Statistical Methods for Analysis of Structural Magnetic Resonance Imaging in Patients with Multiple Sclerosis

by

Elizabeth M. Sweeney

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

February 22, 2016

Abstract

Multiple sclerosis (MS) is an inflammatory disease of the brain and spinal cord characterized by demyelinating lesions. Structural magnetic resonance imaging (sMRI) is a medical imaging technique that is sensitive to these lesions. Quantitive analyses of MRI, such as the number and volume of MS lesions, are essential for diagnosing the disease and monitoring its progression. In addition, the formation of these lesions, a complex process involving inflammation, tissue damage, and repair, is also important for diagnosing and monitoring the disease. While sMRI is sensitive to lesion activity, there is surprisingly poor association between clinical findings and the radiological extent of involvement on MRI using traditional volumetric measures. This phenomenon is referred to as the clinico-radiological paradox.

The work in this thesis is an effort to bridge this clinico-radiological paradox and link the longitudinal findings on structural MRI in patients with MS to disease-modifying treatment and other clinical information. Chapter 2 of the thesis is an introduction to sMRI data. Chapter 3 and 4 of the thesis deal with MS lesion segmentation using multi-sequence structural MRI. Chapter 5 is a culmination of this work. The lesion segmentation technique explored in Chapter 3 and 4 is extended to build a pipeline to extract longitudinal intensity information, or lesion profiles, from lesions in multi-sequence sMRI. A PCA regression model is then introduced to relate the longitudinal lesion profiles to disease-modifying treatment and other clinical information in an attempt to link the information from sMRI to clinical information. In addressing these clinical issue, this thesis also contains a number of biostatistical contributions:

the design and analysis of expert rater trials, data reduction techniques for high dimensional and longitudinal data through principal component analysis (PCA) regression models, and the comparison of supervised learning algorithms.

Thesis Committee

Primary Readers

Peter Calabresi
Director, Division of Neuroimmunology,
Professor of Neurology,
Johns Hopkins School of Medicine

Michelle Carlson
Associate Professor of Mental Health,
Johns Hopkins Bloomberg School of Public Health

Ciprian Crainiceanu (Advisor)
Professor of Biostatistics,
Johns Hopkins Bloomberg School of Public Health

Russell Shinohara (Co-advisor)
Assistant Professor of Biostatistics,
Perelman School of Medicine,
University of Pennsylvania

Alternate Readers

Brian Caffo
Professor of Biostatistics,
Johns Hopkins Bloomberg School of Public Health

Alden Gross
Assistant Professor of Epidemiology,
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I would first like to thank the members of my thesis committee, Dr. Michelle Carlson and Dr. Peter Calabresi, as well as my thesis committee alternates, Dr. Brian Caffo and Dr. Alden Gross. Thank you for taking the time to read my thesis and for your service on my committee. Thank you also to the wonderful coauthors from this thesis (without whom this work would not exist!): my advisor Dr. Ciprian Crainiceanu, my co-advisor Dr. Russell (Taki) Shinohara, Dr. Daniel Reich, Dr. Ani Eloyan, Dr. Dzung Pham, Dr. Navid Shiee, Dr. Farrah Mateen, Dr. Peter Calabresi, Dr. Joshua Vogelstein, Dr. Avni Chudgar, Blake Dewey, Matthew Schindler, and Jennifer Cuzzocreo. Also that you to the Epidemiology and Biostatistics of Aging Training Program that supported my thesis work and my mentors from this program Dr. Karen Bandeen-Roche and Dr. Michelle Carlson.

Thank you to my advisor Ciprian Crainiceanu for encouraging and motivating me to achieve more than I ever thought possible for myself. Working with you for the last five years has been a truly amazing experience. I still remember the day when you suggested that I should reapply to the PhD program. I left your office completely overwhelmed and thought it would be impossible for me to earn a PhD, but you guided me through the entire process. The caring, compassion, and understanding that you gave me during the hardest times of the PhD were above all what made the difference this time. Thank you for always believing in me, investing so much in me, and discovering and nurturing the potential that I could not see in myself. What you have done has truly changed my life.

Thank you to my co-advisor Taki Shinohara for your patience and being a constant source of inspiration and support. Thank you for being there every time I need you, whether it is to talk about the future, to help when I get into a research rut, or just to grab coffee and catch up. Thank you for shaping my research philosophy, for sharing your research ideas with me, and for helping me develop my own. One day we will finish all of the items on our research to-do list and then we will turn around and make more! Thank you for hosting me so many times at Penn. These trips were always where I did my best work (and ate the best food!). Working with you is so much fun and I am so thankful for your mentorship and friendship.

Thank you to Danny Reich for igniting my passion for imaging research and continuing to do so for the last five years. Working with you is extraordinary and every time I go to the NIH to visit your group I leave newly inspired. It is a privilege to work with you and I thank you for your amazing collaboration and mentorship

Thank you to the wonderful group of friends that I made while at Hopkins. You have been a continuous source of fun, support, and encouragement. A special thank you to the friends I also had the pleasure to work with during my PhD: Mandy Mejia, Gina-Maria Pomann, and Jean-Philippe Fortin. Mandy, thank you for being my conference chica, for turning me into a coffee snob, and for the long, sweaty walks into school last summer. Thank you for your amazing friendship and for always being so supportive and positive. Gina, meeting you made this whole graduate school thing worth the trouble. Thank you for making research ridiculously fun (and I mean ridiculously fun...even when it probably should not have been). Jean-Philippe, thank you for being a coding animal with me. You, wooden hand, and the boos really understand who I am. It is a rare thing to find in this life, especially for people as strange as the five of us. Thank you for your friendship and for sharing your time in Baltimore with me. I will miss you.

A very special thank you to my friend and colleague John Muschelli, who I have worked closely with on everything. I looked forward to coming into school everyday so that I could hang out, eat magic beans, and do research with you. Thank you for always being there to talk about the "little details" that end up being the most important parts of the work. Thank you for not killing me while we were working on the ENAR tutorial (I know I almost killed you...). Thank you for telling me to shut up when I was whining and calling me to come into school and work on the days I wanted to do anything but. You helped me through the hardest and ugliest parts of the PhD and for this I will always be thankful. I will miss you the most of all the things at Hopkins.

Most importantly, thank you to my family for their unconditional love and support. Adam, I know you are not an official part of my family, but I have considered you a brother for some time now. Mom, Dad, Kelly, and Adam this thesis is dedicated to you. Thank you for understanding how important this is to me and all of the things that I had to give up to make this possible. Adam, thank you for still loving me after THAT trip to Baltimore. Thank you all for dropping everything each time that I needed you. I could not have done this without any of you. You are the four reasons for my success and I love you.

Table of Contents

Table of Contents **vi**

List of Tables **x**

List of Figures **xii**

1 Introduction **1**

2 A tutorial for multi-sequence clinical structural brain MRI **5**

2.1 Introduction 5

2.1.1 What are these images? 10

2.1.2 How can we handle sMRI? 11

2.1.3 What are some major pitfalls when starting working on sMRI? 13

2.2 Open scientific problems associated with sMRI of the diseased brain 15

2.3 Data structure and intuitive description of associated problems. 18

2.4 Acquisition and reconstruction 20

2.5 Pre-processing 21

2.5.1 Inhomogeneity correction 23
2.5.1.1 Concepts 24
2.5.1.2 Practical approaches, software, and application to data 26
2.5.2 Skull stripping 26
2.5.2.1 Concepts 27
2.5.2.2 Practical approaches, software, and application to data 28
2.5.3 Interpolation 28
2.5.3.1 Concepts 29
2.5.3.2 Practical approaches, software, and application to data 30
2.5.4 Spatial Registration 30
2.5.4.1 Concepts 31
2.5.4.2 Practical approaches, software, and application to data 36
2.5.5 Intensity normalization 41
2.5.5.1 Concepts 42
2.5.5.2 Practical approaches, software, and application to data 43
2.6 Analysis 46
2.6.1 Lesion segmentation 46
2.6.2 Lesion mapping 49
2.6.3 Longitudinal and cross sectional intensity analysis 51
2.7 Conclusions 54

3 OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI 55

3.1 Introduction 55

3.2 Materials and methods 59

3.2.1 Study population 60

3.2.2 Experimental methods 61

3.2.3 Image preprocessing 62

3.2.4 Statistical modeling and spatial smoothing 62

3.2.5 Brain tissue mask 63

3.2.6 Intensity normalization 63

3.2.7 Smoothed volumes 65

3.2.8 OASIS is Automated Statistical Inference for Segmentation 67

3.2.9 OASIS model refinement 70

3.2.10 Probability map and binary segmentation 71

3.2.11 Validation with gold standard: validation set 1 72

3.2.12 Validation with expert rankings: validation set 1 and validation set 2 74

3.3 Experimental results 77

3.3.1 Validation set 1: training with gold standard 77

3.3.2 Validation set 1: neuroradiologist rating results 79

3.3.3 Validation set 2: neuroradiologist rating results 80

3.3.4 Validation set 2: rater agreement with neuroradiologist, neurologist, and radiologist 81

3.4 Discussion 85

4 A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI 92

4.1 Introduction 92

4.2 Experimental methods 95

4.2.1 Study population and experimental methods 95

4.2.2 Image preprocessing 95

4.3 Statistical methods 96

4.3.1 Supervised classification algorithms 97

4.3.2 Feature extraction functions and vectors 101

4.3.2.1 Unnormalized 101

4.3.2.2 Normalized 101

4.3.2.3 Voxel selection 102

4.3.2.4 Smoothed 104

4.3.2.5 Moments 104

4.3.2.6 Smoothed and moments 105

4.3.3 Training and validation studies with manual lesion segmentations 105

4.3.4 Measures of outcome and agreement 108

4.4 Results 108

4.4.1 Classification performance 108

4.4.2 Algorithm agreement 112

4.4.3 Computational time 115

4.4.4 Downsampling the training set: classification performance and computational time 117

4.4.5 Super learner coefficients 118
4.4.6 Interpretability 120
4.5 Discussion 122

5 Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions 128

5.1 Introduction 128
5.2 Material and methods 131
5.2.1 Image acquisition and preprocessing 131
5.2.2 Patient demographics 132
5.2.3 Lesion profiles 133
5.2.4 Principal component analysis and regressions 136
5.2.4.1 Biomarker 137
5.2.4.2 Expert validation of biomarker 138
5.2.4.3 Regression model 139
5.2.5 Function-on-scalar regressions 142
5.3 Results 143
5.3.1 Principal component analysis and regression 143
5.3.1.1 Biomarker 143
5.3.1.2 Expert validation of biomarker 144
5.3.1.3 Regression model 152
5.3.2 Function-on-scalar regression 154
5.4 Discussion 155

Appendices 160

A Appendix to "OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI 161

A.1 OASIS coefficients 161

A.2 Repeated subjects for validation study 1 166

B Appendix to "A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI" 167

B.1 Classification algorithms 167

B.1.1 Logistic regression 168

B.1.2 Linear discriminant analysis 168

B.1.3 Quadratic discriminant analysis 169

B.1.4 Gaussian mixture model 169

B.1.5 Support vector machine with linear kernel 169

B.1.6 Random forest 170

B.1.7 k-nearest neighbors 170

B.1.8 Neural network 170

B.1.9 Super learner 171

C Appendix to "Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions" 172

C.1 Longitudinal profile pipeline 172

C.1.1 Identifying voxels with new lesion formation 173

C.1.2 Intensity normalization 174

C.1.3 Temporal alignment 176

C.1.4 Temporal interpolation 176

C.2 Parametric bootstrapping procedure 177

C.3 Expert validation 178

C.4 Principal component analysis and regression 179

C.5 Function-on-scalar regression 186

List of Tables

3.1 Ranges for validation set 2 scanning parameters 62
3.2 Quantitative validation results for validation set 1 79
3.3 Volume of false positives for validation set 1 79
3.4 Qualitative validation results for validation set 1 80
3.5 Qualitative validation results for validation set 2 83

4.1 Supervised classification algorithms in R 99
4.2 Feature vectors 100
4.3 The training and validation sets 107

5.1 κ coefficients for the ratings of the lesion segmentation and the biomarker. 151

A.1 Coefficients from the first fit 164
A.2 Coefficients from the second fit 165
A.3 Preferred method for the 5 repeated subjects 166

C.1 Coefficients from multivariate PCA regression model 184
C.2 Coefficients from univariate PCA regression model 185

List of Figures

2.1 Multi-sequence MRI data for one subject 8

2.2 Dynamic contrast enhancing (DCE) volume 9

2.3 Inhomogeneity correction of volume 24

2.4 Skull stripping using BET 28

2.5 Steps of registration: a toy example 38

2.6 Nonlinear registration to a template 40

2.7 Intensity normalization methods 44

2.8 Multiple sclerosis lesion segmentation 47

2.9 Multiple sclerosis lesion histograms 51

2.10 Longitudinal intensities within multiple sclerosis lesions 52

3.1 Example of the MRI data 64

3.2 Smoothed MRI volumes 66

3.3 Two stage smoothing of MRI volumes 72

3.4 Partial ROC curve for validation set 1 78

3.5 Boxplot of segmentation ratings 81

3.6 Cerebellum lesion 89

4.1 Scatter plots of the T1-w, T2-w and FLAIR voxel intensities . . 102

4.2 Example of the MRI data 103

4.3 The partial ROC curves for the classification algorithms 109

4.4 The scaled partial AUC for each algorithm 111

4.5 Dice similarity coefficients for all algorithms 114

4.6 Time to fit and make predictions from each algorithm 116

4.7 Downsampling of the training set 119

4.8 Super learner coefficients 121

5.1 Longitudinal MRI studies 134

5.2 Longitudinal MRI studies within lesions 147

5.3 Multi-sequence lesion profiles 148

5.4 Mean profile and first PC 149

5.5 Distributions of the ratings for the two raters 150

5.6 Coefficients from the PCA Regression model 152

5.7 FLAIR coefficient functions from the function-on-scalar regression 155

C.1 SuBLIME and OASIS segmentations 175

C.2 Passed 180

C.3 Passed with minor errors 181

C.4 Some redeeming features 182

C.5 Failed miserably 183

C.6 T2 coefficient functions from the function-on-scalar regression . 187

C.7 PD coefficient functions from the function-on-scalar regression . 188

C.8 T1 coefficient functions from the function-on-scalar regression . 189

Chapter 1

Introduction

Multiple sclerosis (MS) is an inflammatory disease of the brain and spinal cord characterized by demyelinating lesions. Structural magnetic resonance imaging (sMRI) is a medical imaging technique that is sensitive to these lesions [1]. Quantitive analyses of MRI, such as the number and volume of MS lesions, are essential for diagnosing the disease and monitoring its progression [2, 3]. In addition, the formation of these lesions, a complex process involving inflammation, tissue damage, and repair, is also important for diagnosing and monitoring the disease [4]. While sMRI is sensitive to lesion activity, there is surprisingly poor association between clinical findings and the radiological extent of involvement on MRI using traditional volumetric measures. This phenomenon is referred to as the clinico-radiological paradox [5].

The work in this thesis is an effort to bridge this clinico-radiological paradox and link the longitudinal findings on structural MRI in patients with MS to disease-modifying treatment and other clinical information. Chapter 2 of the thesis is an introduction to sMRI data. Chapter 3 and 4 of the thesis deal with MS lesion segmentation using multi-sequence structural MRI. Chapter 5 is a culmination of this work. The lesion segmentation technique explored in Chapter 3 and 4 is extended to build a pipeline to extract longitudinal intensity information, or lesion profiles, from lesions in multi-sequence sMRI. A PCA regression model is then introduced to relate the longitudinal lesion profiles to disease-modifying treatment and other clinical information in an attempt to link the information from sMRI to clinical information. In addressing these clinical issue, this thesis also contains a number of biostatistical contributions: the design and analysis of expert rater trials, data reduction techniques for high dimensional and longitudinal data through principal component analysis (PCA) regression models, and the comparison of supervised learning algorithms.

Chapter 2 of this thesis is an introduction to sMRI data, and includes detailed instructions and code for the preprocessing and analysis of sMRI data. With my advisor Dr. Ciprian Crainiceanu, co-advisor Dr. Russell Shinohara, and Dr. Ani Eloyan, this chapter has been developed into a book chapter in the Handbook of Modern Statistical Methods: Neuroimaging Data Analysis, entitled "A Tutorial for Multi- sequence Clinical Structural Brain MRI" [6]. As part of sMRI analysis education, I also co-developed and taught tutorials on the topic at the Statistics and Applied Mathematical Science Institute (SAMSI) and the Eastern North American Region (ENAR) meeting. More recently I co-developed a Coursera course with Dr. Crainiceanu and fellow PhD student John Muschelli on structural MRI analysis and preprocessing using the statistical software R, that will be released soon.

The majority of my work for this thesis has been dedicated to MS lesions segmentation using multi-sequence sMRI, which is covered in Chapters 3 and

4. While many methods for lesion segmentation exist in the literature, these methods are often difficult to reproduce and do not have publicly available software implementations. Chapter 3 introduces OASIS is Automated Statistical Inference for Segmentation (OASIS), a fully automated, cross-sectional lesion segmentation algorithm [7]. OASIS uses a logistic regression model to create probability maps of lesion presence for multi-sequence MRI studies. In this chapter, I present the OASIS method and perform extensive validation of the method, showing increased performance of OASIS over the previous state of the art lesion segmentation method. The validation of the OASIS method includes qualitative validation, using expert rater trials with a neurologist, neuroradiologist, and radiologist. The OASIS lesion segmentation algorithm is available for public use and is currently implemented as an R package (https://cran.r-project.org/web/packages/oasis/index.html). Chapter 4 compares the logistic regression model from the OASIS algorithm against a number of supervised machine learning algorithms [8]. Here I found that the logistic regression model has performance as well or better than the more complex machine learning models and this work illustrates the importance of model interpretation and parsimony.

Chapter 5 describes the culmination of this thesis work, an attempt to bridge the clinico-radiological paradox. Using an algorithm I previously developed for segmenting new and enlarging MS lesions [9], along with the OASIS algorithm, I developed a pipeline for extracting and normalizing lesion profiles, the longitudinal voxel-level intensities on the multi-sequence sMRI within MS lesions. Using PCA to reduce the dimension of the lesion profiles, I discovered that the score on the first PC is a voxel-level biomarker of lesion repair. The marker is validated by two clinicians in an expert rater trial. The relationship between this biomarker and the clinical measures of interest is modeled using a PCA regression model, and a statistically significant relationship between the biomarker and the use of disease-modifying treatment, steroids, and distance to the boundary of a lesion was found [10].

Chapter 2

A tutorial for multi-sequence clinical structural brain MRI

2.1 Introduction

High resolution structural magnetic resonance imaging (sMRI) is used extensively in clinical practice, as it provides detailed anatomical information of the living organism, is sensitive to many pathologies, and assists in the diagnosis of disease [11]. Applications of sMRI cover essentially every part of the human body from toes to brain and a wide variety of diseases from stroke, cancer, and multiple sclerosis (MS) to internal bleeding and torn ligaments. Since the introduction of MRI in the 1980s, the lack of side effects, the continuously improving resolution of images, and the wide availability of MRI scanners have made sMRI instantly recognizable in the popular literature [12]. Indeed, when one is asked to have an MRI in a clinical context it is almost certainly an sMRI or its close relative, the dynamic contrast enhancing MRI (DCE-MRI) . These images are fundamentally different from functional MRI (fMRI) in size, complexity, measurement target, type of measurement, and intended use. While fMRI aims to study brain activity, sMRI reveals anatomical information. This distinction is important as the scientific problems and statistical techniques for fMRI and sMRI analysis differ greatly [13], yet confusion between the two continues to exist in the literature and among reviewers. Despite the enormous practical importance of sMRI, few statisticians and biostatisticians have made research contributions in this field. This may be due to the subtle aspects of sMRI, the relatively flat learning curve, and the lack of contact between statisticians and biostatisticians and the scientists working in clinical neuroimaging. Our goal is reduce the price of entry, accelerate learning, and provide the information required to progress from novice to initiated sMRI researcher.

This chapter is designed to provide a tutorial for sMRI research, introduce some major unsolved scientific problems in brain imaging of patients with neurological disease, and describe the important technical problems associated with data analysis. Image acquisition and pre-processing, especially as it relates to pre-processing pipelines, will also be discussed. In our experience, it has been impossible to seperate the image pre-processing pipeline from later analysis. The paper is accompanied by sMRI for two subjects with multiple sclerosis at two visits together with the associated R code that can be used to open, visualize, and conduct small statistical analyses. These studies have been pre-processed using the pre-processing steps outlined in this chapter.

An sMRI study typically consists several different sMRI sequences, most commonly the T1-weighted (T1), T2-weighted (T2), Fluid Attenuated Inversion Recovery (FLAIR), and Proton Density (PD). Other sequences are continuously being researched and may become standard in future sMRI studies. Moreover, the type of magnet (1.5T, 3T, or 7T), the brand of MRI scanner, and the choice of scanning parameters may induce major differences between images, even if they are of the same sequence. We refer to the sMRI collection of two or more sequences as multi-sequence sMRI. We will distinguish multi-sequence sMRI from multi-modality imaging, which refers to the combination of at least two different types of imaging, for example sMRI and Computed Tomography (CT), or CT and Positron Emitted Tomography (PET).

From a data perspective, every sequence is a three dimensional array, with each entry representing a voxel, or three dimensional pixel. The size of the voxels depends on the acquisition parameters and provides the resolution of the image. Figure 2.1 displays data from a standard sMRI sequence protocol for three slices shown in the three rows. The voxel size for these images has been interpolated to $1 \times 1 \times 1$ mm (interpolation of sMRI is discussed in detail in Section 2.5.3). Slices are displayed, moving from the inferior to the superior of the brain and are labeled A, B, and C, respectively. Each column corresponds to a different sequence – the FLAIR (A1, B1, C1), T2 (A2, B2, C2), T1 (A3, B3, C3), and PD (A4, B4, C4). An intuitive way to think about the different sequences is that they are slices through the brain seen through different filters. Making such plots in `R` [14] is relatively easy using the `oro.nifti` `R` package [15] package. After setting the working directory to the location of the compressed FLAIR volume, the following lines of code will load the volume and plot one axial slice of the FLAIR image:

```
library(oro.nifti)
flair <- readNIfTI('FLAIRnorm.nii.gz', reorient=TRUE)
```

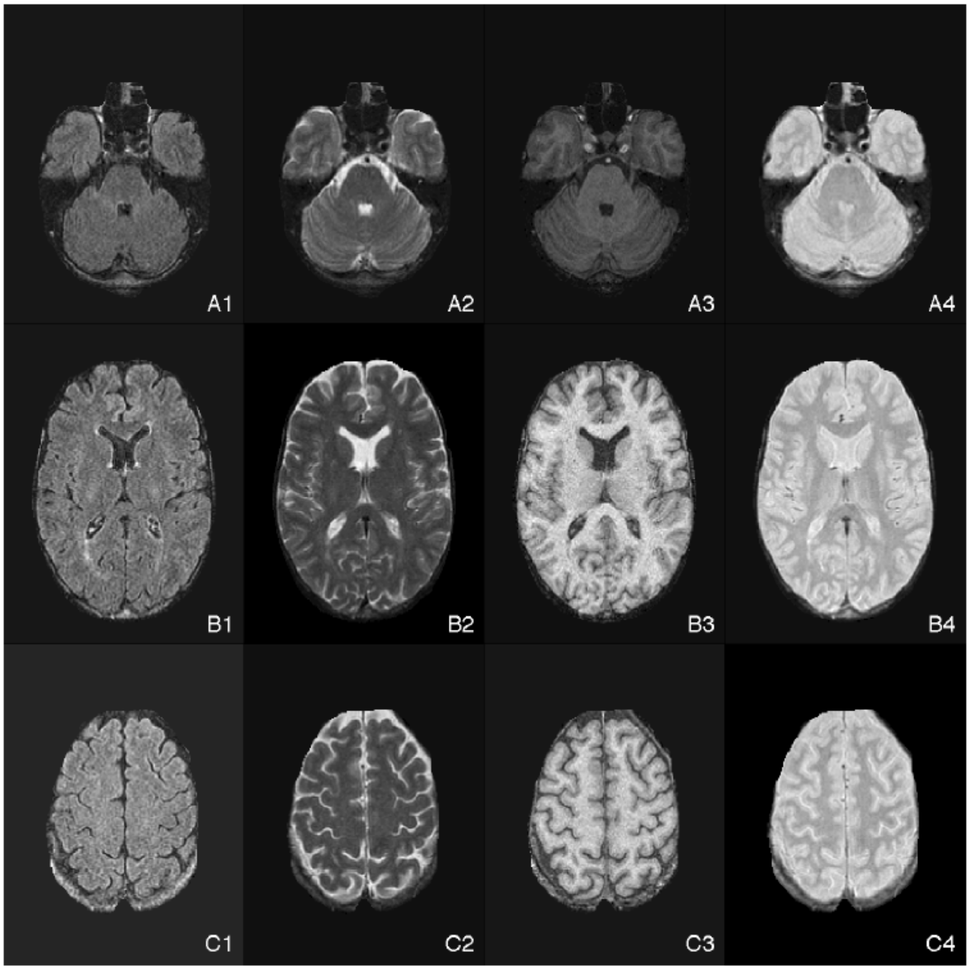


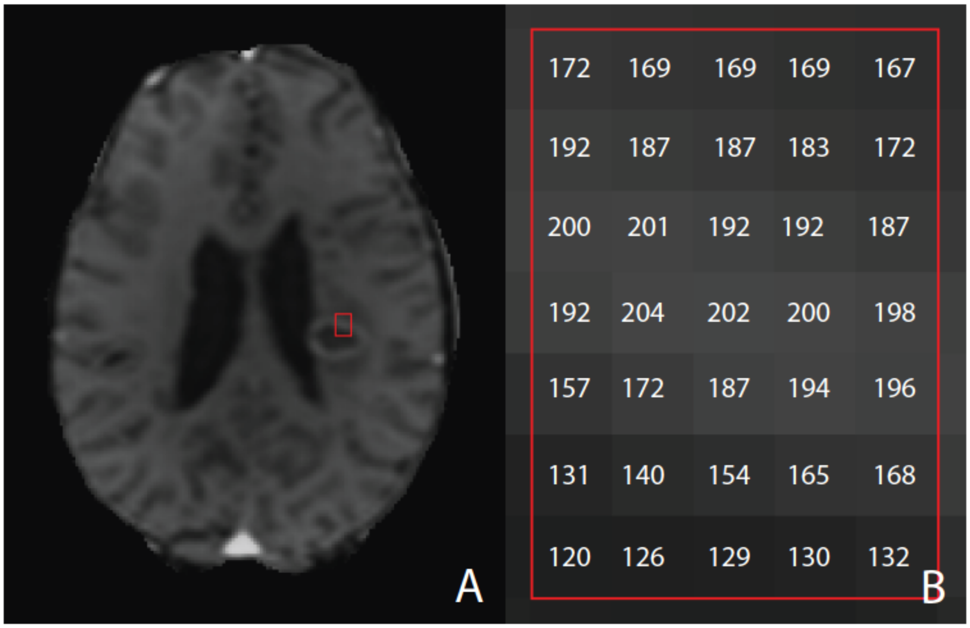
Figure 2.1: Multi-sequence MRI data for one subject. Three axial slices are shown on each row (letters A, B, C indicate a different slice going from the inferior (A) to superior (C) of the brain) indicating FLAIR (A1, B1, C1), T2 (A2, B2, C2), T1 (A3, B3, C3), and PD (A4, B4, C4). A small MS lesion is visible n the A-slice images. Some larger MS lesions are visible closer to the ventricle in the B-slice images.

```
image(flair[,,50])
```

A plot of the sagittal, coronal and axial view from the slices [111, 132, 102] ([sagital slice, coronal slice, axial slice]) can be obtained using the command `orthographic`:

```
orthographic(flair, xyz = c(111, 132, 102))
```

While R packages may change, improve, or become obsolete, we currently like the `oro.nifti` R package, because it is relatively easy to use and allows us to work directly with compressed files. This is a big advantage when working on large studies and/or transferring files. Once the magic of staring of the pictures is gone, some important technical questions remain. Most importantly: 1) what are these images?; 2) how can we handle sMRI?; and 3) what are some major pitfalls when starting working on sMRI? We are now addressing these questions.



172	169	169	169	167
192	187	187	183	172
200	201	192	192	187
192	204	202	200	198
157	172	187	194	196
131	140	154	165	168
120	126	129	130	132

Figure 2.2: A. Dynamic contrast enhancing (DCE) volume after gadolinium injection. B. The numerical data obtained from the red region of interest in (A) from this volume.

2.1.1 What are these images?

At the most basic level, every sMRI volume is a 3 dimensional (3D) array, with dimensions determined according to the acquisition parameters. For example, the FLAIR volume shown in Figure 2.1 is stored as a 3D array and the 50^{th} axial slice (moving from the inferior to the superior of the brain) is stored in `flair[,,50]`. This FLAIR image is interpolated to a voxel size of $1 \times 1 \times 1$mm and is $182 \times 218 \times 182$ voxels, or about 7 million voxels. As shown in Figure 2.1 this MRI study contains 4 sequences, for a total of around 30 million voxels for the entire study. In contrast, fMRI are 4 dimensional (4D) matrices, where time is the fourth dimension. Similarly, dynamic contrast enhanced MRI (DCE-MRI) [16, 17] is also 4D, though here we focus on 3D sMRI.

Figure 2.2 displays an axial slice of the T1 image obtained post-gadolinium injection. Gadolinium chelate is a paramagnetic substance that can be injected in the blood stream and makes blood appear hyper intense in the T1. When a sequence of such images is taken before and after injection, for the purpose of observing and quantifying the blood dynamics into the brain, the sequence is referred to as DCE-MRI. Figure 2.2 shows one time point from a DCE-MRI. Alternatively, only one post-contrast injection image may be acquired and this is referred to as a post-gad T1 image. The image contains an MS lesion surrounded by a hyper intense ring, which indicates blood with a higher concentration of gadolinium. A small red box in Figure 2.2 A is magnified in Figure 2.2 B. Each voxel in the magnification contains both the intensity and the associated numerical data. For example, the largest value in this rectangle is 204, and corresponds to the most hyper intense shade. Images are just representations of data using a particular mapping from real numbers to a gray (or color) scale. Simple manipulations of this mapping can lead to dramatic changes in contrasts, at least in the way they appear to the human eye. The representation appears to be reasonable as the correspondence between known and represented anatomy and pathology are remarkable. Surprisingly, even if one tried to cluster intensities of voxels across the entire brain there is overlap between various tissue classes, simply because the same intensity can easily appear in two different parts of the brain. For example, there are many areas in the normally appearing white matter that have roughly the same intensity with the ring around the lesion.

A natural question then becomes, what are the data units and how comparable are these units across subjects, visits, and studies? Unfortunately, standard sMRI dare unit less. Thus, the size of the units is comparable within the same sequence, though taking the difference between two sequences of the same type is meaningless. Thus, before conducting any sort of analysis on these images, data intensity normalization is a crucial step. We will discuss some methods of intensity normalization in Section 2.5.4.

2.1.2 How can we handle sMRI?

An important characteristic of brain imaging data is that it is big. In most computing environments, loading into memory more than one sMRI study is not recommended. This raises questions about data storage and handling for conducting analyses where data can be accessed one or a few images at a time. We recommend to store data using a folder structure of the type:

```
D:/study_type/subject_id/visit_k/sMRIsequence_name.nii.gz
```

A separate file containing subject identifiers, visit information, covariates and health outcomes can be stored as a master file. Some researchers prefer to have the visit identifier and the subject identifier in the file name to avoid confusion. Regardless of preferences, careful naming and organization of the data is a crucial step towards more sophisticated analyses. As a basic rule, for population level analyses the naming system and directory structure must be consistent, script-friendly, intuitive, and documented.

The compressed files are quite small (around 3Mb), though loading and decompressing hundreds of such files in the computer memory can slow down and even crash computers. We have found that the most robust approach is to upload the minimum number of images necessary for performing the analysis. For example, if one is interested in calculating the mean FLAIR image of spatially registered images then one can simply upload one image at a time and use an iterative formula for calculating the mean. If $\widehat{\mu}_n$ is an estimator of the mean using the first n observations Y_i, $i \geq 1$ then $\widehat{\mu}_n = \frac{n-1}{n}\widehat{\mu}_{n-1} + \frac{1}{n}Y_n$. Similar formulas exist for more complex operations, such as sequential updating of covariance operators.

The `R` computational environment is familiar to statistician and biostatisticians and the R environment has many packages for designed for neuroimaging. However, neuroimaging has been developed primarily outside of statistics, with a distinctly different software and analytic culture. Indeed, in neuroimaging `MATLAB`®, `Python` [18], and `C` are used extensively. Learning these languages is especially useful for direct collaboration. There is an extensive collection of useful neuroimaging software; in this chapter we will cover those which we have found to work particularly well. For example, Medical Image Processing, Analysis, and Visualization (`MIPAV`) is particularly powerful for data visualization, exploratory analysis, spatial inhomogeneity corrections, segmentation, and spatial registration.

2.1.3 What are some major pitfalls when starting working on sMRI?

The biggest mistake in neuroimaging analysis is to look for an application that illustrates a particular biostatistical modeling idea. A "method backwards" approach is problematic in any discipline, but it is especially dangerous in imaging. A reasonably deep understanding of imaging, image pre-processing, and imaging literature can save time, avoid "wheel re-invention", and maintain focus on scientifically relevant and important problems. Thus, we advocate a "problem forward" approach, where biostatisticians and statisticians work directly with collaborators, learn about the details of data acquisition, and identify the most important problems where we can have an impact. Like every technology-intensive field, imaging requires developing a basic set of skills that allows to understand, formulate, and help solve the most important problems. While Biostatisticians "get to play in everyone else's backyard" (John Tukey, Bell Labs, Princeton University), there must be rules about "playing". We have found the Neuroscience community to be incredibly welcoming and open to informed biostatistical and statistical ideas and approaches, when we are open to learning the necessary background for working with neuroimaging data.

Another pitfall is to not understand the dangers that lurk in neuroimaging. Here, we warn of a few. First, there is much biological variation between brains and in addition neurological disease can deform the brain quite dramatically. Therefore, methods that are reasonably well developed for healthy brains tend to fail badly on diseased brains. Second, magnetic coils create spatial inhomogeneities that could be quite large and vary with the subjects and time of the scan. Spatial inhomogeneity corrections, such as N3 [19] or N4ITK[20], work quite well and are reasonably standard in most imaging processing platforms; however, subtle bias fields remain and can strongly affect quantitative analyses. We discuss in detail the inhomogeneity and inhomogeneity correction in Section 2.5.1. Third, for many of the steps in pre-processing, a number of different methods exist and little work has been done to evaluate and compare these methods. For example, it is common to hear statements of the type "my registration method to a template is better" or "this segmentation approach works well". Often there is little evidence supporting such statements, and these judgements are based solely on the qualitative inspection of images. There is a need for validation and replication work as well as understanding human qualitative assessment of images. This is another excellent opportunity for statisticians and biostatisticians to become involved in imaging. A fourth major pitfall is to assume that problems in neuroimaging have been solved. The range, complexity, and diversity of unsolved problems is astonishing. Indeed, registration, intensity normalization, longitudinal co-registration, spatial inhomogeneity, segmentation, population level analyses are all wide open problems. Fifth, quantifying associations between imaging and health is a hard problem that needs to be well understood and addressed. Indeed, brain characteristics are extremely heterogenous across individuals, while longitudinal changes tend to be much smaller. For example, in a study of fractional anisotropy, a measure derived from diffusion weighted imaging, of the corpus callosum [21] the longitudinal variability over 4-5 years only accounted for 2 to 3% of the total observed variability. This raises important problems for biostatisticians and statisticians in many neuroimaging studies where the signal, if it exists, sits under a pile of noise.

2.2 Open scientific problems associated with sMRI of the diseased brain

Given the diversity of diseases and associated scientific questions, it can be difficult to identify important scientific problems. It may be simple to identify segmentation of white matter pathology as a general problem (see Section 2.6.1 for a discussion of lesion segmentation in MS), though we are aware that there are fundamental differences between identifying affected tissues in MS, stroke, cancer, traumatic brain injury (TBI), or Alzheimer Disease (AD.) While recognizing these difficulties, we attempt to provide an overview that is informative. But due to the scope of the problems, we cannot be exhaustive.

From a clinical perspective, the interest is often in subject-specific data. At this level typical scientific questions are related to existence, location, and severity of brain abnormalities that may be clinically relevant. Another set of problems is related to quantifying the volume of white matter, cerebrospinal fluid (CSF), gray matter, and brain. These problems fall under the umbrella of brain tissue segmentation. More subtle problems can also be addressed, such as localized abnormalities. Examples of these are bleeding, gray matter thinning, or quantifying unusual white matter intensity distributions (referred to as "dirty white matter"). For sMRI studies aquired at multiple visits, biological changes between the two visits may be of interest. An example is whether brain abnormalities have disappeared or have worsened, and whether there are quantitative changes in tissue volume or quality between the two visits. A pervasive technical problem is how to align images of the same subject, how to visualize the differences between images when intensities change scale from one visit to the next, and how to eliminate the scanner/visit-specific inhomogeneities. In Section 2.5.1 we will discuss how to address the spatial inhomogeneity correction, in Section 2.5.4 we will present methods for registration (aligning different brains to a template) and co-registration (aligning the sMRI sequences for the same subject's brain from several visits). After applying intensity normalization as described in Section 2.5.5, one can difference the respective sequences for the same subject. Probably the most disappointing part of this exercise on real data is that typically the difference is not zero and reveals the imperfections of the registration and normalization procedures. The most serious problem is the fact that edges and boundaries do not align perfectly and some obvious differences may simply come from the fact that the magnetic signal was stronger in a particular visit than at the other visit. But, where there is disappointment, there is opportunity.

When one is moving from the subject to the population level, a new set of scientific problem arises. Indeed, at the population level one could be interested in mapping the location of lesions on a template brain and studying whether the localization of these lesions is associated with disease severity or progression. Examples of such problems includes mapping the location of MS lesions or of the stroke clot after admission to the Intensive Care Unit (ICU.) Another problem is to quantify differences and changes in brain tissues and their association with health outcomes. For example, how is the size and shape of the ventricles in the brain of a patient infected with HIV related to the duration of the disease or with the type of treatment or with the time from treatment initiation. Another example is to study the association between white matter loss or gray matter thinning and progression to AD. High quality sMRI data for this type of problem is publicly available through the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu/). Another set of problems is to study the population level temporal evolution of lesions or normalized voxel intensity in lesions and their association with health processes and/or treatment. For example, in a stroke trial one may be interested in whether the brain clot is eliminated after surgery, how fast the clot is eliminated and whether faster or slower elimination is better for the patient. In an MS study, one may be interested in analyzing retrospectively whether white matter abnormalities could be used to predict when and where a new lesion will occur. The last set of problems is to study the structure of the data across the population either using unsupervised techniques, such as principal component analysis (PCA) or clustering, or supervised techniques, like regression. For example, one could be interested in analyzing the principal directions of variation in the brain morphology and its association with health outcomes, clustering of subjects according to their image intensities or brain morphology, or identifying locations in the brain that may be strongly associated with cognitive declines related to accelerated aging.

2.3 Data structure and intuitive description of associated problems.

It is useful to describe the data structure and discuss sMRI from a notation perspective. We denote by $Y_{ijm}(v_{ijm})$ the intensity of the mth, $m = 1, \dots, M$, sequence of the sMRI data at the jth study visit, $j = 1, \dots, J_i$, of the ith subject, $i = 1, \dots, I$, at the voxel v_{ijm}. For the data accompanying this chapter $I = J = 2$, and $M = 4$ resulting in a total of 16 images. For those cases when there is only one sMRI per subject (e.g. cross-sectional imaging studies) the index j could be omitted. As the indexes i, j, and k in v_{ijk} indicate, images are typically not registered, in the sense that voxels do not have the same interpretation between the same sMRI sequences, visits, or subjects. A transformation of images that ensures that the voxel depends only on the subject, that is $v_{ijm} = v_i$ is called co-registration. A transformation of the image to a template, $X(v)$, where the voxel does not depend on the subject is called registration to a template or simply registration. While co-registration is less controversial and current software seems to handle it well, registration to a template raises multiple problems, especially in brains affected by disease. We will discuss registration and co-registration in Section 2.5.4

A major problem in imaging is that images may have spatial inhomogeneities. More precisely, this means that the intensity of the image in various tissues (e.g. fat, white matter) varies by the location in the brain. This can be quite obvious when, for example, the inferior part of the brain is brighter than the superior. This can lead to serious problems, as gray matter in the inferior part of the brain may actually be "whiter" than the white matter in the superior part.

Spatial inhomogeneities vary in severity, and can often be very subtle. Such subtle distortions would be discarded by a human observer, but may create serious problems when one tries to analyze data. For example, they have been shown to have a large negative effect on MS lesion segmentation [7]. From a notational perspective, an image with spatial inhomogeneities will have the local intensity distributions in the same tissue vary across locations in the brain. The problem of inhomogeneity correction depends on the definition of "tissue" and requires distribution matching across various tissue types and brain locations. This is a tough problem with imperfect, but reasonable solutions. This is discussed in details in Section 2.5.1. A quick way to diagnose spatial inhomogeneities is to visually identify white matter, fat, gray matter, and bone regions from various parts of the brain and plot the histograms of intensities for each such region separately. A less effective, but faster alternative is to compare the histograms of axial, sagittal, and coronal distributions. Of course, tissue type proportions should vary by slice, but reasonable approximations can be obtained. Another alternative is to use and visualize an aggressive smoother that would hide biological information, but would highlight unusual spatial patterns of image intensity.

Whenever one is interested in analyzing more than one sequence, it is useful for the units in which $Y_{ijm}(v_{ijm})$ is expressed to have the same interpretation and be on the same scale. As we mentioned earlier, this is not the case in sMRI, which can raise fundamental questions related to population level effects. Indeed, if data are not on the same scale even taking the differences between two images does not make sense. A transformation of image intensities from the raw image to an interpretable scale is called image intensity normalization.

This should not be mistaken for image registration, which is also often referred to in practice as "image normalization". In Section 2.5.5 we will discuss the statistical principles of image normalization and we will discuss various ways of conducting image intensity normalization.

2.4 Acquisition and reconstruction

The contrast of an SMRI volume is the relative difference of signal intensities within the volume. When an MRI scan is acquired, changing the scanning parameters changes the contrast of the volume to produce the different sequences, such as FLAIR, T1, T2, and PD. The scanning parameters that contribute to the contrast of an image are the flip angle (FA), the repetition time (TR), the inversion time (TI), and the echo time (TE). A more detailed description of image scanning parameters can be found in [22]. Small changes in the scanning parameters can result in different contrast. For example, two volumes may both be a "FLAIR" volume, but if acquired with different scanning parameters can have different image contrasts. MRI physicist are continually working to develop new imaging techniques in the form of different combinations of these parameters to produce higher quality volumes. It is therefore desirable, but quite difficult, to develop algorithms that are robust to changes in the scanning parameter. Variability in the contrasts can also arise from the strength of the magnet used for imaging. The magnet strength is measure in teslas. Currently, common field strengths for sMRI are 1.5T, 3T, and 7T [22]. Slice thickness and the in-plane resolution of the original volumes is also important, as the volumes may be interpolated during image pre-processing. Information about the scanning parameters, slice thickness, and field strength of the magnet can often be found in the header of the sMRI volume.

During acquisitions, imaging artifacts can arise due to the imaging hardware or from the subject. It is well established that the introduction of artifacts associated with patient motion and the variability associated with scanners can significantly degrade the accuracy of results from further analysis [23] . Therefore volumes from the scanner typically undergo either a manual or automatic quality control to assure that volumes with artifacts are removed before analysis. [23] and [24] both propose automated methods for assessing the quality of an image.

2.5 Pre-processing

After the sMRI data is acquired and reconstructed, data are pre-processed for analysis. It is often hard to define exactly what "pre-processing" means, as it will vary by study, scientist, or even analysis. Indeed, pre-processing and image analysis are closely linked, with pre-processing often having a dramatic impact on the analytic results. Thus, it is important for the biostatisticians and statisticians working with sMRI data to have knowledge of the pre-processing steps and their potential impacts on the downstream analyses. For the purpose of this paper, we divide image pre-processing into four main steps: 1) inhomogeneity correction; 2) spatial interpolation; 3) skull stripping; 4) spatial registration; and 5) intensity normalization. A detailed description of each step with software and data applications is provided in this section. These steps are typically executed in this order in an image pre-processing pipeline and depend on various choices and optimality criteria. A pipeline is a choice of a particular set of image pre-processing steps that can be applied to many images. While we simplify here for understanding, an additional complication is that the order, steps, and algorithm for each step are not agreed upon in the community. Part of the reason for the plurality of pipelines is that it is difficult to quantify the difference in quality between pre-processing pipelines. Developing improved algorithms for image pre-processing and methods for quantifying the quality of pipelines is an area filled with opportunities.

There are many tools for creating image pre-processing pipelines. The choice of tools should be based upon the tools availability, results quality, and computational feasibility for large collections of images. While we do believe that there is no universally best pipeline, a non-exhaustive list of popular sMRI pipelines include the LONI Pipeline Processing Environment [25], the FMRIB Software Library (FSL) [26], and Java Image Science Toolkit (JIST) [27] implemented in Medical Image Processing Analysis and Visualization (MIPAV) [28]. An exciting new tool for `R` users is `ANTsR : Advanced Normalization Tools with R` (http://stnava.github.io/ANTsR/index.html), a pre-processing pipeline that can be run through `R`. Image pipelines often fail on a subset of images, which, left uncorrected, can seriously impact downstream analyses. Therefore, another quality control step must be performed, which often consists of a qualitative visual inspection of the preprocessed images.

2.5.1 Inhomogeneity correction

MRI intensity inhomogeneity is the slow variation of intensities within a tissue class in an image. In the literature intensity inhomogeneity is also referred to as intensity nonuniformity, shading, the bias field or the gain field. Spatial inhomogeneity can be caused by the MRI scanner or by the properties of the object that is being imaged. The latter cause is hard to control and account for, but is relatively small in low magnetic field intensity scanners. Inhomogeneity may raise analytical challenges, because basic assumptions of various models and techniques may be violated. A consequence can be that methods developed for images with no, small, or known inhomogeneity field patterns may fail in heterogeneous imaging studies where inhomogeneity fields can be quite large or have unexpected distributions. For example, many segmentation algorithms use image intensity thresholding on one or more images that are known to discriminate well between specific tissue classes. However, if the same tissue has different intensity distributions at different locations in the brain then segmentation algorithms can be seriously affected. For example, Figure 2.3 A displays a T1 volume from a 7T scanner, while Figure 2.3 B displays the estimated inhomogeneity field, indicating higher intensities in the left-bottom corner. Thus, gray matter in this area has higher intensities than white matter areas in other areas of the brain (gray matter looks “whiter” than white matter.) While resolution and biological details are sharper in higher intensity scanners, spatial inhomogeneity is known to increase with the field strength of the magnet.

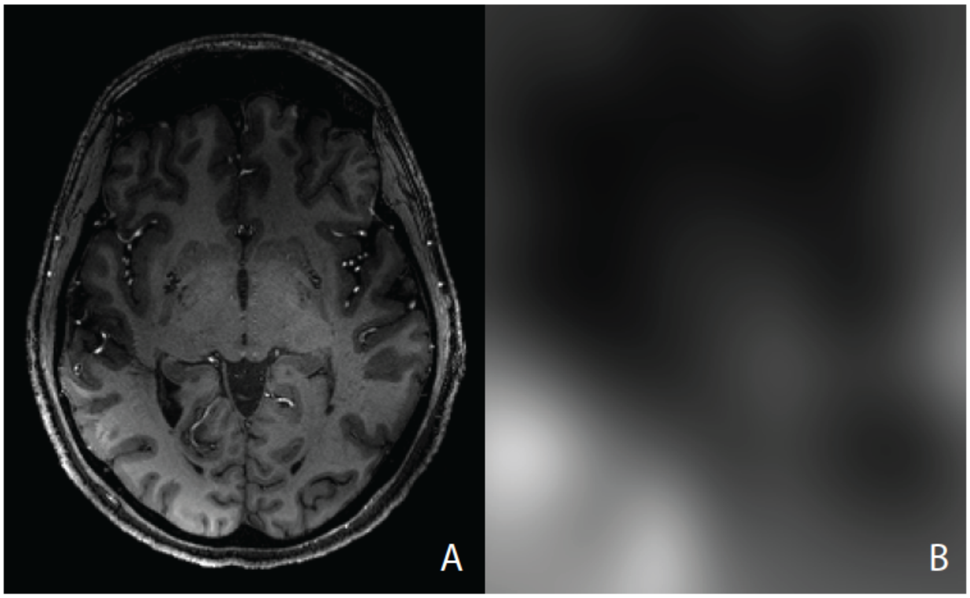


Figure 2.3: A. Axial slice from a T1 -weighted volume obtained from a 7T scanner. Volumes from scanners with a higher magnetic field strength often contain more intensity inhomogenity artifacts, as seen in this image. B. The inhomogenity field for this slice as modeled by the N4 ITK algorithm.

2.5.1.1 Concepts

The inhomogeneity field of an image is commonly modeled multiplicatively. For a voxel v, the observed intensity in an image $Y_m(v)$, where, for simplicity, we have dropped the subject and visit index used in Section 2.3. Conceptually, the observed image is modeled as $Y_m(v) = \alpha(v)X_m(v) + e_v$, where $X_m(v)$ is the true voxel intensity, $\alpha(v)$ is the multiplicative inhomogeneity field, and e_v is an additive error assumed to follow a Gaussian distribution. The additive error is often ignored and data are modeled additively on the log-intensity scale:

$$\log\{Y_m(v)\} = \log\{X_m(v)\} + \log\{\alpha(v)\}. \tag{2.1}$$

[29] and [30] provide comprehensive reviews of methods to correct for image inhomogeneities. The inhomogeneity field can be corrected for prospectively through phantom scans, the use of multiple coils, or special sequences. We focus on retrospective methods to estimate the field from the the data, as these are most relevant to the statistician. Clearly, the deconvolution model 2.1 requires strong assumptions to ensure identifiability. While each method used for estimation makes slightly different technical assumptions, intutitively they all make assumptions about the degree of variation in the $\log\{X_m(v)\}$ and $\log\{\alpha(v)\}$ processes. Typically, though often not explicit, one assumes that $\log\{\alpha(v)\}$ varies spatially much slower than $\log\{X_m(v)\}$. Under this assumption, an aggressive smoother (e.g. 3D kernel smoother with a large window) of the image could be viewed as an estimator of $\log\{\alpha(v)\}$. The majority of inhomogeneity correction methods can be grouped as (1) filtering, (2) surface fitting, and (3) statistical models. In filtering methods, the inhomogeneity field is assumed to be of low spatial frequency and the signal of the anatomical structures in the image of high frequency. The inhomogeneity field can then be removed using a low pass filter. Surface fitting methods use a tissue segmentation first, which is then followed by smoothing within tissue classes. Statistical models assume that the inhomogeneity field follows a particular random process distribution. We will take exception to this nomenclature, as all these approaches are based on statistical models. However, we provide the accepted categorization to help with communication.

When the true inhomogeneity field is not available, criteria used to assess the performance of inhomogeneity correction methods include: 1) variance over the entire image or segmented portions of the image; 2) coefficient of variation over the image; 3) joint coefficient of variation between two tissue classes. When the true inhomogeneity field is available, the mean square error between the derived and true inhomogeneity field is calculated. Other important considerations for assessing these methods are stability, computer requirements, and CPU time [29].

2.5.1.2 Practical approaches, software, and application to data

The most commonly used method for inhomogeneity correction is a statistical model, the nonparametric nonuniform intensity normalization (N3) correction [19]. The method assumes that $f(v) = \log\{\alpha(v)\}$ and $u(v) = \log\{X_m(v)\}$ are two independent random variables with distributions F and U, respectively. The distribution of the sums of these two random variables is the convolution of F and U. N3 searches for the inhomogeneity field to maximize the frequency content of the image intensity distribution and constrains the inhomogeneity field to be modeled as a Gaussian distribution with small variance. Code for this method is publicly available and has been implemented in most pre-processing pipelines. More recently, an improvement and extension of the N3 method has been proposed, the N4ITK [20]. Code for this method is publicly available and the method has already been implemented in many pipelines. Figure 2.3 B shows the inhomogenity field as modeled by the N4 ITK algorithm [20].

2.5.2 Skull stripping

Skull stripping is the process of extracting the brain from an image by removing the background and all other tissue. More specifically, the problem is to estimate $S_{ijm}(v_{ijm}) \in \{0, 1\}$, the indicator variable of brain tissue being contained in voxel v_{ijm} from the images of each subject at each visit, $Y_{ijm}(v_{ijm})$. This process, which may be considered a segmentation task (see also Section 2.6.1), is necessary for the identification of tissue to be studied. Errors in skull stripping can produce both fictitious effects and reduce power if key regions of the brain are erroneously removed.

2.5.2.1 Concepts

While dozens of techniques have been proposed for this task over the past two decades, the most common method remains the brain extraction tool (BET) [31]. BET is a simple technique that aims to iteratively fit a mesh around the surface of the brain, and has been shown to have superior performance to competing methods, although it has documented limitations including a propensity to include extracerebral tissue in the brain mask [32, 33]. While several hybrid methods [34, 35] have been proposed by integrating generative and classifying techniques to produce methods that are robust to differences between scanners and protocols, no solution has satisfactorily solved the problem. Thus, most image analysis groups still resort to manual correction after automatic skull stripping. Recently, multi-atlas label fusion techniques [36, 37] have shown great promise for skull-stripping [38, 39]; these methods use deformable registration to compare the subject under study with a library of other images for which manual skull-stripping images (called atlases) and average (or fuse) these manual labels across atlases. Patch-based techniques [40] have also shown great promise with significantly lower computational burden. As new methods are developed, many authors submit results for active comparison to a validation resource, and comparisons are publicly available [41].

2.5.2.2 Practical approaches, software, and application to data

As BET [31] is so commonly used, we demonstrate its application as an easy-to-use and computationally practical approach. BET was first implemented in FSL [42], but now is available in other image processing packages including MIPAV and JIST [43]. Using MIPAV, skull-stripping can be achieved using BET on a T1-weighted image in less than a minute on a standard personal computer and an example of the results is shown in Figure 2.4.

2.5.3 Interpolation

Interpolation transforms a discrete array of numbers into a continuous image. As we saw in Figure 2.2, sMRI are arrays of intensity values that have been sampled on a grid. When performing operations such as image registration, magnification, image reslicing and resampling, and surface rendering it is desirable to have a continuous image and to know the approximate values of the image at points other than those on the original grid.

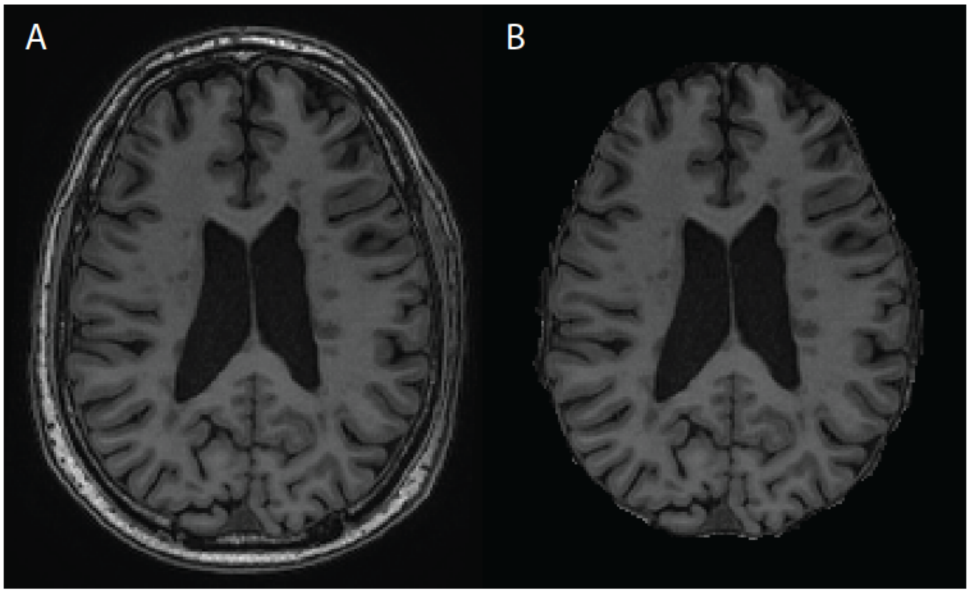


Figure 2.4: An axial slice of from a 3T T1-weighted imaging of a patient with MS before (A, showing $Y_{ijm}(v_{ijm})$) and after (B, showing $Y_{ijm}(v_{ijm})S_{ijm}(v_{ijm})$) skull stripping using BET.

2.5.3.1 Concepts

An extensive review and comparison of interpolation methods in medical image analysis can be found in [44] and [45]. The most common interpolation methods in sMRI analysis are truncated and windowed sinc, nearest neighbors, linear, quadratic, cubic b-splines, cubic, Lagrange, and Gaussian interpolation. Consider the voxel $v = (x_v, y_v, z_v)$ with coordinates (x_v, y_v, z_v) that were not necessarily among the coordinates where data were sampled. Then the image can be interpolated at v as

$$Y_m(v) = Y_m(x_v, y_v, z_v) = \sum_{p,r,s} Y_m(p, r, s)h(x_v - p, y_v - r, z_v - s)$$

where the function $h(\cdot,\cdot,\cdot)$ is the interpolation kernel and the summation is done over all p, r, s where data are observed. To provide some intuition we describe the interpolation kernels for “one nearest neighbor”, “linear” and “windowed sinc”. For one nearest neighbor the interpolation kernel is:

$$h(x) = \begin{cases} 1 & 0 \leq |x| \leq 0.5 \\ 0 & \text{elsewhere} \end{cases}$$

For linear interpolation the interpolation kernel is:

$$h(x) = \begin{cases} 1 - |x| & 0 \leq |x| \leq 0.5 \\ 0 & \text{elsewhere} \end{cases}$$

The sinc function is $\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$. If N denotes the number of supporting points used for interpolation then the windowed sinc interpolation kernel is

$$h(x) = \begin{cases} \frac{\sin(\pi x)}{\pi x} & 0 \leq |x| \leq \frac{N}{2} \\ 0 & \text{elsewhere} \end{cases}$$

2.5.3.2 Practical approaches, software, and application to data

In sMRI pre-processing, interpolation is linked closely with spatial registration. As an image is spatially aligned to another image or a template, the image being registered must be interpolated to determine the values of the registered image in the new coordinate space. Interpolation methods are typically a tuning parameter for registration and are important as they can impact the clarity of the image after registration. Windowed sinc has been shown to produce good results in accordance with the number of supporting points used in the interpolation, but can become quite computationally intensive as the number of supporting points increases [44].

2.5.4 Spatial Registration

Registration is the process of determining the spatial alignment and correspondence between images. Consider the case when one is interested in registering image $Y(p, r, s)$ to $\widetilde{Y}(p, r, s)$, where $p = 1, \ldots, P$, $r = 1, \ldots, R$, and $s = 1, \ldots, S$ are the indexes of the three dimensional arrays. The dimension of both arrays are $P \times R \times S$; when the arrays have different dimensions interpolation, as described in Section 2.5.3, can be applied. If we identify a voxel with its array index $v = (p, r, s)$ then the product of registration is a bijective transformation map, $v \to T(v)$, from one image reference system to another. The registered image in the reference system of $\widetilde{Y}(\cdot, \cdot, \cdot)$ is then $Y\{T(p, r, s)\}$ whereas results or images can be obtained in the "native space" by using the back transformation $v \to T^{-1}(v)$. It is important to note that registration is a transformation of space and does not affect image intensities; however, image intensities can be used to find optimal transformations in a specific class of transformations. In this section we focus on intra-subject registration, also referred to as spatial normalization. Inter-subject registration and registration to a group template image are discussed in the Analysis section. [46] provides a detailed summary of registration methods and we use the notation introduced in this text in the following description.

2.5.4.1 Concepts

There is an infinite number of transformations $v \to T(v)$ and they range from useless to useful. For example, given two images expressed on the same 3D grid one can perfectly transform each individual point from the first image to each individual point in the second image. Such a transformation may or may not respect some order and is characterized by the degree of smoothness (number of degrees of freedom). A random assignment of indexes would have $P \times R \times S!$ degrees of freedom, with most transformations being useless and uninformative. Here we will describe a few useful transformations, including rigid and affine and we will provide the necessary tools for non-linear and diffeomorfic approaches.

Rigid registration is the simplest type of registration and consists of a translation and rotation. Thus, 3D rigid registrations have six-degrees of freedom, 3 associated with the translation vector, $t = (t_x, t_y, t_z)$, in the x, y and z directions and 3 associated with the rotation parameters $\theta = (\alpha, \beta, \gamma)$. For a voxel $v = (i, j, k)$ the rigid transformation can be written as:

$$T_{\text{rigid}}(v) = Rv + t$$

where

$$R = \begin{bmatrix} \cos\beta\cos\gamma & \cos\alpha\sin\gamma + \sin\alpha\sin\beta\cos\gamma & \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma \\ -\cos\beta\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma \\ \sin\beta & -\sin\alpha\cos\beta & \cos\alpha\cos\beta \end{bmatrix}.$$

An affine registration has the same form as the rigid and is of the type $T_{\text{affine}}(v) = Av + t$, with the difference that the matrix A is not constrained to be a rotation matrix. Thus, the total number of degrees of freedom of 3D affine transformation is 12 with 9 degrees of freedom corresponding to the 9 entries of the matrix A and 3 corresponding to the translation vector t.

Choosing the registration class (e.g. rigid or affine) is a crucial step, though one still needs to estimate the parameters of registration in the induced spaces. This is done throughout the minimization of a utility function. There are three main ways of constructing a utility functions using: 1) landmarks; 2) surface fitting; and 3) voxel-similarity metrics. In landmark based registration, fiducial markers or landmark point identified by hand in the images are used as points of reference. A fiducial marker is an object that is placed in the field of view in an image. These landmarks replace the original frame of reference and transformations are applied to reduce a particular distance between them, which could include minimizing the geometric distance or a combination between the geometric distance and the intensities in the image. Either the Root Mean Square (RMS) Error or Fiducial Registration Error (FRE) can be optimized to select the registration parameters. The error of the registration can be assessed by reporting the Target Registration Error (TRE) for the non-landmark areas in the image. Landmark based registration requires the identification of landmarks, which can be time consuming and may be prone to observer error. Therefore, finding landmarks automatically and reliably is an active area of research. An excellent comprehensive description of shape analysis and landmark-based registration can be found in [47].

Surface based registration takes into account the different geometric structures of the brain to improve the intra-subject variability of brain shapes after registration. Several methods exist for surface based registration including methods available in the widely used software FreeSurfer (http://surfer.nmr.mgh.harvard.edu). In surface registration, the idea is to think of the ceberal cortex as a surface which is transformed into the new space so that the gyri and sulci on the cortex are matched.

Voxel-similarity based registration methods are popular, as they do not require the identification of landmarks or segmentation of the image. Here the registration T is optimized by a function of the voxel values in the two images. For images of the same modality, the sum of shared difference (SSD) can be used:

$$SSD = \frac{1}{n}\sum_{v\in\Omega} |Y\{T(v)\} - \widetilde{Y}(v)|$$

where Ω is the image domain of the two images, $Y\{T(v)\}$ is image after the transformation T is applied and $\widetilde{Y}(v)$ is the voxel intensity of the target image. Another popular similarity function is correlation coefficient (CC) between the intensity values in the two images and is defined as

$$CC = \frac{\sum_{v\in\Omega}[Y\{T(v)\} - \bar{Y}^T][\widetilde{Y}(v) - \bar{\widetilde{Y}}]}{\sqrt{\sum_{v\in\Omega}[Y\{T(v)\} - \bar{Y}^T]^2 \cdot \sum_{v\in\Omega}[\widetilde{Y}(v) - \bar{\widetilde{Y}}]^2}},$$

where $\bar{Y}^T = \sum_{v\in\Omega} Y\{T(v)\}/n$ and $\bar{\widetilde{Y}} = \sum_{v\in\Omega} Y(v)/n$. For images that are obtained as different sequences or even different modalities, the intensities of the images can differ quite dramatically. Thus, joint entropy is often used as an alternative methods registration optimization method. For a vector of probabilities $p = (p_1, \ldots, p_K)$ the entropy $H(p) = -\sum_k p_k \log(p_k)$. Entropy can be thought of as a measure of information contained in the image. Maximum entropy, $\log(K)$, is obtained when $p_1 = \ldots = p_K = 1/K$. Minimum entropy, 0, is obtained when $p_1 = 1$ and $p_2 = \ldots = p_K = 0$. Maximum entropy corresponds to a perfectly chaotic environment (e.g. random assignment of shades of gray to an image), whereas minimum entropy corresponds to a perfectly organized system (e.g. assigning the same shade of gray to the entire image.) Registration is often done through minimizing joint entropy,

$$H(A, B) = \sum_a \sum_b p_{AB}(a, b) \log p_{AB}(a, b)$$

where $p^T_{AB}(a, b)$ is the joint probability of the pair of image values a in image A and b in image B being observed at the same voxel. As image intensities can be completely different minimizing $H(A, B)$ directly does not typically work directly. Instead, the histogram of each image intensities can be partitioned into quantiles and one can assess that the two images have the same intensity at voxel v if the image intensities fall within the same inter-quantile interval of the image-specific intensity distribution. In a scatter plot of image intensities, at the corresponding voxels, $H(A, B)$ is a symmetric measure of how far the points are from the identity line. A major problem in this context is that there are many background (non-tissue) voxels, which could dominate the measure. To mitigate the effect of the many background voxels, joint entropy can be replaced by mutual information

$$I(A, B) = H(A) + H(B) - H(A, B)$$

The mutual information is a measure of the mutual dependence of the two images.

In practice the effect of these measures is often not completely understood and minor assumptions can have serious effects on the results of registration. Note, for example, that the quantile transformation that we have introduced above is, essentially, a histogram mathcing approach for signal intensities. Such approaches can be seriously affected if the relative intensities in images have different distributions. For example, a brain with larger ventricles will have a larger number of voxels in the cerebro-spinal fluid (CSF) than a standard template. Similarly, a brain with a large lesion with specific intensity properties will have a histogram with a fundamentally altered shape. Ignoring the effects of pathology and between-subject variability can have large effects on the results of registration and is one of the dirty, unspoken of, secrets of registration.

2.5.4.2 Practical approaches, software, and application to data

We now discuss and visualize some simple examples showing the process of registration. Registration can be thought of as a collection of steps that transform the image into the template space. Recall that registration is a transformation on the voxel location and not of image intensities; image intensities can be used to optimize the transformation using, for example, differences between the transformed image and a template. Suppose that we observe a simple 2-dimensional (2D) image depicted in Figure 2.5 (top left panel) that needs to be transformed to the template space (top right panel). In this example, the template image is a clockwise rotation of the observed image by a $\pi/2$ degree angle and a shift. Hence, we can use the following rotation matrix to transform the observed data into the template space.

$$R = \begin{bmatrix} \cos(\pi/2) & -\sin(\pi/2) \\ \sin(\pi/2) & \cos(\pi/2) \end{bmatrix}$$

Thus, for each pixel with coordinates $x = (x_1, x_2)^T$ we obtain the coordinates in the new space $y = (y_1, y_2)^T$ as $y = Rx$. The resulting image is shown in the left middle panel of Figure 2.5. With a shift in the X coordinate we may completely register the observed image into the template space. The shift can be incorporated in the transformation as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\pi/2) & -\sin(\pi/2) & -101 \\ \sin(\pi/2) & \cos(\pi/2) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$

A comparison between the top-right and top-middle panels in Figure 2.5 indicates that the resulting image is very similar to the image in the template space; in fact, because this is a toy example, they are identical.

A rigid transformation is useful in cases when several images for one subject are acquired over a relatively short period of time, as we expect the images to be similar except that the subject may have changed positions between the image acquisitions. If the acquired image has different voxel dimensions than the template image, we may want to use an affine transformation. In a second example, the observed data is simply a $\pi/4$ degree rotation of the template image while the pixel size is half that of the pixel size of the template image. We may use the following transformation matrix to transform the observed image into the template space.

$$\begin{bmatrix} y_1 \\ y_2 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \cos(-\pi/4) & -\sin(-\pi/4) & 0 \\ \sin(-\pi/4) & \cos(-\pi/4) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$

The image in Figure 2.5 is based on rounding the noninteger coordinates. Thus, as described in Section 2.5.3, interpolation of the intensities is needed to obtain the complete image. As discussed above, once the user chooses the parameterization (e.g. rigid, affine, etc.) and the objective function to be minimized the transformation matrix can be estimated.

Several softwares exist to estimate the transformation matrix for 3D image data, including the Advanced Normalization Tools (ANTS) described by [48], FMRIB Software Library (FSL) (see [26] for a general overview of FSL), Medical Image Processing Analysis and Visualization (MIPAV) (http://mipav.cit.nih.gov) and Statistical Parametric Mapping (SPM) (http://www.fil.ion.ucl.ac.uk/spm/).

The top left image in Figure 2.6 shows one slice of a 3 dimensional T1 image. Suppose that data from multiple subjects need to be registered into the template space shown in the top right panel; this is the Montreal Neurological Institute (MNI) template. The following code in FSL can be used to obtain the affine transformation matrix from the image into the template space.

```
flirt -in Brain.nii.gz -ref Template.nii.gz  -out Brain_affine.nii.gz
      -omat affine.mat
```

where Brain.nii.gz contains the image, Template.nii.gz is the template image. The resulting Brain_affine.nii.gz will be the image transformed into the template

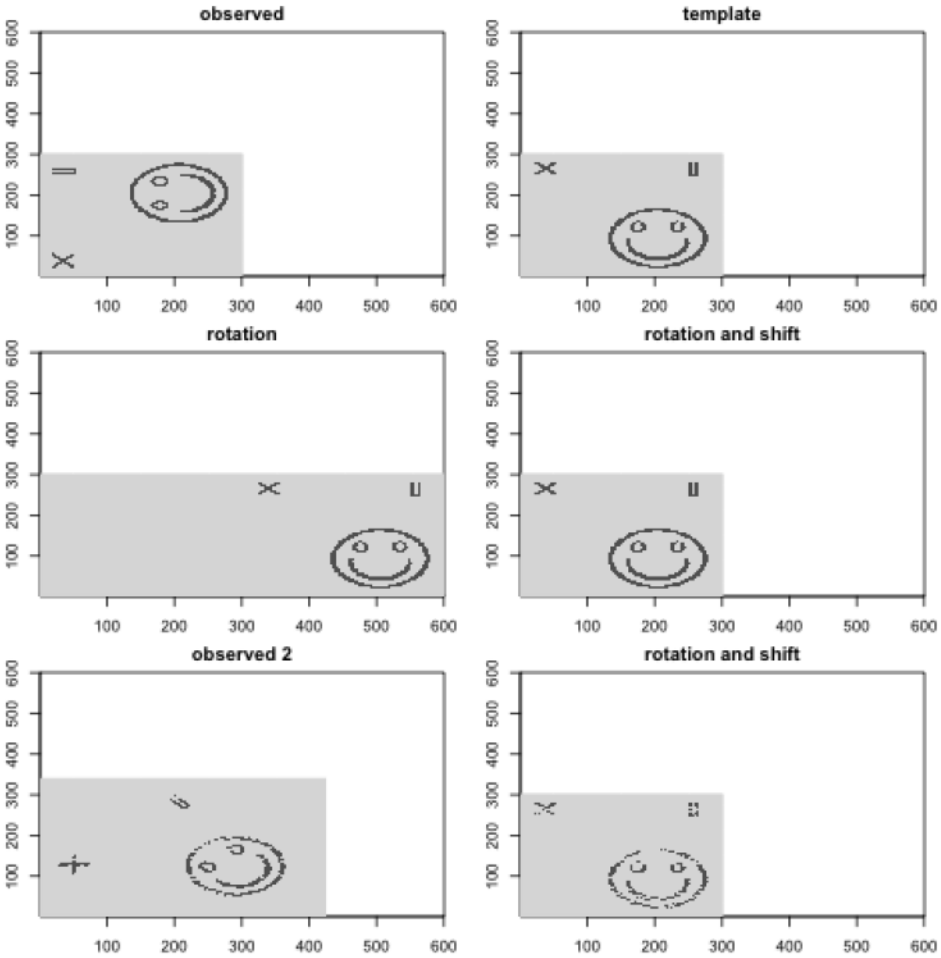


Figure 2.5: Steps of registration: a toy example.

space, finally, the affine.mat will show the transformation matrix.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.089 & 0.001 & 0.025 & 186.19 \\ 0.006 & -1.104 & 0.054 & 224.17 \\ -0.006 & -0.005 & 1.173 & -13.196 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix}$$

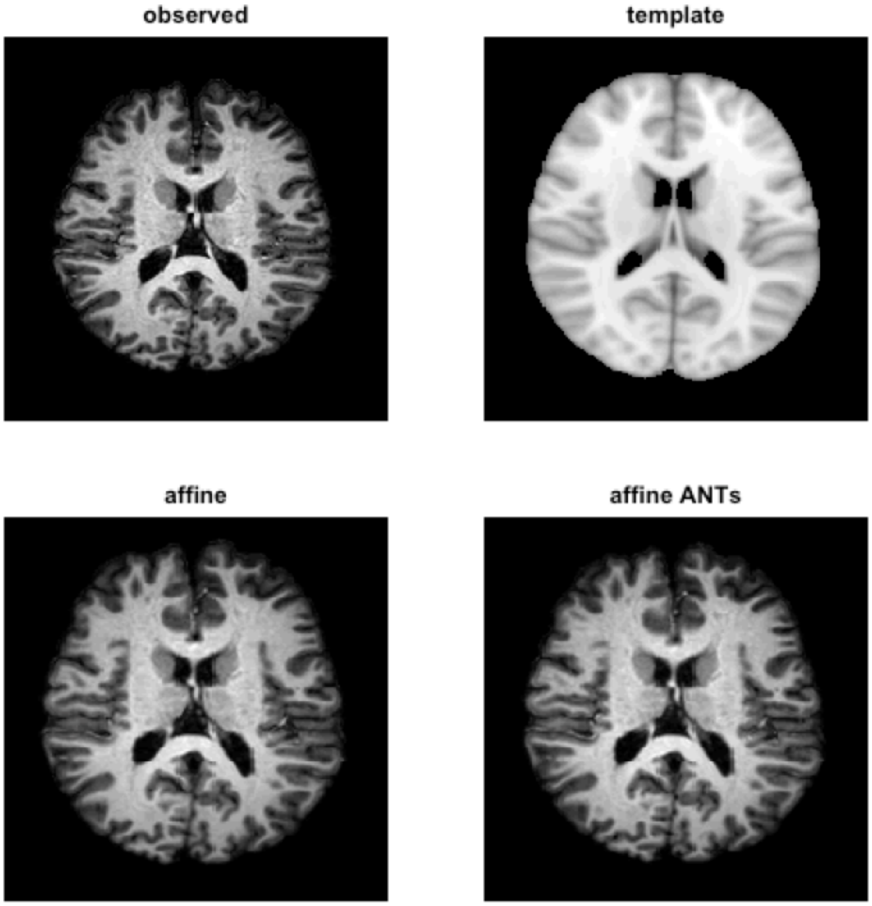


Figure 2.6: Application of two software methods (the function flirt on bottom left and ANTs affine on bottom right) to register a real brain image (top left) to a template (top right).

A similar transformation can be obtained using the ANTs software. However, the function ANTS provides the transformation matrix as an outcome and a second function has to be used to transform the image into the new space based on the estimated matrix from ANTS.

```
ANTS 3 -m MI[Brain.nii.gz, Template.nii.gz, 1, 4] -o img.nii.gz
       -r Gauss[3,0] -i 0


WarpImageMultiTransform 3 Brain.nii.gz BrainWarp.nii.gz  -R
      Template.nii.gz  -i imgAffine.txt
```

If the brain structures are similar up to affine differences then the methods described above will likely produce reasonable results. However, in some cases one area of the brain may need to be transformed more than surrounding tissue or tissue from a different part of the brain. Approaches that go beyond affine transformations are typically referred to as non-linear and diffeomorfic approaches. Here is one parameterization in ANTs to obtain a non-linear transformation:

```
ANTS 3 -m CC[Brain.nii.gz, Template.nii.gz, 1, 8] -o img.nii.gz
       -r Gauss[1,1] -i 30 * 20 * 5 -t SyN[0.25]


WarpImageMultiTransform 3 Brain.nii.gz BrainWarp.nii.gz  -R
      Template.nii.gz  -i imgAffine.txt BrainInverseWarp.nii.gz
```

[48] provides an overview of how one can use ANTs to obtain non-linear transformations of the images from patients with disease, by focusing the optimization function to the area containing the healthy tissue.

2.5.5 Intensity normalization

Conducting any type of population-level analysis or inference on data usually requires that the units of measurement have the same meaning for every subject and visit. Indeed, when units are different, even calculating a simple average is not possible. This is a well-known problem in sMRI image analysis, as these modalities are measured in arbitrary units that depend on many factors including the scanner, protocol, and manual adjustments made by the radiology technician acquiring the images. The importance of intensity normalization has been emphasized by numerous publications in the imaging literature [49, 50, 51]. The normalization process should produce units that: 1) have a common interpretation at the tissue-type level; 2) are replicable; 3) preserve the rank of intensities; 4) have similar distributions for the same tissues of interest across and within patients; 5) are not influenced by biological abnormality or population heterogeneity; 6) are minimally sensitive to noise and artifacts; and 7) do not result in the loss of information either due to pathology or other phenomena. These principles, proposed in [52] and referred to as the statistical principles of image normalization (SPIN), guide the mathematical formulation described in the next section.

2.5.5.1 Concepts

Consider the image intensity $Y_{ij}(v)$ at each voxel v expressed in arbitrary units and measured for subject i at visit j. Normalization is any transformation of the type $Y_{ij}(v) \rightarrow N_{ij}\{Y_{ij}(v)\}$. It is useful to conceptualize the histogram of intensities $Y_{ij}(v)$ as a mixture of densities

$$h_{ij}(x) = \sum_{k=1}^{K} w_{ijk} f_{ijk}(x), \tag{2.2}$$

where $f_{ijk}(x)$ are the subject/visit-specific intensity densities of empty space and known tissues, such as white matter, gray matter, cerebrospinal fluid, bone, skin, and lesions. The weights $w_{ijk} \geq 0$ sum to 1 and represent the relative weights of components $k = 1, \ldots, K$. This includes both cases with and without pathology, as the weight for lesions can be allowed to be 0. Careful inspection of SPIN 4 suggests that after normalization $f_{ijk}(\cdot)$ should be as close to one another as possible for all i and j and for any fixed k. Thus, a natural starting point would be to consider transformations that reduce the distance between the $f_{ijk}(x)$ for any fixed k. Together with SPIN 1, this suggests the existence of the following theoretical model in normalized space for all images: $g_{ij}^{N}(x) = \sum_{k=1}^{K} w_{ijk} g_k(x)$, where the densities $g_k(x)$ are independent of subjects and/or visits, though the weights assigned to these densities depend on subject and visit and may be the measure of interest in medical studies. The fundamental difficulty of normalization is to find a transformation from $h_{ij}(x)$ to $g_{ij}^{N}(x)$ that respects the ordering of distributions and their mutual distances in the normalized space.

2.5.5.2 Practical approaches, software, and application to data

The most widely used image normalization techniques have centered on histogram matching [49, 53, 50]. First, a template histogram is constructed, usually by averaging across a group of subjects in a training set. Then, a nonlinear transformation $N_{ij}(\cdot)$ (often a linear spline) is estimated for each subject at each visit to minimize any deviations between the normalized histogram $N_{ij}\{Y_{ij}(v)\}$ and the template. Although histogram matching methods produce replicable results, they are based on assumptions that are often violated: 1) the tissue-type distribution is the same across subjects and visits; 2) the absence of abnormal pathology; and 3) the absence of technical artifacts. For example, Figure 2.7 shows how the assumption of common distributions of tissue throughout the head can cause mismatching of gray matter (GM) to cerebrospinal fluid (CSF); note how a normal-appearing part of the brain (raw data shown in the top panel) is induced to show erosion of GM by histogram normalization (histogram normalized data shown in the bottom left panel).

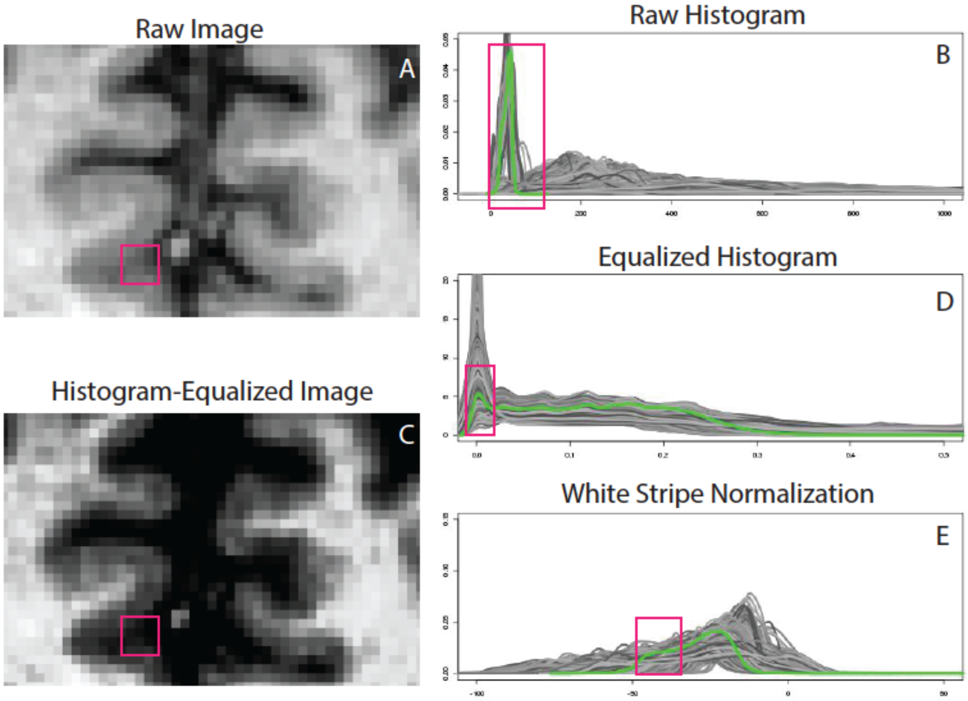


Figure 2.7: First column: region of interest from patient with MCI shown before (A) and after (C) histogram matching. Red square indicates region of gray matter on the unnormalized image that disappears after histogram matching. Second column: histograms (shades of gray indicate different study visits) of the gray matter before (B) and after (D) histogram matching and (E) white stripe normalization for subjects in ADNI. Note the large proportion of gray matter mismatched to background (zero intensity) after histogram matching.

An alternative method [54, 55] that has been proposed is to match the underlying a particular subdistribution consisting of a reference tissue as well as possible. We call this tissue-specific histogram normalization. It is important to note that although the distribution of intensities in the reference tissue are matched, regions of normal tissue with intensities outside the range of the reference tissue are not necessarily comparable; the true normalization function $N_{ij}(\cdot)$ is often nonlinear (due to differences in protocol, etc.) and thus normalizing with respect to white matter may not result in normalized gray matter intensities. Tissue-specific histogram normalization does, however, maintain the natural variability in other tissue types, allowing for the study of pathology.

Assume for the moment that for every subject and visit we have an area of reference tissue (a sub-mask, usually of the white matter). Then we can accurately estimate $f_{ij1}(x)$ (say $k = 1$ for white matter) for each i and j and obtain a normalized estimator that has mean zero and variance one, $f^N_{ij1}(x) = \sigma_{ij1} f_{ijk}(\mu_{ij1} + \sigma_{ij1} x)$, where μ_{ij1} and σ_{ij1} are the mean and standard deviation of $f_{ij1}(x)$, respectively. An estimator of $g_1(x)$ is the average of $f^N_{ij1}(x)$ and linear normalization with respect to the white-matter distribution is

$$h^N_{ij}(x) = \sum_{k=1}^{K} w_{ijk}[\sigma_{ij1} f_{ijk}(\mu_{ij1} + \sigma_{ij1} x)]. \tag{2.3}$$

All units are expressed in multiples of standard deviations, σ_{ij1}, of the white-matter intensities, and zero is the average intensity of white matter. This method relies on the availability of a reference tissue (usually white matter) mask indicating a region of the brain that should be comparable across subjects. This is often unavailable before intensity normalization, however, and

[52] proposed and validated a fully automatic method that avoids this.

Consider a T1 sMRI, $Y_{ij}(v)$, for subject $i = 1, \ldots, n$. [52] use normal-appearing white matter (NAWM) as a reference color, since it is the largest and most contiguous brain tissue. To identify the distribution of NAWM intensities, [52] use a penalized spline smoother [56] to estimate the mode of the intensity histogram in white matter, μ^*_{ij1} (the largest non-background peak). To estimate the variability within NAWM on the raw image, they estimate the standard deviation σ^*_{ij1} of intensities in $\Omega_{i,\tau} = \{v : H^{-1}_{ij}[H(\mu^*_{ij1}) - \tau] < Y_{ij}(v) < H^{-1}_{ij}[H(\mu^*_{ij1}) + \tau]\}$, which is referred to as the white stripe (where $H_{ij}(x) = \int_{-\infty}^{x} h_{ij}(x)\, dx$). Here τ is a quantile tolerance in the original space of intensities. The estimation of μ^*_{ij1} and σ^*_{ij1} was been found to be remarkably robust across several thousand images (failure rate $< 1\%$). If the family of densities $f_{ij1}(v)$ can be parameterized by two parameters then $\mu_{ij1} = \psi_1(\mu^*_{ij1}, \sigma^*_{ij1})$ and $\sigma_{ij1} = \psi_2(\mu^*_{ij1}, \sigma^*_{ij1})$ (proof follows from the method of moments). Thus, matching μ^*_{ij1} and σ^*_{ij1} (estimable directly from the white stripe without prior segmentation) results in matching μ_{ij1} and σ_{ij1}, and this tissue-specific histogram normalization method has been shown to perform well in large multi-center studies.

2.6 Analysis

2.6.1 Lesion segmentation

Segmentation is the labeling of voxels in an image according to particular properties of the voxel (e.g. the type of tissue the voxel contains). Examples include segmenting and comparing the cingulate gyrus of subjects with schizophrenia and healthy controls [57] and segmenting the hippocampus to assess volume loss due to chronic heavy drinking [58]. A review of methods for segmentation of brain sMRI can be found in [59] and [60].

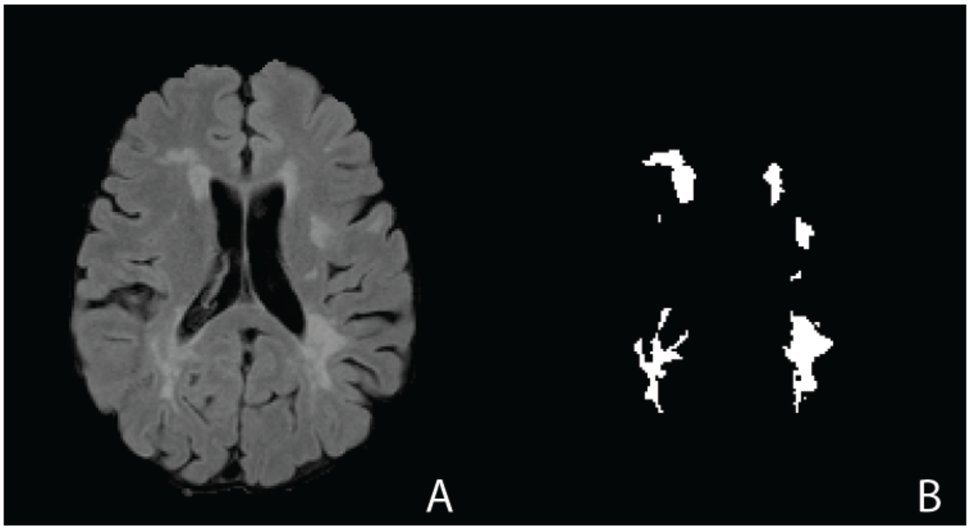


Figure 2.8: A. An axial slice from the FLAIR volume from a patient with multiple sclerosis. B. Manual expert segmentation of the lesions from this slice.

Here we describe the problem of segmentation of brain lesions in multiple sclerosis (MS) from a single sMRI study. MS is an inflammatory disease of the brain and spinal cord characterized by demyelinating lesions that can be observed with sMRI [61]. In MS quantitative analyses of sMRI, such as the number and volume of lesions in an image, are essential for diagnosing the disease and monitoring disease progression [2]. In practice, MS lesions are manually segmented by experts from sMRI. Figure 2.8 shows an example of a manual segmentation of lesion voxels. As manual or semi-automated segmentations of images are time consuming, costly, and prone to large inter- and intra-observer variability [62], development of automated methods of lesion segmentation is an active field of research [63]. Reviews of current lesion segmentation methods can be found in [63], [64], and [65]. LesionTOADs is a readily available software for lesion segmentation that runs in the JIST pipeline enviornment [66]. An excellent resource for lesion segmentation data is the MS lesion segmentation challenge 2008 [67]. This database include sMRI volumes acquired at the Children's Hospital Boston and University of North Carolia along with expert manual segmentations.

Lesions segmentation is a classification problem. In the literature, supervised classifiers are trained on expert manual segmentations of lesion voxels or unsupervised classifiers use clustering methods to identify lesion voxels . The covariates or features in the model are derived from the different imaging sequences Y_m. From these images anatomical information derived from atlases and/or the voxel-level intensity information from an imaging modality $Y_m(v)$ can be used for classification [65]. To illustrate the problem, we provide an overview of a lesion segmentation method from the literature [7], as it is a logistic regression model, a model that is familiar to statistician and biostatisticans. Let $L_i(v) = 1$ if the manual segmentation of voxel v is a lesion for subject i and let $L_i(v) = 0$ otherwise. Concatenate the manual segmentations from each voxel for all subjects into a single vector L. Similarly, let Y denote the design matrix of features derived from the different imaging modalities for all of the voxels and subjects. We can then model the probability that a lesion contains a voxel with the logistic regression model:

$$\text{logit}\{P\{L(v) = 1\} = Y(v)\beta$$

Recently, it has been found that for lesion segmentation the particular classification algorithm is less important than the development of the features [8]. In spite of this, methods continue to be classified according to the method used for classification and not by the feature space. Indeed, the classification method (e.g. random forrest, svm, or logistic regression) offers basically no information on what actually improves prediction. In a seminal paper [68], D.J. Hand notes "the extra performance to be achieved by more sophisticated classification rules, beyond that attained by simple methods, is small". It follows that if aspects of the classification problem are not accurately described (e.g., if incorrect distributions have been used, incorrect class definitions have been adopted, inappropriate performance comparison criteria have been applied, etc.), then the reported advantage of the more sophisticated methods may be incorrect." We subscribe to this view, though we describe it more plastically as "there are very few reasonable ways of cutting the potato." Here, the potato is, of course, the cloud of points in the feature space.

A major source of difficulty in automated segmentation of MRI is due to variable imaging acquisition parameters [65, 63]. Most segmentation methods have tuning parameters that are adjusted to a particular data set and may not generalize to a new data set with different acquisition parameters. Lesion segmentation is also closely intertwined with image pre-processing. Methods using anatomical information rely on proper nonlinear spatial registration to a template image. The use of image intensities requires high quality inhomogeneity correction and intensity normalization – otherwise population level modeling is hopeless. There are even lesion segmentation methods in the literature that iterate between lesion segmentation and inhomogeneity [69].

2.6.2 Lesion mapping

In Section 2.5.4 we discussed spatial registration methods for structural images. There are two main issues in the context of registration in presence of lesions: 1) It can be very hard to find the best way to register images for many subjects into the same brain if their images are distorted due to disease; 2) should the segmentation of lesions be performed before or after registration. If the lesions in the brain have been hand-segmented in the native space, or in the rigidly registered space which is common in MS studies, then one can use the lesion masks to extract the areas that are within the lesion and use the healthy brain for registration. Both ANTs and FSL provide the capability of excluding the lesion areas from the mask when estimating the transformation matrices. For instance in ANTs this can be performed by adding an option in the function call.

```
ANTS 3 -m MI[Brain.nii.gz, Template.nii.gz, 1, 4] -o img.nii.gz
       -r Gauss[3,0] -i 0  -x lesion_mask.nii.gz
```

Another approach to alleviate the effect of the presence in lesions in the brain is to fill in the area of lesions by the average of normal appearing white matter intensities and then estimating the transformation matrix for the resulting images. After obtaining the transformation matrices for each subject their corresponding lesion masks can be transformed into the template space by applying the transformation.

```
WarpImageMultiTransform 3 lesion_mask.nii.gz lesion_maskWarp.nii.gz  -R
      Template.nii.gz  -i imgAffine.txt
```

After registering all the mask maps into the template space we can plot a so-called lesion histogram which shows the prevelence of lesions in certain areas of the brain as shown in Figure 2.9.

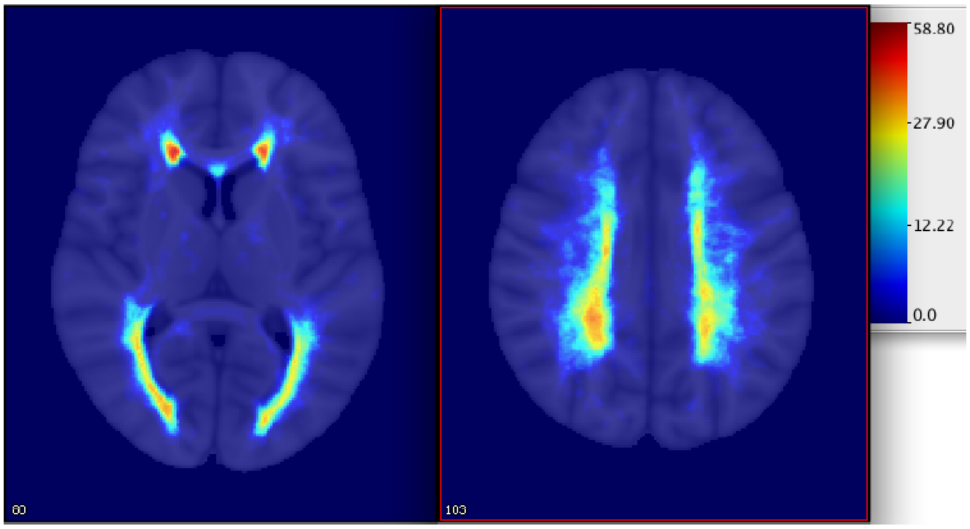


Figure 2.9: Two slices from a histogram of lesions constructed using non-linear registration of observed images in ANTs.

2.6.3 Longitudinal and cross sectional intensity analysis

Another very important type of analysis is to directly analyze the intensities in the image after intensity normalization. Indeed, after intensity normalization, we can compare and quantify the histograms of intensities within each sequence or combined across sequences. This could be done at the brain level, which requires only intensity normalization, or at the tissue level, which would also require a segmentation algorithm. For example, at the population level one may be interested in the intensity distribution in white matter. This could be done by obtaining the histogram intensities, stacking them in a matrix and conducting a PCA or other dimensionality reduction approach. A similar analysis could

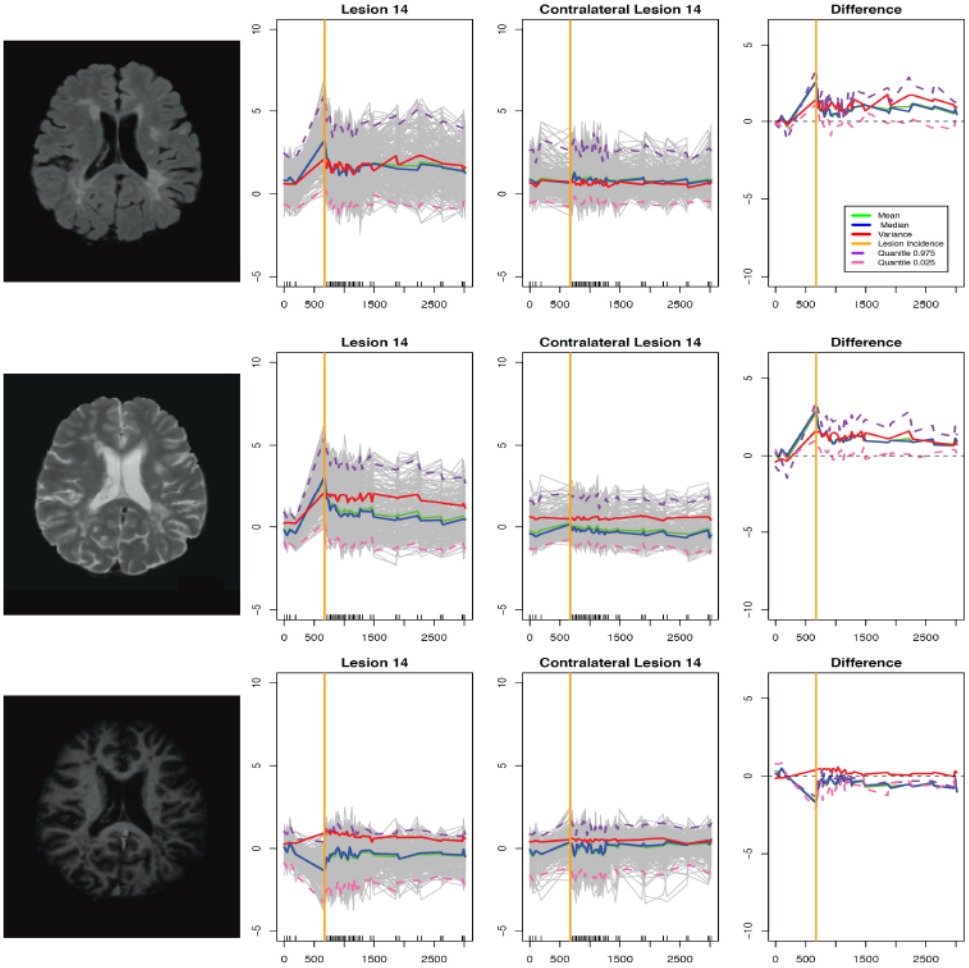


Figure 2.10: Voxel intensities for FLAIR (top panels), T1 (middle panels), and T2 (bottom panels) images in a lesion (labeled lesion 14) for one subject over 8 years at 40 visits.

be conducted at the lesion level if one is interested in cross-sectional differences in lesion intensity distributions. Another possibility is to study the association between these distributions and health outcomes.

To provide a view of what could be achieved, consider a study of natural history of MS conducted by Daniel Reich at the National Institutes of Health (NIH.) Figure 2.10 displays the voxel-specific intensities of the FLAIR (top panels), T1 (middle panels), and T2 (bottom panels) images in a lesion (labeled lesion 14) for one subject over 8 years at 40 visits. The x-axis represents time in days and the y-axis represents image intensities as standard deviations of white matter intensities; see Section 2.5.5 for more details. The first column displays the three sequences along an axial slice, with MS lesions being visible around the ventricles, especially in the FLAIR and T1 images. The second column displays the corresponding voxel intensities, while the third column displays the voxel intensities in the contralateral part of the brain corresponding to the lesion. The contralateral area of the brain is used here as control. The fourth column provides the difference in voxel intensities between each voxel and its contralateral correspondent voxel.

The orange vertical line indicates the first time when lesion 14 was identified using the SuBLIME algorithm, [55] an algorithm designed to segment new and enlarging MS lesion, though the lesion may have occurred earlier than this visit. Indeed, before the lesion is observed there is around a 1.5 year gap between the MRI visits. Visits are indicated as a rug on the x-axis; notice that there are few visits in the beginning followed by many monthly visits after lesion 14 is detected. This is just one example of the data, and similar plots can now be obtained for all lesions. This raisies important scientific questions, such as: 1) how many patterns of intensities are there before and after the lesion is detected? 2) are there subtle changes in image intensities before the lesion is detected that could predict the lesion localization or timing? 3) are certain white matter areas with specific intensity patterns more prone to lesion formation?

2.7 Conclusions

We could have written a book. We probably should have written a book. After finishing this chapter, we realize that we barely scratched the surface of one of the most exciting areas of research in statistics and biostatistics. Far from being exhaustive, this chapter introduces fundamental analytic concepts in neuroimaging that are pertinent both to healthy subjects but, especially, to those who suffer from brain diseases. While the number of biostatisticans and statisticians who work in this area is incredibly small given the importance of the problem, those who "get hooked" become passionate about looking in the most quantitative way possible into the brain and dealing with complex analytic problems. Far from sending the message that this area of research is closed, our own perception is that research is just now starting. Important concepts have already been introduced and progress has been achieved. However, without knowing what has already been achieved the research community will simply reinvent the wheel.

Chapter 3

OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI

3.1 Introduction

Multiple sclerosis (MS) is an inflammatory disease of the brain and spinal cord characterized by demyelinating lesions that are most easily identified, at least on magnetic resonance imaging (MRI) studies, in the white matter of the brain [1]. Quantitive analyses of MRI, such as the number and volume of lesions, are essential for for diagnosing the disease and monitoring its progression [2, 3]. MRI measures are also a common primary endpoint in phase II immunomodulatory drug therapy trials [70]. In these trials, either manual or semi-automated segmentations are used to compute the total number of lesions and the total lesion volume [65]. Manual delineation is challenging as three-dimensional information from several MRI modalities must be integrated [71]. Manual assessment of MRI is also prone to large inter- and intra- observer variability [62]. While semi-automated methods have been found to decrease inter- and intra- rater variability, they still require manual reader input and are time consuming [63]. Therefore a sensitive and specific automated method to detect lesions in the brain is essential for the analysis of studies with a high numbers of MS patients.

A comprehensive review of currently available automated, cross-sectional MS lesion segmentation methods, or methods used to identify lesions from a single MRI study, is provided in [71]. We divide these methods into four categories: supervised classifier with an atlas, supervised classifier with no atlas, unsupervised classifier with an atlas, and unsupervised classifier with no atlas. We focus on supervised methods without atlases, as the method we propose is in this category. Supervised methods without atlases train on manually segmented images annotated by experts and use image intensities of MRI to classify lesions [71]. Supervised classification algorithms are applied to the volumes: artificial neural networks [72], spatial clustering [73], k-nearest neighbors [74, 75, 76], Parzen window [77], Parzen window and morphological grayscale reconstruction [78], Bayes [79], AdaBoost [80], simulated annealing and Markov random fields [81], and graph cuts [82]. All of the aforementioned methods except [76] use multi-modality MRI information to classify lesions. The most widely-used feature across all segmentation methods is voxel intensity, which derives strength from a multi-modality approach [71].

The method we propose uses a logistic regression model to assign voxel-level probabilities of lesion presence in structural MRI of patients with MS. Logistic regression models have been used for segmentation of brain tissues and pathology in MRI [83, 84, 85]. For applications to MS, logistic regression has been used for detection of gadolinium enhancing lesions [86], prediction of gadolinium enhancing lesions without administering contrast agents [87], and for segmentation of new and enlarging MS lesions [9]. To our knowledge logistic regression has not been used in cross-sectional segmentation of MS lesions in structural MRI.

One difficulty in automated segmentation of MRI is due to variable imaging acquisition parameters [71]. All of the segmentation methods reviewed in [71] have tuning parameters that are adjusted to a particular data set and may not generalize to a new data set with different acquisition parameters. These parameters are not informed by the data and therefore must be tuned empirically, often with little to no interpretability of the parameter. Application to a new data set may require several iterations of segmentations to adjust the tuning parameters to values that produce acceptable segmentations. A method in which the tuning parameters are informed by the data and for which adjustments are intuitive and simple would therefore be valuable.

A second difficulty in intensity-based segmentation is that MRI data are acquired in arbitrary units; units can vary widely between and within imaging centers. These variations are attributed to scanner hardware, interactions between hardware and patients, and variations in acquisition parameters [62].

Therefore, proper intensity normalization is essential in developing a generalizable segmentation method. Many of the segmentation methods use intensity-normalized volumes [65], but these methods do not demonstrate the generalizability of the normalization procedure to changes in imaging acquisition parameters and imaging centers. In [63] the authors performed a PubMed and Google Scholar search for MS lesion segmentation papers. Of the 47 papers that met their search criteria, only 13 of these papers used multicenter data for validation, and the largest database used for validation consisted of 41 subjects. To show generalizability, methods must be validated on multicenter data with many subjects.

A third difficulty is intensity inhomogeneity, the slow spatial intensity variations of the same tissue within an MRI volume. Inhomogeneity can significantly reduce the accuracy of image segmentation [29], and therefore some form of spatial normalization is necessary for accurate lesion segmentation. Most lesion segmentation methods assume that these inhomogeneities have been corrected during image preprocessing, but we have found strong spatial patterns within tissue type even after the N3 inhomogeneity correction algorithm [19] is applied.

To address these and related problems, we propose OASIS is Automated Statistical Inference for Segmentation (OASIS), a fully automated, generalizable, and novel statistical method for cross-sectional MS lesion segmentation. Using intensity information from multiple modalities of MRI, a logistic regression model assigns voxel-level probabilities of lesion presence. After training on manual segmentations, the OASIS model produces interpretable results in the form of regression coefficients that can be applied to imaging studies quickly and easily. OASIS uses intensity-normalized brain MRI volumes, enabling the model to generalize to changes in scanner and acquisition sequence. OASIS also adjusts for intensity inhomogeneities that preprocessing bias field correction procedures do not remove, using smoothed volumes. This allows for more accurate segmentation of brain areas that are highly distorted by inhomogeneities, such as the cerebellum. One of the most practical properties of OASIS is that the method is fully transparent, easy to explain and implement, and simple to modify for new data sets.

To illustrate the generalizability of OASIS to changes in imaging acquisition parameters, we evaluated the performance of the algorithm on a total of 300 MRI studies from two separate imaging centers with varying acquisition parameters. This is a crucial criterion for assessing the generalizability and utility of the method.

3.2 Materials and methods

In this section we introduce OASIS, a method inspired by Subtraction Based Inference for Modeling and Estimation (SuBLIME), an automated method for the longitudinal segmentation of incident and enlarging MS lesions [9]. Before the OASIS logistic regression model is fit, a brain tissue mask is created, all MRI volumes are intensity normalized, and smoothed volumes are created to capture local spatial information and adjust for remaining field inhomogeneities. The OASIS method involves two iterations of model fitting: the first to perform an initial lesion segmentation and the second to use this initial lesion segmentation to remove lesions, which can distort the smoothed volumes. After the final model is fit, the regression coefficients are applied to produce three dimensional maps of voxel-level probabilities of lesion presence.

We evaluate the performance of OASIS on MRI volumes of the brain acquired with various acquisition protocols. We use data sets from two different imaging centers for validation, which we refer to as validation set 1 and validation set 2. validation set 1 has manual lesion segmentations. We trained the OASIS method on a subset of the studies in this dataset, and tested on the remaining studies. An expert evaluated the segmentations from validation set 1. validation set 2 is used to demonstrate generalizability to changes in image acquisition parameters. We applied the coefficients from the model trained on validation set 1 to the studies in validation set 2, and experts evaluated the OASIS lesion segmentations.

3.2.1 Study population

validation set 1 contains a total of 131 MRI studies from 131 subjects. Of these studies, 98 are from patients with MS and 33 are healthy volunteer scans. Of the 98 patients with MS, the median age is 44 years (IQR: [33, 54]), 72 are female (26 male), and the median EDSS is 3.5 (IQR:[2, 6]). The median age of the healthy volunteers is 34 (IQR: [28, 42]) and 19 are female (14 male).

validation set 2 contains a total of 169 MRI studies from 149 subjects. Twenty subjects in validation set 2 have baseline and follow-up scans. The mean time between baseline and follow-up for these 20 subjects is 132 days (IQR: [51, 182]). The subjects in the validation set are a mixture of healthy volunteers and patients: 110 of the patients have MS, 38 have other neurological diseases, and one is a healthy volunteer. The median age of the MS patients is 42 (IQR: [33,50]); 54 are female (56 male); 68 have relapsing remitting MS, 31 have primary progressive MS, and 11 have secondary progressive MS. The median age of the patients with other neurological diseases is 41 years, (IQR: [35, 51]) and 8 are female (30 male). The healthy volunteer is a 28 year old female.

3.2.2 Experimental methods

T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR) and proton density (PD) volumes were acquired for all subjects at each study, and all imaging protocols were approved by local institutional review boards. For validation set 1, 3D T1-MPRAGE images (repetition time (TR) = 10 ms; echo time (TE) = 6ms; flip angle (FA) $\alpha = 8°$; inversion time (TI) = 835 ms, resolution = 1.1 mm × 1.1 mm × 1.1 mm), 2D T2-weighted pre-contrast FLAIR images (TR = 11000 ms; TE = 68 ms; TI = 2800 ms; in-plane resolution = 0.83 mm × 0.83 mm; slice thickness = 2.2 mm), T2-weighted and PD images (TR = 4200 ms; TE =12/80 ms; resolution = 0.83 mm × 0.83 mm × 2.2 mm) were acquired on a 3 tesla MRI scanner (Philips Medical Systems, Best, The Netherlands).

For validation set 2, the 3D T2-weighted post-contrast FLAIR was acquired using a variable flip angle sequence, the 2D PD and T2-weighted volumes using a dual-echo fast-spin-echo sequence, and the 3D T1-weighted volume using an inversion-prepared fast spoiled gradient-echo sequence. These studies were acquired on a single 3 tesla MRI scanner (Signa Excite HDxt; GE Healthcare, Milwaukee, Wisconsin). Table 3.1 contains the ranges for the validation set 2 scanning parameters.

Table 3.1: Ranges for validation set 2 scanning parameters

	FA (degrees)	TR (ms)	TE (ms)	TI (ms)
FLAIR	90	(4800, 8802)	(124.3, 151.4)	(1481, 2200)
T2-weighted	90	5317	(116.2, 124.2)	NA
PD	90	5317	(16.0, 23.7)	NA
T1-weighted	(6,13)	(8.7, 9.1)	(3.2, 3.6)	(450, 725)

3.2.3 Image preprocessing

Before building our statistical model for the lesion segmentation, we preprocessed the images from validation set 1 and validation set 2 using the tools provided in Medical Image Processing Analysis and Visualization (MIPAV) [28], TOADS-CRUISE (http://www.nitrc.org/projects/toads-cruise/), and Java Image Science Toolkit (JIST) [27] software packages. We first rigidly aligned the T1-weighted image of each subject into the Montreal Neurological Institute (MNI) standard space (voxel resolution 1mm^3). We then registered the FLAIR, PD, and T2-weighted images of each subject to the aligned T1-weighted images. We also applied the N3 inhomogeneity correction algorithm [19] to all images and removed extracerebral voxels using SPECTRE, a skull-stripping procedure [88].

3.2.4 Statistical modeling and spatial smoothing

We performed all statistical modeling in the R environment (version 2.12.0, R Foundation for Statistical Computing, Vienna, Austria) with the packages AnalyzeFMRI [89], biglm [90], ff [91], and ROCR [92]. We used the FSL tool fslmaths (http://www.fmrib.ox.ac.uk/fsl) for the three dimensional spatial smoothing of the volumes.

3.2.5 Brain tissue mask

The first step in OASIS is to create a mask of the brain that excludes cerebrospinal fluid (CSF). CSF is excluded because it disrupts the capture of the inhomogeneity field and distorts the representation of the local cerebral features when creating smoothed volumes. To make this mask, we used the extracerebral voxel removal mask described in the Image Preprocessing Section and excluded voxels in the mask that appear hypointense in the FLAIR volume. Because CSF is hypointense in the FLAIR, we empirically found that excluding voxels falling below the bottom 15th percentile of FLAIR intensities over the extracerebral voxel removal mask removes CSF outside of the brain and in the ventricles. We refer to this mask as the brain tissue mask. Figure 3.1B shows a slice of the brain tissue mask for a particular subject for illustration.

3.2.6 Intensity normalization

We used intensities from the FLAIR, PD, T2-weighted, and T1-weighted volumes to identify the presence of MS lesions. We denote the observed intensity of voxel v, for subject i by:

$$M_i^0(v), M = FLAIR, PD, T2, T1$$

where M indicates the imaging sequence.

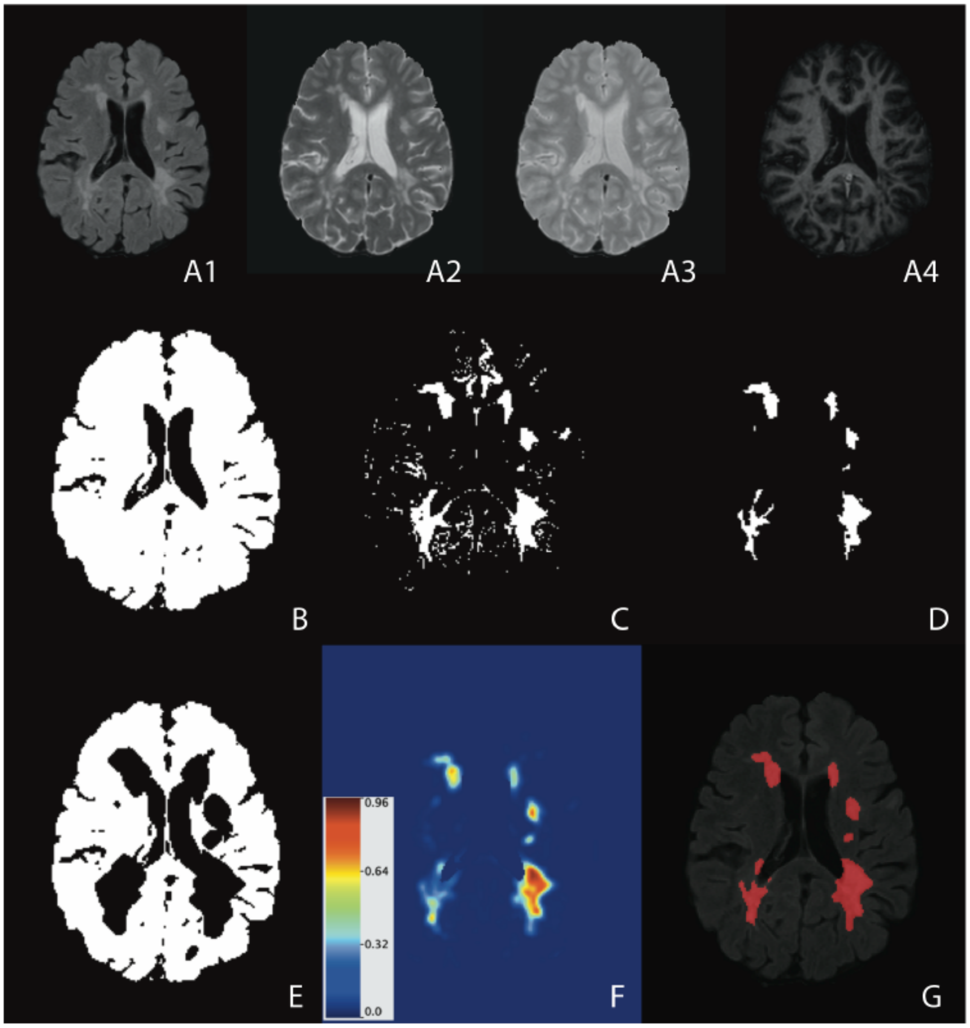


Figure 3.1: A. Axial slice from different modalities of intensity normalized brain MRI of a single subject: A1. FLAIR image. A2. T2-weighted image. A3. PD image. A4. T1-weighted image. B. Brain tissue mask of an axial slice of the brain. C. Axial slice of select voxels for OASIS modeling. D. Manual lesion segmentation of an axial slice of the brain. E. Axial slice of brain tissue mask with dilated lesion mask made at a false positive rate of 1% removed. F. Axial slice of the smoothed probability map with intensity scale. G. Binary segmentation of the probability map from the OASIS model at false positive rate of .005 overlaid on the FLAIR image.

MRI volumes are acquired in arbitrary units. Analyzing images across subjects and imaging centers requires that images be normalized so that voxel intensities have common interpretations. For normalization, we adapt the normalization method from [93] to normalization with respect to the brain tissue mask. The normalized intensity of voxel v, for subject i is denoted by:

$$M_i^N(v) = \frac{M_i^0(v) - \mu_{i,M}^0}{\sigma_{i,M}^0}$$

where $\mu_{i,M}^0$ and $\sigma_{i,M}^0$ are the mean and standard deviation of the observed voxel intensities in the brain tissue mask of subject i, from sequence M. The normalized voxel intensities are standard scores of the brain tissue mask. Figure 3.1A shows a slice of the normalized images from all four modalities from a single subject with MS: FLAIR, T2-weighted, PD, and T1-weighted.

3.2.7 Smoothed volumes

To account for intensity inhomogeneities that remain after initial inhomogeneity correction, we use a sequence of multiresolution smoothed volumes, obtained using different levels of smoothing. The smoothed volumes are created by three dimensional smoothing of the normalized volume from each modality over the brain tissue mask. A Gaussian smoother with relatively large kernel window size is used to smooth over the features in the brain and capture the pattern of the remaining inhomogeneity.

For subject i and imaging modality M, let k be the size of the kernel window. Then the intensity in voxel v of the smoothed volume of imaging modality M is expressed as $\mathcal{G}M_i^N(v,k)$. The smoothed volumes are used in the OASIS model to incorporate spatial information and to account for inhomogeneities in the brain that persist after N3 correction. For OASIS we use smoothed volumes as covariates with kernel window sizes of 10 and 20 voxels, which were found empirically on validation set 1 to work well. Figure 3.2 shows the smoothed volumes for both kernel window sizes of 10 and 20 from each modality. The kernel window size of 20 smooths over the anatomical features almost completely, while the kernel window size of 10 still preserves some of these features, such as the hyperintesities of the gray matter in the FLAIR, T2-weighted, and PD volumes and hypointensities of the gray matter in the T1-weighted volume.

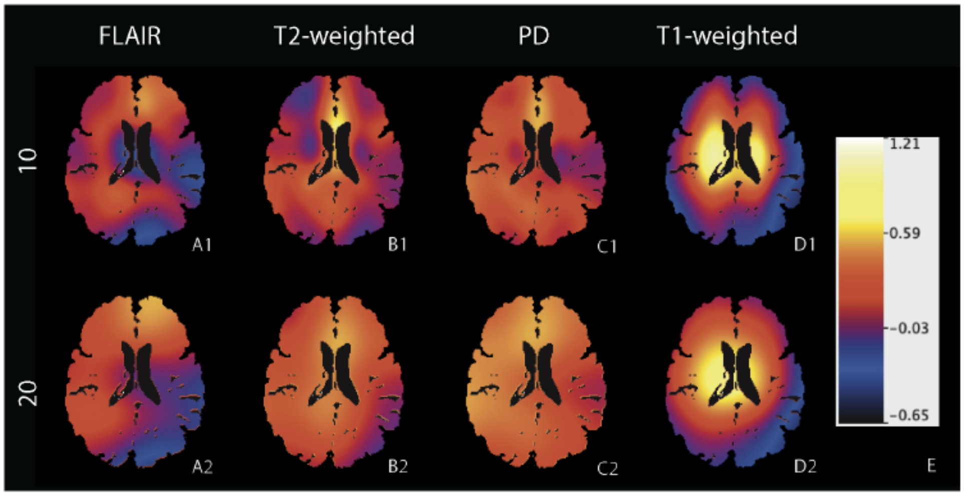


Figure 3.2: Axial slice from a a single subject of the smoothed volumes from all modalities. Row one contains the smoothed volumes with kernel window size of 10 and row two contains the smoothed volumes with kernel window size of 20. Column A contains the FLAIR images, B contains the T2-weighted images, C contains the PD images and D contains the T1-weighted images. To link the figure with the notation used in this paper: A1. $\mathcal{G}FLAIR_i^N(v, 10)$; A2. $\mathcal{G}FLAIR_i^N(v, 20)$; B1. $\mathcal{G}T2_i^N(v, 10)$; B2. $\mathcal{G}T2_i^N(v, 20)$; C1. $\mathcal{G}PD_i^N(v, 10)$; C2. $\mathcal{G}PD_i^N(v, 20)$; D1. $\mathcal{G}T1_i^N(v, 10)$; D2. $\mathcal{G}T1_i^N(v, 20)$; E. Scale of intensities in the smoothed volumes.

3.2.8 OASIS is Automated Statistical Inference for Segmentation

In this section we introduce the OASIS model. OASIS uses logistic regression to model the probability that a voxel is part of a lesion. We choose logistic regression because it is extremely simple and easy to interpret. We model lesions at the voxel level using FLAIR, PD, T2-weighted, and T1-weighted intensities as well as the intensities from the smoothed volumes of each modality with kernel window sizes of 10 and 20 voxels. The model must be trained on a gold standard measure of lesion presence. Figure 3.1D is an example of manual lesion segmentation, which is an appropriate gold standard measure for the OASIS model. The result of our model is a collection of coefficients that can be used to create three-dimensional maps of the probabilities of lesion presence. OASIS obtains the estimated logit of the probability of each voxel being part of a lesion by weighting these 12 images (the four imaging modalities and smoothed volumes for each modality) with the coefficients.

The first step of the modeling procedure is to select candidate voxels to minimize false positives and computation time. Lesions appear as hyperintensities in the FLAIR volume. The brain tissue mask was applied to the FLAIR volume, and the 85th percentile and above of voxels in the brain tissue mask were selected as candidate voxels for lesion presence. In validation set 1, there were a total of 1,093,394 lesion voxels (a volume of 1,093 cm^3). The voxel selection procedure excluded 64,556 (6%) of these voxels, but lowered the searchable area to 15% of the original size. This procedure also decreases the number of potential false positive voxels. Using this threshold also significantly decreases the number of voxels the model must be fit on, allowing for a much faster fit. Figure 3.1C shows a slice of the voxel selection mask for a single subject.

We then fit a voxel-level logistic regression model over the candidate voxels. In the OASIS model, the probability that a voxel is part of a lesion is represented as $P\{L_i(v) = 1\}$, where L is a random variable denoting voxel-level lesion presence. If there is a lesion in voxel v for subject i, then $L_i(v) = 1$. Otherwise, $L_i(v) = 0$. The probability that a voxel v contains lesion incidence is modeled with the following logistic regression model:

$$\begin{array}{lclcll}
logit[P\{L_i(v)=1\}] & = & \beta_0 & + & & \\
\beta_1 FLAIR_i^N(v) & + & \beta_2 \mathcal{G}FLAIR_i^N(v,10) & + & \beta_3 \mathcal{G}FLAIR_i^N(v,20) & + \\
\beta_4 PD_i^N(v) & + & \beta_5 \mathcal{G}PD_i^N(v,10) & + & \beta_6 \mathcal{G}PD_i^N(v,20) & + \\
\beta_7 T2_i^N(v) & + & \beta_8 \mathcal{G}T2_i^N(v,10) & + & \beta_9 \mathcal{G}T2_i^N(v,20) & + \\
\beta_{10} T1_i^N(v) & + & \beta_{11} \mathcal{G}T1_i^N(v,10) & + & \beta_{12} \mathcal{G}T1_i^N(v,20) & \quad [1] \\
 & + & \beta_{13} FLAIR_i^N(v) * \mathcal{G}FLAIR_i^N(v,10) & + & \beta_{14} FLAIR_i^N(v) * \mathcal{G}FLAIR_i^N(v,20) & \\
 & + & \beta_{15} PD_i^N(v) * \mathcal{G}PD_i^N(v,10) & + & \beta_{16} PD_i^N(v) * \mathcal{G}PD_i^N(v,20) & \\
 & + & \beta_{17} T2_i^N(v) * \mathcal{G}T2_i^N(v,10) & + & \beta_{18} T2_i^N(v) * \mathcal{G}T2_i^N(v,20) & \\
 & + & \beta_{19} T1_i^N(v) * \mathcal{G}T1_i^N(v,10) & + & \beta_{20} T1_i^N(v) * \mathcal{G}T1_i^N(v,20) &
\end{array}$$

The effect of magnetic field inhomogeneities are thought to be multiplicative, so we use the interactions between the normalized volume and the smoothed volume in the model.

3.2.9 OASIS model refinement

The second iteration of the OASIS model fitting is done to reduce the influence of lesions in the smoothed volumes. First, we fit the model and use the estimated coefficients to create maps of the estimated probability of lesion presence at each voxel. To incorporate spatial information of the neighboring voxels and reduce noise, we smooth the estimated probabilities from the model using a Gaussian kernel with window size of 3 mm. This kernel size was empirically chosen and found to perform well. The resulting probability maps were then thresholded using a liberal false positive rate of 1% (threshold value of 0.10), which resulted in model based hard segmentations of lesions. These lesions masks were then dilated by 5 voxels to ensure that the entire lesion was captured and removed from the brain tissue mask. Figure 3.1E shows the brain tissue mask with the lesions removed. New smoothed volumes were created by applying a Gaussian smoother with kernel window sizes of 10 and 20 to the normalized image from each modality over the brain tissue mask with the lesions removed. We inpainted the smoothed volumes to fill the places where lesions were removed with the values we would expect in this area if it were occupied by normal, healthy tissue.

The intensity in voxel v of the normalized image after the second Gaussian smoother has been applied is labeled as, $\mathcal{G}^2 M_i^N(v,k)$. Figure 3.3 shows an axial slice for a subject of the FLAIR volume and the smoothed volume for this image with kernel window sizes of 10 and 20 before and after the lesions were removed. To link the figure with the notation, Figure 3.3A shows $FLAIR_i^N(v)$, Figure 3.3B shows a scale of intensities in the smoothed volumes, Figure 3.3C1 shows $\mathcal{G}FLAIR_i^N(v,10)$, Figure 3.3C2 shows $\mathcal{G}^2FLAIR_i^N(v,10)$, Figure 3.3D1 shows $\mathcal{G}FLAIR_i^N(v,20)$, and Figure 3.3D2 shows $\mathcal{G}^2FLAIR_i^N(v,20)$. The lesions are captured in the first smoothed volume, especially with the kernel size of 10, but are not captured in the second smoothed volume. The model [1] was refit over the same voxels using the second smoothed volume to obtain the final coefficients that are used to create the final probability maps. Again, the final estimated probabilities are smoothed using a Gaussian kernel with window size of 3 mm. Figure 3.1F shows a slice of the probability map for a subject and a scale of intensities. Red indicates areas with a higher probability of being a lesion and blue indicates areas with a lower probability of being a lesion.

3.2.10 Probability map and binary segmentation

Using this fitted model to generate a probability map for the entire brain from a set of new images takes about 30 minutes for each study using a standard workstation. The Gaussian smoothing is the slowest step of the algorithm and takes approximately one minute for each volume. These computations can be parallelized to take substantially less time; the entire algorithm can be run in approximately 5 minutes with 8 cores. To make a probability map for a new study, the two sets of regression coefficients, a brain mask, and the FLAIR, PD, T2-weighted, and T1-weighted volumes are required. Using population-level

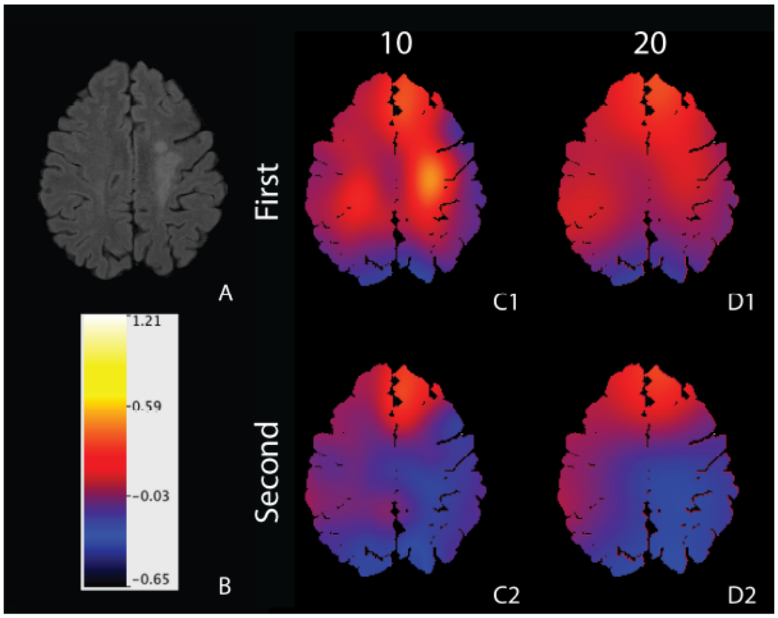


Figure 3.3: Axial slice of the FLAIR volume and the first and second smoothed volumes created from the FLAIR image for a single subject. To link the figure with the notation used in this paper: A. $FLAIR_i^N(v)$ B. Scale of intensities in the smoothed volumes C1. $\mathcal{G}FLAIR_i^N(v,10)$; C2. $\mathcal{G}^2FLAIR_i^N(v,10)$; D1. $\mathcal{G}FLAIR_i^N(v,20)$; D2. $\mathcal{G}^2FLAIR_i^N(v,20)$.

thresholds, the probability maps from OASIS can be used to create hard segmentations of lesion presence. Figure 3.1G shows a slice of a hard segmentation overlaid on the FLAIR image.

3.2.11 Validation with gold standard: validation set 1

validation set 1, described in detail in the Materials section, consists of 131 MRI studies: 98 studies from MS subjects and 33 studies from healthy subjects. To fit the model and to measure performance, we required a set of data in which the outcome is assessed using a gold standard measure. The gold standard was obtained using manual segmentation by a technologist with more than

10 years of experience in delineating white matter lesions. The technologist spent between 30 minutes to an hour segmenting each study, depending on the lesion load and distribution. The majority of the studies had at least moderate pathology and therefore took between 45 minutes to an hour. The segmentations were made from the FLAIR and T1-weighted volumes. Figure 3.1D shows a manually segmented slice for a subject. The mean volume of lesions for MS subjects in validation set 1 is 11.2 cm^3 (IQR: [1.7 cm^3, 16.6 cm^3]). It was assumed that the healthy subjects did not have any lesions.

To evaluate performance of our model within validation set 1, we trained the model [1] on 20 randomly selected subjects (15 MS subjects and 5 healthy subjects) and tested on the remaining 111 subjects (83 MS subjects and 28 healthy subjects). We used only the studies from the 111 subjects in this test set to estimate the voxel-level receiver operator characteristic (ROC) curve and area under the curve (AUC). These performance measures are known to be susceptible to instability. To account for this, we nonparametrically bootstrapped with replacement the subjects to the training and testing sets. We then fit the model on the training set and observed the performance of the model in the testing set.

It is known that the full AUC summarizes test performance over regions of the ROC space that are not clinically relevant for lesion segmentation [9]. Once a test has been able to distinguish well between disease and not disease, the performance of the test for particular applications must be evaluated, in which case one may be interested in only a small portion of the ROC curve [94]. In this particular application we are interested in using the lesion segmentation to identify lesions and to provide accurate estimations of lesion volumes. The mean lesion volume of manual lesion segmentations from validation set 1 is 11.2 cm^3 (IQR [1.7 cm^3, 16.6 cm^3]). For the entire brain, a false positive rate of .01 would correspond to a volume of 12.8 cm^3 of healthy brain being falsely identified as lesion, which is more than the mean lesion volume in validation set 1. Therefore we examined only false positive rates below 1%. We provide the partial ROC curve with bootstrapped 95% confidence bands for clinically relevant false positive rates of 1% and below.

3.2.12 Validation with expert rankings: validation set 1 and validation set 2

For the studies in validation set 2, gold standard segmentations were not available. To evaluate the performance of OASIS on validation set 2, three experts (a neuroradiologist, neurologist, and radiologist) compared OASIS segmentations to those from LesionTOADS, an open-source lesion segmentation software (http://www.nitrc.org/projects/toads-cruise/), [95, 96, 66]. validation set 2, described in detail in the Materials section, consists of 169 MRI studies of 149 subjects, 20 of whom had follow-up visits. These studies were acquired using a variety of imaging protocols.

For the OASIS algorithm, the only parameter that must be tuned when moving to a new dataset is the population-level threshold. For validation set 2 we used the coefficients that were trained on validation set 1 and then empirically adjusted the population level threshold for validation set 2. To adjust this threshold, we randomly sampled 10 subjects from validation set 2. We applied thresholds between 0.10 and 0.50 (by increments of 0.05) to the probability maps, examined the segmentation, and empirically chose a threshold of 0.35 for validation set 2. This threshold adjustment is very fast and transparent. We ran the segmentations for the 10 subjects in parallel, and each segmentation took less than 5 minutes. Next, we thresholded the probability maps at the 9 different thresholds, which took only seconds. Last, we looked through the segmentations and the original images to select the optimal (most reasonable) threshold, which took only about a minute for each subject. The entire process of tuning the threshold took less than an hour and involved only 10 minutes of manual image examination. This procedure only needs to be performed once when moving to a new imaging center or study. For the segmentation comparison, we presented the three experts with segmentations at the threshold value of 0.35 on all of the images in validation set 2 as well as at the threshold from validation set 1 with a false positive rate of 0.005, a threshold value of 0.16. We will refer to the threshold value of 0.35 as the empirically adjusted threshold and the threshold value of 0.16 as the validation set 1 threshold.

We compared both OASIS segmentations to the segmentations produced by the open source software LesionTOADS. We ran LesionTOADS with T1-weighted and FLAIR inputs and the default parameters. We adjusted the smoothing parameter from 0.2 to 0.4 because we empirically found this to improve the quality of the segmentations. It is important to note that LesionTOADS not only segments lesions, but also segments the other tissue classes of the brain. For this analysis, we only used the lesion segmentations.

We designed an image rating system to evaluate the performance of the two segmentation algorithms. For each of the 169 studies, we had three segmentations: the LesionTOADS segmentation, the OASIS segmentation with the threshold from validation set 1, and the OASIS segmentation with the empirically adjusted threshold. We also randomly selected 20 of the MRI studies and created duplicates of these to assess rating reliability, for a total of 189 studies. We randomized the order in which the segmentations were presented to the experts and randomly assigned each segmentation a letter: A, B, or C, so as to blind the rater to the segmentation algorithm.

We presented each of the 189 MRI studies to an experienced MS neuroradiologist. For each study, the neuroradiologist examined the set of three segmentations along with the original FLAIR, PD, T1-weighted, and T2-weighted volumes. The neuroradiologist then scored the performance of each of the segmentations on a continuous scale from 0 to 100, with 0 being an unusable lesion segmentation and 100 being a perfect segmentation. The neuroradiologist was presented all three segmentations simultaneously, so that scores were assigned relative to one another. Fifty of the studies were selected to be scored with the same system by a neurologist with a subspecialty in MS and a general radiologist in order to assess rater agreement among the three raters. The 50 studies were comprised of 45 randomly selected studies with 5 of the studies repeated to assess rater reliability.

The neuroradiologist also compared and scored the OASIS and Lesion-TOADS segmentations from the studies for the 98 MS patients in validation set 1. This allows for comparison of the performance of the segmentations on validation set 1 and validation set 2.

3.3 Experimental results

3.3.1 Validation set 1: training with gold standard

The OASIS model has an estimated full AUC of 98% (95% CI; [96%, 99%]) and a partial AUC for clinically relevant false positive rates of 1% and below of 0.59% (95% CI; [0.50%, 0.67%]) in the test set. Figure 3.4 shows the voxel-level partial ROC curve for the test set with bootstrapped 95% confidence bands for clinically relevant false positive rates. The probability map threshold that corresponds to a false positive rate of 1% is 0.10. The vertical axis of the partial ROC curve shows the true positive rate (sensitivity) for thresholds between 0 and 0.10 of the probability map and the horizontal axis shows the false positive rate (1 - specificity) for these thresholds.

The coefficients from fitting the logistic model [1] over all 131 studies in validation set 1, a total of 24 million voxels, are reported in the Appendix. The coefficients from the first and second fit of the model are provided. We also assessed the variation in the coefficients by nonparametrically bootstrapping the subjects with replacement. The bootstrapped 95% confidence intervals for the coefficients can be found in the Appendix. The variance of these coefficients is large in comparison to the estimates of the coefficients. The instability in the coefficients does not impact the performance of OASIS, as illustrated in the stability of the partial ROC curve.

Choosing a final threshold value after the second probability maps are made is a tradeoff between sensitivity and specificity. OASIS is flexible, and the

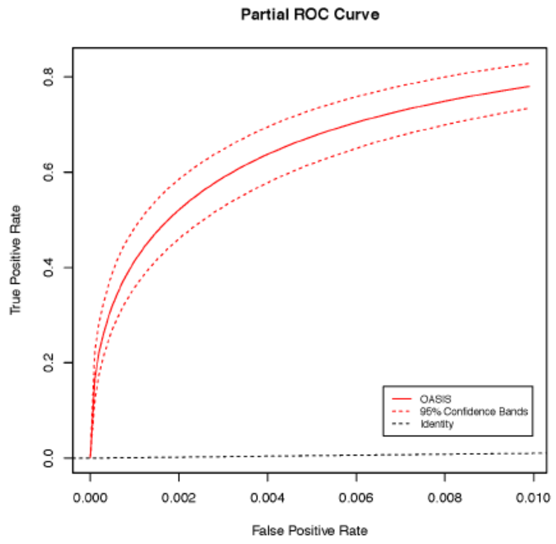


Figure 3.4: Partial ROC curve for the voxel-level detection of lesions in the testing set of validation set 1 for different thresholds of the probability maps produced from OASIS for clinically relevant false positive rates of 1% and below. Bootstrapped 95% confidence bands are also provided. The vertical axis of the partial ROC curve shows the true positive rate (sensitivity) for a given threshold of the probability map and the horizontal axis shows the false positive rate (1 - specificity) for this threshold.

appropriate false positive rate may be selected for a particular application. Table 3.2 shows the threshold values, sensitivity, and dice similarity coefficient [97] for four different false positive rates for the model fit over all of the studies in validation set 1. OASIS detected lesions in many of the healthy subjects. Table 3.3 shows the mean volume of false positive lesions detected in the healthy and MS subjects for the four threshold values from Table 3.2. The volume of false positives for both the MS and healthy subjects is comparable.

Table 3.2: Binary segmentation thresholds with false positive rate, sensitivity and DSC for validation set 1

False Positive Rate	Sensitivity	Threshold Value	DSC
1 %	80%	0.10	0.55
0.75%	76%	0.12	0.58
0.5 %	69%	0.16	0.61
0.25%	58%	0.23	0.59

Table 3.3: Volume of false positive lesion in healthy volunteers and MS subjects from validation set 1 (in cm^3); the actual mean lesion volume is 0 cm^3 for healthy volunteers and 11.2 cm^3 (IQR: [1.7 cm^3, 16.6 cm^3]) for MS subjects

Threshold Value	Healthy Mean (IQR)	MS Mean (IQR)
0.10	8.6 (4.6, 10.6)	10.9 (7.6, 13.6)
0.12	6.7 (3.1, 8.2)	8.0 (5.2, 10.3)
0.16	4.3 (1.5, 5.7)	5.2 (3.0, 7.0)
0.23	2.2 (.7, 2.8)	2.5 (1.2, 3.5)

3.3.2 Validation set 1: neuroradiologist rating results

For the neuroradiologist rankings of the OASIS and LesionTOADS segmentations for the 98 MS subjects in validation set 1, we performed a paired t-test to assess the difference in the means of the OASIS segmentations and the LesionTOADS scores. This difference was found to be 12.6, with a 95% confidence interval of (9.6, 15.8), p-value $< 10^{-12}$. The OASIS empirical threshold was ranked higher than LesionTOADS segmentation in 73 (95% CI: [64, 81]) of the 98 studies or 74% (95% CI: [65%, 82%]). We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the rankings.

3.3.3 Validation set 2: neuroradiologist rating results

Table 3.4 contains summary statistics for the scores from the neuroradiologist ratings of the three segmentations for all 189 studies. The OASIS validation set 1 threshold segmentations and the LesionTOADS segmentations have a much lower first quantile than the OASIS empirical threshold segmentations. For this analysis we focus mainly on the difference between the OASIS empirical threshold and the LesionTOADS segmentation, as the OASIS validation set 1 threshold did not perform well on this new data set. This was expected, as the probability map threshold needs to be adjusted to maintain the same false positive rate when moving to a new data set. We performed a paired t-test to assess the difference in the means of the OASIS empirical threshold scores and the LesionTOADS scores. This difference was found to be 16.6, with a 95% confidence interval of (13.3, 20.0), p-value $< 10^{-14}$. The OASIS empirical threshold was ranked higher than LesionTOADS segmentation in 146 (95% CI: [135, 157]) of the 189 cases or 77% (95% CI: [71%, 83%]). We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the rankings.

Table 3.4: Summary statistics of image ratings of validation set 1 for neuroradiologist on 189 studies

	OASIS	OASIS	LesionTOADS
	validation set 1 Threshold	Empirical Threshold	
Minimum	3.7	3.7	2.7
1st Quantile	27.3	55.7	21.7
Median	42.0	68.3	51.0
Mean	43.2	64.1	47.5
3rdQuantile	57.7	76.3	71.0
Maximum	99.3	99.0	97.3

To assess rater reliability among the 20 duplicated MRI studies, we calculated the intraclass correlation coefficient: 0.61 (95% CI: [0.69, 0.81]). The rankings for the LesionTOADS images and the OASIS empirical threshold were preserved in the duplicate rankings for 17 of the 20 images (95% CI: [14, 20]). We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for both the intraclass correlation coefficients and the rankings.

3.3.4 Validation set 2: rater agreement with neuroradiologist, neurologist, and radiologist

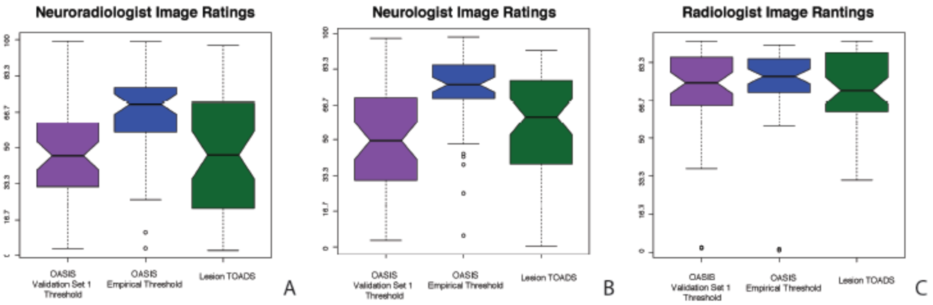


Figure 3.5: Notched box plot of the results from the neuroradiologist, neurologist, and radiologist image ratings for segmentations of the 50 MRI studies from validation set 2: the OASIS validation set 1 threshold segmentations, the OASIS empirically adjusted threshold segmentations, and the LesionTOADS segmentations.

Table 3.5 contains summary statistics for the scores from the neuroradiologist, neurologist, and radiologist ratings of the three segmentations for the set of 50 studies selected to asses rater reliability. Figure 3.5 shows a notched box plot for each rater of these findings. From the box plot we see that there is a statistically significant difference between the medians for all three segmentations for the neuroradiologist and neurologist. There was not a statistically significant difference in the medians of the scores for the three segmentations by the radiologist. Moreover, all three raters indicated that the OASIS validation set 1 segmentations and the LesionTOADS segmentations have a much lower first quantile than the OASIS empirical threshold segmentations. The outliers in the boxplots can be explained as either errors in processing, such as registration or bad artifacts, or as studies that none of the segmentation methods performed well on. We did not remove these studies from the analysis, because

Table 3.5: Mean and standard deviation of the rating from the neuroradiologist, neurologist, and radiologist for OASIS validation set 1 threshold, OASIS empirical threshold and LesionTOADS on 50 studies from validation set 2; mean difference between OASIS empirical threshold and LesionTOADS and percentage of times OASIS was ranked higher than LesionTOADS on these images

	OASIS	OASIS	LesionTOADS	Mean	Percentage
	validation set 1	Empirical	Mean (SD)	Difference	Rank
	Mean (SD)	Mean (SD)		(95% CI)	(95% CI)
Neuroradiologist	46.3 (22.0)	66.1 (20.2)	47.3 (27.2)	18.7 (11.2, 26.3)	76% (64%, 88%)
Neurologist	48.7 (24.3)	73.1 (18.5)	56.6 (26.0)	16.5 (7.0, 25.9)	66% (52%, 78%)
Radiologist	71.6 (19.6)	74.1 (17.9)	71.8 (16.5)	2.3 (-4.2, 8.8)	52% (38%, 66%)

we want to assess the performance of OASIS in the setting of an image processing pipeline, where images may not be properly registered or may contain artifacts.

Again, we will focus mainly on the difference between the OASIS empirical threshold and the LesionTOADS segmentation. We performed a paired t-test to assess the difference in the means of the OASIS empirical threshold scores and the LesionTOADS scores. These differences can be found in Table 5. The mean for the OASIS empirical threshold was greater than the mean for the LesionTOADS scores for all three raters. This difference was found to be statistically significant for both the neuroradiologist and neurologist, (p-values $< 10^{-4}$ and $< 10^{-3}$, respectively), but not for the radiologist, (p-value 0.5). The neuroradiologist and the neurologist tended to spread their scores more, and this allowed better comparison of the segmentation algorithms. Table 5 also shows the percentage of time the OASIS empirical threshold was ranked higher than LesionTOADS segmentation in the 50 studies. We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the rankings.

To assess rater reliability among the 5 duplicated MRI studies, we calculated the intraclass correlation coefficient and the number of times the rankings for the LesionTOADS images and the OASIS empirical threshold were preserved. We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for both the intraclass correlation coefficients and the rankings. For the neuroradiologist, the intraclass correlation coefficient for the 5 repeated studies is 0.55 (0.21, 0.82) and the number of preserved rankings is 4 (2,5). For the neurologist, 0.32 (-0.10, 0.68) and 4 (2,5). For the radiologist, -0.38 (-0.35, 0.71) and 2 (0,4). The repeated rankings for each rater for the 5 subjects is reported in the Appendix.

We calculated the rater agreement for the ranking of the OASIS empirical threshold versus LesionTOADS. We decided to use the rankings of the scores to assess rater agreement rather than the scores themselves, because, as shown from the intraclass correlation coefficient, the scores are not very reliable, while the order in which the observers rank the segmentations, on the other hand, is quite reliable. We calculated the kappa statistic to assess the reliability of the rankings for each pair of raters and nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the kappa statistics. The kappa statistic for the rater agreement between the neuroradiologist and the neurologist was 0.47 (0.20, 0.75), the neuroradiologist and radiologist 0.02 (-0.26, 0.30) and the neurologist and radiologist -0.09 (-0.37, 0.19).

3.4 Discussion

OASIS may be used to assist or even replace manual segmentation of MS lesions in the brain. After training and adjustment of the population level threshold, our fully automatic method does not require human input and avoids the variability introduced by manual segmentation. Using the explicit form of the statistical model, OASIS can easily be adapted and trained for cases where more or fewer imaging sequences are available.

With the OASIS model, a recalibration of the population-level segmentation threshold is necessary for each new data set but can be done on a fairly limited number of subjects, as in the example from this paper. a recalibration of the population-level segmentation threshold is necessary for each new data set but can be done on a fairly limited number of subjects, as in the example from this paper. A set of subjects is required to tune this population level threshold, therefore fully automatic segmentation of a single study from a new imaging center may not feasible with the OASIS model. However, in these cases the threshold can be adjusted very quickly manually (2-5 minutes) by visual inspection of 3-4 slices by adjusting just one parameter. When using an ROC curve for classification, thresholds for subpopulations with different covariate values may need to be defined differently in order to keep false positive rates the same across those subpopulations [98]. Therefore, it was expected that the ROC threshold would need to be adjusted to maintain the same false positive rate from validation set 1 in validation set 2. This threshold is the only tuning parameter in OASIS that must be adjusted when moving to a new data set, and this adjustment is very fast and intuitive to make and does not require multiple iterations of segmentations. We believe that OASIS holds promise for use in multicenter MRI studies, with adjustment of the population level threshold for each site.

Future work includes further validation of OASIS under changes in imaging center and protocol and to also show the reproducibility of the OASIS segmentations. One resource for this is the MS Lesion Segmentation Challenge [67], a common database for MS lesion segmentation algorithms. We plan to do further validation with this database as well as with volumes from additional imaging centers. For this analysis we did not have scan-rescan MRI available. These are crucial for assessing the reproducibility of the method, and we plan to acquire these in the future.

In contrast to many automatic segmentation techniques, OASIS is computationally fast. While training the model on the 131 studies from validation set 1 takes five hours on a standard workstation, this process is only conducted once. The results from this are summarized as the two sets of 21 coefficients in model [1]. Also, the model may be trained on fewer studies, as shown in the partial ROC analysis within validation set 1; the performance of the model remains stable when trained on subsets of 20 studies. Using this fitted model to generate a probability map of the entire brain from a set of new images takes only 30 minutes. These times are for standard workstations and are expected to drop dramatically with multi-core parallel computing and improved technologies. The Gaussian smoothing is the slowest step of the algorithm, and these computations can be parallelized to substantially decrease the time of the entire algorithm to approximately 5 minutes.

After making the image ratings for validation set 2, the neuroradiologist was unblinded and reviewed the three segmentations, providing comments about the strengths and weaknesses of each. The OASIS empirical threshold performed much better than the OASIS validation set 1 threshold. The neuroradiologist reported a preference for the smoothness of the OASIS segmentations in contrast to the LesionTOADS segmentation, which often appeared speckled. The OASIS segmentations often had artifacts in the pineal glands and the choroid plexus of the ventricles. This may be explained by the fact that OASIS was trained on FLAIR images acquired before a gadolinium-based contrast agent was administered to the patient, while the validation was done with FLAIR images that were acquired after gadolinium administration. Voxels in the choroid plexus and pineal glands, which enhance with gadolinium, were brighter and were thus misclassified as lesion. LesionTOADS does not make a similar error, as it imposes topological constraints that preclude these structures from being identified as lesions. Further refinements of OASIS may account for such complex changes of protocol. The LesionTOADS segmentations were more variable than those of OASIS and did not perform well on cases with low lesion load. The OASIS segmentation had systematic errors in the medial frontal cortex and the brainstem. On the other hand, LesionTOADS avoided false positives in the brainstem because it only segments lesions in the cerebrum. Figure 3.6 shows a slice from a subject with an example of a lesion that OASIS segments in the cerebellum. Figure 3.6A shows a single slice of the FLAIR volume, Figure 3.6B shows a single slice of the T1-weighted volume, Figure 3.6C shows the LesionTOADS segmentation of the slice, and Figure 3.6D shows the OASIS segmentation of the slice. LesionTOADS does not segment the cerebellum, whereas OASIS does not restrict the areas that it segments and is able to find the lesion in this slice.

OASIS is not an atlas-based method and therefore does not take into account anatomical information during segmentation, such as tissue class. Further incorporation of anatomical information, such as the tissue class segmentations from LesionTOADS, may help to avoid lesions false positives in areas where we have prior knowledge that lesion presence is low and where OASIS made systematic false positives, such as the medial frontal cortex and the brainstem. Also, this could be used to help with the false positives in the pineal glands and the choroid plexus of the ventricles in the post-contrast FLAIR as these are areas where lesions do not occur in MS.

The smoothed images used in OASIS are similar to the use of smoothed

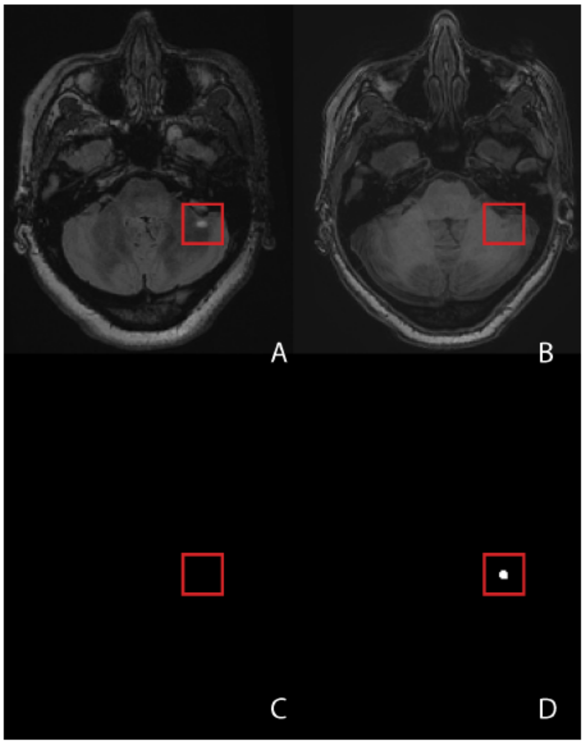


Figure 3.6: Example of a cerebellum lesion classified using OASIS in validation set 2: A. FLAIR volume; B. T1-weighted volume; C. LesionTOADS segmentation; D. OASIS empirically adjusted threshold segmentation.

images for inhomogeneity correction in MRI. For inhomogeneity correction, an image is smoothed to suppress the details of the image and then the original image is divided by this smoothed image in order to correct the image inhomogeneity [99]. Our method differs from this in that we do not divide the original image by the smoothed volume. Instead we use the smoothed volume as a covariate in our model. We also use multiresolution smoothed volumes, in contrast to just one smoothed volume for correction.

Other methods of capturing inhomogeneities may be used in the OASIS model as an alternative to the smoothed volumes. Alternative smoothers may be used instead of the Gaussian kernel and may be more appropriate in other applications. We decided to use the Gaussian filter because it is widely used, can be applied to any image, and is relatively computationally fast. The OASIS modeling framework is very flexible, however, and can be adapted for other methods of capturing the bias field and regional intensity variation.

We used the 15th percentile of FLAIR intensities in the brain to create the brain tissue mask. Other segmentations can be used to remove the CSF. We used the 15th percentile of FLAIR intensities because it is fast and performed well in this application.

Lesions that are hypointense on FLAIR, because of high free water content, are not detected by OASIS. The method models only candidate voxels, the top 15 percent of voxels in the cerebral matter-masked FLAIR volume, to minimize the number of false positives. In the FLAIR volume, such lesions are characterized by hypointensities in the center of a lesion and hyperintensities around the edges. Therefore the center of the lesions is excluded from the candidate voxels. Future work includes expanding the OASIS model to segment these lesions. This could be done by fitting another OASIS model trained only on lesion voxels that appear hypointense in FLAIR lesions. The binary segmentations from the original OASIS model and this model could then be combined to produce a complete lesion segmentation.

Like other voxel-based methods, OASIS is sensitive to major misregistrations within an MRI study. However, in part because it incorporates spatial smoothing, OASIS is not sensitive to minor errors in registration. By simultaneously comparing data from multiple sequences and only considering candidate voxels, OASIS is able to distinguish between artifacts and lesion.

OASIS uses a voxel-level model for assessing the outcome. The assumption of independence between voxels is imperfect, as lesions consist of clusters of voxels. In this work we use smoothing in the smoothed volumes and smoothing of the predicted probabilities of the model to incorporate the spatial nature of the data. Nevertheless, further incorporation of neighboring voxel information is warranted.

Chapter 4

A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI

4.1 Introduction

Machine learning is a popular perspective for mining and analyzing large collections of medical data [100, 101, 102]. We focus on the extent to which the choice of machine learning or classification algorithm and the feature extraction function impact performance in one problem from medical research – supervised multiple sclerosis (MS) lesion segmentation in structural magnetic resonance imaging (MRI). The evaluation of the classification algorithms employed in supervised lesion segmentation methods is not only a function of classification accuracy. Depending on the application, computational efficiency and interpretability may be valued at the cost of classification accuracy. Therefore, our evaluation also includes the computational time and resources required by each algorithm and the interpretability of the results produced by the algorithm. Comparison of machine learning techniques has been performed in other applications[103, 104, 105],but not to our knowledge in multiple sclerosis lesion segmentation. Also many of the currently available comparisons do not consider computational time.

MS is a life-long chronic disease of the central nervous system that is diagnosed primarily in young adults who will have a near normal life expectancy. Because of this, the burden of the disease is great, with large economic, social and medical costs. Between 250,000 and 400,000 people in the United States have been diagnosed with MS, and the estimated annual cost of the disease is over six billion dollars. There is currently no cure for MS, but many therapies exist for treating symptoms and delaying accumulation of permanent disability (http://www.ninds.nih.gov/disorders/multiple_ sclerosis/detail_multiple_sclerosis.htm). MS is characterized by demylinating lesions that are predominately located in the white matter of the brain, and MRI of the brain is sensitive to these lesions [1]. Quantitative MRI metrics, such as the number and volume of lesions, are important clinical tools for research into the pathophysiology and natural history of MS [65]. In practice, lesion burden is determined by manual or semi-automated examination and delineation of MRI, which is time consuming, costly, and prone to large inter- and intra- observer variability [62]. Therefore development of automated MS lesion segmentation methods is an active research field [106, 65]. The problem of automated MS lesion segmentation must be addressed by a method that is both sensitive and specific to white matter lesions, and which generalizes across subjects and imaging centers.

Many machine learning algorithms have been developed for automated segmentation of MS lesions in structural MRI. Over 80 papers have been published on the topic in the last 15 years, and yet no solution to this problem has emerged as superior to other methods [106]. Each lesion segmentation method in the literature is the composition of a classification algorithm and feature extraction function applied to one or many MRI modalities. As different methods use different data sets and performance metrics, the extent to which the classification algorithm, the feature extraction function, and the interplay between the classification algorithm and feature extraction function impacts the performance of these methods is unknown. To investigate this, we examine which factors improve classification performance through the composition of nine supervised classification algorithms with six feature extraction functions. We use voxel intensities form the T1-weighted (T1-w), T2-weighted (T2-w) and fluid-attenuated inversion recovery (FLAIR) MRI modalities to train and validate performance of the combinations of classifiers and feature extraction functions. We are not proposing a new lesion segmentation method. Rather than searching for the optimal method, we explore the problem and present our insights into the tools and methods of approach.

Our findings are that, for the employed feature extraction methods, the particular classification algorithm is much less important than the careful development of the features. These findings are not unique to this problem. Hand (2006) asserts that, in practice, in classification, simple classifiers typically yield performance almost as good or better than more sophisticated classifiers [107] . This is attributed to sources of uncertainty in data that are generally not considered in the classical supervised classification paradigm. Hand refers to using complex classification algorithms as “the illusion of progress”. Our findings support this characterization.

4.2 Experimental methods

4.2.1 Study population and experimental methods

We consider MRI studies with T1-w, T2-w and FLAIR volumes from 98 patients with MS. The 3D T1-MPRAGE images (repetition time (TR) = 10 ms; echo time (TE) = 6ms; flip angle (FA) $\alpha = 8°$; inversion time (TI) = 835 ms, resolution = 1.1 mm × 1.1 mm × 1.1 mm), 2D T2-w FLAIR images (TR = 11000 ms; TE = 68 ms; TI = 2800 ms; in-plane resolution = 0.83 mm × 0.83 mm; slice thickness = 2.2 mm) and T2-w volumes(TR = 4200 ms; TE = 80 ms; resolution = 0.83 mm × 0.83 mm × 2.2 mm) were acquired on a 3 tesla MRI scanner (Philips Medical Systems, Best, The Netherlands) equipped with an 8-channel phased array head coil.

4.2.2 Image preprocessing

We preprocessed the MRI images using the tools provided in Medical Image Processing Analysis and Visualization (MIPAV) [28], TOADS-CRUISE (http://www.nitrc.org/proje cts/toads-cruise/), and Java Image Science Toolkit (JIST) [27] software packages. We first rigidly aligned the T1-w image of each subject into the Montreal Neurological Institute (MNI) standard space (voxel resolution 1 mm^3). We then registered the FLAIR and T2-w images of each subject to the aligned T1-w images. We also applied the N3 inhomogeneity correction algorithm [19] to all images and removed extracerebral voxels using SPECTRE, a skull-stripping procedure [88]. A brain tissue mask is created from the extracerebral voxel removal mask by removing voxels falling below the 15th percentile of the FLAIR intensities over the mask [7]. This brain tissue mask removes cerebrospinal fluid in the ventricles and outside the brain.

4.3 Statistical methods

Our aim is to examine the extent to which the classification algorithm and the feature extraction function impacts the performance of lesion segmentation methods. To do so, we compare compositions of supervised classification algorithms and feature extraction functions for the classification of lesion voxels versus healthy tissue in structural MRI studies. A feature extraction function is a function that acts on the observed intensities from the MRI modalities and produces a feature vector. For each voxel of the brain, we have a “silver standard” manual lesion segmentation used to train and validate the lesion segmentation method. Manual lesion segmentations are considered a silver standard, as opposed to ground truth, as there is much inter- and intra- observer variability amongst expert segmentations. We examine lesion segmentation methods that are the composition of a classification algorithm and a feature extraction function. Here, our goal is not to find an the optimal lesion segmentation method; we instead search for insight by evaluating the performance over a set of possible classifiers and extraction functions.

4.3.1 Supervised classification algorithms

Voxels within a brain are spatially correlated. To include spatial information, we also include functions of voxel neighborhoods as features in our voxel-level classifiers. The extent to which the residuals are still correlated after these features are included or to what extent this information can be used to further improve prediction remain open problems.

We provide a short description of the super learner, as this is not a standard classification technique employed in the neuroimaging literature. The super learner is a method for combining class estimations from different classification algorithms, by weighting the classifiers according to their prediction performance using a cross-validation loss function [108, 109, 110, 111], which is referred to in the literature as ensemble learning, model stacking, or super learning. The super learner assigns each classification algorithm a coefficient weight, $\alpha_i \in (0, 1)$, with $\sum_i \alpha_i = 1$. A more detailed description of the super learner and of the other supervised classification algorithms used in this analysis can be found in the Appendix.

Table 4.1 shows a summary of the classification algorithms applied to each of the feature vectors, including the R package used to apply the algorithm and, when applicable, the tuning parameter values that were searched over. We performed all modeling in the R environment (version 2.15.3, R Foundation for Statistical Computing, Vienna, Austria) with the packages AnalyzeFMRI [112], ROCR [113], MASS [39], class [39], nnet [39], mclust [114, 115], e1071 [116], randomForest [117], and SuperLearner [118].

Table 4.1: A summary of the supervised classification algorithms. Values for the tuning parameters for each algorithm were selected using 10-fold cross validation on the voxels in the training set and validation of the algorithms was performed on a separate set.

Algorithm	R package	Tuning Parameters
Logistic Regression		defaults
Linear Discriminant Analysis	MASS	defaults
Quadratic Discriminant Analysis	MASS	defaults
Gaussian Mixture Model	mclust	defaults
Support Vector Machine (with linear kernel)	e1071	cost: 1/8, 1/4, 1/2, 1, 2, 4, and 8
Random Forest	randomForest	number of trees = 500 mtry = 1: dimension of the feature vector
k -Nearest Neighbor	class	k = 1,10, 100
Neural Network	nnet	size = 1, 5, 10
Super Learner	SuperLearner	all algorithms

Table 4.2: A summary of the six feature vectors

Feature Vector	Voxel Intensities (T1-w, T2-w, FLAIR)	Voxel Selection Procedure	Dimension
Unnormalized	Observed		3
Normalized	Normalized		3
Voxel Selection	Normalized	X	3
Smoothed	Normalized	X	9
	Smoothed volumes, $n \in \{21, 41\}$		9
Moments	Normalized	X	21
	j^{th} Local moment volumes , $j \in \{1, 2, 3\}$, $n \in \{21, 41\}$,		
Smoothed and	Normalized	X	27
Moments	Smoothed volumes, $n \in \{21, 41\}$		
	j^{th} Local moment volumes, $j \in \{1, 2, 3\}$, $n \in \{21, 41\}$		

4.3.2 Feature extraction functions and vectors

In this section we introduce the six feature extraction functions and the vectors these functions produce. In Figure 4.1, we examine 3-dimensional scatter plots of the intensities of the features that form the feature vectors. Note that while many of the feature vectors are in a higher dimension, in this figure we are only able to show 3-dimensions. Scatter plots of the T1-w, T2-w and FLAIR intensities and functions of these intensities for 10,000 randomly sampled voxels of 5 randomly sampled subject's MRI studies (a total of 50,000 voxels) are shown; each point in the plot represents a single voxel from a MRI study. We will refer to this figure throughout this section. Table A.3 contains a summary of the feature vectors introduced in this section.

4.3.2.1 Unnormalized

The unnormalized feature vector contains the observed voxel intensities (after image pre-processing) for a voxel v from the T1-w, T2-w and FLAIR volumes. Figure 4.1D shows a plot of the observed voxel intensities for the five subjects for the three volumes. Lesions appear as hyperintensities on the FLAIR and T2-w volumes and as hypointensities on the T1-w volume.

4.3.2.2 Normalized

We normalize voxel intensities by transforming them into standard scores over the brain tissue mask [93]. The normalized feature vector contains the normalized voxel intensities from the three imaging modalities. Figure 4.1E shows a plot of the normalized voxel intensities from the three modalities. Figure 4.2A shows a slice of the FLAIR volume after the normalization procedure and Figure 4.2B shows the manual lesion segmentation for this slice.

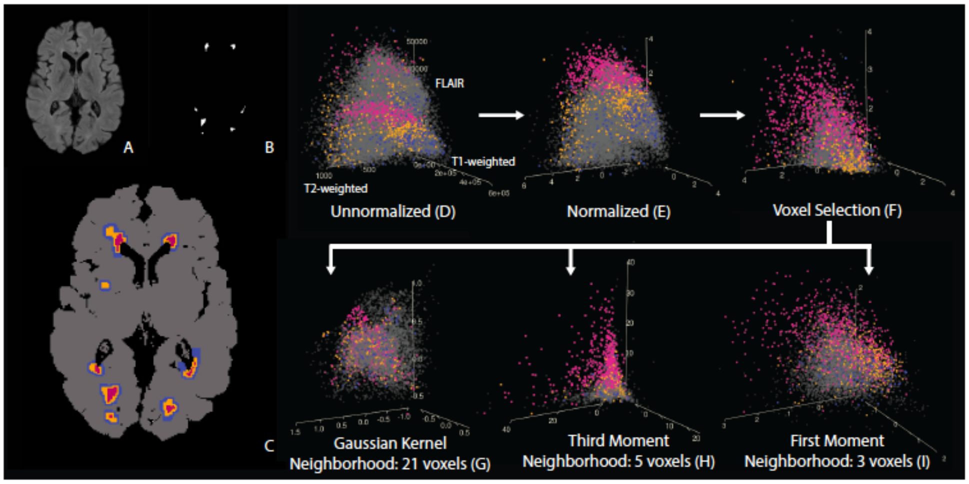


Figure 4.1: Scatter plots of the T1-w, T2-w and FLAIR voxel intensities and functions of these intensities for 10,000 randomly sampled voxels from 5 randomly sampled subject's MRI studies. Each point in the plot represents a single voxel from a study. (A-C) Color key for these plots: (A) the FLAIR volume for an axial slice from a single subject's MRI study, (B) the technician's manual segmentation for this slice and (C) the colors that are used in the plots corresponding to this slice. Lesion voxels are pink, voxels within 1 mm of a lesion voxel are orange, voxels within 2 mm of a lesion voxel are blue and all other voxels in the brain are colored grey. The arrows in the figure indicate the order that the features are created. For the unnormalized intensities there is no plane that can separate lesion voxels from non-lesion voxels, but after normalization and with the addition of features that include neighborhood information, a plane is able to separate lesion and non-lesion voxels with improved accuracy.

4.3.2.3 Voxel selection

We select candidate lesion voxels to lower computational time and restrict the modeling space. As most lesion voxels appear as hyperintensities in the FLAIR volume, we apply the brain tissue mask to the FLAIR volume and select the

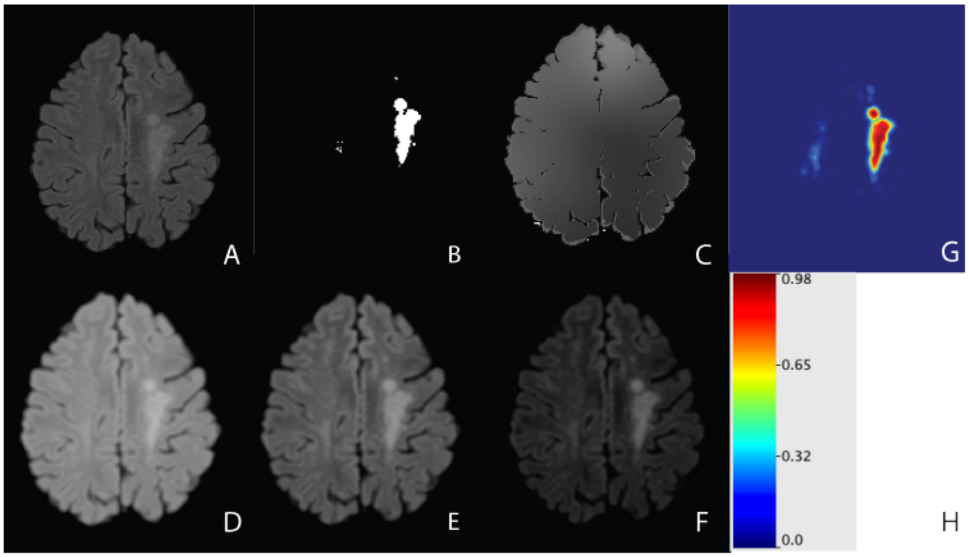


Figure 4.2: An axial slice from a single subject of the FLAIR volume with the normalization procedure, manual lesion segmentation, neighborhood functions of the FLAIR volume, and an example of a classification result. (A) FLAIR volume; (B) manual lesion segmentation; (C) FLAIR smoothed volume with a neighborhood n = 41; (D) FLAIR first local moment volume with neighborhood n = 5; (E) FLAIR second local moment volume with neighborhood n =5; (F) FLAIR third local moment volume with neighborhood n = 5. The smoothed volumes act on large neighborhoods, while the local moment volumes act over smaller neighborhoods; (G) The probability map produced for the logistic regression classifier on the smoothed feature vector; (H) The scale of intensities in the probability map.

85th percentile and above of voxels in the brain tissue mask as candidate voxels for lesion presence. Figure 4.1F shows the normalized voxel intensities for only the candidate voxels. The voxel selection procedure is not needed for simple methods but is needed for more complex algorithms.

4.3.2.4 Smoothed

The smoothed feature vector contains voxel intensities from the smoothed volumes, which are functions of each voxel and its neighbors. The smoothed volumes act on relatively large neighborhoods and are designed to capture anatomical information and residual intensity inhomogeneities (for a detailed discussion see [7]). Let the neighborhood of size n be the $n \times n \times n$ neighborhood of voxels centered at v. Two smoothed volumes for each imaging modality are created by 3-dimensional Gaussian smoothing of the normalized intensities for the modality over the neighborhood of size n within the brain tissue mask; in this application we chose $n = 21$ and 41. We used the FSL tool fslmaths (http://www.fmrib.ox.ac.uk/fsl) to create the smoothed volumes. Due to the high dimension of this feature vector, we cannot visualize intensities from the entire vector. Figure 4.1G shows the voxel intensities of the smoothed volumes for the three modalities for a neighborhood of 21 voxels for candidate voxels from the voxel selection procedure. Figure 4.2C shows a slice of the smoothed volume for the FLAIR, neighborhood of $n = 41$.

4.3.2.5 Moments

In contrast to the smoothed volumes, which act on relatively large neighborhoods, the local moment volumes are designed to incorporate spatial information from nearby neighboring voxels, as a single lesion is typically comprised of multiple clustered voxels. The local moment volumes are created by taking the j^{th} sample moment for each voxel v over the neighborhood of size n of the normalized volume over the brain tissue mask (the j^{th} sample moment for a set of r points is defined to be $\frac{1}{r}\sum_{i=1}^{r} X_i^j$). Here we use $j \in \{1,2,3\}$ for the sample moments and $n = 3$ and 5 for the neighborhood size. Due to the high dimension of this feature vector, we again cannot visualize intensities from the entire vector. Figure 4.1H shows the voxel intensities of the third local moment volumes for the three modalities with a neighborhood of $n = 5$ voxels for candidate voxels. Figure 4.1I shows the voxel intensities of the first local moment volumes for the three modalities with a neighborhood of $n = 3$ voxels for candidate voxels. Figure 4.2 shows an axial slice from a subject of the local moment volume for the FLAIR with a neighborhood of n = 5; Figure 4.2D shows the first local moment volume, Figure 4.2E the second, and Figure 4.2F the third. In the local moment volumes the contrast between lesion and other tissue is increased.

4.3.2.6 Smoothed and moments

The smoothed and moments feature vector contains the intensities from the smoothed feature vector and the moments feature vector. This vector contains the normalized voxel intensities and intensities from the smoothed volumes and local moment volumes for candidate voxels from the voxel selection procedure.

4.3.3 Training and validation studies with manual lesion segmentations

For each of the 98 MRI studies we have a manual lesion segmentation made by a technician with more than 10 years of experience in delineating white matter lesions. A neuroradiologist with more than 10 years of experience in MRI of MS patients reviewed the manual segmentations and found them to be acceptable. The manual segmentations were made only for white matter lesions and were made using the FLAIR and T1-w volumes. To assess the performance of the algorithms on each of the feature vectors, we randomly assigned 49 of the MRI studies to a training set and the remaining 49 studies were used for validation. We fit the algorithms on a set of 500 voxels sampled from each of the training MRI studies (for a total of 24,500 voxels). This was done in order to reduce the time needed to fit each algorithm, as many of the algorithms with tuning parameters required a substantial amount of time to fit on larger sets of voxels. We investigate the effect on performance and computational time of this and further downsampling in the Results section. Values for the tuning parameters for each algorithm were selected using 10-fold cross validation on the voxels in the training set. We validated on the entire brain volume for each of 49 studies in the validation set. Table 4.3 shows a summary of the training set, the training set after the voxel selection procedure and the validation sets.

Table 4.3: A summary of the training set, training set after the voxel selection procedure has been applied, and the validation set. Subjects were randomly assigned to the training or validation set. All training, including tuning of algorithm parameters with 10-fold cross validation, was performed on the training set.

	Training Set	Training Set with Voxel Selection	Validation Set
Number of Subjects	49	49	49
Number of Lesion Voxels (%)	288 (1.2%)	273 (7.5%)	458,407 (0.8%)
Total Number of Voxels	24500	3664	54,751,501

4.3.4 Measures of outcome and agreement

In the Results section, we report the partial Receiver Operating Characteristic (pROC) curve and scaled partial Area Under the Curve (pAUC) as measures of algorithm accuracy. The pROC and pAUC are calculated for false positive rates of 10% and below on the validation set, using the manual lesion segmentations as a "silver standard" of truth. The scaled pAUC is computed by dividing the pAUC by the false positive rate of 10% [119] and takes values between 0 and 1, with a value of 1 corresponding to perfect classification of lesion and non-lesion voxels. We also report the Dice similarity coefficient (DSC) as a measure of agreement amongst the binary segmentations produced by each algorithm. To calculate the DSC we threshold the probability map for each algorithm at a false positive rate of 0.5% in the validation set. The DSC is a measure of the overlap of two sets [97], with a DSC of 0 corresponds to no overlap and a DSC of 1 corresponds to perfect overlap. The efficiency of the algorithms, in the form of computational time for fitting the algorithms on the training set and making predictions on the validation are reported. We also examine the effect of downsampling the training set on fit time, performance and the super learner coefficient weights.

4.4 Results

4.4.1 Classification performance

Figure 4.3 shows the pROC curves for false positive rates up to 10% for the nine classification algorithms organized by the feature vectors, with a plot for each

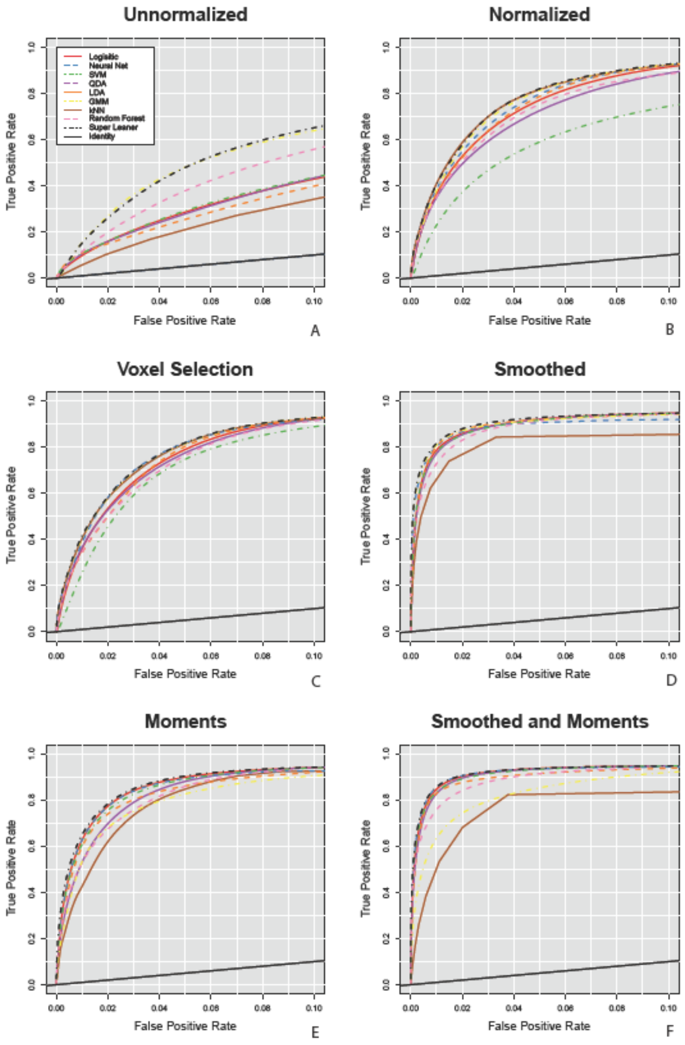


Figure 4.3: The partial Receiver Operating Characteristic (pROC) curves for the classification algorithms for false positive rates up to 10% in the validation set. The diagonal line is shown on each plot in black for reference, and represents a classifier that performs as well as chance. A plot is presented for each of the six feature vectors: (A) unnormalized , (B) normalized, (C) voxel selection, (D) smoothed, (E) moments, and (F) smoothed and moments. The performance of the simpler classification algorithms on the feature vectors with features including spatial information are superior to that of the more complex classifiers on the original features on the unnormalized feature vector.

of the six vectors: unnormalized (Figure 4.3A), normalized (Figure 4.3B), voxel selection (Figure 4.3C), smoothed (Figure 4.3D), moments, (Figure 4.3E) and smoothed and moments (Figure 4.3F). We also report scaled pAUC for false positive rates of 10% for each feature vector in Figure 4.4. The spread in the pROC curves in Figure 4.3 and the scaled pAUC in Figure 4.4 comes more from the different feature vectors than the classification algorithms. When fitting the algorithms on the unnormalized feature vector, algorithms perform better and worse relative to one another. In Figures 4.3 and Figure 4.4 we see that more complex algorithms, such as the super learner, Gaussian mixture model, neural network, and random forest, perform better than simpler algorithms. We define complexity as a function of an algorithm's decision boundary – for example algorithms with linear decision boundaries are relatively simpler than algorithm with nonlinear decision boundaries [120]. After normalization and the addition of features that include neighboring information, the performance of the algorithms begin to converge. All of the algorithms perform worse on the moments feature vector than the smoothed feature vector, even thought the dimension of the moments feature vector is much higher than that of the smoothed feature vector (21 versus 9). In this application and in terms of predictive performance, the smoothed volumes are better features for classifying lesions than the local moment volumes. Also, two of the algorithms, the Gaussian mixture model and the k-nearest neighbors, have a decreased performance relative to other algorithms on the validation set in the smoothed and moments feature vector.

The performance of the simpler classification algorithms on the developed feature vectors is superior to that of the more complex classifiers on the observed MRI intensities in the unnormalized feature vector. In Figure 4.3 we see

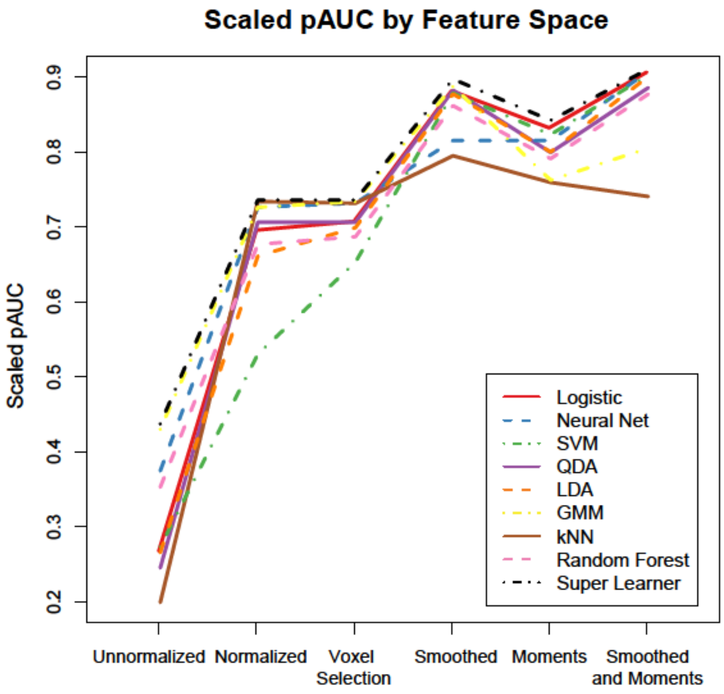


Figure 4.4: The scaled partial Area Under the Curve (pAUC) for each algorithm on each feature vector. The differences in scaled pAUC comes more from differences in feature vectors than differences in classification algorithms. The scaled pAUC of the simpler classification algorithms in the developed feature vectors are larger than that of the more complex classifiers on the original features in the unnormalized feature vector.

that the best performing algorithms on the unnormalized feature vector, the super learner and Gaussian mixture model, exhibit inferior performance to all algorithms on the developed feature vectors. An intuition for this result from the data is provided in the plots of Figure 4.1. Figure 4.1D shows a plot of the intensities from the unnormalized T1-w, T2-w and FLAIR volumes. There is no plane that can separate lesion voxels from non-lesion voxels. This may be the reason as to why algorithms with nonlinear decision boundaries perform better on the unnormalized feature vector. After normalization in Figure 4.1E, a plane is able to separate lesion and non-lesion voxels with improved accuracy. With the addition of the smoothed and moment volumes (Figure 4.1G-H), classification accuracy is further improved.

Note that many of the algorithms required a range of user-supplied tuning parameters (Table 4.3). We invested much time in deciding which tuning parameters to allow the algorithms to search over, as using the incorrect parameters greatly diminished the performance of these algorithms. As there are an infinite number of parameters than can be searched, we can never be certain that we have decided upon the optimal parameters, and therefore we prefer using algorithms that are completely informed by the data.

4.4.2 Algorithm agreement

To assess the agreement in the class label assigned to a voxel by each algorithm, we report the DSC for all pairs of binary segmentations from the classification algorithms and manual segmentations on each feature vector. We found these results to be robust to the choice of false positive rate threshold. Figure 4.5 shows the DSC for each of these pairs, with a plot for each of the six feature vectors: unnormalized (Figure 4.5A), normalized (Figure 4.5B), voxel selection (Figure 4.5C), smoothed (Figure 4.5D), moments, (Figure 4.5E) and smoothed and moments (Figure 4.5F). The DSC is shown as shades of gray in the plots, with darker shades indicating a DSC value closer to one. For the unnormalized feature vector, there is poor overlap between the class labels assigned by most algorithms and there is poor overlap with all algorithms and the manual segmentation. On this vector, the DSC for all pairs of the logistic regression, linear discriminant analysis, quadratic discriminant analysis and support vector machine algorithms are large. Also the super learner and Gaussian mixture model as well as the super learner and the random forest also have relatively large DSC, which is to be excepted as the super learner assigns relatively high coefficient weights to both of these algorithms on the unnormalized feature vector. On the normalized and voxel selection feature vectors, we see high DSC for all pairs of the algorithms, excluding the support vector machine, but see a low DSC with all algorithms and the manual lesion segmentation. On the smoothed, moments, and smoothed and moments feature vectors we see that the DSC for the manual lesion segmentations and all algorithms are large. Also the DSC for all pairs of the algorithms are also large (excluding the Gaussian mixture model and k-nearest neighbors algorithms on the smoothed and moments feature vector).

This plot reiterates that the performance of the simpler classification algorithms on the developed feature vectors are superior to that of the more complex classifiers on the observed MRI intensities in the unnormalized feature vector; the DSC for the class labels produced by all algorithms and the manual lesion segmentations are much larger on the moments, smoothed, and smoothed and moments feature vectors than those on the unnormalized feature vectors. The plot also shows that on the developed feature vectors, the class labels assigned to the voxels for each algorithm are similar. Not only are overall performance of the methods similar, but the segmentations produced from each method are also similar for these vectors.

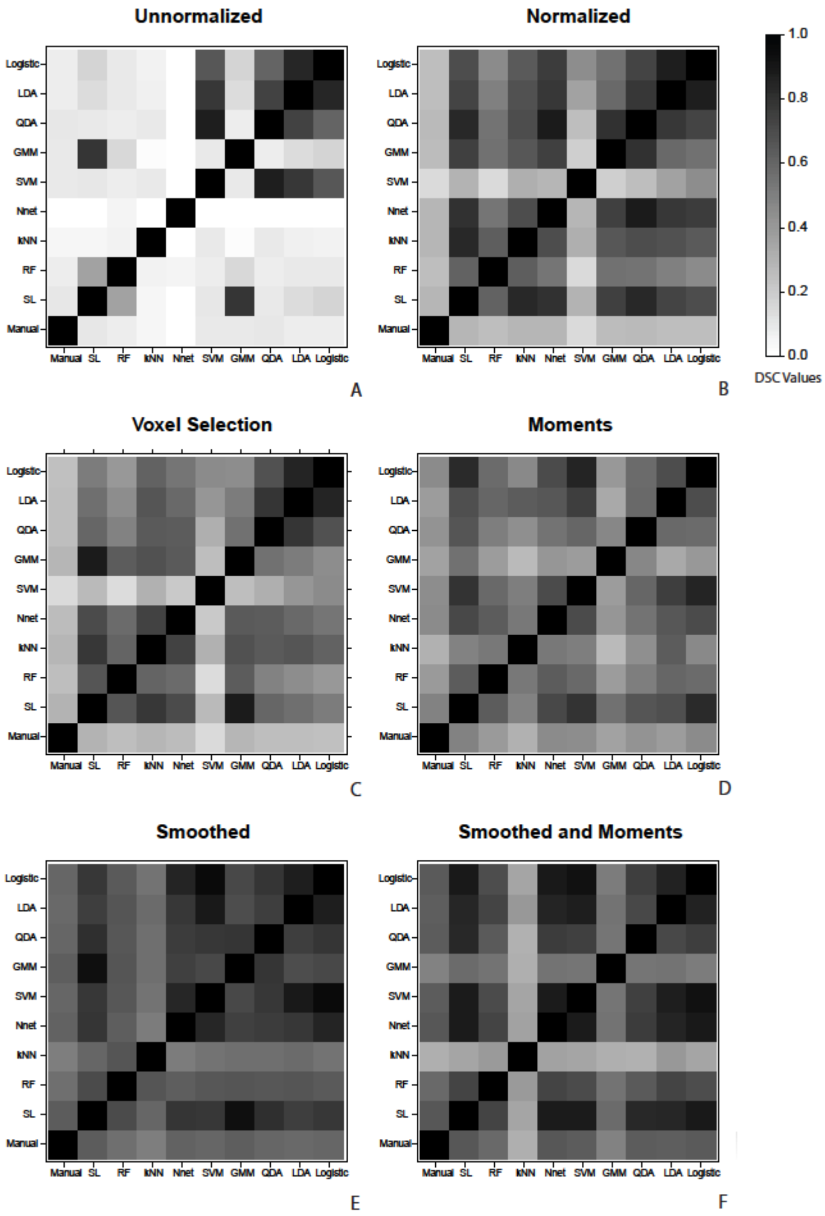


Figure 4.5: The Dice similarity coefficient (DSC) for all pairs of classification algorithm segmentations and manual segmentations. The binary segmentations for each classification algorithm are at a threshold of false positive rate = 0.5% in the validation set. A plot is presented for each of the six feature vectors: (A) unnormalized, (B) normalized, (C) voxel selection, (D) smoothed, (E) moments, and (F) smoothed and moments. On the developed feature vectors, the class labels assigned to the voxels for each algorithm are similar. This shows that not only are the overall predictive performances of the methods similar on these vectors, but the resulting segmentations from each method are also similar.

4.4.3 Computational time

The bar plots in Figure 4.6 show the time to fit each of the classification algorithms and to make predictions on a new MRI study; Figure 4.6A reports the time in hours required to fit the algorithm on each feature vector and Figure 4.6B reports the time in minutes required to make a prediction for a single MRI study from the fitted algorithms. All of the classification algorithms were fit and made predictions on a single core to allow for accurate comparison of computing times, although many of these algorithms can be run in a parallel computing environment to decrease computation time. In Figure 4.6 we see that simpler algorithms without tuning parameters, such as logistic regression, linear discriminant analysis, and quadratic discriminant analysis, require significantly less computational time. These methods take under a minute to make predictions and on the feature vectors that include the smoothed and local moment volumes and these methods have comparable performance to the more complex methods. This suggests that for this application, the computational burden of the super learner is not justifiable. Note that the relative fit time for the random forest is larger than the super learner in the smoothed, moments, and smoothed and moments feature spaces. This is attributed to the mtry tuning parameter, the number of variables sampled at each split of the decision tree, which searches over a number of parameters equal to the dimension of the feature vector. The random forest takes longer to tune than the super learner in this space and this difference can be attributed to the respective packages implementation of model tuning; the SuperLearner package calls the randomForest package to fit the random forest, but selects tuning parameters internally.

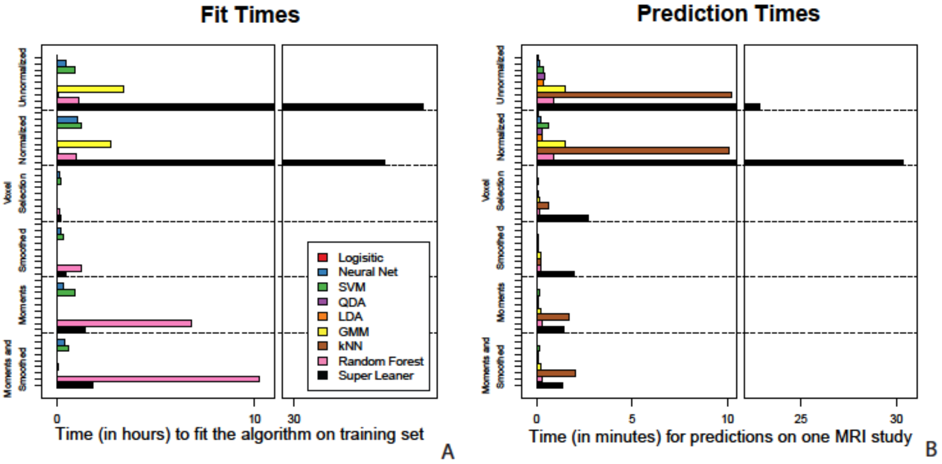


Figure 4.6: (A) The time in hours required to fit the algorithm on each feature vector and (B) the time in minutes required to make a prediction for a single MRI study from the fitted algorithms. Both of the bar plots are partitioned into the six feature vectors on the horizontal axis. The simpler algorithms without tuning parameters require significantly less computational time than more complex methods.

Computational time to fit the algorithm on the training data and make predictions for a new MRI study is an important consideration when choosing the appropriate classification algorithm in the application of lesion segmentation. This is especially important for making predictions for a new MRI study (Figure 4.6B). The algorithm may only need to be trained once and, as shown in this application, and can be trained on a relatively small number of voxels to reduce computational time. In research or clinical settings, fast algorithm predictions are desired, as lesion segmentations may be required for hundreds or thousands of studies. The two algorithms with the best predictive performance on the normalized feature vector, the super learner and the k –nearest neighbors, would not scale well; the algorithms take 10 and 30 minutes respectively to make predictions for a new MRI study. Even on the feature vectors that use the voxel selection procedure, k-nearest neighbors and the super learner often take between two to three minutes to make predictions for a single MRI study, while many of the other algorithms take only a few seconds to make these predictions.

4.4.4 Downsampling the training set: classification performance and computational time

In Figure 4.7 we investigate the impact of the number of voxels used to fit the algorithms on prediction accuracy and computation time. We examine the scaled pAUC for false positive rates up to 10% and the time to fit the algorithm on the unnormalized and smoothed and moments feature vectors. We sample 1,000 to 24,000 voxels (without replacement) by increments of 1,000 from the 24,500 voxels in the training set and fit the algorithms on each of these samples. Figure 4.7A shows the scaled pAUC versus the number of voxels the algorithm is fit on for the unnormalized feature vector and Figure 4.7B shows the time to fit the algorithm. Figure 4.7C and Figure 4.7D show the same for the smoothed and moments feature vector. At 10 hours we cut off the algorithm fit time on the plots; the super learner on the unnormalized feature vector took approximately 30 hours to fit on 24,000 voxels. The vertical axis for the smoothed and moments feature vector shows the number of voxels before the candidate voxel selection procedure is performed, as this procedure is part of the feature vector space, so the actual number of voxels the algorithm is fit on is around 15% of this size.

In Figure 4.7 we see the effectiveness of downsampling the training set to reduce computational time, without impacting performance. The performance of the algorithms is not impacted and the computational time is significantly lowered. On the unnormalized feature vector (Figure 4.7B) the performance of some algorithms varies as the number of voxels increases, especially the support vector machine and the neural network. On the smoothed and moments feature vector the performance is much more consistent. After the algorithm is fit on around 3000 voxels, the scaled pAUC for the validation set is stable as the number of voxel the algorithm is fit on increases. This shows that downsampling the training set can be an effective tool for reducing computation time without loss of performance on a well developed feature vector. In Figure 4.7A and C we see that in the unnormalized feature vector space the fit times for the k-nearest neighbors, quadratic discriminant analysis, linear discriminant analysis, and logistic regression appear to stay relatively constant as the number of voxels on which the algorithm is fit increases from 1000 to 24,000. The random forest, support vector machine, and neural network appear to increase linearly in computation time and the super learner and Gaussian mixture algorithm both appear to increase exponentially. The required time for the super learner, support vector machine, and neural network increase linearly and the others stay relatively constant.

4.4.5 Super learner coefficients

We examined the coefficient weights from the super learner algorithm in Figure 4.8. Figure 4.8A shows the weights, as shades of gray (darker indicating higher weight), on the unnormalized feature vector and Figure 4.8B shows the

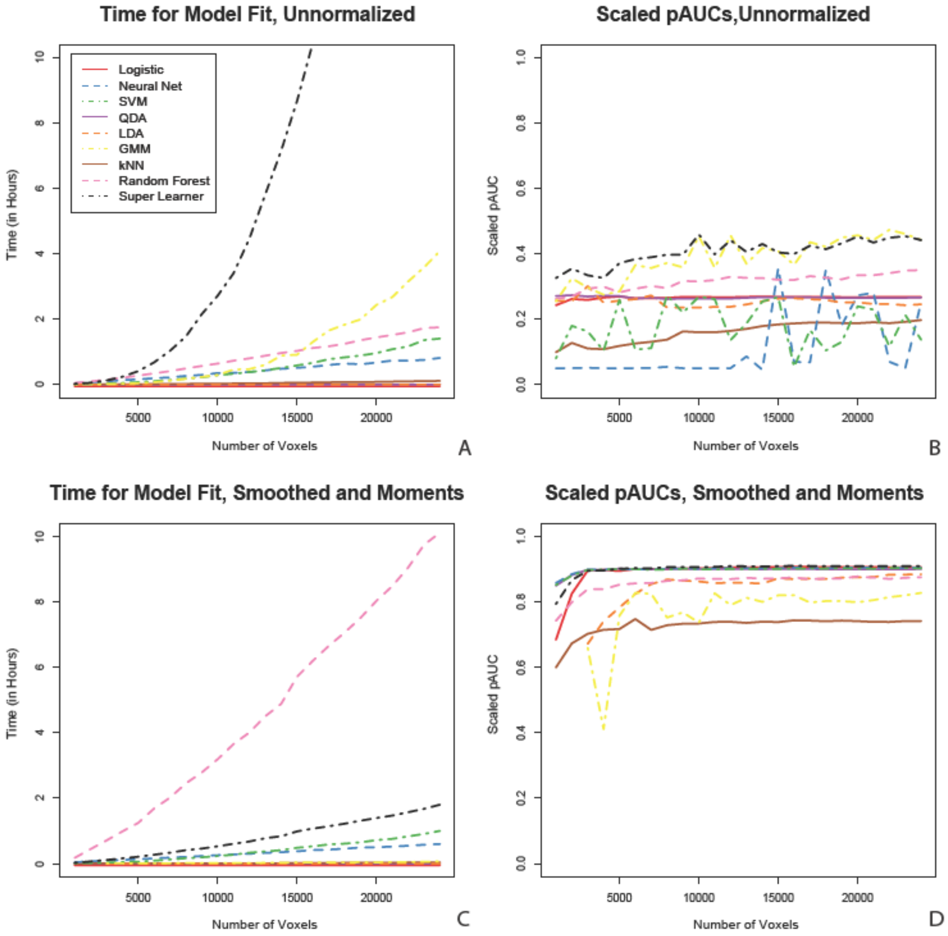


Figure 4.7: The impact of downsampling the training set on computational time and classification performance. Time in hours to fit the algorithm (left column) and scaled pAUC for false positive rates up to 10% (right column) versus the number of voxels the algorithm is fit on for the unnormalized (A,B) and smoothed and moments feature vectors (C,D). Here we see the effectiveness of downsampling the training set as the performance of the algorithms is not impacted and the computational time is significantly lowered.

weights for the smoothed and moments feature vector. The weights are examined versus the number of voxels fit by the algorithm, as in Figure 4.7. The horizontal axis of the plots show the algorithms and tuning parameters and the vertical axis shows the number of voxels on which the algorithm was fit. The super learner is designed to select and combine the classification algorithms that perform best, by cross-validation. In Figure 4.8, we see that, within the same feature vector space, as the number of voxels used to fit the algorithm changes, the super learner consistently assigns large weights on the same small number of algorithms. For the unnormalized feature vector, high weights are selected for the logistic regression, one of the random forest tuning parameters, and the Gaussian mixture model. For the smoothed and moments feature vector, the super learner favors the less complex algorithms, with corresponding explicit statistical models: logistic regression, the quadratic discriminant analysis, and the linear discriminant analysis. Some weight is also assigned to the Gaussian mixture model and the random forest, although the random forest tuning parameter is unstable as the number of voxels increases.

4.4.6 Interpretability

Interpretability of the algorithm is also desirable. Understanding which features improve the performance of the algorithms provides an insight into the problem of lesion segmentation and how the method may generalize; not all algorithms have the ability to provide this information. The logistic regression produces a set of coefficients with an explicit statistical interpretation, as these coefficients inform which features are most important in the context of lesion segmentation. A neural network with only one hidden layer has similar properties, but with the addition of more hidden layers, the meaning of the neural network coefficients becomes less clear. The support vector machine produces a hyperplane decision

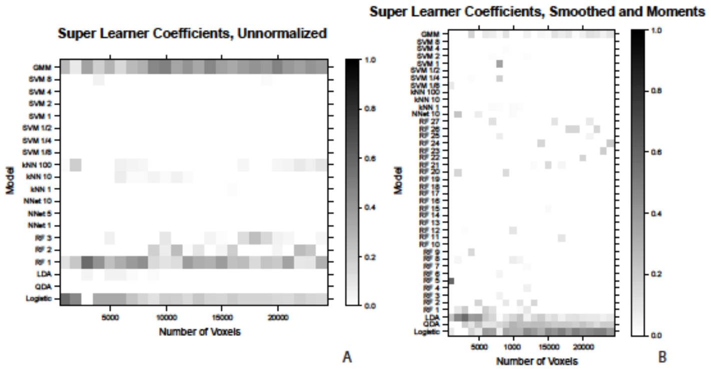


Figure 4.8: The super learner coefficient versus the number of voxels the algorithm is fit on for the (A) unnormalized and the (B) smoothed and moments feature vectors. As the number of voxels used to fit the algorithm changes, the super learner consistently assigns large weights to the same small number of algorithms. For the unnormalized feature vector, high coefficient weights are selected for the logistic regression, one of the random forest tuning parameters, and the Gaussian mixture model. On the smoothed and moments feature vector, the super learner favors the less complex algorithms: logistic regression, the quadratic discriminant analysis, and the linear discriminant analysis. Some weight is also assigned to the Gaussian mixture model and the random forest.

boundary for lesion and non-lesion voxels. Like a neural network with many hidden layers, the support vector machine does not provide much insight into the underlying classification problem, although recent advancements have been made to provide analytical tools for statistical inference in this framework [121]. The linear discriminant analysis produces a mean vector for each class and a shared covariance matrix for the two classes. Quadratic discriminant analysis produces a mean vector for each class and a covariance matrix for each class. The mean vector and covariance matrix or matrices from these algorithms can also be interpreted and provide information about the classification problem. The Gaussian mixture model has a similar interpretability as the linear discriminant analysis and the quadratic discriminant analysis, in that it fits a mixture of normal distributions to each class. But, in the smoothed and moments feature space, the Gaussian mixture model performs worse than many of the other algorithms. The k-nearest neighbors algorithm, random forest, and super learner all provide little intuition about the underlying classification problem, and only provide computationally complex rules for making predictions. The k-nearest neighbors algorithm also has diminished performance in the smoothed and moments feature space. While there is insight to be gained from the algorithms that the super learner assigns weights to (Figure 8), the super learner is not a practical algorithm in this context because of the computational time.

4.5 Discussion

In this paper, we investigated the extent to which the classification algorithm, the feature extraction function, and the interplay between classification algorithm and feature extraction function impacts the performance of a lesion segmentation method. We did not search for the optimal classification method, but instead evaluated performance over a set of possible of segmentation methods, each consisting of a feature extraction function and supervised classification algorithm, to gain insight into the problem. Our findings are, for this problem, that after careful development of the feature vectors with the addition of thoughtfully designed features, the difference in classification algorithm performance disappears. The spread in the pROC curves in Figure 4.3, the scaled pAUC in Figure 4.4, and the DSC in Figure 4.5 are attributed to the differences in feature vectors rather than differences in classification algorithms. The performance of the simpler classification algorithms on the developed feature vectors is superior to that of the more complex classifiers on the observed MRI intensities in the unnormalized feature vector. The resulting lesion segmentation on the feature vector with additional features that incorporate spatial information are also very similar. We therefore limit our assessment of the lesion segmentation methods to those feature vectors with the best predictive performance. And as predictive performance of the majority of the algorithms on these vectors is almost indistinguishable, choosing the appropriate algorithm is a function of (1) time to fit the algorithm, (2) time to make predictions for a new MRI study, (3) interpretability of the algorithm. Using these criteria, algorithms such as logistic regression, linear discriminant analysis, and quadratic discriminant analysis and working on development of the feature vector yields the best performance.

We have shown that the development of the feature vectors greatly increases the predictive performance in this application. Much of this improvement is explained by features using intensity information over the entire brain, not just the information at the voxel-level. The normalization procedure transforms intensities into standard scores of the brain tissue mask, using the sample mean and standard deviation across the mask. The smoothed volumes and local moment volumes also use additional information from neighboring voxels. In the feature vectors containing intensities from the smoothed volumes and local moment volumes, the dimension of the feature vector is increased. The addition of features can often improve the performance of a classifier, especially when very few original features are available[122]. All of the algorithms perform better on the smoothed, moments and smoothed and moments feature vectors, than they do on the feature vectors with lower dimension. It is also useful to note, that in this application, the smoothed volumes which use a much larger neighborhood are a better features for classifying lesions than the local moment volumes, which use information in a smaller neighborhood. While one might expect performance to be higher on the moment volumes than the smoothed volumes, as the dimension is higher in the space of the moment volumes, it has been shown that the addition of noisy features can diminish classification performance. [122] A number of image smoothers can be used to incorporate spatial information. We chose the smoothed volumes and local moment volumes to illustrate a smoother over a relatively large neighborhood and relatively small neighborhood, respectively. We refer the reader to the literature on image smoothing in both computer vision and neuroimaging for a more complete discussion of smoothers, a review of which can be found in "The Image Processing Handbook" [123].

We performed all modeling in the R environment, using the standard implementations of each of the classification algorithms. If a classification algorithm with a high computational time exhibited an improvement in classification accuracy over other algorithms, this algorithm could be tailored to reduce computational time. There are many techniques to reduce computational time, such as more computationally efficient training and testing implementations [124, 125, 126], the use of parallel computation or graphics processing units[127], and programming in other languages such as Python [128]. In the application of lesion segmentation, we did not observe an improvement in classification accuracy that would merit improving computation efficiency.

Another concern is the behavior of the classifiers on unbalanced data. Many of the classification algorithms employed in the analysis have been shown to perform poorly on unbalanced training data, such as k-nearest neighbors, neural networks, and support vector machines [129, 130]. In Table 3 we show the distribution of the lesion and non-lesion voxels in the training set. For feature extraction functions without the voxel selection procedure, 1.2% of the voxels contain lesions. For feature extraction function with voxel selection, 7.5% of the voxels contain lesion. While voxel selection was performed to lower computation time, it can also be thought of as a method of balancing the training data set. It is similar to the method of down-sizing [130], where random elements of the over-sized class are randomly eliminated from the training set. One difference is that in the voxel selection procedure, voxels were not removed at random, but were removed using a threshold on the FLAIR volumes. Other methods for balancing the training set could also be applied to lesion segmentation.

There is a large literature on supervised machine learning algorithms for segmentation of MS lesions in structural MRI. From this literature it is difficult to determine the extent to which the classification algorithm and the feature extraction function impacts the performance of the lesion segmentation algorithms. Each of the methods reports different performance metrics on different datasets. The majority of supervised machine learning algorithms in the literature are a composition of a single classification algorithm and feature extraction function. Most of the feature vectors from lesion segmentation methods in the literature contain intensity normalized voxel intensities, the most popular of which is histogram matching [131]. Other common features are functions of neighborhoods of an image voxel [132, 133, 134, 135, 80, 7, 136] or location information from an anatomical atlas [132, 137, 138, 139, 135]. The supervised classification algorithm used by these methods include neural networks [133, 72, 140, 141, 142], k–nearest neighbors [132, 143], bayesian classifiers [138, 79], principal component analysis classification [135], Parzen window classifiers [77], model stacking [80, 136], classification based on Markov Random Fields [134, 81], supervised learning of optimal spectral gradients [82] , and random forests [137]. Kamber (1995) does compare four different classification methods: a minimum distance classifier, a Bayesian classifier, an unpruned decision tree, and a pruned decision tree and found that the Bayesian classifier performed best. [139] The only features used in the method proposed by Kamber are the MRI voxel intensities and atlas derived prior probabilities of a voxel containing lesion; the method uses neither intensity normalization nor functions of the image intensities. Many of the aforementioned lesion segmentation methods use different image pre-processing steps. In this work we evaluate the impact of classification algorithms and feature extraction functions on classification performance. Our data is resliced to 1mm isotropic resolution, a common resolution for many image processing algorithms, even for data acquired at a lower nominal resolution. As pre-processing steps may also influence classification results, future work is needed to investigate this impact.

We investigate the choice of supervised classification algorithm and feature extraction function on the performance of lesion segmentation methods. Our findings are that the particular classification algorithm is less important than the careful development of the feature vectors. For the employed feature extraction methods, classification algorithms with a linear decision boundary (logistic regression and linear discriminant analysis) performed equally well as classifiers with nonlinear decision boundaries. Also, the performance of the simpler classification algorithms with the feature vectors containing additional features is superior to that of the more complex classifiers on the original features.

Chapter 5

Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions

5.1 Introduction

Structural magnetic resonance imaging (MRI) can be used to detect lesions in the brains of multiple sclerosis (MS) patients. The formation of these lesions is a complex process involving inflammation, tissue damage, and repair – to all of which MRI has been shown to be sensitive. The McDonald criteria for diagnosis of MS emphasize the key role of dissemination of lesions in the central nervous system on MRI not only in space, but also in time [4]. Characterizing the longitudinal behavior of lesions on structural MRI is therefore likely to be important for monitoring disease progression and response to therapy and for understanding the etiology of the disease. Surprisingly, there is poor association between clinical findings and the radiological extent of involvement on MRI using traditional volumetric measures, a phenomenon referred to as the clinico-radiological paradox [5]. Here we address this paradox by modeling the association between the longitudinal behavior of lesions after incidence on MRI and clinical covariates and disease-modifying treatment.

Previous work to characterize the longitudinal behavior of lesions on structural MRI and to further relate these changes to clinical information has only involved single structural MRI sequences. In the work of [144, 145] and [146], longitudinal lesion behavior is characterized only on the intensity normalized proton density (PD) volume, using bi-weekly MRI studies. Although they did not relate these changes to clinical covariates, it was found that the maximal insult within a lesion occurred at the center of the lesion, that lower initial intensity within a lesion was predictive of repair, and that most lesion activity did not last beyond 10 weeks. More recently, [147] examined the change over a 2-year period in normalized T1-weighted (T1) intensity within new lesions, and compared these changes in pediatric and adult-onset MS patients. A generalized linear mixed-effects model was used to relate clinical covariates, such as disease duration and treatments, to changes in intensity in the MRI. The only statistically significant relationship was that the T1 intensity in lesions increased between incidence and 1-year follow-up, and this recovery was more pronounced in children. Work has also been done to relate longitudinal changes in lesion intensity to sample size calculations for clinical trials. [148] used the change in the 25th percentile of intensity-normalized PD signal within a lesion over time to estimate necessary sample sizes for clinical trials of differing lengths. The 25th, 50th, and 75th percentiles of multiple MRI sequences were assessed, and it was found that the 25th percentile of the normalized PD yielded the smallest sample size requirements. A limitation of these studies is that each uses only one MRI sequence to characterize the behavior of the lesions, which ignores information known to be available in the other sequences [149].

Here, we describe two models to understand the relationship between clinical covariates and the longitudinal intensity profiles in lesion tissue from the T1, T2, T2-weighted fluid-attenuated inversion recovery (FLAIR), and PD sequences. The first is a principal component analysis (PCA) and regression model and the second consists of function-on-scalar regression models [150]. We use multi-sequence MRI studies acquired at the National Institute of Neurological Disease and Stroke (NINDS), with subjects being scanned on average once every 37 days (sd 52.3, range [13, 889]) yielding an average of 21 scans per subject (sd 8.0, range [10, 37]). In the PCA and regression model, we first reduce the data to a scalar, voxel-level biomarker for identifying slow and persistent, long-term intensity changes (which we will refer to from this point on as intensity changes for simplicity) within lesion tissue. The ability of the biomarker to identify these changes is then validated in an expert rater trial with two raters, a neuroradiologist and a neurologist. After this validation, we relate the biomarker to clinical information in a voxel-level mixed-effects regression framework. In the function-on-scalar regression, we directly relate the entire longitudinal trajectories from each sequence to the clinical covariates. This allows for assessment of how the clinical information relates to the intensity points at the post-lesion incidence time periods at which these associations occur, unlike in the PCA regression model.

5.2 Material and methods

In this section, we first describe the image acquisition and preprocessing, followed by the patient demographics. Next, we briefly describe the longitudinal lesion intensity profiles in the subsection *Lesion Profiles*, with a more complete description of the pipeline for extracting these profiles provided in the Appendix. We then introduce two models for studying the relationship between the lesion profiles and the clinical information in the subsections *Principal Component Analysis and Regression* and *Function-on-Scalar Regressions*. The subsection *Principal Component Analysis and Regression* also includes the expert rater trial of the voxel-level biomarker for identifying intensity changes within lesion tissue. All analysis, except for image preprocessing, was performed in the R environment [151] using the R package oro.nifti [152].

5.2.1 Image acquisition and preprocessing

Whole-brain 2D FLAIR, PD, T2, and 3D T1 volumes were acquired in a 1.5 tesla (T) MRI scanner (Signa Excite HDxt; GE Healthcare, Milwaukee, Wisconsin) using the body coil for transmission. The 2D FLAIR, PD, and T2 volumes were acquired using fast-spin-echo sequences, and the 3D T1 volume was acquired using a gradient-echo sequence. The PD and T2 volumes were acquired as short and long echoes from the same sequence. The scanning parameters were clinically optimized for each acquired image.

For image preprocessing, we use Medical Image Processing Analysis and

Visualization (http:// mipav.cit.nih.gov) and the Java Image Science Toolkit (http://www.nitrc.org/projects/jist) [27]. We interpolate all images for each subject at each visit to a voxel size of 1 mm^3 and rigidly co-register all volumes longitudinally and across sequences to the Montreal Neurological Institute standard space [153]. We remove extracerebral voxels using a skull-stripping procedure [154]. We automatically segment the entire brain using the T1 and FLAIR images [66] to produce a mask of normal-appearing white matter (NAWM), or white matter excluding lesions. After preprocessing, studies were manually quality controlled by a researcher with over four years experience with structural MRI (EMS). Studies with motion or other artifacts were removed.

5.2.2 Patient demographics

For this analysis, we use 60 subjects scanned at the NINDS, with the earliest scan performed in 2000 and the most recent scan performed in 2008. Three subjects were excluded during the expert validation because it was found that the longitudinal registration had failed, causing overall poor segmentation of lesion tissue. After exclusion of these subjects and subjects that did not have voxels with incident lesions that met a pre specified inclusion criteria (subjects scanned at least once within 40 days of lesion incidence and at least once 200 days after lesion incidence), there were 34 subjects left in the analysis. The 34 subjects included in the analysis had an average of 21 scans each (sd 8.0, range [10, 37]). Figure 5.1 shows the time points at which each of the 34 subjects was scanned. Each row of the plot corresponds to a subject, and each point in the plot represents an MRI study, with time from the subject's baseline visit in years along the horizontal axis. The total follow-up time per subject was on average 2.2 years (sd 1.2, range [0.9, 5.5]). The mean age of the subjects at baseline was 37 years (sd 10.1, range [18,60]). At baseline, there were 30 subjects with relapsing-remitting MS (RRMS) and 4 subjects with secondary-progressive MS (SPMS) . There were 20 females and 14 males,14 subjects on disease-modifying treatment, and 2 subjects who received steroids at the baseline visit. The disease-modifying treatments and use of steroids for many of these subjects changed at subsequent follow-up visits.

5.2.3 Lesion profiles

Figure 5.2 shows an example of the longitudinal, multi-sequence MRI studies used for this analysis. For our analysis, we use intensity profiles from voxels that are detected during a subject's follow-up period. The first row of Figure 5.2 shows the multiple MRI sequences at one time point (from left to right, the FLAIR, T2, PD, and T1 sequences). In each sequence, a red box shows an area with a lesion that develops during the follow-up period. The subsequent 4 rows of the figure show the longitudinal behavior within this red box. Each column of the figure shows a different MRI study, starting at 98 days after baseline in the far left column and going until 343 days after baseline. The lesion in the red box is first observed 175 days after baseline.

The pipeline for extracting the longitudinal voxel-level lesion profiles from the collection of multi-sequence structural MRI is divided into four steps: (1) identifying voxels with new lesion formation, (2) intensity normalization, (3) temporal alignment, and (4) temporal interpolation. We briefly describe these

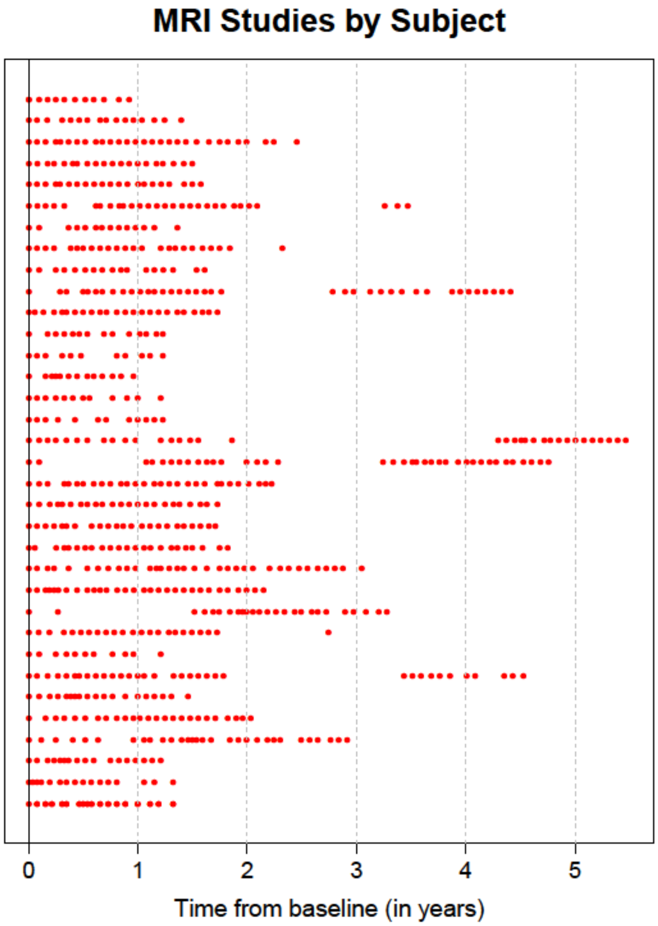


Figure 5.1: The time points at which each of the 34 subjects included in the analysis was scanned. Each row of the plot is a subject, and each point in the plot represents an MRI study. The horizontal axis represents the time from the subject's baseline visit in years.

steps here and include an extended description of all steps in this pipeline in the Appendix. For the first step of identifying the lesion tissue, we distinguish between areas that contain vasogenic edema (which we will refer to simply as "edema") and actual lesion tissue, which both manifest as areas of abnormal signal intensity, especially on the T2-weighted sequences. For this analysis, we are interested only in areas with tissue damage, as opposed to the neighboring edema. We combine two previously developed lesion segmentation methods, SuBLIME and OAISIS, to find new lesion voxels and distinguish between edema and lesion tissue [9, 7]. The row labeled “Segmentation” in Figure 5.2 shows the edema and lesion tissue segmentation for each study at the time point in which the lesion was detected. The subsequent analysis is performed only on the lesion tissue in new lesion voxels.

For intensity normalization, we put the units from each imaging sequence into standard deviations about the mean of intensities within the NAWM mask [66] for the sequence, using the methodology of [93] and [155]. After segmentation and normalization, the intensity normalized longitudinal profiles from the lesion in Figure 5.2 for all four sequences can be seen in the first column of Figure 5.3. From top to bottom in the first column of Figure 5.3 we have the profiles from 150 randomly sampled voxels from the lesion in Figure 5.2 for the FLAIR, T2, PD, and T1 sequences. Each line in the plot represents the longitudinal profile from a single voxel. The x-axis shows the time in days from the baseline visit, with the point of lesion incidence denoted by a vertical dashed line, and the y-axis shows the normalized sequence intensities.

In this work, we are interested in the lesion dynamics only after lesion incidence, so we perform linear interpolation within the window after lesion incidence and up to 200 days post-incidence. We select the end point of 200 days, as it has been previously found that new T2 lesions show the most dramatic changes in intensity for three to four months [146], and we opt to be conservative and include some data beyond this reported stabilization point. Voxels are selected for the analysis if the subject has at least one visit 200 days or more after lesion incidence, and at least one visit within 40 days of incidence. Of the 60 subjects in this analysis, 34 have voxel profiles meeting this inclusion criteria, after removing the three subjects for poor longitudinal registration. We linearly interpolate over a grid of 0 to 200 days in increments of 5 days so that all profiles are observed on the same time grid. We denote the vector of observations from a voxel over this time grid for sequence S in voxel v for subject i in lesion l at registered study time t' (since lesion incidence) as $S_{ilv}^{N}(t')$, for $S =$ FLAIR, T1, T2, and PD. Then we let S_{ilv}^{N} be the longitudinal collection of these interpolated values, namely the 1×41 vector $S_{ilv}^{N} = \{S_{ilv}^{N}(t') : t' \in (0, 5, ..., 200)\}$. The second column of Figure 5.3 shows the temporally registered and linearly interpolated profiles, S_{ilv}^{N}, over the period of 0 to 200 days for the lesion in Figure 5.2 for the same 150 randomly sampled voxels as shown in the first column.

5.2.4 Principal component analysis and regressions

In this section, we outline the PCA and regression modeling approach for studying the relationship between the longitudinal lesion profiles and demographics, disease, and treatment. We begin by describing the voxel-level biomarker for identifying intensity changes within lesion tissue. Next we describe the validation of this biomarker with an expert rater trial with two raters, a neuroradiologist and a neurologist. Last, we describe a mixed-model regression framework for relating the voxel-level biomarker to clinical covariates.

5.2.4.1 Biomarker

We begin by describing the voxel-level biomarker for identifying intensity changes within lesion tissue. The biomarker is the score on the first principal component (PC), after performing PCA on the longitudinal lesion profiles. To perform PCA on the longitudinal lesion profiles, we first concatenate the profiles for each voxel from the four sequences together. For each sequence and at each voxel, we have a 1×41 vector of longitudinal intensities, S_{ilv}^{N}. Let I_{ilv} denote the 1×164 dimensional vector of the four concatenated profiles, S_{ilv}^{N}, from subject i lesion l and voxel v. More precisely,

$$I_{ilv} = \{FLAIR_{ilv}^{N}, T1_{ilv}^{N}, T2_{ilv}^{N}, PD_{ilv}^{N}\} \tag{5.1}$$

and we index the entries $I_{ilv}(j)$, where $j = 1, \ldots, 164$ is the j^{th} entry of the concatenated vector. Note that we first remove the mean from the concatenated profiles and then perform a PCA on these concatenated profiles. Let ϕ_k denote the k^{th} PC, where ϕ_k is also indexed by j. The relationship between the score on the k^{th} PC, the one-dimensional value $\xi_{ilv}(k)$, and the observed trajectory for $I_{ilv}\,(j)$ is:

$$I_{ilv}\,(j) = \sum_{k=1}^{K} \xi_{ilv}(k)\phi_k\,(j)\,. \tag{5.2}$$

We focus on the first PC, ϕ_1, and the score on this component, $\xi_{ilv}(1)$. The first PC is found to identify intensity changes at the voxel-level within lesions. Positive values of $\xi_{ilv}(1)$ correspond to a return of the voxel to intensity values closer to that of normal-appearing tissue and negative values of $\xi_{ilv}(1)$ correspond to the voxels maintaining intensity values closer to those at lesion incidence. This biomarker, $\xi_{ilv}(1)$, collapses the full profiles at each voxel from the four sequences into a single scalar. We use the score on the first PC in this analysis, as the other PCs explain only 25% of the variation in the datA, were not found to be associated with any biological processes, and are thus likely due to scanner-related and other noise. To assess the variability in both the mean and the first PC, we bootstrap this procedure by resampling subjects with replacement 1000 times [156].

5.2.4.2 Expert validation of biomarker

We use expert validation to determine the quality of the lesion tissue segmentation (excluding edema) as well as the ability of the biomarker to identify areas of slow, long-term intensity change. For this validation we use two raters, a neuroradiologist with 11 years of research experience in MS (DSR) and a neurologist with 4 years of research experience in MS (MKS). For each lesion, we first determine the axial slice of the image that contains the largest number of voxels with abnormal signal intensity. Then for each lesion the two raters are presented the following: (1) the full axial slice for the FLAIR, T2, PD, and T1 volumes that contains the largest number of voxels with abnormal signal intensity; (2) the entire collection of longitudinal scans for a box containing the abnormal signal intensity in the FLAIR, T2, PD, and T1 volumes for this axial slice; (3) the segmentation of the lesion and edema tissue within this box; (4) the biomarker for the voxels segmented as lesion tissue within this box and a scale for the intensities within this image; (5) the entire collection of longitudinal scans for the FLAIR, T2, PD, and T1 weighted volumes within this box with the score for the first PC overlaid on the images for each scan after lesion incidence. The raters are then asked to rate the quality of the lesion tissue segmentation and the biomarker for identifying areas of intensity changes on an integer scale from 1 to 4, with each rating corresponding to the following: (1) failed miserably; (2) some redeeming features; (3) passed with minor errors; and (4) passed. Examples of the images presented to the raters for each lesion that received a rating of 1 through 4 for the score on the first PC by both raters are provided in the Appendix. Forty-seven lesions are selected at random to be repeated in the analysis to assess intra-rater reliability.

We report the median of the ratings of the lesion segmentation and the biomarker for each rater over all lesions. To assess between-rater and within-rater reliability, we report the Cohen's κ coefficients over all of the lesions and for the set of repeated lesions respectively, for both the rating of the biomarker and the lesion segmentation. We also report κ for the rating of the lesion segmentation and the biomarker for all lesions, for each rater, to determine if the quality of the segmentation and the quality of the biomarker are related. We nonparametrically bootstrap by subject with replacement 1000 times to produce the confidence intervals for the median of the ratings for each rater and the κ coefficients.

5.2.4.3 Regression model

The clinical information for each subject that we consider at each study visit consists of MS disease subtype, age, sex, an indicator of treatment with steroids, an indicator of disease-modifying treatment, and distance to the boundary of an area of abnormal signal intensity. An example of distance to the boundary of an area of abnormal signal intensity can be seen in the seventh row of the Figure 5.2. We center age at the mean age of 36 years over all of the voxel-level observations. During the observation period, many of the subjects were enrolled in clinical trials at NINDS to test various experimental therapies. Our indicator of disease-modifying treatment indicates treatment with any of the Food and Drug Administration-approved treatments, including interferon beta 1-a (intramuscular or subcutaneous), interferon beta 1-b, and glatiramer acetate, as well as experimental therapy. As many of the covariates change over time, we model the relationship between the lesion profiles and the value of the covariate at the time of lesion incidence for the particular profiles. For the following analysis, we have a total of 57,908 voxels from 315 lesions in 34 subjects.

We now introduce a linear mixed-effects model to relate the biomarker, that is the score on the first PC, to the clinical covariates [157]. We use the value of the covariate at the time of lesion incidence for the particular profiles, which can vary within subject. Thus, for added precision, the covariates that change over time are indexed by the subject index i, lesion index l and voxel index v, as voxels from the same lesion may have different times of incidence. For example, the sex of the subject does not change by time of lesion incidence, so it is only indexed by i. In contrast, age of the subject changes with voxel lesion incidence and is indexed by i, l and v. We also add random effects for subject and lesion, which we denote by b_i and b_l, respectively, with both following a normal distribution: $b_i \sim N(0, \sigma_i^2)$ and $b_l \sim N(0, \sigma_l^2)$, where σ^2 denotes the variance of the random effects. We consider the following basic model for the association between the biomarker, $\xi_{ilv}(1)$, and the covariates:

$$\xi_{ilv}(1) = \beta_0 + \beta_1 \text{SPMS}_{ilv} + \beta_2 \text{ Distance}_{ilv} + \beta_3 \text{Age}_{ilv} + \beta_4 \text{ (Age} - 4)_{+ilv}$$

$$+ \beta_5 \text{Steroids}_{ilv} + \beta_6 \text{Male}_i + \beta_7 \text{Treatment}_{ilv} + b_i + b_l + \epsilon_{ilv}$$

We assume that the error terms are independent and identically distributed, with each following a normal distribution, $\epsilon_{ilv} \sim N\left(0, \sigma_\epsilon^2\right)$. In the model, the term SPMS is an indicator of being diagnosed with SPMS where the comparison group is being diagnosed with RRMS. Note that the age term has been centered at the mean age of 36 years. The term $(\text{Age} - 4)_{+ilv} = \text{Age}_{ilv} \cdot 1(\text{Age}_{ilv} > 4)$ is a spline term for centered age over 4 years (or age over 40 years), which was included in the model after visualizing the relationship between the biomarker and age. We also investigated simpler models with the same mixed-effects structure, but where we considered each covariate separately.

To test for associations, we use two procedures. First, we perform a parametric bootstrapping procedure [156], and second we calculate p-values using a normal approximation for the distribution of the fixed-effects in the mixed-effects model [158]. We use 1000 bootstrap samples for the bootstrap procedure. We perform the parametric bootstrap because steroid use and disease subtype of SPMS did not always appear in nonparametric bootstrap samples. A complete description of this procedure is found in the Appendix. We also use the normal approximation, as this approximation has been found to be a reasonable approximation for the distribution of the fixed-effects in most settings [158].

5.2.5 Function-on-scalar regressions

The previous model is an attempt to collapse the information from the four profiles (across sequences and time) into a single scalar at each voxel. As an alternative, we also fit a two-step function-on-scalar regression model [150], where we can investigate the relationship between the covariates of interest and the profile at each time point. We fit a function-on-scalar regression model for each sequence separately. For simplicity of notation, we now use t for the registered time, as opposed to t'. The outcome in the model is the full lesion intensity profile:

$$\begin{aligned} S^N_{ilv}(t) = \beta'_0(t) + \beta'_1(t)\text{SPMS}_{ilv} + \beta'_2(t)\ \text{Distance}_{ilv} + \beta'_3(t)\text{Age}_{ilv} \\ + \beta'_4(t)\ (\text{Age} - 4\ )_{+ilv} + \beta'_5(t)\text{Steroids}_{ilv} + \beta'_6(t)\text{Male}_i \\ + \beta'_7(t)\text{Treatment}_{ilv} + \epsilon_{ilv}\,(t) \end{aligned}$$

for $S =$ FLAIR, T1, T2, and PD. To fit the model, we use a two-step function-on-scalar regression implemented in the R package refund [159]. The procedure first fits a scalar-on-scalar regression at each individual time point. Then the resulting coefficient functions are smoothed over time using a cubic spline basis with an automatically selected penalty on the second derivative.

To assess the variability in the coefficient functions and provide bootstrapped, point-wise 95% confidence intervals, we non-parametrically bootstrap by subject using 1000 resampled datasets. When samples do not contain subjects with a covariate, for example the indicator of steroids, we remove this sample from the bootstrap and replace it with another sample. The difference between the function-on-scalar regression and the PCA regression model is that PCA collapses the entire temporal intensity profile of the voxel into a scalar. By contrast, the function-on-scalar regression investigates the association at every time point. While function-on-scalar regression is more comprehensive and interpretable, it is more appropriate when there are strong functional effects that are not captured by a small number of principal components, due to the potential for decreased statistical power

5.3 Results

5.3.1 Principal component analysis and regression

5.3.1.1 Biomarker

In Figure 5.4 A we show the mean profiles for each sequence over the registered 200 day period, and in Figure 5.4 B we show the first PC, ϕ_1, for each sequence over the registered 200 day period, where the first PC is divided into different sequences for purposes of presentation. The subfigures for both the mean and the first PC show the bootstrapped 95% confidence intervals. The first PC explains 75% (95% CI: [72%, 76%]) of the variation in the concatenated longitudinal profiles.

To interpret the PCs, we recall that the normalization procedure puts the volumes into units of standard deviations above the mean of the NAWM. Therefore a value of 0 on the image corresponds to the average value of NAWM from the particular MRI scan. The mean profiles for the FLAIR, T2, and PD are all above 0 throughout the time course, as lesions are hyperintense on these sequences. In contrast, the mean profile for the T1 sequence is below 0, as lesions are hypointense on this sequence. The first PC for the FLAIR, T2, and PD is negative throughout the time course, with values closer to 0 at lesion incidence (time 0). Positive scores on this PC indicate a decrease in the signal in these sequences, which corresponds to a return of the voxel to intensity values closer to that of normal-appearing tissue. In contrast, negative scores indicate the voxel maintaining intensity values closer to those at lesion incidence, with more hypointensity than the average profile. Similarly, for T1 the first PC is positive throughout the time course, with values closer to 0 at lesion incidence. Positive scores on this PC indicate increased signal on the T1. As lesions are hypointense on the T1, this also indicates a return of the voxel to intensity values closer to that of normal-appearing tissue. Negative scores again correspond to the voxels maintaining intensity values closer to those at lesion incidence.

We therefore consider the score on the first PC to be a biomarker of intensity changes within the lesion at the voxel level. In the last row of Figure 5.2 we see the PC scores or the biomarker from the lesion that is shown in the figure. We see that the positive scores indicate areas of the lesion that return to values of normal-appearing tissue, while the negative scores show areas that remain at the intensity values at lesion incidence.

5.3.1.2 Expert validation of biomarker

We use expert validation to determine the quality of the lesion segmentation (excluding edema tissue) and the ability of the biomarker to identify areas of slow, long-term intensity change. The distributions of the ratings for the two raters for both the lesion segmentation and the biomarker are shown in Figure 5.5. The first row of plots in Figure 5.5 shows the distribution of the ratings for the lesion segmentation and the second row shows the ratings for the biomarker. Plots in the left column are ratings by the neuroradiologist, and plots on the right column are ratings by the neurologist. The median rating for both the lesion segmentation and the biomarker by the neuroradioloist are 4 (95% CI: [4,4]), which is a rating of passed, the highest possible rating. The median rating for both the lesion segmentation and the biomarker by the neurologist are 3 (95% CI: [3,3]), which is a rating of passed with minor errors. Note that criteria for assigning scores were not discussed between the two raters prior to their respective analyses.

The κ coefficients for the within- and between-rater agreement for both the lesion segmentation and the scores on the biomarker are shown in Table 5.1. The values for the κ coefficient range between 0 and 1, with a value of 1 indicating total agreement and 0 indicating no agreement. The within-rater agreement for the lesion segmentation and the score on the biomarker are higher for the neuroradiologist than the neurologist. There is only modest agreement between the neuroradiologist and neurologist on both ratings, with a κ coefficient of 0.29 (95% CI: [0.18, 0.41]) for the lesion segmentation and 0.24 (95% CI: 0.11, 0.39) for the score on the biomarker. This is due, in part, to the fact that the neurologist spread ratings of the studies between 3 and 4, while the neuroradiologist gave more ratings of 4.

The κ coefficient for the agreement between the rating of the lesion segmentation and the biomarker is 0.97 (95% CI: 0.93, 1.00) for the neuroradiologist and 0.68 (95% CI: 0.58, 0.78) for the neurologist. The high correlation between these ratings, especially for the neuroradiologist, indicates that the quality of the segmentation impacts the quality of the rating of the biomarker. Comments from the raters mirrored this finding, as many of the low scores for both the lesion segmentation and the biomarker were due to: (1) missing the first time point of lesion incidence and segmenting it as new lesion at a later time point; (2) not segmenting the entire lesion; and (3) parts of the same lesion being segmented (unnecessarily) at different time points. As both the ratings for the lesion segmentation and the score on the biomarker were high, the quality of the lesion segmentation does not appear to be negatively impacting the method.

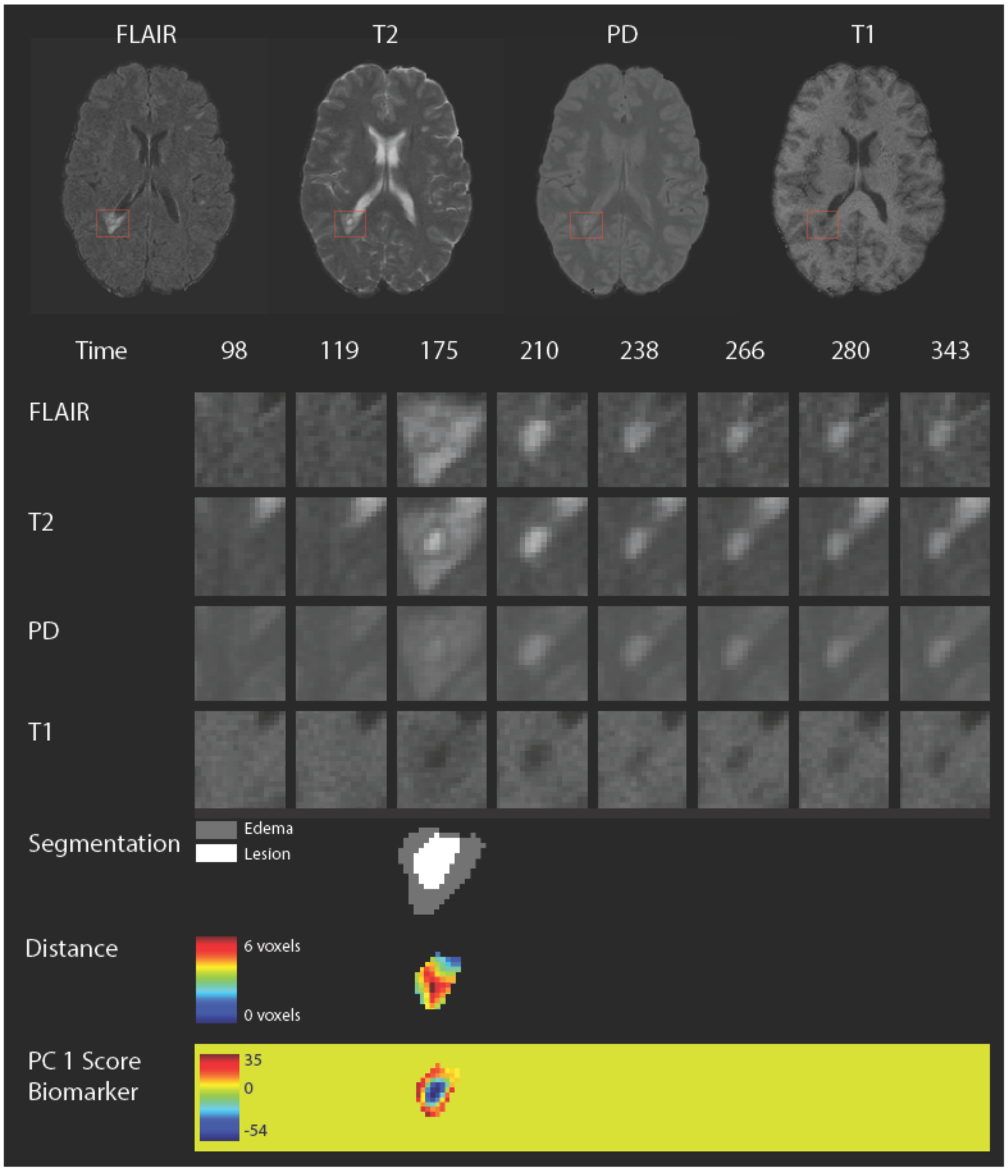


Figure 5.2: The first row of the figure shows an axial slice from the multiple MRI sequences, 175 days after baseline (from left to right, the FLAIR, T2, PD, and T1 sequences). In each sequence, a red box shows an area with a lesion that develops during the follow-up period. In the subsequent rows of the figure, we show the longitudinal behavior within this red box. Each column of the figure represents a different MRI study, starting at 98 days after baseline in the far left column and going until 343 days after baseline. A lesion is first identified in this area at 175 days. The first four rows show the longitudinal behavior of the FLAIR, T2, PD, and T1 sequences. The next row shows the segmentation of the edema and lesion tissue. The following row shows the distance to the boundary of abnormal MRI signal. The last row shows the score on the first PC, which identifies areas of lesion repair and permanent damage.

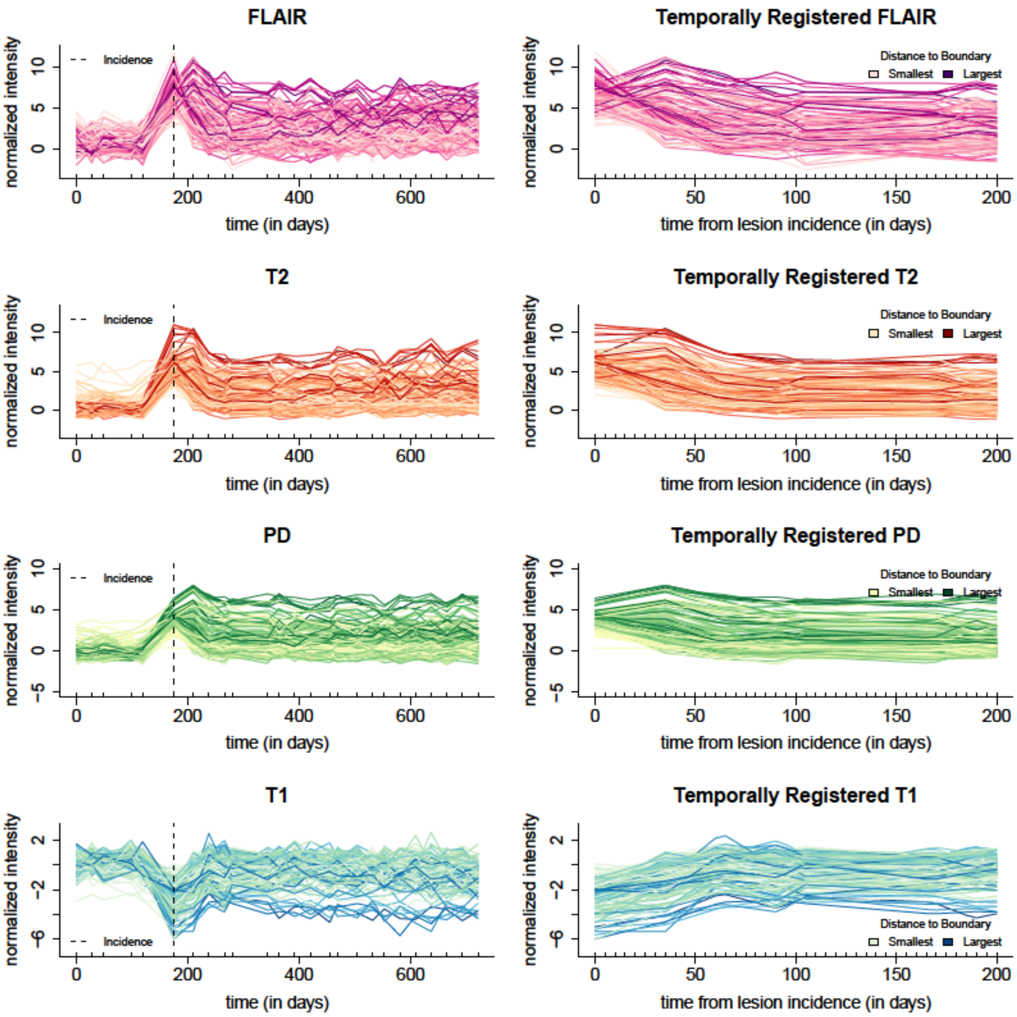


Figure 5.3: The first column of the figure shows the full longitudinal profiles from all four sequences (from top to bottom, the FLAIR, T2, PD, and T1 sequences). The profiles are from 150 randomly sampled voxels from the lesion in Figure 5.2, and for display purposes the periods between each study have been linearly interpolated. Each line in the plot represents the longitudinal profile from a single voxel. The x-axis shows the time in days from the subject's baseline visit, the time of lesion incidence is denoted by a dashed line, and the y-axis shows the normalized sequence intensities. The second column shows the same voxels after temporal alignment and linear interpolation over the 200 day period after incidence, the time period used in this analysis. The profiles are colored by distance to the boundary of abnormal MRI signal.

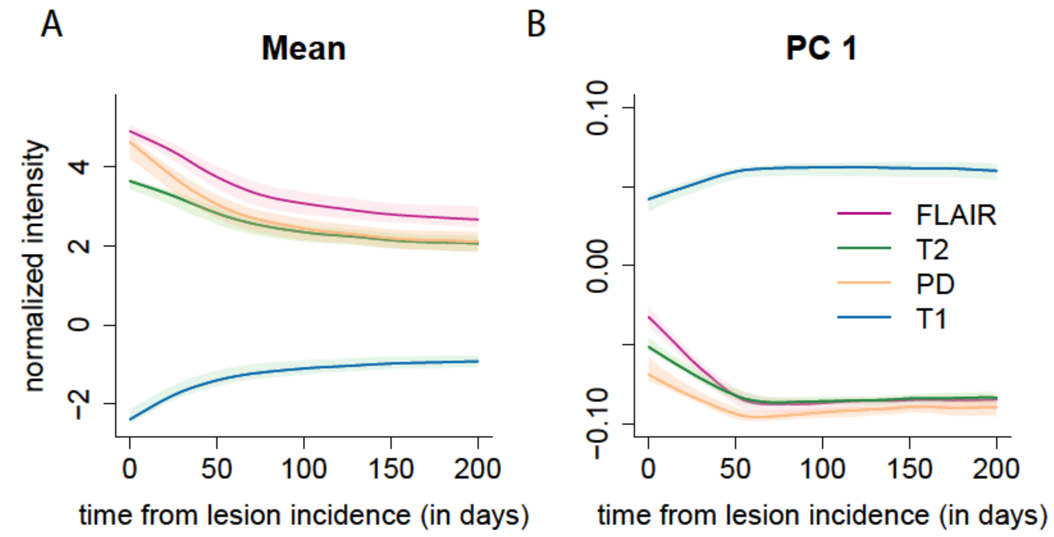


Figure 5.4: Panel A of the figure shows the mean profiles for each of the imaging sequences over the registered 200 day period, and panel B shows the first PC for each of the imaging sequences. The first PC explains 75% of the variation in the concatenated longitudinal profiles. Along the x-axis for both plots is plotted the time in days since lesion detection. The 95% confidence intervals in both panels are obtained using 1000 nonparametric bootstraped samples.

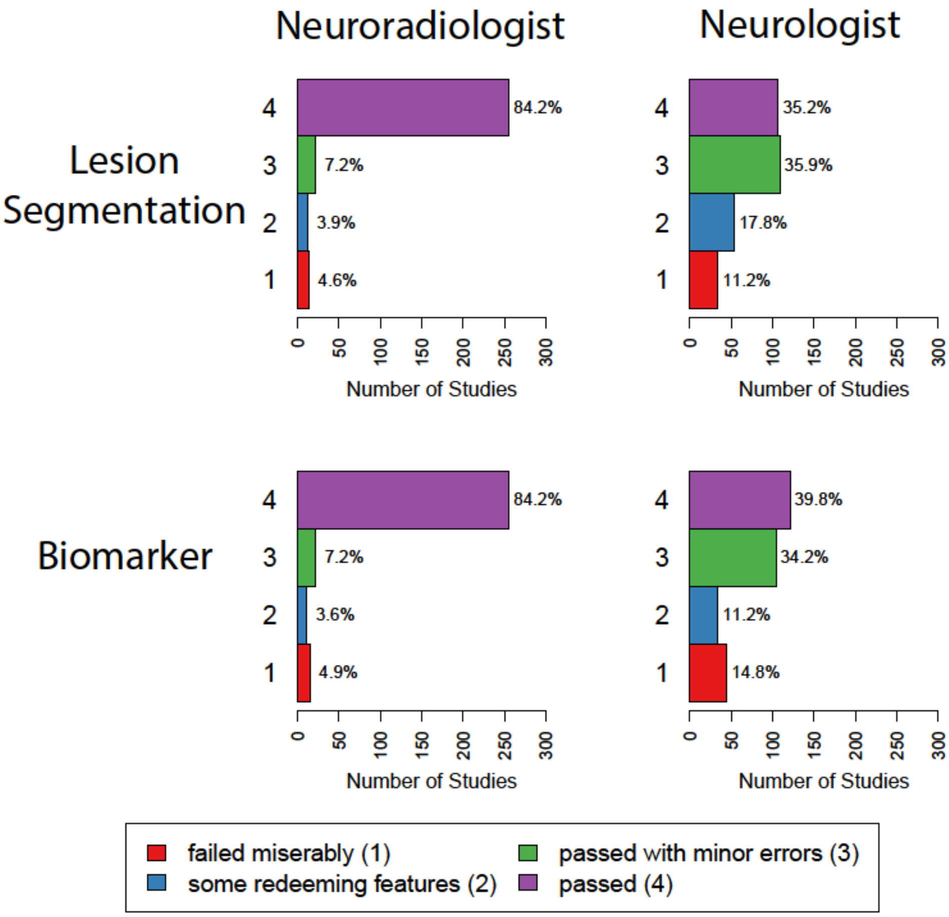


Figure 5.5: The first row of plots shows the distributions of the ratings for the lesion segmentation, and the second row shows the ratings for the biomarker. Plots in the left column are ratings by the neuroradiologist, and plots on the right column are ratings by the neurologist. Each plot shows the number of studies that failed miserably (1), had some redeeming features (2), passed with minor errors (3), and passed (4) along with the percentage of each rating.

Lesion Segmentation			Biomarker	
	Neuroradiologist	Neurologist	Neuroradiologist	Neurologist
Neuroradiologist	0.92; (0.76,0.99)	0.29; (0.18, 0.41)	0.92; (0.76, 0.99)	0.24; (0.11, 0.39)
Neurologist		0.75; (0.62, 0.86)		0.72; (0.51, 0.86)

Table 5.1: The table on the left shows the κ coefficients for the lesion segmentation, and the table on the right shows the same for the biomarker. The between-rater agreement is reported using all lesions. The within-rater agreement is reported using only the forty-seven repeated lesions.

5.3.1.3 Regression model

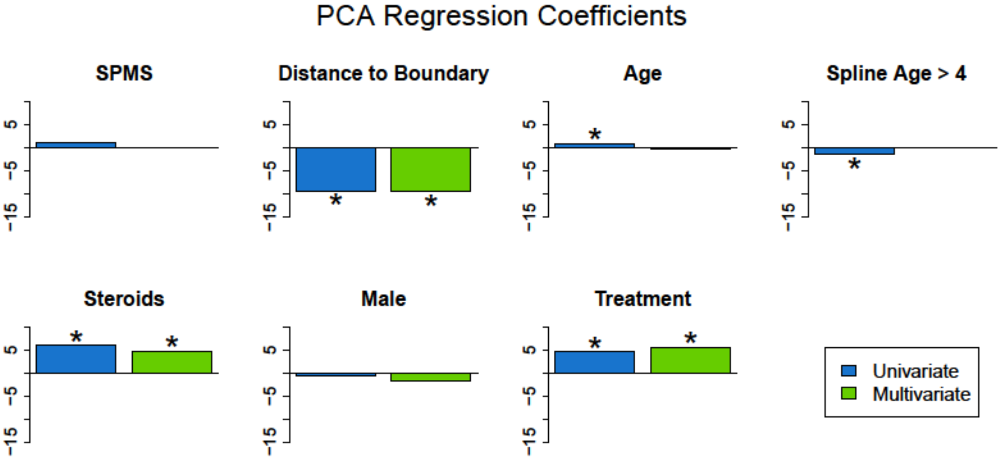


Figure 5.6: This figure shows bar plots of the coefficient estimates from the univariate and multivariate mixed-effects models with the biomarker as an outcome. The results from the univariate model are shown in blue, and the results from the multivariate model are shown in green. Asterisks indicate significance at the 5% level. In both the univariate and multivariate models, disease-modifying therapy, steroids, and age were found to be significant.

We fit both univariate and multivariate mixed-effects models to investigate the relationship between the covariates and the biomarker. The estimates of the coefficients from both models are shown in the bar plots in Figure 5.6, with asterisks indicating statistical significance at the 5% level using the bootstrapped 95% confidence intervals. Tables containing the coefficient estimates, standard errors, t-statistics, p-values using the normal approximation, and 95% bootstrapped confidence intervals can be found in the Appendix for both the univariate and the multivariate models. There are no differences in the conclusions determined by the normal approximation and the bootstrapped 95% confidence intervals. For continuous covariates, such as age, the coefficient is interpreted as the expected change in the biomarker for a one unit increase in the covariate. For binary variables, such as disease subtype, the coefficient is interpreted as the difference in the expected change in the biomarker in the specified group. Therefore, positive coefficients are indicative of the voxel returning to intensity values closer to normal-appearing tissue with an increase in the covariate, while negative coefficients are indicative of the voxel maintaining the intensities at lesion incidence with an increase in the covariate (or in some rare cases having intensities that have an increasing departure from those of normal-appearing tissue over time with an increase in the covariate). The results indicate that voxels that are farther away from the boundary have increased risk for maintaining abnormal signal intensity. In this model, the coefficient for distance to the boundary has a value of -9.4 (95% CI: [-9.6, -9.3]), indicating that for a one voxel (or 1 mm) increase in distance away from the boundary (toward the center of the lesion) the average value of the biomarker decreases by 9.4, adjusting for the other coefficients and the random effects. In the last row of Figure 5.2, we see this spatial relationship between the biomarker and the distance to the lesion boundary, with positive scores near the boundary and negative scores near the center of the lesion. In both models, we found the use of disease-modifying treatment and steroids to be associated with return of a voxel to the value of normal-appearing tissue. The coefficient for treatment has a value of 5.4 (95% CI: [4.7, 6.1]), indicating that when subjects are on treatment the average value of the biomarker increases by 5.4, adjusting for the other coefficients and the random effects. The use of steroids has a similar interpretation, with a coefficient value of 4.3 (95% CI: [2.7, 5.9]).

5.3.2 Function-on-scalar regression

The resulting coefficient functions from the function-on-scalar regression with bootstrapped, point-wise 95% confidence intervals with the FLAIR profile as the outcome are shown in Figure 5.7. Similar figures for models with the T2, PD, and T1 profiles are provided in the Appendix. The coefficient functions for continuous variables in the function-on-scalar regression model are interpreted as the change in the expected profile at each time point for a one unit increase in the covariate. Similarly, for binary variables, the coefficient function is interpreted as the change in the expected profile for the specified group. For the FLAIR profiles, the coefficient functions corresponding to distance to the boundary and age have bootstrapped 95% confidence intervals that do not overlap with 0 across any of the time points, and are therefore statistically significant at the .05 level. The coefficient function for distance to the boundary is greater than 0 throughout the entire trajectory, indicating that the farther away from the boundary the voxel is, the more the FLAIR hyperintensity is maintained within the voxel. For a one voxel (or 1 mm) increase in distance away from the boundary (toward the center of the lesion) the average normalized intensity of the trajectory increases by around 0.5 at all time points, adjusting for the other coefficients and the random effects. The result for distance from the boundary agrees with the results from the PCA regression model.

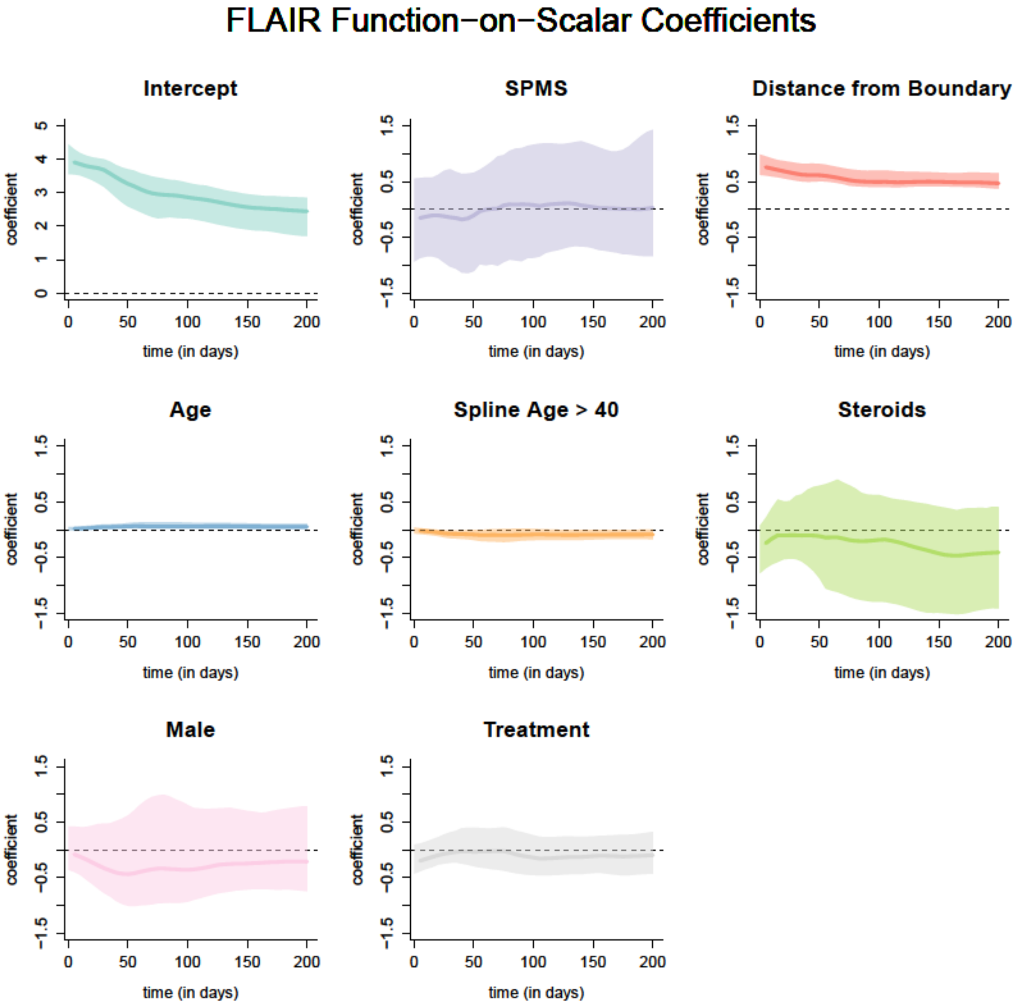


Figure 5.7: Each dark line represents the coefficient function, and the shaded area represents a bootstrapped, point-wise 95% confidence interval. Along the x-axis of each plot is the time in days from lesion incidence. Along the y-axis is the value of the coefficient function at each time point. Only distance from the boundary and age were found to be different for 0 at any point along the profile.

5.4 Discussion

We introduce two models to relate clinical information to the longitudinal intensity profiles in lesion tissue from conventional MRI sequences. The first model is the PCA regression model, where we collapse the longitudinal, multi-sequence MRI information into a biomarker of slow, long-term intensity changes within the lesion at the voxel-level and then relate this to clinical information. We validate the ability of the biomarker to detect these intensity changes using an expert rater trial. The second model is the function-on-scalar regression model, which relates each longitudinal intensity profiles separately to the clinical information and allows for assessment of the time points in which the clinical information is impacting the profiles. The methodology presented here shows promise for both understanding the time course of tissue damage in MS and for evaluating the impact of neuroprotective or reparative treatments for the disease. The biomarker may be particularly useful in clinical trial settings, as it is sensitive to the effects of disease-modifying therapies and shows impressive performance in expert visual validation. Reliable methods to evaluate such treatments, which are currently under development, are lacking at present. In contrast to prior studies of change in lesion intensity in clinical trials, our work is focused on voxel-level analysis, and therefore it can provide spatial information about intensity recovery and does not artificially reduce the size of the data set. This may have implications on the sample size calculations for clinical trials. These methods are also broadly applicable to other imaging modalities and disease areas, in which longitudinal intensity profiles may lead to more sensitive and specific biomarkers.

In the PCA and regression model, we observe a statistically significant relationship between the biomarker and the use of disease-modifying therapy and steroids. Both treatment and steroids were associated with a return of a voxel to intensity values closer to that of normal-appearing tissue. The inference from both models in regards to disease-modifying treatment should only be taken as a proof-of-concept for the relationship between the imaging and the clinical covariates. The models may suffer from confounding by indication, which arises when individuals who are on a treatment are different from those who do not receive treatment, due to unobserved considerations. In the multivariate model, we adjust for age, sex, and disease subtype, but unobservable differences related to treatment choice may cause biased conclusions. However, bias in terms of treatment effect would most plausibly result in underestimation of improvements, as more aggressive therapies are commonly given to subjects with more aggressive or refractory disease. Thus, our findings might underestimate what would be observed in a randomized trial of disease-modifying therapy.

One limitation of this study was the relatively small number of subjects. Future work will involve deploying the methodology and models on a larger number of subjects (n = 34), in both observational studies and randomized clinical trials. While many of the coefficient functions from the function-on-scalar regression are not found to be statistically different from 0, this model may have more power with more subjects. For the bootstrap procedure we only have 34 subjects, resulting in wide confidence intervals for the estimated coefficient functions. In contrast, the regression using score outcomes identifies strong associations between specific covariates and multisequence longitudinal patterns of longitudinal intensities.

The two models presented in this work are fit voxel-wise and therefore may be sensitive to major misregistration within a study and between longitudinal studies for the same subject. The models are also sensitive to local displacement of tissue due to transient swelling in and around lesions or resorption of lesion tissue. We therefore do not call the slow changes in intensity within the voxels that are observed "tissue repair," as we cannot be certain that the change is not due to misregistration or displacement of tissue from the lesions themselves. We do observe a relationship between the return of voxels to the intensity of normal-appearing tissue and both disease-modifying treatment and treatment with steroids, and therefore find this measure useful and deserving of further study. We also see an association with the distance to the boundary of the lesion and slow, long-term intensity changes – with voxels near the boundary of the lesion returning to baseline intensity and voxels near the center of the lesion maintaining abnormal signal intensity. Future work to assess tissue repair may involve investigating a nonlinear registration within individual lesions.

The methods described here use only conventional clinical imaging for patients with MS, namely FLAIR, T2, PD, and T1. While this is beneficial for using the methodology in a clinical trial setting or for analysis of retrospective imaging studies, one could also incorporate advanced imaging into the method. For example, magnetization transfer ratio imaging [160], quantitative T1-weighted imaging [161], and diffusion tensor imaging [162] have been studied in MS lesions. The longitudinal dynamics of lesions on these images could be incorporated into our framework to better understand the behavior of lesions over time and the impact of disease-modifying therapies on this behavior.

For this analysis, all MRI studies are acquired on a single 1.5 T MRI scanner at one imaging center. Similar analysis could be performed at higher field strength, but for this analysis we use a 1.5 T dataset for the availability of the large retrospective cohort study over a long period of time. Although different scanning parameters were used for the acquisitions, further investigation is warranted into the robustness of the methods to changes in scanner, changes in magnetic field strength, as well changes in the imaging center.

Appendices

Appendix A

Appendix to “OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI

A.1 OASIS coefficients

Table 1 and Table 2 display the coefficients from the two fits of the OASIS model. The coefficients for each modality are the log odds ratios of lesion incidence for voxels corresponding to an increased intensity of one normalized unit, fixing all other predictors. Lesions are characterized by hyperintensities in the FLAIR, PD, and T2-weighted images and hypointensities in the T1-weighted image, although not all lesions will appear on the T1-weighted image. As expected, the signs of coefficients for the FLAIR and T2-weighted intensities are positive. The coefficient of negative sign for the PD and positive sign for the T1-weighted are interpreted in the context where all other predictors are fixed. Due to the strong correlation among the imaging modalities, it is difficult to visualize the impact that a one standard deviation increase in the normalized intensity in an image has on the log odds ratio when all other modalities are kept constant. The coefficients for the smoothed volumes are more difficult to interpret. They are again the log odds ratios of lesion incidence for voxels corresponding to an increased intensity of one normalized unit, fixing all other predictors. But, it is very difficult to conceptualize what a one unit increase in the smoothed FLAIR with kernel window size of 10, fixing the FLAIR and the smoothed FLAIR with kernel window size of 20. The intensities in these three images are highly correlated and this may explain, for example, why the signs for the smoothed FLAIR image are opposite, when intuitively you would expect both to be positive. Also, the signs on the coefficients from each modality from the first and second fit are all the same, except for the coefficient for the smoothed T2-weighted with kernel window size of 10, the smoothed T2-weighted with kernel window size of 20 and the interaction between the T2-weighted image and the smoothed T2-weighted with kernel window size of 20. Again, this may be explained by the correlation among the images. We also assessed the variation in the coefficients by nonparametrically bootstrapping the subjects with replacement. The bootstrapped 95% confidence intervals for the coefficients are also provided in Table 1. The variance of these coefficients is large in comparison to the estimates of the coefficients. This may be explained by the strong correlation among the imaging modalities and the smoothed images. But, the instability in the coefficients does not impact the performance of OASIS, as illustrated in the stability of the partial ROC curve.

Table A.1: Regression coefficients from first and second fit of the logistic regression model on all of the MRI studies in validation set 1

Fit 1	Coefficient	Standard Error	Bootstrapped Mean (95% CI)
Intercept	-6.44	0.01	-6.68 (-7.86, -5.84)
FLAIR	1.98	0.00	2.06 (1.61, 2.45)
$\mathcal{G}FLAIR(10)$	7.73	0.03	6.62 (1.53, 11.83)
$\mathcal{G}FLAIR(20)$	-6.87	0.06	-4.87 (-14.47 , 5.70)
PD	-0.46	0.00	-0.50 (-0.87, -0.09)
$\mathcal{G}PD(10)$	-2.18	0.02	-1.91 (-5.10 , 1.27)
$\mathcal{G}PD(20)$	2.00	0.03	1.69 (-3.88, 7.11)
T2	0.64	0.00	0.65 (0.38, 0.98)
$\mathcal{G}T2(10)$	0.53	0.02	0.41 (-2.80, 3.50)
$\mathcal{G}T2(20)$	-0.03	0.04	-0.45 (-5.16, 4.05)
T1	1.52	0.00	1.56 (1.27, 1.89)
$\mathcal{G}T1(10)$	8.68	0.02	8.84 (6.19, 11.66)
$\mathcal{G}T1(20)$	-15.44	0.03	-15.52 (-20.20, -10.72)
$FLAIR * \mathcal{G}FLAIR(10)$	0.10	0.01	0.50 (-1.12, 2.50)
$FLAIR * \mathcal{G}FLAIR(20)$	-5.02	0.03	-5.75 (-9.34, -2.53)
$PD * \mathcal{G}PD(10)$	-1.76	0.01	-2.01 (-3.29, -0.67)
$PD * \mathcal{G}PD(20)$	1.80	0.02	2.16 (0.30, 4.30)
$T2 * \mathcal{G}T2(10)$	-0.10	0.01	-0.06 (-1.41, 1.13)
$T2 * \mathcal{G}T2(20)$	-1.47	0.02	-1.45 (-3.29, 0.50)
$T1 * \mathcal{G}T1(10)$	-0.64	0.01	-0.72 (-1.62, 0.25)
$T1 * \mathcal{G}T1(20)$	0.41	0.02	0.46 (-1.70, 2.25)

Table A.2: Regression coefficients from the second fit of the logistic regression model on all of the MRI studies in validation set 1

Fit 2	Coefficient	Standard Error	Bootstrapped Mean (95% CI)
Intercept	-5.66	0.01	-5.98 (-7.09, -5.12)
FLAIR	1.87	0.00	1.96 (1.52, 2.48)
$\mathcal{G}^2FLAIR(10)$	3.96	0.04	2.82 (-2.81,10.31)
$\mathcal{G}^2FLAIR(20)$	-3.19	0.06	-1.44 (-15.18, 10.25)
PD	-0.78	0.00	-0.80 (-1.21, -0.29)
$\mathcal{G}^2PD(10)$	-1.20	0.03	-0.97 (-4.71, -0.29)
$\mathcal{G}^2PD(20)$	3.16	0.04	2.83 (-2.96, 7.97)
T2	0.91	0.00	0.91 (0.56, 1.28)
$\mathcal{G}^2T2(10)$	-2.63	0.03	-2.51 (-6.33,1.47)
$\mathcal{G}^2T2(20)$	0.10	0.04	-0.37 (-8.44, 6.05)
T1	1.90	0.00	1.92 (1.49, 2.40)
$\mathcal{G}^2T1(10)$	5.38	0.02	5.83 (2.61, 9.20)
$\mathcal{G}^2T1(20)$	-11.57	0.03	-12.11 (-17.91, -6.78)
$FLAIR * \mathcal{G}^2FLAIR(10)$	0.10	0.02	0.58 (-1.27, 2.40)
$FLAIR * \mathcal{G}^2FLAIR(20)$	-4.97	0.03	-5.71 (-9.81, -1.98)
$PD * \mathcal{G}^2PD(10)$	-2.26	0.01	-2.36 (-2.36, -1.00)
$PD * \mathcal{G}^2PD(20)$	2.73	0.02	2.95 (0.36, 5.26)
$T2 * \mathcal{G}^2T2(10)$	1.59	0.02	1.55 (0.15, 2.68)
$T2 * \mathcal{G}^2T2(20)$	-4.52	0.02	-4.43 (-7.04, -2.37)
$T1 * \mathcal{G}^2T1(10)$	-0.82	0.01	-0.86 (-1.70, 0.07)
$T1 * \mathcal{G}^2T1(20)$	0.33	0.02	0.28 (-1.63, 2.01)

A.2 Repeated subjects for validation study 1

Table 3 displays the preferred method amongst the OASIS empirically adjusted threshold and LesionTOADS segmentation from the neuroradiologist, neurologist, and radiologist for the five subjects with repeated studies from validation set 2. There was a tie between the two methods for the Radiologist's second ranking of Subject 5.

Table A.3: Preferred method from neuroradiologist, neurologist, and radiologist for the 5 repeated subjects in validation set 2

Neuroradiologist	First Ranking	Second Ranking	Ranking Preserved
Study 1	OASIS	OASIS	Yes
Study 2	LesoinTOADS	LesionTOADS	Yes
Study 3	OASIS	OASIS	Yes
Study 4	OASIS	OASIS	Yes
Study 5	LesionTOADS	OASIS	**No**
Neurologist			
Study 1	OASIS	OASIS	Yes
Study 2	LesionTOADS	LesionTOADS	Yes
Study 3	OASIS	OASIS	Yes
Study 4	OASIS	OASIS	Yes
Study 5	OASIS	LesionTOADS	**No**
Radiologist			
Study 1	OASIS	OASIS	Yes
Study 2	OASIS	LesionTOADS	**No**
Study 3	OASIS	OASIS	Yes
Study 4	OASIS	LesionTOADS	**No**
Study 5	LesionTOADS	tie	**No**

Appendix B

Appendix to "A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI"

B.1 Classification algorithms

Voxel-level lesion segmentation is a binary classification problem with two classes, voxels either contain lesion or do not contain lesion. We use the notation $y_v \in \{0, 1\}$ to be a class label indicating whether or not voxel v of a brain image from MRI study contains a lesion. Here we provide a brief summary of each classification algorithm used in our analysis and the tuning parameters associated with the algorithm. A complete overview of these algorithms can be found in The Elements of Statistical Learning: Data Mining, Inference and Prediction [163]. For each classification algorithm, Table 4.1 in the Methods section shows the R package used to fit the algorithm and the values for the tuning parameters. In fitting each of the algorithms on the training set, 10-fold cross validation was used to optimize all tuning parameters.

B.1.1 Logistic regression

Logistic regression is a special case of the generalized linear model, with a logistic link function. The generalized linear model with logistic link function, t, is $t\{P(y_v = 1|X = x)\} = \beta_0 + \beta^T x$, where $t(s) = \log(\frac{s}{1-s})$. The resulting decision boundary for logistic regression is linear.

B.1.2 Linear discriminant analysis

Linear discriminant analysis models each class as a multivariate Gaussian with probability distribution functions $f_0(v)$ and $f_1(v)$ with means μ_0 and μ_1 respectively. Linear discriminant analysis makes the assumption that both classes have a common covariance matrix Σ. Prior probabilities of belonging to a class are calculated from the training data and are denoted as π_0 and π_1. Then, the probability of voxel being part of a lesion is defined as $P(y_v = 1|X = x) = \frac{f_1(x)\pi_1}{f_0(x)\pi_0}$. The resulting decision boundary is linear.

B.1.3 Quadratic discriminant analysis

Quadratic discriminant analysis is similar to linear discriminant analysis, but allows for the two classes to have different covariance matrices, Σ_0 and Σ_1. The resulting decision boundary is a quadratic surface.

B.1.4 Gaussian mixture model

In a Gaussian mixture model, each class k is modeled as a mixture of normal distributions with density $P(X = x|y_v = k) = \sum_{r=1}^{R_k} \psi_{kr}\phi(X; \mu_{kr}, \Sigma), k = 0, 1$. The mixing proportions for each class, ψ_{kr}, must sum to one. Then, $P(y_v = 1|X = x) = \frac{P(X=x|y_v=1)\pi_1}{P(X=x|y_v=1)\pi_1+P(X=x|y_v=0)\pi_0}$, where π_0 and π_1 are the prior probabilities of belonging to each class, calculated from the training data. The decision boundary for the Gaussian mixture model is multi-modal

B.1.5 Support vector machine with linear kernel

In a support vector machine with linear kernel, we aim to find the hyperplane that separates the two classes with the largest margin. The hyperplane that separates the two class is defined as, $\{x : f(x) = \beta_0 + \beta^t x\}$. A cost C is assigned to voxels v that are classified incorrectly, and $\sum \psi_i \leq C$, where ψ_i is the amount which the voxel is on the incorrect side of the margin. We search over the $C = 1/8$, 1/4, 1/2, 1, 2, 4, and 8. The decision boundary for the support vector machine is linear in the implicit kernel space, but nonlinear in the original space.

B.1.6 Random forest

A random forest combines multiple classification trees and these classification trees vote on which class a new voxel should be classified as. In the implementation of the random forest, we use 500 classification trees. The parameter we tune the algorithm over is the mtry parameter, the number of variables sampled at each split of the decision tree. We search over mtry = 1 to the dimension of the feature space. The decision boundary for each tree is piecewise linear.

B.1.7 k-nearest neighbors

In k-nearest neighbors to classify a new voxel, v, with features $X = x$, we calculate the distance of its features from the features of the voxels in the labeled training set, using Euclidean distance. The k voxels in the training set with the smallest distance to the new voxel vote on the class of the new voxel. The tuning parameter for the k-nearest neighbor algorithm is the number of voxels to be used. We search over k =1, 10, and 100 neighbors.

B.1.8 Neural network

In a single hidden layer neural network, derived features, Z, are created from linear combinations of the input features, $X : Z_p = \sigma(\alpha_{op}+\alpha_p^T X), p \in \{1, ..., P\}$. These derived features Z are called hidden units. We elected to use a sigmoid activation function σ, $\sigma(s) = \frac{1}{(1+\exp{(-s)})}$, as this is typical. The outcome, y_v, is then modeled as a function of the hidden units: $y_v = t(\beta_0 + \beta^T z)$. We use a single hidden layer neural network with sigmoid activation function; 10-fold cross validation is used to select the number of hidden units in the algorithm and search over P = 1, 5, and 10.

B.1.9 Super learner

The super learner is a method for combining class estimations from different classification algorithms,by weighting the classifiers according to their prediction performance using a cross-validation loss function. We use the super leaner algorithm with 10-fold cross validation with mean squared error loss, but any other cross validation or loss function may be used. The super learner requires a library of K supervised classification algorithms. To train the super learner with 10-fold cross validation, the training set, T, of size n voxels is partitioned into 10 equal samples. For each of these 10 samples, the K algorithms are trained on the remaining data in the training set. A prediction for each voxel v in the reserved sample is then made for the k^{th} classification algorithm, h_k, denoted as $\varphi_k(v), k \in \{1, \dots, K\}$. After this is performed for all 10 samples, a coefficient α_k for each classification algorithm is selected, to minimize the mean square error: $\frac{1}{n}\sum_{v \in T}[y_v - \sum_k \alpha_k \varphi_k(v)]^2$, under the constraint $\sum_k \alpha_k = 1$. The fitted classifiers with the coefficients α_k can then used to make predictions for new samples. We use the SuperLearner R package based on all of the other supervised classification algorithms and tuning parameters used in this analysis. To decrease prediction time for new data, algorithms with α_k coefficients close to zero are dropped during the prediction phase (the onlySL parameter is set to TRUE when making predictions; the default for this parameter is FALSE) .

Appendix C

Appendix to "Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions"

C.1 Longitudinal profile pipeline

Here we provide a more complete description of the procedure for extracting the longitudinal voxel-level lesion profiles, which is divided into four steps: (1) identifying voxels with new lesion formation, (2) intensity normalization, (3) temporal alignment, and (4) temporal interpolation. All voxels in this analysis are part of incident or enlarging lesions detected during the subject's follow-up period. All voxels that are part of lesions that existed at baseline are excluded from the analysis.

C.1.1 Identifying voxels with new lesion formation

When identifying voxels with new lesion formation, we distinguish between areas that contain vasogenic edema (which we will refer to simply as "edema") and actual lesion, which both manifest as areas of abnormal signal intensity, especially on the T2-weighted sequences. For this analysis, we are interested in areas with tissue damage, as opposed to edema. To identify areas with new lesion formation, we first find areas in the MRI with new abnormal signal intensity, which includes both edema and lesion. We then segment lesions by analyzing subsequent visit data.

SuBLIME segmentation of voxel-level lesion incidence and enlargement is a method for detecting voxels that are part of an area of new abnormal signal intensity between two MRI studies [9]. For each subject, we produce SuBLIME maps between the respective sets of consecutive MRI studies. We exclude all abnormal signal intensity areas that contained fewer than 27 voxels, as these areas could be artifact or noise. We then produce cross-sectional lesion segmentations using OASIS segmentation of abnormal signal presence [7]. As the signal from edema disappears rapidly from the MRI after lesion formation, we locate the incident abnormal signal voxels using SuBLIME, but only include the voxels that are detected by OASIS at the following study visit, as these voxels should not contain edema. Therefore, only voxels that have an MRI study within 40 days after SuBLIME detects the area of abnormal signal intensity, where the intensity remains in the OASIS maps, are considered as lesion tissue and used in this analysis, as by this time edema would subside. We use expert validation by a neuroradiologist and a neurologist, both with experience in MS imaging, to confirm that this method is identifying lesion tissue, which we describe in detail in the section *Expert Validation*. The figure below shows the SuBLIME segmentation for each study and the OASIS segmentation for each study, corresponding to Figure 5.2. The row corresponding to the SuBLIME segmentation is further divided into edema and lesion voxels using the method described above. Only voxels that are part of lesion tissue are used in the analysis.

C.1.2 Intensity normalization

Structural MRI is acquired in arbitrary units. Therefore, in addition to pulse sequence similarity, intensity normalization is paramount for comparing intensities in a voxel over time within subject and for comparing voxel intensities between subjects. We normalize each sequence separately on each scan by calculating the mean and standard deviation over a mask of the normal-appearing white matter (NAWM) from the brain segmentation described in the section *Image Acquisition and Preprocessing* [66]. We then subtract the mean from the intensity in each voxel and divide by the standard deviation [93, 155]. Let $S_{ilv}(t)$ be the observed intensity from imaging sequence S in voxel v for subject i in lesion l at study time t, with $S =$ FLAIR, T1, T2, and PD. Let $\mu_{Si}(t)$ and $\sigma_{Si}(t)$ be the mean and standard deviation, respectively, over the NAWM mask for sequence S at scan time t for subject i . Then the normalized intensity in voxel v in lesion l for subject i at scan time t is:

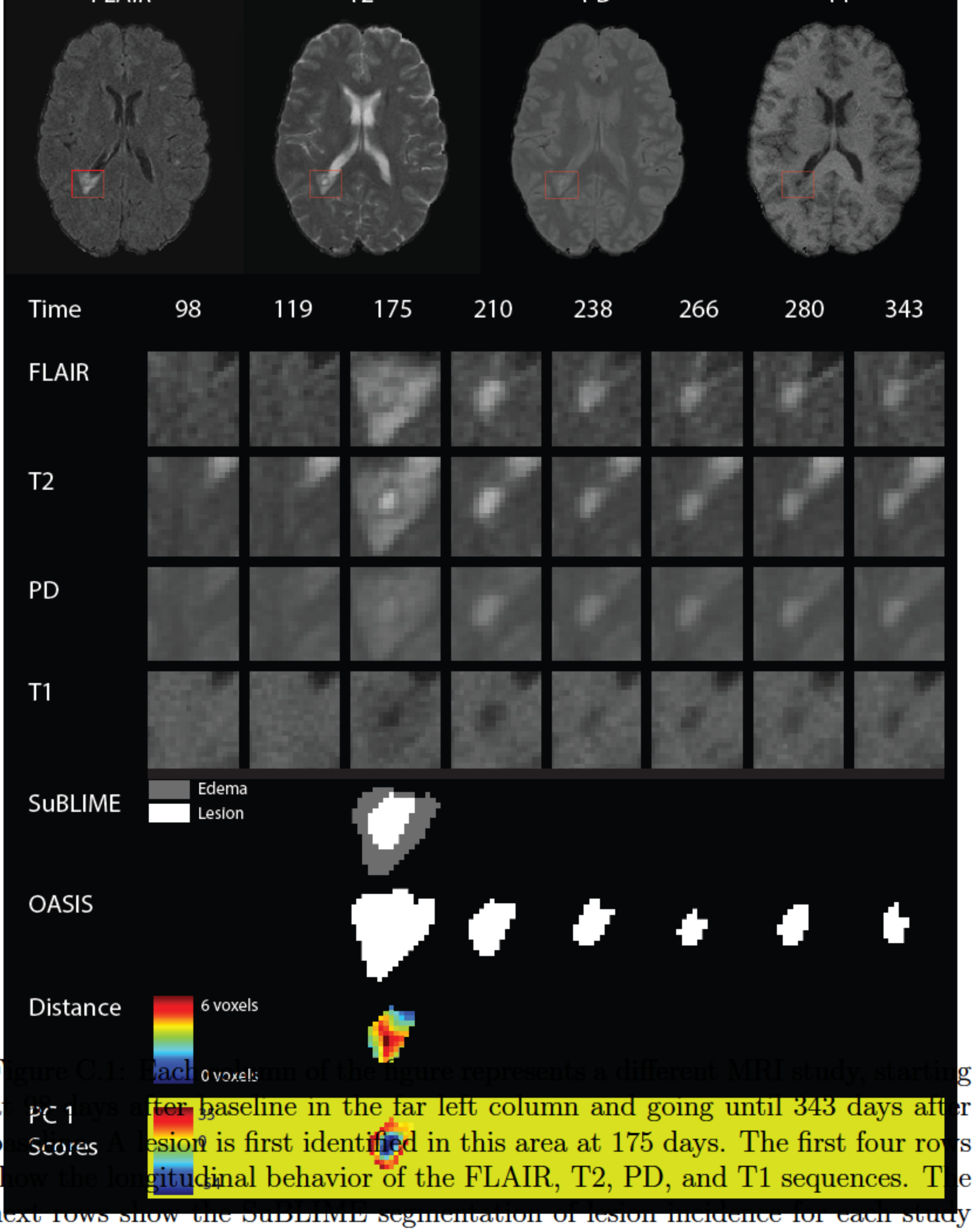


Figure C.1: Each column of the figure represents a different MRI study, starting at 98 days after baseline in the far left column and going until 343 days after baseline. A lesion is first identified in this area at 175 days. The first four rows show the longitudinal behavior of the FLAIR, T2, PD, and T1 sequences. The next rows show the SuBLIME segmentation of lesion incidence for each study and the OASIS segmentation of lesion presence in each study. The SuBLIME segmentation has been further divided into areas of edema and lesion.

$$S_{ilv}^{N}(t) = \frac{S_{ilv}(t) - \mu_{Si}(t)}{\sigma_{Si}(t)}$$

Thus, all image intensities are expressed as a departure, in multiples of standard deviation of white matter intensities, from the subject's mean normal-appearing white matter (NAWM) in each imaging sequence.

C.1.3 Temporal alignment

The date of the study visit at which SuBLIME detects the lesion voxels is considered the time of incidence for this voxel. If a voxel is determined to be a new or enlarging lesion by SuBLIME more than once over the follow-up time, the first occurrence is considered to be the time of lesion incidence for that voxel. Voxel profiles from incident lesions during the follow-up of each subject are aligned in time, using the time of incidence as time 0, therefore any observations before incidence have a negative time and after lesion incidence have a positive time. Let t' denote this aligned time scale. Then we have $S^N_{ilv}(t')$, where $S^N_{ilv}(0)$ indicates the intensity in sequence S at the time of lesion incidence.

C.1.4 Temporal interpolation

Next we perform a temporal linear interpolation so that all voxels are observed on the same time grid. In this work, we are interested in the lesion dynamics only after lesion incidence, therefore we perform the linear interpolation within the window after lesion incidence and up to 200 days post-incidence. The end point of 200 days is selected as it has been previously found that new T2 lesions show the most dramatic changes in intensity for three to four months [146], and we opt to be conservative and include data beyond this reported stabilization point. Voxels are selected for the analysis if the subject has at least one visit 200 days or more after lesion incidence. Of the 60 subjects in this analysis, 34 have voxel profiles meeting this inclusion criteria, after removing the three subjects for poor longitudinal registration. We linearly interpolate over a grid of 0 to 200 days by increments of 5 days so that all profiles are observed on the same time grid. We denote the vector of observations from a voxel over this time grid for sequence S as S^N_{ilv}, where S^N_{ilv} is a 1×41 vector.

C.2 Parametric bootstrapping procedure

Let B be the number of bootstrap samples to be performed and let b index these B samples. Let Y_{ilv} be the outcome for an observation indexed by i, l, and v. Let $\boldsymbol{X}$ be the design matrix and $\boldsymbol{\beta}$ be the vector of the coefficients. For this analysis we have a model of the form:

$$Y_{ilv} = \boldsymbol{X}\boldsymbol{\beta} + b_i + b_l + \epsilon_{ilv}$$

where $b_i \sim N\left(0, \sigma_i^2\right)$ and $b_l \sim N\left(0, \sigma_l^2\right)$ are random intercepts, and $\epsilon_{ilv} \sim N\left(0, \sigma_\epsilon^2\right)$ is an error term. For the parametric model, we fit the above mixed-effect model to get an estimate of $\boldsymbol{\beta}$, which we denote as $\hat{\boldsymbol{\beta}}$. We then fix this estimate, and keep $\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Using the fitted variances, $\hat{\sigma}_i^2$, $\hat{\sigma}_l^2$ and $\hat{\sigma}_\epsilon^2$, we generate a random intercept for each lesion from a $N\left(0, \hat{\sigma}_i^2\right)$ distribution, a random intercept for each subject from a $N\left(0, \hat{\sigma}_l^2\right)$, and random noise for each voxel from a $N\left(0, \hat{\sigma}_\epsilon^2\right)$. We then add the random intercepts and noise to $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ for the corresponding observation and use this as our outcome to refit the model and get out bootstrapped coefficient vector $\boldsymbol{\beta}_b^*$. To obtain the bootstrap sample, we repeat this procedure B times.

C.3 Expert validation

Examples of the set of evaluation images presented to the experts for each lesion are shown in Figures C.2, C.3, C.4, and C.5. The first row of the figures shows the full axial slice for the FLAIR, T2, PD, and T1 volumes that contains the largest number of voxels with abnormal signal intensity. The second through fourth rows show the entire collection of longitudinal scans for a box containing the abnormal signal intensity in the FLAIR, T2, PD, and T1 weighted volumes at the baseline time point for this axial slice. The scans are displayed in chronological order, from first time point to last time point, from left to right. The fifth row shows the segmentation of the lesion and edema tissue within this box at each time point. The sixth row shows the score on the first PC for the voxels segmented as lesion tissue, displayed at the time of lesion incidence for each voxel. The seventh through tenth row show the entire collection of longitudinal scans for the FLAIR, T2, PD, and T1 weighted volumes within this box, with the score for the first PC overlaid on the images for each scan after lesion incidence. The last row shows the scale for the score on the first PC. The figures show examples of the four different ratings for the score on the first PC. Both raters rate the scans as either (1) failed miserably, (2) some redeeming features, (3) passed with minor errors, or (4) passed.

C.4 Principal component analysis and regression

Table C.1 shows the coefficient estimates, standard errors, t-statistics, the p-values using the normal approximation, and the 95% bootstrapped confidence intervals for the multivariate PCA regression model. Table C.2 shows the same for the individual univariate PCA regression models.

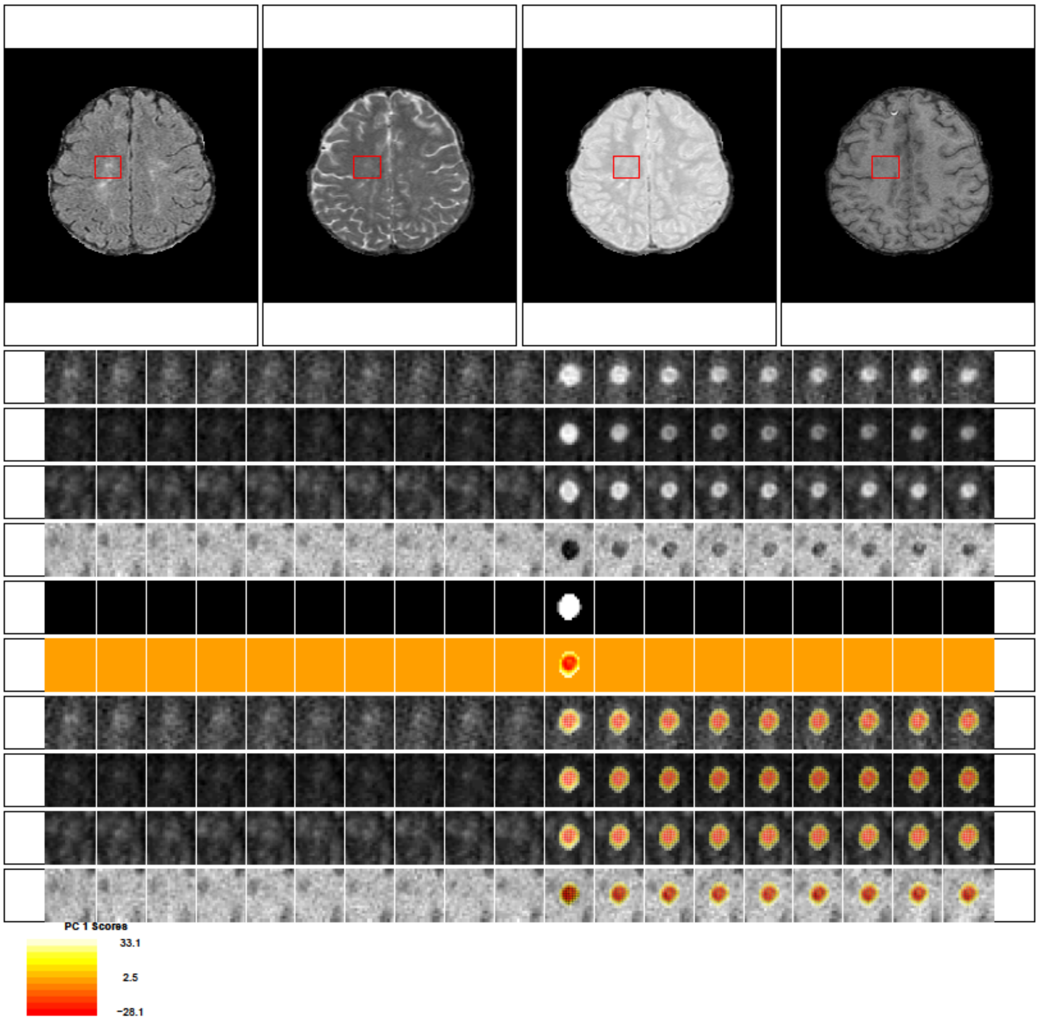


Figure C.2: This scan received a rating of 4 for the score on the first PC from both raters. Both raters also gave a rating of 4 for the lesion segmentation.

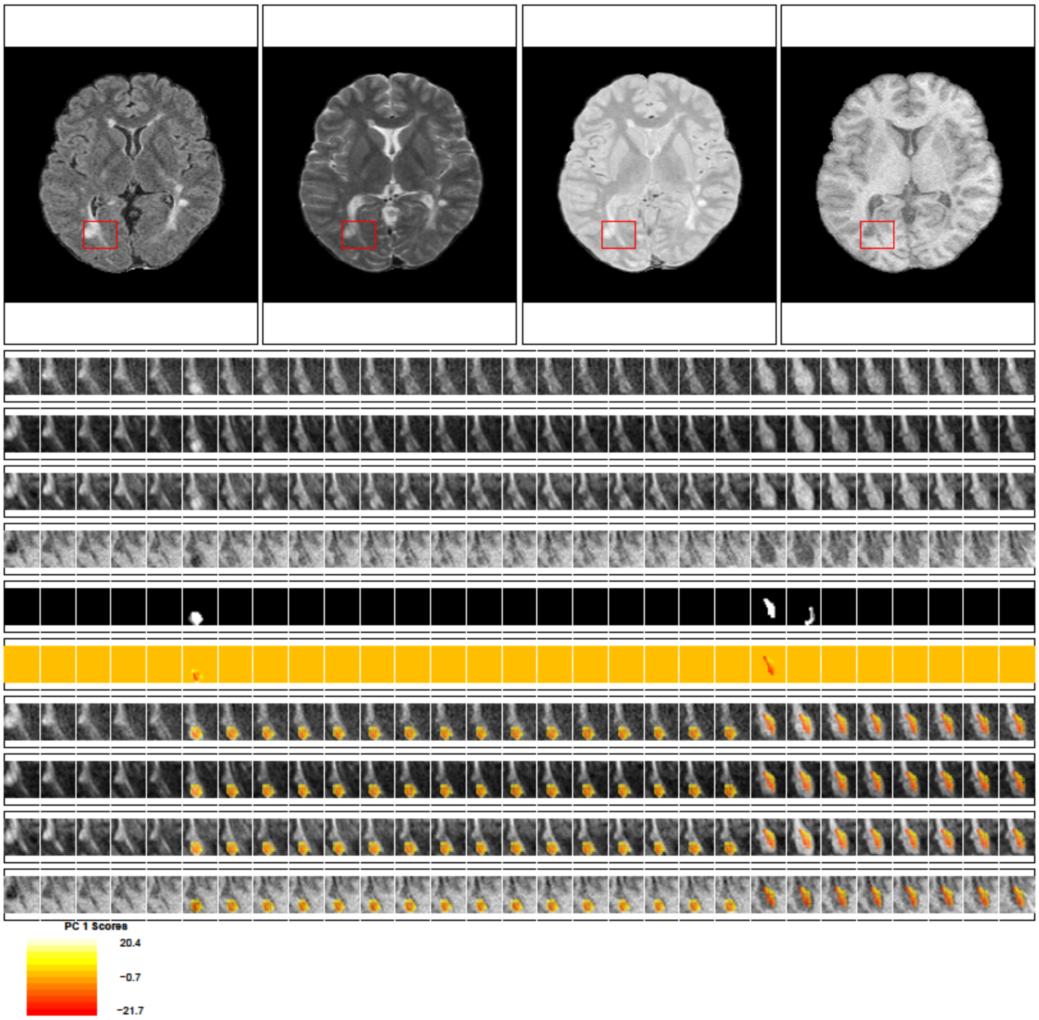


Figure C.3: This scan received a rating of 3 for the score on the first PC from both raters. Both raters also gave a rating of 3 for the lesion segmentation. Note that at the 23rd time point new lesion voxels are segmented, but the score for the first PC is not produced for this time point, as the voxels did not meet the scanning criteria for being included in the analysis.

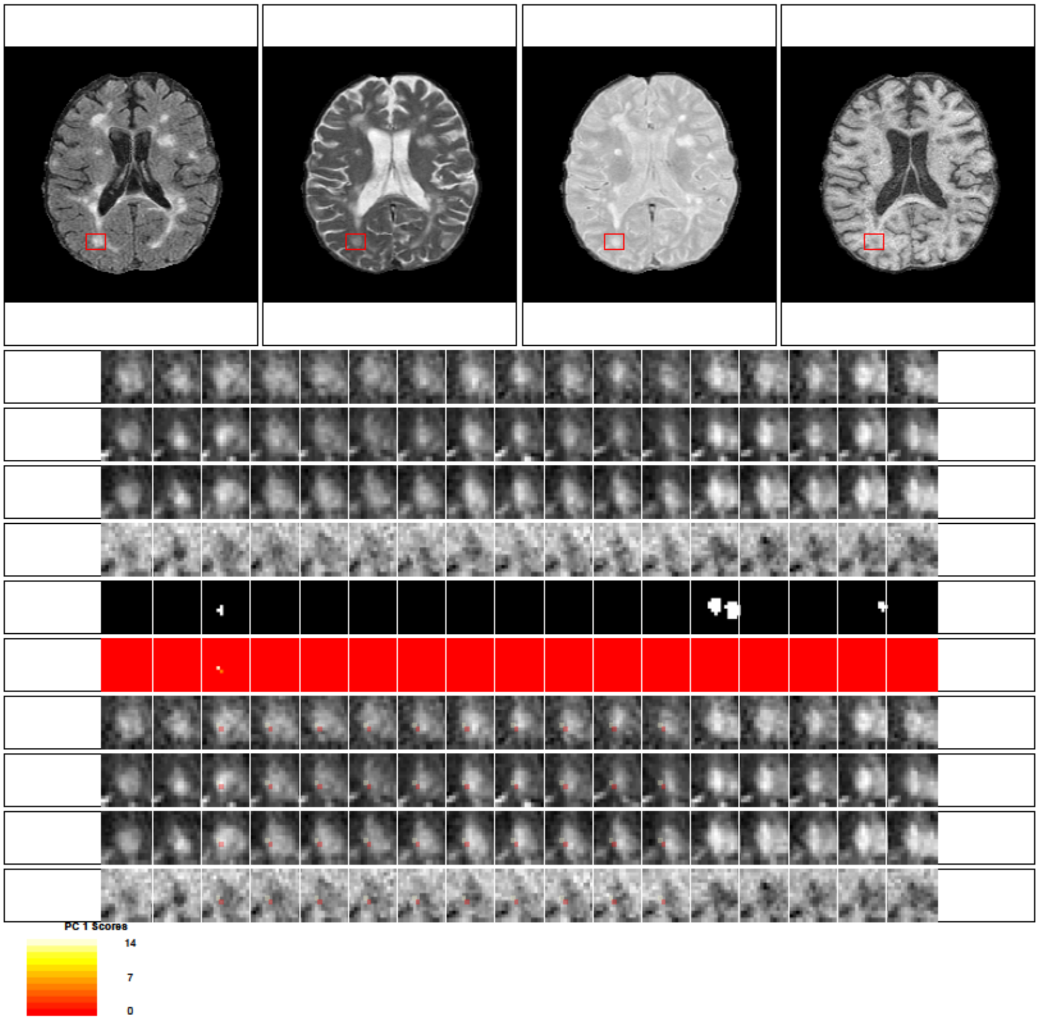


Figure C.4: This scan received a rating of 2 for the score on the first PC from both raters. Both raters also gave a rating of 2 for the lesion segmentation. The neuroradioloigst commented that this scan received a low rating because it was not clear that the segmented portion for time point 3 was lesion.

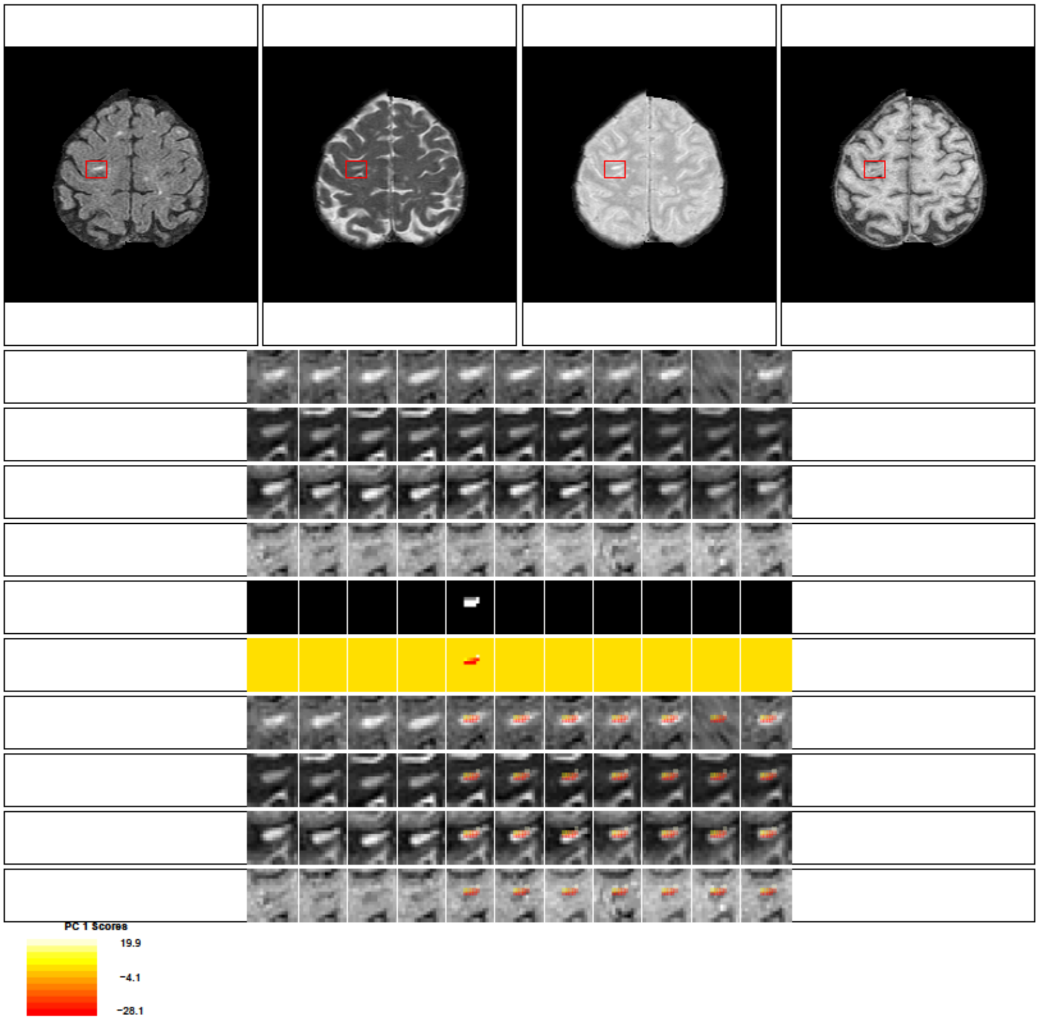


Figure C.5: This scan received a rating of 1 for the score on the first PC from both raters. Both raters also gave a rating of 1 for the lesion segmentation. Both raters commented that the low rating was because the lesion had existed in all time points and was not a new lesion.

	Estimate	Standard Error	t-value	p-value	95% Bootstrapped CI
SPMS	2.15	4.41	0.49	0.63	(-6.19, 10.93)
Distance to Boundary	-9.39	0.08	-123.74	0.00	(-9.56, -9.25)
Age	-0.21	0.18	-1.16	0.25	(-0.57, 0.13)
$(\text{Age} - 4)_+$	-0.10	0.23	-0.42	0.68	(-0.54, 0.35)
Steroids	4.26	0.79	5.42	0.00	(2.67, 5.85)
Male	1.16	2.55	0.45	0.65	(-3.94, 6.61)
Treatment	5.39	0.36	15.03	0.00	(4.67, 6.08)
Intercept	8.89	1.92	4.64	0.00	(5.17, 12.85)

Table C.1: Coefficient estimates, standard errors, t-statistics, p-values, and bootstrapped 95% confidence intervals for the multivariate PCA regression model.

	Estimate	Standard Error	t-value	p-value	95% Bootstrapped CI
SPMS	0.65	4.11	0.16	0.88	(-7.71, 9.18)
Distance to Boundary	-9.37	0.08	-123.18	0.00	(-9.52, -9.22)
Age	0.89	0.19	4.58	0.00	(0.51, 1.23)
$(\text{Age} - 4)_+$	-1.55	0.24	-6.40	0.00	(-1.95, -1.14)
Steroids	6.03	0.78	7.77	0.00	(4.55, 7.59)
Male	0.43	2.43	0.18	0.86	(-4.32, 4.97)
Treatment	4.48	0.38	11.76	0.00	(3.67, 5.25)

Table C.2: Coefficient estimates, standard errors, t-statistics, p-values, and bootstrapped 95% confidence intervals for the univariate PCA regression model.

C.5 Function-on-scalar regression

The coefficient functions from the function-on-scalar regression with bootstrapped 95% confidence intervals with the T2, PD, and T1 profile as the outcome are shown below. Similar to using the FLAIR profile as the outcome, only the distance to the boundary and age were found to be different from 0 at any point along the profile.

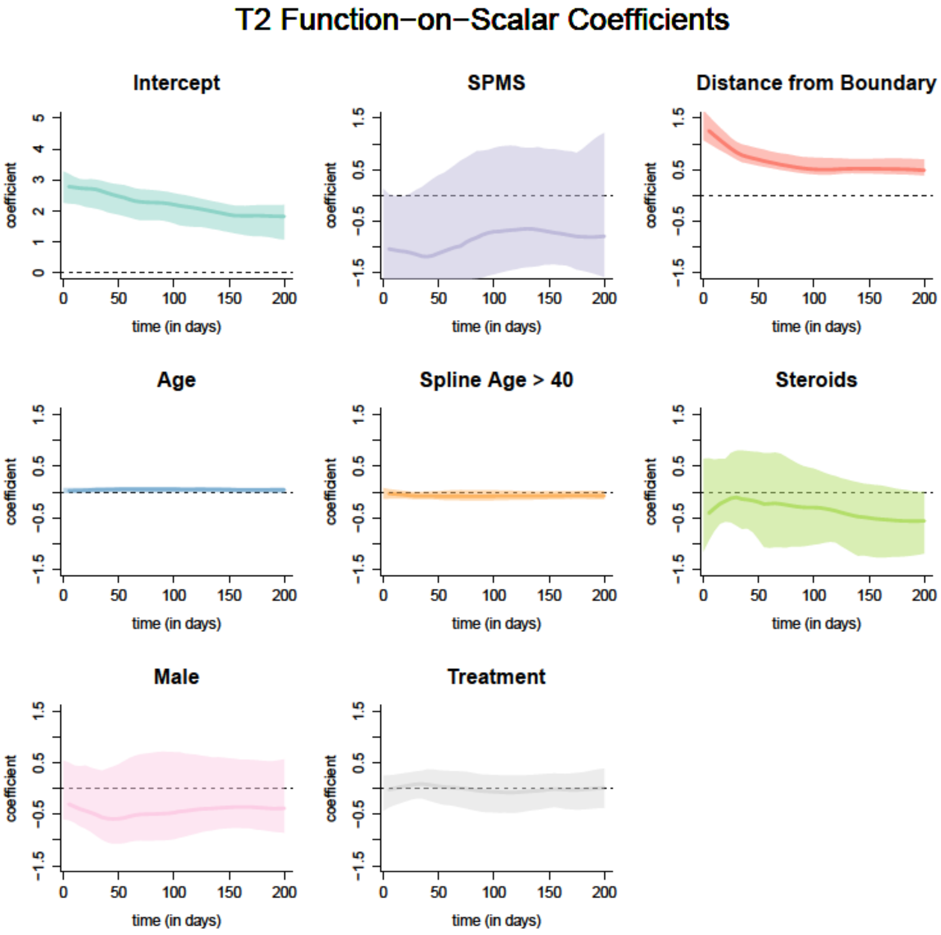


Figure C.6: Each dark line represents the coefficient function, and the shaded area represents a bootstrapped, point-wise 95% confidence interval. Along the y-axis is the value of the coefficient function at each time point. Only distance from the boundary and age were found to be different from 0 at any point along the profile.

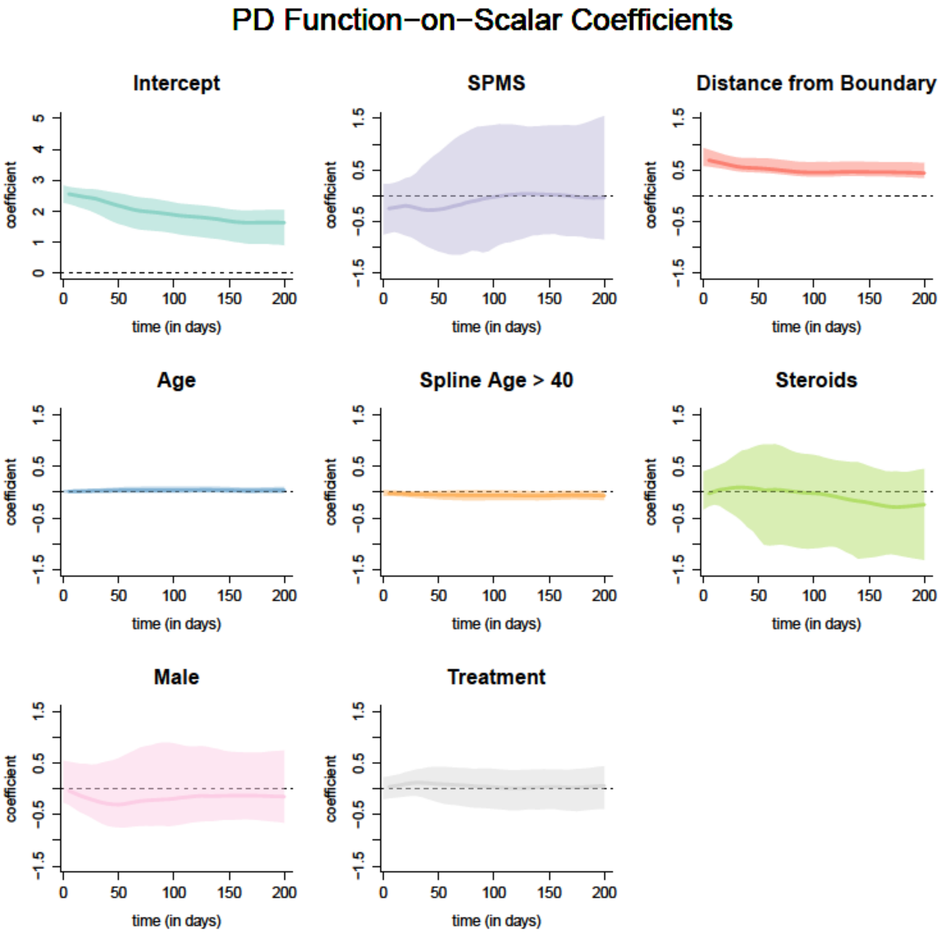


Figure C.7: Each dark line represents the coefficient function, and the shaded area represents a bootstrapped, point-wise 95% confidence interval. Along the x-axis of each plot is the time in days from lesion incidence. Along the y-axis is the value of the coefficient function at each time point. Only distance from the boundary and age were found to be different from 0 at any point along the profile.

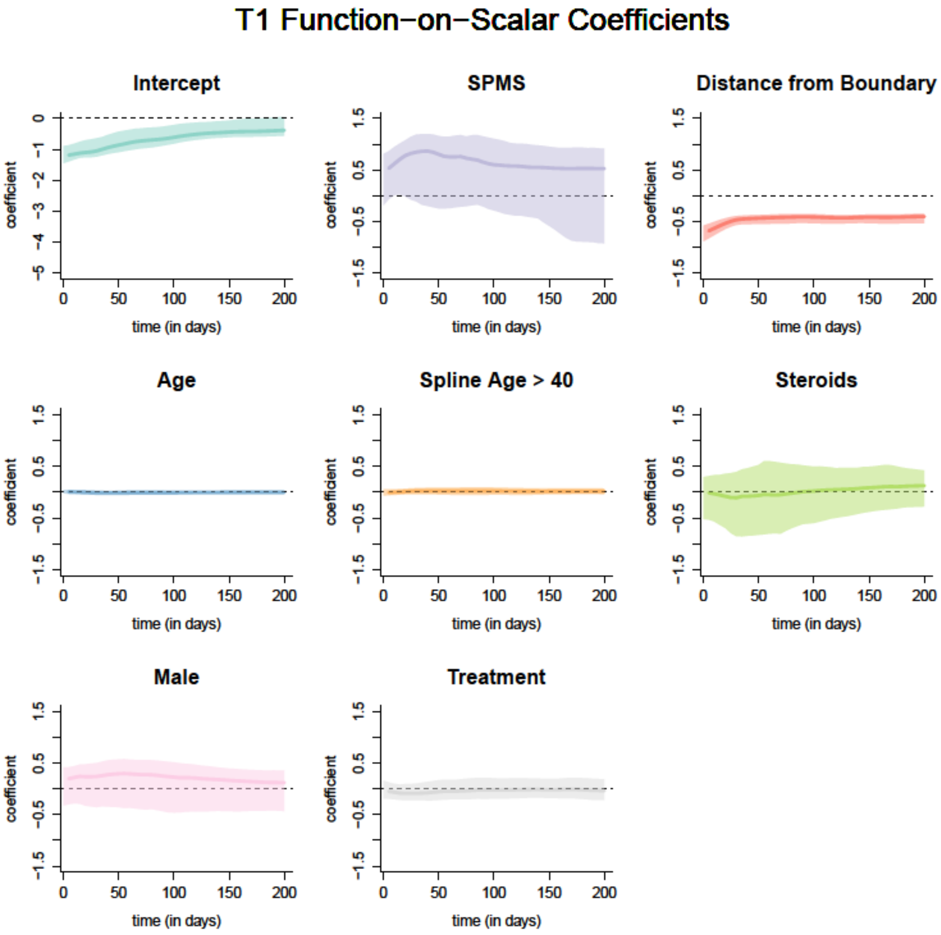


Figure C.8: Each dark line represents the coefficient function, and the shaded area represents a bootstrapped, point-wise 95% confidence interval. Along the x-axis of each plot is the time in days from lesion incidence. Along the y-axis is the value of the coefficient function at each time point. Only distance from the boundary and age were found to be different from 0 at any point along the profile.

Bibliography

[1] M. A. Sahraian and E.-W. Radue, *MRI atlas of MS lesions.* Springer Science & Business Media, 2007.

[2] À. Rovira and A. León, "Mr in the diagnosis and monitoring of multiple sclerosis: an overview," *European journal of radiology*, vol. 67, no. 3, pp. 409–414, 2008.

[3] À. Rovira, J. Swanton, M. Tintoré, E. Huerga, F. Barkhof, M. Filippi, J. L. Frederiksen, A. Langkilde, K. Miszkiel, C. Polman *et al.*, "A single, early magnetic resonance imaging study in the diagnosis of multiple sclerosis," *Archives of Neurology*, vol. 66, no. 5, pp. 587–592, 2009.

[4] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos *et al.*, "Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria," *Annals of neurology*, vol. 69, no. 2, pp. 292–302, 2011.

[5] F. Barkhof, "The clinico-radiological paradox in multiple sclerosis revisited," *Current opinion in neurology*, vol. 15, no. 3, pp. 239–245, 2002.

[6] E. M. Sweeney, A. Eloyan, R. T. Shinohara, and C. M. Crainiceanu, "A tutorial for multisequence clinical structural brain MRI," in *Handbook of Modern Statistical Methods: Neuroimaging Data Analysis*, to appear.

[7] E. M. Sweeney, R. T. Shinohara, N. Shiee, F. J. Mateen, A. A. Chudgar, J. L. Cuzzocreo, P. A. Calabresi, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, "OASIS is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI," *NeuroImage: clinical*, vol. 2, pp. 402–413, 2013.

[8] E. M. Sweeney, J. T. Vogelstein, J. L. Cuzzocreo, P. A. Calabresi, D. S. Reich, C. M. Crainiceanu, and R. T. Shinohara, "A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural mri," *PloS one*, vol. 9, no. 4, p. e95753, 2014.

[9] E. Sweeney, R. Shinohara, C. Shea, D. Reich, and C. Crainiceanu, "Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal mri," *American Journal of Neuroradiology*, vol. 34, no. 1, pp. 68–73, 2013.

[10] E. M. Sweeney, R. T. Shinohara, B. E. Dewey, M. K. Schindler, J. Muschelli, D. S. Reich, C. M. Crainiceanu, and A. Eloyan, "Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions," *NeuroImage: Clinical*, vol. 10, pp. 1–17, 2016.

[11] S. W. Atlas, *Magnetic resonance imaging of the brain and spine.* Lippincott Williams & Wilkins, 2009, vol. 1.

[12] B. Kevles, *Naked to the bone: Medical imaging in the twentieth century.* Rutgers University Press, 1997.

[13] M. A. Lindquist *et al.*, "The statistical analysis of fmri data," *Statistical Science*, vol. 23, no. 4, pp. 439–464, 2008.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[15] B. Whitcher, V. J. Schmid, and A. Thornton, "Working with the DICOM and NIfTI data standards in R," *Journal of Statistical Software*, vol. 44, no. 6, pp. 1–28, 2011. [Online]. Available: http://www.jstatsoft.org/v44/i06/

[16] R. Shinohara, C. Crainiceanu, B. Caffo, M. Gaitn, and D. Reich, "Population-wide principal component-based quantification of blood-brain-barrier dynamics in multiple sclerosis," *Neuroimage*, vol. 57, no. 4, pp. 1430–1446, 2011.

[17] R. Shinohara, A. Goldsmith, F. Mateen, C. Crainiceanu, and D. Reich, "Predicting breakdown of the blood-brain barrier in multiple sclerosis without contrast agents," *American Journal of Neuroradiology*, vol. 33, no. 8, pp. 1586–1590, 2012.

[18] G. van Rossum and J. de Boer, "Interactively testing remote servers using the python programming language," *CWI Quarterly*, vol. 4, no. 4, pp. 283–303, 1991. [Online]. Available: http://www.jstatsoft.org/v44/i06/

[19] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *Medical Imaging, IEEE Transactions on*, vol. 17, no. 1, pp. 87–97, 1998.

[20] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 6, pp. 1310–1320, 2010.

[21] S. Greven, C. M. Crainiceanu, B. Caffo, and D. Reich, "Longitudinal functional principal component analysis," *Electronic Journal of Statistics*, vol. 4, pp. 1022–1054, 2010.

[22] M. A. Bernstein, K. F. King, and X. J. Zhou, *Handbook of MRI pulse sequences*. Elsevier, 2004.

[23] E. L. Gedamu, D. Collins, and D. L. Arnold, "Automated quality control of brain mr images," *Journal of Magnetic Resonance Imaging*, vol. 28, no. 2, pp. 308–319, 2008.

[24] T. Ihalainen, O. Sipilä, and S. Savolainen, "Mri quality control: six imagers studied using eleven unified image quality parameters," *European radiology*, vol. 14, no. 10, pp. 1859–1865, 2004.

[25] D. E. Rex, J. Q. Ma, and A. W. Toga, "The loni pipeline processing environment," *Neuroimage*, vol. 19, no. 3, pp. 1033–1048, 2003.

[26] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[27] B. C. Lucas, J. A. Bogovic, A. Carass, P.-L. Bazin, J. L. Prince, D. L. Pham, and B. A. Landman, "The java image science toolkit (jist) for rapid prototyping and publishing of neuroimaging software," *Neuroinformatics*, vol. 8, no. 1, pp. 5–17, 2010.

[28] M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, "Medical image processing, analysis and visualization in clinical research," in *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on.* IEEE, 2001, pp. 381–386.

[29] Z. Hou, "A review on mr image intensity inhomogeneity correction," *International Journal of Biomedical Imaging*, vol. 2006, 2006.

[30] U. Vovk, F. Pernus, and B. Likar, "A review of methods for correction of intensity inhomogeneity in mri," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 3, pp. 405–421, 2007.

[31] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[32] C. Fennema-Notestine, I. B. Ozyurt, C. P. Clark, S. Morris, A. Bischoff-Grethe, M. W. Bondi, T. L. Jernigan, B. Fischl, F. Segonne, D. W. Shattuck *et al.*, "Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location," *Human brain mapping*, vol. 27, no. 2, pp. 99–113, 2006.

[33] J.-M. Lee, U. Yoon, S. H. Nam, J.-H. Kim, I.-Y. Kim, and S. I. Kim, "Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error," *Computers in biology and medicine*, vol. 33, no. 6, pp. 495–507, 2003.

[34] A. Carass, J. Cuzzocreo, M. B. Wheeler, P.-L. Bazin, S. M. Resnick, and J. L. Prince, "Simple paradigm for extra-cerebral tissue removal: algorithm and analysis," *NeuroImage*, vol. 56, no. 4, pp. 1982–1992, 2011.

[35] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 9, pp. 1617–1634, 2011.

[36] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain mri segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[37] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 7, pp. 903–921, 2004.

[38] K. K. Leung, J. Barnes, M. Modat, G. R. Ridgway, J. W. Bartlett, N. C. Fox, and S. Ourselin, "Brain maps: an automated, accurate and robust brain extraction technique using a template library," *Neuroimage*, vol. 55, no. 3, pp. 1091–1108, 2011.

[39] J. Doshi, G. Erus, Y. Ou, B. Gaonkar, and C. Davatzikos, “Multi-atlas skull-stripping,” *Academic radiology*, vol. 20, no. 12, pp. 1566–1576, 2013.

[40] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, and D. L. Collins, “Beast: Brain extraction based on nonlocal segmentation technique,” *NeuroImage*, vol. 59, no. 3, pp. 2362–2373, 2012.

[41] D. W. Shattuck, G. Prasad, M. Mirza, K. L. Narr, and A. W. Toga, “Online resource for validation of brain segmentation methods,” *NeuroImage*, vol. 45, no. 2, pp. 431–439, 2009.

[42] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney *et al.*, “Advances in functional and structural mr image analysis and implementation as fsl,” *Neuroimage*, vol. 23, pp. S208–S219, 2004.

[43] B. C. Lucas, J. A. Bogovic, A. Carass, P.-L. Bazin, J. L. Prince, D. L. Pham, and B. A. Landman, “The java image science toolkit (jist) for rapid prototyping and publishing of neuroimaging software,” *Neuroinformatics*, vol. 8, no. 1, pp. 5–17, 2010.

[44] T. M. Lehmann, C. Gonner, and K. Spitzer, “Survey: Interpolation methods in medical image processing,” *Medical Imaging, IEEE Transactions on*, vol. 18, no. 11, pp. 1049–1075, 1999.

[45] K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny, *Statistical parametric mapping: The analysis of functional brain images: The analysis of functional brain images.* Academic Press, 2005.

[46] J. V. Hajnal and D. L. Hill, *Medical image registration.* CRC press, 2010.

[47] I. Dryden and K. Mardia, *Statistical shape analysis.* John Wiley & Sons, Chichester, 1998.

[48] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ants)," *Insight J*, 2009.

[49] L. G. Nyul and J. K. Udupa, "On standardizing the mr image intensity scale," *Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.

[50] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Evaluating intensity normalization on mris of human brain with multiple sclerosis," *Medical Image Analysis*, vol. 15, no. 2, pp. 267–282, 2011.

[51] N. Weisenfeld and S. Warfield, "Normalization of joint image-intensity statistics in mri using the kullback-leibler divergence," in *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on.* IEEE, 2004, pp. 101–104.

[52] R. Shinohara, E. Sweeney, J. Goldsmith, N. Shiee, F. Mateen, P. Calabresi, S. Jarso, D. Pham, D. Reich, and C. Crainiceanu,

"Normalization techniques for statistical inference from magnetic resonance imaging," in *UPenn Biostatistics Working Papers.* http://biostats.bepress.com/upennbiostat/art36, 2013, p. Working Paper 36.

[53] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[54] R. T. Shinohara, C. M. Crainiceanu, B. S. Caffo, M. I. Gaitán, and D. S. Reich, "Population-wide principal component-based quantification of blood-brain-barrier dynamics in multiple sclerosis." *NeuroImage*, 2011.

[55] E. M. Sweeney, R. T. Shinohara, C. D. Shea, D. S. Reich, and C. M. Crainiceanu, "Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal mri," *American Journal of Neuroradiology*, 2012.

[56] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression.* Cambridge University Press, 2003.

[57] C. E. Priebe, M. I. Miller, and J. Tilak Ratnanather, "Segmenting magnetic resonance images via hierarchical mixture modelling," *Computational statistics & data analysis*, vol. 50, no. 2, pp. 551–567, 2006.

[58] T. P. Beresford, D. B. Arciniegas, J. Alfers, L. Clapp, B. Martin, Y. Du, D. Liu, D. Shen, and C. Davatzikos, "Hippocampus volume loss due to chronic heavy drinking," *Alcoholism: Clinical and Experimental Research*, vol. 30, no. 11, pp. 1866–1870, 2006.

[59] D. L. Pham, C. Xu, and J. L. Prince, “Current methods in medical image segmentation 1,” *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[60] M. A. Balafar, A. R. Ramli, M. I. Saripan, and S. Mashohor, “Review of brain mri image segmentation methods,” *Artificial Intelligence Review*, vol. 33, no. 3, pp. 261–274, 2010.

[61] M. A. Sahraian, E.-W. Radue, and A. Gass, *MRI atlas of MS lesions.* Springer, 2008.

[62] J. Simon, D. Li, A. Traboulsee, P. Coyle, D. Arnold, F. Barkhof, J. Frank, R. Grossman, D. Paty, E. Radue *et al.*, “Standardized mr imaging protocol for multiple sclerosis: Consortium of ms centers consensus guidelines,” *American Journal of Neuroradiology*, vol. 27, no. 2, pp. 455–461, 2006.

[63] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, “Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging,” *Medical image analysis*, vol. 17, no. 1, pp. 1–18, 2013.

[64] X. Lladó, O. Ganiler, A. Oliver, R. Martí, J. Freixenet, L. Valls, J. C. Vilanova, L. Ramió-Torrentà, and À. Rovira, “Automated detection of multiple sclerosis lesions in serial brain mri,” *Neuroradiology*, vol. 54, no. 8, pp. 787–807, 2012.

[65] X. ó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira, “Segmentation of multiple

sclerosis lesions in brain mri: a review of automated approaches," *Information Sciences*, vol. 186, no. 1, pp. 164–185, 2012.

[66] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions," *NeuroImage*, vol. 49, no. 2, pp. 1524–1535, 2010.

[67] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield, "3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation," *MIDAS Journal*, vol. 2008, pp. 1–6, 2008.

[68] D. Hand, "Classifier technology and the illusion of progress," *Statistical Technology*, vol. 21, no. 1, pp. 1–15, 2006.

[69] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 8, pp. 677–688, 2001.

[70] M. P. Sormani, L. Bonzano, L. Roccatagliata, G. R. Cutter, G. L. Mancardi, and P. Bruzzi, "Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: A meta-analytic approach," *Annals of neurology*, vol. 65, no. 3, pp. 268–275, 2009.

[71] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images,"

Computer methods and programs in biomedicine, vol. 104, no. 3, pp. e158–e177, 2011.

[72] D. Goldberg-Zimring, A. Achiron, S. Miron, M. Faibel, and H. Azhari, "Automated detection and characterization of multiple sclerosis lesions in brain mr images," *Magnetic resonance imaging*, vol. 16, no. 3, pp. 311–318, 1998.

[73] B. Alfano, A. Brunetti, M. Larobina, M. Quarantelli, E. Tedeschi, A. Ciarmiello, E. M. Covelli, and M. Salvatore, "Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis," *Journal of Magnetic Resonance Imaging*, vol. 12, no. 6, pp. 799–807, 2000.

[74] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in mr imaging," *NeuroImage*, vol. 21, no. 3, pp. 1037–1044, 2004.

[75] P. Anbeek, K. L. Vincken, G. S. Van Bochove, M. J. Van Osch, and J. van der Grond, "Probabilistic segmentation of brain tissue in mr imaging," *Neuroimage*, vol. 27, no. 4, pp. 795–804, 2005.

[76] P. Anbeek, K. L. Vincken, and M. A. Viergever, "Automated ms-lesion segmentation by k-nearest neighbor classification," *MIDAS J*, 2008.

[77] B. R. Sajja, S. Datta, R. He, M. Mehta, R. K. Gupta, J. S. Wolinsky, and P. A. Narayana, "Unified approach for multiple sclerosis lesion segmentation on brain mri," *Annals of biomedical engineering*, vol. 34, no. 1, pp. 142–151, 2006.

[78] S. Datta, B. R. Sajja, R. He, J. S. Wolinsky, R. K. Gupta, and P. A. Narayana, “Segmentation and quantification of black holes in multiple sclerosis,” *Neuroimage*, vol. 29, no. 2, pp. 467–474, 2006.

[79] M. Scully, V. Magnotta, C. Gasparovic, P. Pelligrimo, D. Feis, and H. Bockholt, “3d segmentation in the clinic: a grand challenge ii at miccai 2008–ms lesion segmentation,” in *MIDAS Journal-MICCAI 2008 Workshop*, 2008, pp. 1–9.

[80] J. Morra, Z. Tu, A. Toga, and P. Thompson, “Automatic segmentation of ms lesions using a contextual model for the miccai grand challenge,” *Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge*, pp. 1–7, 2008.

[81] N. Subbanna, M. Shah, S. Francis, S. Narayanan, D. Collins, D. Arnold, and T. Arbel, “Ms lesion segmentation using markov random fields,” in *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, London, UK*, 2009.

[82] J. Lecoeur, J.-C. Ferré, and C. Barillot, “Optimized supervised segmentation of ms lesions from multispectral mris,” in *MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)*, 2009.

[83] E. T. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M. J. Brammer, “Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images

of the brain," *Medical Imaging, IEEE Transactions on*, vol. 18, no. 1, pp. 32–42, 1999.

[84] T. A. Dinh, T. Silander, C. T. Lim, and T.-Y. Leong, "An automated pathological class level annotation system for volumetric brain images," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1201.

[85] C.-H. Lee, M. Schmidt, A. Murtha, A. Bistritz, J. Sander, and R. Greiner, "Segmenting brain tumors with conditional random fields and support vector machines," in *Computer vision for biomedical image applications*. Springer, 2005, pp. 469–478.

[86] Z. Karimaghaloo, M. Shah, S. J. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain mri using conditional random fields," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 6, pp. 1181–1194, 2012.

[87] R. Shinohara, J. Goldsmith, F. Mateen, C. Crainiceanu, and D. Reich, "Predicting breakdown of the blood-brain barrier in multiple sclerosis without contrast agents," *American Journal of Neuroradiology*, vol. 33, no. 8, pp. 1586–1590, 2012.

[88] A. Carass, J. Cuzzocreo, M. B. Wheeler, P.-L. Bazin, S. M. Resnick, and J. L. Prince, "Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis," *NeuroImage*, vol. 56, no. 4, pp. 1982–1992, 2011.

[89] C. Bordier, M. Dojat, and P. L. De Micheaux, "Analyzefmri: an r package to perform statistical analysis on fmri datasets," in *useR! 2009-The R User Conference*, 2009, p. 25.

[90] T. Lumley, "biglm: bounded memory linear and generalized linear models. r package version 0.7," 2009.

[91] D. Adler, C. Glaeser, O. Nenadic, J. OehlschlŠgel, and W. Zucchini, "ff: memoryefficient storage of large data on disk and fast access functions. r package version 2.2–2," 2011.

[92] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, T. Sing, and O. Sander, "Visualizing the performance of scoring classifiers," *Package ROCR Version 1.0*, vol. 4, 2009.

[93] R. T. Shinohara, C. M. Crainiceanu, B. S. Caffo, M. I. Gaitán, and D. S. Reich, "Population-wide principal component-based quantification of blood–brain-barrier dynamics in multiple sclerosis," *NeuroImage*, vol. 57, no. 4, pp. 1430–1446, 2011.

[94] N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology 1," *Radiology*, vol. 229, no. 1, pp. 3–8, 2003.

[95] N. Shiee, P.-L. Bazin, J. L. Cuzzocreo, D. S. Reich, P. A. Calabresi, and D. L. Pham, "Topologically constrained segmentation of brain images with multiple sclerosis lesions," 2008.

[96] N. Shiee, P. Bazin, and D. Pham, "Multiple sclerosis lesion segmentation using statistical and topological atlases," *Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge*, pp. 1–10, 2008.

[97] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[98] M. S. Pepe, *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press, 2003.

[99] L. Axel, J. Costantini, and J. Listerud, "Intensity correction in surface-coil mr imaging," *American Journal of Roentgenology*, vol. 148, no. 2, pp. 418–420, 1987.

[100] E. Mjolsness and D. DeCoste, "Machine learning for science: state of the art and future prospects," *Science*, vol. 293, no. 5537, pp. 2051–2055, 2001.

[101] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[102] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez *et al.*, "Machine learning in bioinformatics," *Briefings in bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[103] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine learning*, vol. 36, no. 1-2, pp. 105–139, 1999.

[104] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 5, pp. 5–16, 2006.

[105] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning.* ACM, 2008, pp. 96–103.

[106] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and L. D. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical Image Analysis*, 2012.

[107] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical Science*, vol. 21, pp. 1–14, 2006.

[108] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[109] M. LeBlanc and R. Tibshirani, "Combining estimates in regression and classification," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1641–1650, 1996.

[110] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, pp. 1–21, 2007.

[111] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[112] C. Bordier, M. Dojat, and P. L. de Micheaux, "Temporal and Spatial Independent Component Analysis for fMRI Data Sets Embedded in the AnalyzeFMRI R Package," *Journal of Statistical Software*, vol. 44, no. 9, pp. 1–24, 2011. [Online]. Available: http://www.jstatsoft.org/v44/i09/

[113] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, *ROCR: Visualizing the performance of scoring classifiers.*, 2009, r package version 1.0-4. [Online]. Available: http://CRAN.R-project.org/package=ROCR

[114] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, 2012.

[115] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[116] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012, r package version 1.6-1. [Online]. Available: http://CRAN.R-project.org/package=e1071

[117] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[118] E. Polley and M. van der Laan, *SuperLearner: Super Learner Prediction*, 2012, r package version 2.0-9. [Online]. Available: http://CRAN.R-project.org/package=SuperLearner

[119] S. Walter, "The partial area under the summary roc curve," *Statistics in medicine*, vol. 24, no. 13, pp. 2025–2040, 2005.

[120] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[121] B. Gaonkar and C. Davatzikos, "Deriving statistical significance maps for svm based image classification and group comparisons," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. Springer, 2012, pp. 723–730.

[122] G. Trunk, "A problem of dimensionality: A simple example," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 1, no. 3, pp. 306–307, 1979.

[123] J. C. Russ, *The image processing handbook*. CRC press, 2006.

[124] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*. Citeseer, 2000, pp. 839–846.

[125] M. T. Hagan and M. B. Menhaj, “Training feedforward networks with the marquardt algorithm,” *Neural Networks, IEEE Transactions on*, vol. 5, no. 6, pp. 989–993, 1994.

[126] E. D. Karnin, “A simple procedure for pruning back-propagation trained neural networks,” *Neural Networks, IEEE Transactions on*, vol. 1, no. 2, pp. 239–242, 1990.

[127] V. Garcia, E. Debreuve, and M. Barlaud, “Fast k nearest neighbor search using gpu,” in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference on.* IEEE, 2008, pp. 1–6.

[128] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[129] H. He and E. A. Garcia, “Learning from imbalanced data,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009.

[130] N. Japkowicz *et al.*, “Learning from imbalanced data sets: a comparison of various strategies,” in *AAAI workshop on learning from imbalanced data sets*, vol. 68. Menlo Park, CA, 2000.

[131] L. G. Nyu and J. K. Udupa, “On standardizing the MR image intensity scale,” *Image*, vol. 1081, 1999.

[132] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, “Automatic segmentation of different-sized white matter lesions by voxel probability estimation,” *Medical Image Analysis*, vol. 8, no. 3, pp. 205–215, 2004.

[133] A. Cerasa, E. Bilotta, A. Augimeri, A. Cherubini, P. Pantano, G. Zito, P. Lanza, P. Valentino, M. C. Gioia, and A. Quattrone, “A cellular neural network methodology for the automated segmentation of multiple sclerosis lesions,” *Journal of Neuroscience Methods*, vol. 203, no. 1, pp. 193–199, 2012.

[134] B. Johnston, M. S. Atkins, B. Mackiewich, and M. Anderson, “Segmentation of multiple sclerosis lesions in intensity corrected multispectral mri,” *Medical Imaging, IEEE Transactions on*, vol. 15, no. 2, pp. 154–169, 1996.

[135] D.-J. Kroon, v. E. Oort, and C. Slump, “Multiple sclerosis detection in multispectral magnetic resonance images with principal components analysis,” 2008.

[136] M. Wels, M. Huber, and J. Hornegger, “Fully automated segmentation of multiple sclerosis lesions in multispectral MRI,” *Pattern Recognition and Image Analysis*, vol. 18, no. 2, pp. 347–350, 2008.

[137] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, “Spatial decision forests for MS lesion segmentation in multichannel magnetic resonance images,” *NeuroImage*, vol. 57, no. 2, pp. 378–390, 2011.

[138] R. Harmouche, L. Collins, D. Arnold, S. Francis, and T. Arbel, “Bayesian MS lesion classification modeling regional and local spatial information,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 984–987.

[139] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, and A. C. Evans, “Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images,” *Medical Imaging, IEEE Transactions on*, vol. 14, no. 3, pp. 442–453, 1995.

[140] A. Hadjiprocopis and P. Tofts, “An automatic lesion segmentation method for fast spin echo magnetic resonance images using an ensemble of neural networks,” in *Neural Networks for Signal Processing, 2003. NNSP’03. 2003 IEEE 13th Workshop on*. IEEE, 2003, pp. 709–718.

[141] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, “Morphometric analysis of white matter lesions in MR images: method and validation,” *Medical Imaging, IEEE Transactions on*, vol. 13, no. 4, pp. 716–724, 1994.

[142] A. P. Zijdenbos, R. Forghani, and A. C. Evans, “Automatic” pipeline” analysis of 3-d mri data for clinical trials: application to multiple sclerosis,” *Medical Imaging, IEEE Transactions on*, vol. 21, no. 10, pp. 1280–1291, 2002.

[143] S. Vinitski, C. F. Gonzalez, R. Knobler, D. Andrews, T. Iwanaga, and

M. Curtis, “Fast tissue segmentation based on a 4D feature map in characterization of intracranial lesions,” *Journal of Magnetic Resonance Imaging*, vol. 9, no. 6, pp. 768–776, 1999.

[144] D. S. Meier and C. R. Guttmann, “Time-series analysis of mri intensity patterns in multiple sclerosis,” *NeuroImage*, vol. 20, no. 2, pp. 1193–1209, 2003.

[145] ——, “Mri time series modeling of ms lesion development,” *Neuroimage*, vol. 32, no. 2, pp. 531–537, 2006.

[146] D. S. Meier, H. L. Weiner, and C. R. Guttmann, “Time-series modeling of multiple sclerosis disease activity: a promising window on disease progression and repair potential?” *Neurotherapeutics*, vol. 4, no. 3, pp. 485–498, 2007.

[147] R. Ghassemi, R. Brown, B. Banwell, S. Narayanan, D. L. Arnold *et al.*, “Quantitative measurement of tissue damage and recovery within new t2w lesions in pediatric-and adult-onset multiple sclerosis,” *Multiple Sclerosis Journal*, p. 1352458514551594, 2014.

[148] D. S. Reich, R. White, I. C. Cortese, L. Vuolo, C. D. Shea, T. L. Collins, and J. Petkau, “Sample-size calculations for short-term proof-of-concept studies of tissue protection and repair in multiple sclerosis lesions via conventional clinical imaging,” *Multiple Sclerosis Journal*, p. 1352458515569098, 2015.

[149] H. McFarland, F. Barkhof, J. Antel, and D. Miller, “The role of mri as a surrogate outcome measure in multiple sclerosis,” *Multiple Sclerosis*, vol. 8, no. 1, pp. 40–51, 2002.

[150] J. Fan and J.-T. Zhang, “Two-step estimation of functional linear models with applications to longitudinal data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 2, pp. 303–322, 2000.

[151] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[152] B. Whitcher, V. J. Schmid, and A. Thornton, “Working with the DICOM and NIfTI data standards in R,” *Journal of Statistical Software*, vol. 44, no. 6, pp. 1–28, 2011. [Online]. Available: http://www.jstatsoft.org/v44/i06/

[153] V. Fonov, A. Evans, R. McKinstry, C. Almli, and D. Collins, “Unbiased nonlinear average age-appropriate brain templates from birth to adulthood,” *NeuroImage*, vol. 47, p. S102, 2009.

[154] A. Carass, M. B. Wheeler, J. Cuzzocreo, P.-L. Bazin, S. S. Bassett, and J. L. Prince, “A joint registration and segmentation approach to skull stripping,” in *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on.* IEEE, 2007, pp. 656–659.

[155] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, “Statistical normalization techniques for magnetic resonance imaging,” *NeuroImage: Clinical*, 2014.

[156] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap.* CRC press, 1994.

[157] C. E. McCulloch and J. M. Neuhaus, *Generalized linear mixed models.* Wiley Online Library, 2001.

[158] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, “Random effects structure for confirmatory hypothesis testing: Keep it maximal,” *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.

[159] C. Crainiceanu, P. Reiss, J. Goldsmith, L. Huang, L. Huo, and F. Scheipl, *refund: Regression with Functional Data*, 2014, r package version 0.1-11. [Online]. Available: http://CRAN.R-project.org/package=refund

[160] J. Van Waesberghe, W. Kamphorst, C. J. De Groot, M. A. Van Walderveen, J. A. Castelijns, R. Ravid, G. Lycklama a Nijeholt, P. Van Der Valk, C. H. Polman, A. J. Thompson *et al.*, “Axonal loss in multiple sclerosis lesions: magnetic resonance imaging insights into substrates of disability,” *Annals of neurology*, vol. 46, no. 5, pp. 747–754, 1999.

[161] M. Filippi, G. Iannucci, M. Cercignani, M. A. Rocca, A. Pratesi, and G. Comi, “A quantitative study of water diffusion in multiple sclerosis

lesions and normal-appearing white matter using echo-planar imaging," *Archives of neurology*, vol. 57, no. 7, pp. 1017–1021, 2000.

[162] M. Filippi, M. Cercignani, M. Inglese, M. Horsfield, and G. Comi, "Diffusion tensor magnetic resonance imaging in multiple sclerosis," *Neurology*, vol. 56, no. 3, pp. 304–311, 2001.

[163] T. Hastie, R. Tibshirani, and J. Friedman, *Linear Methods for Regression.* Springer, 2009.

Curriculum Vitae
Elizabeth Sweeney

The Johns Hopkins University
Bloomberg School of Public Health
Department of Biostatistics
615 North Wolfe Street
Baltimore MD 21205

Email: emsweene@jhsph.edu

Education

Ph.D. Biostatistics, The Johns Hopkins Bloomberg School of Public Health
Advisor: Dr. Ciprian Crainiceanu, Co-Advisor: Dr. Russell Shinohara,
expected May 2016

Sc.M. Biostatistics, The Johns Hopkins Bloomberg School of Public Health,
Advisor: Dr. Ciprian Crainiceanu, 2012

B.S. Mathematics, with honor of distinction, Purdue University, Indianapolis, 2010

Peer Reviewed Publications

Under Revision

Pomann, G. M., Staicu, A. M., Lobaton, E., Mejia A.F., Dewey B. E., Reich, D. S., **Sweeney, E. M.**, and Shinohara, R. T. A lag functional linear model for prediction of magnetization transfer ration in multiple sclerosis lesions. *Under Revision at Annals of Applied Statistics.*

In Press

Sweeney, E. M., Crainiceanu, C. M., and Gertheiss, J. Testing Differentially Expressed Genes in Dose-Response Studies and with Ordinal Phenotypes. *Statistical Applications in Genetics and Molecular Biology*

Published

Fortin, J. P., **Sweeney, E. M.**, Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., and the Alzheimer's Disease Neuroimaging Initiative. *Removing inter-subject technical variability in magnetic resonance imaging studies.* NeuroImage.

Mejia A.F., **Sweeney E. M.**, Dewey B. E., Nair G., Sati P., Shea C. D., Reich D. S. and Shinohara R. T. Statistical estimation of T1 relaxation time using conventional magnetic resonance imaging. *NeuroImage.*

George, I. C., Sati, P., Absinta, M., Cortese, I. C., **Sweeney, E. M.**, Shea C. D., and Reich, D. S (2016). Clinical 3-tesla FLAIR* MRI improves diagnostic accuracy in multiple sclerosis. *Multiple Sclerosis Journal.*

Sweeney, E. M., Shinohara, R. T., Dewey, B. E., Schindler, M. K., Muschelli, J., Reich, D. S., Crainiceanu, C. M., and Eloyan, A. (2015) Relating multi-sequence longitudinal intensity profiles and clinical covariates in new multiple sclerosis lesions. *NeurouImage:Clinical.*

Roy, S., He, Q., **Sweeney, E. M.**, Carass, A., Reich, D. S, Prince, J. L., and Pham, D. L. (2015). Subject specific sparse dictionary learning for atlas based brain MRI segmentation. *IEEE Journal of Biomedical and Health Informatics.*

Muschelli, J., Ullman, N. L., **Sweeney, E. M.**, Eloyan, A., Martin, N., Vespa, P., Awad, I., Hanley, D. F., and Crainiceanu, C. M. (2015). Quantitative Intracerebral Hemorrhage Localization. *Stroke.*

Muschelli, J., **Sweeney, E. M.**, Lindquist, M. A., and Crainiceanu, C. M. (2015). fslr: Connecting the FSL Software with R. *The R Journal* .

Pomann, G. M., **Sweeney, E. M.**, Reich, D. S., Staicu, A. M., and Shinohara, R. T. (2015). Scan-stratified case-control sampling for modeling blood-brain barrier integrity in multiple sclerosis. *Statistics in Medicine.*

Eloyan, A., Shou, H., Shinohara, R. T., **Sweeney, E. M.**, Nebel, M. B., Cuzzocreo, J. L., Reich, D. S., Lindquist, M. A., and Crainiceanu, C. M. (2014) Health effects of lesion localization in multiple sclerosis: Spatial registration and confounding adjustment. *PLoS ONE.*

Shinohara, R. T., **Sweeney, E. M.**, Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi P. A., Jarso, S., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2014). Statistical normalization technique for magnetic resonance imaging. *NeurouImage:Clinical.*

Sweeney, E. M., Thakur, K. T, Lyons, J. T., Smith, B. R., Wiley, J. Z., Cervantes-Arslanian, A. M., Schwamm, L. H., Elkind, M. S. V., Shinohara, R. T., Mateen, F. J. (2014). IV-TPA for acute ischemic stroke in HIV-infected adults. *European Journal of Neurology.*

Muschelli, J., **Sweeney, E. M.**, and Crainiceanu, C. M. (2014). BrainR: Interactive 3 and 4d images of high resolution neuroimage data. *The R Journal* .

Sweeney, E. M., Vogelstein, J. T., Cuzzocreo, J. L., Calabresi P. A., Reich, D. S., Crainiceanu, C. M., and Shinohara, R. T. (2014). A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI. *PLoS ONE* 9(4): e95753.

Sweeney, E. M., Shinohara, R. T., Shiee, N., Mateen, F. J., Chudgar, A. A., Cuzzocreo, J. L., Calabresi P. A., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2013). OASIS is Automated Statistical Inference for Segmentation with applications to multiple sclerosis lesion segmentation in MRI. *NeuroImage Clinical,* 2, 402-413.

Sweeney, E. M., Shinohara, R. T., Shea, C. D., Reich, D. S., and Crainiceanu, C. M. (2013). Automatic lesion incidence estimation and detection using multi-modality longitudinal MRIs. *American Journal of Neuroradiology* 34(1), 68-73.

Book Chapters

In Press

Sweeney, E. M., Eloyan, A., Shinohara, R. T., and Crainiceanu, C. M. The Statistical Analysis of Structural MRI Neuroimaging Data. Handbook of Modern Statistical Methods: Neuroimaging Data Analysis. CRC Press, 2016.

Professional Experience

College of Engineering,
North Carolina State University,
Raleigh, North Carolina

2016 - Present	Associated Researcher with Courtesy Appointment

Department of Public Health,
American University of Armenia,
Yerevan

November 2015	Visiting Faculty in the School of Public Health

Center on Aging and Health,
Johns Hopkins,
Baltimore, Maryland

2014 - Present	Training Grant on the Epidemiology and Biostatistics of Aging (T32AG021334) Mentors: Dr. Michelle Carlson and Dr. Karen Bandeen-Roche

Department of Biostatistics and Epidemiology,
University of Pennsylvania,
Philadelphia, Pennsylvania

2012 - Present	Associated Researcher with Courtesy Appointment

Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health,
Baltimore, Maryland

2012 - 2013	Biostatistics Researcher working with Dr. Ciprian M. Crainiceanu

National Institute of Neurological Disorders and Stroke (NINDS),
Bethesda, Maryland

2013 - 2015	Special Volunteer, Neuroimmunology Branch
2011 - 2013	Technical Intramural Research Training Award Trainee, Neuroimmunology Branch
Summer 2011	Summer Internship Program, Neuroimmunology Branch

Invited Presentations

2016	“Modeling spatial correlation within Multiple Sclerosis lesions on structural MRI to improve estimation of the relationship between treatment and brain MRI ” [Invited Session] Joint Statistical Meeting 2016, Chicago, IL.
2015	“Relating multi-sequence longitudinal data from MS lesions on structural MRI to clinical covariates and treatment” T3 seminar, Neuroimmunology Branch, NINDS, Bethesda, MD.

2015 “Relating Multi-Sequence Longitudinal Data from MS Lesions on Structural MRI to Clinical Covariates” Duke University, Biostatistics Core, Durham, NC.

2014 “Multiple Sclerosis Lesion Segmentation from Structural MRI” Ludwig Maximilian University of Munich, Department of Statistics, Munich, Germany.

2014 “Analysis of Longitudinal Structural MRI in Multiple Sclerosis” Indiana University School of Medicine, Department of Biostatistics, Indianapolis, IN.

2014 “SuBLIME and OASIS for Multiple Sclerosis Lesion Segmentation in Structural MRI” 2014 Eastern North America Region of the International Biometric Society Invited Poster Session, Baltimore, MD.

2013 “Health effects of lesion localization in Multiple Sclerosis: Spatial registration and confounding adjustment”. SAMSI Neuroimaging Data Analysis Workshop, Registration Working Group, Research Triangle Park, NC.

2013 “OASIS is Automated Statistical Inference for Segmentation with applications to multiple sclerosis lesion segmentation in MRI”. University of California, Berkeley Biostatistics Department, Berkeley, CA.

Contributed Oral Presentations

2015 “Relating Multi-Sequence Longitudinal Data from MS Lesions on Structural MRI to Clinical Covariates”. [Topics Contributed Session] Joint Statistical Meeting 2015, Seattle, WA.

2014 “Analysis of Multi-Sequence Time Series Data from MS Lesions on Structural MRI”. International Conference on Advances in Interdisciplinary Statistics and Combinatorics 2014, Greensboro, NC.

2014 “Analysis of Multi-Sequence Time Series Data from MS Lesions on Structural MRI”. Joint Statistical Meeting 2014, Boston, MA.

2013 “OASIS is Automated Statistical Inference for Segmentation with applications to multiple sclerosis lesion segmentation in MRI”. Joint Statistical Meeting 2013, Montreal, Canada.

2013 “Do not use a cannon to kill a mosquito: a comparison of supervised classification algorithms in the context of MS lesion segmentation in structural MRI”. 28th International Workshop on Statistical Modeling, Palermo, Italy.

2012 “Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal magnetic resonance images”. Functional Data Analysis Workshop, Bristol, United Kingdom.

Extended Department Research Visits

University of Göettingen

Department of Animal Sciences
Biometrics & Bioinformatics
Göettingen, Germany
working with Dr. Jan Gertheis
July 2014

Ludwig Maximilian University of Munich

Department of Statistics
Munich, Germany
working with Dr. Sonja Greven
June 2014

University of Pennsylvania

Department of Biostatistics and Epidemiology
Philadelphia, PA
working with Dr. Russell T. Shinohara
January 2013, 2014, and 2015, July 2014, September - October 2015

Software

R packages

oasis	*R package implementation of OASIS is Automated Statistical Inference for Segmentation (OASIS) segmentation of Multiple Sclerosis lesions in the brain on structural MRI, hosted on R CRAN.*

Teaching

American University of Armenia Yervan, Armenia

2015	Instructor for Public Health 321: Inferential Biostatistics, 17 students, (Student Evaluations: Course: 4.7/5, Instructor: 4.8/5) *Sole instructor for graduate–level course in introductory biostatistics for the masters of public health program. Responsible for preparing and teaching daily lectures, lab sessions, and STATA lab sessions.*

Coursera

2015 Instructor, Coursera, Neurohacking with R
Co-developed a massive open online course (MOOC) for Coursera on neuroimage data, pre-processing and statistical analysis performed completely within the statistical software R. Developed half of the code and slides for the course and recorded lectures delivering slides.

Tutorials

2015 Crainiceanu, C. M., **Sweeney, E. M.**, and Muschelli, J. "A Tutorial for Multisequence Clinical Structural Brain MRI ". 2015 Eastern North America Region of the International Biometric Society Meeting, Miami, FL.

2013 **Sweeney, E. M.**"MIPAV + R for analysis of structural MRI data ". SAMSI Neuroimaging Data Analysis Workshop, Multisequence Structural MRI Working Group, Research Triangle Park, NC.

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Baltimore, MD

Teaching assistant responsibilities included grading student homework and exams, answering student emails, and holding weekly office hours unless noted otherwise.

2016 Lead Teaching Assistant for Statistical Methods in Public Health III, 500 students, (Prof: Dr. Marie Diener-West), *Responsible for teaching a weekly lab session*

2015 Teaching Assistant for Biostatistics in Medical Product Regulation, Online Course, 20 students (Prof: Dr. Simon Day and Dr. Mary Foulkes)

2015 Teaching Assistant for Statistical Methods in Public Health IV, 300 students, (Prof: Dr. James Tonascia)

2015 Teaching Assistant for Statistics for Laboratory Scientists I, 50 students, (Prof Dr. Ingo Ruczinski)

2014 Teaching Assistant for Methods in Biostatistics I & II, 50 students, (Prof: Dr. Ciprian Crainiceanu), *Responsible for writing and teaching a weekly lab session.*

2012 Teaching Assistant for Statistical Methods in Public Health III & IV, 500 students, (Prof: Dr. Marie Diener-West)

2011 Teaching Assistant for Statistical Reasoning in Public Health I & II, 100 students, (Prof: Dr. John McGready)

Peer Review

Biomedical Signal Processing and Control, Biometrics, Computer Methods and Programs in Biomedicine, International Journal for Quality in Health Care, Journal of Computational and Graphical Statistics, Journal of the Neurological Sciences, Multiple Sclerosis Journal, NeuroImage, NeuroImage:Clinical, Neuroradiology

Honors and Awards

2016	Jane and Steve Dykacz Award, Johns Hopkins Biostatistics Department *Best student paper in medical statistics for "Relating multi-sequence longitudinal intensity profiles and clinical covariates in new multiple sclerosis lesions"*
2015	American Statistical Association Gertrude M. Cox Scholarship *Award to encourage women to enter statistically oriented professions*
2014	Young Researcher Travel Award, International Conference on Advances in Interdisciplinary Statistics and Combinatorics
2014	Jane and Steve Dykacz Award, Johns Hopkins Biostatistics Department *Best student paper in medical statistics for "OASIS Is Automated Statistical Inference for Segmentation with Applications to Multiple Sclerosis Lesion Segmentation in MRI"*
2012	Johns Hopkins Imaging Conference Poster Session Peer Choice Award
2011	Exceptional Summer Student, NINDS
2008, 2009	Yuri Abramovich Memorial Mathematics Scholarship, IUPUI Mathematics Department
2007, 2008, 2009	Anna K Suter Mathematics Scholarship, IUPUI Mathematics Department

Academic Service

2016 Invited Session Organizer, Statisticians and Multiple Sclerosis Research, Joint Statistical Meeting 2016, Chicago, IL.

2015 Peer Facilitator, Johns Hopkins University Data Science Hackathon, Baltimore, MD.

2015 - 2016 Gerontology Interest Group Co-Leader, Johns Hopkins School of Public Health, Baltimore, MD.

2015 Session Chair, 2015 Eastern North America Region of the International Biometric Society Meeting, Miami, FL.

2014 - 2105 Student Journal Club Co-Organizer with John Muschelli, Biostatistics Department, Johns Hopkins School of Public Health, Baltimore, MD.

2014 Volunteer, Eastern North America Region of the International Biometric Society Meeting 2014, Baltimore, MD.

2013 Session Chair, Joint Statistical Meeting 2013, Montreal, Canada.

Computing Proficiency

Statistical Software: R, STATA, Matlab [working knowledge]

Programming: Shell scripting

Neuroimaging Software: MIPAV, FSL, SPM

Document Preparation: LaTeX, MS Office

Other: Adobe Illustrator and Photoshop

Last updated: March 30, 2016