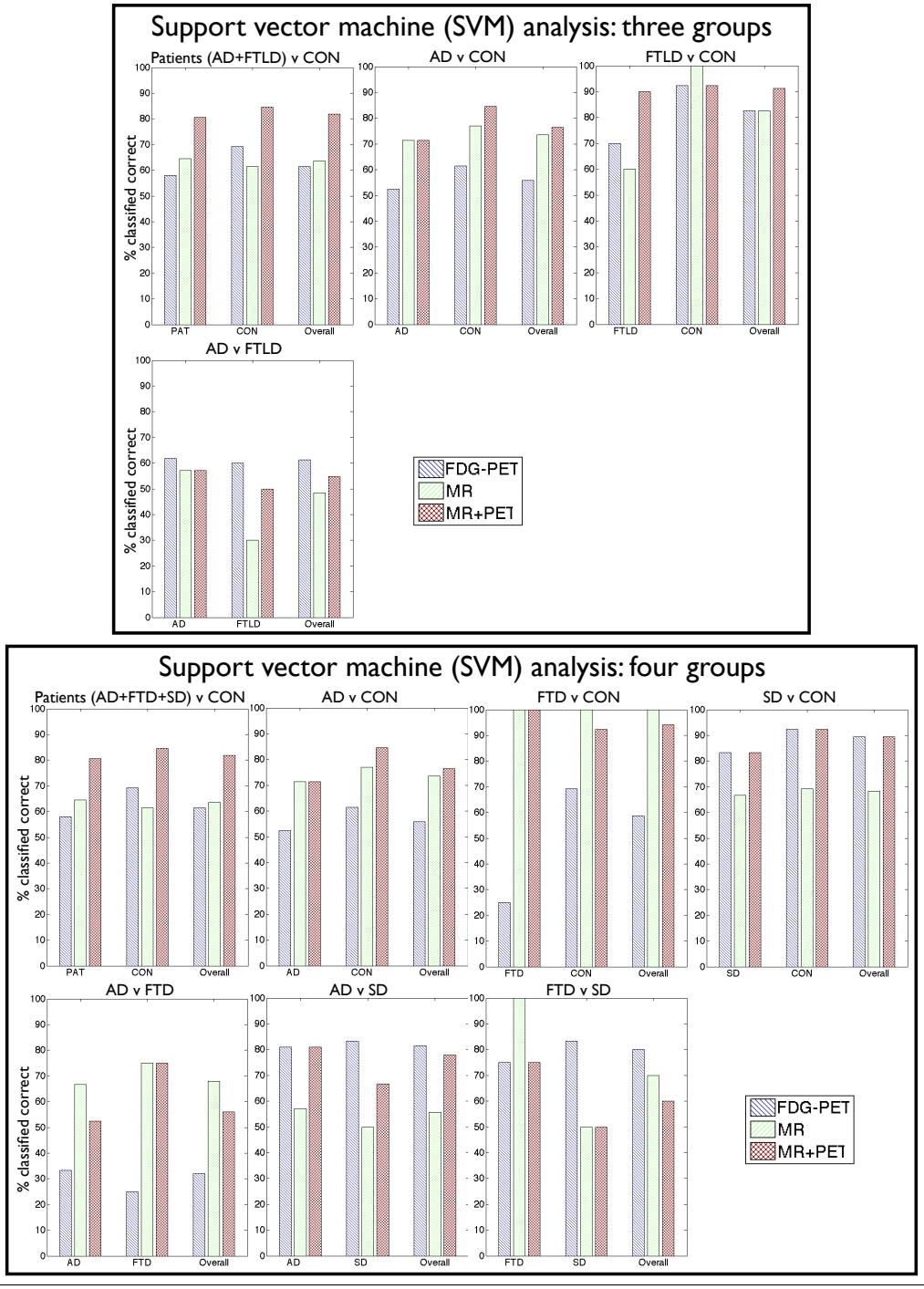
trols and dementia), combined MR+PET measures should be used, but when diagnosis of a specific neurodegenerative dementia is required (i.e., to discriminate between dementias), single modality predictors may be better.

We were certainly surprised to find that the best single-modality predictors frequently outperformed the multi-modality approach when discriminating between patient groups (in both the three- and four-group scheme) as we expected that the LDA would be able to maintain performance in the presence of additional, uninformative data. We therefore validated our results against those of another discriminant analysis method, support vector machines (SVM), implemented in MATLAB (www.mathworks.com). Once again, the best single-modality predictors frequently achieved improved discrimination between patient groups, and improved classification accuracies (shown in Figure 6.8), relative to the multi-modality approach. Furthermore, the identity of the best single-modality predictors was the same using SVM or LDA: FDG metabolism if the analysis involved SD patients (i.e., AD v. FTLD, AD v. SD, FTD v. SD); hippocampal volume for AD v. FTD. We are therefore confident that the presented LDA results are accurate for this clinical dataset. For the four-group scheme, however, we should be mindful that the discrimination between FTD and SD patients involves very low subject numbers, which may be driving the unusual behaviour of the discriminant analysis (Figure 6.5: bottom panel, bottom right plot). It would therefore be interesting to expand this work in the future using a bigger data cohort.

6.4.4 Hippocampal subfield metabolism

For the three-level hippocampal subfield analyses, left-sided hypometabolism (and left<right asymmetry) was observed for all subfields in FTLD, whilst there was selective left-sided hypometabolism (identified by a left<right asymmetry) for the CA23 subfield in AD. Subsequent four-level group analysis of subfield metabolism demonstrated that the left-sided hypometabolism (and left<right asymmetry) of all subfields in FTLD, was mainly driven by SD

Figure 6.8 *Support vector machine (SVM) analysis.* Percentage of cases (subjects) correctly classified using hippocampal FDG-PET metabolism (blue left-slanted lines, left bar of each triplet), MR volume (green right-slanted lines, middle bar of each triplet), and combined MR+PET measures (red criss-crossing lines, right bar of each triplet) for SVM training, in the three- (top panel) and four- (bottom panel) group scheme. For each plot, classification accuracy for each group (first and second triplet) is followed by the overall classification accuracy (third triplet).



patients, but there also appeared to be left-side hypometabolism in the CA23 and DGH subfields for FTD patients (whilst the CA1 subfield was relatively spared). For the FTD group, significance may not be achieved due to low subject numbers, so we recommend further investigation of these trends on a larger clinical dataset. As for the three-level group analysis, there was selective left-sided hypometabolism (identified by a left<right asymmetry) for the CA23 subfield in AD. We therefore hypothesize that the left CA23 subfield is one of the earliest sites for hippocampal metabolic dysfunction in AD — possibly a downstream effect of dysfunction in the entorhinal cortex. Furthermore, since this CA23 hypometabolism is only observed on the left side, whilst atrophy is bilateral and symmetric, these results reinforce the proposal that synaptic compensatory mechanisms may be stronger on the right side in AD patients.

6.4.5 Limitations

An obvious limitation of this study is the relatively low number of FTD and SD patients, however, the validity of the statistical analysis is maintained throughout. Furthermore, inclusion of these groups builds upon previous studies exploring relationships between hippocampal atrophy and hypometabolism in AD patients [Villain et al., 2008; Chételat, 2003; Chételat et al., 2008a; De Santi et al., 2001], and the remaining subject numbers are in accord with these studies. Additionally, this analysis of hippocampal subfield metabolism is somewhat exploratory because standard FDG-PET data processing does not allow such relatively small image features to be resolved. Findings of glucose utilization within the hippocampal subfields will therefore benefit from further validation when higher resolution methods are available. We do, however, believe that our proposed state-of-the-art, native-space processing of FDG-PET data facilitates the valid investigation of hippocampal subfield metabolism. To arrive at this conclusion, we first remember that the smallest resolvable feature in a PET image is dictated by the point spread function (PSF) of the scanner, which is typically 4-5mm (FWHM) for high-resolution FDG-PET data. Secondly, standard PET data processing involves spatial normalization to a common

reference or template space, followed by smoothing at 8-14mm FWHM — effectively introducing a 2-3 fold increase in the PSF of the PET data, and a corresponding increase in the size of the image feature than can be successfully resolved — however, the whole hippocampus is acceptably resolved with standard PET data processing. Consequently, we can infer that image features 2-3 times *smaller* than the hippocampus can be acceptably resolved with native-space analysis of PET data (i.e., without the image degradation typically associated with standard PET processing). Coupled with PVEC of the native-space PET data, we further increase the effective anatomical resolution of the underlying PET signal, such that we can resolve even smaller image features. We therefore believe that our proposed state-of-the-art, native-space ROI analysis provides sufficient resolution for investigation of the metabolic activity in the three largest hippocampal subfields (CA1, CA23, DGH), where the smallest dimension is typically 2mm.

Finally, the partial volume correction method requires homogeneous activity within the GM for total recovery of the GM concentration in the corrected FDG-PET data, which is generally not the case. Despite this, however, the algorithm does provide an exceedingly good estimate of the true concentration within the GM — simulation and phantom experiments using the original Müller-Gärtner method (with less accurate WM concentration estimates than the modified Müller-Gärtner method applied herein) have demonstrated excellent recovery of the tracer concentration in both normal neocortex and the hippocampus [Müller-Gärtner et al., 1992]. Furthermore, since an assumption of homogeneous hippocampal activity would tend to equalize any subfield differences, we can infer that any metabolic differences observed here for the hippocampal subfields will be less than would be obtained from a more advanced PVEC method (i.e., considering heterogeneous GM uptake). So whilst caution is required in the interpretation of our results of hippocampal subfield metabolism, reported subfield differences will most likely be underestimated rather than overestimated, so there is less reason to be concerned about false positive findings. To further cement our confidence in the cPET measures of hippocampal metabolism presented herein, we also provide the uncorrected measures of wholefield

hippocampal metabolism (derived from the uncorrected PET data) in Appendix C. As is evident from Figure C.1, the findings are extremely similar for the PET and cPET measures: with the exception of a slight rise in right-side metabolism for AD patients, we generally observed the same laterality and groupwise trends, but there were generally more significant metabolic differences after correction for PVE. This is in accord with the aforementioned higher sensitivity of PVEC measures for detecting hippocampal hypometabolism in amnestic MCI patients [Mevel et al., 2007].

6.4.6 Conclusion

In conclusion, this analysis of MRI and FDG-PET data for both AD and FTLD patients (including the SD and FTD subtypes) makes several distinct contributions to the field. Firstly, the subfield analysis and vertex analysis, respectively, provide novel insight into the local metabolic and atrophic hippocampal changes that occur in these dementias, the subtle differences between them, and as such, provide insight into the possible mechanistic differences between these neurodegenerative diseases. Secondly, compared to previous single-modality studies of hippocampal changes in AD and FTLD, and the few reports examining MR+PET features in AD, this work also contributes combined MR+PET analysis of both AD and FTLD; investigating the success of both single- and multi-modality hippocampal measures to discriminate dementia patients from controls and other dementias.

To fully assess the clinical utility of the aforementioned findings, follow-up studies will need to be performed. For example, with the use of gold standard, autopsy-confirmed diagnoses, it will be interesting to assess the extent to which knowledge of these hippocampal measures and predicted classifications can be used to support successful clinical diagnosis of both controls and patients. Furthermore, with longitudinal data for the control group, it will be of utmost importance to investigate whether the image processing techniques herein can help identify those controls that later progress to dementia, and to which specific dementia. It is entirely

possible to incorporate each step of our image processing procedure into a fully-automated pipeline, that only requires data (image) input to produce the predicted classifications. In this sense, a user-friendly version of the proposed methodology could be made available for the clinic, providing supplementary information for medical doctors.

We do appreciate that FDG-PET is currently available only in some imaging centers, so it may not yet be possible to clinically utilize the full complement of measures available with this procedure. Where possible, however, we believe that both structural and functional information should be utilized to fully appreciate the mechanisms at play in these complex neurodegenerative diseases. We also appreciate that single structures such as the hippocampus do not work in isolation; it is their interplay with other subcortical and cortical structures that provides the full story of degeneration. As such, improved classification of single subjects (and better prediction of future decline) will most likely be achieved by combining information across multiple structures and imaging modalities (e.g., the modalities used here plus others such as task-based and resting-state functional MRI), in addition to other indicators of neurodegeneration such as beta-amyloid burden as imaged by PET [Barthel et al., 2011; Chételat et al., 2011] or detected in CSF. We do, however, hope that this study provides useful insight into the differential mechanisms of degeneration in both AD and FTLD, to be used as a platform for future multi-modality and multi-group studies, and better diagnostic decision support for individual dementia patients.

7

Conclusions and Future Work

Here we conclude this thesis by reviewing all that has been attempted and achieved in the preceding chapters, together with our major findings, their broader implications for the field, and a discussion of the interesting projects that may subsequently stem from this work. We begin by motivating and revisiting the original goals of the thesis, before summarizing how the work undertaken in each chapter has contributed towards our success in achieving these goals. Collectively, these achievements include: development of fully-automated, accurate and robust hippocampal segmentation tools for clinical MR data; knowledge of the relative strengths and weaknesses of current automated methods (as well as the issues faced when trying to make a fair comparison of such methods); a novel label-warping approach for both generation of a consensus, hippocampal labelling protocol, and subsequent re-training of atlas- and model-based segmentation methods; and finally, a demonstration of the state-of-the-art, clinically-relevant information that can be derived from our robust and accurate hippocampal delineations. Naturally, the work herein has also uncovered issues that will need to be addressed at a later date. We outline these future work proposals in Section 7.2. Finally, we close the chapter (and the thesis) with some concluding remarks in Section 7.3.

7.1 Thesis summary and key contributions

As we have stressed throughout this thesis, the need to accurately and robustly segment the hippocampus in brain images is motivated by a clinical need to understand the degenerative and dysfunctional processes of associated neurological diseases, ultimately for more reliable patient diagnosis, monitoring, and assessment of treatment efficacy. Thus far, however, it seems that several factors have limited the widespread use of image-derived hippocampal measures in the clinic, including: (i) a scarcity of fully-automated, accurate and reliable hippocampal segmentation methods for clinical MR data; (ii) sparse knowledge of the relative performance of the automated methods that are available; and (iii) uncertainty over the value of the information that can be derived from the hippocampal segmentations (both in the research laboratories and in the clinic). These study limitations underpinned the core goals of this thesis, initially outlined in Section 1.3. Upon reflection now, we have succeeded in achieving both our main goal — developing novel hippocampal segmentation tools for reliable and precise delineation of the hippocampus in structural MR images — and our subsequent goals that focused on establishing the academic and potential clinical use of the derived hippocampal measures.

Firstly, the evaluation of several automated hippocampal segmentation methods (FIRST, FS, CFNS, CFRS and CFRS-NMI) on two clinical datasets (CMA and BPSA), allowed us to determine the relative performance of current methods and how they might be improved. For example, knowledge of the modes of failure of FIRST (namely misalignment of the query image with the reference MNI152 image before model fitting) fuelled the development of FIRSTv3, which uses a more robust, structure-specific, three-step alignment process to achieve significantly improved segmentation of the hippocampus in clinical MR data. Furthermore, by highlighting the strengths and weaknesses of the automated segmentation methods, this work allowed us to infer the pros and cons of their underlying methodology — the potential superiority of tailored, atlas-based approaches like the CF methods, but also the susceptibility of segmentation techniques to inaccuracies and possible bias in their prior information. An example of the latter was

the surprising (and somewhat concerning) output produced by FIRST when the query image was clearly misregistered with the reference (MNI152) image and the prior information therein. Due to the strong dependency of FIRST on its prior information, the algorithm proceeded to generate a seemingly reasonable hippocampal segmentation, but one which was at odds with the underlying image anatomy. If these erroneous FIRST segmentations had been evaluated by considering only their summary statistics, such as the hippocampal surface area or volume, it may have been wrongly assumed that all was well with the FIRST segmentations. We therefore reiterate the importance of visualizing the output segmentations produced by any given automated segmentation method.

Armed with the insights of the previous chapter, Chapter 4 then presented the development of FMASH: a novel region-growing method for automated segmentation of the hippocampus on structural MR data. FMASH uses a 3D Sethian Fast Marching (FM) technique to efficiently propagate a hippocampal region from an automatically-defined seed point in the MR image. The propagating hippocampal region is driven by a potential function comprised of both subject-specific intensity features and a probabilistic shape prior, so the method is more specifically identified as a combined region-growing and atlas-based approach. The novelty of FMASH arises from the automated initialization of the propagating front using a single-voxel hippocampal seed derived from the structural MR data, and the potential function itself. Thorough validation of the method, in both Chapters 4 and 5, revealed its superior performance over other methods (FIRST and FS), across a range of clinical datasets (BPSA, CMA, and ADNI), with differing image quality and demographics. We believe that the success of FMASH is due to both precise alignment of the prior information and the dependency of the method on subject-specific intensity features in the (to-be-segmented) query image. Firstly, FMASH uses the hippocampus-specific, three-stage spatial normalization protocol (originally developed for FIRSTv3 and the CF methods) for alignment of the probabilistic spatial prior to the target query image, thus providing robust and accurate initialization of the method. Secondly, when

sufficient intensity information is present, FMASH is primarily driven by subject-specific intensities in the MR image, so the method is capable of capturing image features that are unique to the particular query subject — something that is not possible with more constrained atlas- and model-based methods. Finally, the dependency of FMASH on the image-derived intensity features means that the method has an in-built mechanism to detect gross misalignment of the spatial prior and the derived hippocampal seed: if the initialization is very poor (e.g., with the seed point on or outside the boundary of the target hippocampus), then the growth of the propagating front will be restricted, and the resultant segmentation will be highly unreasonable. In contrast to FIRST, summary statistics (e.g., hippocampal surface area or volume) could inform the user of the success or failure of the automated segmentation. We will further discuss this attribute of FMASH in Section 7.2.

In the form of FIRSTv3 and FMASH, robust and accurate, automated hippocampal segmentation methods have thus been developed, and the relative performance of such methods has been established, thereby addressing both issues (i) and (ii) above. In Chapter 5, we additionally investigated the previously unknown impact of misinformed (i.e., inaccurate or biased) prior information, using both FMASH and the thresholded probabilistic prior as exemplar atlas-based segmentation methods. By establishing spatial information from one or both of two clinical datasets (CMA and ADNI), with the same demographic makeup of AD patients and controls, but with differences in data (i.e., MR scan parameters) and manual labelling protocol, we were able to calculate the detrimental effect of using either a partially-independent *Mixed* (CMA+ADNI) prior or a fully-independent cross-dataset *CMA* prior for atlas-based segmentation of another ADNI dataset, without the confounding effects of demographic differences. What we found was a strong dependency of the atlas-based segmentation methods on the data and labelling protocol used to generate their probabilistic spatial prior. On average, the fully-independent, cross-dataset *CMA* prior resulted in an 8-13% decrease in (Dice) segmentation performance compared to our reference *ADNI* prior. However, prior difference maps revealed

that the observed differences in segmentation performance were largely due to differences in the spatial distribution of the priors, and therefore differences in the manual labelling protocol used to generate them — particularly evident in the head of the hippocampus. Consequently, we proposed warping the cross-dataset *CMA* prior onto a representative sample of labels from the reference ADNI data (i.e., the *ADNI* prior) to account for differences arising from the manual labelling protocols. What we found was significantly improved segmentation performance for this *Warped* prior compared to the original, cross-dataset *CMA* prior, and comparable segmentation performance compared to the aforementioned partially-independent *Mixed* prior. This novel label-warping approach was therefore able to account for the detrimental effects of using cross-dataset priors in atlas-based segmentation, and the resultant prior information was akin to that established from direct knowledge of the reference data. Furthermore, since the original cross-dataset *CMA* prior requires no prior knowledge of the reference ADNI data, this label-warping approach allows any cross-dataset prior to be generated and then post-processed so that it conforms to a specific reference prior (and therefore a specific data and labelling standard). This work therefore contributes a means of "re-training" previous atlas- and model-based segmentation methods, ensuring that all prior information conforms to a specified standard, without the tedious and laborious task of manually re-labelling all of the training data. In doing so, this label-warping approach allows atlas- and model-based segmentation methods to be fairly evaluated without the confounding effects of differences in the manual labels used to establish their prior information. A question arises, then, over what standard or reference prior to choose. In order for all segmentation methods to be fairly evaluated against the same standard, a reference segmentation protocol must first be agreed upon; yet with so many differing datasets and manual labelling protocols, this is no simple task. Here, we propose that our label-warping approach also be used to generate a single, consensus labelling protocol, and from this, a unique reference prior for the required "re-training" of atlas- and model-based hippocampal segmentation methods. Alternatively, our label-warping procedure allows users to transform any output segmentation to their preferred protocol, without the need for additional training data. All that

would be required would be a probabilistic prior for each labelling protocol. A bank of priors could be built-up by the community and become a common resource, allowing users to take the segmentation output from various methods and transform them to their desired protocol. The potential use and contribution of our label-warping approach is thus threefold.

A particularly intriguing outcome from the *FMASH ADNI Validation* (Chapter 5: Part 1) was that all three of the evaluated segmentation methods (FMASH, FIRST, FS) overestimated the head of the hippocampus compared to the gold standard ADNI manual labels. From our newly-acquired knowledge of the spatial differences between the CMA and the ADNI manual labels (established from their prior difference maps), and the fact that all three of these automated methods use CMA data as part of their training, it is now evident that the repeatedly-observed false positive (FP) findings and the relatively poor segmentation accuracy for all three automated methods (FIRST, FS, and FMASH) was primarily due to differences in the prior information provided by the CMA and the ADNI manual labels. This observation raised the interesting point that the *perceived* performance of evaluated segmentation methods is completely dependent on the gold standard to which the resultant segmentation estimates are compared, irrespective of whether the automated segmentations or the gold standard labels are the best representation of the underlying *true* hippocampal anatomy. The fact that the segmentations provided by our three automated methods appear more representative of the underlying subject anatomy only reinforces the need for a consenus labelling protocol, the need for consistent training of atlas- and model-based hippocampal segmentation methods, and therefore, the value of our proposed label-warping approach.

For the final part of this work (in Chapter 6), we focused on addressing issue (iii) above; demonstrating how information derived from hippocampal segmentations can improve our understanding of the structural and functional hippocampal changes that occur in neurodegenerative dementias, and how this complementary information may be best utilized for improved patient diagnosis and monitoring in the clinic. From the MR image-derived hippocampal segmentations,

morphological (size and shape) properties of the hippocampus were computed, and measures of tissue glucose utilization were extracted from the corresponding FDG-PET data. Novel aspects and contributions of this work were: (i) multi-modal analysis of wholefield hippocampal measures in both AD and FTLD patients (including FTD and SD subtypes) compared to controls, including linear discriminant analysis to investigate the classification accuracies of single- and multi-modality hippocampal measures; (ii) the combination of subject-specific, native-space analysis (of all imaging data) and state-of-the-art image processing, enabling us to detect clear differences in the local (i.e., subfield) metabolic profile of AD and FTLD patients compared to controls. We believe that this is the first time that such measures of hippocampal subfield metabolism have been reported, so this work contributes both novel methodology and a more detailed understanding of the local hippocampal changes that occur in neurodegenerative dementias. Statistical analysis of the hippocampal measures revealed some intriguing results; identifying disease-related regional vulnerability and compensatory mechanisms of the human hippocampus, as well as demonstrating the superiority of multi-modal data analysis for dementia versus non-dementia diagnoses. These initial results should be confirmed in a larger clinical cohort before being incorporated into routine clinical practice, but they are nonetheless intriguing findings here, and provide a basis for future investigations.

In summary, the work presented in this thesis provides four key contributions. Firstly, our newly-developed, automated segmentation tools provide a means of accurately and robustly segmenting the hippocampus in clinical MR data, so the methods can be used with confidence in a wider range of neuroimaging studies. Secondly, our investigations provide an appreciative understanding of the difficulties faced when trying to compare and evaluate automated segmentation methods — the bias introduced to atlas- and model-based segmentation methods by their prior information, as well as the dependence of the computed performance metrics on the chosen gold standard, reference labels — thus identifying a crucial need for a universal standard labelling protocol, and in turn, fair comparison of developed automated methods. Thankfully,

this issue is addessed by the third main contribution of this work. Our proposed label-warping approach provides a means of accounting for the detrimental effects of inaccurate or biased prior information, making it ideal for numerous applications: (i) generating a single, consensus labelling protocol against which all other segmentation methods can be consistently and fairly compared; (ii) fast and efficient "re-training" of previous atlas- and model-based segmentation methods so that the spatial distribution of their training data conforms to the consensus (or indeed any other) labelling protocol, thereby eliminating such bias from their segmentations; and (iii) transforming output segmentations to any desired labelling protocol. Finally, the clinical application study shows that valuable information and insight can be derived from accurate and robust hippocampal segmentations. Collectively then, this work contributes both computational tools and valuable knowledge for further neurological investigations in both the academic and the clinical field.

7.2 Future work

In this penultimate section of the thesis, we begin by suggesting extensions and improvements to our newly-developed FMASH tool, whilst the remaining future work proposals extend from the discussion above.

7.2.1 Extensions and improvements to FMASH

Although our novel FMASH tool already combines the desirable attributes of both (intensity-driven) region growing and (knowledge-driven) atlas-based methods, and already provides relatively superior hippocampal segmentations, we believe that there are improvements to be made. For example, rather than estimating the prior information from the spatial distribution of all labels in the training data, and combining them with equal weighting to generate the probabilistic spatial prior, a type of CF approach could be implemented to select the best-matched training data (i.e., the top-ranked classifiers), with the fusion of these training samples weighted by

their match (e.g., the image similarity measure) with respect to the query image. This proposed extension to FMASH is a type of 'multi-approach' or 'combinatorial' method that combines different, desirable methodologies in order to develop a more advanced and accurate hippocampal segmentation method.

An alternative combinatorial approach for improved hippocampal segmentation would be the development of a post-processing technique that performs a known spatially-varying weighting of several segmentation estimates to arrive at the most accurate, consensus segmentation for any given query image. This spatially-varying weighting of segmentation estimates could be derived from our acquired knowledge of the spatial distribution of FP and FN findings for each of the evaluated segmentation methods. For example, we have established that (despite excellent definition of most hippocampal boundaries) FMASH has a tendency to underestimate the tail of the hippocampus, whilst other methods perform comparably well in this region. If the FMASH segmentation was down-weighted (or completely ignored) in the tail region of the hippocampus, and replaced by another, then a final, consensus segmentation could be achieved, with even better segmentation accuracy than either of the constituent segmentations alone.

Finally, we mentioned above that FMASH has an in-built mechanism to detect gross misalignment of the spatial prior and the derived hippocampal seed, because (unlike FIRST) an unreasonable segmentation is achieved when the initialization is extremely poor (e.g., with the seed point on or outside the boundary of the target hippocampus). We proposed that summary statistics (e.g., hippocampal surface area or volume) could inform the user of the success or failure of the automated segmentation. Better still, the method could be further developed to automatically detect these misalignments and rectify them. For example, if a highly unreasonable hippocampal segmentation was produced, indicating poor initial alignment of the spatial prior and resultantly poor seed point selection, the algorithm could be re-run with a shift of the spatial prior and selection of an alternative seed point.

7.2.2 Other proposals

For more general improvement of hippocampal segmentation techniques, and true acceptance of hippocampal measures in the clinical domain, we also propose the following

1. Generation of a universal, standard labelling protocol for hippocampal segmentation. An initiative (led by Dr Giovanni Frisoni) is already underway to establish a harmonized protocol for manual segmentation of the hippocampus on MRI data (www.hippocampal-protocol.net), but our label-warping method presented herein could perhaps contribute to that work.

2. Re-training of current hippocampal segmentation methods so that the spatial distribution of their training data conforms to the agreed, universal standard.

3. Enforced evaluation of all hippocampal segmentation methods against the same, universal standard.

4. Enforced reporting of a standard subset of performance metrics, to be uploaded to a freely-accessible web-based platform for direct comparison of the relative performance of differing segmentation methods.

5. Further confirmation of our findings on larger clinical datasets.

7.3 Concluding remarks

Accurate and robust hippocampal segmentation methods, together with their derived hippocampal measures, are valuable tools in our quest for better understanding of associated neurological diseases and our search for disease-related imaging biomarkers. In combination with other markers of disease, such as blood and CSF measures, these tools could ultimately improve the reliability of patient diagnosis, monitoring, and assessment of treatment efficacy in the clinic. Thus far, however, it seems that methodological inaccuracies and study limitations have pre-

vented the widespread use of hippocampal segmentation techniques in the clinical domain. In this thesis, we have developed more reliable and accurate tools for hippocampus delineation in clinical MR images. In doing so, we have highlighted some further issues associated with the training and evaluation of hippocampal segmentation methods, which has led to the additional development of a novel label-warping procedure to account for such issues. We have also shown how state-of-the-art image processing and derived hippocampal measures can be used to better understand the degenerative and dysfunctional processes of dementia, and potentially improve the clinical diagnosis of dementia patients. The results presented herein have also raised a number of interesting research questions, which motivate future investigations. By addressing the original methodological inaccuracies and study limitations, it is therefore hoped that the comprehensive investigations and computational developments in this thesis have made a valuable contribution to the field of computational neuroimaging, and in particular, to the study of hippocampal degeneration and dysfunction in neurological diseases.

A

Computation of the FM Arrival Time Surface

This appendix presents that methodology required to compute the (minimal action) arrival time surface U from Equation 4.5 of Chapter 4.

A.1 Methods

If we denote $\{A_1, A_2\}$, $\{B_1, B_2\}$ and $\{C_1, C_2\}$ the three couples of opposite neighbours surrounding our trial point (i, j, k), such that $U_{A_1} \leq U_{A_2}$, $U_{B_1} \leq U_{B_2}$, $U_{C_1} \leq U_{C_2}$ and $U_{A_1} \leq U_{B_1} \leq U_{C_1}$, the following three cases need to be sequentially examined

1. If $u \geq U_{C_1} \geq U_{B_1} \geq U_{A_1}$ the derived quadratic equation is

$$(u - U_{A_1})^2 + (u - U_{B_1})^2 + (u - U_{C_1})^2 = \tilde{P}^2. \tag{A.1}$$

   * Compute u, selecting the largest solution to equation A.1.
   * If u is indeed $> U_{C_1}$, output $u = U_{i,j,k}$.
   * If $u < U_{C_1}$, consider case 2.

2. If $u \geq U_{B_1} \geq U_{A_1}$ and $u < U_{C_1}$ the derived quadratic equation is

$$(u - U_{A_1})^2 + (u - U_{B_1})^2 = \tilde{P}^2. \tag{A.2}$$

- Compute u, selecting the largest solution to equation A.2.
- If $u > U_{B_1}$, output $u = U_{i,j,k}$.
- If $u < U_{B_1}$, consider case 3.

3. If $u \geq U_{A_1}$ and $u < U_{B_1}$ the derived quadratic equation is simply

$$(u - U_{A_1})^2 = \tilde{P}^2. \tag{A.3}$$

- Compute u, selecting the largest solution to equation A.3.
- If $u > U_{A_1}$, output $u = U_{i,j,k}$.
- If $u < U_{A_1}$, there are no solutions to the quadratic equation.

Considering each of the above cases in turn, the Matlab 'roots' function was used to return a column vector whose elements are the roots of the respective polynomial, p. The function takes as input a row vector c containing the coefficients of the polynomial, ordered in descending powers. For each of the above three cases (Equations A.1-A.3), the row vector c was constructed as follows

1. $c = [\ 3 \quad -(2U_{A_1} + 2U_{B_1} + 2U_{C_1}) \quad (U_{A_1}^2 + U_{B_1}^2 + U_{C_1}^2 - \tilde{P}^2)\ ]$
2. $c = [\ 2 \quad -(2U_{A_1} + 2U_{B_1}) \quad (U_{A_1}^2 + U_{B_1}^2 - \tilde{P}^2)\ ]$
3. $c = [\ 1 \quad -(2U_{A_1}) \quad (U_{A_1}^2 - \tilde{P}^2)\ ]$

and $U_{i,j,k}$ was computed by selecting the largest solution to the quadratic equation (i.e., the largest root).

B

Leipzig Data: Scanner and Sequence Variability

This appendix explores group-wise differences in MR scanner and sequence for the Leipzig data used in Chapter 6 (Clinical Application), and describes any effect of scanner/sequence on grey matter (GM) estimates and atrophy measures.

B.1 Materials and methods

Table B.1 shows the proportion of MR scanner and sequence for each group (CON, AD, FTLD).

Table B.1 *Distribution of MR scanner and sequence for the Leipzig data.*

	Scanner		**Sequence**	
	Bruker	Siemens	MDEFT	MP-RAGE
CON	2	11	8	5
AD	6	15	13	8
FTLD	6	4	7	3

Here, group comparisons for MR scanner and sequence were evaluated using a Chi-square test for independent samples (SPSS statistics software, version 19.0). Additionally, our collaborators at the Max-Planck-Institute in Leipzig previously evaluated scanner and sequence effects on GM volume using voxel-based morphometry.

B.2 Results

Results of the aforementioned chi-square test confirmed that there was no significant between-group difference in scanner type (χ^2(2, N=44)= 5.81; p=0.07) or sequence (χ^2(2, N=44)= 0.23; p=0.893). Our Leipzig collaborators previously observed differences in GM volume (controlling for sequence, clinical diagnosis, age and gender) between the scanners, with the Siemens scanner detecting more GM volume in the lateral and medial parietal and precuneal regions. The reverse contrast did not reveal any differences. Sequence comparisons revealed that MP-RAGE gave higher GM volume in ventral areas (inferior-temporal, occipitotemporal and rectal gyrus), mainly at the border of GM and cerebrospinal fluid and in the left posterior putamen. Lowering the significance threshold to 0.01 (uncorrected) did not reveal any differences regarding scanner in the hippocampus or parahippocampal regions.

B.3 Conclusions

These results, together with the fact that the distribution of the scanner type and sequence were balanced between groups, suggest that scanner type and sequence are unlikely to have contributed to the hippocampal results reported herein. We did not check separately for a scanner effect on glucose metabolism as the FDG-PET data were acquired on a single scanner and using the same protocol. If there are some scanner- or sequence-specific effects introduced by the partial volume effect correction (PVEC), these should have been observable in the original MRI data (on which this PVEC is based).

C

Hippocampal Metabolism from Uncorrected PET Data

This appendix presents the measures of wholefield hippocampal metabolism derived from the uncorrected Leipzig PET data and the corresponding FIRSTv3 segmentations, for comparison with the PVE-corrected PET (cPET) measures presented in Chapter 6 (Figure 6.2, bottom row).

C.1 Statistical analysis

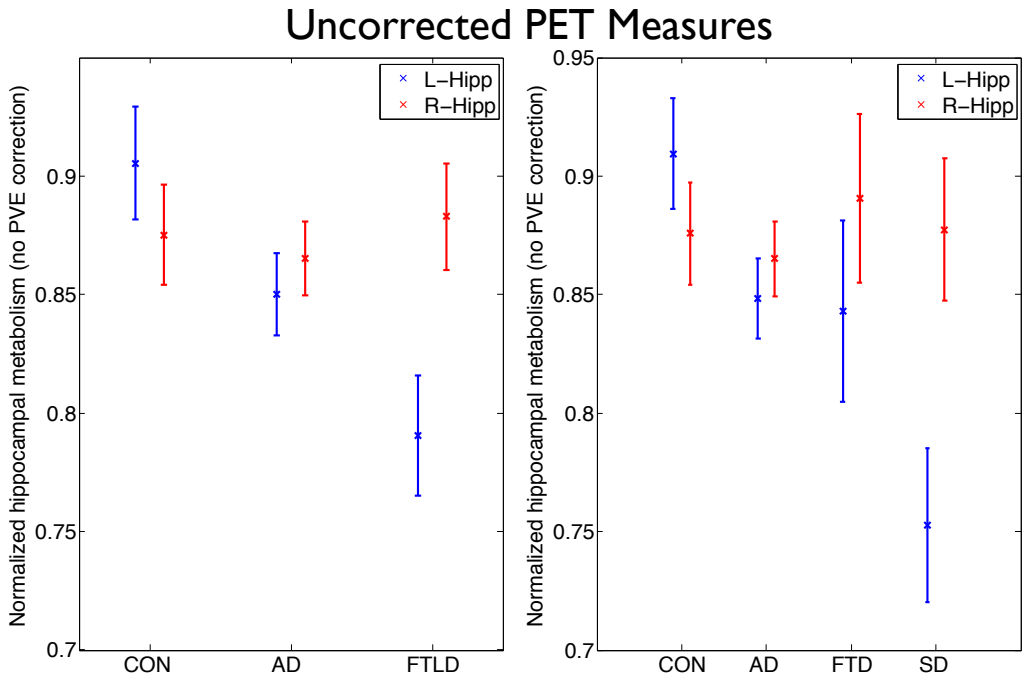
SPSS statistics software (Version 19.0) was used for all statistical analysis of uncorrected PET measures. Similar to the analysis of the cPET wholefield hippocampal measures, analysis of variance (ANOVA) was performed on the uncorrected PET measures. Side was always considered as a within-subjects factor with two levels (left, right), group/disease was either a three-level (CON, AD, FTLD) or four-level (CON, AD, SD, FTD) between-subjects factor, and age and sex were added as covariates throughout.

Initially, 2*3 and 2*4 (side*group) mixed ANOVA designs were employed to test for main effects of side or group and any side-by-group interactions. Following this, two-way ANOVA designs investigated the effect of side for each individual group, whilst three- and four-way ANOVA designs tested for a between-subjects effect of group for each side. Paired-sample t-tests with Bonferroni correction for multiple comparisons were also performed.

C.2 Results

The uncorrected PET measures of wholefield hippocampal metabolism are given in Figure C.1. Additionally, when describing the results of the ANOVA designs below, we state whether the computed statistical significance is greater than or less than that previously reported for our cPET data analysis using the terms +cPET and –cPET, respectively.

Figure C.1 *Measures of wholefield hippocampal metabolism from uncorrected PET data.* Estimated marginal means of normalized hippocampal metabolism across group, for the three- and four-level group analysis (left and right panel, respectively). Each duplet of error bars show the standard errors for the left (left) and right (right) hippocampal measures. Metabolism is normalized by cerebellar FDG uptake.



Three-level group analysis

Similar to the results of the cPET data analysis, a side-by-group interaction was found for the 2*3 (side*group) RM-ANOVA design of wholefield hippocampal metabolism (p=0.001, +cPET), and no main effects. Pairwise comparisons revealed a similar left $<$ right asymmetry (p=0.007, –cPET), driven primarily by the FTLD group (p$<$0.0005, +cPET). For the left side, there was still a significant main effect of group for the left hippocampus (p=0.011, –cPET),

with pairwise comparisons revealing left-sided hypometabolism for FTLD patients compared to CON (p=0.005, –cPET). The only real difference for the uncorrected measures of whole-hippocampus metabolism (compared to the cPET measures), was a lack of lateral asymmetry for AD patients (p=0.395, –cPET), owing to reduced right-side metabolism in AD patients (Figure C.1).

Four-level group analysis

The 2*4 (side*group) RM-ANOVA and subsequent pairwise comparisons found a side-by-group interaction ($p<0.0005$, +cPET) and left $<$ right asymmetry ($p<0.001$, same as cPET) for the uncorrected measures of wholefield hippocampal metabolism, which was similar to the aforementioned cPET analysis. However, there was no (bilateral) main effect of group (p=0.235, –cPET), and no pairwise group differences for the uncorrected measures, which constituted the major difference compared to the analysis of the cPET measures. There was, however, still a main effect of group for the left hippocampus (p=0.008, –cPET), driven by hypometabolism in SD patients compared to both CON and AD patients (p=0.002 and 0.033, respectively; both –cPET).

C.3 Conclusions

Taking the results presented herein together with those of Chapter 6 (Figure 6.2, bottom row), we conclude that there is generally little difference between the PET and cPET measures of hippocampal glucose metabolism, because we observed similar laterality and groupwise differences. The only noticeable change was a drop in right-side hippocampal metabolism for AD patients using the uncorrected PET data. In general, there was also lower sensitivity of the uncorrected measures for detecting metabolic differences, both between dementia patients and controls (in accord with previous findings [Mevel et al., 2007]), and between different dementias.

Bibliography

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *Neuroimage*, 46:726–738.

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J. V., and Rueckert, D. (2007). Classifier selection strategies for label fusion using large atlas databases. In *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 10, pages 523–531.

Andersson, J., Smith, S., and Jenkinson, M. (2008). FNIRT - FMRIB's Non-linear Image Registration Tool. *Hum Brain Mapp*, pages 15–19.

Apostolova, L. G., Dutton, R. A., Dinov, I. D., Hayashi, K. M., Toga, A. W., Cummings, J. L., and et al. (2006). Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. *Arch Neurol*, 63(5):693–699.

Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3-d point sets. *IEEE transactions on pattern analysis and machine intelligence*, 9(5):698–700.

Ashton, E., Parker, K., Berg, M., and Chen, C. (1997). A Novel Volumetric Feature Extraction Technique with Applications to MR Images. *IEEE Trans. Med. Imag.*, 16(4):365–371.

Avants, B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., and Gee, J. (2010). The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*, 49(3):2457–2466.

Barnea, D. I. and Silverman, H. F. (1972). A Class of Algorithms for Fast Digital Image Registration. *IEEE Transactions on Computers*, c-21(4):179–186.

Barnes, J., Foster, J., Boyes, R., Pepple, T., Moore, E., Schott, J., Frost, C., Scahill, R., and Fox, N. (2008). A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. *Neuroimage*, 40:1655–1671.

Barthel, H., Gertz, H.-J., Dresel, S., Peters, O., Bartenstein, P., Buerger, K., and et al. (2011). Cerebral amyloid- PET with florbetaben (18F) in patients with Alzheimer's disease and healthy controls: a multicentre phase 2 diagnostic study. *Lancet Neurol*, 10(5):424–435.

Becker, J. T., Davis, S. W., Hayashi, K. M., Meltzer, C. C., Toga, A. W., Lopez, O. L., and et al. (2006). Three-dimensional patterns of hippocampal atrophy in mild cognitive impairment. *Arch Neurol*, 63(1):97–101.

Bishop, C. A., Jenkinson, M., Andersson, J., Declerck, J., and Merhof, D. (2011). Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH): Method and validation on clinical data. *Neuroimage*, 55(3):1009–1019.

Bishop, C. A., Jenkinson, M., Declerck, J., and Merhof, D. (2010). Evaluation of hippocampal

segmentation methods for healthy and pathological subjects. In *Eurographics Workshop Vis Comput Biomed*, pages 1–8.

Bobinski, M., Wegiel, J., Tarnawski, M., Bobinski, M., Reisberg, B., De Leon, M., and et al. (1997). Relationships between regional neuronal loss and neurofibrillary changes in the hippocampal formation and duration and severity of Alzheimer disease. *J. Neuropathol. Exp. Neurol.*, 56:414–420.

Bobinski, M., Wegiel, J., Wisniewski, H., Tarnawski, M., Bobinski, M., Reisberg, B., and et al. (1996). Neurofibrillary pathology: Correlation with hippocampal formation atrophy in Alzheimer disease. *Neurobiol. Aging*, 17:909–919.

Carmichael, O., Aizenstein, H., Davis, S., J.T., B., Thompson, P., C.C., M., and Y., L. (2005). Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 27:979–990.

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., and et al. (2001). Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Ann Neurol*, 49(4):433–442.

Chen, M. (1999). *3-D Deformable Registration Using a Statistical Atlas with Applications in Medicine.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Chételat, G. (2003). Dissociating atrophy and hypometabolism impact on episodic memory in mild cognitive impairment. *Brain*, 126(9):1955–1967.

Chételat, G., Desgranges, B., de la Sayette, V., Viader, F., Eustache, F., and Baron, J. C. (2003). Mild cognitive impairment: Can FDG-PET predict who is to rapidly convert to Alzheimer's disease? *Neurol*, 60(8):1374–1377.

Chételat, G., Desgranges, B., Landeau, B., Mézenge, F., Poline, J. B., de la Sayette, V., and et al. (2008a). Direct voxel-based comparison between grey matter hypometabolism and atrophy in Alzheimer's disease. *Brain*, 131(1):60–71.

Chételat, G., Fouquet, M., Kalpouzos, G., Denghien, I., de la Sayette, V., Viader, F., and et al. (2008b). Three-dimensional surface mapping of hippocampal atrophy progression from MCI to AD and over normal aging as assessed using voxel-based morphometry. *Neuropsychologia*, 46(6):1721–1731.

Chételat, G., Villemagne, V. L., Pike, K. E., Ellis, K. A., Bourgeat, P., Jones, G., and et al. (2011). Independent contribution of temporal beta-amyloid deposition to memory decline in the pre-dementia phase of Alzheimer's disease. *Brain*, 134(3):798–807.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., and the Alzheimer's Disease Neuroimaging Initiative (2009a). Fully Automatic Hippocampus Segmentation and Classification in Alzheimer's Disease and Mild Cognitive Impairment Applied on Data from ADNI. *Hippocampus*, 19:579–587.

Chupin, M., Hammers, A., Liu, R., Colliot, O., Burdett, J., Bardinet, E., Duncan, J., Garnero, L., and Lemieux, L. (2009b). Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation. *Neuroimage*, 46:749–761.

Chupin, M., Mukuna-Bantumbakulu, A., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnehun, S., Lemieux, L., Dubois, B., and Garnero, L. (2007). Anatomically constrained region deformation for the automated segmentation of the hippocampus and amygdala: Method and validation on controls and patients with Alzheimer's disease. *Neuroimage*, 34:996–1019.

Cohen, L. and Kimmel, R. (1997). Global Minimum for Active Contour Models: A Minimal Path Approach. *Int. J. Comp. Vis.*, 24:57–78.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., and Marchal, G. (1995). Automated Multi-Modality Image Registration Based On Information Theory. In *Proc. XIVth Int. Conf. Information Processing in Medical Imaging*, pages 263–274.

Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., and Lehéricy, S. (2008). Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. *Radiology*, 248:194–201.

Csernansky, J. G., Wang, L., Joshi, S., Miller, J. P., Gado, M., Kido, D., and et al. (2000). Early DAT is distinguished from aging by high-dimensional mapping of the hippocampus. Dementia of the Alzheimer type. *Neurol*, 55(11):1636–1643.

De Leon, M. J., Convit, A., Wolf, O. T., Tarshish, C. Y., DeSanti, S., Rusinek, H., and et al. (2001). Prediction of cognitive decline in normal elderly subjects with 2-[(18)F]fluoro-2-deoxy-D-glucose/poitron-emission tomography (FDG/PET). *Proc Natl Acad Sci U S A*, 98(19):10966–10971.

De Santi, S., de Leon, M. J., Rusinek, H., Convit, A., Tarshish, C. Y., Roche, A., and et al. (2001). Hippocampal formation glucose metabolism and volume losses in MCI and AD. *Neurobiol Aging*, 22(4):529–539.

Delacourte, A., David, J. P., Sergeant, N., Buée, L., Wattez, A., Vermersch, P., and et al. (1999). The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. *Neurol*, 52(6):1158–1165.

Deschamps, T. and Cohen, L. (2000). Minimal Paths in 3D Images and Application to Virtual Endoscopy. *Lecture Notes in Computer Science*, 1843:543–557.

Diehl, J., Grimmer, T., Drzezga, A., Riemenschneider, M., Förstl, H., and Kurz, A. (2004). Cerebral metabolic patterns at early stages of frontotemporal dementia and semantic dementia. A PET study. *Neurobiol Aging*, 25(8):1051–1056.

Diehl-Schmid, J., Grimmer, T., Drzezga, A., Bornschein, S., Riemenschneider, M., Förstl, H., and et al. (2007). Decline of cerebral glucose metabolism in frontotemporal dementia: a longitudinal 18F-FDG-PET-study. *Neurobiol Aging*, 28(1):42–50.

Drzezga, A., Grimmer, T., Henriksen, G., Stangier, I., Perneczky, R., Diehl-Schmid, J., and et al. (2008). Imaging of amyloid plaques and cerebral glucose metabolism in semantic dementia and Alzheimer's disease. *Neuroimage*, 39(2):619–633.

Dubois, B., Feldman, H. H., Jacova, C., Dekosky, S. T., Barberger-Gateau, P., Cummings, J., and et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol*, 6(8):734–746.

Dukart, J., Mueller, K., Horstmann, A., Vogt, B., Frisch, S., Barthel, H., and et al. (2010). Differential effects of global and cerebellar normalization on detection and differentiation of dementia in FDG-PET studies. *Neuroimage*, 49:1490–1495.

Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., and Dale, A. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33:341–55.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-Mental State (a practical method for grading the state of patients for the clinician). *J Psychiatr Res*, 12:189–198.

Franceschi, M., Anchisi, D., Pelati, O., Zuffi, M., Matarrese, M., Moresco, R. M., and et al. (2005). Glucose metabolism and serotonin receptors in the frontotemporal lobe degeneration. *Ann Neurol*, 57(2):216–225.

Frisoni, G. B., Ganzola, R., Canu, E., Rub, U., Pizzini, F. B., Alessandrini, F., and et al. (2008). Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. *Brain*, 131(12):3266–3276.

Frisoni, G. B., Pievani, M., Testa, C., Sabattoli, F., Bresciani, L., Bonetti, M., and et al. (2007). The topography of grey matter involvement in early and late onset Alzheimer's disease. *Brain*, 130(3):720–730.

Friston, K., Ashburner, J., Frith, C., Poline, J.-B., Heather, J., and Frackowiak, R. (1995). Spatial registration and normalization of images. In *Annual Meeting of the Organization for Human Brain Mapping*, pages 165–189.

Galton, C. J., Patterson, K., Graham, K., Lambon-Ralph, M. A., Williams, G., Antoun, N., and et al. (2001). Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurol*, 57(2):216–225.

Garraux, G., Salmon, E., Degueldre, C., Lemaire, C., Laureys, S., and Franck, G. (1999). Comparison of impaired subcortico-frontal metabolic networks in normal aging, subcortico-frontal dementia, and cortical frontal dementia. *Neuroimage*, 10(2):149–162.

Geuze, E., Vermetten, E., and Bremner, J. D. (2005). MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders. *Molecular psychiatry*, 10(2):160–184.

Haller, J., Banerjee, A., Christensen, G., Gado, M., Joshi, S., Miller, M., Sheline, Y., Vannier, M., and Csernansky, J. (1997). Three-dimensional Hippocampal MR Morphometry with High-dimensional Transformation of a Neuroanatomic Atlas. *Radiology*, 202:504–510.

Hammers, A., Heckemann, R., Koepp, M., Duncan, J., Hajnal, J., Rueckert, D., and Aljabar, P. (2007). Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof-of-principle study. *Neuroimage*, 36:38–47.

Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *Neuroimage*, 33:115–126.

Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frölich, L., and et al. (2002). Discrimination between Alzheimer Dementia and Controls by Automated Analysis of Multicenter FDG PET. *Neuroimage*, 17(1):302–316.

Hogan, R. E., Mark, K. E., Wang, L., Joshi, S., Miller, M. I., and Bucholz, R. D. (2000). Mesial temporal sclerosis and temporal lobe epilepsy: MR imaging deformation-based segmentation of the hippocampus in five patients. *Radiology*, 216(1):291–297.

Ishii, K., Sasaki, M., Kitagaki, H., Yamaji, S., Sakamoto, S., Matsuda, K., and et al. (1997). Reduction of cerebellar glucose metabolism in advanced Alzheimers disease. *J Nucl Med*, 38:925–928.

Ishii, K., Sasaki, M., Yamaji, S., Sakamoto, S., Kitagaki, H., and Mori, E. (1998). Relatively preserved hippocampal glucose metabolism in mild Alzheimer's disease. *Dement Geriatr Cogn Disord*, 9(6):317–322.

Jack, C. R., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., and et al. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurol*, 51(4):993–999.

James, A., Hough, M., James, S., Burge, L., Winmill, L., Nijhawan, S., and et al. (2011). Structural brain and neuropsychometric changes associated with pediatric bipolar disorder with psychosis. *Bipolar disorders*, 13(1):16–27.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med Image Anal*, 5(2):143–156.

Jeong, Y., Cho, S. S., Park, J. M., Kang, S. J., Lee, J. S., Kang, E., and et al. (2005). 18F-FDG PET findings in frontotemporal dementia: an SPM analysis of 29 patients. *J Nucl Med*, 46(2):233–239.

Jobst, K. A., Smith, A. D., Barker, C. S., Wear, A., King, E. M., Smith, A., and et al. (1992). Association of atrophy of the medial temporal lobe with reduced blood flow in the posterior parietotemporal cortex in patients with a clinical and pathological diagnosis of Alzheimer's disease. *J Neurol Neurosurg Psychiatry*, 55(3):190–194.

Jonathan, J. and Frosta, M. (2005). Does Alzheimer's disease affect hippocampal asymmetry? Evidence from a cross-sectional and longitudinal volumetric MRI study. *Dement Geriatr Cogn Disord*.

Kelemen, A., Székely, G., and Gerig, G. (1999). Elastic Model-Based Segmentation of 3-D Neuroradiological Data Sets. *IEEE Trans. Med. Im.*, 18:829–839.

Kennedy, D., Filipek, P., and Caviness, V. (1989). Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. *IEEE Trans. Med. Imag.*, 8:1–7.

Konrad, C., Ukas, T., Nebel, C., Arolt, V., Toga, A., and Narr, K. (2009). Defining the human hippocampus in cerebral magnetic resonance images - An overview of current segmentation protocols. *Neuroimage*, 47:1185–1195.

Laakso, M. P., Frisoni, G. B., Könönen, M., Mikkonen, M., Beltramello, A., Geroldi, C., and et al. (2000). Hippocampus and entorhinal cortex in frontotemporal dementia and Alzheimer's disease: a morphometric MRI study. *Biol Psychiatry*, 47(12):1056–1063.

Leung, K., Barne, J., Ridgway, G., Bartlett, J., Clarkson, M., Macdonald, K., Schuff, N., Fox, N., Ourselin, S., and Initiative, A. D. N. (2010). Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *Neuroimage*, 51:1345–1359.

Lisy, M. E., Jarvis, K. B., DelBello, M. P., Mills, N. P., Weber, W. A., Fleck, D., and et al. (2011). Progressive neurostructural changes in adolescent and adult patients with bipolar disorder. *Bipolar disorders*, 13(4):396–405.

Lotjonen, J., Wolz, R., Koikkalainen, J., Thurfjell, L., Waldemar, G., Soininen, H., and Rueckert, D. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage*, 49(3):2352–2365.

Maroy, R., Boisgard, R., Comtat, C., Frouin, V., Cathier, P., Duchesnay, E., Dolle, F., Nielsen, P., Trebossen, R., and Tavitian, B. (2008). Segmentation of Rodent Whole-Body Dynamic

PET Images: An Unsupervised Method Based on Voxel Dynamics. *IEEE Trans Med Imaging*, 27:342–354.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurol*, 34(7):939–44.

Meguro, K., LeMestric, C., Landeau, B., Desgranges, B., Eustache, F., and Baron, J. C. (2001). Relations between hypometabolism in the posterior association neocortex and hippocampal atrophy in Alzheimer's disease: a PET/MRI correlative study. *J Neurol Neurosurg Psychiatry*, 71(3):315–321.

Mevel, K., Desgranges, B., Baron, J.-C., Landeau, B., De la Sayette, V., Viader, F., and et al. (2007). Detecting hippocampal hypometabolism in Mild Cognitive Impairment using automatic voxel-based approaches. *Neuroimage*, 37(1):18–25.

Minoshima, S., Frey, K., Foster, N., and Kuhl, D. (1995). Preserved pontine glucose metabolism in Alzheimer disease: a reference region for functional brain image (PET) analysis. *J. Comput. Assist. Tomogr.*, 19:541–547.

Minoshima, S., Giordani, B., Berent, S., Frey, K. A., Foster, N. L., and Kuhl, D. E. (1997). Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Ann Neurol*, 42(1):85–94.

Morey, R., Petty, C., Xu, Y., Hayes, J., Wagner II, H., Lewis, D., LaBar, K., Styner, M., and McCarthy, G. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage*, 45.

Morra, J., Zhuowen, T., Apostolova, L., Green, A., Avedissian, C., Madsen, S., Parikshak, N., Hua, X., Toga, A., Jack Jr., C., Weiner, M., Thompson, P., and Initiative, T. A. D. N. (2008). Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. *Neuroimage*, 43:59–68.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurol*, 43(11):2412–2414.

Mosconi, L. (2005). Reduced hippocampal metabolism in MCI and AD: Automated FDG-PET image analysis. *Neurol*, 64(11):1860–1867.

Mosconi, L., Tsui, W. H., Herholz, K., Pupi, A., Drzezga, A., Lucignani, G., and et al. (2008). Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease, and other dementias. *J Nucl Med*, 49(3):390–398.

Müller-Gärtner, H. W., Links, J. M., Prince, J. L., Bryan, R. N., McVeigh, E., Leal, J. P., and et al. (1992). Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. *J Cereb Blood Flow Metab*, 12(4):571–583.

Mummery, C. J., Patterson, K., Wise, R. J., Vandenberghe, R., Vandenbergh, R., Price, C. J., and et al. (1999). Disrupted temporal lobe connections in semantic dementia. *Brain*, 122 (Pt 1):61–73.

Narr, K. L., Thompson, P. M., Szeszko, P., Robinson, D., Jang, S., Woods, R. P., and et al. (2004). Regional specificity of hippocampal volume reductions in first-episode schizophrenia. *Neuroimage*, 21(4):1563–1575.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S., and et al. (1998). Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. pages 1546–1554.

Nestor, P. J., Fryer, T. D., Smielewski, P., and Hodges, J. R. (2003). Limbic hypometabolism in Alzheimer's disease and mild cognitive impairment. *Ann Neurol*, 54(3):343–351.

Pardoe, H., Pell, G., Abbott, D., and Jackson, G. (2009). Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? *Epilepsia*, 50:2586–2592.

Patenaude, B. (2007). *Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation.* PhD thesis, University of Oxford.

Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922.

Pitiot, A., Toga, A., and Thompson, P. (2002). Adaptive elastic segmentation of brain mri via shape-model-guided evolutionary programming. *IEEE Trans. Med. Im.*, 21.

Qiu, A., Fennema-Notestine, C., Dale, A., Miller, M., and Initiative, A. D. N. (2009). Regional shape abnormalities in mild cognitive impairment and Alzheimer's disease. *Neuroimage*, 45:656–661.

Quarantelli, M., Berkouk, K., Prinster, A., Landeau, B., Svarer, C., Balkay, L., and et al. (2004). Integrated software for the analysis of brain PET/SPECT studies with partial-volume-effect correction. *J Nucl Med*, 45(2):192–201.

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., and et al. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*.

Rizzo, G., Pasquali, P., and Gilardi, M. (1991). Multimodal Biomedical Image Registration: Use of a Cross Correlation Technique. In *Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, volume 13, pages 219–220.

Rousset, O. G., Ma, Y., and Evans, A. C. (1998). Correction for partial volume effects in PET: principle and validation. *J Nucl Med*, 39(5):904–911.

Santens, P., De Bleecker, J., Goethals, P., Strijckmans, K., Lemahieu, I., Slegers, G., and et al. (2001). Differential regional cerebral uptake of (18)F-fluoro-2-deoxy-D-glucose in Alzheimer's disease and frontotemporal dementia at initial diagnosis. *Eur. Neurol.*, 45:19–27.

Schroeter, M. L., Raczka, K., Neumann, J., and von Cramon, D. Y. (2008). Neural networks in frontotemporal dementia–a meta-analysis. *Neurobiol Aging*, 29(3):418–426.

Schroeter, M. L., Raczka, K., Neumann, J., and Yves von Cramon, D. (2007). Towards a nosology for frontotemporal lobar degenerations—A meta-analysis involving 267 subjects. *Neuroimage*, 36(3):497–510.

Schroeter, M. L., Stein, T., Maslowski, N., and Neumann, J. (2009). Neural correlates of Alzheimer's disease and mild cognitive impairment: a systematic and quantitative meta-analysis involving 1351 patients. *Neuroimage*, 47(4):1196–1206.

Sethian, J. (1996). A fast marching level set method for monotonically advancing fronts. *Proc. Natl. Acad. Sci. USA*, 93:1591–1595.

Shattuck, D., Prasad, G., Mirza, M., Narr, K., and Toga, A. (2009). Online resource for validation of brain segmentation methods. *Neuroimage*, 45:431–439.

Shi, F., Liu, B., Zhou, Y., Yu, C., and Jiang, T. (2009). Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus*, 19(11):1055–1064.

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum Brain Mapp*, 17(3):143–155.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., and et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23 Suppl 1:S208–19.

Strakowski, S. M., DelBello, M. P., Zimmerman, M. E., Getz, G. E., Mills, N. P., Ret, J., and et al. (2002). Ventricular and periventricular structural volumes in first- versus multiple-episode bipolar disorder. *The American journal of psychiatry*, 159(11):1841–1847.

Studholme, C., Hill, D., and Hawkes, D. (1998). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32:71–86.

Taylor, D. and Barrett, W. (1994). Image segmentation using globally optimal growth in three dimensions with an adaptive feature set. *Visualization in Biomedical Computing*, pages 98–107.

Thompson, P. M., Hayashi, K. M., de Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., and et al. (2004). Mapping hippocampal and ventricular change in Alzheimer disease. *Neuroimage*, 22(4):1754–1766.

Thompson, P. M., Schwartz, C., and Toga, A. W. (1996). High-resolution random mesh algorithms for creating a probabilistic 3D surface atlas of the human brain. *Neuroimage*, 3(1):19–34.

Tondelli, M., Wilcock, G. K., Nichelli, P., De Jager, C. A., Jenkinson, M., and Zamboni, G. (2011). Structural MRI changes detectable up to ten years before clinical Alzheimer's disease. *Neurobiol Aging*.

Van de Pol, L. A. (2006). Hippocampal atrophy on MRI in frontotemporal lobar degeneration and Alzheimer's disease. *J Neurol Neurosurg Psychiatry*, 77(4):439–442.

van der Lijn, F., den Heijer, T., Breteler, M., and Niessen, W. (2008). Hippocampus segmentation in MR images using atlas registration, voxel classification and graph cuts. *Neuroimage*, 43:708–720.

Villain, N., Desgranges, B., Viader, F., de la Sayette, V., Mézenge, F., Landeau, B., and et al. (2008). Relationships between Hippocampal Atrophy, White Matter Disruption, and Gray Matter Hypometabolism in Alzheimer's Disease. *J Neurosci*, 28(24):6174–6181.

Villain, N., Fouquet, M., Baron, J.-C., Mézenge, F., Landeau, B., De la Sayette, V., and et al. (2010). Sequential relationships between grey matter and white matter atrophy and brain metabolic abnormalities in early Alzheimer's disease. *Brain*, 133(11):3301–3314.

Viola, P. and Wells III, W. M. (1997). Alignment by Maximization of Mutual Information. *International Journal of Computer Vision*, 24(2):137–154.

Wang, L., Swank, J. S., Glick, I. E., Gado, M. H., Miller, M. I., and Morris, J. C. e. a. (2003). Changes in hippocampal volume and shape across time distinguish dementia of the Alzheimer type from healthy aging. *Neuroimage*, 20(2):667–682.

Watson, D. R., Anderson, J. M. E., Bai, F., Barrett, S. L., McGinnity, T. M., Mulholland, C. C., Rushe, T. M., and Cooper, S. J. (2012). A voxel based morphometry study investigating brain structural changes in first episode psychosis. *Behavioural brain research*, 227(1):91–99.

Waxman, A. D., Herholz, K., Lewis, D. H., Herscovitch, P., Minoshima, S., Ichise, M., and *et al.* (2009). *Society of Nuclear Medicine Procedure Guideline for FDG-PET Brain Imaging v1.0.* Society of Nuclear Medicine.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G., Joyce, P. R., Karam, E. G., Lee, C. K., Lellouch, J., Lépine, J. P., Newman, S. C., Rubio-Stipec, M., Wells, J. E., Wickramaratne, P. J., Wittchen, H., and Yeh, E. K. (1996). Cross-national epidemiology of major depression and bipolar disorder. *JAMA : the journal of the American Medical Association*, 276(4):293–299.

Wolz, R., Aljabar, P., Hajnal, J., Hammers, A., Rueckert, D., and the Alzheimer's Disease Neuroimaging Initiative (2010). LEAP: Learning embeddings for atlas propagation. *Neuroimage*, 49:1316–1325.

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C. (1998). Automated image registration: I. General methods and intrasubject, intramodality validation. *Journal of computer assisted tomography*, 22(1):139–152.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., and et al. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, 45(1 Suppl):S173–86.

Xu, Y., Valentino, D., Scher, A., Dinov, I., White, L., Thompson, P., Launer, L., and Toga, A. (2008). Age effects on hippocampal structural changes in old men: The HAAS. *Neuroimage*, 40:1003–1015.

Yakushev, I., Landvogt, C., Buchholz, H., Fellgiebel, A., Hammers, A., Scheurich, A., and et al. (2008). Choice of reference area in studies of Alzheimers disease using positron emission tomography with fluorodeoxyglucose-F18. *Psychiatry Res*, 164:143–153.

Yang, J. and Duncan, J. (2004). 3D image segmentation of deformable objects with joint shape-intensity prior models using level sets. *Med. Image Anal.*, 8:285–294.

Yang, J., Staib, L., and Duncan, J. (2004). Neighbor-constrained segmentation with 3D deformable models. *IEEE Trans. Med. Imag.*, 23:940–948.

Yushkevich, P. A., Avants, B. B., Pluta, J., Das, S., Minkoff, D., Mechanic-Hamilton, D., and et al. (2009). A high-resolution computational atlas of the human hippocampus from postmortem magnetic resonance imaging at 9.4 T. *Neuroimage*, 44(2):385–398.