2020

Statistical Methods For Multi-Modal Image Analysis With Applications In Multiple Sclerosis And Neurodevelopment

Alessandra Michelle Valcarcel
University of Pennsylvania

Statistical Methods For Multi-Modal Image Analysis With Applications In Multiple Sclerosis And Neurodevelopment

Abstract

Multi-modal neuroimaging, where several high-dimensional imaging variables are collected, has enabled the visualization and analysis of the brain structure and function in unprecedented detail. Due to methodological and computational challenges, the vast number of imaging studies evaluate data from each modality separately and do not consider information encoded in the relationships between imaging types. In this work, we propose methods that quantify the complex relationships between multiple imaging modalities and map how these relationships vary spatially across different anatomical regions of the brain. In order to understand relationships between several high-dimensional imaging variables, we use novel multi-modal image analysis techniques for feature development and image fusion in conjunction with machine learning techniques to develop automatic approaches for multiple sclerosis lesion detection. Additionally, we use multi-modal image analysis to understand the association between high-dimensional imaging variables with phenotypes of interest to investigate structure-function relationships in development, aging, and pathology of the brain. We find that by leveraging the relationship between imaging modalities, we can more accurately detect neuropathology and delineate brain trajectories to provide complementary characterizations of healthy development. We provide publicly available R packages to allow easy access and implemention of the proposed methods in new data and contexts.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Russell T. Shinohara

Second Advisor

Kristin A. Linn

Subject Categories

Statistics and Probability

STATISTICAL METHODS FOR MULTI-MODAL IMAGE ANALYSIS WITH APPLICATIONS IN MULTIPLE SCLEROSIS AND NEURODEVELOPMENT

Alessandra M. Valcarcel

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation	Co-Supervisor of Dissertation
Russell T. Shinohara	Kristin A. Linn
Associate Professor of Biostatistics	Assistant Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Nandita Mitra, Professor of Biostatistics

Haochang Shou, Assistant Professor of Biostatistics

Amit Bar-Or, Professor of Neurology

STATISTICAL METHODS FOR MULTI-MODAL IMAGE ANALYSIS WITH APPLICATIONS IN MULTIPLE SCLEROSIS AND NEURODEVELOPMENT

ACKNOWLEDGEMENT

First and foremost, I want to thank my family, whose unwavering support and faith in me made this work possible. You all celebrated my successes but more importantly picked me up after each time I fell. To my Mom, you taught me to think critically, optimize everything (even how you shop for groceries), and to feistily pursue my dreams. From my Dad, I have learned to be humble, patient, and to always stay grinding. I would be lost (and most likely naked) without my siblings Titi, Mike, and Andrew. You all were my personal shoppers, stylists, taxi drivers, concert buddies, therapists, and general partners in crime throughout this process. Thank you for for helping me however you could and making sure I showered and ate. To my Grandparents, thank you for giving me a foundation from which to stand on, for always sharing your wisdom, and teaching me knowledge is power. I realize now that I took that advice perhaps too literally but am eternally grateful. To my extended family, thank you for your hospitality and letting me lean on your strength. Thank you to my nieces and nephews for always wanting to play and being my favorite distraction. Lastly, thank you Lily and Hunter for your unconditional love and furry snuggles.

I would also like to extend my deepest gratitude to my friends for always listening to me and, more importantly, for forcing me to shut up and drink a beer. This journey would have been dull without you all. My Hopewell gals, GOAT squad, and Philly crew thank you for the good times even amid personal bad times.

To my advisors, Kristin and Taki, words cannot describe the impact you both have had on me and how truly incredible you both are. It has been an honor to learn from you both over the past five years. You are two of the most intelligent, generous, patient, passionate, and inspiring people I've ever met, and I will always look up to you both. More than your mentorship, I am grateful for your friendship. Our KAT (Kristin, Ali, Taki) meetings always started with at least 10 minutes of just chatting about life, discovering and debating delicious restaurants in Philly, or planning our next adventures. I cherish these small moments and memories and will miss them in my weekly routine more than you know. Thank you both for everything.

I also want to thank my committee chair Nandita Mitra for her invaluable leadership and mentorship throughout my graduate education and as it relates to this work. To my committee members, Amit Bar-Or and Haochang Shou, I would like to extend my sincerest appreciation for sharing your

expertise and valuable time with me.

Thank you to the past and present PennSIVE members for patiently teaching me, generously listening to my ideas, and the endless debate and discussion. Melissa Martin extended a great amount of assistance editing and reviewing my papers. I would also like to specifically thank Simon Vandekar for his collaboration and friendship. One day we will find the mushroom shirt.

I must thank the biostatistics faculty and other biomedical faculty members, from whom I have learned a great deal. I especially want to thank Sarah Ratcliffe, Jinbo Chen, Pam Shaw, Justine Shults, and Yong Chen who selflessly shared their time and were engaging and effective teachers. I without a doubt must thank the biostatistics staff, namely Marissa Fox, Cathy Vallejo, and Eli Elliott for all of their help behind the scenes. I also would like to thank the computing staff, specifically computing wizards Ryan Salomon and Curt Calafut for their personal attention and sometimes constant email correspondence whenever I broke the cluster.

I also want to thank past and present biostatistics graduate students for their unwavering support and strong sense of community. I'm deeply indebted to the students in my cohort: Carrie Caswell, Jordan Dworkin, Rui Duan, Jiaxin Fan, Jiarui Lu, Eric Oh, and Lingjiao Zhang. You all were my personal tutors at some point or another and I would not have made it through the program without each of your investment, kindness, and friendship. Jordan and Eric, thank you both especially for putting up with me as your work bae. Other students who kept my life balanced with good company, laughter, music, beer, and more: Erica Billig, Julie Kobie, Leah Suttner, Lior Rennert, Le Wang, Emily Xu, Carolyn Lou, Nick Illenberger, Danni Tu, Sarah Weinstein, Andrew Chen, and Arman Oganisian.

I would also like to extend my deepest gratitude to all of my amazing collaborators: Matt Schindler, Ted Satterthwaite, Carly O'Donnell, Dzung Pham, Ipek Oguz, John Muschelli, Ciprian Crainiceanu, Peter Calabresi, Rohit Bakshi, and Elizabeth Sweeney.

I hope others who I have not listed by name still know how much I appreciate their help.

Penn and Philly thank you for 5 amazing years which included a Super-Bowl win.

It takes a village and I have dope one.

This work was supported by the PennSIVE Center and the National Institutes of Health, NIH R01 MH112847, NIH R01 NS085211, NIH R21 NS093349, and NIH R01 NS060910. The content is solely the responsibility of the authors and does not necessarily represent theofficial views of the funding agency.

ABSTRACT

STATISTICAL METHODS FOR MULTI-MODAL IMAGE ANALYSIS WITH APPLICATIONS IN MULTIPLE SCLEROSIS AND NEURODEVELOPMENT

Alessandra M. Valcarcel

Russell T. Shinohara

Kristin A. Linn

Multi-modal neuroimaging, where several high-dimensional imaging variables are collected, has enabled the visualization and analysis of the brain structure and function in unprecedented detail. Due to methodological and computational challenges, the vast number of imaging studies evaluate data from each modality separately and do not consider information encoded in the relationships between imaging types. In this work, we propose methods that quantify the complex relationships between multiple imaging modalities and map how these relationships vary spatially across different anatomical regions of the brain. In order to understand relationships between several high-dimensional imaging variables, we use novel multi-modal image analysis techniques for feature development and image fusion in conjunction with machine learning techniques to develop automatic approaches for multiple sclerosis lesion detection. Additionally, we use multi-modal image analysis to understand the association between high-dimensional imaging variables with phenotypes of interest to investigate structure-function relationships in development, aging, and pathology of the brain. We find that by leveraging the relationship between imaging modalities, we can more accurately detect neuropathology and delineate brain trajectories to provide complementary characterizations of healthy development. We provide publicly available R packages to allow easy access and implemention of the proposed methods in new data and contexts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Over the last two decades, the development and use of neuroimaging technologies has grown exponentially (Liu et al., 2015a). The rise of neuroimaging research is attributed to numerous discoveries showing added clinical benefits and insights using imaging data, advancing image device technology, and unprecedented acceleration of computing power at a rapidly decreasing cost (Liu et al., 2015b). Multi-modal imaging data quantify different, yet complimentary, properties of the brain and its activity. When studied jointly, multi-modal imaging data may improve our understanding of the brain.

Common multi-modal data includes, but is not limited to, structural and functional magnetic resonance imaging (MRI), positron emission tomography (PET), electroencephalography (EEG), magnetoencephalography (MEG), and computed tomography (CT). Multi-modal data can be acquired from simulatenous imaging measurements or integration of separate measurements. For example, multiple measures from the same device using differing acquisition sequences or protocol such as T1-weighted (T1), T2-weighted (T2), Fluid Attenuated Inversion Recovery (FLAIR) images from MRI is often considered multi-modal. Multi-modal data may also be comprised of images collected from different imaging devices such as MRI and CT.

Multi-modal image analysis can be used for feature extraction, image fusion, machine learning, and visualization. Commonly, multi-modal analysis seeks to understand the coupled association among the multi-modal data and how these relate to covariates of interest. Many disease areas benefit from multi-modal analysis. Multi-modal neuroimaging has been used to study healthy brain development and aging (Satterthwaite et al., 2014b, 2016; Vandekar et al., 2016), Alzheimer's disease (Petersen et al., 2010), schizophrenia (Kochunov et al., 2014; Sui et al., 2011), epilepsy (Abela et al., 2014), obsessive-compulsive disorder (OCD) (Radua et al., 2014), bipolar disorder (Sui et al., 2011), attention-deficit hyperactivity disorder (ADHD) (Anderson et al., 2014), autism spectrum disorder (ASD) (Stigler et al., 2011), traumatic brain injury (TBI) (Cherubini et al., 2007), stroke (Copen, 2015), multiple sclerosis (Carass et al., 2017a,b; Valcarcel et al., 2018a,b), and brain tumors (Durst et al., 2014). In this work, we propose novel statistical approaches to multi-modal image analysis with applications in multiple sclerosis and neurodevelopment.

Total brain white matter lesion (WML) volume is the most widely established MRI outcome measure in studies of MS (Lublin et al., 2014). To estimate WML volume, there are a number of automatic segmentation methods available, yet manual delineation remains the gold standard approach. The most widely established MRI outcome measure is the volume of hyperintense lesions on T2-weighted images (T2L) (Bakshi et al., 2005, 2008; Zivadinov and Bakshi, 2004). Unfortunately, T2L are non-specific for the level of tissue destruction and show a weak relationship to clinical status (Molyneux et al., 2000). Interest in lesions that appear hypointense on T1-weighted images (T1L) ("black holes") has grown because T1L provide more specificity for axonal loss and a closer link to neurologic disability (Andermatt et al., 2017; Katdare and Ursekar, 2015). The technical difficulty of T1L segmentation has led investigators to rely on time-consuming manual assessments prone to inter- and intra-rater variability (Garcia-Lorenzo et al., 2013; Llado et al., 2012). In Chapter 2 of this dissertation, we develop MIMoSA, a Method for InterModal Segmentation Analysis. Although the majority of statistical techniques for the automated segmentation of WMLs are based on single imaging modalities, recent advances have used multi-modal techniques for identifying WMLs. Complementary modalities emphasize different tissue properties, which help identify interrelated features of lesions. MIMoSA utilizes novel covariance features from inter-modal coupling regression in addition to features that capture the mean image structure to model the probability a lesion is contained in each voxel. We demonstrate MIMoSA's flexibility and utility by accurately automatically segmenting T1L and T2L using data acquired at the Brigham and Women's Hospital (BWH).

Multi-modal automatic segmentation approaches often yield a probability map to which a threshold is applied to create lesion segmentation masks. Unfortunately, few approaches systematically determine the threshold employed; many methods use a manually selected threshold (Sweeney et al., 2013, 2014), thus introducing human error and bias into the automated procedure. In Chapter 3, we propose and validate an automatic thresholding algorithm, Thresholding Approach for Probability Map Automatic Segmentation in Multiple Sclerosis (TAPAS), to obtain subject-specific threshold estimates for segmenting T2L. Using multi-modal MRI, the proposed method applies an automatic segmentation algorithm to obtain probability maps. We obtain the subject-specific threshold that maximizes the Sorensen-Dice similarity coefficient (DSC). These subject-specific thresholds are modeled on a naive estimate of volume using a generalized additive model. Applying this model, we predict a subject-specific threshold in data not used for training. We ran a Monte Carlo-resampled split-sample cross-validation (100 validation sets) using two data sets: the first obtained from the Johns Hopkins Hospital (JHH) on a Philips 3 tesla (3T) scanner (n = 94) and a second collected at BWH using a Siemens 3T scanner (n = 40).

The amount of data produced by a single imaging modality in any given experiment can be quite large. Therefore, multi-modal imaging and analyses results in enormous amounts of data (Kang, Caffo, and Liu, 2016). For example, 7 tesla (7T) imaging produces high contrast images with exceptional resolution and detail and can yield over 30 million measurements in just a single imaging modality (Rutland et al., 2020). Imaging studies may collect multi-modal data from thousands of subjects over multiple years (Kang, Caffo, and Liu, 2016). Therefore, the size of data sets has become so large it cannot be loaded into memory for analysis, even on high-performance computing clusters. In Chapter 4, we develop software tools in R for memory efficient representations of large imaging data.

In Chapter 5, we propose Inter-Modal Coupling (IMCo) analysis. IMCo is a multi-modal data analysis method for studying the complex relationships between multiple imaging modalities and mapping how these relationships vary spatially across different anatomical regions of the brain. Given a particular voxel location in the brain, we regress an outcome image modality on the remaining modalities using all voxels in a local neighborhood of the target voxel. In this work we compare the performance of three estimation frameworks that account for the spatial dependence among voxels in a neighborhood: generalized estimating equations (GEE), linear mixed effects models with varying random effect structures, and weighted least squares. We run a large scale simulation to assess estimator accuracy and efficiency across a number of different generative models.

CHAPTER 2

A DUAL MODELING APPROACH TO AUTOMATIC SEGMENTATION OF CEREBRAL T2 HYPERINTENSITIES AND T1 BLACK HOLES IN MULTIPLE SCLEROSIS

2.1. Introduction

Multiple sclerosis (MS) is an inflammatory and demyelinating autoimmune disease of the central nervous system which typically leads to neurodegeneration (Ahlgren, Odn, and Lycke, 2011; Compston and Coles, 2002; Harbo, Gold, and Tintor, 2013). The inflammatory and demyelinating process causes multifocal lesions and widespread atrophy in white and gray matter, often leading to physical disability, cognitive dysfunction, and unemployment (Rovira and Len, 2008; Tauhid et al., 2015). Structural magnetic resonance imaging (MRI) is a commonly used tool for the diagnosis, longitudinal management, and scientific investigation of MS (Lublin et al., 2014) because it allows for the detection of white matter lesions (WML). Common MRI metrics used to assess disease activity and severity in patient management and clinical trials include WML count and volume, the latter of which particularly relies on accurate segmentation.

Several complementary characterizations of WML are commonly delineated. Gadolinium-enhancing lesions (EL) are closely linked to acute perivascular inflammatory activity due to focal break-down of the blood-brain barrier and typically fade over 26 weeks (Zivadinov and Bakshi, 2004). T2 hyperintense lesions (T2L), which typically start as EL but later remain as non-enhancing lesions, are nonspecific for the severity of underlying pathology (Zivadinov and Bakshi, 2004). That is, T2 sequences are nonspecific for the type and degree of tissue injury such as demyelination, inflammation, edema or axonal loss. This non-specificity is one factor that contributes to modest associations between T2L metrics and clinical status (Molyneux et al., 2000). Approximately 50% of T2L also appear as persistent T1 hypointensities (T1L), commonly referred to as black holes, which are likely to be the most destructive regions with severe demyelination and axonal loss (Andermatt et al., 2017; Katdare and Ursekar, 2015). Furthermore, the T1L/T2L ratio, an index of the destructive potential of lesions, has been shown to be particularly sensitive in tracking MS therapeutic response (Kim et al., 2016). T1L metrics provide high clinical significance but are usually assessed manually in both clinical and trial settings because they are difficult to segment (Bakshi et al., 2005).

Manual segmentation is the gold standard approach for WML quantification and requires an expert to analyze scans visually. Unfortunately, this process is costly, time-consuming, and prone to intra- and inter-rater variability (Garcia-Lorenzo et al., 2013; Llado et al., 2012; Sweeney et al., 2014). Difficulties associated with manual lesion segmentation have led to the development of various segmentation methods with different levels of accuracy and complexity (Sweeney et al., 2014). While many methods are available, no single approach has been shown to perform optimally across multiple lesion assessments and scanning platforms. This is largely due to the challenges associated with heterogeneous lesion characteristics within and across subjects and variability introduced by scanning hardware and acquisition protocols.

The majority of automatic lesion segmentation methods delineate T2L (Dadar et al., 2017; Garcia-Lorenzo et al., 2013; Meier et al., 2018; Shiee et al., 2010; Sweeney et al., 2013; Valcarcel et al., 2018b). In contrast, few studies have investigated a fully automatic segmentation approach for T1L. The sparsity of prior research is in part due to a technical challenge: since T1L and their boundaries appear similar to gray matter (Ceccarelli et al., 2012) and are subtler than the boundaries of T2L, they are much more difficult to segment by manual and automatic methods. Related to the segmentation of T1L, Khayati et al. proposed a method to segment different stages of lesions, including chronic lesions which include T1L as well as other lesional phenotypes (Khayati et al., 2008). The simplest method to segment T1L was proposed by Filippi et al. using an expert-driven semi-automated thresholding approach to estimate lesion volumes (Filippi et al., 1996). Molyneux et al. similarly proposed a semi-automated technique to delineate T1L in a multi-center study where they showed that T1L volume is a consistent and reproducible metric that can be applied to MRI data from various scanners (Molyneux et al., 2000). Following these results, Datta et al. recently developed fully automated methods using fuzzy connectivity modeling (Datta et al., 2006). Wu et al. proposed an algorithm to detect EL, T1L, and T2L using intensity-based statistical k-nearest neighbor classification combined with template-driven segmentation and partial volume artifact correction (Wu et al., 2006). To automatically segment T1L, Spies et al. proposed an approach that used a standard classification algorithm to partition T1-weighted images into gray matter, white matter, and cerebrospinal fluid and then found T1L in the white matter using voxel-wise testing with healthy controls as a reference (Spies et al., 2013). Harmouche et al. proposed a method to segment T1L and T2L jointly by modeling the posterior probability density function (Harmouche et al., 2015).

Unfortunately, none of these approaches provide publicly available software, and the studies were based on relatively small MRI datasets with uniform patient demographics and lesion load (Datta et al., 2006; Filippi et al., 1996; Molyneux et al., 2000; Spies et al., 2013; Wu et al., 2006). Additionally, studies to date have only used a single rater for manual segmentations. Likely due to these limitations, adoption of these previously published methods has been slow, and studies have continued to obtain T1L segmentations manually. A comprehensive, automated technique with readily available software that integrates aspects of WML burden of multiple lesion characterizations in a diverse patient population would thus address an important, unmet need in the radiological assessment of MS lesions.

In our previous work, a Method for Inter-Modal Segmentation Analysis (MIMoSA) was developed and validated as an automatic T2L segmentation method in people with MS (Valcarcel et al., 2018b). MIMoSA has readily available software for implementation in R as a package on Neuroconductor (`https://neuroconductor.org/package/details/mimosa`) with documentation and a vignette available on GitHub (`https://github.com/avalcarcel9/mimosa/blob/master/vignettes/mimosa_git.md`) ("Neuroconductor"; Valcarcel, 2018). In the present study, we applied the MIMoSA method to automatically segment T1L. Since no publicly available software for automatic detection of T1L exists, we automatically segmented T2L using MIMoSA and used these measures as a reference for T1L performance. This was motivated by our findings that MIMoSA is a competitive T2L segmentation approach (Valcarcel et al., 2018b), and all T1L are also seen as T2L (but not vice-versa). Moreover, since the data in this study were acquired under a different protocol than data in the original development of MIMoSA, application of MIMoSA to segment T2L enabled us to validate and assess the robustness of MIMoSA's accuracy across different scanner platforms and protocols. For further comparison, OASIS, another validated T2L lesion segmentation algorithm (Sweeney et al., 2013), was used to automatically segment T1L. Finally, we examined correlations between lesion volume and clinical status measurements in order to determine if automatic lesion segmentation reduced noise and revealed stronger associations with disability.

Here we propose an automatic approach to segmenting T1L with software that is publicly avail-

Table 2.1: Demographic information for subjects in this study are provided. Included were 40 subjects diagnosed with multiple sclerosis (MS) and scanned between 2015 and 2016 at the Brigham and Women's Hospital.

	Mean	Std. Dev.	Min	Max
Age (years)	50.4	9.9	30.4	69.9
Disease duration (years)	14.5	4.6	3.8	21.3
Expanded Disability Status Scale score	2.3	1.6	0.0	7.0
Timed 25-ft walk (seconds)	5.1	2.6	3.0	18.4
T1L manual volume (mL)	7.70	8.33	0.18	35.03
T2L manual volume (mL)	13.57	12.78	0.58	52.04

	%
Male	30
Female	70
Relapsing-remitting MS	80
Secondary progressive MS	20

able for implementation. The ability to segment T1L automatically and quickly has the potential to facilitate tracking of disease activity and lesional damage over time. Additionally, using the same automatic approach to determine T2L and T1L reduces variability in segmentation metrics by eliminating multiple data processing pipelines.

2.2. Materials and methods

2.2.1. Patients and study design

Data were collected at the Brigham and Women's Hospital in Boston, Massachusetts. The Institutional Review Board approved the study and transfer of data to the University of Pennsylvania. Forty patients, all with a clinical diagnosis of MS, were consecutively obtained from MRI scans at the center. Subjects had an examination by an MS specialist neurologist to assess the type of MS, the level of physical disability on the Expanded Disability Status Scale (EDSS), and ambulatory function on the timed 25-ft walk (T25FW). Patient demographics are provided in Table 2.1. Additionally, a scatterplot of lesion count against volume is displayed in Figure 2.1, which shows a wide range of lesion counts and volumes across subjects.

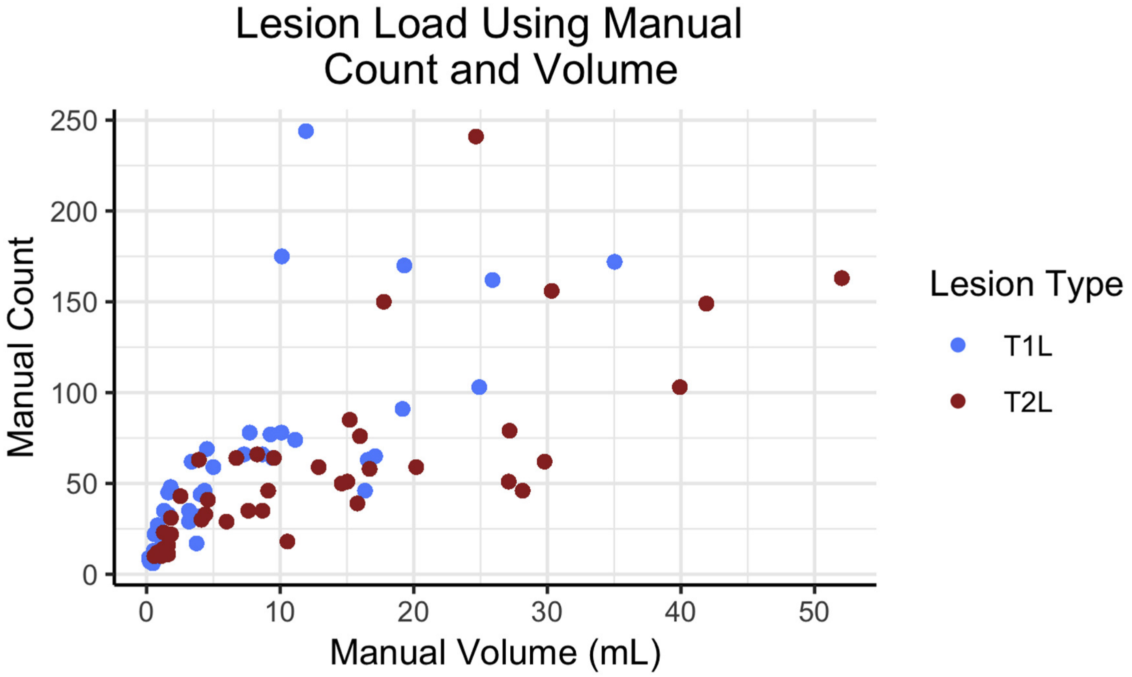


Figure 2.1: Lesion volume (mL) and count for each subject are presented using manual segmentation masks. The lesion number and volume across subjects are both diverse for T1 lesions (T1L) and T2 lesions (T2L).

2.2.2. Image acquisition and preprocessing

High-resolution 3D T1-weighted (T1WI), T2-weighted (T2WI), and fluid-attenuated inversion recovery (FLAIR) volumes of the brain were collected on a Siemens 3 tesla (3T) Skyra instrument using a consistent scan protocol for all subjects. Acquisition details are provided in Table 2.2 and have also been detailed previously (Meier et al., 2018).

2.2.3. Image analysis

All images were preprocessed prior to implementing the MIMoSA model using the R (R Development Core Team, 2018) packages *extrantsr* (*extrantsr: Extra Functions to Build on the ANTsR Package*) and *WhiteStripe* (R. Taki Shinohara and John Muschelli, 2017), as well as Multi-Atlas Skull-Stripping (MASS) (Doshi et al., 2013; *NITRC: CBICA: Multi Atlas Skull Stripping (MASS): Tool/Resource Info*). After N4 inhomogeneity correction Tustison et al., 2010, volumes were co-registered across sequences for each subject using a rigid-body transformation with a Lanczos windowed sinc interpolator. To remove extracerebral voxels, MASS was implemented (Doshi et al.,

Table 2.2: Image acquisition protocol using a 3 tesla (3T) Siemens Skyra scanner at the Brigham and Women's Hospital.

3T Brain MRI Acquisition Protocol			
Scanner Hardware	Siemens Skyra		
Scanner Software	Syngo MR D13		
Coil	20 channel		
MR Acquisition Type	3D		
Orientation	Sagittal		
Number of signal averages	1		
Sequence type	FLAIR	T2WI	T1WI
Number of slices	176	192	176
Voxel size (mm)	1.0×1.0×1.0	0.98×0.98×1.0	1.0×1.0×1.0
TR (ms)	5000	2500	2300
TE (ms)	389	300	2.96
TI (ms)	1800	N/A	900
Flip angle (degrees)	120	120	9
Parallel acceleration	2	4	2
Scan time (minutes)	6:00	3:18	5:09

2013; *NITRC: CBICA: Multi Atlas Skull Stripping (MASS): Tool/Resource Info*). Manually delineated T2L masks were obtained in the FLAIR space, and manual T1L masks were obtained in the T1WI space. To avoid interpolation errors in these masks, analyses of T1L and T2L were conducted in their respective native spaces and no transformations of the segmentation masks nor the primary imaging sequences were applied. First, T1WI and T2WI images were registered to the FLAIR for all T2L modeling; then, separately, T2WI and FLAIR images were registered to the T1WI space for all T1L modeling. As conventional MRI volumes are acquired in arbitrary units, statistical intensity normalization using WhiteStripe (R. Taki Shinohara and John Muschelli, 2017) was applied in order to facilitate modeling of intensities across subjects.

T1L and T2L were manually segmented by a reading panel of two trained observers under the supervision of an experienced observer at the Brigham and Women's Hospital. Each trained observer independently determined the presence or absence of T1L and T2L and then reviewed these results together to form a consensus. In the event of a disagreement, a senior experienced observer was consulted. A WML was categorized as a T2L if it appeared as hyperintense on the FLAIR. T1L, or black holes, were defined as appearing hypointense on T1WI and at least partially hyperintense on the FLAIR volumes. After a consensus of lesions was determined, one observer segmented all T1L and T2L using an edge-finding tool in Jim (v. 7.0) (*Jim* 2014). This process resulted in manually segmented gold standard masks for T1L and T2L for each subject in the study. Figure 2.2 shows examples of preprocessed images and manual T1L and T2L segmentations. All gadolinium-enhancing lesions were excluded from T1L manual segmentations.

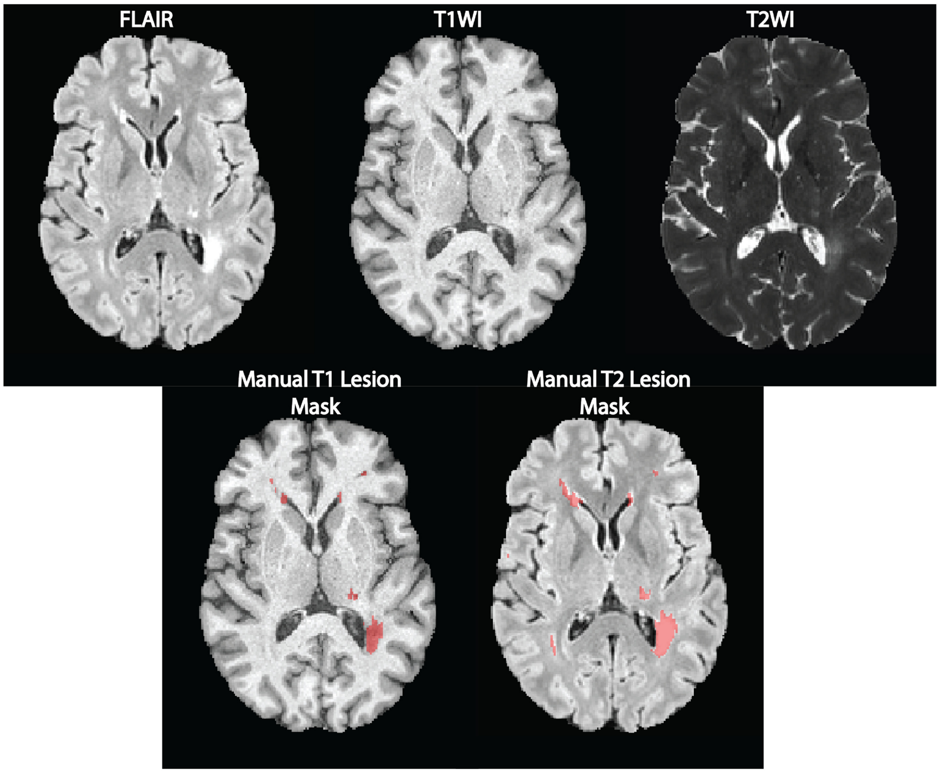


Figure 2.2: Axial slices from an inhomogeneity corrected, registered, and intensity normalized MRI of a single subject are displayed in the top row. In the bottom row, manual lesion segmentation masks are overlaid on T1WI and FLAIR volumes.

2.2.4. Automatic segmentation of T1L and T2L using MIMoSA

To automatically segment both T1L and T2L, MIMoSA (Valcarcel et al., 2018b) was applied using the *mimosa* (Valcarcel, 2018) package in R, which is available on Neuroconductor ("Neuroconductor"). MIMoSA was originally developed to automatically segment T2L. We attempted to modify the MIMoSA paradigm to better tailor it to the T1L segmentation task by introducing: 1) a two-stage model that first segments T2L and then segments T1L, and 2) a modification of the candidate mask procedure. However, these changes did not improve the results over the original MIMoSA method proposed for T2L segmentation. Therefore, the original method was applied with no changes. In this section, we broadly summarize the steps of the approach and elaborate on each step in the sections that follow.

MIMoSA relies on a brain tissue mask that excludes cerebrospinal fluid and extracerebral tissue. Given this mask, MIMoSA first identifies candidate lesion voxels by thresholding hyperintensities on the FLAIR. This step reduces computation time and minimizes false positive detection. Since feature extraction is known to be pivotal for a segmentation algorithm's accuracy and generalizability (Sweeney et al., 2014), MIMoSA relies on features that capture the mean structure of each imaging volume as well as the covariance across volumes. The procedure proceeds by creating these features, which are later used as predictors in a multivariable regression model. Once all relevant features have been calculated, MIMoSA fits a local logistic regression using training data with gold standard manual segmentations of either T1L or T2L. Coefficients from the model fit are then used to produce maps that contain the probability that each voxel location contains lesional tissue. Thresholding can be applied to the probability maps to obtain binary lesion segmentation maps for each patient. MIMoSA also includes a thresholding algorithm that optimizes the similarity of predicted segmentation masks in the training set with manual segmentations based on the Sorensen-Dice coefficient (DSC). The MIMoSA model can then be applied to subjects who were not included in the training set in order to automatically segment lesions.

In this study, MIMoSA was applied to automatically segment T1L and T2L by fitting separate models for each lesion type. One model was fit to segment T2L using preprocessed images registered to the FLAIR space, and a separate model was fit to segment T1L using preprocessed images registered to the T1WI space. This separate fitting procedure is necessary because, while all T1L are seen as T2L, not all T2L are seen as T1L (Sweeney et al., 2014). Specific steps of the MIMoSA method are described in more detail in the following sections and illustrated in Figure 2.3.

2.2.5. MIMoSA candidate voxel selection

The first step in the MIMoSA procedure is to select candidate voxels for lesion presence for a candidate mask. Since WML appear as hyperintensities on the FLAIR volume, the method excludes voxels whose FLAIR intensities are likely not consistent with lesional tissue. Candidate voxels are defined as the 85th percentile and above on the FLAIR volume. This step reduces computation time and restricts the modeling space, which empirically has been found to reduce false positives and leads to an increase in performance measures (Sweeney et al., 2013; Valcarcel et al., 2018b).

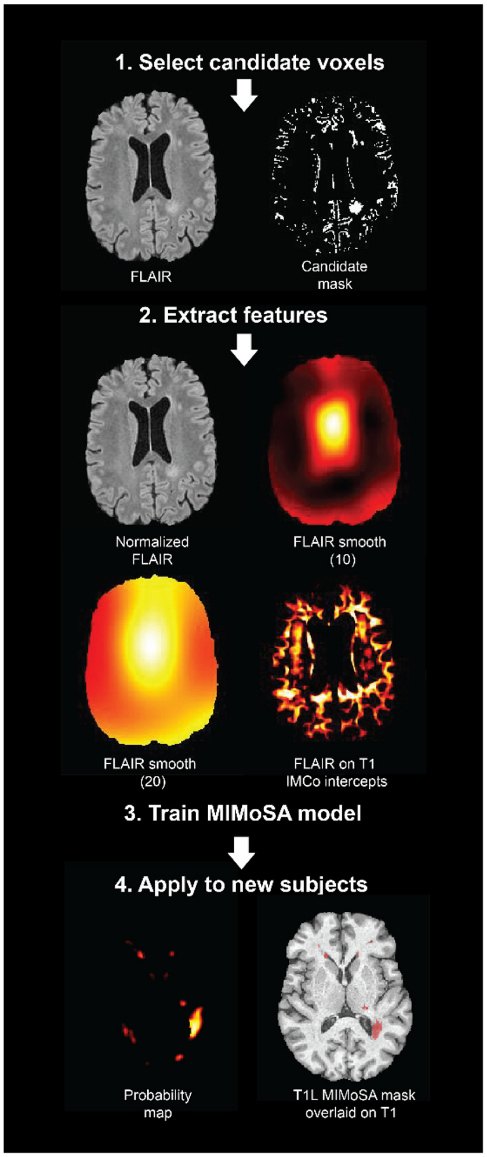


Figure 2.3: The MIMoSA procedure is demonstrated and visualized in an example axial slice. 1. MIMoSA first selected candidate voxels defined as being the 85th quantile or above in intensity on the FLAIR images. 2. Features inside the candidate mask were then extracted (full brain features derived from FLAIR volumes are only shown for simplicity). 3. To obtain T1 lesion (T1L) masks and T2 lesion (T2L) masks, separate models were fit. 4. Training the MIMoSA models on a subset of subjects with manual segmentations yielded segmentation models which were then applied to test subjects not included in the training set.

2.2.6. MIMoSA feature extraction

The next step in the algorithm is to obtain features from the candidate voxels that will be used in the model. MIMoSA utilizes three distinct feature types: (1) normalized images, (2) smoothed images, and (3) inter-modal coupling (IMCo) intercept and slope images (Vandekar et al., 2016). MIMoSA allows for T1WI, T2WI, FLAIR, and Proton Density (PD) MRI modalities as inputs, but it has been shown that only T1WI and FLAIR sequences are required to achieve statistically equivalent performance to the model with all four sequences (Valcarcel et al., 2018b). In this study, PD images were not collected; therefore, only T1WI, T2WI, and FLAIR are used as inputs and subsequently included in the model as features. Since sequences are generally acquired in arbitrary units, MIMoSA utilizes intensity-normalized images to facilitate across-subject modeling of intensities (Sweeney et al., 2013; Valcarcel et al., 2018b). To account for average signal intensities around each voxel, Gaussian smoothers with varying kernel sizes are applied to the intensity-normalized images and also included in the model. The smoothed-image features have been noted to mitigate segmentation artifacts that are due to residual image inhomogeneity after N4 correction (Shinohara et al., 2014) and to incorporate local spatial context. The model incorporates images smoothed with parameters $\sigma = 10mm$ and $\sigma = 20mm$. To further help distinguish the lesional tissue from normal appearing white matter, the MIMoSA model includes features extracted from IMCo regressions, which quantify the local covariance between two image modalities throughout the brain at the subject level.

For a given center voxel, the IMCo features are extracted from a weighted linear regression of one modality on the other in a local neighborhood around the center voxel. The weights are derived from a Gaussian kernel with fixed full width half maximum (FWHM) parameter $(3mm)$. Thus, voxels in the neighborhood are weighted by their distance to the center voxel. MIMoSA estimates the intercept and slope from a weighted linear regression at all voxels in the candidate mask for each pair of imaging modalities. That is, MIMoSA exhausts all possible pairs of the scanning contrasts available for feature extraction. For each pair, IMCo regression is performed twice so that both image types in the pair are used, once as the outcome and once as the predictor. For example, for T1WI and FLAIR images, MIMoSA performs IMCo regression using T1WI intensities as the predictor with FLAIR intensities as the outcome and then repeats the IMCo regression with T1WI as the outcome and FLAIR as the predictor. With our three contrasts, six unique IMCo regressions were performed.

2.2.7. *Fit the MIMoSA model*

After features are calculated, a logistic regression is fit to model the probability that a voxel contains lesional tissue (Walter, 2005). Logistic regression is straightforward to interpret and implement and is commonly used in the segmentation literature (Dadar et al., 2017; Sweeney et al., 2014).

The MIMoSA model is a voxel-level logistic regression that is fit using the candidate voxels. Let $L_i(v)$ be a random variable denoting voxel-level lesion presence at voxel v; if voxel v contains lesional tissue for subject i, then $L_i(v) = 1$, otherwise $L_i(v) = 0$. We model the probability that a voxel v contains lesion $P\{L_i(v) = 1\}$ with the following logistic regression model:

$$\begin{aligned} logit[P(L_i(v) = 1)] = \\ & \beta_0 + \boldsymbol{X}_i^T(v)\boldsymbol{\beta} + \mathcal{G}\boldsymbol{X}_i^T(v, 10)(\boldsymbol{\beta}_{\mathbf{10}} + \boldsymbol{X}_i^T(v) \otimes \boldsymbol{\beta}^*_{10}) \\ & + \mathcal{G}\boldsymbol{X}_i^T(v, 20)(\boldsymbol{\beta}_{20} + \boldsymbol{X}_i^T(v) \otimes \boldsymbol{\beta}^*_{20}) + \mathcal{C}\boldsymbol{X}^T_{i,l}(v)\boldsymbol{\beta}_l + \mathcal{C}\boldsymbol{X}^T_{i,s}(v)\boldsymbol{\beta}_s \end{aligned} \tag{2.1}$$

where we denote the normalized images $\boldsymbol{X}_i(v) = [T_{(1,i)}(v), FLAIR_i(v), T_{(2,i)}(v)]^T$ and express the smoothed images in vector form by $\mathcal{G}\boldsymbol{X}_i(v, \delta) = [\mathcal{G}(T_{(1,i)}(v); N(v, \delta)), \ldots, \mathcal{G}(T_{2,i}(v); N(v, \delta))]^T$, where $\mathcal{G}$ denotes the image smoothing operator with parameter $\delta \in 10mm, 20mm$. We further denote all combinations of intercept and slope IMCo parameters respectively by $\mathcal{C}\boldsymbol{X}^T_{(i,I)}(v)$ and $\mathcal{C}\boldsymbol{X}^T_{(i,S)}(v)$. We use $\otimes$ to represent the Hadamard product. The interaction terms between the normalized volumes and the smoothed volumes, denoted by $\boldsymbol{\beta}^*_{j0}$, contribute to the model by capturing differences between voxel intensities and their local mean intensities. These aid in mitigating artifacts due to residual field inhomogeneity and have generally been shown to improve lesion detection performance (Sweeney et al., 2013, 2014).

The normalized and smoothed volumes allow the MIMoSA model to capture mean structure within modalities and the IMCo features help to capture inter-modal patterns that contain information about lesion presence. The combination of modeling mean structure within an image type and the covariance across image types allows for sensitive and specific delineation of WML. The model is trained using manually segmented gold standard lesion masks. Two separate models are fit for automatically segmenting T1L and T2L using their respective gold standard masks. More specifically, the only difference between the models is whether $L_i(v)$ denotes T1L or T2L. Each model output is a set of coefficients that can be used to obtain lesion probability maps on subjects not included in the training of the model.

2.2.8. Apply the MIMoSA model

To determine where lesions are present, a probability map is obtained using the estimated regression coefficients for each voxel in the candidate mask. To create a binary segmentation, a population-level threshold on the probability map is applied. Any lesion smaller than 8 cubic millimeters is removed (Shinohara et al., 2011). Figure 2.3 shows an example of a probability map and binary segmentation for a subject not included during training of the model.

2.2.9. MIMoSA optimal thresholding algorithm

To make the method fully automated, an optimization strategy for the thresholding is employed to yield binary lesion segmentations. After the model is fit on the training data, probability maps for the subjects in the training set are generated. A threshold is then applied to the probability maps for each subject based on a user-defined grid of possible threshold values to create a set of binary segmentation masks; in this study, the grid selected was 0% to 100% by 1% increments. Using the set of predicted lesion masks for each threshold, DSC is calculated at the subject level. After DSC is calculated for each subject in the training set, the average across subjects for each threshold is collected. The threshold with the highest average DSC score is applied to probability maps estimated for subjects in the test set.

2.2.10. Statistical analyses

Training and testing of the MIMoSA method was conducted using cross-validation. In addition to implementing MIMoSA, a competitive T2L segmentation algorithm, OASIS was also applied (Valcarcel et al., 2018b). OASIS was specifically chosen for the present study because it can be easily trained using publicly available software and there are no publicly available data for benchmarking T1L automatic lesion segmentation. To fit the models and measure performance, 100 iterations of the following procedure were performed. First, 20 subjects were randomly allocated to the training set and the remaining 20 subjects constituted the test set. Thus, every subject was represented in each iteration. MIMoSA and OASIS were then trained to detect T1L and T2L separately using subjects in the training set. After fitting the models, the estimated coefficients were applied to the test set to generate probability maps. To generate lesion masks, the threshold obtained from the optimal thresholding algorithm described above was applied.

In each of the 100 iterations, subject-level DSC, partial AUC (pAUC, up to 1% false positive rate), root mean square error (Root MSE), detection error (DE) (Wack et al., 2012), and outline error (OE) (Wack et al., 2012) were recorded (Sing et al., 2005). pAUC was estimated instead of traditional AUC because it only considers regions of the ROC space that correspond to clinically relevant values of specificity (Walter, 2005). All performance measures were calculated at the subject level and then averaged across subjects and cross-validation folds. Figure 2.4 shows the full cross-validation pipeline. In addition to these summary measures, MIMoSA performance was assessed by estimating the Pearson correlation ($\hat{\rho}$) between manually segmented and MIMoSA-predicted volumes.

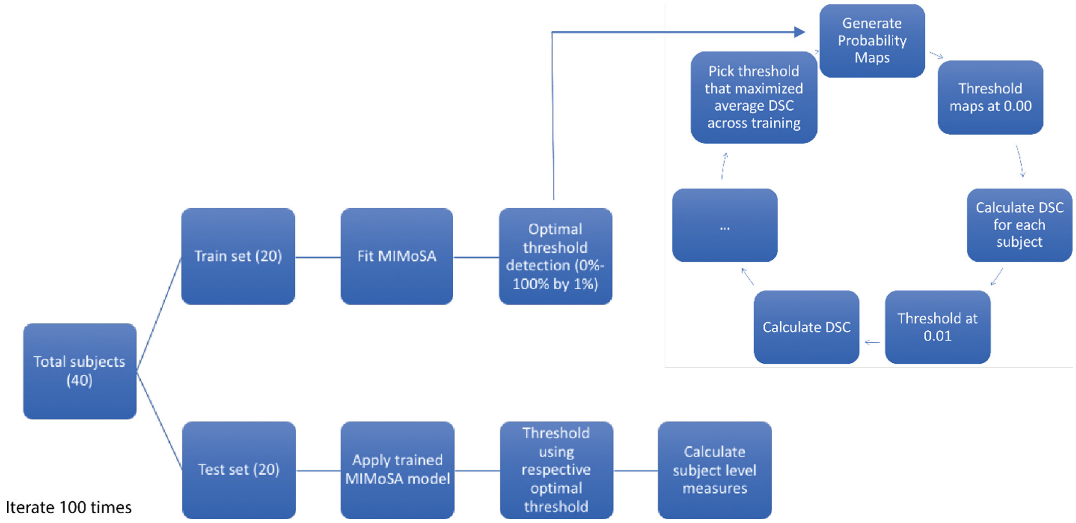

Generate Probability Maps
Pick threshold that maximized average DSC across training
Threshold maps at 0.00
...
Calculate DSC for each subject
Calculate DSC
Threshold at 0.01
Train set (20)
Fit MIMoSA
Optimal threshold detection (0%-100% by 1%)
Total subjects (40)
Test set (20)
Apply trained MIMoSA model
Threshold using respective optimal threshold
Calculate subject level measures
Iterate 100 times

Figure 2.4: Cross-validation scheme used to assess MIMoSA performance on T1 lesions (T1L) and T2 lesions (T2L) is pictured. Subjects were randomized to either the training set or the testing set. The MIMoSA model was fit using subjects in the training set. To identify the optimal threshold, probability maps were generated for subjects. These maps were thresholded along a grid selected from 0% to 100% by 1% and then the Sorensen-Dice coefficient (DSC) was calculated. The threshold that resulted in the maximum DSC across subjects in the training set was applied as the threshold in the test set. This procedure was iterated 100 times. Summary statistics are based only on the test set data. The same analysis was repeated using OASIS as the segmentation approach.

To adjudicate MIMoSA's performance, Pearson correlation coefficients were calculated to assess the relationship between image-derived features (T1L volume, T2L volume, and the T1L/T2L ratio

(Kim et al., 2016) and clinical variables, including clinical status, disease duration (time from first symptoms in years), EDSS score, and T25FW. Manual segmentation-based measures of T1L, T2L, and the T1L/T2L ratio were also computed, and associations with clinical variables were estimated for comparison. To avoid overfitting, correlations were estimated in each cross-validation fold using only subjects in the test set and then averaged across folds. We denote MIMoSA measures by $\hat{\rho}(MIMoSA)$, whereas manual evaluations are represented by $\hat{\rho}(Manual)$. For each measure, p-values were similarly calculated in each fold and averaged across folds. We additionally calculated each measure adjusted for sex and age.

In order to assess the accuracy and variability of the optimal threshold for each subject in the testing set we applied thresholds from 0% to 100% by 1% increments to obtain lesion masks. DSC was then calculated comparing the MIMoSA mask at each threshold with the manual segmentation.

2.3. Results

2.3.1. Segmentation Accuracy

Results are provided in Table 2.3, including average DSC, partial AUC with up to 1% false positive rate, and the correlation coefficient for MIMoSA and OASIS volumes with manual volumes ($\hat{\rho}$). Results in Table 2.3 indicate competitive lesion segmentation performance of both T1L and T2L. DSC and pAUC for T2L lesion segmentation were competitive compared to state-of-the-art automatic methods. DSC and pAUC for T1L were modest compared to those measures for T2L but high compared with previous automated approaches in T1L studies. The MIMoSA performance measures were all greater than the OASIS performance measures, indicating superior automatic segmentation. Specifically, for T1L the 95% confidence interval for DSC was 0.02 to 0.16 and pAUC was 0.03 to 0.13. Since 0 is not contained in these intervals, we can conclude that MIMoSA statistically significantly segmented T1L more accurately than OASIS.

Similarly, the DE, OE, and Root MSE were all lower for MIMoSA segmentations than OASIS, indicating that MIMoSA has less error. The DE for both methods was very small, indicating that the automatic methods detected most of the lesions that were found manually. OE was much higher than DE, indicating that the automatic methods tended to disagree at the boundary of lesions. Root MSE, though very small for both MIMoSA and OASIS, favored MIMoSA and suggested that

Table 2.3: Results from the cross-validation are presented. Sorensen-Dice coefficient (DSC), partial AUC (pAUC) with up to 1% false positive rate, root mean square error (Root MSE), detection error (DE), and outline error (OE) were averaged within each testing set and then across folds. Standard deviation (SD) was calculated within cross-validation folds and then averaged across 100 iterations. DE and OE are presented in mL. The correlation coefficient relating MIMoSA volumes to manual volumes ($\hat{\rho}$) was recorded in each fold and then averaged across folds.

	DSC (**SD**)	**pAUC** (**SD**)	**Root MSE** (**SD**)	**DE** (**SD**)	**OE** (**SD**)	$\hat{\rho}$
MIMoSA T1L	0.53 (0.14)	0.64 (0.12)	0.06 (0.03)	1.02 (0.96)	9.22 (9.63)	0.88
OASIS T1L	0.43 (0.14)	0.55 (0.13)	0.08 (0.04)	1.76 (1.49)	9.85 (1.49)	0.85
MIMoSA T2L	0.66 (0.13)	0.70 (0.10)	0.07 (0.03)	1.41 (1.12)	14.9 (13.8)	0.95
OASIS T2L	0.55 (0.13)	0.62 (0.11)	0.09 (0.04)	2.55 (2.17)	15.6 (15.1)	0.88

MIMoSA had smaller average error.

Change in lesion volume and counts are both important outcomes commonly used in MS clinical trials (Bakshi et al., 2005). The correlation between manual segmentation volume and MIMoSA volume was high for both T1L and T2L. In Figure 2.5, plots of MIMoSA predicted volume are displayed against manual segmentation volume. The trend for both T1L volume and T2L volume were markedly linear and close to the identity line. Subjects with low total lesion volume tended to have accurate MIMoSA volume estimation with small variance. As total lesion volume increases, the standard deviations around the MIMoSA volume estimates increase. Figure 2.5 also provides plots of MIMoSA predicted count versus manual segmentation count. The count estimated by MIMoSA for subjects with smaller lesion volumes (i.e. less than 25mL) was similar to the manual segmentation count. For larger lesion loads, MIMoSA underestimated the count. With a few exceptions, subjects with low lesion counts tended to have small variance around the MIMoSA estimate, but variability of the estimates can be seen to increase along with increasing lesion counts. Although MIMoSA underestimated lesion count for subjects with large manual lesion counts, the MIMoSA volume estimate remained accurate. In follow-up investigations, we found that the joint underestimation of lesion count and accurate estimation of volume by MIMoSA was attributable to generous segmentation of spatially neighboring lesions that resulted in more confluent lesions.

Subject-level DSC and pAUC are presented in Figure 2.6. While DSC tended to be larger for patients with larger manual lesion volume, pAUC tended to be higher for patients with small to moderate manual lesion volume.

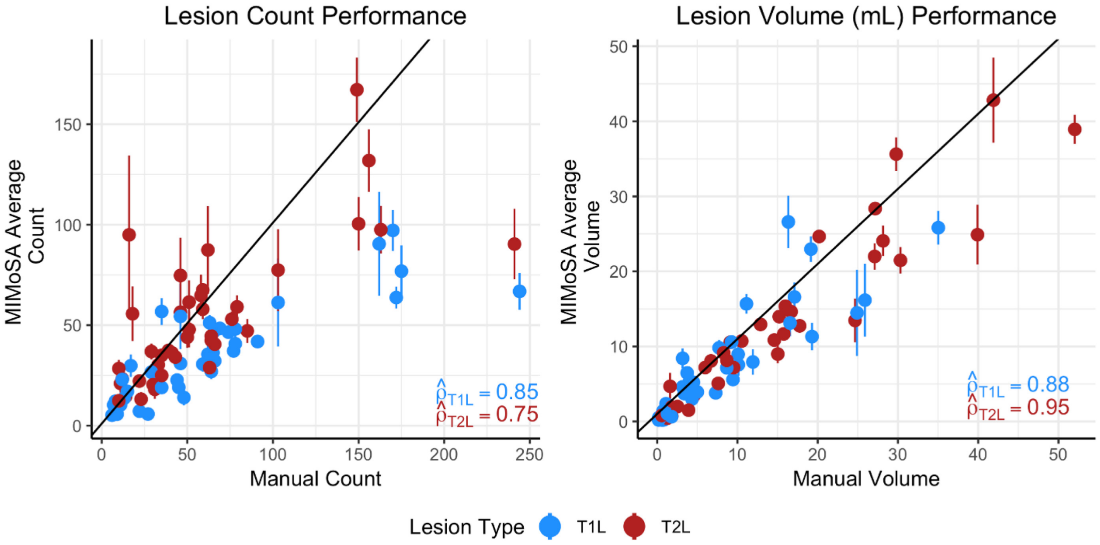


Figure 2.5: Lesion volume and count are presented to compare manual segmentation with MIMoSA segmentation metrics. MIMoSA values were obtained by averaging volume or count for each test subject across cross-validation folds (100). The solid line depicts the $y = x$ line. Vertical lines traversing the points are computed at the subject-level and indicate one standard deviation above and below the mean.

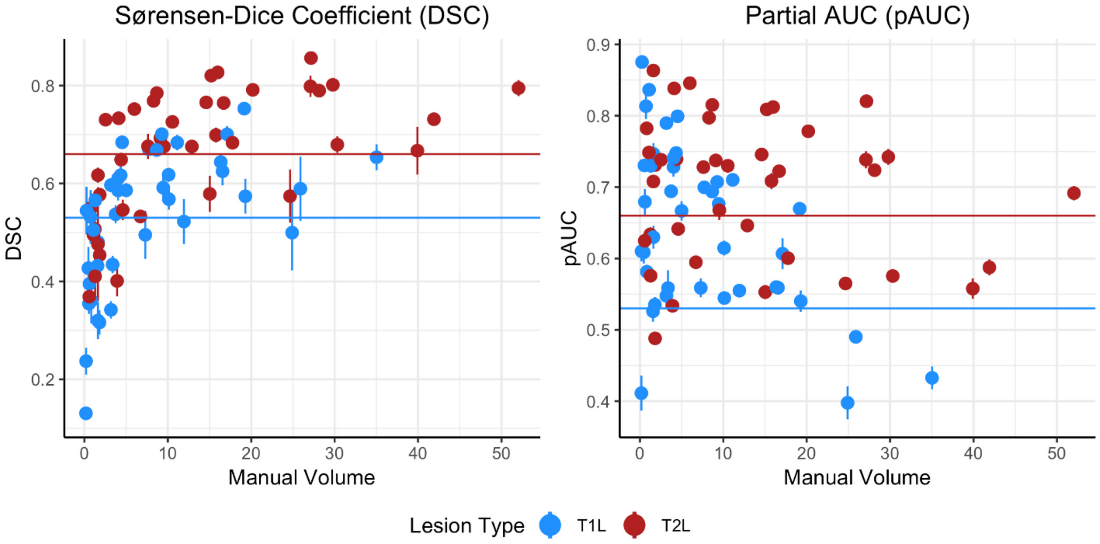


Figure 2.6: To further demonstrate model accuracy, Sorensen-Dice coefficient (DSC) and partial AUC (pAUC) with up to 1% false positive rate were calculated. Results for each subject were averaged across folds and are presented. Horizontal lines are the respective overall averages presented in Table 2.3. Vertical lines traversing the points are computed at the subject-level and indicate one standard deviation above and below the mean.

2.3.2. Correlations with clinical status

In practice, lesion segmentation metrics are commonly used to predict clinical status and evaluate therapeutic efficacy (Bakshi et al., 2005; Zivadinov and Bakshi, 2004). In Table 2.4, clinical

Table 2.4: Clinical-MRI relationships with manual lesion volume, denoted as $\hat{\rho}(Manual)$, or MIMoSA lesion volume, denoted as $\hat{\rho}(MIMoSA)$, was averaged across cross-validation folds. T1 lesion (T1L) volume, T2 lesion (T2L) volume, and the T1L/T2L ratio were correlated separately with Expanded Disability Status Scale (EDSS) score, timed 25-ft walk (T25FW), and disease duration. For each assessment, p-values were calculated and are presented in parentheses beside each measure. The first table presents unadjusted correlations; the second table presents correlations adjusted for sex and age (in years).

		Clinical Correlations		
		EDSS	**T25FW**	**Disease Duration**
T1L	$\hat{\rho}(Manual), (p-value)$	0.32, (0.26)	-0.07, (0.56)	0.12, (0.54)
	$\hat{\rho}(MIMoSA), (p-value)$	0.34, (0.21)	-0.05, (0.58)	0.29, (0.30)
T2L	$\hat{\rho}(Manual), (p-value)$	0.33, (0.24)	-0.07, (0.55)	0.15, (0.52)
	$\hat{\rho}(MIMoSA), (p-value)$	0.32, (0.23)	-0.08, (0.56)	0.23, (0.37)
T1L/T2L ratio	$\hat{\rho}(Manual), (p-value)$	0.33, (0.22)	0.13, (0.56)	0.06, (0.54)
	$\hat{\rho}(MIMoSA), (p-value)$	0.33, (0.22)	0.18, (0.45)	0.40, (0.12)

		Adjusted Clinical Correlations		
		EDSS	**T25FW**	**Disease Duration**
T1L	$\hat{\rho}(Manual), (p-value)$	0.36, (0.23)	-0.03, (0.56)	0.08, (0.59)
	$\hat{\rho}(MIMoSA), (p-value)$	0.34, (0.25)	0.04, (0.58)	0.13, (0.53)
T2L	$\hat{\rho}(Manual), (p-value)$	0.38, (0.21)	-0.02, (0.55)	0.10, (0.57)
	$\hat{\rho}(MIMoSA), (p-value)$	0.34, (0.23)	-0.05, (0.58)	0.14, (0.50)
T1L/T2L ratio	$\hat{\rho}(Manual), (p-value)$	0.36, (0.20)	0.16, (0.53)	0.12, (0.51)
	$\hat{\rho}(MIMoSA), (p-value)$	0.30, (0.31)	0.18, (0.46)	0.26, (0.30)

measures are related to both manual and MIMoSA lesion segmentation metrics. EDSS score and T25FW were correlated with T1L and T2L volume, as well as the T1L/T2L ratio. The correlations displayed in this table show that $\hat{\rho}(MIMoSA)$ tended to be equal to or larger than $\hat{\rho}(Manual)$. The associated p-values in Table 2.4 indicate that MIMoSA and the manual segmentations performed similarly. Age and sex-adjusted results were similar to unadjusted results, with the exception of EDSS. Results are visualized in Figure 2.7. The correlations, whether calculated with the manual or MIMoSA volumes, were modest but consistent with the established literature.

Figure 2.7 also facilitates the comparison of correlations across the T1L and T2L metrics, both marginally and adjusted for age and sex. T1L and T2L tended to have similar correlations with clinical variables. However, the T1L/T2L ratio has similar or higher correlations with clinical measures. Notably, the partial correlations of T1L and T2L with T25FW are small in magnitude, whereas the T1L/T2L ratio is more strongly associated with T25FW.

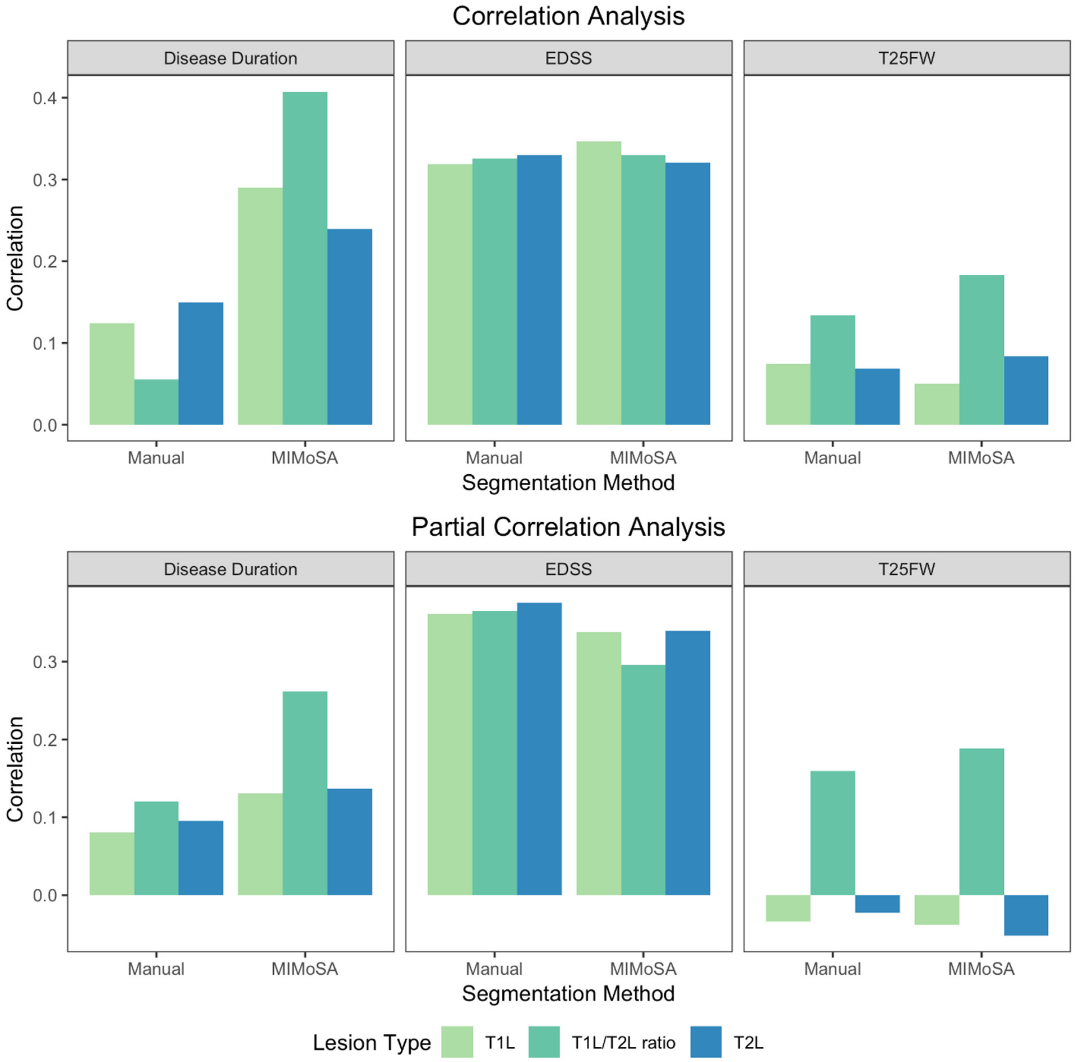


Figure 2.7: Visualization of clinical-MRI relationships. Both manual and MIMoSA segmentations provided T1 lesion (T1L) volume, T2 lesion (T2L) volume, and the T1L/T2L ratio. The value of the vertical axis for disease duration, Expanded Disability Status Scale (EDSS) score, and timed 25-ft walk (T25FW) represents the Pearson correlations between each measure and the MIMoSA or manual segmentation volume. The first row presents unadjusted correlations and the second row presents correlations adjusted for sex and age (in years).

2.3.3. Optimal threshold

To assess the accuracy and variability of the optimal thresholding algorithm, Table 2.5 presents summary measures across the cross-validation iterations. The mean values were slightly larger for T2L compared to T1L, while the standard deviations and range were similar. Figure 2.8 shows average DSC across thresholds applied to subjects in the testing set. For both T1L and T2L,

Table 2.5: Summary measures for the optimal threshold obtained across iterations in the cross-validation are shown for T1 lesions (T1L) and T2 lesions (T2L).

Lesion Type	**Mean**	**Std. Dev.**	**Min, Max**
T1L	0.28	0.05	0.2,0.36
T2L	0.32	0.04	0.25,0.39

the average optimal threshold that was applied, denoted by the colored point, lay close to the peak of each curve, which indicates that the optimal threshold algorithm indeed chose appropriate thresholds to apply to test subjects. Additionally, we note that the relatively flat areas of the curves surrounding the maximum DSC value suggest that slight differences in thresholds did not have a major impact on segmentation accuracy.

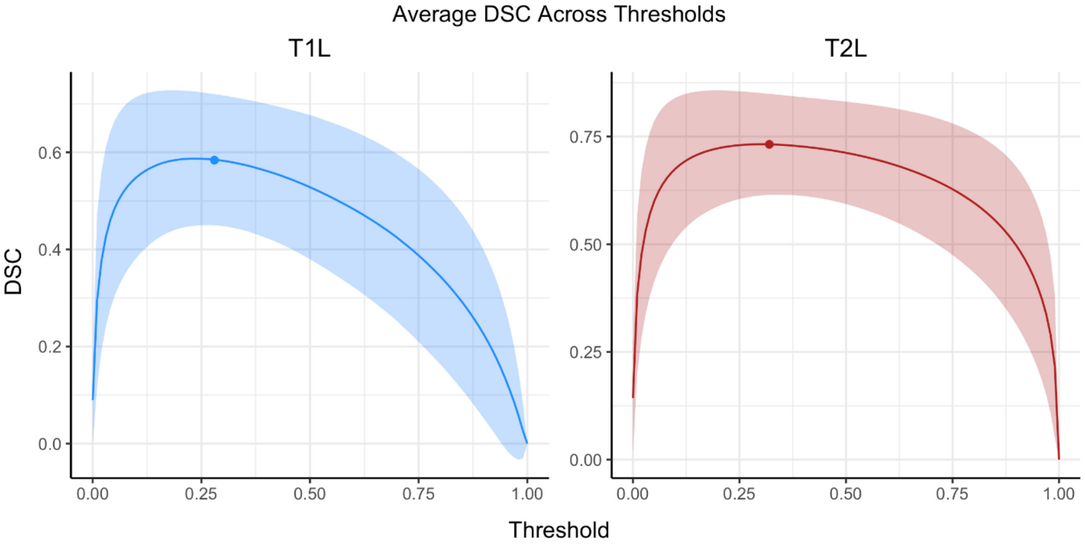


Figure 2.8: To assess the accuracy and variability of the optimal threshold, the average Sorensen-Dice coefficients (DSC) across subjects and iterations are shown across thresholds. Results for T1 lesions (T1L) and T2 lesions (T2L) are presented separately. The solid line represents the average while the filled-in area corresponds to one standard deviation from the mean. The round points on each figure are the average optimal threshold selected.

2.3.4. Qualitative performance

An example of MIMoSA's qualitative performance is provided in Figure 2.9, where an axial slice from a subject chosen at random is provided. MIMoSA masks are overlaid on T1WI and FLAIR volumes respectively for T1L and T2L, and the probability maps used to generate MIMoSA segmentations are shown. Qualitative results were consistent with quantitative performance.

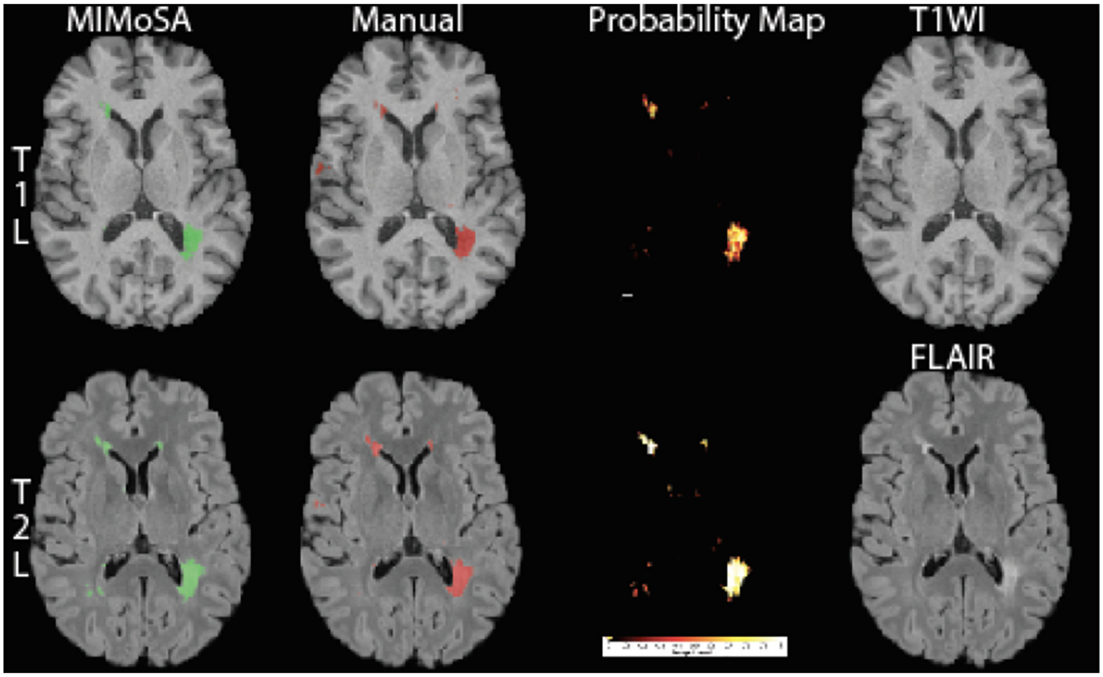


Figure 2.9: Segmented T1 lesions (T1L) and T2 lesions (T2L) in a randomly selected subject and axial slice are pictured. The first row shows T1L segmentations for both MIMoSA and manual assessment, the MIMoSA probability map, and the T1WI volume. In the second row, T2L segmentations for both MIMoSA and manual assessment, the MIMoSA probability map, and the FLAIR volume are displayed. The Sorensen-Dice coefficients (DSC) between the MIMoSA and manual segmentation for T1L and T2L were 0.54 and 0.69, respectively. To elucidate the differences between the T1L and T2L tissue type segmentations for both the MIMoSA and manual segmentations, we provide DSC between the lesion types. The DSC between MIMoSA T1L and T2L was 0.64 and the DSC between the manually segmented T1L and T2L was 0.52.

2.4. Discussion

MIMoSA is a fully automated segmentation method that leverages changes in inter-modal covariance structure that occurs in white matter pathology. It can be used to delineate T1L and T2L accurately, reliably, and efficiently in people with MS. Improvements in accuracy seem to be driven by the inclusion of IMCo regression features, which are features that are not included in OASIS. These measures seem especially useful for detecting T1L, a challenging task since T1L lesions often appear similar to gray matter. MIMoSA does not require human input, which promises to promote stability across a range of lesion delineation tasks. By using the same procedure to automatically segment T1L and T2L, MIMoSA also offers a consistent framework to obtain both metrics. Furthermore, the optimal thresholding algorithm fully automates the MIMoSA segmentation method by using the training subjects and their manual segmentations to provide a threshold that empirically works well in the test set. Results from our cross-validation experiments demonstrate its accuracy and support its use in practice. The MIMoSA model can easily be adapted and trained for cases with different sets of imaging sequences (Sweeney et al., 2013; Valcarcel et al., 2018b). The full modeling procedure is fast and can be easily implemented using software and documentation provided on Neuroconductor ("Neuroconductor"; Valcarcel, 2018).

MIMoSA provides accurate and reliable automatic segmentations of both T1L and T2L. Though T2L DSC and pAUC measures were slightly larger, indicating greater similarity with our manual segmentations, T1L performance was competitive. Simultaneous delineation of T1L and T2L may lead to a better understanding of overall patient status. MIMoSA total lesion volumes were well-correlated with the manual total lesion volumes, suggesting that MIMoSA may provide a promising alternative to manual segmentation in the assessment of new therapies in clinical trials (Valcarcel et al., 2018b). This may be especially useful for multi-center studies with a large number of patients or longitudinal studies with sequences collected over time.

The MIMoSA method was previously implemented on data acquired at a different site using a different scanner and acquisition protocol than data collected in this study (Sweeney et al., 2013; Valcarcel et al., 2018b). The results here indicate that the method performed well using images acquired across scanner manufacturers and protocols when the model was appropriately trained. Previously published experiments indicate that 20 subjects is sufficient for model training (Valcarcel et al., 2018b). Pre-trained models are available for immediate application of the method, but for the best results training on data acquired under the protocol of interest is encouraged ("Neuroconductor"; Valcarcel, 2018). The MIMoSA method should be implemented after appropriate image preprocessing. MIMoSA users should be aware that processing failures in registration, skull-stripping, and normalization may lead to segmentation failures. Quality control should be implemented after each step of preprocessing before applying MIMoSA.

Often lesion volumes are correlated with clinical covariates and disease status in patient management and clinical trials that evaluate therapy effectiveness. Therefore, automatic segmentation approaches should be as sensitive as manual measures. Correlations were provided to compare manual and MIMoSA segmentations with clinically relevant variables. Our results indicate that the relationship between MIMoSA volumetric assessments showed as close or better correlations compared to correlations with manual segmentations. This was likely due to the stability and consistency introduced by an automatic method that requires no operator input. Segmentation of T1L can be challenging since the intensity profile is often indistinguishable from gray matter (Bakshi et al., 2005), especially with respect to delineating boundaries; thus, reliability in these areas may be driving stronger correlation with covariates. For T2L evaluation, correlations seemed to be approximately equal between MIMoSA and manual segmentations. In general, the measurements, whether obtained from manual segmentation or MIMoSA, were similar, advocating for the use of the automated method to reduce cost and time.

In this study, T1L and T2L (Barkhof, 1999) were correlated approximately equally with clinical metrics. While the sample size and cross-validation in this study were powerful enough to evaluate the accuracy of MIMoSA, it did not likely provide sufficient power to show improvement in clinical associations. With a larger clinical cohort, it should be possible to see the increased clinical value of T1L compared to T2L. Additionally, the images were acquired using a gradient echo acquisition which has been shown in the literature to identify T1L more commonly than a spin echo acquisition but with weaker associations to clinical status (Dupuy et al., 2015). The T1L/T2L ratio demonstrated equal or stronger associations with clinical covariates compared to T1L or T2L volumes alone, motivating the advantage of segmenting both T1L and T2L.

In this dataset, two subjects presented with gadolinium enhancing lesions. Unfortunately, without a post-contrast T1 included in the MIMoSA model, we tend to segment these as T1L. In the future, we propose to include post-contrast T1 imaging in the MIMoSA model to assess the capability of MIMoSA to distinguish black holes from contrast-enhancing lesions. We will also evaluate whether MIMoSA improves longitudinal assessment of dynamic lesion evolution and therapeutic response over currently available methods, in particular, when a number of sequences are collected at each visit. Finally, we demonstrated MIMoSA's robustness to multiple scanners and protocols when assessing T2L volume. Thus, MIMoSA may be useful for large, multi-center clinical trials that employ a number of different scanners. In all future work, comparison of MIMoSA T1L and T2L volumes to benchmark manual assessment is warranted.

CHAPTER 3

TAPAS: A THRESHOLDING APPROACH FOR PROBABILITY MAP AUTOMATIC SEGMENTATION IN MULTIPLE SCLEROSIS

3.1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory and degenerative disease of the central nervous system characterized by multifocal demyelinating lesions (Compston and Coles, 2002; Confavreux and Vukusic, 2008) and atrophy in both white and gray matter, which may lead to physical and cognitive disability and poor functional outcomes (e.g. social isolation, unemployment) (Rovira and Len, 2008; Tauhid et al., 2015). In MS research and clinical care, magnetic resonance imaging (MRI) is a commonly used tool for detection and quantification of disease activity and severity (Bakshi et al., 2005; Ge, 2006; Zivadinov and Bakshi, 2004). MRI allows for the detection of T2-weighted (T2) hyperintense white matter lesions. Both lesion volume and count have become important metrics in the clinical and research domain (Dworkin et al., 2018; Ge, 2006). Advanced MRI also allows for cortical lesion detection, one of the new biomarkers integrated in the revised McDonald criteria (Thompson et al., 2018). Typically, total lesion burden (i.e. lesion load), is defined as the volume of total brain matter containing lesions and is a cornerstone for assessing disease severity in MS research and clinical investigations (Calabresi et al., 2014; Popescu et al., 2013; Tauhid et al., 2014).

To quantify lesion burden, different approaches use MRI to identify and segment lesional tissue. Manual segmentation is the gold standard approach and requires a neuroradiologist or imaging expert to inspect scans visually and delineate lesions. Due to difficulties associated with manual segmentation such as cost, time, and large intra- and inter-rater variability, many automatic segmentation methods have been developed (Carass et al., 2017b; Egger et al., 2017; Garcia-Lorenzo et al., 2013; Llado et al., 2012). Unfortunately, since lesions present heterogeneously on MRI scans, automatic segmentation remains a difficult task, though numerous methods have been proposed. No single approach is widely accepted or proven to perform optimally across lesion types, scanning platforms, and centers (Danelakis, Theoharis, and Verganelakis, 2018; Sweeney et al., 2014). A common key step in automatically delineating lesions involves creating a continuous map indicating the degree of lesion likelihood using various imaging modalities (Danelakis, Theoharis, and Verganelakis, 2018; Roy et al., 2015; Sweeney et al., 2013, 2014; Valcarcel et al., 2018b). In these cases, a threshold is then applied to probability maps to obtain binary lesion segmentations, also referred to as lesion masks.

Automatic approaches are susceptible to biases in lesion volume estimation associated with the total lesion load (Commowick et al., 2018); that is, in subjects with few lesions, automated techniques tend to over-segment lesions, and in subjects with higher lesion load, lesions are under-segmented. Bias in lesion volume estimation may also be associated with MRI hardware specifications, differences in protocols, artifacts, or partial volume effects.

To investigate this volume bias, we leveraged the 2015 Longitudinal Lesion Challenge (`https://smart-stats-tools.org/lesion-challenge`) (Carass et al., 2017a,b), a publicly available data set consisting of imaging of five subjects with MS for training and fourteen subjects with MS for testing. In training and testing sets, subjects had at least four imaging visits. The training data contain manual delineations from two expert raters while the testing set does not publicly provide manual delineations; rather, the testing set only consists of volume estimates from each rater. Challengers who wish to compare new segmentation methods can submit their testing set automatic segmentations. The automatic segmentation method is ranked using a weighted average of various similarity measures. A leader board with method performance measures is maintained by challenge organizers and some published work compares top performing methods (Carass et al., 2017b).

We present data from challengers as Bland-Altman plots (Bland and Altman, 2007, 2016) to assess disagreement with manual volumes from the top two performing approaches described in Carass et al., 2017b (see appendix Table C3). Bland-Altman plots are provided in Figure 3.1 to compare the automatically generated and manually delineated volumetric measures. This graphical approach presents the differences between techniques, automatic and manual, against the averages of the two. If no points lie outside the limits of agreement, the mean difference plus and minus 1.96 times the standard deviation of the differences, according to classical guidelines this indicates the difference between techniques is not clinically important and the two methods can be used interchangeably.

The plots in Figure 3.1 show systematic deviations in automatic and manual volumes. Both ranked

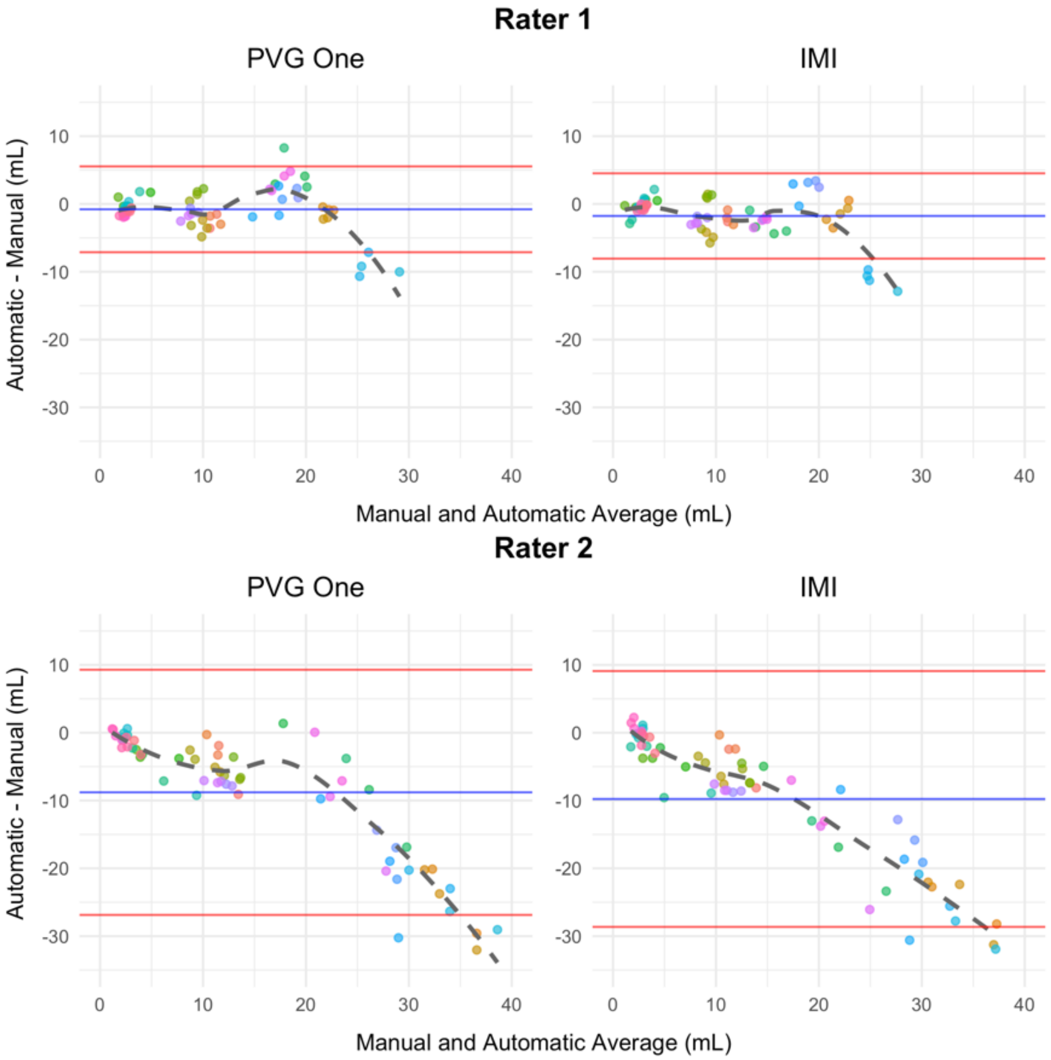


Figure 3.1: Bland-Altman plots using the first (left) and second (right) ranked automatic segmentation methods' volumes from the 2015 Longitudinal Lesion Challenge are presented. We show plots comparing volumes obtained from the automatic and manual methods. The manual volumes were delineated by rater 1 (top) and rater 2 (bottom). Using the differences, we highlight the mean (blue) plus and minus 1.96 times the standard deviation (red). Each subject is represented in a unique color and each point represents a subject-time point. There are fourteen unique subjects with at least four follow-up imaging sessions.

methods show that as lesion load increases, automatic segmentation approaches underestimate volume compared with rater 1 and rater 2. This is evident by the dashed fitted smooth lines which deviate away from the mean and outside the limits of agreement starting around lesion loads larger than 20 mL in all four of the plots. While the direction of over- or under-estimation and magnitude vary for rater 1 and rater 2 across challenge submissions, each approach shows systematic deviation and bias in volume estimates. Bias in manual segmentation may be due to the inability of raters to objectively delineate the diffuse part of lesions. Supervised automatic approaches require manual segmentations for training, and therefore may be biased in focusing only on the focal portions of lesions ignoring regions of diffuse signal abnormalities near the boundaries of lesions.

The bias present in the volumetric estimates from automatic approaches may be related to the thresholding procedure that segmentation methods apply to probability maps in order to create binary lesion masks. Currently, there are no stand-alone automated approaches for choosing thresholds for segmentation. After probability maps are created, experts may inspect each subject and visually determine a threshold to apply that performs well. Likewise, users may pick a single threshold that generally performs well across all subjects (Sweeney et al., 2013). These two thresholding methods, similar to manual segmentation, introduce human bias, cost, and time into the automated procedure. Several recent publications use cross-validation approaches for determining a threshold to apply to all subjects (see Roy et al., 2015; Valcarcel et al., 2018b for example), but most methods do not provide sufficient detail to reproduce the thresholding approach. Further, these methods propose a group-level threshold rather than subject-specific thresholds.

Using probability maps generated by an automatic segmentation method, we fit the subject-specific threshold that yields the maximum expected Sorensen-Dice similarity coefficient (DSC) (Zijdenbos et al., 1994) based on a naive estimate of lesion volume using a generalized additive model. This approach provides a supervised method to detect a subject-specific threshold for lesion segmentation by attempting to estimate a threshold that optimizes DSC and reduces bias. DSC is defined as the ratio of twice the common area to the sum of the individual areas. That is, $DSC = \dfrac{2\#\{A_1 \cap A_2\}}{\#\{A_1\} + \#\{A_2\}} \in [0, 1]$ where $\#\{A\}$ denotes the number of voxels classified as lesion in measurement A. After training on a subset of subjects with manual segmentations, the TAPAS model can be applied to estimate a subject-specific threshold to apply to lesion probability maps in order to obtain automatic segmentations. The TAPAS method is fully transparent, fast to implement, and simple to train or modify for new data sets.

3.2. Materials and methods

3.2.1. Data and preprocessing

The first data set studied (JHH data) was collected at the Johns Hopkins Hospital in Baltimore, Maryland. This data set consists of 98 subjects with MS, four of which were excluded from our analyses due to poor image quality. Whole-brain 3D T1-weighted (T1), 2D T2-weighted fluid attenuated inversion recovery (FLAIR), T2-weighted (T2), and proton density-weighted (PD) images were acquired on a 3 tesla (3T) MRI scanner (Philips Medical Systems, Best, The Netherlands). A more detailed description of the acquisition protocol was provided in previously published work (Sweeney et al., 2013; Valcarcel et al., 2018b). Manual T2 hyperintense lesion segmentations for each subject were delineated by a neuroradiology research specialist with a Bachelor of Arts in Neuroscience trained in manual segmentation of MS lesions with more than 10 years of experience.

All images were N3 bias corrected (Sled, Zijdenbos, and Evans, 1998). The T1 scan for each subject was then rigidly aligned to the Montreal Neurological Institute (MNI) standard template space at 1 mm^3 isotropic resolution. FLAIR, PD, and T2 images were then aligned to the transformed T1 image. Extracerebral voxels were removed from all images using the Simple Paradigm for Extra-Cerebral Tissue Removal: Algorithm and Analysis (SPECTRE) algorithm (Carass et al., 2011). MRI scans were acquired in arbitrary units, and therefore analyzing images across subjects required that images be intensity-normalized. We thus intensity normalized each modality using *WhiteStripe* (Muschelli and Shinohara, 2018; Shinohara et al., 2014). All image preprocessing was conducted using tools provided in Medical Image Processing Analysis and Visualization (MIPAV) (McAuliffe et al., 2001), TOADS-CRUISE (`http://www.nitrc.org/projects/toads-cruise/`), Java Image Science Toolkit (JIST) (Lucas et al., 2010), and Neuroconductor (“Neuroconductor”) R (version 3.5.0) (R Development Core Team, 2018) packages.

We used a second data resource (BWH data) collected at the Brigham and Women’s Hospital in Boston, Massachusetts from 40 subjects with MS. MRI data were consecutively obtained. High-resolution 3D T1, T2, and FLAIR scans of the brain were collected on a Siemens 3T Skyra unit with a 20-channel head coil. The detailed scan parameters have been reported previously in Table 2.2 as well as in Meier et al., 2018; Valcarcel et al., 2018a.

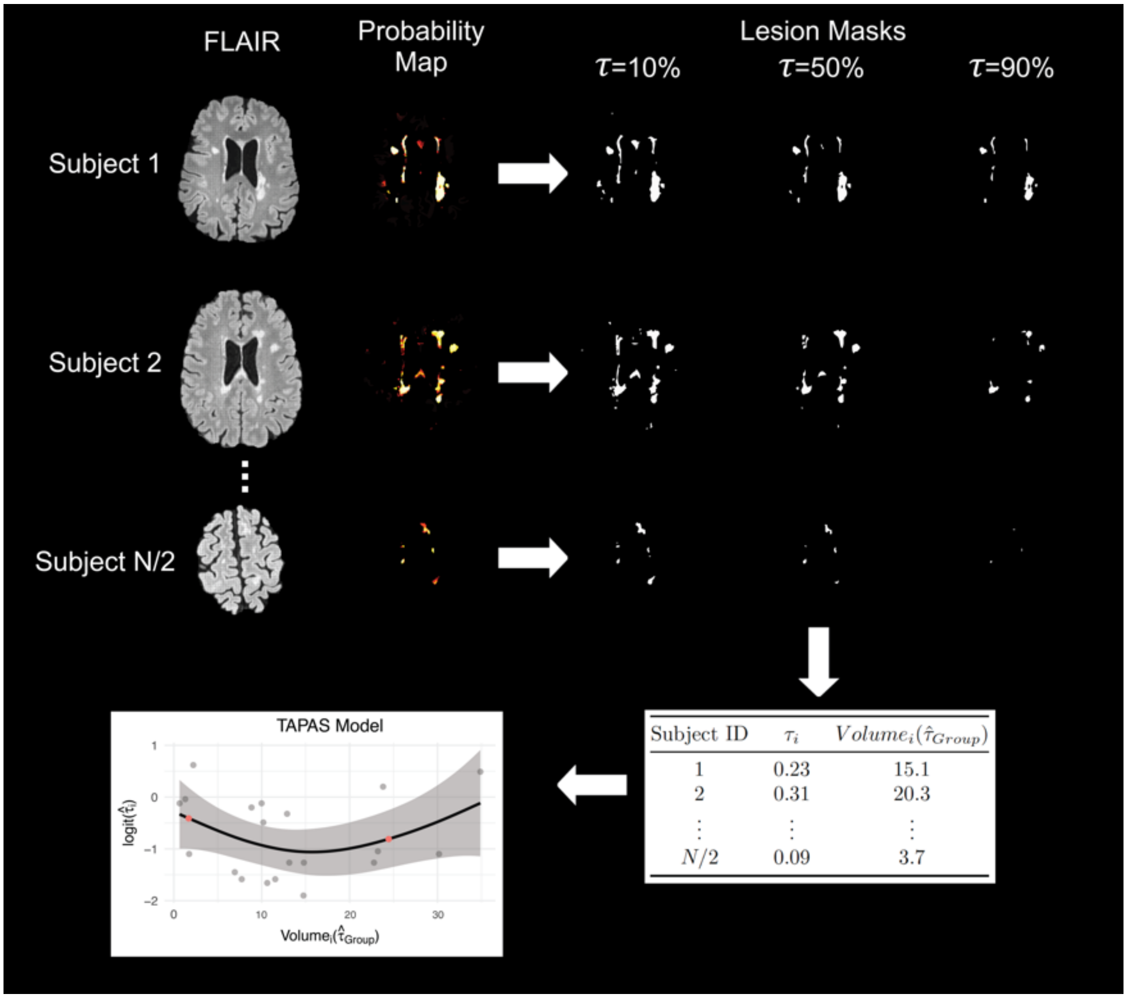
T2 hyperintense lesions were manually segmented by a reading panel of two trained observers, referred to here as rater 1 and rater 2, under the supervision of an experienced observer, referred to as rater 3, at the Brigham and Women's Hospital. A lesion was included if it appeared as hyperintense on the FLAIR. Raters 1 and 2 independently marked all MS lesions and then reviewed these results together to form a consensus. In the event of a disagreement, rater 3 was consulted and resolved any differences. After a consensus of marked lesions was determined, rater 1 segmented all lesions to determine their volume using an edge-finding tool in Jim (*Jim* 2014). This process resulted in a manually segmented gold standard lesion mask for each subject in the study. Rater 3 certified the final lesion delineation. Rater 1 had a neuroscience undergraduate degree as well as three years of work experience evaluating MS lesions on MRI scans as a research assistant. Rater 2 had a medical doctorate and four years of experience working in MS MRI research. Rater 3 had a medical doctor degree as well as more than 10 years of experience in MS MRI, initially as a trained research fellow, then serving as a faculty member and image analyst.

We performed N4 bias correction (Tustison et al., 2010) on all images and rigidly co-registered T1 and T2 images for each participant to the corresponding FLAIR at 1 mm^3 resolution. Extracerebral voxels were removed from the registered T1 images using Multi-Atlas Skull Stripping (MASS) (Doshi et al., 2013) and the brain mask was applied to the FLAIR and T2 scans. We intensity-normalized images to facilitate across-subject modeling of intensities using *WhiteStripe* (Muschelli and Shinohara, 2018; Shinohara et al., 2014). Image preprocessing was applied using software available in R (version 3.5.0) (R Development Core Team, 2018) and from NITRC (`https://www.nitrc.org/projects/cbica_mass/`).

The Institutional Review Boards at the appropriate institutions approved these studies.

3.2.2. TAPAS algorithm

Although the two data sets were processed using different pipelines, the proposed technique is completely independent of the preprocessing pipeline. We applied the BWH preprocessing pipeline to the JHH data and re-ran the analyses; we present these results in section A.2 of the Appendix. TAPAS simply relies on a continuous map of degree or probability of lesion at each voxel in the brain. Maps are generated by an automatic segmentation algorithm in order to predict a subject-level threshold for segmentation. In our experiments, we used the predicted lesion probability maps from a Method for Inter-Modal Segmentation Analysis (MIMoSA) (Valcarcel et al., 2018a,b), an automatic segmentation procedure. We also implemented the lesion prediction algorithm (LPA) (version 2.0.15) using the lesion segmentation tool (LST), an open source toolbox for statistical parametric mapping (SPM) (version 12) in MATLAB R2019a (Schmidt et al., 2012). In section A.3 of the Appendix, we provide results obtained from using LST-LPA as the automatic segmentation algorithm.



Subject ID	τ_i	$Volume_i(\hat{\tau}_{Group})$
1	0.23	15.1
2	0.31	20.3
⋮	⋮	⋮
$N/2$	0.09	3.7

Figure 3.2: The TAPAS procedure is shown using sample axial slices from the data. A set of training scans with manual delineations were used to train and apply MIMoSA in order to obtain probability maps. For each subject's probability map, we applied thresholds at $\tau = 0\%$ to 100% by 1% to create estimated lesion masks. For simplicity, in this example, we have only shown $\tau = 10\%$, 50%, and 90%. Based on Sorensen-Dice similarity coefficient (DSC) calculations within and across subjects we estimated $\hat{\tau}_i$ and $\hat{\tau}_{Group}$. Using $\hat{\tau}_{Group}$ we obtained $volume_i(\hat{\tau}_{Group})$. We fit the TAPAS model and applied it to subjects in the test set to determine $\hat{\tau}_i$. Red points in the plot represent $\hat{\tau}^{0.1}$ and $\hat{\tau}^{0.9}$, or lower and upper bounds at the volume associated with the 10th and 90th percentiles, respectively.

We first divide the data set under study into two parts: the first is used for training TAPAS, and the second we refer to as the test set. In each subject in the training set of size $N/2$, we apply a grid of thresholds $\tau \in \{\tau_1, ..., \tau_J\}$, denoted as $\boldsymbol{\tau}$, to the probability map in order to generate estimated lesion segmentation masks. The estimated lesion segmentation masks are binary masks indicating estimated lesion presence or absence generated for each threshold in $\boldsymbol{\tau}$. Figure 3.2 shows an example of these lesion masks at 10%, 50%, and 90%. For each subject in the training set we initially let $\boldsymbol{\tau}$ vary from $\tau_1 = 0\%$ to $\tau_J = 100\%$ in 1% increments and calculate DSC between each estimated segmentation mask and the corresponding manual segmentation for the image. We then estimate:

1. $\hat{\tau}_{Group} = \underset{\tau \in \{\tau_1, ..., \tau_J\}}{\arg\max} \dfrac{2\sum_{i=1}^{N/2} DSC_i(\tau)}{N}$, and
2. $\hat{\tau}_i = \underset{\tau \in \{\tau_1, ..., \tau_J\}}{\arg\max} \{DSC_i(\tau)\}$ for each subject i.

The threshold estimated by $\hat{\tau}_{Group}$ represents the threshold that produces maximum average DSC across all subjects in the training set, and $\hat{\tau}_i$ is defined as the subject-specific threshold that yields maximum DSC for subject i. In practice, we suggest initially using a threshold grid of $\tau_1 = 0\%$ to $\tau_J = 100\%$ in 1% increments but based on training refine the grid to be more sensitive to the data.

In the event of a tie among thresholds that maximize DSC we first ensure these tied thresholds are adjacent and then select the median threshold. In our analyses all ties were in fact adjacent. If ties are not adjacent, we suggest enlarging the threshold region and repeating the analysis. In addition, we repeat the optimization minimizing absolute error (AE) rather than maximizing DSC since DSC can be biased for patients with low lesion load. These results are presented in section A.1 of the Appendix. It is also possible this step could be implemented using an optimization framework and may result in a reduction in computation time, but we did not validate other optimization approaches.

We apply $\hat{\tau}_{Group}$ to each respective subject and obtain a naive estimate of the volume, $volume_i(\hat{\tau}_{Group})$. We then regress $logit(\hat{\tau}_i)$ on $volume_i(\hat{\tau}_{Group})$ using a generalized additive model with an identity link function and a normal error. The generalized additive model was chosen over linear models after manual inspection of scatter plots indicated non-linear trends. This is evident in the scatter plot displayed in the bottom left panel of Figure 3.2 as the scatter plot presented in this example case does not appear linear but quadratic. This held true for not just this example case but most cross-validation iterations. We use an identity link function since both $\hat{\tau}_i$ and $volume_i(\hat{\tau}_{Group})$

are continuous. The identity link does not bound the outcome $\hat{\tau}_i$ between 0 and 1; so, rather than modeling $\hat{\tau}_i$, we model $logit(\hat{\tau}_i)$ to force $\hat{\tau}_i$ to be between 0 and 1. We implement the generalized additive model using the *gam* function available through the *mgcv* package in R. This function fits the model using a penalized scatter-plot smoother with thin-plate splines and smoothing parameter estimated using generalized cross-validation (*Generalized Additive Models*; Wood, 2003, 2004; Wood, Pya, and Sfken, 2016). More specifically, the following generalized additive model is fit as the TAPAS model:

$$logit(\hat{\tau}_i) = f_1(volume_i(\hat{\tau}_{Group})) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

In the model fitting procedure, we exclude subjects from model training if their $\hat{\tau}_i$ produces an estimated segmentation mask with $DSC < 0.03$. We found this to empirically improve TAPAS performance as it removes subjects for which even the best performing $\hat{\tau}_i$ yields an inaccurate automatic segmentation mask.

After the TAPAS model is fit, we apply the model to subjects in the testing set. For each subject i, we obtain a probability map from an automatic segmentation procedure. We then use $\hat{\tau}_{Group}$ to threshold the probability map in order to estimate $volume_i(\hat{\tau}_{Group})$. We use these predicted volumes in the TAPAS model to estimate the fitted value $logit(\hat{\tau}_i)$, from which we can obtain the estimated subject-specific threshold. The probability maps are then re-thresholded using $\hat{\tau}_i$ to generate the lesion segmentation masks.

When applying the TAPAS model in the testing set, we aim to reduce extrapolation and excessive variability associated with left and right tail behavior of the spline model. Thus, for any volume we obtain using $\hat{\tau}_{Group}$ that is larger than the volume at the 90th percentile, we use the threshold for the subject whose volume is at the 90th percentile, denoted $\hat{\tau}^{0.9}$, rather than the fitted $\hat{\tau}_i$. Similarly, for any volume we obtain from $\hat{\tau}_{Group}$ that is smaller than the volume at the 10th percentile, we use the value of $\hat{\tau}^{0.1}$. Figure 3.2 shows an outline of the full TAPAS procedure and model.

To implement TAPAS, we developed an R package that is available with documentation on GitHub (`www.github.com/avalcarcel9/rtapas`) and Neuroconductor (https://neuroconductor.org/package/rtapas).

3.2.3. Performance Assessment

For the two data sets in this study (JHH and BWH), we ran separate Monte Carlo-resampled split-sample cross-validations. More specifically, we repeatedly randomly sampled subjects (100 times) without replacement to assign half of the subjects in the study to each of the training and testing sets. Each iteration therefore contained a unique set of subjects to train TAPAS and a separate set of subjects to test the algorithm's performance. The Monte Carlo-resampled split-sample cross-validation analysis assures that the proposed algorithm does not provide significantly different lesion volume estimations when different trained regression models are used. In each training set, we applied MIMoSA using the R package *mimosa* (Valcarcel, 2018) available on Neuroconductor (https://neuroconductor.org/package/mimosa) ("Neuroconductor"). After fitting the MIMoSA model using subjects in the training set, we generated probability maps for all subjects in the training and testing sets.

In each split-sample experiment, the training set was used to fit the TAPAS model and the testing set applied the TAPAS model to determine a subject-specific threshold $\hat{\tau}_i$. This subject-specific threshold was used to create binary lesion segmentation masks and calculate lesion volume. In the BWH data, we found using a threshold grid ranging from $\tau_1 = 0\%$ to $\tau_J = 100\%$ in 1% increments to be too wide in initial experiments. Therefore, we refined the threshold grid range from $\tau_1 = 13\%$ to $\tau_J = 54\%$ in 0.4% increments. We compared the TAPAS, group, and manually generated masks and volumes using the subscripts $TAPAS$, $Group$, and $Manual$ respectively. The use of $\hat{\tau}_{Group}$ to threshold probability maps and generate lesion segmentations was previously applied (Valcarcel et al., 2018a,b) and aided in automatic segmentation measures compared to user-defined threshold application. In addition to calculating volume from TAPAS, group, and manual lesion masks we also calculate partial volume denoted with the subscript $Partial$. We define partial volume as the sum of the voxel level probabilities from the probability map generated by MIMoSA. Calculating partial volume does not require thresholding. Rather than applying a hard threshold to estimate lesion volume, we hypothesize that it may be more advantageous to compute total lesion burden using this continuous measures from probability maps. These partial volumes may yield stronger correlations with clinical outcomes.

We provide quantitative comparisons between TAPAS and the group thresholding procedure for subjects in the testing set. First, to assess whether segmentation masks produced using TAPAS or the group thresholding procedure differed in accuracy as measured by DSC, we compared segmentations between lesion masks produced by TAPAS (DSC_{TAPAS}) and those produced by the group thresholding procedure (DSC_{Group}) with manual segmentations. We compared these measures using a paired t-test within each split-sample experiment using subjects in the test set. Second, to assess bias and inaccuracy present in $volume_{TAPAS}$ and $volume_{Group}$ we calculated absolute error defined as $AE = |Threshold\,Volume - Manual\,Volume|$. In order to determine whether AE differed statistically, paired t-tests were conducted between AE_{TAPAS} and AE_{Group} within each split-sample experiment. Third, to adjudicate whether TAPAS yielded volumetrics with similar phenotype associations, we calculated the Spearman's correlation coefficient between $volume_{TAPAS}$, $volume_{Group}$, $volume_{Partial}$, and $volume_{Manual}$ and clinical variables. We denote these correlations by $\hat{\rho}_{TAPAS}$, $\hat{\rho}_{Group}$, $\hat{\rho}_{Partial}$, and $\hat{\rho}_{Manual}$, respectively. We estimated correlations in each split-sample experiment and averaged across experiments.

3.2.4. Expert validation

In addition to the Monte Carlo-resampled split-sample cross-validations, 3 board-certified neurologists with subspecialty training in neuroimmunology compared segmentations produced using TAPAS and the group thresholding approach. For each subject (40 subjects from BWH data and 94 subjects from JHH data), we randomly selected a cross-validation iteration in which they were included as a test set subject and therefore have segmentations produced from TAPAS and the group thresholding procedure to present to the raters. We randomly assigned the order in which the subjects were presented to the expert rater. Additionally, we randomly assigned each segmentation a letter (A or B) so as to blind the rater to the segmentation algorithm.

We presented each of the 134 MRI studies to the experts individually. For each study, the expert rater was presented with the set of two segmentations overlaid onto the FLAIR along with each of the MRI contrasts simultaneously. For BWH data this included FLAIR, T1, and T2 imaging modalities, while for the JHH data this included FLAIR, T1, T2, and PD imaging modalities. The expert was then asked, ”Evaluate how well each of the two segmentations depicts your impression of the extent of the white matter abnormality in the image displayed.” Ratings were given on a scale of 1-to-5 scale with labels of ”1 - Excellent”, ”2 - Good”, ”3 - Fair”, ”4 - Poor”, ”5 - Very Poor”. Ratings

Table 3.1: Demographic information for subjects in this study are provided. We include information from 94 patients imaged at Johns Hopkins's Hospital (JHH) and 40 patients imaged at the Brigham and Women's Hospital (BWH).

	Mean	Std. Dev.	Min	Max
JHH				
Age (years)	43.4	12.3	21.4	67.3
Disease duration (years)	11.3	9.2	0.0	45.0
Expanded Disability Status Scale score	3.9	2.1	0.0	8.0
Lesion volume (mL)	11.5	13.1	0.0	77.0
BWH				
Age (years)	50.4	9.9	30.4	69.9
Disease duration (years)	14.5	4.6	3.8	21.3
Expanded Disability Status Scale score	2.3	1.6	0.0	7.0
Lesion volume (mL)	13.6	12.8	0.6	52.0
Timed 25-ft walk (seconds)	11.5	6.9	1.0	25.0

	%
JHH	
Female	73
Male	27
Clinically isolated syndrome	1
Primary progressive MS	10
Relapsing-remitting MS	64
Secondary progressive MS	26
BWH	
Female	70
Male	30
Relapsing-remitting MS	80
Secondary progressive MS	20

were given independently, with no discussion by raters occurring during the rating process.

3.3. Results

3.3.1. Demographics

JHH and BWH participant demographics are included in Table 3.1. In the JHH data, disease duration was defined as years since diagnosis and participants were examined by a neurologist to assess Expanded Disability Status Scale (EDSS) score. In the BWH data, disease duration was defined as years since first symptoms. In order to assess the level of physical ability and ambulatory function in the BWH data, an MS neurologist examined patients to evaluate Expanded Disability Status Scale (EDSS) and timed 25-foot walk (T25FW) (in seconds).

3.3.2. Volumetric bias assessment

Using Bland-Altman visualization, we compare automatic and manual volumes in addition to the partial volume in Figure 3.3. Subject-level volumes were obtained by averaging each subject's measurement for all split-sample experiments in which it was allocated to the testing set. The JHH data $volume_{Group}$ estimate exhibits systematic bias, evident in Figure 3.3, for volumes exceeding 20 mL. Visually, we observed a moderate inverse relationship in these subjects. This indicates that $volume_{Group}$ under-estimates $volume_{Manual}$ in subjects with larger lesion loads with increasing magnitude. The JHH data $volume_{Partial}$ estimate also exhibits systematic bias using Figure 3.3. For subjects with small lesion load $volume_{Partial}$ over-estimates $volume_{Manual}$ whereas for subjects with moderate and large lesion load $volume_{Partial}$ under-estimates $volume_{Manual}$. Unlike the Group Bland-Altman plot, the TAPAS plot does not exhibit obvious patterns of systematic bias. The cluster of points that begins to negatively deviate from 0 in the Group plot is still centered randomly around 0 in the TAPAS plot. Additionally, the mean and standard deviation for the differences are smaller using $volume_{TAPAS}$ compared to $volume_{Group}$ and $volume_{partial}$. There are four points that lie outside the limits of agreement in both thresholding procedures, but in the TAPAS plot these are closer to 0.

The BWH Bland-Altman plots are nearly identical and almost indistinguishable when comparing the group threshold procedure with the TAPAS outputs. There does not appear to be a systematic bias in either $volume_{Group}$ or $volume_{TAPAS}$ estimates since points are randomly scattered around 0 in the positive and negative directions. This exemplifies TAPAS's propensity to conserve unbiased estimates when systematic bias is absent. The Bland-Altman plot calculated using $volume_{Partial}$ shows all points lie within the limits of agreement but they are not randomly scattered around the mean difference. For small lesion loads, the points cluster above the mean line and show a negative association as in the JHH data.

3.3.3. Absolute error assessment

Scatter plots and their corresponding predicted linear models are presented in Figure 3.4 to compare AE_{TAPAS}, and AE_{Group}, and $AE_{Partial}$ with $volume_{Manual}$. The JHH data plot shows smaller AE estimates associated with $volume_{TAPAS}$ compared to $volume_{Group}$ and $volume_{Partial}$. This is highlighted by the negative shift in AE_{TAPAS} points throughout as well as a smaller slope estimate

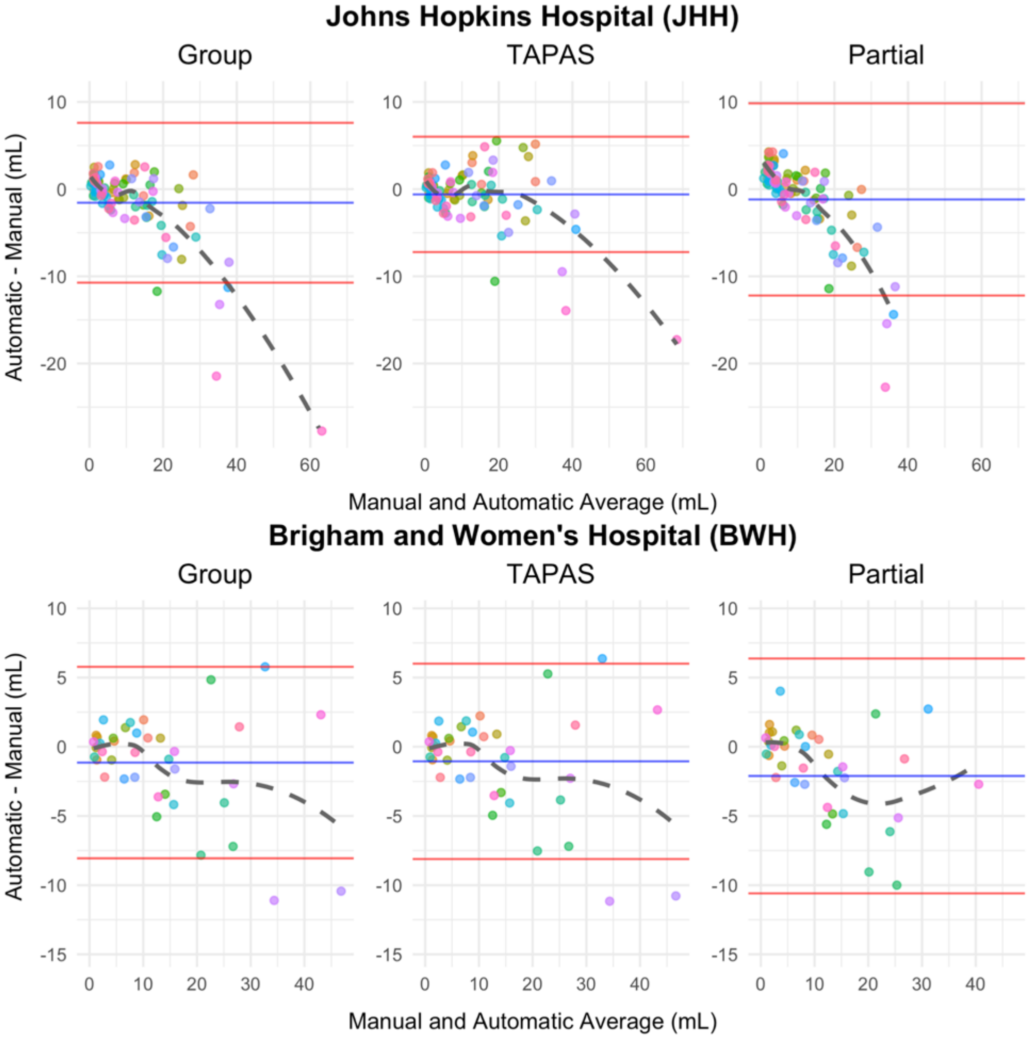


Figure 3.3: Bland-Altman plots comparing $volume_{Manual}$ with volumes obtained using automatic thresholding approaches ($volume_{Group}$, $volume_{TAPAS}$, and $volume_{Partial}$) are shown. The mean of the difference in volume is presented in blue and the mean plus and minus the standard error is shown in red. Each point represents a unique subject. Subject-specific points were obtained by averaging results across test set subjects in each split-sample fold.

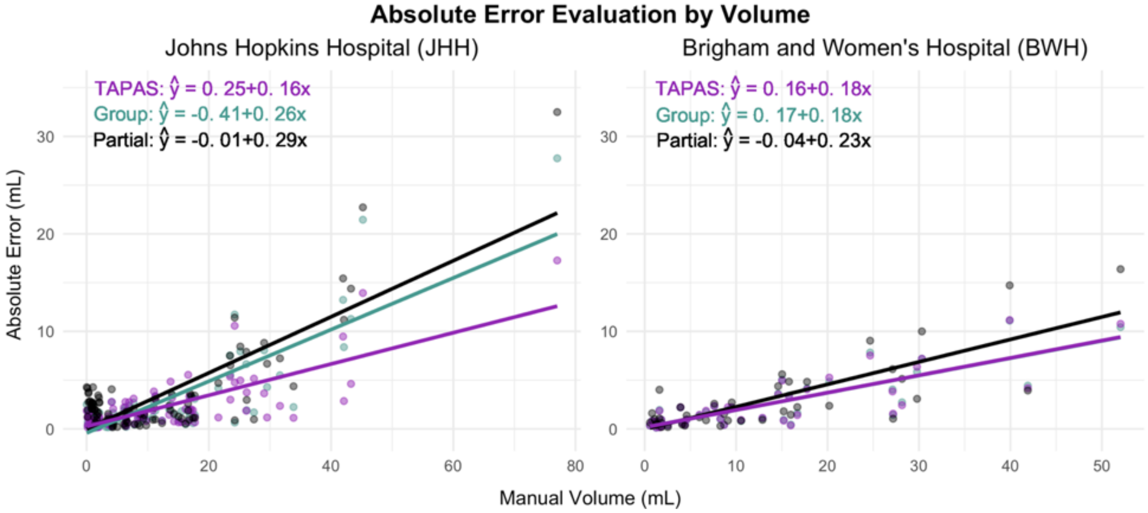


Figure 3.4: Scatter plots with fitted linear models are presented for the subject-level average absolute error ($\hat{y}$) on manual volume (x) in mL. Fitted equations are given in the top left corner.

(provided in the top left corner of the figure). The JHH data points with $volume_{Manual}$ larger than 70 mL are influential for both the group, TAPAS, and partial fitted models. However, removing these points, $volume_{TAPAS}$ still shows larger reductions in AE compared to $volume_{Group}$. The coefficient associated with AE_{Group} is 0.26 while the coefficient associated with AE_{TAPAS} is 0.16. This means that for a 1 mL increase in $volume_{Manual}$, the predicted change in AE is 0.1 mL less when using TAPAS compared to the group thresholding procedure. The reduction in AE associated with using TAPAS over the group thresholding procedure is on the order of magnitude of average differences found in clinical trial evaluations of MS therapies (see, for example, Barkhof et al., 2007). In the BWH data, all values are remarkably similar across TAPAS and the group thresholding approach. The partial volume leads to notably larger predicted absolute error. The results in Figure 3.3 and Figure 3.4 are consistent and indicate that TAPAS performs at least as well as or better than the group thresholding procedure in terms of reducing bias in lesion volume estimates.

Comparing the two thresholding approaches more rigorously we found the average AE_{TAPAS} across subjects in the testing sets and iterations in the JHH data is 2.09 mL compared to 2.62 mL from AE_{Group} and 3.29 mL from $AE_{Partial}$. In the BWH data, average AE_{TAPAS} and AE_{Group} were both found to be 2.62 mL and average $AE_{Partial}$ was 3.17 mL. TAPAS yields equal or reduced average AE. The average DSC_{TAPAS} across subjects in the testing sets and iterations in the JHH data is 0.61 compared to 0.6 from DSC_{Group}. In the BWH data, the average DSC_{TAPAS} is 0.67

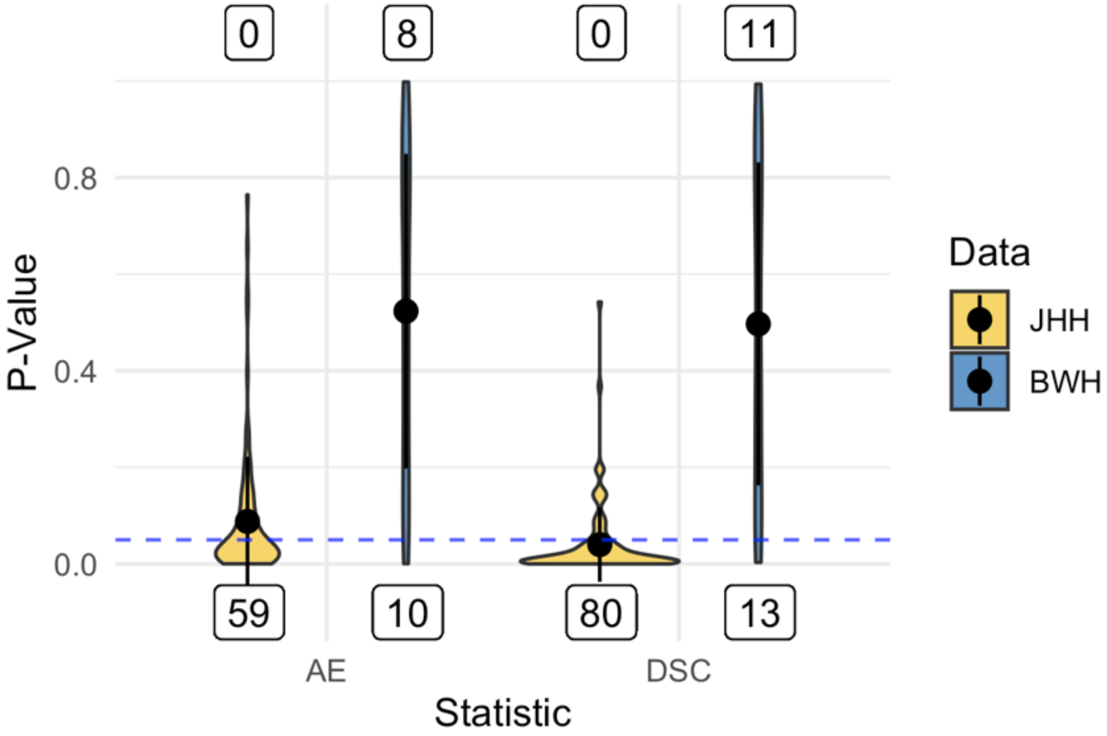


Figure 3.5: Violin plots of p-values from paired t-tests to compare subject-level absolute error (AE) and Sorensen-Dice coefficient (DSC) in each test set are presented. The mean for each statistic and data set is presented as points within each violin plot and the black lines extend the mean by the standard deviation. Labels below represent the number of significant p-values favoring TAPAS performance measures. Labels above represent the number of significant p-values favoring group thresholding performance. The dashed horizontal blue line highlights the $\alpha = 0.05$ cutoff.

while average DSC_{Group} is 0.66. TAPAS yields equal or superior average DSC. We do not report $DSC_{Partial}$ as the partial volume is calculated from the probability maps rather than the lesion segmentation masks and binary segmentations are required to calculate DSC.

To examine this statistically, we employed one-sided paired t-tests to evaluate AE and DSC from TAPAS compared with those obtained from the group thresholding procedure. Figure 3.5 shows violin plots of p-values from both sets of tests for the two data sets. In the JHH data more than half of the split-sample experiments resulted in p-values below the $\alpha = 0.05$ for AE and DSC with no statistically significant results favoring the group thresholding procedure. This indicates superior performance using TAPAS compared to the group thresholding procedure. The BWH data was more uniform with approximately equal statistically significant results favoring TAPAS and the group thresholding procedure.

Table 3.2: Subject-specific volume estimates, $volume_{Manual}$ (Manual), $volume_{TAPAS}$ (TAPAS), $volume_{Group}$ (Group), and $volume_{Partial}$ (Partial), were compared with clinical covariates available from the data collected at the Johns Hopkins Hospital (JHH) and the Brigham and Women's Hospital (BWH) and are represented in this table. Spearman's correlation coefficient ($\hat{\rho}$) was obtained in the testing set for each iteration and averaged across folds. Clinical variables included Expanded Disability Status Scale (EDSS) score, disease duration in years, and timed 25-ft walk (T25FW) in seconds.

	Estimates for $\hat{\rho}$			
	Partial	**Group**	**TAPAS**	**Manual**
JHH				
EDSS	0.32	0.34	0.34	0.29
Disease duration	0.37	0.39	0.39	0.39
BWH				
EDSS	0.42	0.43	0.43	0.45
Disease duration	0.31	0.32	0.32	0.29
T25FW	0.02	0.02	0.02	0.03

3.3.4. Correlation analysis

We assessed the relationship between $volume_{TAPAS}$, $volume_{Group}$, $volume_{Partial}$, and $volume_{Manual}$ with various clinical variables. These results are provided in Table 3.2. All correlations found are modest but align with previously published literature (Barkhof, 1999; Stankiewicz et al., 2011; Tauhid et al., 2014; Valcarcel et al., 2018a). In the JHH data, $\hat{\rho}_{TAPAS}$ and $\hat{\rho}_{Group}$ are indistinguishable from each other and slightly larger than $\hat{\rho}_{Partial}$ $\hat{\rho}_{Manual}$. Similarly, the BWH data show identical $\hat{\rho}_{TAPAS}$ and $\hat{\rho}_{Group}$ nearly equivalent to $\hat{\rho}_{Partial}$ and $\hat{\rho}_{Manual}$. In terms of phenotypic associations $volume_{TAPAS}$ yielded similar correlation estimates as $volume_{Group}$, $volume_{Partial}$, and $volume_{Manual}$.

3.3.5. Threshold evaluation

In Figure 3.6 we present scatter plots of the thresholds predicted in the testing set from both TAPAS and the group threshold procedure. There are a few notable differences between the threshold scatter plots produced from TAPAS and those produced by the group thresholding procedure. In both data sets the subject-specific thresholds have a much wider range than the group thresholds. In the JHH data, the distribution shape is bi-modal for the subject-specific thresholds but uni-modal for the group thresholds. In the BWH data, the distribution shape is similar between the two thresholding approaches.

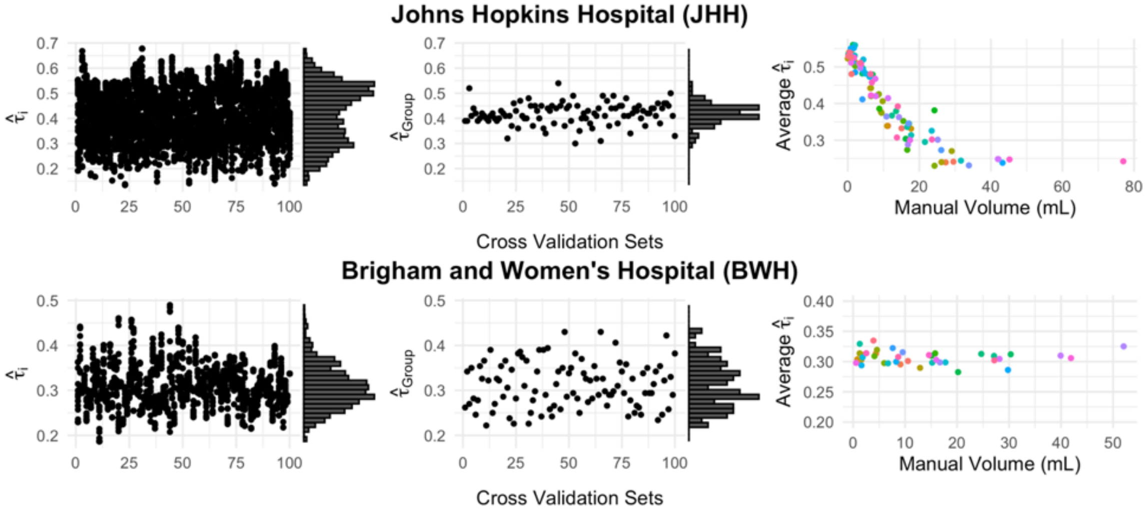


Figure 3.6: Scatter plots of the subject-specific threshold $\hat{\tau}_i$ (TAPAS) and $\hat{\tau}_{Group}$ (group thresholding procedure) on cross-validation number are presented with marginal histograms for both data sets in the first two columns. The third column presents scatterplots of the average subject-specific thresholds from TAPAS and the manually delineated lesion volume.

We also present average subject-specific thresholds plotted against the manual volumes in mL. The JHH data show that as manual volume increases the average TAPAS threshold also decreases. The thresholds plateau after manual volume of 20 mL and similar thresholds are detected for all lesion loads greater than 20 mL. In the BWH data we see the points are randomly scattered and there is no pattern between average subject-specific threshold and manual volume in mL.

3.3.6. Qualitative results

We present segmentations from the TAPAS and the group threshold approach as well as manual delineations in Figure 3.7. This figure shows that TAPAS and the group thresholding procedure generally agree with the manual segmentation. Some tissue was manually segmented and not detected by either thresholding algorithm. The major differences between all the methods are found at the boundaries of lesions, which are known to be difficult to discern for both automatic and manual approaches. Overall, the automatic segmentation algorithm paired with either thresholding approach is able to detect the majority of lesional space with few false positives.

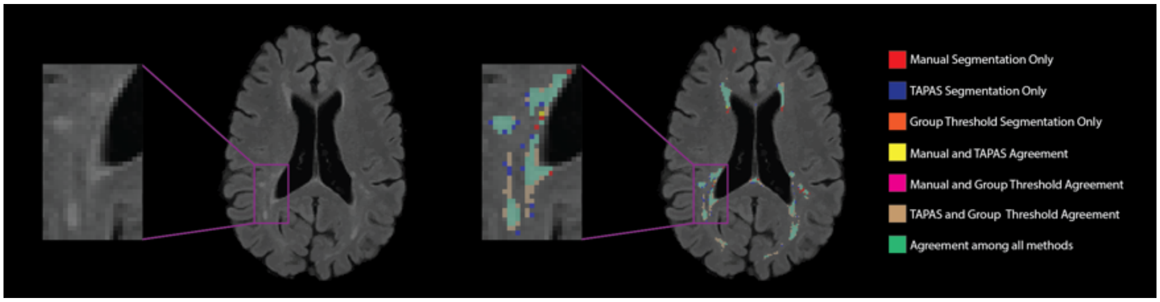

Manual Segmentation Only
TAPAS Segmentation Only
Group Threshold Segmentation Only
Manual and TAPAS Agreement
Manual and Group Threshold Agreement
TAPAS and Group Threshold Agreement
Agreement among all methods

Figure 3.7: T2 hyperintense lesion segmentations from an example axial slice are displayed. The colors represent the different individual or overlapping segmentations obtained from manual, TAPAS threshold, and group threshold masks. The majority of segmented area was in agreement among all lesion masks (green). Both the group thresholding approach and TAPAS missed some area that was manually segmented (red). There was a small area where only TAPAS and manual segmentations agreed (yellow), but almost no area where only the group threshold agreed with the manual segmentation (fuchsia).

3.3.7. Rater study

The mean rating for TAPAS segmentations for each rater was 1.87 (SD=0.81), 2.72 (SD=0.94), and 3.10 (SD=1.14). The mean rating for the group thresholding approach for each rater was 1.92 (SD=0.81), 2.66 (SD=0.97), and 3.10 (SD=1.14). The mean rating across the three raters for both TAPAS and the group thresholding approaches was 2.56 (SD = 1.10). Raters evaluated how well each of the two segmentations depicted the extent of the white matter abnormality in the images displayed. An overall average score between 2 and 3 indicated therefore that the segmentations produced from either method are between fair and good quality. The three raters responded favorably to the segmentations.

77% of the studies resulted in the same rating between TAPAS and the group threshold segmentations. 12% of the studies resulted in raters ranking the TAPAS segmentation more favorably than the group threshold segmentation whereas 11% of the studies resulted in raters favoring group threshold segmentation.”

Though both thresholding approaches were trained using manual segmentations, the gold standard approach, we and our expert raters believe the resulting segmentations from the automatic approaches do in fact capture the extent of white matter abnormality in the brain fairly well.

3.3.8. Computation time

The TAPAS thresholding procedure is easily implemented using the *rtapas* R package available with documentation on GitHub `www.github.com/avalcarcel9/rtapas` and Neuroconductor `https://neuroconductor.org/package/rtapas`. The model is supervised and must be trained. All benchmarking was done on a 2017 MacBook Pro with 3.1 GHz Intel Core i5 and 16GB of memory using a single core. To benchmark, a single subject with voxel size $1\ mm^3$ was used. Before training the TAPAS model, the training data must be generated and takes approximately 20 minutes per subject. This process is parallelizeable through the package to decrease computation time. The model itself takes less than a second to train. After a model has been fit, a single test subject's prediction data and segmentation mask can be generated in about 30 seconds.

3.4. Discussion

Most automatic segmentation algorithms produce continuous maps of lesion likelihood, which are subsequently thresholded to create binary lesion segmentation masks. While a number of automatic approaches exist for lesion segmentation, there are few automatic algorithms available for threshold selection. Thresholds are commonly chosen using cross-validation procedures conducted at the group level, or arbitrarily through subjective human input. This introduces variability and biases in automatic segmentation results. Furthermore, thresholding approaches often apply a single common threshold value to all subjects' probability maps. This lack of subject specificity may lead to inaccuracy in lesion segmentation masks, especially in subjects with the smallest and largest lesion loads.

This study sought to address these issues by introducing a supervised fully automated algorithm for subject-specific threshold prediction that also reduces volumetric bias if present. The TAPAS procedure is easily implemented and performs well on data acquired with different scanning protocols or pre-processed with different pipelines. We validated TAPAS in two unique data sets from different imaging centers using 3T MRI scanners from different vendors. In section A.2 of the Appendix we applied a different preprocessing pipeline to the JHH data and found TAPAS outperforms the group thresholding procedure even under varying processing.

The TAPAS procedure is a supervised fully automated thresholding approach that determines a subject-specific threshold to apply to continuous maps (including predicted probability maps) for automatic lesion segmentation. TAPAS volume estimates are accurate and reduce systematic biases associated with differential total lesion load when present. In the JHH data, we observed such a bias using the MIMoSA algorithm, which was mitigated using TAPAS.

The BWH data used a consensus approach with two trained raters to manually segment lesions consulting a third rater in the event of a disagreement. We believe this approach reduces intra- and inter-rater variability normally present with a single rater and allows for a closer approximation of the ground truth, and, thus, better training of automatic approaches. The Bland-Altman plots in these data indicate unbiased estimation using a group threshold or TAPAS. In this study, we showed that without systematic biases TAPAS preserves the unbiased volumetric estimation of the automated segmentation technique.

In clinical trial evaluations of therapeutic efficacy, associations between clinical variables and lesion volume are of primary interest. This study shows that TAPAS and group threshold volumes resulted in similar correlations to clinical variables as the manual volume. Therefore, the automatic segmentations produced after thresholding, using either TAPAS or group thresholding, should be as sensitive to image-phenotype correlations as manual measures. Correlations were thus estimated to compare $volume_{Manual}$, $volume_{TAPAS}$, and $volume_{Group}$ with clinically relevant variables. The results indicate the correlations between respective volumes and clinical variables are all approximately equal. Agreement across the thresholding methods with manual measures advocates for the use of TAPAS to reduce cost and time while providing a subject-specific threshold.

Currently, available assessments of lesion volume are weakly correlated with clinical outcomes. This may be in part due to discarding voxels with low estimated probability of containing lesion, mostly around the edges of lesions, that may capture signal. The partial volume computed in this analysis was an attempt to include these voxels in the calculation of volume in the hopes of reducing biases and providing a metric that correlates better with clinical assessments. Unfortunately, these partial volumes did not yield stronger correlations with clinical outcomes and showed more bias compared to volumes computed with a threshold. These methods have not been assessed in clinical trials to date, and additional studies and methodological innovations are warranted.

TAPAS is a post-hoc subject-specific threshold detection algorithm built to reduce volumetric bias associated with automatic segmentation procedures. In this study, we optimized TAPAS using DSC in this Chapter and AE in section A.1 of the Appendix provided. Both optimizations favor TAPAS over group thresholding with DSC having more dramatic improvements than AE. Though DSC can be biased or under-estimate true accuracy in subjects with low lesion load, we find it performs well compared to AE. Automatic approaches are constantly being built and improved upon to yield more accurate and robust methods. TAPAS allows for improvement upon even the most accurate and robust automatic segmentation procedures with no observed addition of error. Beyond MS or MRI, this methodology can be used for automatic segmentation of other tissues or body parts using different imaging types after proper validation.

We initially ran all cross-validation settings with a threshold grid ranging from 0% to 100% in 1% increments. In certain settings, these increments were too large which led to sub-optimal TAPAS models. We refined threshold grids in these settings and found improved performance. Due to the iterated nature of cross-validations, we chose to use one threshold grid for the entire set of cross-validation folds (100). In practice, data will likely consist of a training and testing set. We suggest applying the original threshold grid, 0% to 100% by 1% increments, and evaluating model fit through subject-specific threshold selection in the training and testing data in order to inform the selection of a finer grid. The grid should be updated until results are stable. We believe this will lead to optimal performance.

There are several notable limitations to the proposed algorithm. First, the method must be used in conjunction with continuous maps of likelihood of lesion, so investigators must use automatic approaches that generate these maps for adaptive thresholding. Second, since the TAPAS model fits a generalized additive model, training data sets with small sample size, uniform lesion load, or those dissimilar from testing data may result in poor model fit or inappropriate threshold estimation. For example, when we applied TAPAS to the 2015 Longitudinal Lesion Challenge data we found poor model fit associated with fitting a generalized additive model to data that only included 5 unique subjects for training. To apply TAPAS to longitudinally acquired data, such as those presented in the 2015 segmentation challenge, a sufficiently large sample of subjects with variable lesional volume is required.

Atrophy of the brain and spinal cord are key measures of disease progression in MS and may be more closely associated with clinical status than lesion volume (Bakshi et al., 2008; Fisher et al., 2008; Fisniku et al., 2008; Keshavan et al., 2016; Sanfilipo et al., 2006). It is important to note that TAPAS is easily extended or applied to settings in which brain volumes are estimated. Many segmentation methods for structures other than lesions, for example the thalamus which is of key interest in MS presently (Fadda et al., 2019; Neema et al., 2009; Oh et al., 2019), also use thresholding to determine binary segmentations and volumes. Future work will include assessments of whether biases such as those studied in this paper exist for atrophy assessments and techniques for their mitigation.

Future developments will include specialized methods for the analysis of longitudinal lesion volumetrics. Additionally, to investigate the repeatability of this study and stability of the algorithm we will implement the method on scan-rescan data to evaluate reliability of the subject-specific probability and lesion volume estimation. It is possible that the underlying method may benefit from dynamic thresholds for smaller lesions and larger lesions even within the same subject. That is, we may need to move beyond even a subject-specific threshold since, when a subject has larger lesions, the error associated with those lesions contributes more to the DSC metric than the same relative error associated with smaller lesions. There may thus be a tendency of TAPAS to better segment larger lesions at the cost of doing worse on smaller lesions.

CHAPTER 4

MEMORY EFFICIENT COMPUTATIONAL TOOLS FOR IMAGE ANALYSIS IN R

4.1. Introduction

Common MRI acquisition protocols include imaging at 1.5 tesla (1.5T) and 3 tesla (3T). A typical image at 1.5T and 3T consumes approximately 10 and 85 megabytes, respectively, in the R environment. Unfortunately, R is not inherently well suited for big data sets. Medical images, depending on the image dimension and voxel size, can require extensive memory resources (Kang, Caffo, and Liu, 2016). Carrying out image analysis as sample size or number of scans increases becomes challenging even on well-equipped hardware. A standard 2020 MacBook Pro comes with 16 gigabytes of memory. This powerful laptop can handle approximately 1600 images acquired with a voxel resolution of 2mm compared to only 190 images acquired at 1mm. High-performance computing clusters offer more memory than a laptop but may still be insufficient for image analysis of studies with large sample size and multi-modal data. Other computing languages such as C/C++ or Fortran allow for fast and memory-efficient operations on large data. Unfortunately, these languages lack flexibility, are not well-suited for interactive data exploration, and cannot access R's rich package environment required for image analysis.

Further, recent advancement of 7 tesla (7T) imaging has led to increased installation of 7T MRI scanners around the world for both clinical and research applications (Rutland et al., 2020). While still new, 7T scanners are the future of MRI. Imaging on 7T produces high contrast images with exceptional resolution and detail, but the increased precision comes at the cost of larger size and memory. Imaging at 7T will bring resolutions as high as 0.65mm isotropic to standard imaging studies. Table 4.1 provides a comparison of 3D MRI information for images at varying resolutions acquired at 1.5T, 3T, and 7T. This table presents data from a typical image acquired on respective MRI scanners. Images acquired at 7T with 0.65 mm voxel sizes consume substantially more memory both on on-disk and in R compared to 1.5T and 3T. On the same 2020 MacBook Pro with 16 gigabytes of memory only approximately 70 7T images can be loaded into memory.

Due to the high memory consumption of images in R, it becomes difficult to perform simple operations voxel-wise across subjects. For example, calculating the median at the voxel-level across

Table 4.1: 3D MRI information for images at aquired 1.5 tesla (1.5T), 3 tesla (3T), and 7 tesla (7T) are provided. On-disk size was calculated using compressed images.

	1.5T	3T	7T
Pixel Dimension (millimeter)	2 × 2 × 2	1 × 1 × 1	0.65 × 0.65 × 0.65
Image Dimension	94 × 124 × 94	240 × 256 × 176	256 × 312 × 384
On-Disk Image Size (Megabyte)	3	5	20
R Image Size (Megabyte)	10	85	235

1000 images becomes computationally intense since it may not be possible to load all 1000 images into memory at once depending on the acquisition protocol and computational resources. Acquisition of MRI images at 1.5T and 3T in both clinical and reseach settings has become more common and accessible leading to multi-modal imaging studies with large sample sizes. Analyses of data from these studies face memory problems in R and statistical analysis is difficult. While 7T imaging is new and scanning on these machines is limited thusfar, the data size is much larger than 1.5T and 3T. As 7T imaging becomes more commonplace, memory, even in small studies, will quickly be a concern. Therefore, there is a need for more advanced computational tools for image analysis in R.

4.2. Materials and methods

The issues associated with limited memory are also faced in the statistical analysis of genomics data. Bioconductor provides tools for the analysis of genomic data and has existing tools for memory efficient statistical analysis of large genomic data in R (Gentleman et al., 2004). `DelayedArray` is an R package currently hosted on Bioconductor (Pages, Hickey, and Lun, 2020). `DelayedArray` allows common array operations on an object without loading it into memory. In order to reduce memory usage and optimize performance, operations on the object are either delayed or executed using a block processing mechanism. `DelayedMatrixStats` is an R package currently hosted on Bioconductor (Hickey, 2020). `DelayedMatrixStats` contains functions for statistical calculations (i.e. row or column median calculation) using `DelayedArray` efficient block processing on large matrices while keeping local memory usage low. `DelayedMatrixStats` builds on both the `DelayedArray` and `matrixStats` packages to allow for high-performing functions operating on rows and columns of objects of class DelayedMatrix (Bengtsson, 2019). The functions are optimized by data type and for subsetted calculations such that both memory usage and processing time are minimized.

We developed an R package `NiftiArray` to overcome big data limitations in imaging (Muschelli and Valcarcel, 2020). The `NiftiArray` package allows for fast random access of imaging data in NIfTI format and is compatible with existing software for performing common statistical operations without loading data objects into memory (Hickey, 2020; Pages, 2020; Pages, Hickey, and Lun, 2020). The package establishes the NiftiArray class, a convenient and memory-efficient array-like container for on-disk representation of NIfTI images. The NiftiArray class is an extension of the HDF5Array class and converts NIfTI objects on disk to HDF5 files which allow for block processing and memory-efficient representations in R (Pages, 2020).

4.3. Results

In R, multi-modal image analysis requires large quantities of memory. Using `NiftiArray`, images are represented extremely efficiently such that large numbers of images can be analyzed in R. Figure 4.1 compares memory requirements in R for a single image that is dimension 182 by 218 by 182 with voxel size 1mm^3. Traditional imaging packages used to load data such as `neurobase` (Muschelli, 2020) and `RNifti` (Clayden, Cox, and Jenkinson, 2020) result in images consuming approximately 60,000 kilobytes of memory. The same image loaded into R using `NiftiArray` consumes only 9 kilobytes. Using `NiftiArray` results in massive memory efficiency gains.

Speed can be as important as memory during analyses. Ideally, software should be both fast and memory efficient. Time lags due to slow software can cause distraction to the user. Software function calls running in 0.1 second or less will result in no perceived time lag to a user and thus no distraction to user thoughts and tasks. Software functions that take 1.0 second will result in a user observed delay but flow and thought process remain uninterrupted. Software functions running at 10 seconds result in noticeable delay and loss of flow or thought processes (Card, Robertson, and Mackinlay, 1991; “Response time in man-computer conversational transactions”). When software functions run beyond 10 seconds the user may lose track of the task at hand. That is, the users may open Twitter or Instagram and completely lose track of what they were doing. Therefore, function calls that take no more than 1 second are ideal to minimize lag and maximize user attention spans.

The `NiftiArray` and `RNifti` load speeds are both at approximately the 0.1 limit of seamless user

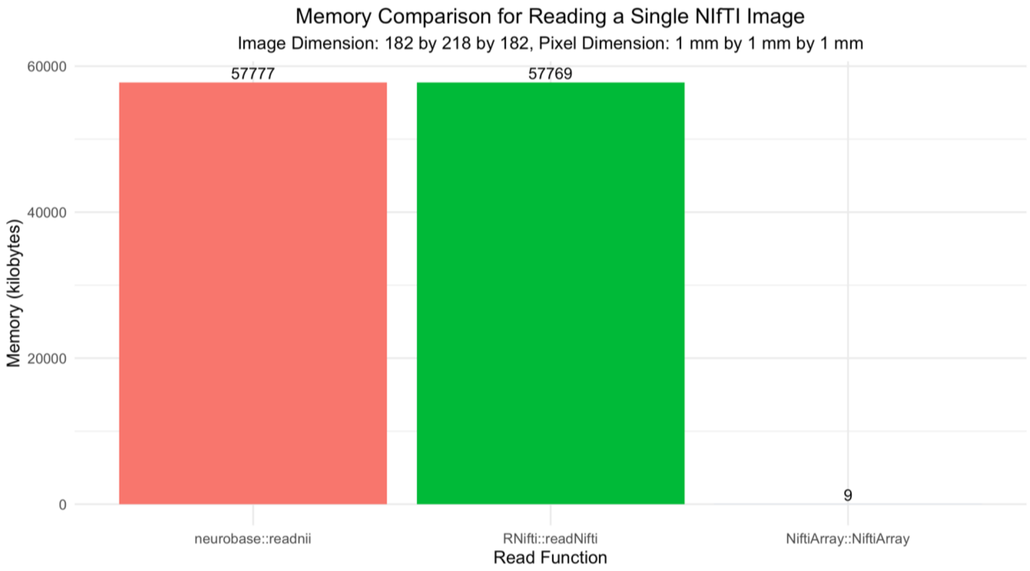


Figure 4.1: Comparison of the local memory consumption in R of a single image that is dimension 182 by 218 by 182 with voxel size 1mm^3. The first two bars (pink and green) load the image into R using `neurobase`'s `readnii` and `RNifti`'s `readNifti`. The last column uses the `NiftiArray` function from the `NiftiArray` package. Memory is compared using kilobytes.

flow. The remaining function from `neurobase` is at the 1 second limit where a users will notice a lag but not lose their train of thought. The `RNifti` R wraps code in C++ and is thus extremely fast. Though coded in R, `NiftiArray` shows competitive speed with `RNifti` and results in no perceived time lag to users with reading speeds at around the 0.1 limit.

4.4. Summary

`NiftiArray` is an R package that allows for memory efficient analysis of NIfTI images (Muschelli and Valcarcel, 2020). Memory conservation is not achieved at cost of speed. Since `NiftiArray` is compatible with `DelayedMatrixStats` quick calculations of common simple statistics (i.e. voxel mean, median, and standard deviation) at the voxel-level are now simple to implement on large imaging datasets (Hickey, 2020). Users can also create their own user-defined calculations to compute other voxel-level analyses quickly utilizing delayed operations since `NiftiArray` is also compatible with `DelayedArray` (Pages, Hickey, and Lun, 2020).

A development version of the `NiftiArray` package is available on GitHub (`https://github.com/muschellij2/NiftiArray`). A stable version is available on Neuroconductor (`https://neuroconductor.org/package/NiftiArray`). A package website was developed for doc-

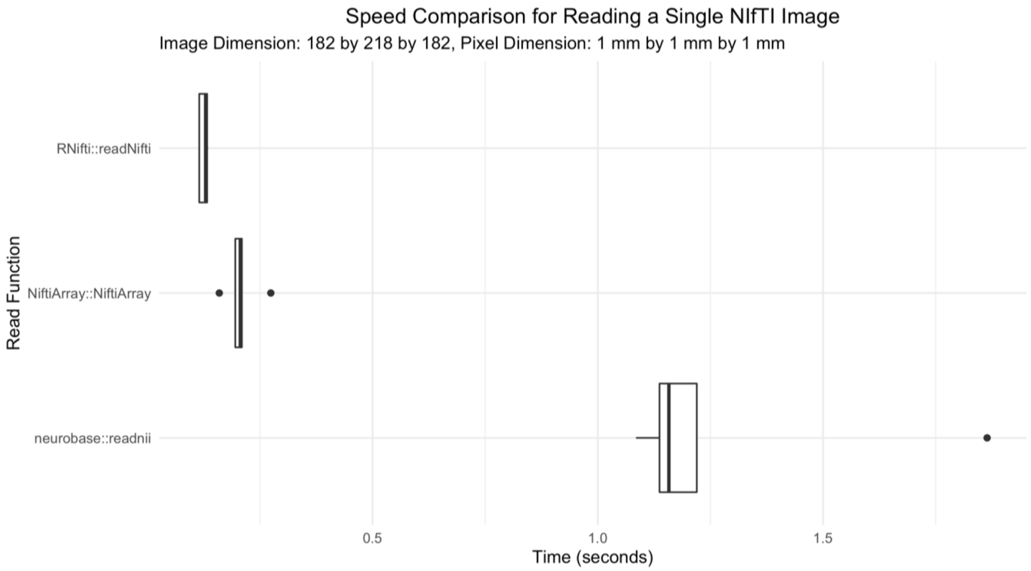


Figure 4.2: Comparison of the time taken to read a single image that is dimension 182 by 218 by 182 with voxel size 1mm^3 using the `neurobase readnii`, `RNifti readNifti`, and `NiftiArray NiftiArray` functions. Speed is compared in seconds for 100 benchmark iterations.

umentation and to host a tutorial. The package website is available through Neuroconductor (`https://neuroconductor.org/help/NiftiArray/`).

CHAPTER 5

Approaches for Modeling Spatially Varying Associations Between Multi-Modal Images

5.1. Introduction

All methods for imaging the brain and measuring its activity (structural magnetic resonance imaging (MRI), functional MRI, diffusion tensor imaging (DTI), computerized tomography (CT), positron emission tomography (PET), electroencephalogram (EEG), and more) have both technical and physiological limitations (Liu et al., 2015a). Multi-modal imaging provides complementary measurements to enhance signal and our understanding of neurobiological processes (Biessmann et al., 2011). When studied jointly, multi-modal imaging data may improve our understanding of the brain. Unfortunately, the vast number of imaging studies evaluate data from each modality separately (voxel- or region-wise) and do not consider information encoded in the relationships between imaging types.

There are a number of existing approaches for multi-modal image analysis, many of which rely on sparsity assumptions and penalization to deal with the massively high dimension of multiple images. Multivariate pattern analysis (MVPA) can be used to integrate information across a set of modalities that are predictive of a phenotype using models such as support vector machines (SVM) (Zhang et al., 2011). MVPA leverages the correlation structure among images in a black box manner for the purpose of prediction rather than studying correlations among imaging features directly. Methods such as independent component analysis (ICA) (Calhoun, Liu, and Adali, 2009) and canonical correlation analysis (CCA) (Correa et al., 2008) can be used to identify common signals from multi-modal images that may provide insight into the correspondence between different types of images. However, the common signals may be spatially distributed throughout the brain and difficult to interpret.

In contrast to high-dimensional predictive models, voxel-wise analyses using the general linear model are primarily used to study associations between a single image modality and demographics, clinical phenotypes, or treatment groups. An exception is biological parametric mapping (BPM)

(Casanova et al., 2007; Yang et al., 2011) which allows for a voxel-wise regression of one image modality on another as well as covariates of interest. However, BPM relies strictly on across-subject information at a single location to estimate the local relationship between different imaging modalities. In contrast, our proposed approach leverages both within- and between-subject information to quantify local relationships between modalities.

In this work, we develop inter-modal coupling (IMCo), a general framework for quantifying how multiple image modalities covary with each other that extends recent work on cortical coupling (Vandekar et al., 2016). IMCo is a regression-based framework that can be used to provide population- and subject-level estimates of spatially varying multi-modal image associations. Within-subject IMCo estimates have been included as features in a competitive automatic segmentation algorithm to delineate multiple sclerosis lesions (Valcarcel et al., 2018a,b). While IMCo estimates have proven useful as features in this predictive model setting, we aim to expand the utility of IMCo as a general framework by providing: 1) valid inference for population-level IMCo parameters and 2) subject-level IMCo estimates with good statistical properties.

Statistical analysis of spatially correlated data requires methods that can properly account for the dependence between observations. Two-stage least squares regression, which has been implemented in analyses of local cortical coupling (Vandekar et al., 2016), is an estimation approach that partitions the between- and within-subject spatial variation. In the first stage, the method estimates a linear association for each subject among the correlated observations and then models the individual-level linear association estimates across subjects in the second stage. Thus, the first stage analysis is restricted to within-subject modeling and the second stage is restricted to between- or across-subject modeling.

Modeling within-subject data in the first stage may result in noisy subject-level estimates and unreliable variance estimates (Diggle et al., 2002). These issues motivate the use of single-stage estimation models such as weighted least squares (WLS) (*Linear Models in Statistics*), linear mixed effects (LME) models (Laird and Ware, 1982), or generalized estimating equations (GEE) (Liang and Zeger, 1986). These models borrow information across subjects while appropriately accounting for spatially correlated data within subjects. A primary distinction between these models is that WLS and LME models are full-likelihood based methods while GEEs rely on partial-likelihood specification.

WLS accounts for spatial correlation by attempting to give each data point its proper amount of influence in the model. WLS estimators are unbiased even when incorrect weights are used. However, inefficiency and unreliable inference can result from incorrect specification of the weights. In two-stage analyses, WLS can be used to provide within-subject estimates, but in single-stage, across-subject estimation, WLS only permits estimation of population-level association parameters.

LME models incorporate both fixed and random effect terms in the linear predictor from which the conditional mean of the response can be estimated (Laird and Ware, 1982). A LME model relies on full specification of the likelihood to estimate parameters, and therefore distributional assumptions are required. In addition to estimating population-level effects, LME models allow for estimation of subject-specific effects.

GEEs estimate the average response over the population (”population-averaged” effects) rather than the effect of changing one or more covariates for a given individual (”conditional” effects). Thus, within-subject estimates cannot be obtained from the GEE approach. The GEE method does not require full specification of the multivariate distribution of the correlated voxel values but rather only the first two moments of the distribution. Instead of attempting to correctly estimate the true within-subject covariance structure, the GEE treats it as a nuisance and relies on specification of a ”working” correlation structure to use for mean parameter estimation. The working correlation structure does not need to be specified correctly in order to obtain unbiased estimates of regression coefficients. However, as the working correlation structure gets further from the population correlation structure, estimators will become increasingly inefficient (standard errors will be large).

Statistical properties, advantages, and disadvantages of two-stage estimation, WLS, LME models, and GEE methods have been well documented in longitudinal data analyses. Often the best model choice depends on the scientific question of interest. In the context of relating multi-modal images where the spatial correlation is complex and unknown, it is unclear how these models may perform. To gain insight on the relative performance of these models for multi-modal imaging studies, we designed and implemented a comprehensive set of Monte Carlo simulations based on real imaging data. We focus primarily on population-level estimation of the association between images, comparing bias and mean squared error of estimators from multiple model specifications.

Table 5.1: Demographic information for 831 subjects scanned as part of the Philadelphia Neurodevelopmental Cohort (PNC). Statistics presented: N (percent); mean (SD, minimum, maximum).

Characteristic	**N=831**
Sex	
Male	353 (42%)
Female	478 (58%)
Age (years)	15.6 (3.4, 8.2, 23.0)

5.2. Materials and methods

5.2.1. Subjects

The Philadelphia Neurodevelopmental Cohort (PNC) is a large-scale study of child development carried out by the University of Pennsylvania and the Center for Applied Genomics at the Children's Hospital of Philadelphia. The PNC consists of rich multi-modal neuroimaging, genetics, and detailed clinical and cognitive phenotyping (Satterthwaite et al., 2014b, 2016). The PNC includes 9,498 participants ages 8-23 at baseline who underwent thorough cognitive and psychiatric evaluation. Of these adolescents, 1,601 underwent multi-modal neuroimaging including T1-weighted structural neuroimaging, diffusion tensor imaging, perfusion neuroimaging using arterial spin labeling, functional imaging tasks of working memory and emotion identification, and resting state imaging of functional connectivity. All subjects were imaged on a single scanner under the same imaging protocol. The study design and data are described in detail by Satterthwaite et al., 2016 and Satterthwaite et al., 2014b.

In the current study, we exclude subjects who were taking psychoactive medication, had any medical problems that could impact brain function, had a history of psychiatric hospitalization, or had any abnormalities of brain structure or distortions of brain anatomy as determined by review of the T1-weighted image by a neuroradiologist. Participants with unusable T1-weighted images are also excluded because the T1-weighted image is necessary for registration. Participants with poor quality images are also excluded based on modality-specific quality assurance. These exclusions are summarized in Figure 5.1. The final sub-sample for our analysis consists of 831 adolescents (478 females) aged 8–23 (mean = 15.6, sd = 3.4) at time of first scan. Demographics are included in Table 5.1.

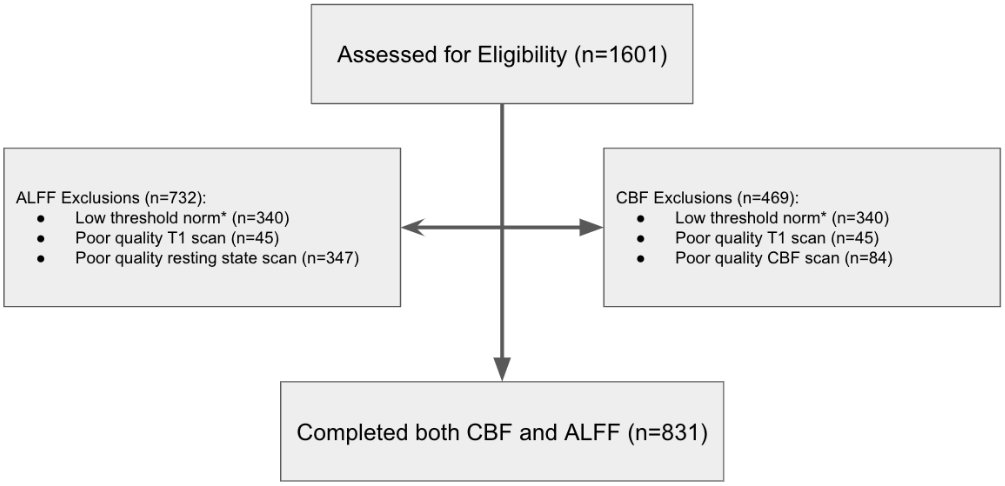

Assessed for Eligibility (n=1601)
ALFF Exclusions (n=732):
Low threshold norm* (n=340)
Poor quality T1 scan (n=45)
Poor quality resting state scan (n=347)
CBF Exclusions (n=469):
Low threshold norm* (n=340)
Poor quality T1 scan (n=45)
Poor quality CBF scan (n=84)
Completed both CBF and ALFF (n=831)

Figure 5.1: Exclusion criteria for the current analysis of PNC data. *Indicates no medical co-morbidities, no abnormal brain structure on radiology read, not currently using psychoactive or psychiatric medications, and no inpatient hospitalizations.

5.2.2. Image acquisition and preprocessing

In this work, we apply our proposed methodology to study spatially varying relationships between local functional connectivity quantified by Amplitude of Low Frequency Fluctuation (ALFF) images and cerebral blood flow (CBF). ALFF quantifies the amplitude of low-frequency oscillations over time and space from resting-state BOLD scans to determine correlated activity between brain regions. We choose ALFF rather than other resting-state functional connectivity measures since prior work shows abnormal resting-state low-frequency fluctuations are associated with neurodevelopment and psychopathology (Bing et al., 2013; Hoptman et al., 2010; Liu et al., 2014; Liu et al., 2018; Wang et al., 2019; Zang et al., 2007; Zhou et al., 2015).

Imaging data were acquired on a single Siemens TIM Trio 3 tesla scanner with a 32-channel head coil using identical sequencing protocol (Satterthwaite et al., 2014b). Image acquisition, processing, and quality assurance for the full set of multi-modal imaging data collected as part of the PNC have been previously described (Satterthwaite et al., 2014b, 2016).

CBF was calculated from brain perfusion imaging using a custom written pseudo-continuous arterial spin labeling (pCASL) sequence (Satterthwaite et al., 2014a,b; Wu et al., 2007). ALFF was computed from resting-state fMRI imaging as the sum over frequency bins in the low-frequency

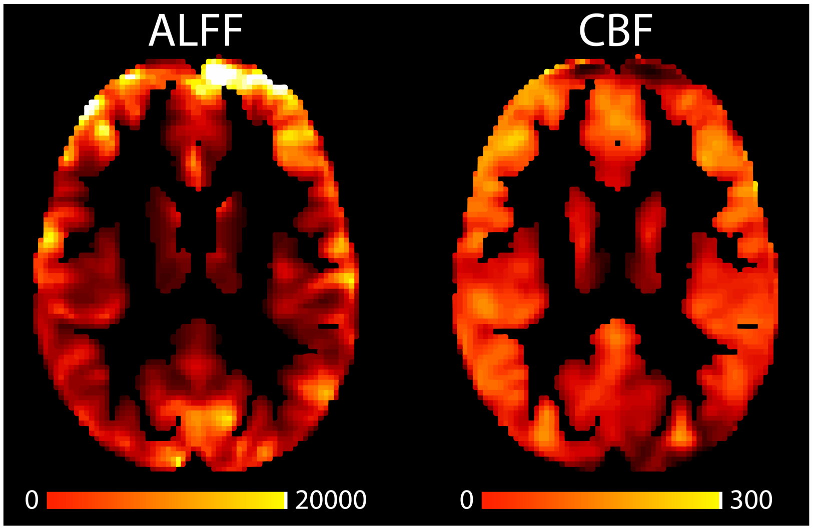


Figure 5.2: Axial slices of ALFF and CBF images from a randomly selected subject from the PNC. The brain mask includes voxels within the gray matter (minimum 10% probability according to an atlas prior) that had adequate image coverage for both resting-state functional MRI and ASL.

(0.01-0.08) band of the power spectrum (Kaczkurkin et al., 2019; Zang et al., 2007). Figure 5.2 displays an axial slice of volumetric ALFF and CBF imaging modalities from a randomly selected subject from the PNC.

5.2.3. Inter-modal coupling

Inter-modal coupling is our proposed general framework for quantifying the relationship between two imaging modalities, denoted by X and Y, and mapping how the relationship varies spatially across the brain. We assume all images have been registered to a common template space. Let v_0 denote a specific voxel in the brain, where $v_0 = 1, ..., V$ indexes all voxels in the brain mask. Given a particular location in the brain, we regress the outcome image modality Y on the remaining modality X using data from all voxels in a local neighborhood of the target voxel v_0. We repeat this procedure at all target voxels, $v_0 = 1, ..., V$. We call the spatially varying relationship between two imaging modalities inter-modal coupling (IMCo), which can be estimated at the subject (within-subject) or population (across-subject) level (Valcarcel et al., 2018a,b; Vandekar et al., 2016). With this general set up, we now specify several models that could be used for the regressions at each target voxel.

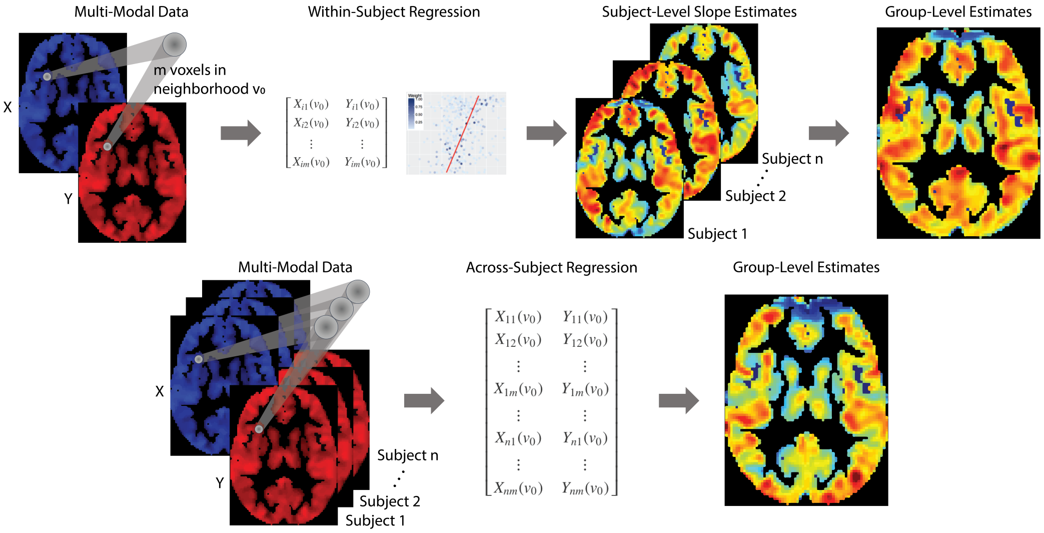


Figure 5.3: Schema of the within-subject (row 1) and across-subject (row 2) inter-modal coupling (IMCo) frameworks. We let m denote voxels within the neighborhood of the target voxel v_0. Within-subject IMCo is implemented in two stages using weighted least squares regression (WLS) to produce IMCo maps which are then regressed across subject on covariates of interest. Across-subject coupling is implemented in one stage using generalized estimating equations (GEE), linear mixed effects models (LME), or population-level WLS regression.

5.2.4. Within-subject WLS

The within-subject (WS) version of IMCo first regresses image modality Y on X at the individual subject level. For subject i, let $\boldsymbol{N}(v_0)$ denote a neighborhood of voxels centered at the target voxel v_0 that includes v_0 itself. We index voxels in $\boldsymbol{N}(v_0)$ as v_j, $j = 1, ..., J$. Thus, $|\boldsymbol{N}(v_0)| = J$. Let $\boldsymbol{X}_i(v_0)$ denote a design matrix that includes a column of 1's for the intercept and subject i's vectorized neighborhood of voxels surrounding v_0 from the independent imaging modality X. Let $\boldsymbol{Y}_i(v_0)$ denote subject i's vectorized neighborhood of voxels surrounding v_0 from the dependent imaging modality Y. We use one modeling approach to fit the within-subject framework.

Assume the following underlying true model for the data from subject i, neighborhood $\boldsymbol{N}(v_0)$:

$$\boldsymbol{Y}_i(v_0) = \boldsymbol{X}_i(v_0)\boldsymbol{\beta}_i^{IMCo}(v_0) + \boldsymbol{\epsilon}_i(v_0), \tag{5.1}$$

where we assume $E[\boldsymbol{\epsilon}_i(v_0)] = \boldsymbol{0}$ $(\dim(\boldsymbol{\epsilon}_i) = J \times 1)$ and $Var(\boldsymbol{\epsilon}_i(v_0)) = \Sigma_i(v_0)$ $(\dim(\Sigma_i) = J \times J)$. In order to reduce the parameter space, in this working model, we assume a common covariance structure within neighborhood $\boldsymbol{N}(v_0)$ for all subjects i and at all voxels v_0 in the brain mask. That is, we assume $Var(\boldsymbol{\epsilon}_i(v_0)) = \Sigma$ for all i and v_0. The IMCo parameter vector $\boldsymbol{\beta}_i^{IMCo}(v_0)$ consists of an intercept and slope. The within-subject IMCo framework seeks estimated intercept and slope brain maps for each subject (henceforth referred to as IMCo maps).

To fit model (5.1), we use weighted least squares (WLS), where the weights depend on the neighboring voxels' distance from the central voxel. Let $d(v_0, v_j)$ denote a measure of distance from voxel $v_j \in \boldsymbol{N}(v_0)$ to v_0. We use Euclidean distance for $d(v_0, v_j)$ but other distance measures could be specified. We then assign weights according to an isotropic Gaussian kernel,

$$w_j(v_0) = \exp\Big\{ - \frac{d(v_0, v_j)^2}{2\sigma^2} \Big\}, \tag{5.2}$$

where the parameter σ determines the smoothness of the estimated IMCo measures across the brain. We specify σ based on the full-width half-maximum (FWHM) parameter of a Gaussian distribution, with FWHM units measured in millimeters. A larger FWHM specification results in a larger neighborhood size and smoother IMCo maps since σ grows with the FWHM value. We define $W = \text{diag}(\boldsymbol{w}(v_0))$ where $\boldsymbol{w}(v_0)$ is a $J \times 1$ vector of weights and all off-diagonal elements in W are equal to 0. Then, W is used as the weight matrix in a standard weighted least squares fit.

Population-level analyses can proceed in a second modeling step using subject-level IMCo maps. For example, it might be of interest to test whether the relationship between image modalities differs by sex or is associated with age. We refer to the estimation of subject-level IMCo parameters as stage 1 modeling. In stage 2 modeling, we regress the individual IMCo maps ($\hat{\beta}_{i0}^{IMCo}$ or $\hat{\beta}_{i1}^{IMCo}$) from the stage 1 modeling on covariates of interest using a voxel-wise analysis.

5.2.5. Across-subject modeling

The across-subject (AS) modeling framework regresses image modality Y on X using local neighborhoods as defined in the within-subject approach by stacking observations in neighborhood $\boldsymbol{N}(v_0)$ across subjects $i = 1, ..., n$.

We propose three models for the across-subject IMCo framework, which we will compare using simulated data:

1. weighted least squares using across-subject estimation (WLS-AS),
2. generalized estimating equations (GEE), and
3. linear mixed effects models (LME).

5.2.6. Across-subject WLS

We assume the following model:

$$\boldsymbol{Y}_i(v_0) = \boldsymbol{X}_i(v_0)\boldsymbol{\beta}^{IMCo}(v_0) + \boldsymbol{\epsilon}_i(v_0). \tag{5.3}$$

Notice, in Equation (5.1) the parameter vector $\boldsymbol{\beta}_i(v_0)$ is subject-specific, whereas in Model (5.3) $\boldsymbol{\beta}(v_0)$ is a population-level parameter vector that represents the expected relationship between image modalities. The assumptions on the errors and covariance structure are the same as in the WLS-WS model in Equation (5.1). We fit Model (5.3) using weighted least squares, stacking the vectors $\boldsymbol{Y}_i(v_0)$ and design matrices $\boldsymbol{X}_i(v_0)$ from each subject. Weights for each subject's observations are assigned in the same way as WLS-WS.

5.2.7. LME

Linear mixed effects models incorporate both fixed and random effects in a linear predictor from which the conditional mean of the response can be estimated (Laird and Ware, 1982). A fixed effect is a parameter that does not vary. In contrast, random effects are parameters that are themselves random variables. LME relies on specification of the full likelihood to estimate parameters. Therefore, distributional assumptions are required.

We assume the following model:

$$\boldsymbol{Y}_i(v_0) = \boldsymbol{X}_i(v_0)\boldsymbol{\beta}(v_0) + Z_i(v_0)\boldsymbol{b}_i(v_0) + \boldsymbol{\epsilon}_i(v_0). \tag{5.4}$$

The design matrix for the random effects, $Z_i(v_0)$, is an $n \times k$ matrix. We use $\boldsymbol{b}_i(v_0)$ to denote the k random effects where $\boldsymbol{b}_i(v_0) \sim N(\boldsymbol{0}, \psi(v_0))$ and $\psi(v_0)$ is the $k \times k$ covariance matrix for the random effects.

The design matrix for the fixed effects $\boldsymbol{X}_i(v_0)$ and vector of fixed effects, $\boldsymbol{\beta}(v_0)$, are defined in the same way as those in Equations (5.3) and (5.9). We use $\boldsymbol{\epsilon}_i(v_0)$ to denote the error associated with the fixed effects where $\boldsymbol{\epsilon}_i(v_0) \sim N(\mathbf{0}, \sigma^2\Lambda_i(v_0))$ and $\sigma^2\Lambda_i(v_0)$ is the $n \times n$ covariance matrix for the errors associated with subject i. More concisely, the assumptions surrounding the model in Equation (5.4) can be written as:

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{\epsilon}_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \psi & \mathbf{0} \\ \mathbf{0} & \sigma^2\Lambda_i \end{bmatrix} \right). \tag{5.5}$$

In order to make the conditional nature of the LME model obvious we re-write (5.5) as a two level hierarchical model:

$$\boldsymbol{Y}_i | \boldsymbol{b}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \sigma^2\Lambda_i), \tag{5.6}$$

and

$$\boldsymbol{b}_i \sim N(\mathbf{0}, \sigma^2\Lambda_i). \tag{5.7}$$

5.2.8. GEE

Assume $E[\boldsymbol{Y}_i(v_0)] = \boldsymbol{\mu}_i(v_0)$ where $\boldsymbol{\mu}_i(v_0)$ $(\dim(\boldsymbol{\mu}_i(v_0)) = J \times 1)$ for subject i as the mean vector. We assume a marginal model to relate the mean response vector and the covariates:

$$g(\boldsymbol{\mu}_i(v_0)) = \boldsymbol{X}_i(v_0)\boldsymbol{\beta}(v_0), \tag{5.8}$$

where g is a known link function. The estimating equation then depends on the regression parameters $\boldsymbol{\beta}(v_0)$, and variance structure, $V_i(v_0)$ $(\dim(V_i(v_0)) = J \times J)$, via:

$$U(\boldsymbol{\beta}(v_0)) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i(v_0)}{\partial \boldsymbol{\beta}(v_0)} V_i^{-1}(v_0)\{\boldsymbol{Y}_i(v_0) - \boldsymbol{\mu}_i(\boldsymbol{\beta}(v_0))\}. \tag{5.9}$$

5.2.9. Simulation

In this section we evaluate the performance of the IMCo modeling approaches using simulations. MRI data sets are large and characterized by complex dependence structures that reflect the brain's anatomy and neurophysiology. In order to incorporate the sophisticated spatial structure of neuroimaging data, we first create a synthetic outcome imaging modality, denoted as Y^*, from the CBF images collected as part of the PNC. We induce additional known spatial correlation and dependence in order to evaluate our methods with respect to these parameters.

We simulate data using the following model:

$$\begin{aligned} Y_i^*(v_0) =& \beta_0 + \beta_1 \times CBF_i(v_0) + \beta_2 \times CBF_i(v_0) \times I(GM(v_0) = 1) + \\ & \beta_3 \times Sex_i + \beta_4 \times Sex_i \times CBF_i(v_0) + b_{0i} + q_i(v_0) + \epsilon_i(v_0), \end{aligned} \tag{5.10}$$

where $b_{0i} \sim N(0, \sigma_b^2)$ and $\epsilon_i(v_0) \sim N(0, \sigma_\epsilon^2)$. We let $I(GM(v_0) = 1)$ denote an indicator variable for whether a voxel is in the gray matter. This allows for a different effect of CBF in gray matter compared to white matter. In order to induce an exchangeable correlation structure in portions of the brain, we first split the brain into quadrants. Each quadrant is assigned a quadrant ID from 1 to 4. All the voxels in each respective quadrant are then simulated from a mean zero normal distribution with the standard deviation assigned as the quadrant ID. More formally, $q_i(v_0) \sim N(0, \sigma_Q^2)$ where Q ranges from 1 to 4. Voxels within a quadrant will have exchangeable correlation structure since they are generated from the same distribution.

In an effort to reduce computation time and conserve memory, we select a set of adjacent axial slices to decrease the total number of voxels where we need to estimate IMCo parameters. We use axial slices 46 to 52 from the CBF images in the simulation. Analyses are carried out in a brain mask created by intersecting pCASL and resting-state fMRI masks. Additionally, we do not run IMCo in neighborhoods that are missing more than 50% of voxels.

To assess how the quality of IMCo estimates vary with sample size, we run the simulation with $n = 30$ and $n = 100$. We sample n CBF subject images randomly from the available 831 PNC participants in each iteration of the simulation. We simulate Y^* for each subject from the model in Equation (5.10) using the n randomly sampled CBF images. This process is repeated for $k = 100$

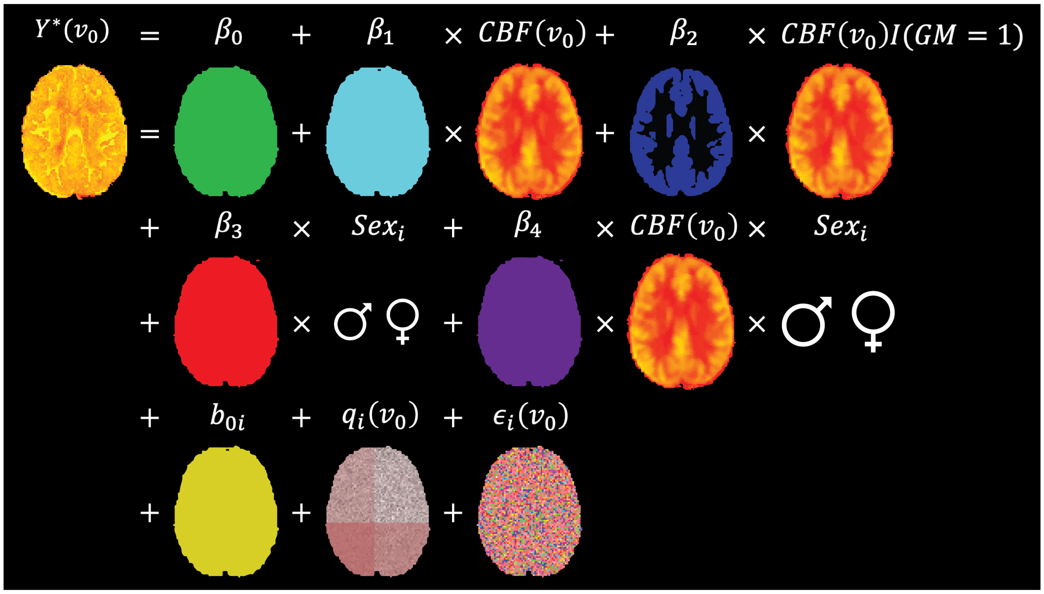

$Y^*(v_0) = \beta_0 + \beta_1 \times CBF(v_0) + \beta_2 \times CBF(v_0)I(GM=1)$
$+ \beta_3 \times Sex_i + \beta_4 \times CBF(v_0) \times Sex_i$
♂ ♀
$+ b_{0i} + q_i(v_0) + \epsilon_i(v_0)$

Figure 5.4: Axial slices of simulation parameters are displayed. Colors are arbitrarily chosen to emphasize whether the parameter is assigned at the voxel or image level.

Term	Parameter	Value
Sample size	n	30, 100
Number of Iterations	k	100
Intercept	β_0	222.26
$CBF(v_0)$	β_1	2.58
$CBF \times I(GM = 1)$	β_2	-1.73
Sex_i	β_3	46.3
$CBF \times Sex_i$	β_4	-0.03
$\epsilon_i(v_0) \sim N(0, \sigma_\epsilon^2)$	σ_ϵ^2	81
$b_{0i} \sim N(0, \sigma_b^2)$	σ_b^2	466489
$q_i(v_0) \sim N(0, \sigma_1^2)$ Quadrant 1	σ_1^2	1
$q_i(v_0) \sim N(0, \sigma_1^2)$ Quadrant 2	σ_2^2	4
$q_i(v_0) \sim N(0, \sigma_3^2)$ Quadrant 3	σ_3^2	9
$q_i(v_0) \sim N(0, \sigma_4^2)$ Quadrant 4	σ_4^2	16

Table 5.2: Parameter values assigned to generate simulation data.

Monte Carlo iterations. To inform realistic parameters values for Model (5.10), we ran an exploratory analysis using all 831 PNC subjects' ALFF and CBF images. We regressed ALFF on CBF, sex, and a CBF by sex interaction using a linear mixed effects model with a random intercept. All simulation parameters and assigned values are provided in Table 5.2.

After generating Y^* for n subjects, we fit the following IMCo models:

1. GEE with an exchangeable working correlation structure (GEE-Exch),

2. GEE with an independent working correlation structure (GEE-Ind),

3. Linear mixed effects model with a random intercept (LME-RI),

4. Linear mixed effects model with a random slope and intercept (LME-RSI),

5. Weighted least squares across-subject approach (WLS-AS), and

6. Weighted least squares within-subject approach (WLS-WS).

It is important to note that the LME-RI model is correctly specified based on the data generating model shown in (5.10). The GEE-Exch model is also expected to perform well as the correlation structure and model specifications are correct. Since the WLS-WS approach uses a two stage model fitting framework, the results from the second stage analysis where we regress sex on the IMCo slope maps across subjects at the voxel-level are only comparable to the CBF by Sex interaction term in the other models. We are unable to estimate the other parameters in this model.

Only the LME-RI, LME-RSI, and WLS-WS models are capable of estimating subject-specific parameters. The true model in Equation (5.10) includes a subject-specific intercept b_{0i}. We calculate bias and MSE for the subject-specific intercept estimate for these models only.

To assess the simulation performance after models are fit, we calculate voxel-level bias and mean square error (MSE) across simulation iterations. We calculate average bias and MSE in four brain regions: full brain, gray matter only, white matter only, and locations proximal to where the white and gray matter IMCo neighborhoods overlap. We present results for white and gray matter separately as the true underlying model has distinct effects in white and gray matter. Furthermore, as previous IMCo studies have shown artifacts in boundary regions (Vandekar et al., 2016), it is of interest to assess model performance in neighborhoods at the boundary of white and gray matter.

To carry out these simulations we use tools built in Chapter 4.

5.3. Results

5.3.1. Bias

We present overall bias estimates from the simulation setting with $n = 30$ in Table 5.3. We visualize these results in Figure 5.5.

To better understand model accuracy in each tissue class and at the boundary of tissue classes, we compare bias across the modeling approaches in each brain region. The average bias is generally similar for full brain, boundary, white, and gray matter regions. In general, at the boundary of white and gray matter where neighborhoods include voxels from both tissue classes, we do not observe larger biases. There are two notable exceptions. The CBF parameter estimates show larger biases across all models in the brain white matter. Additionally, the WLS-WS approach shows increases in bias at the boundary white and gray matter.

We compare parameter bias for each model to assess overall accuracy. The LME-RI model has the minimum bias across all parameter estimates, with the exception of CBF where the WLS-AS model yields the smallest bias. Results from the LME-RSI and GEE-Exch models are similar and show bias equal to or only slightly elevated compared to the LME-RI model. With a sample size of only 30, the GEE-Ind model results in larger biases than the LME-RI, LME-RSI, and GEE-Exch models. We are only able to compare bias from the WLS-WS model to the other models for the CBF by sex interaction term. While the WLS-WS model yields larger biases compared to the LME-RI, LME-RSI, and GEE-Exch models, the bias is smaller than both GEE-Ind and WLS-AS models. The WLS-AS model consistently yields the largest bias compared to the other modeling approaches.The models tend to over-estimate the intercept and CBF by sex interaction while they underestimate the CBF and sex parameters.

We repeat the simulation using $n = 100$ and present results in Table 5.5. These results are visualized in Figure 5.6. Generally, results are consistent with those presented in the $n = 30$ setting except for a few notable differences. Within the gray matter models are not consistently over- or under-estimating the parameters. Notice, in the CBF by sex interaction bias panel in Figure 5.6, bias for GEE-Ind and WLS-AS is negative whereas the remaining models show positive bias. The GEE-Ind model still produces larger bias than the LME-RI, LME-RSI, and GEE-Exch models but with $n = 100$ differences are not as far as in the $n = 30$ simulation setting. With larger sample size, the GEE-Ind model seems to be approaching the performance of the LME-RI, LME-RSI, and GEE-Exch models. All models tend to over-estimate the intercept and interaction term but under-estimate sex and CBF terms.

Comparing the two simulation settings, $n = 30$ with $n = 100$, we find that the estimates for the intercept and sex are less biased in the $n = 100$ setting across all modeling approaches. The remaining parameter estimates, CBF and the interaction term, show similar biases. Noticeably, the GEE-Ind model yields substantially less bias with $n = 100$ compared to $n = 30$.

Table 5.7 presents bias results for the subject-specific intercept estimates in the $n = 30$ simulation setting. These results are visualized in Figure 5.7. The bias is similar across models for the different brain regions (full, boundary, white, and gray matter). The WLS-WS model results in the smallest bias closely followed by the LME-RI and LME-RSI models. The results for the $n = 100$ simulation setting are presented in Table 5.9 and Figure 5.8. Findings are similar to those in the $n = 30$ simulation. Compared to using a sample size of $n = 30$, the subject-specific bias estimates are substantially lower for the $n = 100$ simulation setting.

5.3.2. MSE

We present MSE estimates from the simulation setting with $n = 30$ in Table 5.4. In order to better visualize findings, these results are also displayed as a bar chart in Figure 5.5.

To better understand model accuracy in each tissue class and at the boundary of tissue classes, we compare MSE across the modeling approaches in each brain region. The average MSE is generally similar for full brain, boundary, white, and gray matter regions. In general, at the boundary of white and gray matter where neighborhoods include voxels from both tissue classes, we do not observe larger MSE.

We compare MSE for parameters across models to assess estimation accuracy. The LME-RI model results in the smallest MSE across parameters but these estimates are often equal to the LME-RSI and GEE-Exch models. MSE for the WLS-AS model is extremely large for each parameter. The GEE-Ind model shows larger MSE compared to the LME-RI, LME-RSI, and GEE-Exch models but is smaller than the WLS-AS model. We are only able to compare MSE from the WLS-WS model to the other models for the CBF by sex interaction term. While the WLS-WS model yields larger MSE compared to the LME-RI, LME-RSI, and GEE-Exch models, the MSE is smaller than both GEE-Ind and WLS-AS models.

We repeat the simulation using $n = 100$ and present these results in Table 5.6. These results are visualized in Figure 5.7. The MSE findings using a sample size of $n = 100$ are the same as those in the $n = 30$ setting. As expected, the MSE is smaller using $n = 100$ compared to $n = 30$.

Table 5.10 presents MSE results for the subject-specific intercept in the $n = 30$ simulation setting. These results are visualized in Figure 5.8. We find no differences in MSE performance across the brain regions (full, boundary, white, and gray matter). The LME-RSI model results in the smallest MSE. The results for the $n = 100$ simulation setting are presented in Table 5.10 and Figure 5.8 and this setting yields the same findings as in $n = 30$. The MSE findings using a sample size of $n = 100$ are the same as those in the $n = 30$ setting. As expected, the MSE is smaller using $n = 100$ compared to $n = 30$.

5.3.3. Bias and MSE maps

The table and bar charts presenting bias and MSE do not elicit patterns related to model accuracy across brain regions. Displaying bias and MSE results in brain maps does elucidate performance differences in certain brain regions across modeling approaches. We present axial slices of bias and MSE for simulation settings $n = 30$ and $n = 100$ in Figures 5.9 and 5.10. Below the average bias and MSE maps for this axial slice we present a histogram summarizing the full brain average bias or MSE.

Both simulation settings show that the WLS-WS model bias is most severe in the boundary regions where IMCo neighborhoods include voxels from both the white and gray matter. The GEE-Ind and WLS-AS models show larger average biases in the brain white matter.

In terms of MSE, both simulation settings also show that the WLS-WS model has increased MSE in the boundary regions where IMCo neighborhoods include voxels from both the white and gray matter. Across both simulation settings, the GEE-Ind and WLS-AS average MSE is largest in the brain white matter.

All models show increased bias and MSE at the edge of the brain. Neighborhoods at the edge of

Table 5.3: Across-subject parameter bias estimates from the n = 30 (100 iterations) simulation setting are presented. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
Intercept	**GEE-Exch**	81.56	79.37	85.37	101.03
Intercept	**GEE-Ind**	98.85	103.74	90.35	101.79
Intercept	**LME-RI**	81.56	79.37	85.37	101.03
Intercept	**LME-RSI**	86.44	84.61	89.61	107.77
Intercept	**WLS-AS**	98.65	108.79	81.04	92.50
CBF	**GEE-Exch**	-0.90	-0.45	-1.68	-1.49
CBF	**GEE-Ind**	-1.20	-0.79	-1.90	-1.47
CBF	**LME-RI**	-0.90	-0.45	-1.68	-1.49
CBF	**LME-RSI**	-0.97	-0.52	-1.76	-1.59
CBF	**WLS-AS**	-1.07	-0.96	-1.27	-1.10
Sex	**GEE-Exch**	-101.08	-101.04	-101.13	-98.53
Sex	**GEE-Ind**	-118.76	-124.83	-108.22	-109.92
Sex	**LME-RI**	-101.08	-101.04	-101.13	-98.53
Sex	**LME-RSI**	-101.32	-101.32	-101.33	-98.85
Sex	**WLS-AS**	-134.21	-146.27	-113.25	-121.66
CBF by Sex Interaction	**GEE-Exch**	0.03	0.03	0.03	0.03
CBF by Sex Interaction	**GEE-Ind**	0.26	0.33	0.13	0.18
CBF by Sex Interaction	**LME-RI**	0.03	0.03	0.03	0.03
CBF by Sex Interaction	**LME-RSI**	0.03	0.03	0.03	0.03
CBF by Sex Interaction	**WLS-AS**	0.48	0.64	0.22	0.40
CBF by Sex Interaction	**WLS-WS**	0.11	0.08	0.17	0.35

the brain will only include voxel intensities from the brain and will have fewer repeated observations per subject.

The histograms created using average bias and MSE from the full brain provided in Figures 5.9 and 5.10 show that LME-RI, LME-RSI, and GEE-Exch models perform similarly and the best in terms of minimizing bias and MSE. The GEE-Ind and WLS-WS both perform reasonably well in terms of minimizing bias and MSE. Across all the models, the WLS-AS model has unstable MSE estimates and performs the worst in terms of bias.

5.4. Discussion

In this work, we provide a unified framework to study the complex relationships between multiple imaging modalities. We quantify the relationship between multiple imaging modalities and map how these relationships vary spatially across different anatomical brain regions. Inter-modal cou-

Table 5.4: Across-subject parameter MSE estimates from the n = 30 (100 iterations) simulation setting are presented. We present average MSE calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
Intercept	**GEE-Exch**	29944.26	29622.64	30503.10	33055.74
Intercept	**GEE-Ind**	178091.05	228269.48	90904.58	116534.35
Intercept	**LME-RI**	29944.25	29622.63	30503.10	33055.73
Intercept	**LME-RSI**	30855.70	30580.61	31333.68	34447.96
Intercept	**WLS-AS**	423153.80	535501.11	227947.11	308948.87
CBF	**GEE-Exch**	1.68	0.46	3.81	3.26
CBF	**GEE-Ind**	59.31	51.55	72.78	39.79
CBF	**LME-RI**	1.68	0.46	3.81	3.26
CBF	**LME-RSI**	1.85	0.54	4.12	3.55
CBF	**WLS-AS**	186.61	134.72	276.77	152.61
Sex	**GEE-Exch**	76493.33	76434.28	76595.91	74720.31
Sex	**GEE-Ind**	414384.38	523410.18	224948.91	268711.84
Sex	**LME-RI**	76493.33	76434.29	76595.91	74720.30
Sex	**LME-RSI**	76491.44	76451.37	76561.08	74692.44
Sex	**WLS-AS**	1064767.87	1325011.53	612587.00	784131.75
CBF by Sex Interaction	**GEE-Exch**	0.04	0.04	0.05	0.06
CBF by Sex Interaction	**GEE-Ind**	144.37	119.13	188.23	91.05
CBF by Sex Interaction	**LME-RI**	0.04	0.04	0.05	0.06
CBF by Sex Interaction	**LME-RSI**	0.03	0.03	0.04	0.05
CBF by Sex Interaction	**WLS-AS**	497.56	337.80	775.15	399.07
CBF by Sex Interaction	**WLS-WS**	0.74	0.56	1.04	1.59

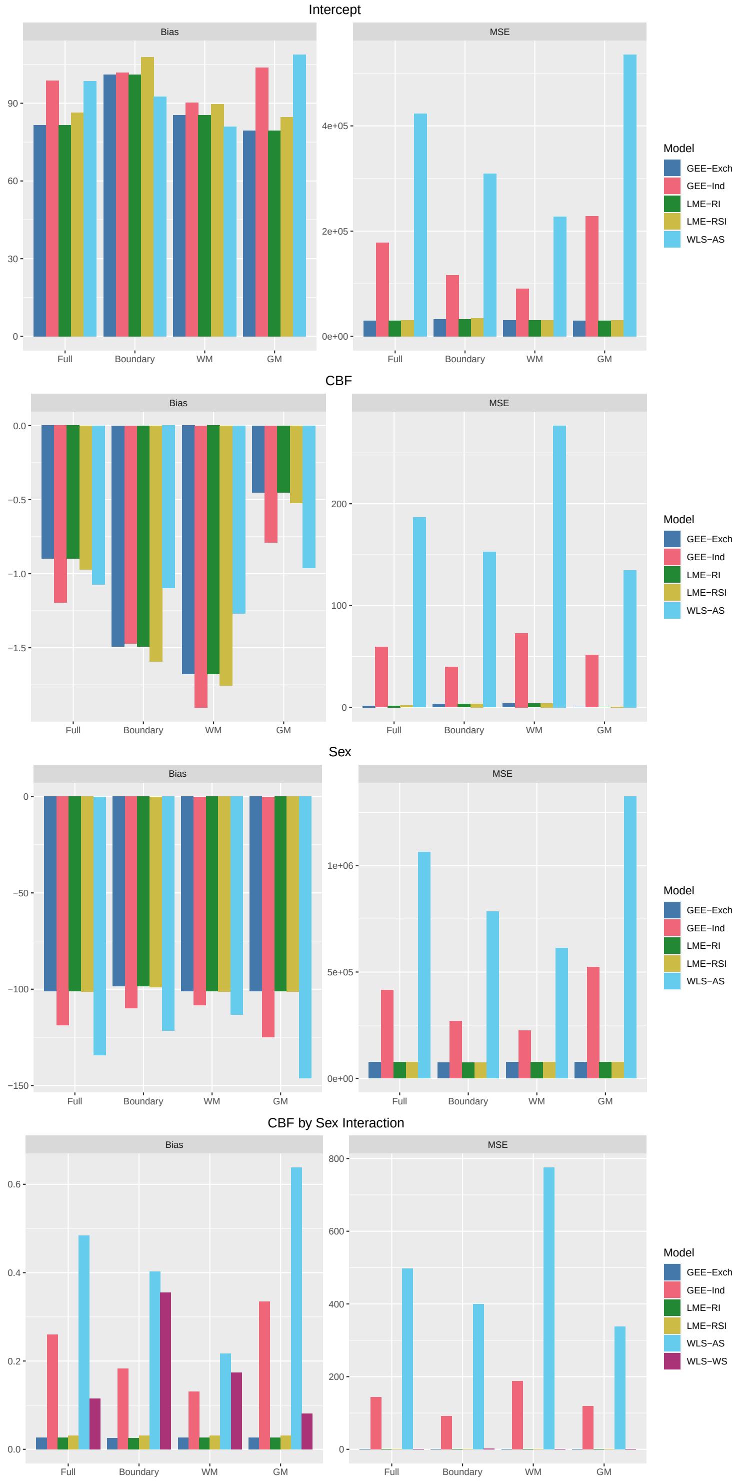


Figure 5.5: Across-subject parameter bias and MSE estimates from the n = 30 (100 iterations) simulation setting are presented in a bar chart. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

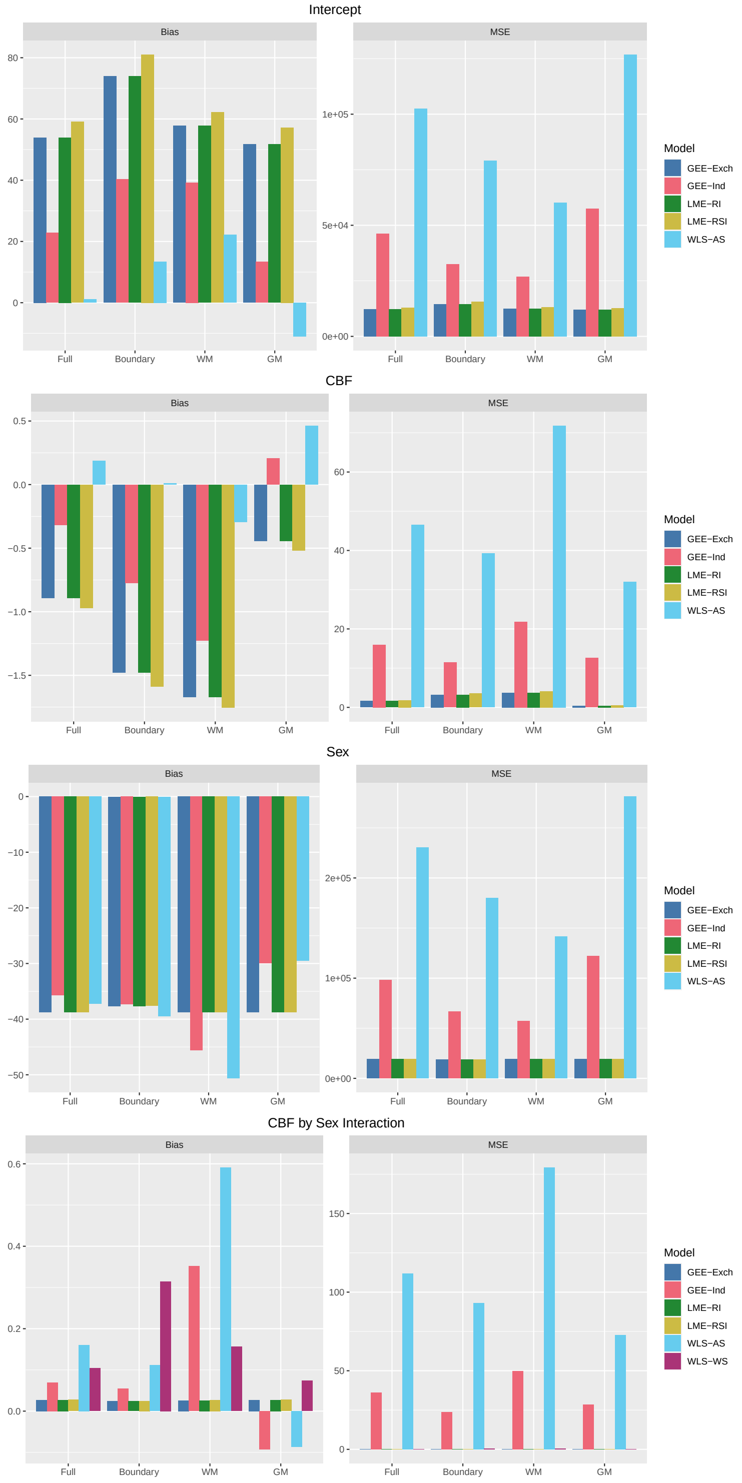


Figure 5.6: Across-subject parameter bias and MSE estimates from the n = 100 (100 iterations) simulation setting are presented in a bar chart. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Table 5.5: Across-subject parameter bias estimates from the n = 100 (100 iterations) simulation setting are presented. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
Intercept	**GEE-Exch**	53.93	51.73	57.75	73.96
Intercept	**GEE-Ind**	22.84	13.38	39.26	40.28
Intercept	**LME-RI**	53.93	51.73	57.75	73.96
Intercept	**LME-RSI**	59.05	57.23	62.20	81.02
Intercept	**WLS-AS**	1.11	-11.00	22.16	13.45
CBF	**GEE-Exch**	-0.89	-0.44	-1.67	-1.48
CBF	**GEE-Ind**	-0.32	0.21	-1.22	-0.77
CBF	**LME-RI**	-0.89	-0.44	-1.67	-1.48
CBF	**LME-RSI**	-0.97	-0.52	-1.75	-1.59
CBF	**WLS-AS**	0.19	0.46	-0.29	0.01
Sex	**GEE-Exch**	-38.79	-38.80	-38.77	-37.65
Sex	**GEE-Ind**	-35.69	-29.98	-45.61	-37.35
Sex	**LME-RI**	-38.79	-38.80	-38.77	-37.65
Sex	**LME-RSI**	-38.75	-38.75	-38.75	-37.60
Sex	**WLS-AS**	-37.23	-29.50	-50.64	-39.45
CBF by Sex Interaction	**GEE-Exch**	0.03	0.03	0.03	0.02
CBF by Sex Interaction	**GEE-Ind**	0.07	-0.09	0.35	0.05
CBF by Sex Interaction	**LME-RI**	0.03	0.03	0.03	0.02
CBF by Sex Interaction	**LME-RSI**	0.03	0.03	0.03	0.02
CBF by Sex Interaction	**WLS-AS**	0.16	-0.09	0.59	0.11
CBF by Sex Interaction	**WLS-WS**	0.10	0.07	0.16	0.32

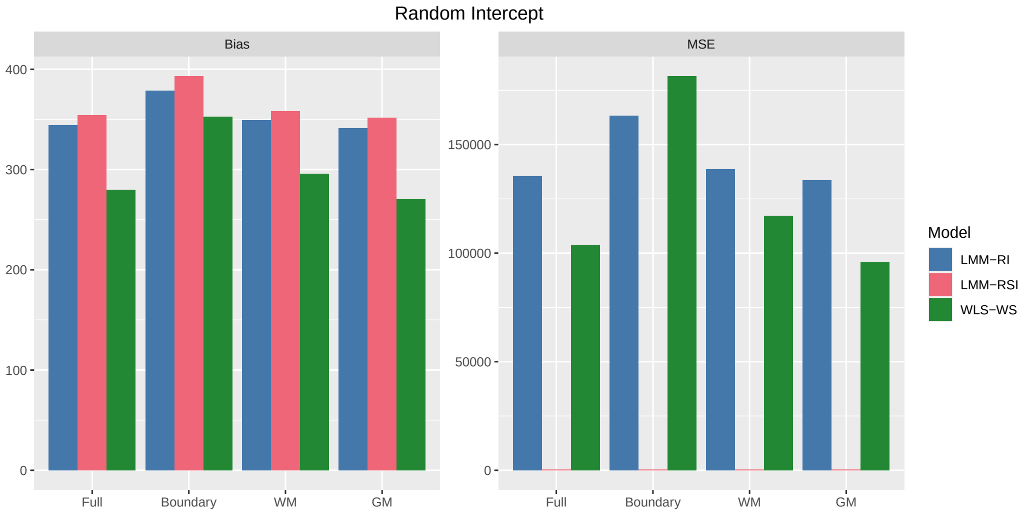


Figure 5.7: Within-subject parameter bias and MSE estimates from the n = 30 (100 iterations) simulation setting are presented in a bar chart. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Table 5.6: Across-subject parameter MSE estimates from the n = 100 (100 iterations) simulation setting are presented. We present average MSE calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
Intercept	**GEE-Exch**	12221.94	12033.42	12549.52	14576.54
Intercept	**GEE-Ind**	46292.44	57449.22	26907.22	32561.95
Intercept	**LME-RI**	12221.95	12033.42	12549.52	14576.55
Intercept	**LME-RSI**	12910.58	12746.65	13195.42	15667.65
Intercept	**WLS-AS**	102510.52	126826.41	60260.96	79211.22
CBF	**GEE-Exch**	1.65	0.44	3.76	3.21
CBF	**GEE-Ind**	15.98	12.61	21.82	11.45
CBF	**LME-RI**	1.65	0.44	3.76	3.21
CBF	**LME-RSI**	1.83	0.52	4.10	3.52
CBF	**WLS-AS**	46.53	31.97	71.84	39.26
Sex	**GEE-Exch**	19178.84	19189.88	19159.66	18743.40
Sex	**GEE-Ind**	98302.82	121889.12	57320.95	66620.21
Sex	**LME-RI**	19178.85	19189.88	19159.66	18743.41
Sex	**LME-RSI**	19180.52	19190.31	19163.51	18748.01
Sex	**WLS-AS**	230227.46	281216.30	141632.88	179773.68
CBF by Sex Interaction	**GEE-Exch**	0.01	0.01	0.02	0.02
CBF by Sex Interaction	**GEE-Ind**	36.17	28.38	49.72	23.62
CBF by Sex Interaction	**LME-RI**	0.01	0.01	0.02	0.02
CBF by Sex Interaction	**LME-RSI**	0.01	0.01	0.01	0.02
CBF by Sex Interaction	**WLS-AS**	111.67	72.78	179.23	93.05
CBF by Sex Interaction	**WLS-WS**	0.25	0.19	0.37	0.57

Table 5.7: Within-subject parameter bias estimates from the n = 30 (100 iterations) simulation setting are presented. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
b_{0i}	**LME-RI**	343.99	341.07	349.07	378.85
b_{0i}	**LME-RSI**	354.21	351.80	358.40	393.12
b_{0i}	**WLS-WS**	279.64	270.32	295.73	352.85

Table 5.8: Within-subject parameter MSE estimates from the n = 30 (100 iterations) simulation setting are presented. We present average MSE calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
b_{0i}	**LME-RI**	135401.85	133580.50	138566.5	163389.87
b_{0i}	**LME-RSI**	354.21	351.80	358.4	393.12
b_{0i}	**WLS-WS**	103823.53	96102.65	117157.7	181491.34

Table 5.9: Within-subject parameter bias estimates from the n = 100 (100 iterations) simulation setting are presented. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
b_{0i}	**LME-RI**	647.76	643.34	655.44	677.79
b_{0i}	**LME-RSI**	656.59	652.83	663.12	689.78
b_{0i}	**WLS-WS**	279.93	270.09	296.94	354.88

Table 5.10: Within-subject parameter MSE estimates from the n = 100 (100 iterations) simulation setting are presented. We present average MSE calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

Term	Model	Full	GM	WM	Boundary
b_{0i}	**LME-RI**	424920.64	419413.84	434488.86	475399.63
b_{0i}	**LME-RSI**	656.59	652.83	663.12	689.78
b_{0i}	**WLS-WS**	104397.85	95737.03	119355.31	184454.98

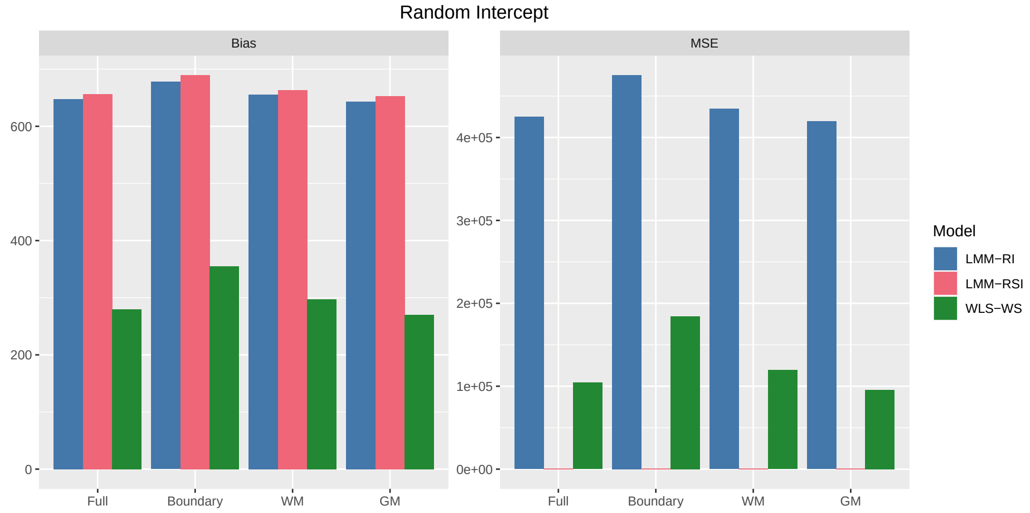


Figure 5.8: Within-subject parameter bias and MSE estimates from the n = 30 (100 iterations) simulation setting are presented in a bar chart. We present average bias calculated from voxels in the full brain (Full), gray matter only (GM), white matter only (WM), and where the white and gray matter IMCo neighborhoods overlap (Boundary).

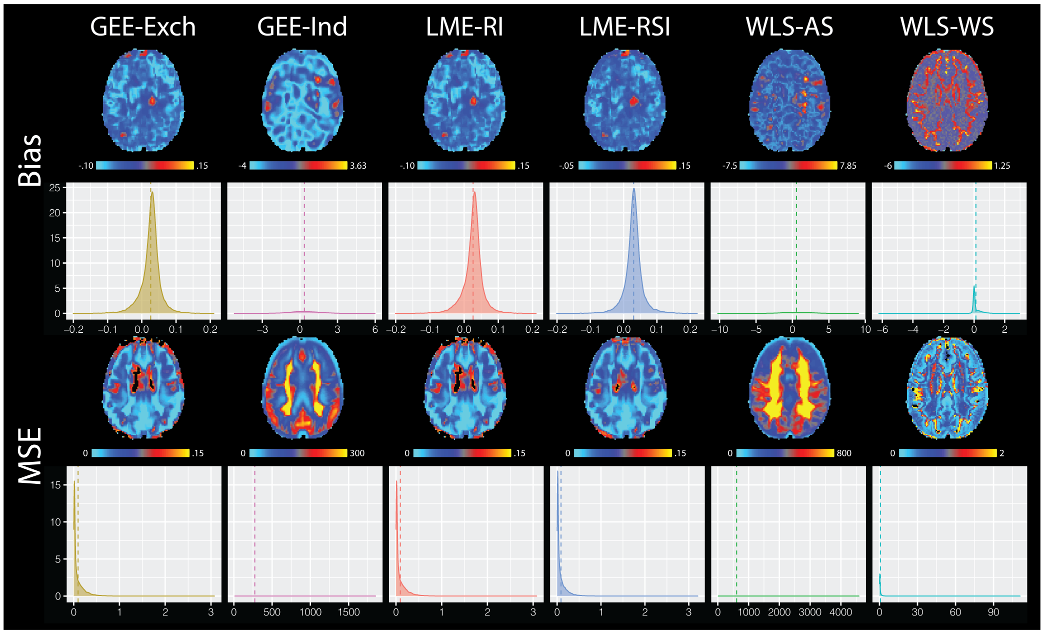


Figure 5.9: Axial slices of bias (row 1) and MSE (row 3) maps are presented for across-subject model estimates. Below the bias and MSE maps we present histograms of full brain bias and MSE. Results are from simulation setting n = 30 (100 iterations). The color bars for each map and x-axis for the histogram representations use different scales across models.

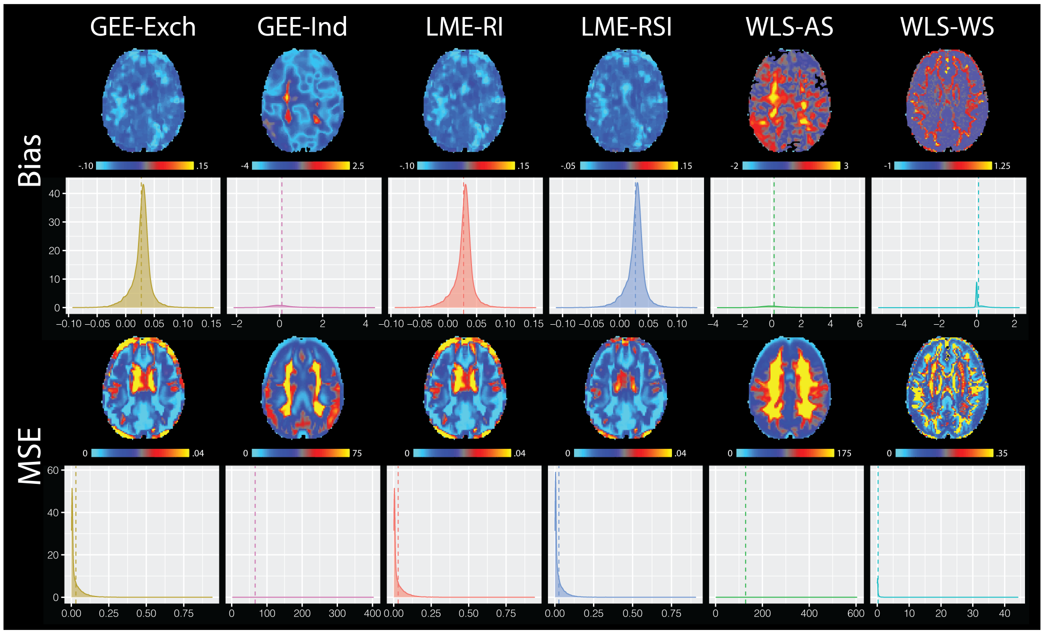


Figure 5.10: Axial slices of bias (row 1) and MSE (row 3) maps are presented for across-subject model estimates. Below the bias and MSE maps we present histograms of full brain bias and MSE. Results are from simulation setting n = 100 (100 iterations). The color bars for each map and x-axis for the histogram representations use different scales across models.

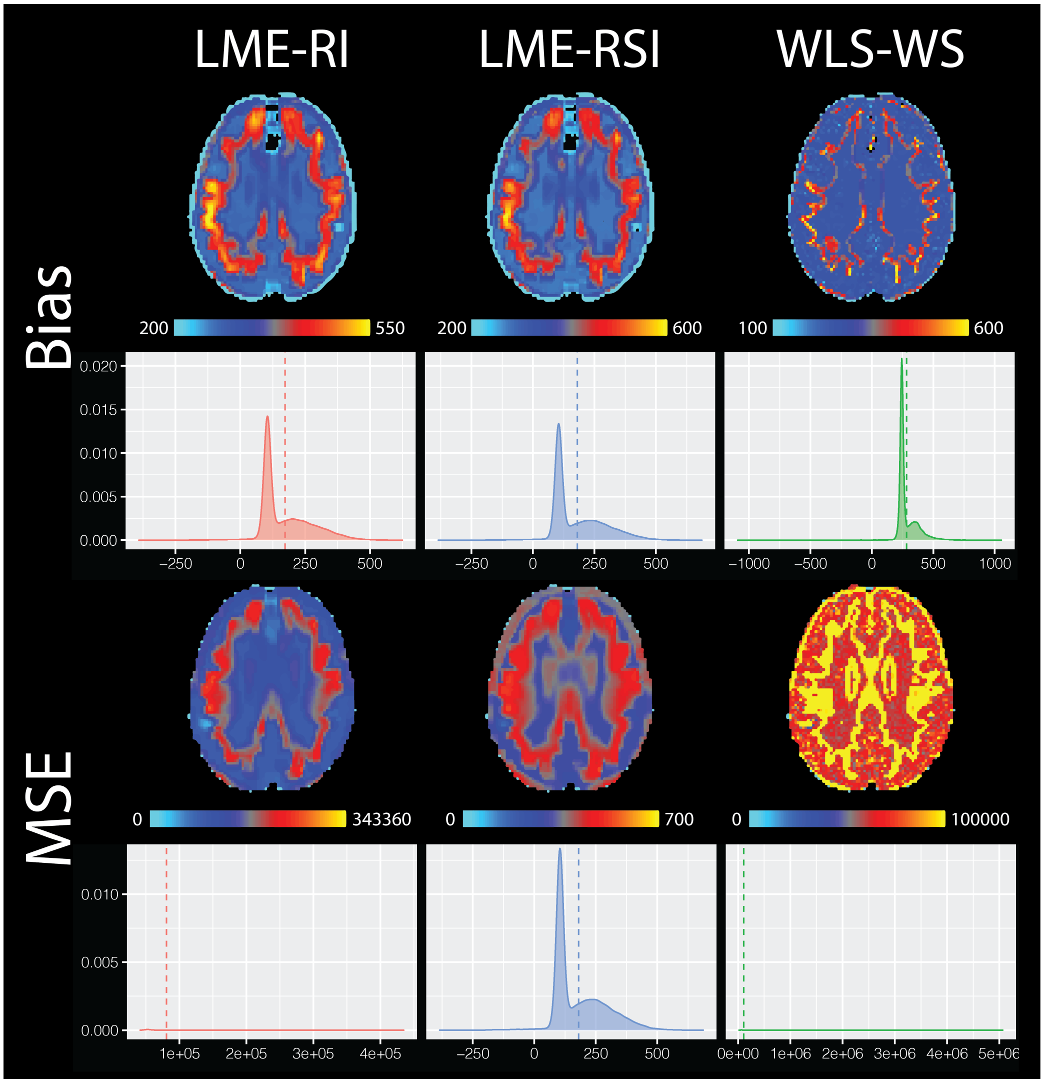


Figure 5.11: Axial slices of bias (row 1) and MSE (row 3) maps are presented for within-subject model estimates. Below the bias and MSE maps we present histograms of full brain bias and MSE. Results are from simulation setting n = 30 (100 iterations). The color bars for each map and x-axis for the histogram representations use different scales across models.

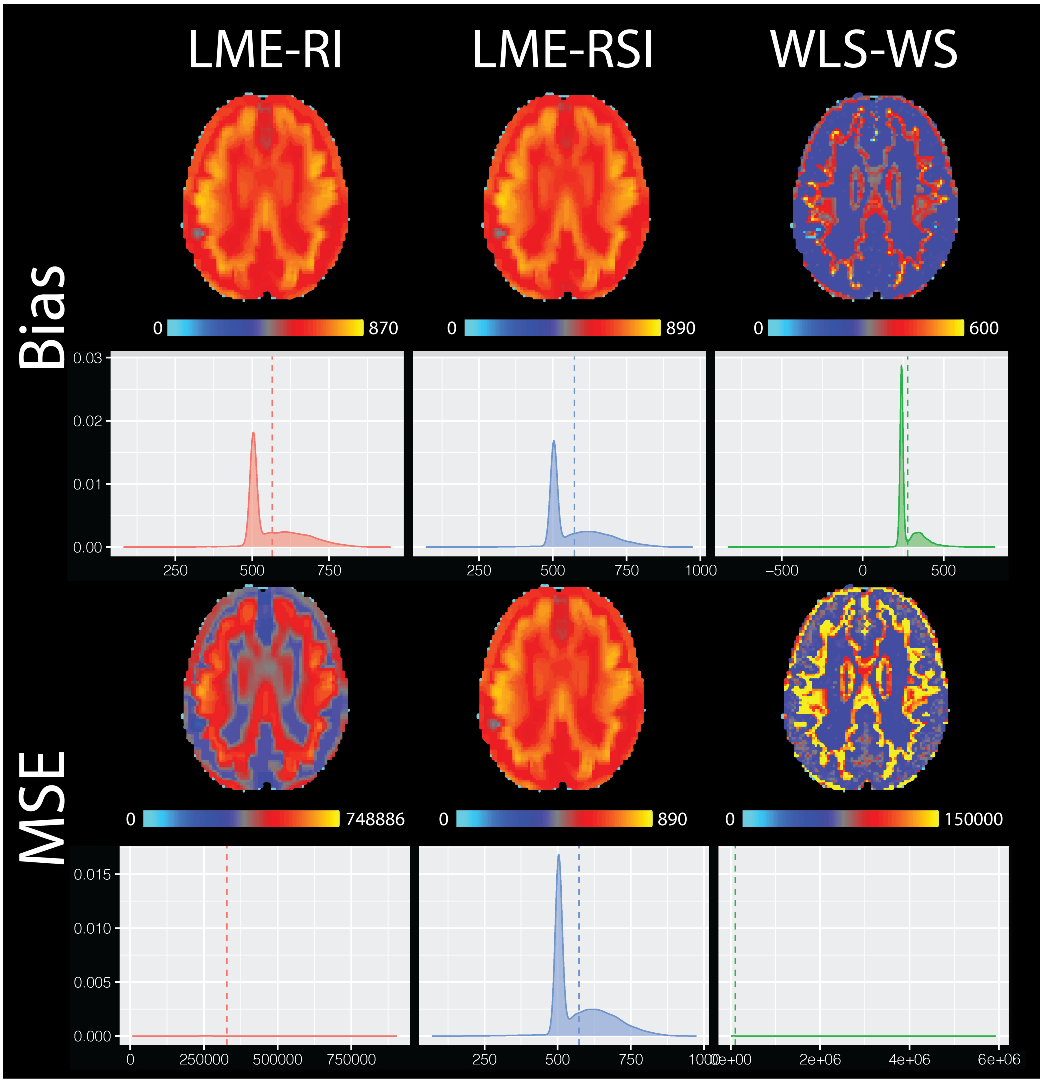


Figure 5.12: Axial slices of bias (row 1) and MSE (row 3) maps are presented for within-subject model estimates. Below the bias and MSE maps we present histograms of full brain bias and MSE. Results are from simulation setting n = 100 (100 iterations). The color bars for each map and x-axis for the histogram representations use different scales across models.

pling (IMCo) utilizes local neighborhoods of voxels to relate multi-modal imaging data. In this way, we leverage spatial information encoded in each imaging modality to improve signal estimation and detection. All proposed IMCo estimation approaches allow for across-subject or group-level estimation and inference. A subset of IMCo methods allows for estimation of subject-specific measures. Subject-specific measures can be used for novel feature development, image fusion, or to simply explore and better understand subject-level effects. The WLS-WS subject-level estimates have already been included as features in a competitive automatic segmentation algorithm to delineate multiple sclerosis lesions (Valcarcel et al., 2018a,b). The inclusion of IMCo features improved segmentation quality.

The estimation approaches leveraged in this work have been extensively investigated and validated in the longitudinal data analysis field. Though the theoretical assumptions of each estimation approach (LME, GEE, WLS-AS, and WLS-WS) seem plausible in our neuroimaging application, prior to this study, the proposed methods had not been rigorously examined. In this work, we carry out a simulation study to assess estimation accuracy and efficiency for a set of estimation approaches where we fit a variety of models using a sample size of $n = 30$ and $n = 100$.

Predominately, models closest to correct specification (LME-RI, LME-RSI, and GEE-Exch) perform best. The performance of incorrectly specified models like the WLS-WS and GEE-Ind are reasonable. The WLS-AS model is biased and unstable and we caution its use in practice unless a priori information is known about a suitable weight. Comparing the results of the two sample sizes, $n = 30$ versus $n = 100$, we find decreases in bias and MSE across models when a larger sample size is used. Many of the proposed methods rely on asymptotic normality so this is expected. These results provide insights for real multi-modal image analysis. We suggest comprehensive data exploration and using a priori information prior to model fitting. Additionally, it may be useful to carry out model selection using multiple IMCo models in order to determine a single best performing model or to combine results across a few estimation approaches.

There are several notable limitations to the proposed work. First, in order to carry out across-subject IMCo analyses, such as those performed here, all images must be registered to a common template. IMCo performance will likely suffer if implemented on images that were poorly registered or when registration has failed. Second, in order to run IMCo, we must first load both image modalities for a subject into memory, extract the local neighborhoods for each target voxel, and then run the models at every voxel location. This process is memory intensive and slow. Depending on the scanner strength, resolution, and acquisition protocol, brain masks may include millions of voxels that consume memory and require modeling. We have developed a publicly available R package, NiftiArray (Muschelli and Valcarcel, 2020), to reduce memory consumption and are currently working on developing methods to increase computation speeds. Lastly, in real multi-modal data analysis the local spatial relationship between modalities may be endogenous. That is, the independent modality at one location within a local neighborhood may affect the dependent modality of a voxel in a different location within the neighborhood. In the presence of endogenous variables, GEE models can result in biased estimates with all working correlation structures except for independence (Pepe and Anderson, 1994).

Future developments will include simulation studies with both larger sample sizes and increased number of iterations. Due to computational challenges and limitations, we only repeated sampling across 100 iterations. Additionally, comparing the simulation settings with $n = 30$ and $n = 100$ we notice only inconsistent reductions in bias and MSE. To elucidate the effect of sample size, we will re-run analyses with larger sample sizes. At present, analyses are carried out in the full brain mask even though the data are generated with distinct effects in brain white and gray matter. Neighborhoods at the boundary of the tissue class use voxels from both classes in the estimation procedure. While LME and GEE methods do not show increased bias levels at the boundary between white and gray matter, the WLS-WS does have increased bias in these regions. We would like to extend methods and carry out the estimation procedures regionally (i.e. in the white and gray matter separately) to better understand boundary effects. In future studies, we will explore the impact of neighborhood size on estimation accuracy and efficiency. To minimize memory consumption, we used only a neighborhood size of $3 \times 3 \times 3$ voxels. The methods presented in this work include only two imaging modalities but easily extend to accommodate multivariate regression using more than one independent modality. Lastly, we would like to better characterize inferential performance for the IMCo parameter estimates.

CHAPTER 6

DISCUSSION

Multi-modal neuroimaging approaches are used to better understand patterns of healthy brain development as well as detection, diagnosis, and prognosis of many disorders. The continued growth and development of novel imaging contrasts and sequences coupled with the advancement of ultra-high resolution scanners is leading to visualizations of the brain in unprecedented detail. The increasing complexity and dimensionality due to these cutting-edge neuroimaging techniques will demand parallel advances of computational and statistical approaches to understand such large and complex data. In this paper, we present new methods for multi-modal neuroimaging analysis as well as applications of multi-modal imaging techniques to study multiple sclerosis and neurodevelopment. As new imaging contrasts and sequences are developed and scanners become more powerful, these methods will remain applicable and allow for unmatched discoveries of brain structure and function.

In Chapter 2 we propose a fully automated segmentation method, MIMoSA, that utilizes the changes in inter-modality covariance structure that occur in white matter pathology, and can be used to assist in white-matter lesion detection or replace manual segmentation. MIMoSA avoids the variability associated with manual and semi-automated lesion segmentation. The model can be easily adapted and trained in cases where fewer imaging sequences are available. Though originally designed to segment T2L, MIMoSA performs well at segmenting T1 black holes in patients with multiple sclerosis.

In Chapter 3 we introduce a statistical technique, TAPAS, for reducing the volumetric bias between probabilistic lesion segmentation algorithm and a manual rater by optimizing similarity metrics among raters. TAPAS reduces bias in brain lesion volume estimates with automatic segmentation approaches. The proposed pipeline allows a more accurate estimation of lesion volume at the subject-level compared to traditional thresholding of probability maps for lesion segmentation which only offer group-level threshold estimates.

R is a computing environment which contains free and open source software for statistical analysis and visualization. As such, R is a necessary tool for carrying out statistical methodological research in imaging. Unfortunately, R is not well suited for big data. Imaging data can be enormous due to resolution, repeated measures, and large sample size. In Chapter 4, we develop `NiftiArray` which allows for memory efficient representations of imaging data in R. This work leverages and adapts a number of statistical tools built in the genomics field to overcome big data limitations of R. The `NiftiArray` package is compatible with `DelayedMatrixStats` and therefore can quickly calculate voxel-wise statistics quickly. To carry out more complex statistical analyses, `NiftiArray` is compatible with `DelayedArray` which reduces memory usage by applying either delaying operations on the object or executing operations using a block processing mechanism.

In Chapter 5, we propose a novel multi-modal analysis framework which we refer to as IMCo analysis. Current multi-modal imaging techniques do not leverage the spatially correlated data structure naturally present within an image. In this chapter, we quantify the covariability across image modalities by regressing local neighborhoods from the images onto each other. To estimate the covariability between the modalities, we use linear mixed effects models, generalized estimating equations, and weighted least squares (one- and two-stage). Each method accounts for the spatial correlation of voxels within a neighborhood. We assess the accuracy of each estimation approach using a large simulation study. Generally, we found all estimation approaches to perform well in terms of minimizing bias and mean square error when models are correctly specified. Under misspecification, certain estimation approaches did not provide accurate estimation. The `NiftiArray` software package built in Chapter 4 enabled research related to IMCo to be completed within a reasonable time. Without the memory saving tool, the large scale simulation would not have been possible due to memory and time constraints associated with these analyses.

In an effort to facilitate reproducible and replicable research, publicly available software packages with documentation and tutorials accompany each method discussed in this work. Resources are available on GitHub and Neuroconductor. Details on software are provided in Appendix sections B.1, B.2, B.3, and B.4.

APPENDIX A

CHAPTER 3 EXTENSIONS

Please take note that Figure A.5, Figure A.10, and Figure A.15 in this appendix are re-creations of Figure 3.7 provided in Chapter 3. These figures include example slices from the various analyses. Each figure is made using the same test subject within the same cross-validation fold. Approximately the same slice was used for each representation, but the exact slice was not possible due to differing processing pipelines yielding images in slightly different spaces.

A.1. Training TAPAS with Absolute Error

In this section of the appendix we present results using the same data and processing pipeline described in Chapter 3. We first apply MIMoSA to obtain probability maps and then implement the thresholding algorithm as described in section 3.2.2 of the paper except we re-define $\hat{\tau}_{Group}$ and $\hat{\tau}_i$ as $\hat{\tau}^{AE}_{Group}$ and $\hat{\tau}^{AE}_i$, respectively, to emphasize that the optimization approach involves the minimization of absolute error (AE) rather than maximization of the Sorensen-Dice coefficient (DSC).

1. $\hat{\tau}^{AE}_{Group} = \underset{\tau \in \{\tau_1,\dots,\tau_J\}}{\arg\min} \dfrac{2\sum_{i=1}^{N/2} AE_i(\tau)}{N}$, and
2. $\hat{\tau}^{AE}_i = \underset{\tau \in \{\tau_1,\dots,\tau_J\}}{\arg\min} \{AE_i(\tau)\}$ for each subject i.

We choose a threshold grid of $\tau_1 = 0\%$ to $\tau_J = 100\%$ in 1% increments. We use the exact Monte Carlo-resampled split-sample cross-validation method. That is, the same subjects assigned to training and testing sets across iterations in Chapter 3 are used here. The only difference is that in the TAPAS algorithm we optimize using AE.

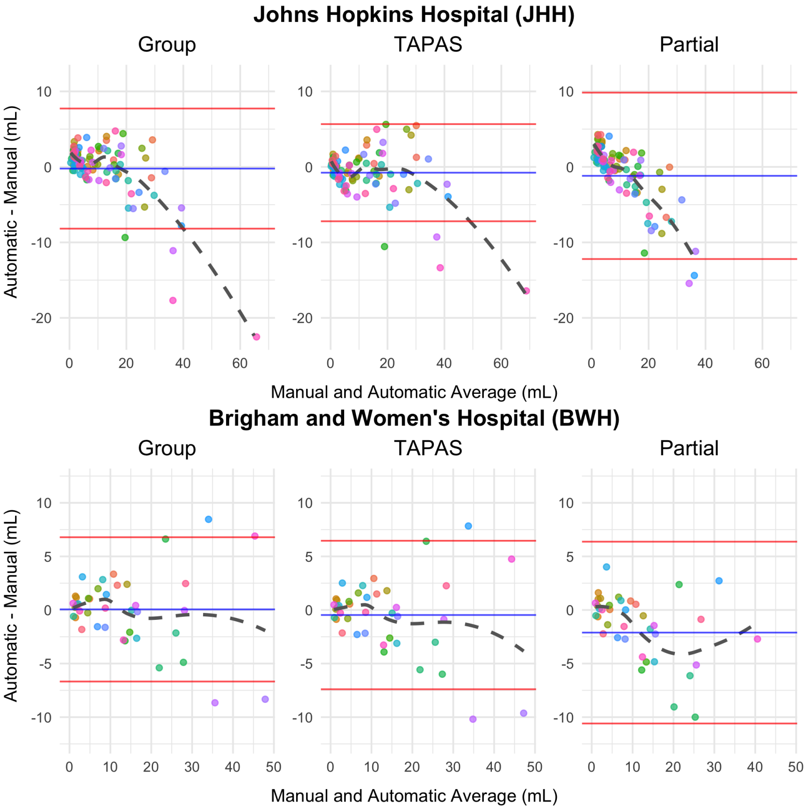


Figure A.1: Bland-Altman plots comparing $volume_{Manual}$ with volumes obtained using automatic thresholding approaches ($volume_{Group}$, $volume_{TAPAS}$, and $volume_{Partial}$) are shown. TAPAS is trained using absolute error (AE) rather than the Sorensen-Dice coefficient (DSC). The mean of the difference in volume is presented in blue and the mean plus and minus the standard error is shown in red. Each point represents a unique subject. Subject-specific points were obtained by averaging results across test set subjects in each split-sample fold.

The results presented in Figure A.1 are re-created from the Chapter 3 Figure 3.3. The Bland-Altman plots in these two figures are nearly identical.

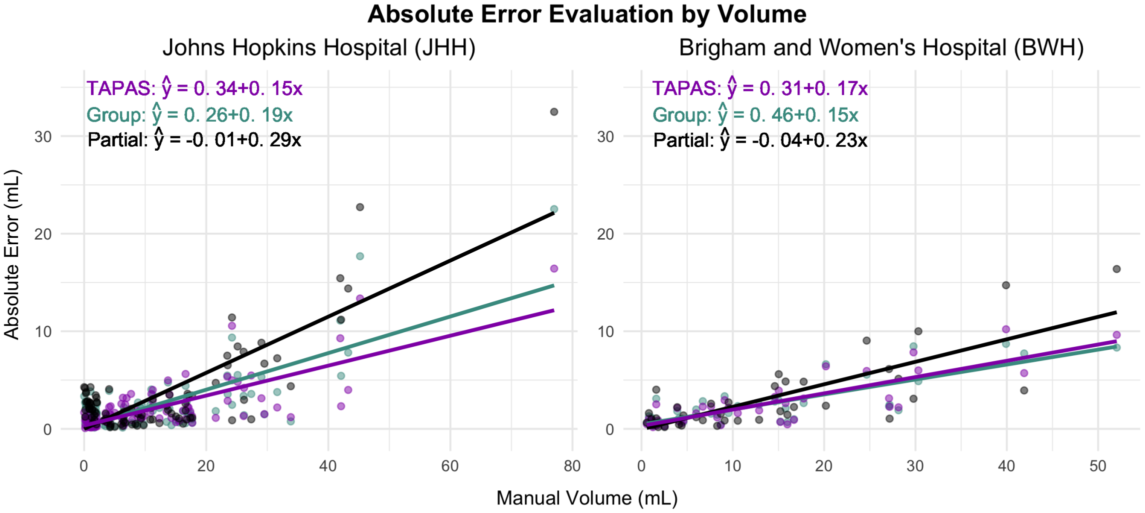


Figure A.2: Scatter plots with fitted linear models are presented for the subject-level average absolute error ($\hat{y}$) on manual volume (x) in mL in the Johns Hopkins Hospital (JHH) and Brigham and Women's Hospital (BWH) data sets. Fitted equations are given in the top left corner. The TAPAS algorithm was optimized using AE rather than the Sorensen-Dice coefficient (DSC).

The results presented in Figure A.2 are re-created from Chapter 3 Figure 3.4. The scatter plots presented here are slightly different than those presented in Chapter 3 Figure 3.4. The slope of the Johns Hopkins Hospital (JHH) group fitted line is smaller while the TAPAS slope is approximately the same. The slope of the Brigham and Women's Hospital (BWH) group and TAPAS fitted lines is similar to those presented in Chapter 3. These results indicate that TAPAS performs as well as or better than the group thresholding procedure at reducing AE.

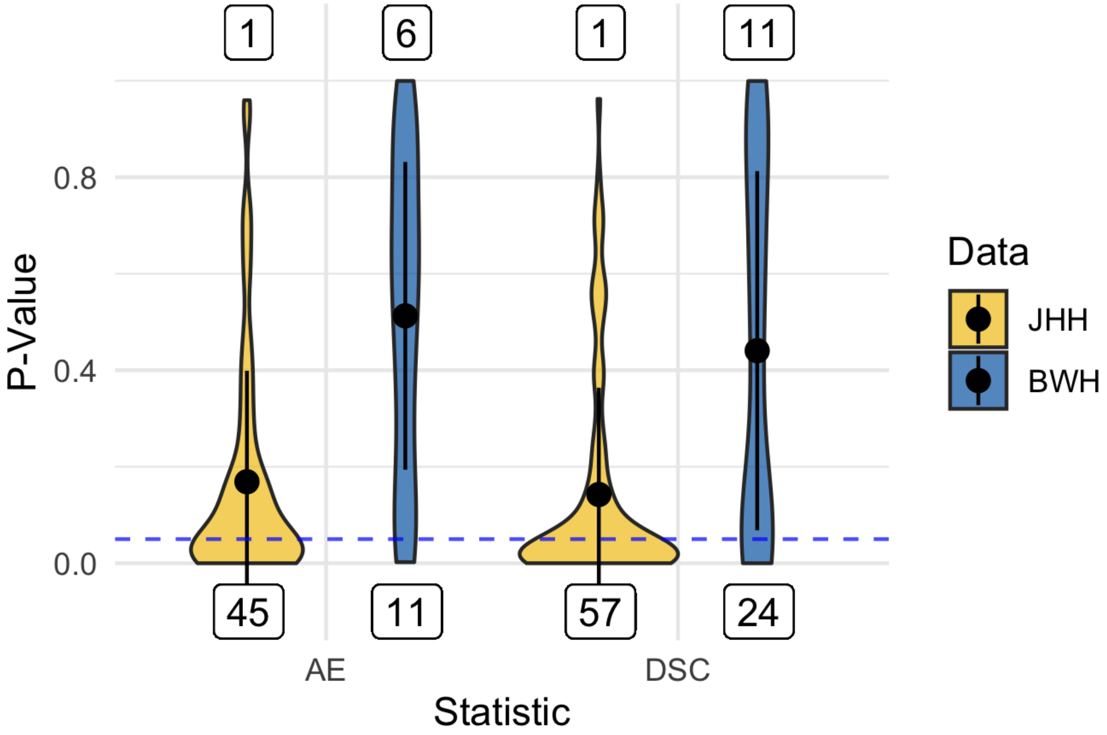


Figure A.3: Violin plots of p-values from paired t-tests to compare subject-level absolute error (AE) and Sorensen-Dice coefficient (DSC) in each test set are presented. The mean for each statistic and data set is presented as points within each violin plot and the black lines extend the mean by the standard deviation. Labels below represent the number of significant p-values favoring TAPAS performance measures. Labels above represent the number of significant p-values favoring group thresholding performance measures. The dashed horizontal blue line highlights the $\alpha = 0.05$ cutoff. The TAPAS algorithm was optimized using AE rather than DSC.

The results presented in Figure A.3 are re-created from Chapter 3 Figure 3.5. We employed one-sided paired t-tests to evaluate AE and DSC from TAPAS compared with those obtained from the group thresholding procedure. Figure A.3 shows violin plots of p-values from both sets of tests for the two data sets. In the JHH data approximately half of the split-sample experiments resulted in p-values below the $\alpha = 0.05$ for AE and DSC with only one statistically significant result favoring the group thresholding procedure. This indicates superior performance using TAPAS compared to the group thresholding procedure. The BWH data is more uniform with statistically significant results favoring TAPAS over the group thresholding procedure only slightly. The number of statistically significant improvements in AE and DSC in the JHH data are somewhat smaller compared to findings in Chapter 3. Generally, TAPAS thresholding still results in as good or better performance compared with group thresholding.

Table A.1: Subject-specific volume estimates, $volume_{Manual}$ (Manual), $volume_{TAPAS}$ (TAPAS), $volume_{Group}$ (Group), and $volume_{Partial}$ (Partial), were compared with clinical covariates available in each data set and are represented in this table. Spearman's correlation coefficient ($\hat{\rho}$) was obtained in the testing set for each iteration and averaged across folds. Clinical variables included Expanded Disability Status Scale (EDSS) score, disease duration in years, and timed 25-ft walk (T25FW) in seconds. The TAPAS algorithm was optimized using absolute error (AE) rather than the Sorensen-Dice coefficient (DSC).

	Estimates for $\hat{\rho}$			
	Partial	**Group**	**TAPAS**	**Manual**
JHH				
EDSS	0.32	0.34	0.33	0.29
Disease duration	0.37	0.39	0.39	0.39
BWH				
EDSS	0.42	0.43	0.42	0.45
Disease duration	0.31	0.30	0.31	0.29
T25FW	0.02	0.01	0.02	0.03

The results presented in Table A.1 are re-created from Chapter 3 Table 3.2. The results presented in this table are nearly identical to those in Chapter 3. The Chapter 3 correlation estimates tend to be slightly stronger for some clinical covariates.

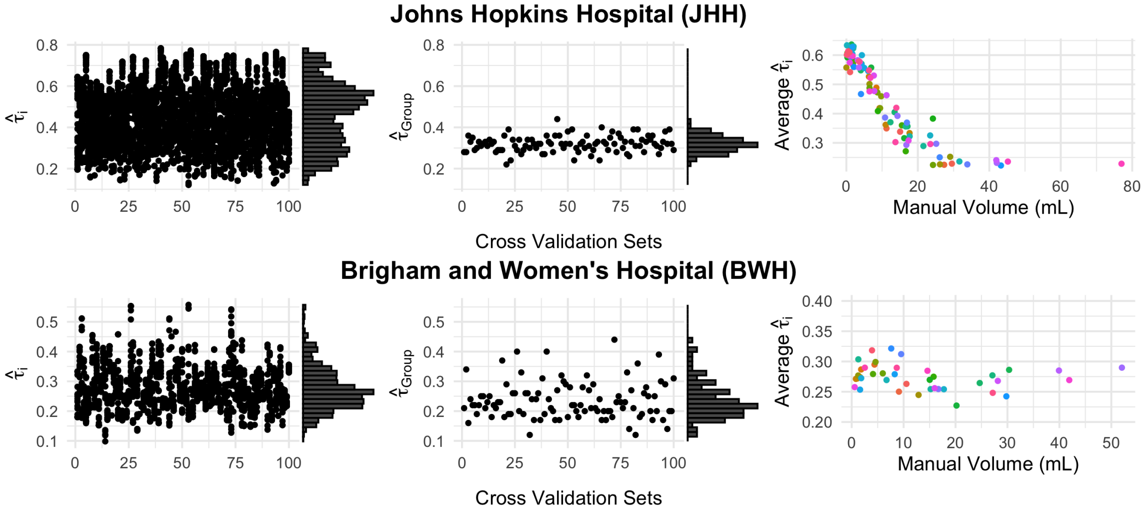


Figure A.4: Scatter plots of the subject-specific threshold $\hat{\tau}_i$ (TAPAS) and $\hat{\tau}_{Group}$ (group thresholding procedure) on cross-validation number are presented with marginal histograms for both data sets in the first two columns. The third column presents scatterplots of the average subject-specific thresholds from TAPAS and the manually delineated lesion volume. The TAPAS algorithm was optimized using absolute error (AE) rather than the Sorensen-Dice coefficient (DSC).

The results presented in Figure A.4 are re-created from Chapter 3 Figure 3.6. Generally, these figures are similar in shape and spread. The group threshold scatter plots for both JHH and BWH data have a smaller range of selected $\hat{\tau}_{Group}$.

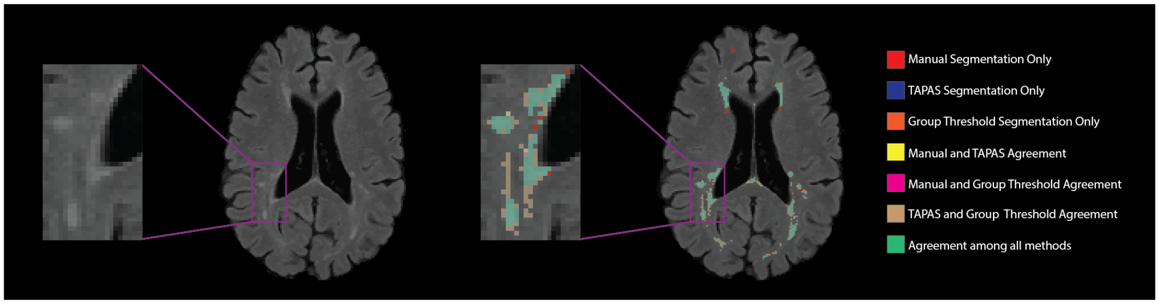

Manual Segmentation Only
TAPAS Segmentation Only
Group Threshold Segmentation Only
Manual and TAPAS Agreement
Manual and Group Threshold Agreement
TAPAS and Group Threshold Agreement
Agreement among all methods

Figure A.5: T2 hyperintense lesion segmentations from an example axial slice are displayed. The colors represent the different individual or overlapping segmentations obtained from manual, TAPAS threshold, and group threshold masks. The majority of segmented area was in agreement among all lesion masks (green). Both the group thresholding approach and TAPAS missed some area that was manually segmented (red). There was a small area where only TAPAS and manual segmentations agreed (yellow), but almost no area where only the group threshold agreed with the manual segmentation (fuchsia). The TAPAS algorithm was optimized using absolute error (AE) rather than the Sorensen-Dice coefficient (DSC).

The results presented in Figure A.5 are re-created from Chapter 3 Figure 3.7.

A.2. Training TAPAS with the Sorensen-Dice coefficient using Brigham and Women's Hospital Preprocessing Pipeline

In this section of the appendix we implement the TAPAS algorithm and replicate the cross-validation as described in section 3.2.2 and section 3.2.3, respectively, of Chapter 2. We preprocessed the JHH data using the same processing pipeline as the BWH data described in the section 3.2.1 of Chapter 3. That is, we performed N4 bias correction (Tustison et al., 2010) on all images and rigidly co-registered T1 and T2 images for each participant to the FLAIR at 1 mm^3 resolution. Extracerebral voxels were removed from the registered T1 images using Multi-Atlas Skull Stripping (MASS) (Doshi et al., 2013) and the brain mask was applied to the FLAIR and T2 scans. We intensity-normalized images to facilitate across-subject modeling of intensities using *WhiteStripe* (Muschelli and Shinohara, 2018; Shinohara et al., 2014). Image preprocessing was applied using software available in R (version 3.5.0) R Development Core Team, 2018 and from NITRC (`https://www.nitrc.org/projects/cbica_mass/`).

Additionally, we implement the algorithm as described in section 3.2.2 of Chapter 3 with a minor change to the threshold grid. In this analysis, we refine the grid of thresholds applied to range from 11% to 52% in $.4\%$ increments. We use the same Monte Carlo-resampled split-sample cross-validations described in section 3.2.3 of Chapter 3.

Only the JHH data was processed differently so the BWH data and results in this section are the same as Chapter 3 and excluded from this section of the appendix.

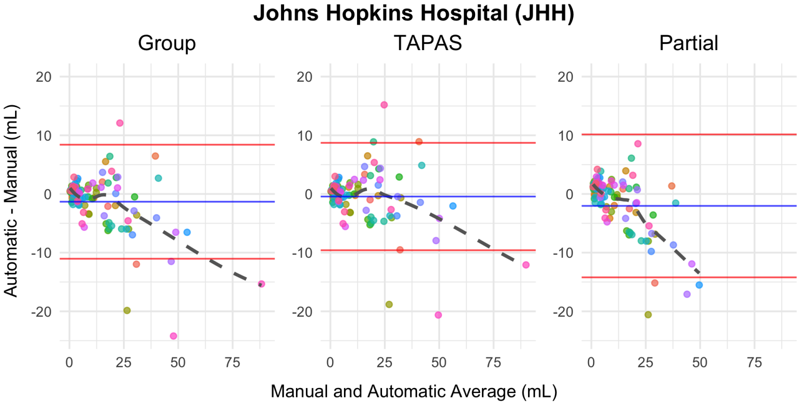


Figure A.6: Bland-Altman plots comparing $volume_{Manual}$ with volumes obtained using automatic thresholding approaches ($volume_{Group}$, $volume_{TAPAS}$, and $volume_{Partial}$) are shown. The mean of the difference in volume is presented in blue and the mean plus and minus the standard error is shown in red. Each point represents a unique subject. Subject-specific points were obtained by averaging results across test set subjects in each split-sample fold. The JHH data were processed using the same pipeline as the BWH data in Chapter 3.

The results presented in Figure A.6 are re-created from Chapter 3 Figure 3.3. The Bland-Altman plots presented for the JHH data are different here than in the original work. The JHH data begins to exhibit systematic bias for volumes exceeding 20 mL, similar to the findings in Chapter 3, but the plots presented here shows that the systematic deviation is not as dramatic as in Figure 3.3 in Chapter 3. Comparing the group and TAPAS plots on the left and right, the TAPAS plot still shows a less steep fitted line and therefore less systematic bias associated with the volumes compared to the group thresholding approach.

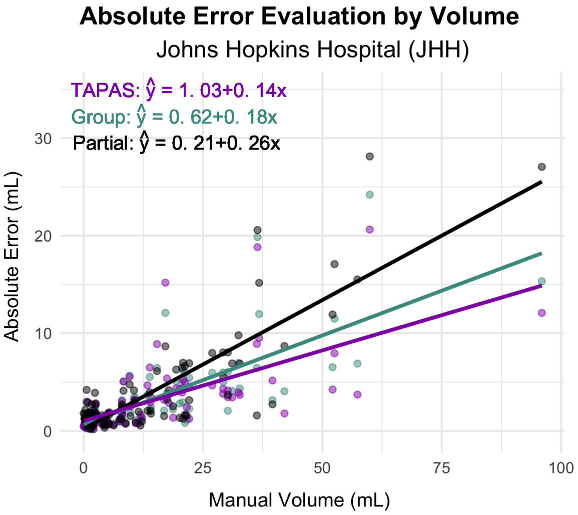


Figure A.7: Scatter plot with fitted linear models are presented for the subject-level average absolute error ($\hat{y}$) on manual volume (x) in mL. Fitted equations are given in the top left corner. The Johns Hopkins Hospital (JHH) data are processed using the same pipeline as the BWH data in Chapter 3.

The results presented in Figure A.7 are re-created from Chapter 3 Figure 3.4. The results presented in this figure are consistent with those presented in Chapter 3. The line of best fit for TAPAS using the JHH data is slightly less steep than the line presented in Chapter 3. TAPAS still shows reduced AE compared to the group thresholding procedure.

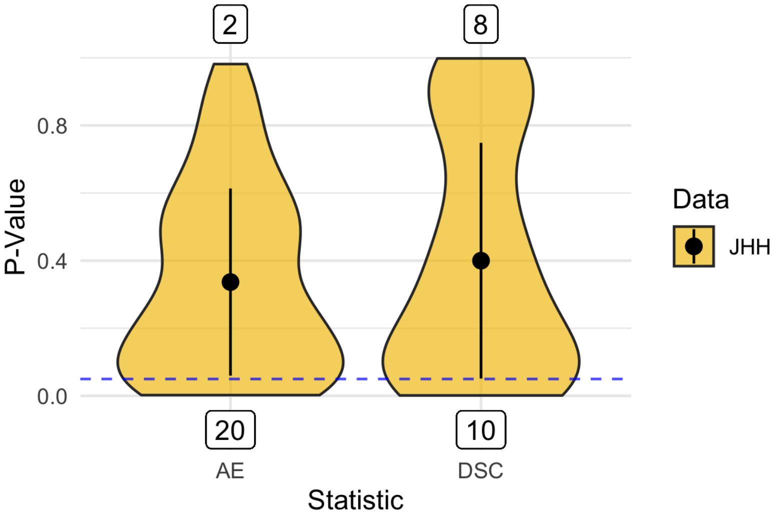


Figure A.8: Violin plots of p-values from paired t-tests to compare subject-level absolute error (AE) and Sorensen-Dice coefficient (DSC) in each test set are presented. The mean for each statistic and data set is presented as points within each violin plot and the black lines extend the mean by the standard deviation. Labels below represent the number of significant p-values favoring TAPAS performance measures. Labels above represent the number of significant p-values favoring group thresholding performance. The dashed horizontal blue line highlights the $\alpha = 0.05$ cutoff. The Johns Hopkins Hospital (JHH) data are processed using the same pipeline as the Brigham and Women's Hospital (BWH) data in Chapter 3.

The results presented in Figure A.8 are re-created from Chapter 3 Figure 3.5. We employed one-sided paired t-tests to evaluate AE and DSC from TAPAS compared with those obtained from the group thresholding procedure. Figure A.8 shows violin plots of p-values from both sets of tests for the JHH data set using the BWH data processing pipeline. More split-sample experiments resulted in p-values below the $\alpha = 0.05$ for AE and DSC with only a few statistically significant results favoring the group thresholding procedure. This indicates superior performance using TAPAS compared to the group thresholding procedure. The number of statistically significant improvements in AE and DSC in the JHH data are somewhat smaller compared to those in Chapter 3 but still favor TAPAS in a large number of cross-validation iterations.

Table A.2: Subject-specific volume estimates, $volume_{Manual}$ (Manual), $volume_{TAPAS}$ (TAPAS), $volume_{Group}$ (Group), and $volume_{Partial}$ (Partial), were compared with clinical covariates available in each data set and are represented in this table. Spearman's correlation coefficient ($\hat{\rho}$) was obtained in the testing set for each iteration and averaged across folds. Clinical variables included Expanded Disability Status Scale (EDSS) score, disease duration in years, and timed 25-ft walk (T25FW) in seconds. The Johns Hopkins Hospital (JHH) data are processed using the same pipeline as the Brigham and Women's Hospital (BWH) data in Chapter 3.

	Estimates for $\hat{\rho}$			
	Partial	**Group**	**TAPAS**	**Manual**
JHH				
EDSS	0.36	0.34	0.34	0.29
Disease duration	0.43	0.43	0.43	0.39

The results presented in Table A.2 are re-created from Chapter 3 Table 3.2. Correlation estimates are similar to those presented in Chapter 3.

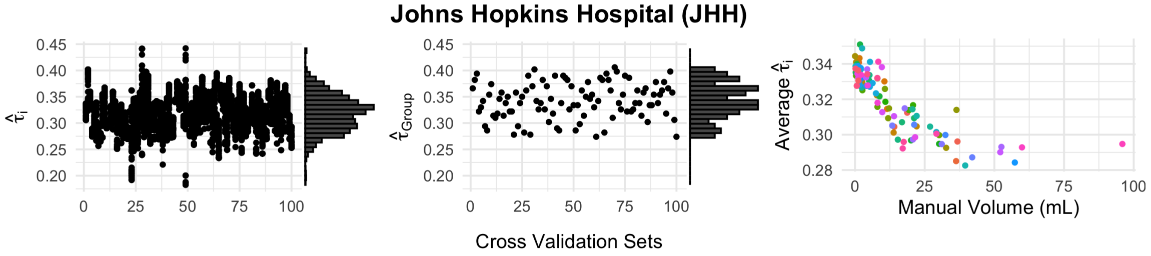


Figure A.9: Scatter plots of the subject-specific threshold $\hat{\tau}_i$ (TAPAS) and $\hat{\tau}_{Group}$ (group thresholding procedure) on cross-validation number are presented with marginal histograms for both data sets in the first two columns. The third column presents scatterplots of the average subject-specific thresholds from TAPAS and the manually delineated lesion volume. The JHH data are processed using the same pipeline as the Brigham and Women's Hosptial (BWH) data in Chapter 3.

The results presented in Figure A.9 are re-created from Chapter 3 Figure 3.6. The JHH TAPAS scatter plot shows that the range of thresholds selected at the subject level are much more narrow compared to Chapter 3 findings. Further, the bi-modal pattern in the Chapter 3 histogram is no longer present in these data and an approximately normal shape is formed. The group threshold scatter and histogram plots are similar to those in Chapter 3.

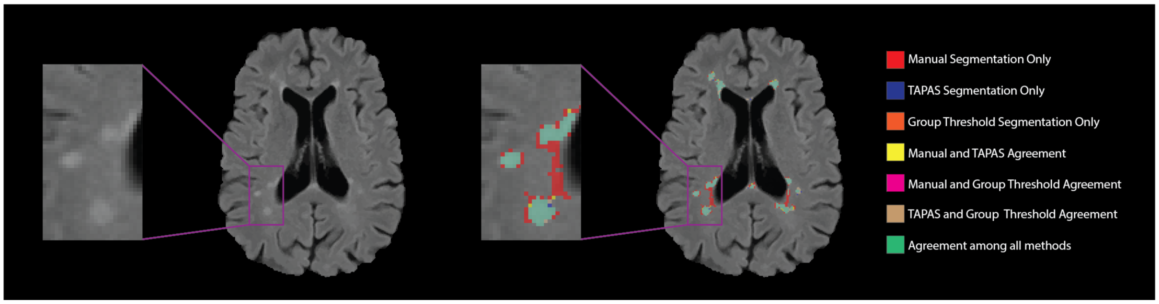

🟥 Manual Segmentation Only
🟦 TAPAS Segmentation Only
🟧 Group Threshold Segmentation Only
🟨 Manual and TAPAS Agreement
🟪 Manual and Group Threshold Agreement
🟫 TAPAS and Group Threshold Agreement
🟩 Agreement among all methods

Figure A.10: T2 hyperintense lesion segmentations from an example axial slice are displayed. The colors represent the different individual or overlapping segmentations obtained from manual, TAPAS threshold, and group threshold masks. The majority of segmented area was in agreement among all lesion masks (green). Both the group thresholding approach and TAPAS missed some area that was manually segmented (red). There was a small area where only TAPAS and manual segmentations agreed (yellow), but almost no area where only the group threshold agreed with the manual segmentation (fuchsia). The Johns Hopkins Hospital (JHH) data were processed using the same pipeline as the Brigham and Women's Hospital (BWH) data in Chapter 3.

The results presented in Figure A.10 are re-created from Chapter 3 Figure 3.7.

A.3. Training TAPAS with the Sorensen-Dice coefficient using Lesion Segmentation Tool's Lesion Prediction Algorithm

In this section of the appendix we present results using Lesion Segmentation Tool's Lesion Prediction Algorithm (LST-LPA) as the automatic segmentation method. The LST-LPA method implements its own processing pipeline as part of the algorithm. Therefore, no processing was carried out on the images before implementing LST-LPA. The processing pipeline that LST-LPA uses can be found in their documentation (`https://www.applied-statistics.de/LST_documentation.pdf`) and in their original work (Schmidt et al., 2012). We implement the algorithm as described in section 3.2.2 of Chapter 3 with a minor change to the threshold grid. In this analysis, we refine the grid of thresholds originally applied. In the JHH data we use a grid from 0% to 22% in 0.2% increments and in the BWH data we use a grid from 12% to 55% in $.4\%$ increments. We use the same Monte Carlo-resampled split-sample cross-validations described in section 3.2.3 of Chapter 3.

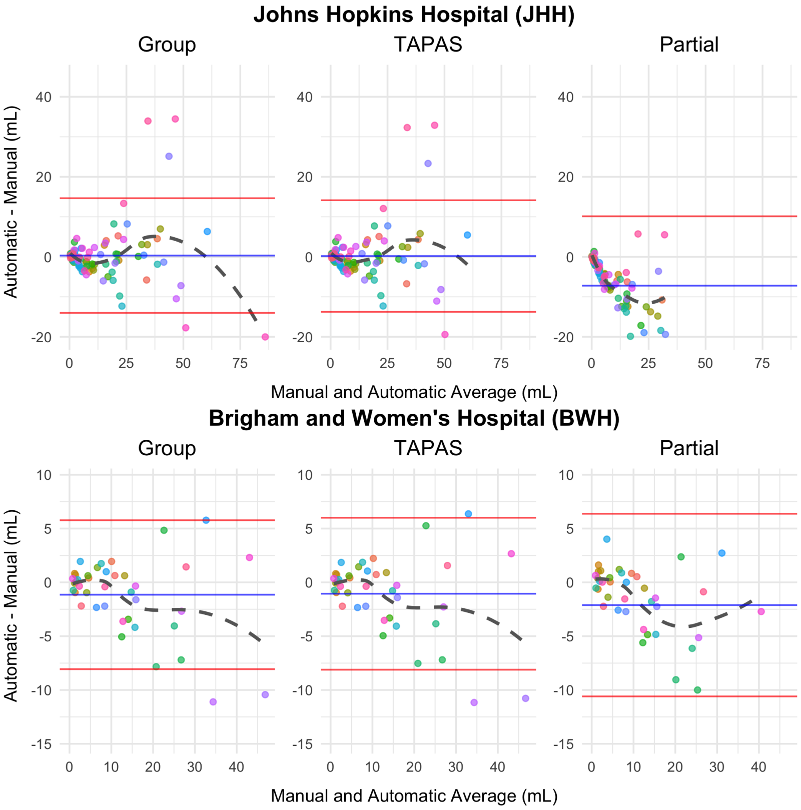


Figure A.11: Bland-Altman plots comparing $volume_{Manual}$ with volumes obtained using automatic thresholding approaches ($volume_{Group}$, $volume_{TAPAS}$, and $volume_{Partial}$) are shown. The mean of the difference in volume is presented in blue and the mean plus and minus the standard error is shown in red. Each point represents a unique subject. Subject-specific points were obtained by averaging results across test set subjects in each split-sample fold.

The results presented in Figure A.11 are re-created from Chapter 3 Figure 3.3. The Bland-Altman plots presented for the BWH data are nearly identical to those presented in Chapter 3. The plots using the JHH data though are different than those in Chapter 3 and show both the group and TAPAS thresholding procedures perform similarly.

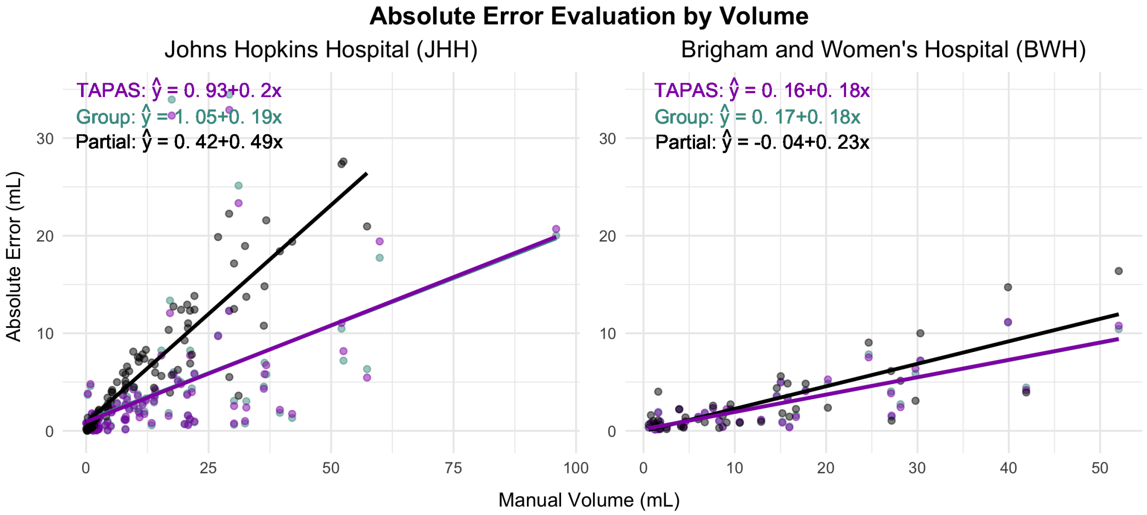


Figure A.12: Scatter plots with fitted linear models are presented for the subject-level average absolute error ($\hat{y}$) on manual volume (x) in mL. Fitted equations are given in the top left corner.

The results presented in Figure A.12 are re-created from Chapter 3 Figure 3.4. The scatter plots presented here are show similar results using either TAPAS or the group thresholding procedure.

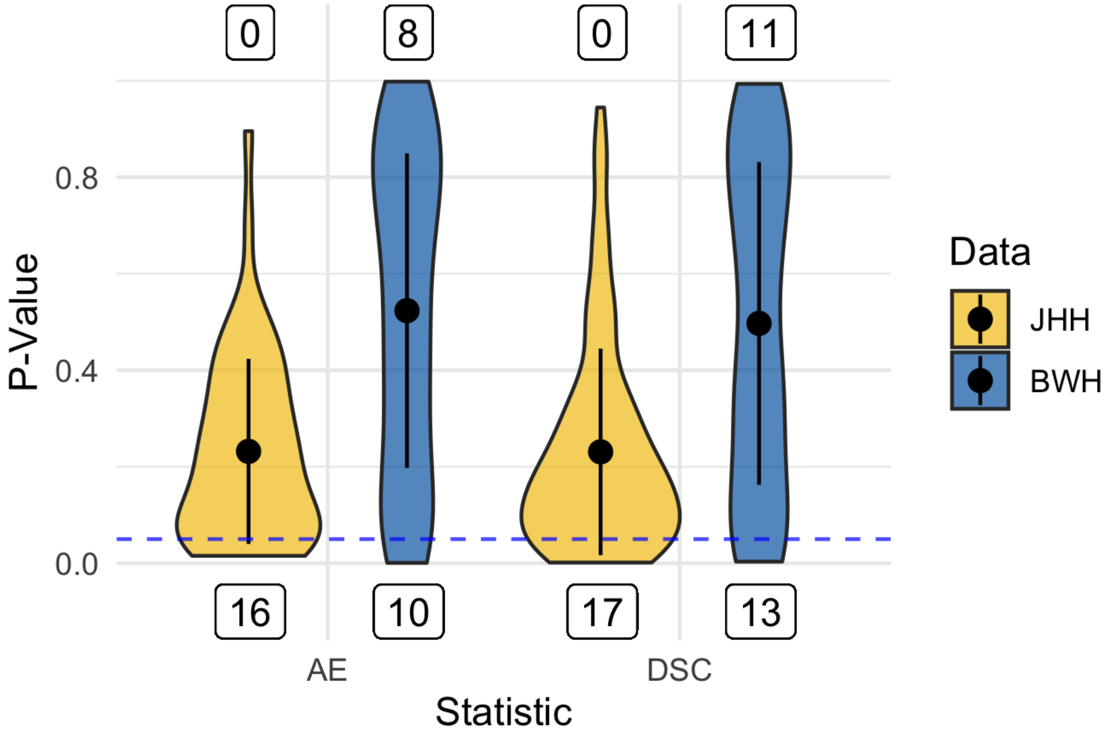


Figure A.13: Violin plots of p-values from paired t-tests to compare subject-level absolute error (AE) and Sorensen-Dice coefficient (DSC) in each test set are presented. The mean for each statistic and data set is presented as points within each violin plot and the black lines extend the mean by the standard deviation. Labels below represent the number of significant p-values favoring TAPAS performance measures. Labels above represent the number of significant p-values favoring group thresholding performance. The dashed horizontal blue line highlights the $\alpha = 0.05$ cutoff.

The results presented in Figure A.13 are re-created from Chapter 3 Figure 3.5. We employed one-sided paired t-tests to evaluate AE and DSC from TAPAS compared with those obtained from the group thresholding procedure. Figure A.13 shows violin plots of p-values from both sets of tests for the two data sets. The labels beneath each violin show the number of p-values less than $\alpha = 0.05$ that favor the TAPAS measure (i.e. a reduction in AE and an increase in DSC). The labels above each violin show the number of p-values less than $\alpha = 0.05$ that favor the group measure. In the JHH data p-values for AE and DSC favor the TAPAS thresholding procedure compared with the group thresholding procedure. The BWH data is more uniform.

Table A.3: Subject-specific volume estimates, $volume_{Manual}$ (Manual), $volume_{TAPAS}$ (TAPAS), $volume_{Group}$ (Group), and $volume_{Partial}$ (Partial), were compared with clinical covariates available in each data set and are represented in this table. Spearman's correlation coefficient ($\hat{\rho}$) was obtained in the testing set for each iteration and averaged across folds. Clinical variables included Expanded Disability Status Scale (EDSS) score, disease duration in years, and timed 25-ft walk (T25FW) in seconds.

	Estimates for $\hat{\rho}$			
	Partial	**Group**	**TAPAS**	**Manual**
JHH				
EDSS	0.35	0.38	0.38	0.29
Disease duration	0.43	0.43	0.43	0.39
BWH				
EDSS	0.42	0.43	0.43	0.45
Disease duration	0.31	0.32	0.32	0.29
T25FW	0.02	0.02	0.02	0.03

The results presented in Table A.3 are re-created from Chapter 3 Table 3.2. The correlations presented are similar to those presented in Chapter 3. The JHH data correlation estimates are slightly increased using LST-LPA here compared to MIMoSA in Chapter 3.

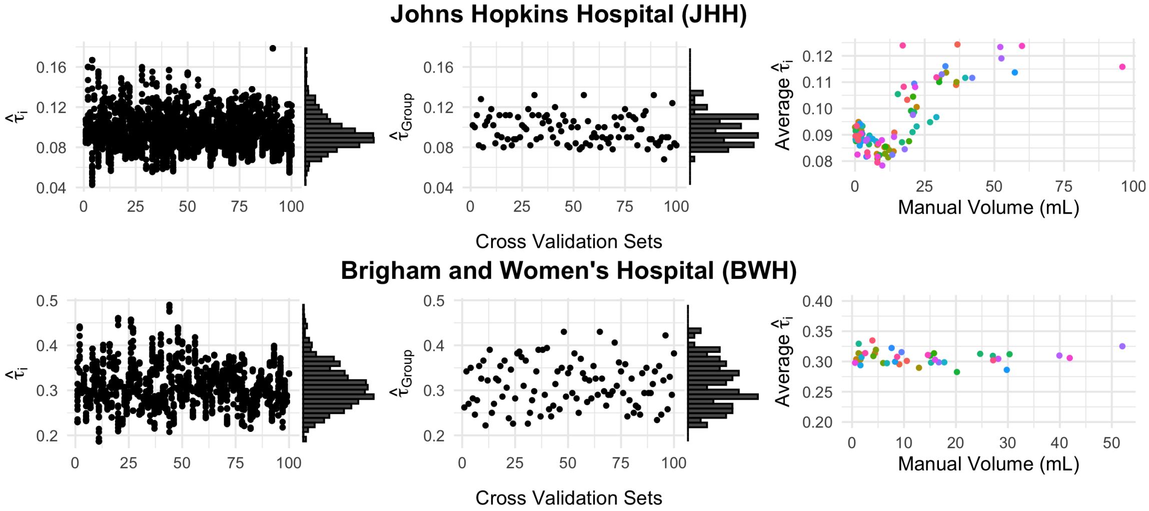


Figure A.14: Scatter plots of the subject-specific threshold $\hat{\tau}_i$ (TAPAS) and $\hat{\tau}_{Group}$ (group thresholding procedure) on cross-validation number are presented with marginal histograms for both data sets in the first two columns. The third column presents scatterplots of the average subject-specific thresholds from TAPAS and the manually delineated lesion volume.

The results presented in Figure A.14 are re-created from Chapter 3 Figure 3.6. The BWH data scatter and histogram plots are similar to those in Chapter 3 for both thresholding procedures. The JHH plots presented here are quite different than those in Chapter 3 for both the group and TAPAS thresholding procedures. The group thresholding procedure thresholds are centered around 0.10 here whereas in Chapter 3 they centered around 0.40. The TAPAS subject-specific threshold distribution is very different than Chapter 3 findings. The distribution here is centered around 0.10 whereas Chapter 3 is centered around 0.40. The spread here is also much more narrow than Chapter 3 findings.

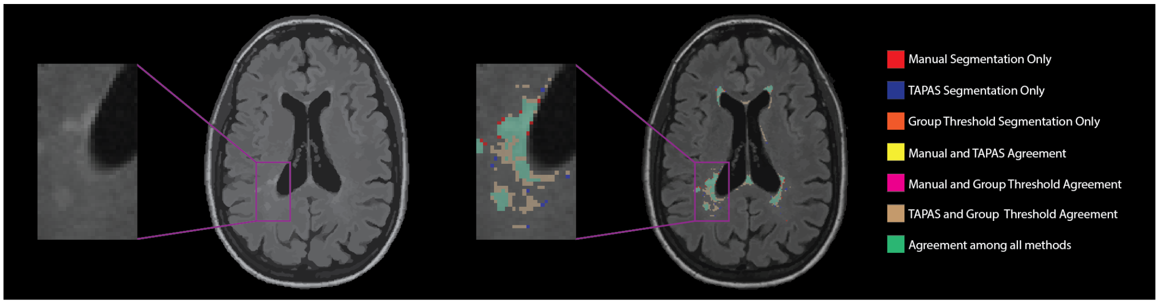

🟥 Manual Segmentation Only
🟦 TAPAS Segmentation Only
🟧 Group Threshold Segmentation Only
🟨 Manual and TAPAS Agreement
🟪 Manual and Group Threshold Agreement
🟫 TAPAS and Group Threshold Agreement
🟩 Agreement among all methods

Figure A.15: T2 hyperintense lesion segmentations from an example axial slice are displayed. The colors represent the different individual or overlapping segmentations obtained from manual, TAPAS threshold, and group threshold masks. The majority of segmented area was in agreement among all lesion masks (green). Both the group thresholding approach and TAPAS missed some area that was manually segmented (red). There was a small amount of area where only TAPAS and manual segmentations agreed (yellow), but almost no area where only the group threshold agreed with the manual segmentation (fuchsia).

The results presented in Figure A.15 are re-created from Chapter 3 Figure 3.7.

APPENDIX B

SOFTWARE

B.1. mimosa

Software to implement the MIMoSA method is available through the mimosa R package. The package takes NIfTI images as inputs in order to train the MIMoSA model and then predicts lesion segmentations. Additionally, we have made some pre-trained models available in the event gold standard manual segmentations are not available for training. We suggest when possible to train the model. The development version of the package is available on GitHub https://github.com/avalcarcel9/mimosa. A stable version is available on Neuroconductor https://neuroconductor.org/package/mimosa. Package documentation is available on GitHub https://avalcarcel9.github.io/mimosa/.

B.2. rtapas

Software to implement the TAPAS method is available through the rtapas R package. This package creates data structures necessary for training the TAPAS model from NIfTI inputs. After training, the model can be used to predict subject-specific thresholds to use on probability maps for automatic lesion segmentation. The development version of the package is available on GitHub https://github.com/avalcarcel9/rtapas. A stable version is available on Neuroconductor https://neuroconductor.org/package/rtapas. Package documentation is available on GitHub https://avalcarcel9.github.io/rtapas/.

B.3. NiftiArray

NiftiArray is an R package that allows for convenient and memory-efficient containers for on-disk representation of NIfTI objects. The package also allows for all operations supported by DelayedArray. Operations on NiftiArray objects can be either delayed or block-processed. The development version of the package is available on GitHub https://github.com/muschellij2/NiftiArray. A stable version is available on Neuroconductor https://neuroconductor.org/package/NiftiArray. Package documentation is available on GitHub https://neuroconductor.org/help/NiftiArray/.

B.4. Inter-Modal Coupling

There are currently two privately distributed R packages to implement inter-modal coupling analysis. Please contact Alessandra Valcarcel if you would like access to the private package. After in progress manuscripts are published the packages will be publicly available on GitHub and Neuroconductor. Check `https://github.com/avalcarcel9` for updates.

BIBLIOGRAPHY

Abela, E, Rummel, C, Hauf, M, Weisstanner, C, Schindler, K, and Wiest, R (2014). Neuroimaging of Epilepsy: Lesions, Networks, Oscillations. en. *Clin Neuroradiol* 24.1, 5–15. ISSN: 1869-1447. DOI: 10.1007/s00062-014-0284-8. URL: https://doi.org/10.1007/s00062-014-0284-8 (visited on 04/14/2020).

Ahlgren, C, Odn, A, and Lycke, J (2011). High nationwide prevalence of multiple sclerosis in Sweden. en. *Mult Scler* 17.8, 901–908. ISSN: 1352-4585. DOI: 10.1177/1352458511403794. URL: https://doi.org/10.1177/1352458511403794 (visited on 12/19/2018).

Andermatt, S, Papadopoulou, A, Radue, E-W, Sprenger, T, and Cattin, P (2017). Tracking the Evolution of Cerebral Gadolinium-Enhancing Lesions to Persistent T1 Black Holes in Multiple Sclerosis: Validation of a Semiautomated Pipeline. eng. *J Neuroimaging* 27.5, 469–475. ISSN: 1552-6569. DOI: 10.1111/jon.12439.

Anderson, A, Douglas, PK, Kerr, WT, Haynes, VS, Yuille, AL, Xie, J, Wu, YN, Brown, JA, and Cohen, MS (2014). Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. en. *NeuroImage*. Multimodal Data Fusion 102, 207–219. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2013.12.015. URL: http://www.sciencedirect.com/science/article/pii/S1053811913012196 (visited on 04/14/2020).

Bakshi, R, Minagar, A, Jaisani, Z, and Wolinsky, JS (2005). Imaging of multiple sclerosis: Role in neurotherapeutics. en. *Neurotherapeutics* 2.2, 277–303. ISSN: 1545-5343. DOI: 10.1602/neurorx.2.2.277. URL: https://link.springer.com/article/10.1602/neurorx.2.2.277 (visited on 12/19/2018).

Bakshi, R, Thompson, AJ, Rocca, MA, Pelletier, D, Dousset, V, Barkhof, F, Inglese, M, Guttmann, CRG, Horsfield, MA, and Filippi, M (2008). MRI in multiple sclerosis: current status and future prospects. eng. *Lancet Neurol* 7.7, 615–625. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(08)70137-6.

Barkhof, F (1999). MRI in multiple sclerosis: correlation with expanded disability status scale (EDSS). en. *Mult Scler* 5.4, 283–286. ISSN: 1352-4585. DOI: 10.1177/135245859900500415. URL: https://doi.org/10.1177/135245859900500415 (visited on 08/23/2018).

Barkhof, F, Polman, CH, Radue, E-W, Kappos, L, Freedman, MS, Edan, G, Hartung, H-P, Miller, DH, Montalbn, X, Poppe, P, Vos, Md, Lasri, F, Bauer, L, Dahms, S, Wagner, K, Pohl, C, and Sandbrink, R (2007). Magnetic Resonance Imaging Effects of Interferon Beta-1b in the BENEFIT Study: Integrated 2-Year Results. en. *Arch Neurol* 64.9, 1292–1298. ISSN: 0003-9942. DOI: 10.1001/archneur.64.9.1292. URL: https://jamanetwork.com/journals/jamaneurology/fullarticle/794481 (visited on 11/07/2019).

Bengtsson, H (2019). *matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)*. URL: https://CRAN.R-project.org/package=matrixStats.

Biessmann, F, Plis, S, Meinecke, FC, Eichele, T, and Mller, K-R (2011). Analysis of multimodal neuroimaging data. eng. *IEEE Rev Biomed Eng* 4, 26–58. ISSN: 1941-1189. DOI: 10.1109/RBME.2011.2170675.

Bing, X, Ming-guo, Q, Ye, Z, Jing-na, Z, Min, L, Han, C, Yu, Z, Jia-jia, Z, Jian, W, Wei, C, Han-jian, D, and Shao-xiang, Z (2013). Alterations in the cortical thickness and the amplitude of low-frequency fluctuation in patients with post-traumatic stress disorder. en. *Brain Research* 1490, 225–232. ISSN: 0006-8993. DOI: 10.1016/j.brainres.2012.10.048. URL: http://www.sciencedirect.com/science/article/pii/S000689931201726X (visited on 12/19/2019).

Bland, JM and Altman, DG (2007). Agreement Between Methods of Measurement with Multiple Observations Per Individual. *Journal of Biopharmaceutical Statistics* 17.4, 571–582. ISSN: 1054-3406. DOI: 10.1080/10543400701329422. URL: https://doi.org/10.1080/10543400701329422 (visited on 12/18/2018).

Bland, JM and Altman, DG (2016). Measuring agreement in method comparison studies: en. *Statistical Methods in Medical Research*. DOI: 10.1177/096228029900800204. URL: https://journals.sagepub.com/doi/10.1177/096228029900800204 (visited on 12/18/2018).

Calabresi, PA, Radue, E-W, Goodin, D, Jeffery, D, Rammohan, KW, Reder, AT, Vollmer, T, Agius, MA, Kappos, L, Stites, T, Li, B, Cappiello, L, Rosenstiel, Pv, and Lublin, FD (2014). Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. English. *The Lancet Neurology* 13.6, 545–556. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(14)70049-3. URL: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(14)70049-3/abstract (visited on 12/19/2018).

Calhoun, VD, Liu, J, and Adali, T (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. eng. *Neuroimage* 45.1 Suppl, S163–172. ISSN: 1095-9572. DOI: 10.1016/j.neuroimage.2008.10.057.

Carass, A, Cuzzocreo, J, Wheeler, MB, Bazin, P-L, Resnick, SM, and Prince, JL (2011). Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis. *NeuroImage* 56.4, 1982–1992. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.03.045. URL: http://www.sciencedirect.com/science/article/pii/S1053811911003235 (visited on 11/13/2018).

Carass, A, Roy, S, Jog, A, Cuzzocreo, JL, Magrath, E, Gherman, A, Button, J, Nguyen, J, Bazin, P-L, Calabresi, PA, Crainiceanu, CM, Ellingsen, LM, Reich, DS, Prince, JL, and Pham, DL (2017a). Longitudinal multiple sclerosis lesion segmentation data resource. *Data in Brief* 12, 346–350. ISSN: 2352-3409. DOI: 10.1016/j.dib.2017.04.004. URL: http://www.sciencedirect.com/science/article/pii/S2352340917301361 (visited on 12/20/2018).

Carass, A, Roy, S, Jog, A, Cuzzocreo, JL, Magrath, E, Gherman, A, Button, J, Nguyen, J, Prados, F, Sudre, CH, Jorge Cardoso, M, Cawley, N, Ciccarelli, O, Wheeler-Kingshott, CAM, Ourselin, S, Catanese, L, Deshpande, H, Maurel, P, Commowick, O, Barillot, C, Tomas-Fernandez, X, Warfield, SK, Vaidya, S, Chunduru, A, Muthuganapathy, R, Krishnamurthi, G, Jesson, A, Arbel, T, Maier, O, Handels, H, Iheme, LO, Unay, D, Jain, S, Sima, DM, Smeets, D, Ghafoorian, M, Platel, B, Birenbaum, A, Greenspan, H, Bazin, P-L, Calabresi, PA, Crainiceanu, CM, Ellingsen, LM, Reich, DS, Prince, JL, and Pham, DL (2017b). Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage* 148, 77–102. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2016.12.064. URL: http://www.sciencedirect.com/science/article/pii/S1053811916307819 (visited on 01/31/2019).

Card, SK, Robertson, GG, and Mackinlay, JD (1991). The information visualizer, an information workspace. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

CHI '91, 181–186. DOI: 10.1145/108844.108874. URL: https://doi.org/10.1145/108844.108874 (visited on 04/14/2020).

Casanova, R, Srikanth, R, Baer, A, Laurienti, PJ, Burdette, JH, Hayasaka, S, Flowers, L, Wood, F, and Maldjian, JA (2007). Biological parametric mapping: A statistical toolbox for multimodality brain image analysis. eng. *Neuroimage* 34.1, 137–143. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2006.09.011.

Ceccarelli, A, Jackson, JS, Tauhid, S, Arora, A, Gorky, J, Dell'Oglio, E, Bakshi, A, Chitnis, T, Khoury, SJ, Weiner, HL, Guttmann, CRG, Bakshi, R, and Neema, M (2012). The Impact of Lesion In-Painting and Registration Methods on Voxel-Based Morphometry in Detecting Regional Cerebral Gray Matter Atrophy in Multiple Sclerosis. en. *American Journal of Neuroradiology* 33.8. Publisher: American Journal of Neuroradiology Section: Brain, 1579–1585. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A3083. URL: http://www.ajnr.org/content/33/8/1579 (visited on 03/30/2020).

Cherubini, A, Luccichenti, G, Pran, P, Hagberg, GE, Barba, C, Formisano, R, and Sabatini, U (2007). Multimodal fMRI tractography in normal subjects and in clinically recovered traumatic brain injury patients. en. *NeuroImage* 34.4, 1331–1341. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2006.11.024. URL: http://www.sciencedirect.com/science/article/pii/S1053811906011025 (visited on 04/14/2020).

Clayden, J, Cox, B, and Jenkinson, M (2020). *RNifti: Fast R and C++ Access to NIfTI Images*. URL: https://CRAN.R-project.org/package=RNifti.

Commowick, O, Istace, A, Kain, M, Laurent, B, Leray, F, Simon, M, Pop, SC, Girard, P, Amli, R, Ferr, J-C, Kerbrat, A, Tourdias, T, Cervenansky, F, Glatard, T, Beaumont, J, Doyle, S, Forbes, F, Knight, J, Khademi, A, Mahbod, A, Wang, C, McKinley, R, Wagner, F, Muschelli, J, Sweeney, E, Roura, E, Llad, X, Santos, MM, Santos, WP, Silva-Filho, AG, Tomas-Fernandez, X, Urien, H, Bloch, I, Valverde, S, Cabezas, M, Vera-Olmos, FJ, Malpica, N, Guttmann, C, Vukusic, S, Edan, G, Dojat, M, Styner, M, Warfield, SK, Cotton, F, and Barillot, C (2018). Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. en. *Sci Rep* 8.1, 1–17. ISSN: 2045-2322. DOI: 10.1038/s41598-018-31911-7. URL: https://www.nature.com/articles/s41598-018-31911-7 (visited on 11/15/2019).

Compston, A and Coles, A (2002). Multiple sclerosis. English. *The Lancet* 359.9313, 1221–1231. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(02)08220-X. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(02)08220-X/abstract (visited on 12/19/2018).

Confavreux, C and Vukusic, S (2008). The Clinical Epidemiology of Multiple Sclerosis. *Neuroimaging Clinics of North America*. Multiple Sclerosis, Part I: Background and Conventional MRI 18.4, 589–622. ISSN: 1052-5149. DOI: 10.1016/j.nic.2008.09.002. URL: http://www.sciencedirect.com/science/article/pii/S1052514908000877 (visited on 12/19/2018).

Copen, WA (2015). Multimodal Imaging in Acute Ischemic Stroke. en. *Curr Treat Options Cardio Med* 17.3, 10. ISSN: 1534-3189. DOI: 10.1007/s11936-015-0368-z. URL: https://doi.org/10.1007/s11936-015-0368-z (visited on 04/14/2020).

Correa, NM, Li, Y-O, Adal, T, and Calhoun, VD (2008). Canonical Correlation Analysis for Feature-Based Fusion of Biomedical Imaging Modalities and Its Application to Detection of Associative

Networks in Schizophrenia. eng. *IEEE J Sel Top Signal Process* 2.6, 998–1007. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2008.2008265.

Dadar, M, Maranzano, J, Misquitta, K, Anor, CJ, Fonov, VS, Tartaglia, MC, Carmichael, OT, Decarli, C, Collins, DL, and Alzheimer's Disease Neuroimaging Initiative (2017). Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. eng. *Neuroimage* 157, 233–249. ISSN: 1095-9572. DOI: 10.1016/j.neuroimage.2017.06.009.

Danelakis, A, Theoharis, T, and Verganelakis, DA (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. en. *Computerized Medical Imaging and Graphics* 70, 83–100. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2018.10.002. URL: http://www.sciencedirect.com/science/article/pii/S0895611118303227 (visited on 11/04/2019).

Datta, S, Sajja, BR, He, R, Wolinsky, JS, Gupta, RK, and Narayana, PA (2006). Segmentation And Quantification Of Black Holes In Multiple Sclerosis. *Neuroimage* 29.2, 467–474. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2005.07.042. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1808226/ (visited on 12/12/2017).

Diggle, P, Diggle, DoMaSPJ, Diggle, PJ, Heagerty, P, Liang, K-Y, Heagerty, PJ, Zeger, S, and Zeger, BaBDS (2002). *Analysis of Longitudinal Data*. en. Google-Books-ID: kKLbyWycRwcC. OUP Oxford. ISBN: 978-0-19-852484-7.

Doshi, J, Erus, G, Ou, Y, Gaonkar, B, and Davatzikos, C (2013). Multi-Atlas Skull-Stripping. *Academic Radiology* 20.12, 1566–1576. ISSN: 1076-6332. DOI: 10.1016/j.acra.2013.09.010. URL: http://www.sciencedirect.com/science/article/pii/S1076633213004182 (visited on 11/13/2018).

Dupuy, SL, Tauhid, S, Kim, G, Chu, R, Tummala, S, Hurwitz, S, and Bakshi, R (2015). MRI detection of hypointense brain lesions in patients with multiple sclerosis: T1 spin-echo vs. gradient-echo. en. *European Journal of Radiology* 84.8, 1564–1568. ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2015.05.004. URL: http://www.sciencedirect.com/science/article/pii/S0720048X15002326 (visited on 03/31/2020).

Durst, CR, Raghavan, P, Shaffrey, ME, Schiff, D, Lopes, MB, Sheehan, JP, Tustison, NJ, Patrie, JT, Xin, W, Elias, WJ, Liu, KC, Helm, GA, Cupino, A, and Wintermark, M (2014). Multimodal MR imaging model to predict tumor infiltration in patients with gliomas. en. *Neuroradiology* 56.2, 107–115. ISSN: 1432-1920. DOI: 10.1007/s00234-013-1308-9. URL: https://doi.org/10.1007/s00234-013-1308-9 (visited on 04/14/2020).

Dworkin, JD, Linn, KA, Oguz, I, Fleishman, GM, Bakshi, R, Nair, G, Calabresi, PA, Henry, RG, Oh, J, Papinutto, N, Pelletier, D, Rooney, W, Stern, W, Sicotte, NL, Reich, DS, Shinohara, RT, and Cooperative, tNAIiMS (2018). An Automated Statistical Technique for Counting Distinct Multiple Sclerosis Lesions. en. *American Journal of Neuroradiology*. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A5556. URL: http://www.ajnr.org/content/early/2018/02/22/ajnr.A5556 (visited on 12/19/2018).

Egger, C, Opfer, R, Wang, C, Kepp, T, Sormani, MP, Spies, L, Barnett, M, and Schippling, S (2017). MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? *NeuroImage: Clinical* 13, 264–270. ISSN: 2213-1582. DOI: 10.1016/

j.nicl.2016.11.020. URL: http://www.sciencedirect.com/science/article/pii/S2213158216302285 (visited on 02/04/2019).

Fadda, G, Brown, RA, Magliozzi, R, AubertBroche, B, O'Mahony, J, Shinohara, RT, Banwell, B, Marrie, RA, Yeh, EA, Collins, DL, Arnold, DL, and BarOr, A (2019). A surface-in gradient of thalamic damage evolves in pediatric multiple sclerosis. en. *Annals of Neurology* 85.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.25429, 340–351. ISSN: 1531-8249. DOI: 10.1002/ana.25429. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25429 (visited on 03/16/2020).

Filippi, M, Rovaris, M, Campi, A, Pereira, C, and Comi, G (1996). Semi-automated thresholding technique for measuring lesion volumes in multiple sclerosis: effects of the change of the threshold on the computed lesion loads. en. *Acta Neurologica Scandinavica* 93.1, 30–34. ISSN: 1600-0404. DOI: 10.1111/j.1600-0404.1996.tb00166.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0404.1996.tb00166.x/abstract (visited on 12/21/2017).

Fisher, E, Lee, J-C, Nakamura, K, and Rudick, RA (2008). Gray matter atrophy in multiple sclerosis: A longitudinal study. en. *Annals of Neurology* 64.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.21436, 255–265. ISSN: 1531-8249. DOI: 10.1002/ana.21436. URL: https://onlinelibrary.wiley. com/doi/abs/10.1002/ana.21436 (visited on 03/16/2020).

Fisniku, LK, Chard, DT, Jackson, JS, Anderson, VM, Altmann, DR, Miszkiel, KA, Thompson, AJ, and Miller, DH (2008). Gray matter atrophy is related to long-term disability in multiple sclerosis. en. *Annals of Neurology* 64.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.21423, 247–254. ISSN: 1531-8249. DOI: 10.1002/ana.21423. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.21423 (visited on 03/16/2020).

Garcia-Lorenzo, D, Francis, S, Narayanan, S, Arnold, DL, and Collins, DL (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. en. *Medical Image Analysis* 17.1, 1–18. ISSN: 1361-8415. DOI: 10.1016/j.media.2012.09.004. URL: http://www.sciencedirect.com/science/article/pii/S1361841512001338 (visited on 03/30/2020).

Ge, Y (2006). Multiple Sclerosis: The Role of MR Imaging. en. *American Journal of Neuroradiology* 27.6, 1165–1176. ISSN: 0195-6108, 1936-959X. URL: http://www.ajnr.org/content/27/6/1165 (visited on 12/19/2018).

Gentleman, RC, Carey, VJ, Bates, DM, Bolstad, B, Dettling, M, Dudoit, S, Ellis, B, Gautier, L, Ge, Y, Gentry, J, Hornik, K, Hothorn, T, Huber, W, Iacus, S, Irizarry, R, Leisch, F, Li, C, Maechler, M, Rossini, AJ, Sawitzki, G, Smith, C, Smyth, G, Tierney, L, Yang, JY, and Zhang, J (2004). Bioconductor: open software development for computational biology and bioinformatics. en. *Genome Biol* 5.10, R80. ISSN: 1474-760X. DOI: 10.1186/gb-2004-5-10-r80. URL: https://doi.org/10.1186/gb-2004-5-10-r80 (visited on 04/11/2020).

Harbo, HF, Gold, R, and Tintor, M (2013). Sex and gender issues in multiple sclerosis: en. *Therapeutic Advances in Neurological Disorders*. Publisher: SAGE PublicationsSage UK: London, England. DOI: 10.1177/1756285613488434. URL: https://journals.sagepub.com/doi/10.1177/1756285613488434 (visited on 03/30/2020).

Harmouche, R, Subbanna, NK, Collins, DL, Arnold, DL, and Arbel, T (2015). Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood information. eng. *IEEE Trans Biomed Eng* 62.5, 1281–1292. ISSN: 1558-2531. DOI: 10.1109/TBME.2014.2385635.

Hickey, P (2020). *DelayedMatrixStats: Functions that Apply to Rows and Columns of 'DelayedMatrix' Objects*. URL: https://bioconductor.org/packages/release/bioc/html/DelayedMatrixStats.html.

Hoptman, MJ, Zuo, X-N, Butler, PD, Javitt, DC, D'Angelo, D, Mauro, CJ, and Milham, MP (2010). Amplitude of low-frequency oscillations in schizophrenia: A resting state fMRI study. en. *Schizophre-nia Research* 117.1, 13–20. ISSN: 0920-9964. DOI: 10.1016/j.schres.2009.09.030. URL: http://www.sciencedirect.com/science/article/pii/S0920996409004745 (visited on 12/19/2019).

Jim (2014). URL: http://www.xinapse.com/Manual (visited on 11/30/2017).

John Muschelli. *extrantsr: Extra Functions to Build on the ANTsR Package*.

Kaczkurkin, AN, Sotiras, A, Baller, EB, Barzilay, R, Calkins, ME, Chand, GB, Cui, Z, Erus, G, Fan, Y, Gur, RE, Gur, RC, Moore, TM, Roalf, DR, Rosen, AFG, Ruparel, K, Shinohara, RT, Varol, E, Wolf, DH, Davatzikos, C, and Satterthwaite, TD (2019). Neurostructural Heterogeneity in Youths With Internalizing Symptoms. en. *Biological Psychiatry*. ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2019.09.005. URL: http://www.sciencedirect.com/science/article/pii/S0006322319317032 (visited on 12/16/2019).

Kang, J, Caffo, B, and Liu, H (2016). Editorial: Recent Advances and Challenges on Big Data Analysis in Neuroimaging. English. *Front. Neurosci.* 10. Publisher: Frontiers. ISSN: 1662-453X. DOI: 10.3389/fnins.2016.00505. URL: https://www.frontiersin.org/articles/10.3389/fnins.2016.00505/full (visited on 04/14/2020).

Katdare, A and Ursekar, M (2015). Systematic imaging review: Multiple Sclerosis. en. *Annals of Indian Academy of Neurology* 18.5. Company: Medknow Publications and Media Pvt. Ltd. Distributor: Medknow Publications and Media Pvt. Ltd. Institution: Medknow Publications and Media Pvt. Ltd. Label: Medknow Publications and Media Pvt. Ltd. Publisher: Medknow Publications, 24. ISSN: 0972-2327. DOI: 10.4103/0972-2327.164821. URL: http://www.annalsofian.org/article.asp?issn=0972-2327;year=2015;volume=18;issue=5;spage=24;epage=29;aulast=Katdare;type=0 (visited on 03/30/2020).

Keshavan, A, Paul, F, Beyer, MK, Zhu, AH, Papinutto, N, Shinohara, RT, Stern, W, Amann, M, Bakshi, R, Bischof, A, Carriero, A, Comabella, M, Crane, JC, D'Alfonso, S, Demaerel, P, Dubois, B, Filippi, M, Fleischer, V, Fontaine, B, Gaetano, L, Goris, A, Graetz, C, Grger, A, Groppa, S, Hafler, DA, Harbo, HF, Hemmer, B, Jordan, K, Kappos, L, Kirkish, G, Llufriu, S, Magon, S, Martinelli-Boneschi, F, McCauley, JL, Montalban, X, Mhlau, M, Pelletier, D, Pattany, PM, Pericak-Vance, M, Cournu-Rebeix, I, Rocca, MA, Rovira, A, Schlaeger, R, Saiz, A, Sprenger, T, Stecco, A, Uitdehaag, BMJ, Villoslada, P, Wattjes, MP, Weiner, H, Wuerfel, J, Zimmer, C, Zipp, F, Hauser, SL, Oksenberg, JR, and Henry, RG (2016). Power estimation for non-standardized multisite studies. en. *NeuroImage* 134, 281–294. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2016.03.051. URL: http://www.sciencedirect.com/science/article/pii/S1053811916002561 (visited on 03/16/2020).

Khayati, R, Vafadust, M, Towhidkhah, F, and Nabavi, SM (2008). A novel method for automatic determination of different stages of multiple sclerosis lesions in brain MR FLAIR images. eng.

Comput Med Imaging Graph 32.2, 124–133. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2007.10.003.

Kim, G, Tauhid, S, Dupuy, SL, Tummala, S, Khalid, F, Healy, BC, and Bakshi, R (2016). An MRI-defined measure of cerebral lesion severity to assess therapeutic effects in multiple sclerosis. eng. *J. Neurol.* 263.3, 531–538. ISSN: 1432-1459. DOI: 10.1007/s00415-015-8009-8.

Kochunov, P, Chiappelli, J, Wright, SN, Rowland, LM, Patel, B, Wijtenburg, SA, Nugent, K, McMahon, RP, Carpenter, WT, Muellerklein, F, Sampath, H, and Elliot Hong, L (2014). Multimodal white matter imaging to investigate reduced fractional anisotropy and its age-related decline in schizophrenia. en. *Psychiatry Research: Neuroimaging* 223.2, 148–156. ISSN: 0925-4927. DOI: 10.1016/j.pscychresns.2014.05.004. URL: http://www.sciencedirect.com/science/article/pii/S0925492714001206 (visited on 04/14/2020).

Laird, NM and Ware, JH (1982). Random-Effects Models for Longitudinal Data. *Biometrics* 38.4. Publisher: [Wiley, International Biometric Society], 963–974. ISSN: 0006-341X. DOI: 10.2307/2529876. URL: https://www.jstor.org/stable/2529876 (visited on 04/07/2020).

Liang, K-Y and Zeger, SL (1986). Longitudinal data analysis using generalized linear models. en. *Biometrika* 73.1. Publisher: Oxford Academic, 13–22. ISSN: 0006-3444. DOI: 10.1093/biomet/73.1.13. URL: https://academic.oup.com/biomet/article/73/1/13/246001 (visited on 04/07/2020).

Liu, J, Ren, L, Womer, FY, Wang, J, Fan, G, Jiang, W, Blumberg, HP, Tang, Y, Xu, K, and Wang, F (2014). Alterations in amplitude of low frequency fluctuation in treatment-nave major depressive disorder measured with resting-state fMRI. en. *Human Brain Mapping* 35.10, 4979–4988. ISSN: 1097-0193. DOI: 10.1002/hbm.22526. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.22526 (visited on 12/19/2019).

Liu, S, Cai, W, Liu, S, Zhang, F, Fulham, M, Feng, D, Pujol, S, and Kikinis, R (2015a). Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders. en. *Brain Inf.* 2.3, 167–180. ISSN: 2198-4026. DOI: 10.1007/s40708-015-0019-x. URL: https://braininformatics.springeropen.com/articles/10.1007/s40708-015-0019-x (visited on 10/27/2019).

Liu, S, Cai, W, Liu, S, Zhang, F, Fulham, M, Feng, D, Pujol, S, and Kikinis, R (2015b). Multimodal neuroimaging computing: the workflows, methods, and platforms. eng. *Brain Inform* 2.3, 181–195. ISSN: 2198-4018. DOI: 10.1007/s40708-015-0020-4.

Liu, X, Chen, J, Shen, B, Wang, G, Li, J, Hou, H, Chen, X, Guo, Z, and Mao, C (2018). *Altered Intrinsic Coupling between Functional Connectivity Density and Amplitude of Low-Frequency Fluctuation in Mild Cognitive Impairment with Depressive Symptoms*. en. Research article. DOI: 10.1155/2018/1672708. URL: https://www.hindawi.com/journals/np/2018/1672708/abs/ (visited on 12/19/2019).

Llado, X, Oliver, A, Cabezas, M, Freixenet, J, Vilanova, JC, Quiles, A, Valls, L, Rami-Torrent, L, and Rovira, (2012). Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. *Information Sciences* 186.1, 164–185. ISSN: 0020-0255. DOI: 10.1016/j.ins.2011.10.011. URL: http://www.sciencedirect.com/science/article/pii/S0020025511005548 (visited on 12/19/2018).

Lublin, FD, Reingold, SC, Cohen, JA, Cutter, GR, Srensen, PS, Thompson, AJ, Wolinsky, JS, Balcer, LJ, Banwell, B, Barkhof, F, Bebo, B, Calabresi, PA, Clanet, M, Comi, G, Fox, RJ, Freedman, MS, Goodman, AD, Inglese, M, Kappos, L, Kieseier, BC, Lincoln, JA, Lubetzki, C, Miller, AE, Montalban, X, O'Connor, PW, Petkau, J, Pozzilli, C, Rudick, RA, Sormani, MP, Stve, O, Waubant, E, and Polman, CH (2014). Defining the clinical course of multiple sclerosis. *Neurology* 83.3, 278. DOI: 10.1212/WNL.0000000000000560. URL: http://n.neurology.org/content/83/3/278.abstract.

Lucas, BC, Bogovic, JA, Carass, A, Bazin, P-L, Prince, JL, Pham, DL, and Landman, BA (2010). The Java Image Science Toolkit (JIST) for Rapid Prototyping and Publishing of Neuroimaging Software. en. *Neuroinform* 8.1, 5–17. ISSN: 1559-0089. DOI: 10.1007/s12021-009-9061-2. URL: https://doi.org/10.1007/s12021-009-9061-2 (visited on 11/13/2018).

McAuliffe, MJ, Lalonde, FM, McGarry, D, Gandler, W, Csaky, K, and Trus, BL (2001). "Medical Image Processing, Analysis and Visualization in clinical research". In: *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, 381–386. DOI: 10.1109/CBMS.2001.941749.

Meier, DS, Guttmann, CRG, Tummala, S, Moscufo, N, Cavallari, M, Tauhid, S, Bakshi, R, and Weiner, HL (2018). Dual-Sensitivity Multiple Sclerosis Lesion and CSF Segmentation for Multichannel 3T Brain MRI. en. *Journal of Neuroimaging* 28.1, 36–47. ISSN: 1552-6569. DOI: 10.1111/jon.12491. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jon.12491 (visited on 01/31/2019).

Miller, RB. Response time in man-computer conversational transactions. *Proceedings of the December 9-11, 1968, fall joint computer conference, part I.* URL: https://dl.acm.org/doi/10.1145/1476589.1476628 (visited on 04/11/2020).

Molyneux, PD, Brex, PA, Fogg, C, Lewis, S, Middleditch, C, Barkhof, F, Sormani, MP, Filippi, M, and Miller, DH (2000). The precision of T1 hypointense lesion volume quantification in multiple sclerosis treatment trials: a multicenter study. eng. *Mult. Scler.* 6.4, 237–240. ISSN: 1352-4585. DOI: 10.1177/135245850000600405.

Muschelli, J (2020). *neurobase: 'Neuroconductor' Base Package with Helper Functions for 'nifti' Objects*. URL: https://CRAN.R-project.org/package=neurobase.

Muschelli, J and Shinohara, RT (2018). *White Matter Normalization for Magnetic Resonance Images using WhiteStripe*. URL: https://neuroconductor.org/package/WhiteStripe.

Muschelli, J and Valcarcel, A (2020). *NiftiArray: HDF5 Delayed Array for Nifti Objects*. URL: https://github.com/muschellij2/NiftiArray.

Muschelli, J, Gherman, A, Fortin, J-P, Avants, B, Whitcher, B, Clayden, JD, Caffo, BS, and Crainiceanu, CM. Neuroconductor: an R platform for medical imaging analysis. en. *Biostatistics*. DOI: 10.1093/biostatistics/kxx068. URL: https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxx068/4791943 (visited on 11/19/2018).

Neema, M, Arora, A, Healy, BC, Guss, ZD, Brass, SD, Duan, Y, Buckle, GJ, Glanz, BI, Stazzone, L, Khoury, SJ, Weiner, HL, Guttmann, CRG, and Bakshi, R (2009). Deep Gray Matter Involvement on Brain MRI Scans Is Associated with Clinical Progression in Multiple Sclerosis. en. *Journal of Neuroimaging* 19.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1552-

6569.2008.00296.x, 3–8. ISSN: 1552-6569. DOI: 10.1111/j.1552-6569.2008.00296.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1552-6569.2008.00296.x (visited on 03/16/2020).

NITRC: CBICA: Multi Atlas Skull Stripping (MASS): Tool/Resource Info. URL: https://www.nitrc.org/projects/cbica_mass/ (visited on 10/11/2017).

Oh, J, Ontaneda, D, Azevedo, C, Klawiter, EC, Absinta, M, Arnold, DL, Bakshi, R, Calabresi, PA, Crainiceanu, C, Dewey, B, Freeman, L, Gauthier, S, Henry, R, Inglese, M, Kolind, S, Li, DK, Mainero, C, Menon, RS, Nair, G, Narayanan, S, Nelson, F, Pelletier, D, Rauscher, A, Rooney, W, Sati, P, Schwartz, D, Shinohara, RT, Tagge, I, Traboulsee, A, Wang, Y, Yoo, Y, Yousry, T, Zhang, Y, Sicotte, NL, and Reich, DS (2019). Imaging outcome measures of neuroprotection and repair in MS. *Neurology* 92.11, 519. DOI: 10.1212/WNL.0000000000007099. URL: http://n.neurology.org/content/92/11/519.abstract.

Pages, H (2020). *HDF5Array: HDF5 backend for DelayedArray objects*. URL: https://bioconductor.org/packages/release/bioc/html/HDF5Array.html.

Pages, H, Hickey, P, and Lun, A (2020). *DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets*. URL: https://bioconductor.org/packages/release/bioc/html/DelayedArray.html.

Pepe, MS and Anderson, GL (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* 23.4. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610919408813210, 939–951. ISSN: 0361-0918. DOI: 10.1080/03610919408813210. URL: https://doi.org/10.1080/03610919408813210 (visited on 04/13/2020).

Petersen, RC, Aisen, PS, Beckett, LA, Donohue, MC, Gamst, AC, Harvey, DJ, Jack, CR, Jagust, WJ, Shaw, LM, Toga, AW, Trojanowski, JQ, and Weiner, MW (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology* 74.3, 201–209. ISSN: 0028-3878. DOI: 10.1212/WNL.0b013e3181cb3e25. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2809036/ (visited on 04/14/2020).

Popescu, V, Agosta, F, Hulst, HE, Sluimer, IC, Knol, DL, Sormani, MP, Enzinger, C, Ropele, S, Alonso, J, Sastre-Garriga, J, Rovira, A, Montalban, X, Bodini, B, Ciccarelli, O, Khaleeli, Z, Chard, DT, Matthews, L, Palace, J, Giorgio, A, Stefano, ND, Eisele, P, Gass, A, Polman, CH, Uitdehaag, BMJ, Messina, MJ, Comi, G, Filippi, M, Barkhof, F, Vrenken, H, and Group, obotMS (2013). Brain atrophy and lesion load predict long term disability in multiple sclerosis. en. *J Neurol Neurosurg Psychiatry* 84.10, 1082–1091. ISSN: 0022-3050, 1468-330X. DOI: 10.1136/jnnp-2012-304094. URL: https://jnnp.bmj.com/content/84/10/1082 (visited on 12/19/2018).

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. URL: https://www.R-project.org/.

R. Taki Shinohara and John Muschelli (2017). *WhiteStripe: White Matter Normalization for Magnetic Resonance Images using Whitestripe*.

Radua, J, Grau, M, Heuvel, OA van den, Schotten, M Thiebaut de, Stein, DJ, Canales-Rodrguez, EJ, Catani, M, and Mataix-Cols, D (2014). Multimodal Voxel-Based Meta-Analysis of White Matter Abnormalities in ObsessiveCompulsive Disorder. en. *Neuropsychopharmacology* 39.7.

Number: 7 Publisher: Nature Publishing Group, 1547–1557. ISSN: 1740-634X. DOI: 10.1038/npp.2014.5. URL: https://www.nature.com/articles/npp20145 (visited on 04/14/2020).

Rencher, A and G. Bruce, S. *Linear Models in Statistics*. en-us. Second. John Wiley & Sons. URL: https://www.wiley.com/en-us/Linear+Models+in+Statistics%2C+2nd+Edition-p-9780471754985 (visited on 04/11/2020).

Rovira, and Len, A (2008). MR in the diagnosis and monitoring of multiple sclerosis: An overview. English. *European Journal of Radiology* 67.3, 409–414. ISSN: 0720-048X, 1872-7727. DOI: 10.1016/j.ejrad.2008.02.044. URL: https://www.ejradiology.com/article/S0720-048X(08)00169-1/abstract (visited on 12/19/2018).

Roy, S, He, Q, Sweeney, E, Carass, A, Reich, DS, Prince, JL, and Pham, DL (2015). Subject Specific Sparse Dictionary Learning for Atlas Based Brain MRI Segmentation. *IEEE J Biomed Health Inform* 19.5, 1598–1609. ISSN: 2168-2194. DOI: 10.1109/JBHI.2015.2439242. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4562064/ (visited on 12/19/2018).

Rutland, JW, Delman, BN, Gill, CM, Zhu, C, Shrivastava, RK, and Balchandani, P (2020). Emerging Use of Ultra-High-Field 7T MRI in the Study of Intracranial Vascularity: State of the Field and Future Directions. en. *American Journal of Neuroradiology* 41.1. Publisher: American Journal of Neuroradiology Section: Adult Brain, 2–9. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A6344. URL: http://www.ajnr.org/content/41/1/2 (visited on 04/14/2020).

Sanfilipo, MP, Benedict, RH, Weinstock-Guttman, B, and Bakshi, R (2006). Gray and white matter brain atrophy and neuropsychological impairment in multiple sclerosis. *Neurology* 66.5, 685. DOI: 10.1212/01.wnl.0000201238.93586.d9. URL: http://n.neurology.org/content/66/5/685.abstract.

Satterthwaite, TD, Shinohara, RT, Wolf, DH, Hopson, RD, Elliott, MA, Vandekar, SN, Ruparel, K, Calkins, ME, Roalf, DR, Gennatas, ED, Jackson, C, Erus, G, Prabhakaran, K, Davatzikos, C, Detre, JA, Hakonarson, H, Gur, RC, and Gur, RE (2014a). Impact of puberty on the evolution of cerebral perfusion during adolescence. en. *PNAS* 111.23, 8643–8648. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1400178111. URL: https://www.pnas.org/content/111/23/8643 (visited on 12/02/2019).

Satterthwaite, TD, Elliott, MA, Ruparel, K, Loughead, J, Prabhakaran, K, Calkins, ME, Hopson, R, Jackson, C, Keefe, J, Riley, M, Mensh, FD, Sleiman, P, Verma, R, Davatzikos, C, Hakonarson, H, Gur, RC, and Gur, RE (2014b). Neuroimaging of the Philadelphia Neurodevelopmental Cohort. *Neuroimage* 86, 544–553. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2013.07.064. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3947233/ (visited on 10/27/2019).

Satterthwaite, TD, Connolly, JJ, Ruparel, K, Calkins, ME, Jackson, C, Elliott, MA, Roalf, DR, Hopson, R, Prabhakaran, K, Behr, M, Qiu, H, Mentch, FD, Chiavacci, R, Sleiman, PMA, Gur, RC, Hakonarson, H, and Gur, RE (2016). The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. eng. *Neuroimage* 124.Pt B, 1115–1119. ISSN: 1095-9572. DOI: 10.1016/j.neuroimage.2015.03.056.

Schmidt, P, Gaser, C, Arsic, M, Buck, D, Frschler, A, Berthele, A, Hoshi, M, Ilg, R, Schmid, VJ, Zimmer, C, Hemmer, B, and Mhlau, M (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage* 59.4, 3774–3783. ISSN:

1053-8119. DOI: 10.1016/j.neuroimage.2011.11.032. URL: http://www.sciencedirect.com/science/article/pii/S1053811911013139 (visited on 06/11/2019).

Shiee, N, Bazin, P-L, Ozturk, A, Reich, DS, Calabresi, PA, and Pham, DL (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. en. *NeuroImage* 49.2, 1524–1535. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2009.09.005. URL: http://www.sciencedirect.com/science/article/pii/S1053811909009823 (visited on 03/30/2020).

Shinohara, RT, Crainiceanu, CM, Caffo, BS, Gaitn, MI, and Reich, DS (2011). Population-wide principal component-based quantification of bloodbrain-barrier dynamics in multiple sclerosis. *NeuroImage* 57.4, 1430–1446. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.05.038. URL: http://www.sciencedirect.com/science/article/pii/S1053811911005465 (visited on 11/30/2018).

Shinohara, RT, Sweeney, EM, Goldsmith, J, Shiee, N, Mateen, FJ, Calabresi, PA, Jarso, S, Pham, DL, Reich, DS, and Crainiceanu, CM (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* 6, 9–19. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2014.08.008. URL: http://www.sciencedirect.com/science/article/pii/S221315821400117X (visited on 11/13/2018).

Sing, T, Sander, O, Beerenwinkel, N, and Lengauer, T (2005). ROCR: visualizing classifier performance in R. en. *Bioinformatics* 21.20. Publisher: Oxford Academic, 3940–3941. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti623. URL: https://academic.oup.com/bioinformatics/article/21/20/3940/202693 (visited on 03/30/2020).

Sled, JG, Zijdenbos, AP, and Evans, AC (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. eng. *IEEE Trans Med Imaging* 17.1, 87–97. ISSN: 0278-0062. DOI: 10.1109/42.668698.

Spies, L, Tewes, A, Suppa, P, Opfer, R, Buchert, R, Winkler, G, and Raji, A (2013). Fully automatic detection of deep white matter T1 hypointense lesions in multiple sclerosis. eng. *Phys Med Biol* 58.23, 8323–8337. ISSN: 1361-6560. DOI: 10.1088/0031-9155/58/23/8323.

Stankiewicz, JM, Glanz, BI, Healy, BC, Arora, A, Neema, M, Benedict, RHB, Guss, ZD, Tauhid, S, Buckle, GJ, Houtchens, MK, Khoury, SJ, Weiner, HL, Guttmann, CRG, and Bakshi, R (2011). Brain MRI Lesion Load at 1.5T and 3T versus Clinical Status in Multiple Sclerosis. en. *Journal of Neuroimaging* 21.2, e50–e56. ISSN: 1552-6569. DOI: 10.1111/j.1552-6569.2009.00449.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1552-6569.2009.00449.x (visited on 12/19/2018).

Stigler, KA, McDonald, BC, Anand, A, Saykin, AJ, and McDougle, CJ (2011). Structural and functional magnetic resonance imaging of autism spectrum disorders. en. *Brain Research*. The Emerging Neuroscience of Autism Spectrum Disorders 1380, 146–161. ISSN: 0006-8993. DOI: 10.1016/j.brainres.2010.11.076. URL: http://www.sciencedirect.com/science/article/pii/S0006899310025898 (visited on 04/14/2020).

Sui, J, Pearlson, G, Caprihan, A, Adali, T, Kiehl, KA, Liu, J, Yamamoto, J, and Calhoun, VD (2011). Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. en. *NeuroImage*. Special Issue: Educational Neuroscience 57.3, 839–855. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.05.055. URL: http://www.sciencedirect.com/science/article/pii/S1053811911005635 (visited on 04/14/2020).

Sweeney, EM, Shinohara, RT, Shiee, N, Mateen, FJ, Chudgar, AA, Cuzzocreo, JL, Calabresi, PA, Pham, DL, Reich, DS, and Crainiceanu, CM (2013). OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI. *NeuroImage: Clinical* 2, 402–413. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2013.03.002. URL: http://www.sciencedirect.com/science/article/pii/S2213158213000235 (visited on 11/12/2018).

Sweeney, EM, Vogelstein, JT, Cuzzocreo, JL, Calabresi, PA, Reich, DS, Crainiceanu, CM, and Shinohara, RT (2014). A Comparison of Supervised Machine Learning Algorithms and Feature Vectors for MS Lesion Segmentation Using Multimodal Structural MRI. en. *PLOS ONE* 9.4, e95753. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0095753. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0095753 (visited on 12/19/2018).

Tauhid, S, Neema, M, Healy, BC, Weiner, HL, and Bakshi, R (2014). MRI phenotypes based on cerebral lesions and atrophy in patients with multiple sclerosis. *Journal of the Neurological Sciences* 346.1, 250–254. ISSN: 0022-510X. DOI: 10.1016/j.jns.2014.08.047. URL: http://www.sciencedirect.com/science/article/pii/S0022510X14005796 (visited on 12/19/2018).

Tauhid, S, Chu, R, Sasane, R, Glanz, BI, Neema, M, Miller, JR, Kim, G, Signorovitch, JE, Healy, BC, Chitnis, T, Weiner, HL, and Bakshi, R (2015). Brain MRI lesions and atrophy are associated with employment status in patients with multiple sclerosis. en. *J Neurol* 262.11, 2425–2432. ISSN: 1432-1459. DOI: 10.1007/s00415-015-7853-x. URL: https://doi.org/10.1007/s00415-015-7853-x (visited on 12/19/2018).

Thompson, AJ, Banwell, BL, Barkhof, F, Carroll, WM, Coetzee, T, Comi, G, Correale, J, Fazekas, F, Filippi, M, Freedman, MS, Fujihara, K, Galetta, SL, Hartung, HP, Kappos, L, Lublin, FD, Marrie, RA, Miller, AE, Miller, DH, Montalban, X, Mowry, EM, Sorensen, PS, Tintor, M, Traboulsee, AL, Trojano, M, Uitdehaag, BMJ, Vukusic, S, Waubant, E, Weinshenker, BG, Reingold, SC, and Cohen, JA (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. English. *The Lancet Neurology* 17.2, 162–173. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(17)30470-2. URL: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(17)30470-2/abstract (visited on 07/26/2019).

Tustison, NJ, Avants, BB, Cook, PA, Zheng, Y, Egan, A, Yushkevich, PA, and Gee, JC (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging* 29.6, 1310–1320. ISSN: 0278-0062. DOI: 10.1109/TMI.2010.2046908.

Valcarcel, A (2018). *mimosa: 'MIMoSA': A Method for Inter-Modal Segmentation Analysis*. URL: https://github.com/avalcarcel9/mimosa.

Valcarcel, AM, Linn, KA, Khalid, F, Vandekar, SN, Tauhid, S, Satterthwaite, TD, Muschelli, J, Martin, ML, Bakshi, R, and Shinohara, RT (2018a). A dual modeling approach to automatic segmentation of cerebral T2 hyperintensities and T1 black holes in multiple sclerosis. *NeuroImage: Clinical* 20, 1211–1221. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2018.10.013. URL: http://www.sciencedirect.com/science/article/pii/S2213158218303231 (visited on 11/12/2018).

Valcarcel, AM, Linn, KA, Vandekar, SN, Satterthwaite, TD, Muschelli, J, Calabresi, PA, Pham, DL, Martin, ML, and Shinohara, RT (2018b). MIMoSA: An Automated Method for Intermodal Segmentation Analysis of Multiple Sclerosis Brain Lesions. en. *Journal of Neuroimaging* 28.4, 389–398. ISSN: 1552-6569. DOI: 10.1111/jon.12506. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jon.12506 (visited on 11/13/2018).

Vandekar, SN, Shinohara, RT, Raznahan, A, Hopson, RD, Roalf, DR, Ruparel, K, Gur, RC, Gur, RE, and Satterthwaite, TD (2016). Subject-level measurement of local cortical coupling. en. *NeuroImage* 133, 88–97. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2016.03.002. URL: http://www.sciencedirect.com/science/article/pii/S105381191600197X (visited on 12/02/2019).

Wack, DS, Dwyer, MG, Bergsland, N, Di Perri, C, Ranza, L, Hussein, S, Ramasamy, D, Poloni, G, and Zivadinov, R (2012). Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates. *BMC Medical Imaging* 12.1, 17. ISSN: 1471-2342. DOI: 10.1186/1471-2342-12-17. URL: https://doi.org/10.1186/1471-2342-12-17 (visited on 08/17/2018).

Walter, SD (2005). The partial area under the summary ROC curve. en. *Statistics in Medicine* 24.13. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2103, 2025–2040. ISSN: 1097-0258. DOI: 10.1002/sim.2103. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2103 (visited on 03/30/2020).

Wang, P, Yang, J, Yin, Z, Duan, J, Zhang, R, Sun, J, Xu, Y, Liu, L, Chen, X, Li, H, Kang, J, Zhu, Y, Deng, X, Chang, M, Wei, S, Zhou, Y, Jiang, X, Wang, F, and Tang, Y (2019). Amplitude of low-frequency fluctuation (ALFF) may be associated with cognitive impairment in schizophrenia: a correlation study. *BMC Psychiatry* 19.1, 30. ISSN: 1471-244X. DOI: 10.1186/s12888-018-1992-4. URL: https://doi.org/10.1186/s12888-018-1992-4 (visited on 12/19/2019).

Wood, SN. *Generalized Additive Models: An Introduction with R*. en. 2nd ed. Chapman and Hall/CRC. URL: https://www.crcpress.com/Generalized-Additive-Models-An-Introduction-with-R/Wood/p/book/9780429093159 (visited on 12/12/2018).

Wood, SN (2003). Thin Plate Regression Splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 65.1, 95–114. ISSN: 1369-7412. URL: https://www.jstor.org/stable/3088828 (visited on 12/12/2018).

Wood, SN (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* 99.467, 673–686. ISSN: 0162-1459. DOI: 10.1198/016214504000000980. URL: https://doi.org/10.1198/016214504000000980 (visited on 12/12/2018).

Wood, SN, Pya, N, and Sfken, B (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association* 111.516, 1548–1563. ISSN: 0162-1459. DOI: 10.1080/01621459.2016.1180986. URL: https://doi.org/10.1080/01621459.2016.1180986 (visited on 12/12/2018).

Wu, W-C, Fernndez-Seara, M, Detre, JA, Wehrli, FW, and Wang, J (2007). A theoretical and experimental investigation of the tagging efficiency of pseudocontinuous arterial spin labeling. eng. *Magn Reson Med* 58.5, 1020–1027. ISSN: 0740-3194. DOI: 10.1002/mrm.21403.

Wu, Y, Warfield, SK, Tan, IL, Wells, WM, Meier, DS, Schijndel, RA van, Barkhof, F, and Guttmann, CRG (2006). Automated segmentation of multiple sclerosis lesion subtypes with multichannel

MRI. eng. *Neuroimage* 32.3, 1205–1215. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2006.04.211.

Yang, X, Beason-Held, L, Resnick, SM, and Landman, BA (2011). Biological Parametric Mapping WITH Robust AND Non-Parametric Statistics. *Neuroimage* 57.2, 423–430. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.04.046. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114289/ (visited on 10/27/2019).

Zang, Y-F, He, Y, Zhu, C-Z, Cao, Q-J, Sui, M-Q, Liang, M, Tian, L-X, Jiang, T-Z, and Wang, Y-F (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. eng. *Brain Dev.* 29.2, 83–91. ISSN: 0387-7604. DOI: 10.1016/j.braindev.2006.07.002.

Zhang, D, Wang, Y, Zhou, L, Yuan, H, and Shen, D (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. en. *NeuroImage* 55.3, 856–867. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.01.008. URL: http://www.sciencedirect.com/science/article/pii/S1053811911000267 (visited on 10/27/2019).

Zhou, J, Yao, N, Fairchild, G, Zhang, Y, and Wang, X (2015). Altered Hemodynamic Activity in Conduct Disorder: A Resting-State fMRI Investigation. en. *PLOS ONE* 10.3, e0122750. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0122750. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122750 (visited on 12/19/2019).

Zijdenbos, A, Dawant, B, Margolin, R, and Palmer, A (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging* 13.4, 716–724. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/42.363096.

Zivadinov, R and Bakshi, R (2004). Role of MRI in multiple sclerosis I: inflammation and lesions. eng. *Front. Biosci.* 9, 665–683. ISSN: 1093-9946.