A neuroinformatics framework for the collection, curation, and visualization of imaging biomarkers in multiple sclerosis

by

Anisha Keshavan

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

[redacted]

Copyright 2017

by

Anisha Keshavan

Dedication and Acknowledgements

I dedicate this work to my parents, Nandita Raja and Raja Keshavan, and my partner Brandon O'Hern. I thank my advisor Roland Henry, and my friends Kesshi Jordan, Bagrat Amirbekian, and Salem Kimble. Thanks to Arno Klein, Satra Ghosh, and Riley Bove for their mentorship.

The text of chapter 3 in this dissertation is a reprint of the material as it appears in Keshavan, Anisha, et al. "Power estimation for non-standardized multisite studies." *NeuroImage* 134 (2016): 281-294. The senior author listed in this publication, Dr. Roland Henry, directed and supervised the research.

The text of chapter 5 in this dissertation is a reprint of the material as it appears in Keshavan, Anisha, et al. "Mindcontrol: A Web Application for Brain Segmentation Quality Control" *NeuroImage* (2017).

A neuroinformatics framework for the collection, curation, and visualization of imaging biomarkers in multiple sclerosis

Anisha Keshavan

Imaging biomarkers from magnetic resonance images have provided insights into the progression of multiple sclerosis (MS). As neuroimaging datasets grow in size to accommodate multidimensional association studies, traditional methods for data collection and analysis are too imprecise and inefficient on a large scale. This dissertation addresses the challenges associated with collecting datasets from multiple scanners with non-standardized acquisition protocols, presents software for time and space efficient image processing, and software for collaborative quality control. Finally, a visualization framework is proposed to gain better intuition and understanding of high-dimensional imaging datasets through a web-based interactive data exploration tool. This dissertation lays the groundwork for large, multivariate studies in MS, and translational tools for the use of imaging biomarkers in the MS clinic.

Contents

List of Tables

List of Figures

1 Introduction

Multiple sclerosis (MS) is a central nervous system disease that affects 2.5 million people worldwide, and has no cure. **Precision medicine** is a new disease management approach that accounts for individual variability in genetics, environment, and clinical phenotypes, to tailor more effective, personalized, treatments [1]. A precision medicine approach to treat MS requires a large and deeply phenotyped reference dataset of MS patients, in order to quantify the differences in biology that make each patient unique. Finally, effective methods of presenting large amounts of data are needed, in order to empower clinicians to tailor treatments based on an individual's specific biological makeup.

Researchers at UCSF have been collecting a deeply phenotyped dataset of MS patients. Metrics include genetic markers of risk burden, clinical evaluations, and MRI images. MRI imaging plays an important role in the diagnosis of MS, and researchers have found promising MRI biomarkers that may track with disease progression. Some examples include the location and size of lesions, cortical atrophy patterns, spinal cord gray matter volume, and various quantitative white matter metrics. While these imaging phenotypes have the potential to inform a precise treatment plan for an individual with MS, there exist three major challenges: **data collection** from heterogeneous sources, **data curation** and **reproducible processing**, and effective **data visualization** of large datasets.

In order to collect enough data to provide a reference dataset for precision medicine, we must be able to combine MRI biomarkers from different hospitals and research institutions. Biomarkers derived from a variety of MRI scanners are subject to biases from differences in scanner hardware and software. Accurate measurement of these biases, along with a statistical model that accounts for them, can help researchers plan an efficient data collection scheme. In "Power estimation for non-standardized multisite studies" [2], I proposed a **statistical model** to estimate power for the collection of a large MRI dataset across various imaging sites. This framework enables rapid data collection by easing the traditional standardization requirements of data acquisition.

Collecting a large MRI dataset brings about challenges that are traditionally associated with “big data”. These include curation, quality checking, and the implementation of robust, reproducible analyses. Proper data curation provides easy accessibility by researchers not directly involved in data collection, minimization of processing errors, optimized use of analysis software, and aids in the development of automated tools[3]. Ensuring the quality of processing pipeline inputs and outputs is a vital aspect of biomarker development, and often requires manual intervention, especially with pathological MRI scans. The curation of the automated pipelines and their associated manual edits is implemented in PBRAINand Mindcontrol, respectively. **PBRAIN** is an open-source framework for the curation of neuroimaging analysis pipelines, organization of input and output data, and the storage data provenance. **Mindcontrol** is a web-based application that consists of a dashboard to organize, quality control, annotate, edit, and collaborate on neuroimaging processing results [4, 5].

MS Imaging biomarkers, with promising predictive value, are not readily available to MS neurologists due to many obstacles in the translation of biomarkers to the clinic. One such obstacle is the difficulty in comprehending large amounts of data in a short time span. One way to address this is to reduce 3D/4D data into scalar metrics that describe the image, such as brain volume, or lesion volume. However, this loss of spatial information deprives the clinician of the crucial context necessary to determine the nature and progression of the disease for any given individual. One solution is to develop effective data visualizations. The information extracted from 4D imaging data can be abstracted to highlight the important features of our data, in the same way a subway map highlights connections and line routes, maintains spatial information, and hides unnecessary information like the terrain. The final project of my dissertation is the MINDMELD project, which combines 3D brain visualization with clinical and imaging scalar metrics in an interactive, web-based application.

Figure 1.1 shows the components of my research, which consists of four projects relating to the collection, curation, and visualization of neuroimaging biomarkers in MS. My thesis concludes with a proposal for future directions, which was funded by the Gordon and Betty Moore Foundation, Alfred P. Sloan Foundation, and the Washington Research Foundation. I propose a crowd-sourced quality control application to improve the precision of segmentation measures and at the same time, engage, educate and excite the public to help advance cutting edge neuroscience research. In the age of rapid data collection, biomarker discovery, and the development of new therapies, collaborative informatics methods are vital to help us understand the complex disease mechanisms of multiple sclerosis.

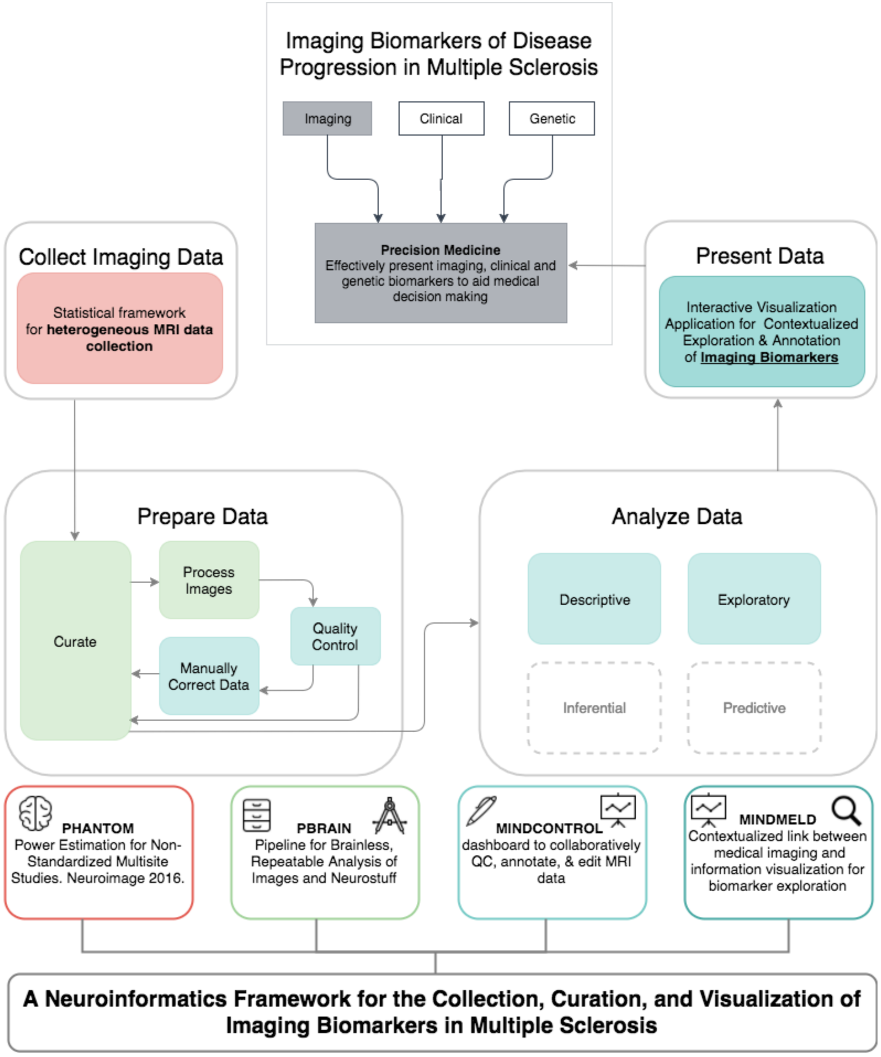

Imaging Biomarkers of Disease Progression in Multiple Sclerosis
Imaging
Clinical
Genetic
Precision Medicine
Effectively present imaging, clinical and genetic biomarkers to aid medical decision making
Collect Imaging Data
Statistical framework for heterogeneous MRI data collection
Present Data
Interactive Visualization Application for Contextualized Exploration & Annotation of Imaging Biomarkers
Prepare Data
Curate
Process Images
Quality Control
Manually Correct Data
Analyze Data
Descriptive
Exploratory
Inferential
Predictive
PHANTOM
Power Estimation for Non-Standardized Multisite Studies. Neuroimage 2016.
PBRAIN
Pipeline for Brainless, Repeatable Analysis of Images and Neurostuff
MINDCONTROL
dashboard to collaboratively QC, annotate, & edit MRI data
MINDMELD
Contextualized link between medical imaging and information visualization for biomarker exploration
A Neuroinformatics Framework for the Collection, Curation, and Visualization of Imaging Biomarkers in Multiple Sclerosis

Figure 1.1: Diagram for Anisha Keshavan's dissertation. The overall goal is to develop precision medicine tools that present imaging, clinical, and genetic biomarkers to MS neurologists. For my PhD dissertation, I have focused on the imaging aspect of this tool. It begins with the large-scale collection of imaging data, outlined by the **PHANTOM** project. Next, data must be processed and curated, by the **PBRAIN** project, and quality controlled through the **MINDCONTROL** project. Finally, the aggregation of all imaging biomarkers is presented in the **MINDMELD** project, which provides descriptive, exploratory, and interactive data analyses of imaging biomarkers in MS along with three-dimensional scientific data visualization.

2 Background

2.1 Multiple Sclerosis

Multiple Sclerosis (MS) is an immune modulated demylinating disease of the central nervous system. The cause of MS is unknown, and is likely an interplay between environmental factors (such as viral [6]) and genetic factors [7, 8, 9]. MS is a debilitating disease, where symptoms include a loss of vision (optic neuritis), sensory loss, muscle cramping and spasticity, bladder and bowel dysfunction, fatigue, and cognitive problems that include a short attention span, memory loss, and even depression [10]. The incidence has been increasing, particularly for women, and prevalence of MS has been increasing over time due to longer survival rates [11] from recent advances in therapeutics. Disease modifying therapies, such as β-interferon [12] are designed with the goal of postponing long-term disability; however, the length of a clinical trial is not sufficient to evaluate long-term outcomes. In order to evaluate the efficacy of a new treatment, clinical trials use short-term clinical scores along with disease burden based on magnetic resonance imaging (MRI) [12].

2.2 Magnetic Resonance Imaging

Magnetic resonance imaging is based on quantum properties of atoms. Hydrogen protons have a property called spin, which can be perturbed by strong magnetic fields and a sequence of magnetic gradients and radiofrequency (RF) pulses. After an RF pulse is applied, a proton releases energy in the form an electromagnetic signal, which is recorded by receiver coils. Local tissue properties affect the signal from protons on fat and water molecules, resulting in contrast between gray matter, white matter, and CSF in the brain. Different types of contrast can be generated based on the timing of the RF pulses, such as the T1-weighted image and the T2-weighted structural image. The variety of contrasts available in MR, combined with the lack of ionizing radiation during image acquisition, gives MR a clear advantage over computed tomography (CT), which was previously used in the diagnostic criteria of MS.

2.3 MR Imaging in MS

MRI revolutionized the diagnostic criteria for MS, enabling clinicians to diagnose and treat MS at earlier stages. Markers of inflammation and demyelination from multiple sclerosis manifest as focal, demyelinated scar tissue, called lesions or plaques, in the white matter of the brain and spinal cord. The first evidence of the utility of MRI over CT in detecting MS plaques was shown in 1981 [13]. In comparing CT and MR images, [13] found that 1) MS lesions were better delineated compared to CT 2) there were more lesions in MR that could not be seen on CT and 3) the pattern of lesion distribution in MRI more closely matched post-mortem images. Within a decade, MR was incorporated into the diagnostic criteria for the disease [14].

2.3.1 MS Lesions

There are three characteristic types of lesions seen in MR images based on their signal intensity in T1 and T2 contrasts : 1) Gadolinium (Gd+) enhancing lesions, 2) T2 hyperintense lesions, and 3) T1 hypointense "black holes". MS-related inflammation causes breakdown of the blood-brain barrier, allowing the Gd+ MR contrast agent to enter the lesion, appearing hyperintense on the T1-weighted image. These lesions are sensitive and specific to MS [15]. On the other hand, T2 hyperintensities are used as markers of overall disease burden but lack neurological specificity [15]. T2 lesions are used in the McDonald criteria for diagnosis of MS [16], with guidelines on the location and shape characteristics more specific to MS [15]. T1 "black holes", or hypointensities on the T1-weighted image, are indicative of permanent axonal damage and loss [15, 17, 18, 19].

2.3.2 The Clinico-Radiological Paradox

While focal white matter lesions seen on MRI largely characterize multiple sclerosis, lesion volumes are not strongly correlated with clinical disability [20, 21, 22]. This disconnect between clinical disability in multiple sclerosis (MS) and structural damage seen on MRI is called the clinico-radiological paradox [23]. Instead, many groups have found that gray matter atrophy is a better predictor of disability[24, 25, 26, 27].

2.4 Volumetric Analysis of MRI

Automated estimation of brain atrophy has been possible due to advances in computing and image processing algorithms. Tissue classification algorithms label each voxel (a volumetric unit of MR signal intensity) as belonging to different tissue classes, such as white matter, gray matter, CSF, or an MS lesion. More advanced algorithms use brain atlases to parcellate different regions of brain, providing regional volume and thickness estimates of different cortical regions, subcortical structures, and the cerebellum.

2.4.1 Brain Tissue Segmentation

There are many tissue classification algorithms and software packages available to the community (see FSL's FAST [28], FIRST [29], ANTS' Atropos [30], and Freesurfer [31]). Early segmentation algorithms could classify between gray matter, white matter, and CSF, based purely on voxel intensities. This was computed using Markov Random Field theory (MRF) [32]. MRF models the image as a graph, with voxels modeled as vertices and edges modeled as a relationship (such as the euclidean distance) between the voxel and its neighbors[33]. For example, the probability of voxel V_1 belonging to tissue class c_1 given that neighboring voxel V_2 belonging to class c_2 can be computed based on prior, manually segmented images [34, 31]. One of the most frequently used segmentation packages is FreeSurfer [31], because it delineates between different regions of gray matter. For delineating structures within the gray matter, the standard MRF does not work; MRFs are isotropic, meaning that these probability distributions are equal in all directions, and stationary, meaning that probability distributions are the same, regardless of spatial location in the brain [34, 31]. The freesurfer team encoded spatial context into the MRF, removing the stationary and isotropic constraints for the standard MRF, such that the amygdala, which is always anterior and superior to the hippocampus, could be segmented from the hippocampus even though the two structures have similar intensity distributions [34, 31].

2.4.2 Gray Matter Atrophy in MS

The automated segmentation of brain regions has led to the discovery of gray matter atrophy patterns and their relationship to MS disease progression. The gray matter pathology of MS has been widely studied in post-mortem brains (for a review, see [35]), and myelin immuno-staining has shown widespread demyelination in the cortex that is independent of white matter lesion load [36]. In vivo, decreased regional gray matter thickness in patients compared to healthy controls has been observed by many groups, both cross-sectionally [37, 38, 39], and longitudinally [40]. In addition, regional thickness moderately correlates with clinical disability [37]. Patients with later stages of the disease have more advanced cortical thinning [41], and patients in the early stages of the disease showed cortical thinning that was correlated with mild cognitive impairment [42].

The atrophy of the thalamus, in particular, has been extensively studied. Researchers found that the thalamic volume of MS patients was 16% lower than healthy controls [43], that thalamus and putamen atrophy was related with slower information processing speed [44], and that thalamic atrophy and ventricular size is associated with the transition from a clinically isolated syndrome to clinically definite MS [45]. Furthermore, MRI thalamic atrophy estimates were compared to histo-pathological neuronal loss measures, and a 30-35% loss in MS was found [46]. In addition to the thalamus, researchers found a distinct pattern of atrophy in the fronto-temporal regions of the rapidly advancing form of MS, called secondary progressive (SPMS), which also related to cognitive impairment [47, 38]. Atrophy of the anterior cingulate cortex was strongly related to both lesion volume and clinical disability, showing signs even in early stages of disease [48], while the later stages showed atrophy in the motor cortex [38]. Parcellation of the hippocampus subregions showed that atrophy in the CA1 region was a feature of MS, with substantially worse atrophy for SPMS patients [49].

2.4.3 MS Genetics and Gray Matter Atrophy

Understanding the relationship between MS genetics and MR phenotypes, such as brain atrophy, is a crucial aspect of precision medicine; it could help us predict how the disease progresses in a given individual and affect treatment decisions. An association between the HLA alleles on the major histocompatibility complex (MHC) and T2 lesion load was found in patients with primary progressive disease in a small cohort of 41 patients [50]. Genotype-phenotype correlations were found on a larger cohort (N>500) showing a link between HLA alleles, T2 lesion load, and normalized brain volume [51]. Recently, our group found that the same HLA genetic markers which are related to an increased risk for MS are also associated with subcortical gray matter atrophy in women, and an earlier age of onset (N=586) [52]. While these findings are promising, their small sample sizes limit what we can discover. Larger sample sizes would enable us to discover and replicate associations between more genetic loci and imaging phenotypes. However, these studies are limited by the inherent difficulties of collecting, processing, and ensuring the accuracy of biomarkers from large imaging datasets. Study of the field of neuroinformatics could help address these bottlenecks.

2.5 Neuroinformatics

Neuroinformatics is the study of the organization, curation, and computational analysis models and methods for neurological data. There are major roadblocks that prevent the efficient computation of biomarkers from large imaging datasets, which are critical for precision medicine approaches. Collecting a large dataset from a single institution takes a very long time; a collaborative, multisite approach is more time efficient but results in noisier data from heterogeneity in scanner hardware and acquisition protocols. Processing large amounts of data requires advanced distributed computing clusters and software that can seamlessly distribute tasks to the computing grid [53]. Big data studies must employ many researchers, but collaboration and knowledge transfer can be difficult in such a highly specialized and multidisciplinary environment, which leads to inefficiencies and increased likelihood of errors [3]. Gray matter segmentation algorithms perform adequately on healthy control data, but often error on pathological brains and require manual intervention to fix, which can take an exorbitant amount of time on very large datasets. Solutions to these neuroinformatics problems are critical to the success of large-scale genotype-phenotype studies, and for translation of imaging biomarkers to the MS clinic.

This body of work contributes novel tools to the neuroinformatics field by addressing the need to compute and validate imaging biomarkers *on a large scale*. First, a statistical model is proposed to address the problem of scanner biases for large, multisite MRI datasets. The results of this study were used to plan a multisite MS genotype/imaging phenotype study. Next, a computational framework for the efficient, reproducible, collaborative, and automated processing and curation of biomarkers is presented. Finally, a quality control tool for visual inspection and manual intervention for large datasets is proposed, followed by a visualization application to efficiently comprehend high-dimensional neuroimaging data. These tools provide a framework for running efficient, reproducible, and precise large-scale imaging studies to advance precision medicine.

3 Power Estimation for Non-Standardized Multisite Studies

Abstract

A concern for researchers planning multisite studies is that scanner and T1-weighted sequence-related biases on regional volumes could overshadow true effects, especially for studies with a heterogeneous set of scanners and sequences. Current approaches attempt to harmonize data by standardizing hardware, pulse sequences, and protocols, or by calibrating across sites using phantom-based corrections to ensure the same raw image intensities. We propose to avoid harmonization and phantom-based correction entirely. We hypothesized that the bias of estimated regional volumes is scaled between sites due to the contrast and gradient distortion differences between scanners and sequences. Given this assumption, we provide a new statistical framework and derive a power equation to define inclusion criteria for a set of sites based on the variability of their scaling factors. We estimated the scaling factors of 20 scanners with heterogeneous hardware and sequence parameters by scanning a single set of 12 subjects at sites across the United States and Europe. Regional volumes and their scaling factors were estimated for each site using Freesurfer's segmentation algorithm and ordinary least squares, respectively. The scaling factors were validated by comparing the theoretical and simulated power curves, performing a leave-one-out calibration of regional volumes, and evaluating the absolute agreement of all regional volumes between sites before and after calibration. Using our derived power equation, we were able to define the conditions under which harmonization is not necessary to achieve 80% power. This approach can inform choice of processing pipelines and

outcome metrics for multisite studies based on scaling factor variability across sites, enabling collaboration between clinical and research institutions.

3.1 Introduction

The pooled or meta-analysis of regional brain volumes derived from T1-weighted MRI data across multiple sites is reliable when data is acquired with similar acquisition parameters [54, 55, 56]. The inherent scanner- and sequence-related noise of MRI volumetrics under heterogeneous acquisition parameters has prompted many groups to standardize protocols across imaging sites [54, 57, 58]. However, standardization across multiple sites can be prohibitively expensive and requires a significant effort to implement and maintain. At the other end of the spectrum, multisite studies without standardization can also be successful, albeit with extremely large sample sizes. The ENIGMA consortium, for example, combined scans of over 10,000 subjects from 25 sites with varying field strengths, scanner makes, acquisition protocols, and processing pipelines. The unusually large sample size enabled this consortium to provide robust phenotypic traits despite the variability of non-standardized MRI volumetrics and the power required to run a genome wide association study (GWAS) to identify modest effect sizes [59]. These studies raise the following question: Is there a middle ground between fully standardizing a set of MRI scanners and recruiting thousands of subjects across a large number of sites? Eliminating the harmonization requirement for a multisite study would facilitate inclusion of retrospectively acquired data, and data from sites with ongoing longitudinal studies that would not want to adjust their acquisition parameters.

Towards this goal, there is a large body of literature addressing the correction of geometric distortions that result from gradient non-linearities. These corrections fall into two main categories: phantom-based deformation field estimation and direct magnetic field gradient measurement-based deformation estimation; the latter requires extra hardware and spherical harmonic information from the manufacturer [60]. Calibration phantoms, such as the Alzheimer's Disease Neuroimaging

Initiative (ADNI) [61] and LEGO® phantoms [62], have been used by large multisite studies to correct for these geometric distortions, which also affect regional volume measurements. These studies have outlined various correction methods that significantly improve deformation field similarity between scanners. However, the relationship between the severity of gradient distortion and its effect on regional volumes, in particular, remains unclear. In the case of heterogeneous acquisitions, correction becomes especially difficult due to additional noise sources. Gradient hardware differences across sites are compounded with contrast variation due to sequence parameter changes. In order to properly evaluate the reproducibility of brain segmentation algorithms, these phantoms should resemble the human brain in size, shape, and tissue distribution. Droby and colleagues evaluated the stability of a post-mortem brain phantom and found similar reproducibility of volumetric measurements to that of a healthy control [63]. In this study, we propose to measure between-site bias through direct calibration of regional volumes by imaging 12 common healthy controls at each site. Quantifying regional bias allows us to overcome between-site variability by increasing sample size to an optimal amount, rather than employing a phantom-based voxel-wise calibration scheme that corrects both contrast differences and geometric distortions.

We hypothesized that all differences in regional contrast and geometric distortion result in regional volumes that are consistently and linearly scaled from their true value. For a given region of interest (ROI), two mechanisms simultaneously impact the final boundary definition: (1) gradient nonlinearities cause distortion and (2) hardware (including scanner, field strength, and coils) and acquisition parameters modulate tissue contrast. Based on the results of Tardiff and colleagues, who found that contrast-to-noise ratio and contrast inhomogeneity from various pulse sequences and scanner strengths cause regional biases in VBM[64, 65], we hypothesized that each ROI will scale differently at each site. Evidence for this scaling property can also be seen in the overall increase of gray matter volume and decrease of white matter volume of the ADNI-2 compared to the ADNI-1 protocols despite attempts to maintain compatibility between these protocols [66]. It was also observed that hippocampal volumes were 1.17% larger on 3T scanners compared to the 1.5T scanners in the ADNI study [67]. By imaging 12 subjects in 20 different scanners using varying acquisition schemes, we were able to estimate the scaling factor for each regional volume at each site. We also defined a framework for calculating the power of a multisite study as a function of the scaling factor variability between sites. This enables us to power a cross-sectional study, and to outline the conditions under which harmonization could be replaced by sample size optimization. This framework can also indicate which regional volumes are sufficiently reliable to investigate using a multisite approach.

Regional brain volumes are of interest in most neurological conditions, including healthy aging, and typically indicates the degree of neuronal degeneration. In this study, we investigate a number of well-defined regional brain volumetrics related to multiple sclerosis disease progression. Even though focal white matter lesions seen on MRI largely characterize multiple sclerosis (MS), lesion volumes are not strongly correlated with clinical disability [20, 21, 22]. Instead, global gray matter atrophy correlates better with clinical disability (for a review, see [68]), along with white matter volume, to a lesser extent [69]. In addition, regional gray matter atrophy measurements, such as thalamus [70, 71, 72, 73] and caudate [74, 75] volumes, appear to be better predictors of disability [24, 25, 26, 27].

3.2 Theory

Linear mixed models are common in modeling data from multisite studies because metrics derived from scanner, protocol, and population heterogeneity may not have uncorrelated error terms when modeled in a general linear model (GLM), which violates a key assumption [76]. In fact, Fennema-Notestine and colleagues found that a mixed model, with scanner as a random effect, outperformed pooling data via GLM[77] on a study on hippocampal volumes and aging. Since we are only interested in the effect of scanner-related heterogeneity, we assume that the relationship between the volumetrics and clinical factors of interest are the same at each site. This causes error terms to cluster by scanner and sequence type due to variation in field strengths, acquisition parameters, scanner makes, head coil configurations, and field inhomogeneities, to name a few [54]. Linear mixed models, which include random effects and hierarchical effects, appropriately integrate observation-level data based on their clustering characteristics [76]. The model we propose in this study is similar to a mixed model, with a multiplicative effect instead of an additive effect. Our goal is to incorporate an MRI bias-related term in our model in order to optimize sample sizes.

We first defined the true, unobserved model for subject i at site j as:

$$U_{ij} = \beta_{00} + \beta_{10} X_{i,j} + \beta_{20} Z_{i,j} + \epsilon_{i,j} \tag{3.1}$$

Where $U_{i,j}$ is the unobserved value of the regional brain volume of interest (without any effects from the scanner), and β_{00}, β_{10} and β_{20} are the true, unobserved, effect sizes. The covariates are $Z_{i,j}$, residuals are $\epsilon_{i,j}$, and the contrast vector, $X_{i,j}$, is given the weights $X_{high}, X_{low} = 0.5, -0.5$ so that β_{10} is computed as the average difference between the high and low groups. For this derivation we assume an equal number of subjects observed at each site in the high and low groups with balanced covariates. ϵ is normally distributed with mean 0 and standard deviation σ_0.

We defined a site-level model using the notation of [78], to express the relationship between a brain metric that is scaled by a_j as $Y_{i,j} = a_j * U_{ij}$ and high or low disease group $X_{i,j}$ for subject $i = 1, \ldots, n$ at site j as

$$Y_{i,j} = b_{0j} + b_{1,j} X_{i,j} + b_{2,j} Z_{i,j} + r_{i,j} \tag{3.2}$$

The site mean, disease effect, and covariate effect randomly vary between sites so the intercept and slope coefficients become dependent variables [78] and we assume:

$$b_{k,j} = a_j * \beta_{k,0} \tag{3.3}$$

where the true underlying coefficient, $\beta_{k,0}$ for $k = 0, 1, 2$ is scaled randomly by each site. The major contributors to brain structure region of interest (ROI) boundary variability are contrast differences and gradient distortions, both of which adjust the boundary of the whole ROI rather than add a constant term. To reflect this property, we modeled the systematic error from each MRI sequence as a multiplicative ($Y_{i,j} = a_j * Y_i$) rather than additive ($Y_{ij} = Y_i + a_j$) error term. Similarly, the residual term is also scaled by site, $r_{i,j} \sim N(0, a_j^2 \sigma_0^2)$, and the scaling factor, a_j, is sampled from a normal distribution with mean μ_a and variance σ_a^2.

$$a_j \sim N(\mu_a, \sigma_a^2) \tag{3.4}$$

For identifiability, let $\mu_a = 1$. The mean disease effect estimate, $\beta_{1,j}$ is defined as the mean brain metric volume difference in the high and low groups.

$$D_{Y,j} = \overline{Y_{H_j}} - \overline{Y_{L_j}} \tag{3.5}$$

The unconditional variance of the disease effect estimate at site j is can be written in terms of the unobserved difference between groups before scaling, $D_{U,j} = D_{Y,j}/a_j$:

$$var[D_{Y,j}] = var[D_{U,j}a_j] = var[D_{U,j}]var[a_j] + var[D_{U,j}]E[a_j]^2 + var[a_j]E[D_{U,j}]^2 \tag{3.6}$$

Where we are assuming that $D_{U,j}$ and a_j are independent, meaning that MRI-related biases are independent of the biological effects being studied. For the derivation of this formula, see the Appendix. Given the distribution of scaling factors and the variance of the true disease effect, $var[D_{U,j}] = 4\sigma_0^2/n$, the equation simplifies to

$$var[D_{Y,j}] = \frac{4\sigma_0^2}{n}\mu_a^2 + \frac{4\sigma_0^2}{n}\sigma_a^2 + \sigma_a^2\beta_{10}^2 \tag{3.7}$$

We standardize the equation by defining the coefficient of variability for the scaling factors as $CV_a^2 = (\frac{\sigma_a}{\mu_a})^2$, and the standardized true effect size as $\delta = \frac{\beta_{10}}{\sigma_0}$.

$$var[D_{Y,j}] = \mu_a^2\sigma_0^2\Big(\frac{4}{n} + CV_a^2\Big(\frac{4}{n} + \delta^2\Big)\Big) \tag{3.8}$$

Finally, the coefficients are averaged over J sites to produce the overall estimate $\hat{\beta_{10}} = \frac{1}{J}\sum_{j=1}^{J} D_{Y,j}$, and

$$E[\hat{\beta_{10}}] = \frac{1}{J}\sum_{j=1}^{J} E[D_{Y,j}] = \frac{\beta_{10}}{J}\sum_{j=1}^{J} E[a_j] = \beta_{10}\mu_a \tag{3.9}$$

Note that this estimator is asymptotically normally distributed when the number of centers, J, is fixed, because it is the average of asymptotically normal estimators. When the number of subjects per site is not equal, the maximum likelihood estimator is the average of the site-level estimates weighted by the standard error, and this is shown in the Appendix. The variance of the overall estimate can be expressed as

$$var[\hat{\beta_{10}}] = \frac{1}{J^2}\sum_{j=1}^{J} var[D_{Y,j}] = \frac{\sigma_0^2\mu_a^2\Big(\frac{4}{n} + CV_a^2(\frac{4}{n} + \delta^2)\Big)}{J} \tag{3.10}$$

In order to test the average disease effect under the null hypothesis that $\beta_1 = 0$, the non-central F distribution, $F(1, J-1; \lambda)$ [78] is applied, with the non-centrality parameter defined as

$$\lambda = \frac{E[\hat{\beta_{10}}]^2}{var[\hat{\beta_{10}}]} = \frac{J\delta^2}{\frac{4}{n} + CV_a^2(\frac{4}{n} + \delta^2)} \tag{3.11}$$

Figure 3.9 shows power curves for small to medium effect sizes ($\delta = 0.2, 0.3$, defined in [78]), and a false positive rate of $\alpha = 0.002$, which allows for 25 comparisons under Bonferroni correction, where the corrected $\alpha = 0.05$. Power increases for larger λ and maximizes at $\lambda = \frac{Jn\delta^2}{4}$ as CV_a approaches 0. In this case, the power equation is dominated by the total number of subjects, as is the case for the GLM. However, even as the number of subjects per site, n, approaches infinity and for non-negligible CV_a, λ is still bounded by $\frac{J}{CV_a^2}$. At this extreme, the power equation is largely influenced by the number of sites. This highlights the importance of the site-level sample size (J) in addition to the subject-level sample size (n) for power analyses, especially when there is larger variability between sites for metrics of interest. In the methods section, the acquisition protocols and the standard processing pipelines that were used to calculate CV_a values of relevant regional brain volumes for MS are described, though this framework could be applied to any MRI derived metric.

We emphasize that the use of phantom subjects does not directly contribute to the power equation in Figure 1, as it does not account for any sort of calibration or scaling. However, it requires an estimate for CV_a, which is the variability of scaling biases between sites. The goal of this study is to provide researchers with estimates of CV_a from our set of calibration phantoms and our set of non-standardized MRI acquisitions. For a standardized set of scanners, the values of CV_a may be considered an upper bound.

3.3 Methods

3.3.1 Acquisition

T1-weighted 3D-MPRAGE images were acquired from 12 healthy subjects (3 Male, 9 Female, ages 24-57) in 20 scanners across Europe and the United States. Institutional approval was acquired and signed consent was obtained for each subject at each site. These scanners varied in make and model, including all three major manufacturers: Siemens, GE, Philips. Two scans were acquired from each subject, where the subject got out of the scanner between scans for a couple minutes, and was repositioned and rescanned by the scanning technician of that particular site. Previously, Jovicich and colleagues showed that reproducible head positioning along the z axis significantly reduced image intensity variability across sessions [56]. By repositioning in our study, a realistic measure of test-retest variability, which includes the repositioning consistency of each site's scanning procedure, was captured. Because gradient distortion effects correspond to differences in z-positioning [62], the average translation in the Z-direction between the two runs of each subject at each site was estimated with a rigid body registration.

Tables 1 through 3.4 show the acquisition parameters for all 20 scanners. Note that the definitions of repetition time (TR), inversion time (TI) and echo time (TE) vary by scanner make. For example, the TR in a Siemens scanner is the time between preparation pulses, while for Philips and GE, the TR is the time between excitation pulses. We decided to report the parameters according to the scanner make definition, rather than trying to make them uniform, because slightly different pulse programming rationales would make a fair comparison difficult. In addition, a 3D-FLASH sequence (TR=20ms, TE=4.92ms, flip angle=25 degrees, resolution=1mm isotropic) was acquired on healthy controls and MS patients at site 12, in order to compare differences in scaling factor estimates between patients and healthy controls.

3.3.2 Processing

A neuroradiologist reviewed all images to screen for major artifacts and pathology. The standard Freesurfer [79] version 5.3.0 cross-sectional pipeline (recon-all) was run on each site's native T1-weighted protocol, using the RedHat 7 operating system on IEEE 754 compliant hardware. Both 1.5T and 3T scans were run with the same parameters (without using the -3T flag), meaning that the non-uniformity correction parameters were kept at the default values. All Freesurfer results were quality controlled by evaluating the cortical gray matter segmentation and checking the linear transform to MNI305 space which is used to compute the estimated total intracranial volume [80]. Scans were excluded from the study if the cortical gray matter segmentation misclassified parts of the cortex, or if the registration to MNI305 space was grossly innaccurate. Three scans were excluded for misregistration. Two exclusions were because of data transfer errors. Because of time constraints, some subjects were not able to be scanned. One of the 12 subjects could not travel to all the sites, and that subject was replaced by another of the same age and gender. The details of this are provided in the supplemental materials and the total number of scans is shown in tables 1 - 3.4. 46 Freesurfer ROIs, including the left and right subcortical ROIs, from the aparc.stats tables, were studied. In this study we report on the ROIs relevant to the disease progression of MS, which include the gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV), and the volumes of the lateral ventricle (LV), amygdala (amyg), thalamus (thal), hippocampus (hipp), caudate (caud). The remaining ROIs are reported in the supplemental materials.

Test-retest reliability, defined as ICC(1,1) [81], was computed across each site and protocol for the selected metrics using the "psych" package in R [82]. The between-site ICC(2,1) values were computed following the procedure from previous studies on multisite reliability [81, 54]. Variance components were calculated for a fully crossed random effects model for subject, site, and run using the "lme4" package in R. Using the variance components, between site ICC was defined as

$$ICC_{BW} = \frac{\sigma^2_{subject}}{\sigma^2_{subject} + \sigma^2_{site} + \sigma^2_{run} + \sigma^2_{subject\times site} + \sigma^2_{unexplained}} \tag{3.12}$$

and an overall within-site ICC was defined as

$$ICC_{WI} = \frac{\sigma^2_{subject} + \sigma^2_{site} + \sigma^2_{subject\times site}}{\sigma^2_{subject} + \sigma^2_{site} + \sigma^2_{run} + \sigma^2_{subject\times site} + \sigma^2_{unexplained}} \tag{3.13}$$

Scaling factors between sites were estimated using ordinary least squares from the average of the scan-rescan volumes, referenced to average scan-rescan volumes from the UCSF site. The OLS was run with the intercept fixed at 0. CV_a for each metric was calculated from the sampling distribution of scaling factor estimates $\hat{a}$ as follows:

$$CV_a = \frac{std(\hat{a})}{mean(\hat{a})} \tag{3.14}$$

3.3.3 Scaling Factor Validation

Scaling factor estimates were validated under the assumption of scaled, systematic error, in 2 ways: first, by simulating power curves that take into account the uncertainty of the scaling factor estimate, and second, by a leave-one-out calibration. For the simulation, we generate data for each of the 20 sites included in this study. Subcortical gray matter volumes (scGMV) for each site were generated for two subject groups based on a small standardized effect size (Cohen's d) of 0.2, which reflects the effect sizes seen in genomics studies. Age and gender were generated as matched covariates, where age was sampled from a normal distribution with mean and standard deviation set at 41 and 10 years, respectively. Gender was sampled from a binomial distribution with a probability of 60% female to match typical multiple sclerosis cohorts.

Coefficients were set on the intercept as 63.135 cm^3, β_{10} as -.95 cm^3, covariates Z_{Age} as -.25 cm^3/year and Z_{Gender} as 4.6 cm^3. scGM volumes were generated in a linear model using these coefficients and additional noise was added from the residuals, which were sampled from a normal distribution with zero mean and standard deviation 5.03 cm^3. Next, the scGM volumes were scaled by each site's calculated scaling factor and gaussian noise from the residuals of the scaling factor fit of that particular site were added.

$$scGMV_{site_j} = scGMV_{true_j} * a_j + N(0, \sigma^2_{fit_j}) \tag{3.15}$$

The simulated dataset of each individual site was modeled via OLS, and an F score on X_{Group} was calculated following our proposed statistical model:

$$F_{X_{Group}} = \frac{(\frac{1}{J}\sum_{j=1}^{20}\hat{\beta}_j)^2}{\frac{1}{J^2}\sum_{j=1}^{20}\sigma_j^2} \tag{3.16}$$

A power curve was constructed by running the simulation 5000 times, where power for a particular p-value was defined as the average number of F values greater than the critical F for a set of false positive rates ranging from $1e^{-4} - 1e^{-2}$. The critical F was calculated with degrees of freedom of the numerator and denominator as 1 and 19 respectively. The simulated power curve was compared to the derived theoretical power curve to evaluate how scaling factor uncertainty influences power estimates. If the scaling factors of each site, which were calculated from the 12 subjects, were not accurate, then the added residual noise from the scaling factor estimate would result in the simulated power curve deviating largely from the theoretical curve.

The scaling factors were also validated by calibrating the regional volumes of each site in a leave-one-out cross-validation. The calibrated volume for a particular subject i and site j was scaled by the scaling factor estimated from all subjects excluding subject i. Within- and between-site ICC's were calculated for the calibrated volumes. If the scaling factor estimates were inaccurate, the between-site ICCs of calibrated regional volumes would be worse than the between-site ICCs of the original regional volumes. Additionally, the between-site ICC's after calibration should be similar to those found for harmonized studies, such as [54].

Finally, to address the concern about whether these scaling factors could apply to a disease population, we calculated scaling factors from 12 healthy controls and 14 MS patients between 2 different sequences (3D-MPRAGE versus 3D-FLASH) at the UCSF scanner (site 12). The patients had a mean age of 51 years with standard deviation of 11 years, mean disease duration of 15 years with a standard deviation of 12 years, and mean Kurtzke Expanded Disability Status Scale (EDSS) [83] score of 2.8 with a standard deviation of 2.2.

The accuracy of our scaling factor estimates depends on the accuracy of tissue segmentation, but the lesions in MS specifically impact white matter and pial surface segmentations. Because of the effect of lesions on Freesurfer's tissue classification, all images were manually corrected for lesions on the T1-weighted images by a neurologist after editing by Freesurfer's quality assurance procedure, which included extensive topological white matter corrections, white matter mask edits, and pial edits on images that were not lesion filled. These manual edits altered the white matter surface so that white matter lesions were not misclassified as gray matter or non-brain tissue. The errors in white matter segmentations most typically occurred at the border of white matter and gray matter and around the ventricles. The errors in pial surface segmentations most typically occurred near the eyes (orbitofrontal) and the superior frontal or medial frontal lobes. Images that were still misclassified after thorough edits were removed from the analysis, because segmentations were not accurate enough to produce realistic scaling factor estimates.

3.4 Results

Scan-rescan reliability for the 20 scanners is shown in tables 1 through 3.4. The majority of scan-rescan reliabilities were greater than 80% for the selected Freesurfer-derived volumes, which included gray matter volume (GMV), cortical white matter volume (cWMV), cortex volume (cVol), lateral ventricle (LV), thalamus (thal), amygdala (amyg), caudate (caud), hippocampus (hipp), and estimated total intracranial volume (eTIV). However, the thalamus at sites 3 and 16 had low scan-rescan reproducibility, below 70%. The left hippocampus and amygdala at site 5 were also below 70%, and the left amygdala at site 16 was also low, at 55%. In addition, the average translation in the Z-direction across all sites was $3.5mm \pm 3.7mm$, which falls within the accuracy range reported by [62]. The repositioning Z-translation measurements for each site separately is reported in the supplemental materials.

Between- and within-site ICC's are plotted with the calibrated ICC's in Figure 3.2. The between-site ICC's of the 46 ROIs improved, with the exception of the right lateral ventricle, which did not change after calibration, and the fifth ventricle, which had very low scan-rescan reliability, and is shown in the supplemental materials. The within-site ICC's of the thalamus, hippocampus, and amygdala decreased slightly after calibration. Both calibrated and uncalibrated within-site ICC's were greater than 90% for the MS related ROIs listed in this paper. For the full set of within- and between- site ICC's of the Freesurfer aseg regions, see the Supplemental Materials.

Simulation results are shown in Figure 3.3. The simulated and theoretical curves align closely when power is equal to 80%, but the simulated curve is slightly lower than the theoretical curve for power below 80%. This is probably due to the uncertainty in our scaling factor estimates.

Table 3.5 shows the scaling factor variability (CV_a) for the selected ROIs, which range from 2 to 8 %. The full distribution of CV_a for all the Freesurfer ROIs is shown in Figure 3.7. To derive the maximum acceptable CV_a for 80% power, the theoretical power equation was solved at various subject and site sample sizes with the standardized effect size we detected in our local single center cohort (0.2). The distribution of CV_a across all ROIs was plotted adjacent to the power curves (Figure 3.7) to understand how many ROIs would need to be calibrated for each case. Finally, figures 3.4, 3.5, and 3.6 show the scaling factors from the calibration between two scanners with different sequences at UCSF. Scaling factors derived from the healthy controls (HC) and MS subjects were identical for subcortical gray matter volume (1.05) and very similar for cortical gray matter volume (1, 1.002 for HC, MS) and white matter volume (.967, .975 for HC, MS).

3.5 Discussion

In this study we proposed a statistical model based on on the physics of MRI volumetric biases using the key assumption that biases between sites are scaled linearly. Variation in scaling factors could explain why a study may estimate different effect sizes based on the pulse sequence used. For example, [84] found significant effects of RF head coils, pulse sequences, and resolution on VBM results. The estimation of scaling factors in our model depends on good scan-rescan reliability. In our study, scan-rescan reliabilities for each scanner were generally > 0.8 for Freesurfer-derived regional volumes. Volumes of cortex, cortical gray, subcortical gray, and cortical white matter parcellation had greater than 90% reliability for all 20 sites. The subcortical regions and estimated total intracranial volume had an average reliability of over 89%, however, some sites had much lower scan-rescan reliability. For example, the thalamus at sites 3 and 16 had test-retest reliabilities between 41 and 63 %. This could be explained by the visual quality control process of the segmented images, which focused on the cortical gray matter segmentation and the initial standard space registration only, due to time restrictions. Visually evaluating all regional segmentations may be unrealistic for a large multisite study. On the other hand, Jovicich and colleagues [85] reported a low within-site ICC of the thalamus across sessions (0.765 ± 0.183) using the same freesurfer cross-sectional pipeline as this study. The poor between-site reliability (61%) of the thalamus is consistent with findings from [86], in which a multisite VBM analysis showed poor consistency in that region. Other segmentation algorithms may be more robust for subcortical regions in particular. Using FSL's FIRST segmentation algorithm, Cannon and colleagues [54] report a between-site ICC of the thalamus of 0.95, compared to our calibrated between-site ICC of 0.78. FSL's FIRST algorithm [87] uses a Bayesian model of shape and intensity features to produce a more precise segmentation. Nugent and colleagues reported the reliability of the FIRST algorithm across 3 platforms. Their study of subcortical ROIs found a good scan-rescan reliability of 83%, but lower between-site ICCs ranging from 57% to 93% [88]. The LEAP algorithm proposed by Wolz and colleagues [89] was shown to be extremely reliable with strong ICCs > 0.97 for hippocampal segmentations [67]. Another factor not accounted for in our segmentation results was the effect of partial voluming, which adds uncertainty to tissue volume estimates. In [90], researchers developed a method to more accurately estimate partial volume effects using only T1-weighed images from the ADNI dataset. This approach resulted in higher classification accuracy between Alzheimer's disease (AD) patients and mild cognitively impaired (MCI) patients from normal controls (NL). Designing optimized pipelines that are robust for each site, scanner make, and metric, is outside the scope of this paper. However, Kim and colleagues have developed a robust technique for tissue classification of heterogeneously acquired data that incorporates iterative bias field correction, registration, and classification [91]. Wang and colleagues developed a method to reduce systematic errors of segmentation algorithms relative to manual segmentations by training a wrapper method that learns spatial patterns of systematic errors [92]. Methods such as those employed by Wang and colleagues may be preferred over standard segmentation pipelines when data acquisition is not standardized. Due to its wide range of acquisition parameters and size of the dataset, our approach could be used to evaluate such generalized pipelines in the future.

The above derivation of power for a multisite study defines hard thresholds for the amount of acceptable scaling factor variability (CV_a) using scaled, systematic error from MRI. Many factors contribute to the CV_a cut-off, such as the total number of subjects, total number of sites, effect size, and false positive rate. In Figure 3.7, we show the distribution of experimental CV_a values across all Freesurfer aseg ROIs to reference while comparing power curves of various sample sizes. The maximum CV_a value is 9% which, with enough subjects and sites, falls well below the maximum acceptable CV_a value. However, with the minimum number of subjects and sites, the power curves of figure 3.7 show that the maximum acceptable CV_α must be below 5% for 80% power. If we minimize the total number of subjects to 2260 for the 20 sites in our study, the CV_a of the amygdala does not meet this requirement (see table 3.5). One option to address this is to harmonize protocols, which may reduce CV_a values below those estimated from our sites such that they satisfy the maximum CV_a requirement. The other option is to recruit more subjects per site. The number of additional subjects needed to overcome a large CV_a can be estimated using our power equation. In the case of the parameters defined in figure 3.7 (a small effect size of 0.2, false positive rate of 0.002), 40 additional subjects beyond the initial 2260 are needed to adequately power the study. This is easily visualized in figure 3.7: the point on the curve for the initial 2260 subjects over 20 sites lies below the harmonization zone, while that of 2300 total subjects lies above. The number of additional subjects needed to achieve an adequately powered multisite study depends on effect sizes, false positive rates, power requirements, and site-level sample size.

We have validated our scaling factors by demonstrating that a leave-one-out calibration resulted in increased absolute agreement between sites compared to the original, uncalibrated values for 44 out of 46 ROIs studied. Tables 3.6 and 3.7 compare these calibrated and original values to the ICC findings of other harmonization efforts. Table 3.6 compares our between-site ICCs before and after scaling factor calibration to those of [54]. [54] used a cortical pattern matching segmentation algorithm [93] for the cortical ROIs and FSL's FIRST algorithm for the subcortical ROIs. The between-site ICC for gray matter volume (GMV) for our study was 0.78 while [54] reported an ICC of 0.85. This difference could be explained by the harmonization of scanners in [54]. After using the scaling factors to calibrate GMV, the between-site ICC increased to 0.96, indicating that the estimated CV_a of GMV (4%) is an accurate representation of the true between-site bias variability. Scaling calibration of the hippocampus also outperformed the between-site ICC of [54] (0.84 versus 0.79), validating the CV_a estimate of 3% for both hemispheres. For the amygdala and caudate volumes, scaling calibration showed improvement to nearly the same value as [54]. The amygdala increased from 0.54 to 0.74 (versus 0.76 in the [54]), and the ICC of the caudate increased from 0.82 to 0.91 (versus 0.92 in the [54]). The CV_a of the left and right amygala were the highest in our study, at 7 and 9 percent, respectively. The most extreme asymmetry in the scaling factors was between the left and right caudate (2% and 7%, respectively), which demonstrates regional contrast to noise variation. Even after scaling factor calibration, the between-site ICC produced by our approach varied widely from that of [54] in two ROIs. The between-site ICC of white matter volume (WMV) was very high (0.96 versus 0.774) and that of thalamus volume was very low (.61 versus .95), compared to [54]. This could be due to differences algorithm differences (FIRST vs. Freesurfer). It should also be noted that the scan-rescan reliability of the thalamus was particularly low in some sites, which propagated errors to scaling factor estimates. Therefore, the 5% CV_a estimate for the thalamus in both hemispheres may not be reproducible and would need to be recalculated using a different algorithm.

Table 3.7 shows comparisons of our within-site ICCs to the average within-site ICCs reported by [85]. Similar to our study, scanners were not strictly standardized and the freesurfer cross-sectional algorithm was run. All within site ICCs (both before and after scaling factor calibration) fall within the range described by [85], including the thalamus. Our last attempt to validate this statistical model and accompanying scaling factor estimates was to simulate multisite data using scaling factor estimates and their residual error from the estimate. We found that the power curves align closely, and match when power is at least 80%. We believe that the small deviations from the theoretical model result from scaling factor estimation error and a non-normal scaling factor distribution due to a relatively small sampling of scaling factors (J = 20 sites).

The data acquisition of our study is similar to that of [94], in which the researchers acquired T1-weighted images from 8 consistent human phantoms across 5 sites with non-standardized protocols. These scanners were all 1.5T except for one 1T scanner. [94] calibrated the intensity histograms of the images before segmentation with a calibration factor estimated based on the absolute agreement of volumes to the reference site (ICC). After applying their calibration method, the ICC of the lateral ventricle was ≥ 0.96, which is similar to our pre- and post- calibrated result of 0.97. The ICC for the intensity calibrated gray matter volume in [94] was ≥ 0.84, compared to our calibrated between-site ICC of 0.78 (uncalibrated), and 0.96 (calibrated). Our between-site ICCs for white matter volume (0.96 and 0.98 for the pre- and post- calibrated volumes, respectively) were much higher than those of the intensity calibrated white matter volume in [94] (≥ $.78$). This could be explained by the fact that our cohort of sites is a consortium studying multiple sclerosis, which is a white matter disease, so there may be a bias toward optimizing scan parameters for white matter. Most importantly, the calibration method of [94] requires re-acquisition of a human phantom cohort at each site for each multisite study. Alternatively, multisite studies employing our approach can use the results of our direct-volume calibration (the estimates of CV_a for each ROI) to estimate sample sizes based on our proposed power equation and bias measurements without acquiring their own human phantom dataset to use in calibration.

To our knowledge, this is the first study measuring scaling factors between sites with non-standardized protocols using a single set of subjects, and deriving an equation for power that takes this scaling into account via mixed modeling. This study builds on the work of [77], which investigated the feasibility of pooling retrospective data from three different sites with non-standardized sequences using standard pooling, mixed effects modeling, and fixed effects modeling. [77] found that mixed effects and fixed effects modeling outperformed standard pooling. Our statistical model specifies how MRI bias between sites affects the cross-sectional mixed effects model, so it is limited to powering cross-sectional study designs. Jones and colleagues have derived sample size calculations for longitudinal studies acquired under heterogeneous conditions without the use of calibration subjects [95]. This can be useful for studies measuring longitudinal atrophy over long time periods, during which scanners and protocols may change. For the cross-sectional case, the use of random effects modeling enables us to generalize our results to any protocol with acquisition parameters similar to those described here (primarily MPRAGE). If protocols change drastically compared to our sample of 3D MPRAGE-type protocols, a small set of healthy controls should be scanned before and after any major software, hardware, or protocol change so that the resulting scaling factors can be compared to the distribution of scaling factors (CV_a) reported in this study. A large CV_a can severely impact the power of a multisite study, so it is important not to generalize the results in this study to non-MPRAGE sequences without validation. Potentially, new 3D-printed brain-shaped phantoms with similar regional contrast to noise ratios as human brains may become an excellent option for estimating CV_a.

A limitation of our model is the assumption of independence between the unobserved effect ($D_{U,j}$) at a particular site , j, with the scaling factor of that site (a_j). This assumption does not hold if patients with more severe disease have tissue with different properties that, when scanned, shows different regional contrast than that of healthy controls. As shown in the Appendix, the calculation of the unconditional variance of the observed estimate (equation 3.7) can get quite complicated. We addressed this issue for multiple sclerosis patients by showing that the scaling factors from healthy controls are very similar to those derived from an MS population. The largest difference in scaling factors between healthy controls and multiple sclerosis patients was in white matter volume, where $a_{MS} = 0.967$ and $a_{HC} = 0.975$. A two-sample T test between the scaling factors produced a p-value of 0.88, showing that we could not detect a significant difference between scaling factors of HC and

MS. This part of the study was limited in that we only scanned MS patients at two scanners, while the healthy controls were scanned at 20, so we could not estimate a patient-derived CV_a (the direct input to the power equation). However, the similarity between scaling factors for the subcortical gray matter, cortical gray matter, and white matter volumes between the MS and HC populations suggests that, given careful editing of volumes in the disease population, the independence assumption holds for MS. We recommend that researchers studying other diseases validate our approach by scanning healthy controls and patients before and after an upgrade or sequence change to test the validity of the independence assumption.

Even though we did not standardize the protocols and scanners within this study, the consortium is unbalanced in that there are 16 3T scanners, 11 of which are Siemens. Of the Siemens 3T scanners, there is little variability in TR, TE, and TI, however, there is more variance in the use of parallel imaging, the number of channels in the head coil (12, 20 or 32), and the field of view. Similar to the findings of [96], we could not detect differences in scan-rescan reliability between field strengths. Wolz and colleagues could not detect differences in scan-rescan reliabilities of the hippocampus volumes estimated by the LEAP algorithm, but they detected a small bias between field strengths. They found that the hippocampus volumes measured from the 3T ADNI scanners were 1.17 % larger than those measured from the 1.5T [67]. A two-sample T-test with unequal variances was run between the scaling factors of the 1.5T versus 3T scanners. This test could not detect differences in any ROI except for the left- and right- amydgala. We found that the scaling factors were lower for the 1.5T scanners than for the 3T scanners (0.9 versus 1.02), suggesting that the amygdala volume estimates from the 1.5T were larger than those of the 3T. It should be noted that this interpretation is limited due to the small sample size of 1.5T scanners in this consortium.

Another limitation of this study is that we were under-powered to accurately estimate both the scaling and intercept for a linear model between two sites, and that we did not take the intercept into account when deriving power. We excluded the intercept from our analysis for two reasons: (1) we believe that the nature of systematic error from MRI segmentation is not additive, meaning that offsets in metrics between sites for different subjects is scaled with ROI size instead of a constant additive factor and (2) the model becomes more complicated if site-level effects are both multiplicative and additive. The other limitation of this study is that we assumed that subjects across all sites will come from the same population, and that stratification occurs solely from systematic errors within each site. In reality, sites may recruit from different populations and the true disease effect will vary even more. For example, in a comparison study between the matched ADNI cohort and a matched Mayo Clinic Study of Aging cohort, researchers found different rates of hippocampal atrophy even though no differences in hippocampal volume was detected [97]. This could be attributed to sampling from two different populations. This added site-level variability requires a larger site-level sample size, for an example of modeling this, see [98].

In this study, we reported reliability using both between-site ICC and CV_a because these two metrics have complementary advantages. ICC depends on the true subject-level variability studied. Since we scanned healthy controls, our variance component estimates of subject variability may be lower than that of our target population (patients with multiple sclerosis related atrophy). As a result, ICCs may be lower than expected in MS based on the results of healthy controls. We tried to address this issue by scanning subjects in a large age range, capturing the variability in gray and white matter volume due to atrophy from aging. On the other hand, CV_a is invariant to true subject variability, but is limited by the accuracy of between-site scaling estimates. Both between-site ICC and CV_a should be reported when evaluating multisite reliability datasets to understand a given algorithm's ability to differentiate between subjects (via the ICC) and the magnitude of systemic error between sites (via the CV_a), which could be corrected using harmonization.

3.6 Conclusion

When planning a multisite study, there is an emphasis on acquiring data from more sites because the estimated effect sizes from each site are sampled from a distribution and averaged. Understanding how much of the variance in the distribution is due to scanner noise as opposed to population heterogeneity is an important part of powering a study. For the purposes of this study, we estimated the effect size variability of Freesurfer-derived regional volumes, but this framework could be generalized to any T1-weighted segmentation algorithm, and any modality for which systematic errors are scaled. Scaling factor calibration of metrics resulted in higher absolute agreement of metrics between sites, which showed that the scaling factor variabilities for the ROIs in this study were accurate. The equation for power we outlined in this study along with our measurements of variability between sites should help researchers undestand the trade-off between protocol harmonization and sample size optimization, along with the choice of outcome metrics. Our statistical model and bias measurements enables collaboration between research institutions and hospitals when hardware and software adaptation are not feasible. We provide a comprehensive framework for assessing and making informed quantitative decisions for MRI facility inclusion, pipeline and metric optimization, and study power.

3.7 Acknowledgements

We thank the study participants, MR technicians, and acknowledge Stephane Lehéricy, Eric Bardinet, Frédéric Humbert and Antoine Burgos from the ICM IRM facility (CENIR) and the CIC Pitié-Salpêtrière for their expertise. Funding was provided by R01 NS049477. Additional support was provided by ICM F-75013 Paris, INSERM and IHU-A-ICM (ANR-10-IAIHU-06). BD is a Clinical Investigator of the Research Foundation Flanders (FWO-Vlaanderen). AG and BD are supported by the Research Fund KU Leuven (OT/11/087) and the Research Foundation Flanders (G0A1313N). MRI acquisitions in Hospital Clinic of Barcelona were funded by a "Proyecto de Investigación en Salut" grant (PI15/00587. PIs Albert Saiz and Sara Llufriu) from the Instituto de Salud Carlos III.

3.8 Tables

	1	2	3	4
TR (ms)	8.18	7.10	2130	2080
TE (ms)	3.86	3.20	2.94	3.10
Strength (T)	1.50	1.50	1.50	1.50
TI (ms)	300	862.90	1100	1100
Flip Angle (°)	20	8	15	15
Make	GE	Ph	Si	Si
Voxel Size (mm)	.94x.94x1.2	1x1x1	1x1x1	.97x.97x1
Distortion Correction	N	N	N	Y
Parallel Imaging	-	2	2	-
FOV (mm)	240x240x200	256x256x160	256x256x176	234x250x160
Read Out Direction	HF	AP	HF	HF
Head coil # channels	2*	8	20	12
Model	Signa LX	Achieva	Avanto	Avanto
OS	11x	2.50	VD13B	B17A
Acq. Time (min)	06:24	05:34	04:58	08:56
orientation	sag	sag	sag	sag
# scans	24/24	24/24	24/24	18/18
Amyg (L)	.93	.89	.61	.96
Amyg (R)	.93	.90	.83	.88
Caud (L)	.96	.96	.98	.99
Caud (R)	.96	.97	.90	.96
GMV	.96	.99	.98	.99
Hipp (L)	.94	.95	.89	.93
Hipp (R)	.93	.91	.94	.95
Thal (L)	.77	.93	.59	.82
Thal (R)	.91	.90	.76	.82
cVol	.95	.99	.97	.99
cWMV	.99	1	.99	.99
eTIV	1	1	1	1
scGMV	.98	.97	.98	.93

Table 3.1: Top: Acquisition parameters for the four 1.5T scanners. Si = Siemens, Ph = Philips, GE= General Electric. Bottom: Test-retest reliabilities for selected ROIs, processed by Freesurfer. The ROIs are gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV), and the volumes of the lateral ventricle (LV), amygdala (Amyg), thalamus (Thal), hippocampus (Hipp), caudate (Caud), and finally the estimated total intracranial volume (eTIV). Test-retest reliability is computed as within-site ICC(1,1). * signifies a quadrature coil

	5	6	7	8	9
TR (ms)	8.21	7.80	9	8.21	6.99
TE (ms)	3.22	2.90	4.00	3.81	3.16
Strength (T)	3	3	3	3	3
TI (ms)	450	450	1000	1016.30	900
Flip Angle (°)	12	12	8	8	9
Make	GE	GE	Ph	Ph	Ph
Voxel Size (mm)	.94x.94x1	1x1x1.2	1x1x1	1x1x1	1x1x1
Distortion Correction	N	Y	Y	Y	Y
Parallel Imaging	2	2	3	2	-
FOV (mm)	240x240x172	256x256x166	240x240x170	240x240x160	256x256x204
Read Out Direction	HF	FH	AP	FH	FH
Head coil # channels	8	8	16	32	8
Model	MR750	Signa HDxt	Achieva	Achieva TX	Intera
OS	DV24	HD23.0_V01_1210a	3.2.3.2	5.1.7	3.2.3
Acq. Time (min)	5:02	7:11	05:55	05:38	08:30:00
orientation	sag	sag	sag	sag	sag
# scans	24/24	24/24	24/24	24/24	21/22
Amyg (L)	.67	.89	.66	.85	0.97
Amyg (R)	.88	.79	.91	.94	0.94
Caud (L)	.96	.98	.98	.97	0.98
Caud (R)	.95	.96	.98	.93	0.96
GMV	1	.99	.99	.98	0.99
Hipp (L)	.51	.97	.83	.90	0.99
Hipp (R)	.95	.96	.93	.96	0.99
Thal (L)	.97	.81	.94	.80	0.88
Thal (R)	.70	.87	.96	.96	0.97
cVol	.99	.99	.98	.98	0.99
cWMV	1	.99	1	1	1.00
eTIV	1	1	1	.92	0.99
scGMV	.98	.99	.96	.98	0.99

Table 3.2: Top: Acquisition parameters for the 3T Philips and GE scanners. Ph = Philips, GE= General Electric. Bottom: Test-retest reliabilities for selected ROIs, processed by Freesurfer. The ROIs are gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV), and the volumes of the lateral ventricle (LV), amygdala (Amyg), thalamus (Thal), hippocampus (Hipp), caudate (Caud), and finally the estimated total intracranial volume (eTIV). Test-retest reliability is computed as within-site ICC(1,1)

	10	11	12	13	14	15
TR (ms)	2300	2300	2300	2300	2300	2000
TE (ms)	2.96	2.98	2.98	2.96	2.96	3.22
Strength (T)	3	3	3	3	3	3
TI (ms)	900	900	900	900	900	900
Flip Angle (°)	9	9	9	9	9	8
Make	Si	Si	Si	Si	Si	Si
Voxel Size (mm)	1x1x1	1x1x1.1	1x1x1	1x1x1	1x1x1	1x1x1
Distortion Correction	Y	N	Y	Y	Y	N
Parallel Imaging	2	-	2	2	2	2
FOV (mm)	256x256x176	240x256x176	240x256x176	240x276x156	256x256x176	256x208x160
Read Out Direction	HF	RL	HF	HF	HF	RL
Head coil # channels	20	32	20	20	20	32
Model	Prisma	Prisma fit	Skyra	Skyra	Skyra	Skyra
OS	D13D	VD13D	VD13	VD13	VD13C	VD13
Acq. Time (min)	05:09	07:46	05:12	05:12	05:09	04:56
orientation	sag	sag	sag	sag	sag	ax
# scans	22/22	24/24	25/25	23/24	23/24	22/22
Amyg (L)	.83	.89	.80	.85	.98	.89
Amyg (R)	.94	.92	.93	.85	.93	.84
Caud (L)	.99	.99	.98	.99	.98	.98
Caud (R)	.99	.96	.95	.95	.98	.97
GMV	.99	.98	.99	1	.99	.97
Hipp (L)	.94	.98	.99	.95	.97	.98
Hipp (R)	.91	.94	.97	.98	.95	.96
Thal (L)	.92	.87	.87	.76	.91	.89
Thal (R)	.74	.93	.80	.91	.93	.89
cVol	.99	.98	.98	1	.99	.96
cWMV	1	1	1	1	1	.97
eTIV	1	1	1	1	1	.97
scGMV	.98	.99	.98	.98	.99	.99

Table 3.3: Top: Acquisition parameters for the 3T Siemens (Si) Skyra and Prisma scanners. Bottom: Test-retest reliabilities for selected ROIs, processed by Freesurfer. The ROIs are gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV), and the volumes of the lateral ventricle (LV), amygdala (Amyg), thalamus (Thal), hippocampus (Hipp), caudate (Caud), and finally the estimated total intracranial volume (eTIV). Test-retest reliability is computed as within-site ICC(1,1)

	16	17	18	19	20
TR (ms)	2300	2150	1900	1900	1800
TE (ms)	2.98	3.40	3.03	2.52	3.01
Strength (T)	3	3	3	3	3
TI (ms)	900	1100	900	900	900
Flip Angle (°)	9	8	9	9	9
Make	Si	Si	Si	Si	Si
Voxel Size (mm)	1x1x1	1x1x1	1x1x1	1x1x1	.86x.86x.86
Distortion Correction	N	N	N	N	N
Parallel Imaging	2	2	2	2	2
FOV (mm)	256x256x176	256x256x192	256x256x176	256x256x192	220x220x179
Read Out Direction	HF	RL	AP	FH	FH
Head coil # channels	12	12	12	32	32
Model	Trio	Trio	Trio	Trio	Trio
OS	MRB17	VB17	VB17A	VB17	MRB19
Acq. Time (min)	05:03	04:59	04:26	05:26	06:25
orientation	sag	ax	sag	sag	sag
# scans	24/24	23/24	23/24	24/24	24/24
Amyg (L)	.55	.88	.77	.88	.91
Amyg (R)	.85	.93	.81	.94	.93
Caud (L)	.99	.95	.97	.97	.97
Caud (R)	.97	.92	.98	.91	.95
GMV	.99	.99	.98	.99	1
Hipp (L)	.71	.96	.94	.93	.96
Hipp (R)	.94	.94	.92	.83	.96
Thal (L)	.45	.85	.80	.80	.88
Thal (R)	.61	.95	.85	.96	.79
cVol	.99	.98	.96	.99	1
cWMV	1	.99	.99	1	1
eTIV	.97	1	1	1	1
scGMV	.98	.98	.98	.98	.98

Table 3.4: Top: Acquisition parameters for 3T Siemens (Si) Trio scanners. Bottom: Test-retest reliabilities for selected ROIs, processed by Freesurfer. The ROIs are gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV), and the volumes of the lateral ventricle (LV), amygdala (Amyg), thalamus (Thal), hippocampus (Hipp), caudate (Caud), and finally the estimated total intracranial volume (eTIV). Test-retest reliability is computed as within-site ICC(1,1)

	CVa
variable	
LV (L)	0.03
LV (R)	0.03
cWMV	0.02
cVol	0.04
scGMV	0.02
GMV	0.04
Caud (L)	0.02
Caud (R)	0.07
Amyg (R)	0.09
Amyg (L)	0.07
Hipp (L)	0.03
Hipp (R)	0.03
Thal (L)	0.05
Thal (R)	0.05

Table 3.5: Coefficient of variability (CV_a) values for selected ROIs. CV_a was defined in equation 3.14. The ROIs are gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV, which does not include cerebellar white matter), and the volumes of the lateral ventricle (LV), amygdala (Amyg), thalamus (Thal), hippocampus (Hipp), caudate (Caud), and finally the estimated total intracranial volume (eTIV)

ROI	ICC BW	ICC BW Cal	[54] ICC BW
GMV	.78	.96	.854
WMV	.96	.98	.774
Thal	.61	.73	.95
Hipp	.75	.84	.79
Amyg	.56	.74	.76
Caud	.82	.91	.92

Table 3.6: Between-site ICC comparison to the study by [54], where MRI sequences were standardized and subcortical segmentation was performed using FIRST, and cortical segmentation using cortical pattern matching. ICC BW and ICC BW Cal were calculated using our multisite healthy control data, where ICC BW Cal was calculated as the between site ICC of volumes after applying the scaling factor from a leave-one-out calibration. Other than the thalamus (Thal), we found that the between-site ICCs were comparable to [54] for the amygdala (Amyg), caudate (Caud), and even higher for the hippocampus (Hipp), gray matter volume (GMV) and white matter volume (WMV)

ROI	ICC WI	ICC WI Cal	[85] ICC WI Average
LV	1	1	.998 ± 0.002
Thal	.86	.84	0.765 ± .183
Hipp	.93	.93	0.878 ± .132
Amyg	.89	.86	0.761 ± .134
Caud	.97	.97	0.909 ± 0.092

Table 3.7: Comparing the within-site ICC before and after leave-one-out scaling factor calibration with the cross-sectional freesurfer results of [85], where scanners were standardized, and the average within-site ICC is shown. The within-site ICCs of our study fall within the range of [85], which shows the that sites in this study are as reliable as those in [85].

3.9 Figures

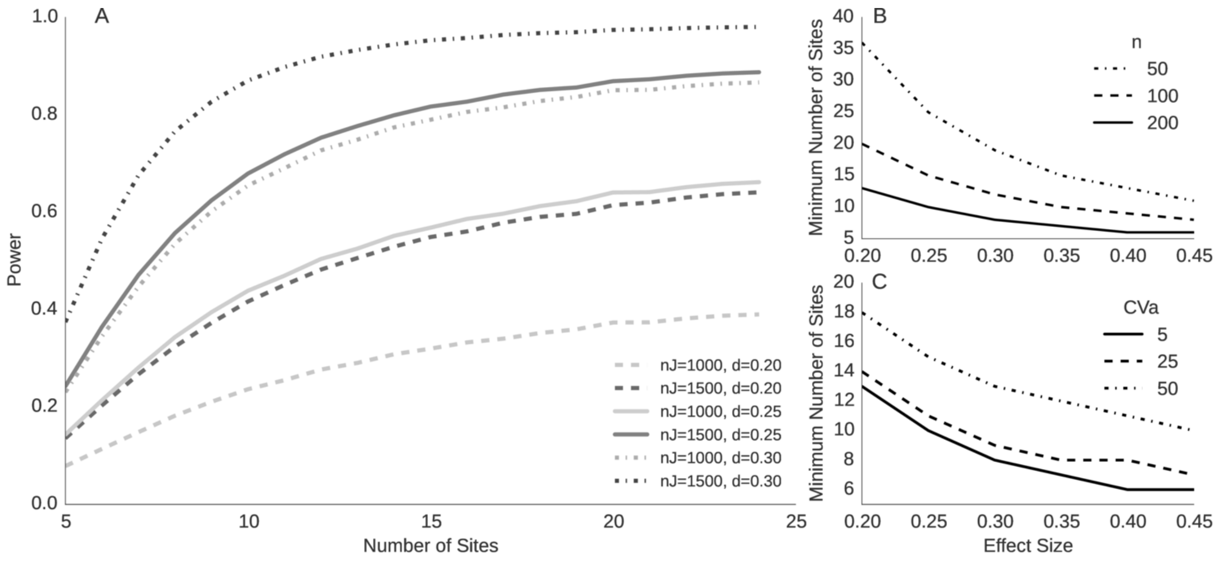


Figure 3.1: **A.** Power contours for total number of subjects (nJ) over various effect sizes (d), p= 0.002, CV_a = 5%. **B.** # of sites required for effect sizes and # subjects per site (n). **C** effect of CV_a on # sites for various effect sizes, where n = 200 subjects per site

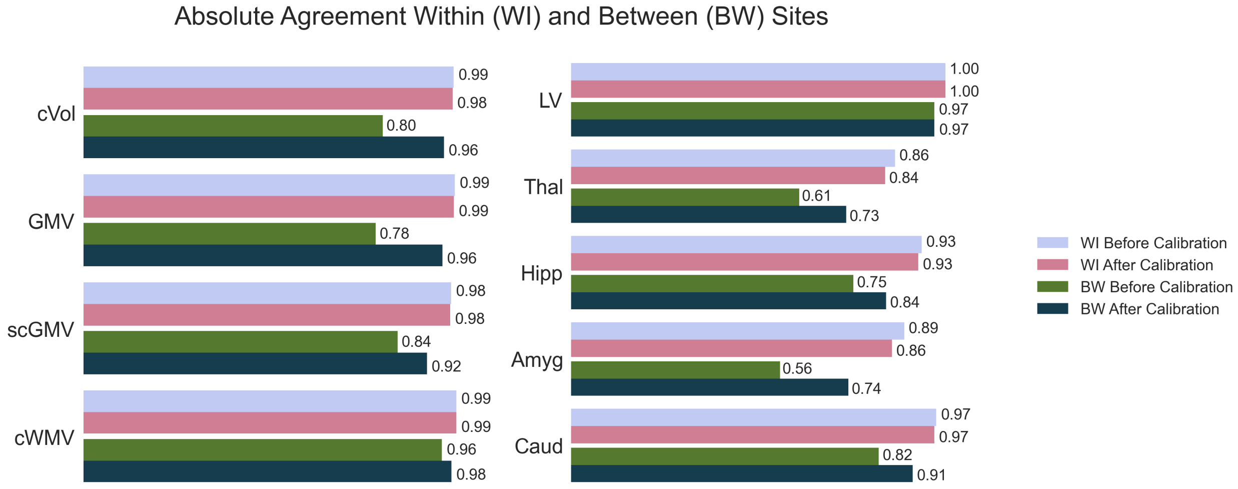


Figure 3.2: Leave-one-out calibration improvement on within- (WI) and between- (BW) site ICCs for gray matter volume (GMV), subcortical gray matter volume (scGMV), cortex volume (cVol), cortical white matter volume (cWMV), lateral ventricle (LV), Thalamus (Thal), Hippocampus (Hipp), Amygdala (Amyg), Caudate (Caud)

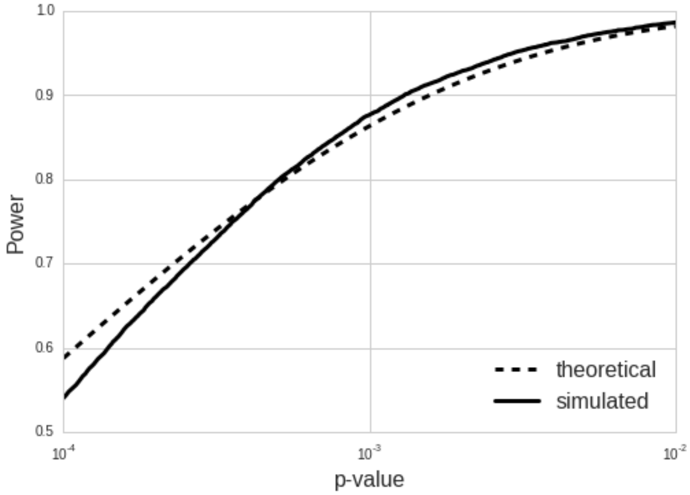


Figure 3.3: Theoretical power vs. simulated power with scaling factor uncertainty

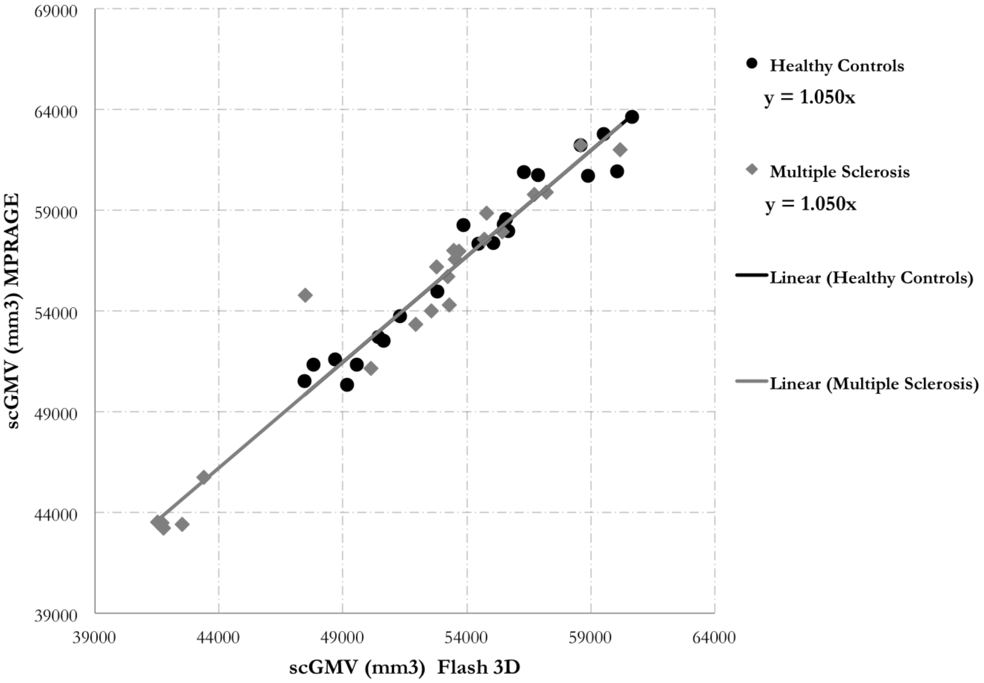


Figure 3.4: Sub-cortical gray matter volume (scGMV) calibration between 2 scanners/sequences at UCSF. The trendline fit shows the slopes (scaling factors) are identical for the healthy control and MS populations

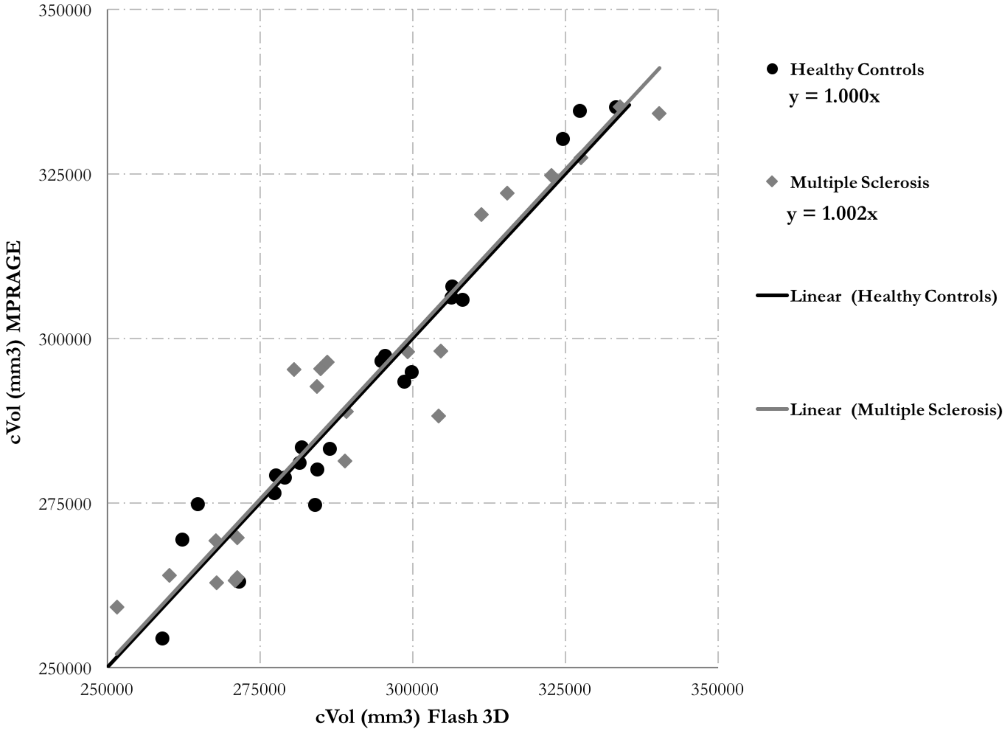


Figure 3.5: Cortex gray matter volume (cVol) calibration between 2 scanners/sequences at UCSF. The trendline fit shows the slopes (scaling factors) are very close for the healthy control and MS populations

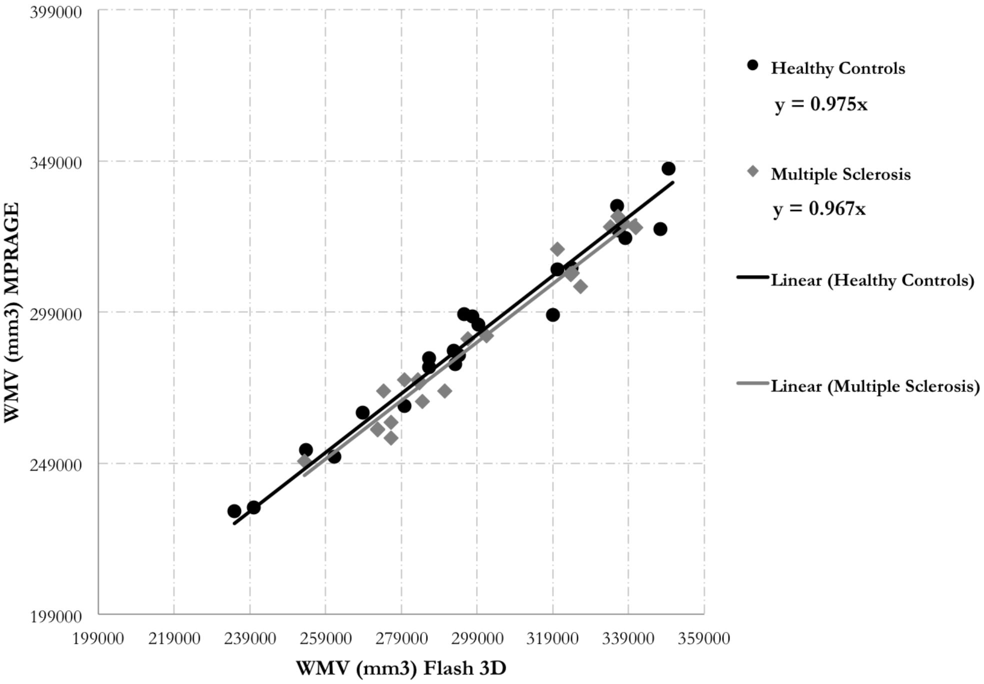


Figure 3.6: White matter volume (WMV) calibration between 2 scanners/sequences at UCSF. The trendline fit shows the slopes (scaling factors) are very close for the healthy control and MS populations

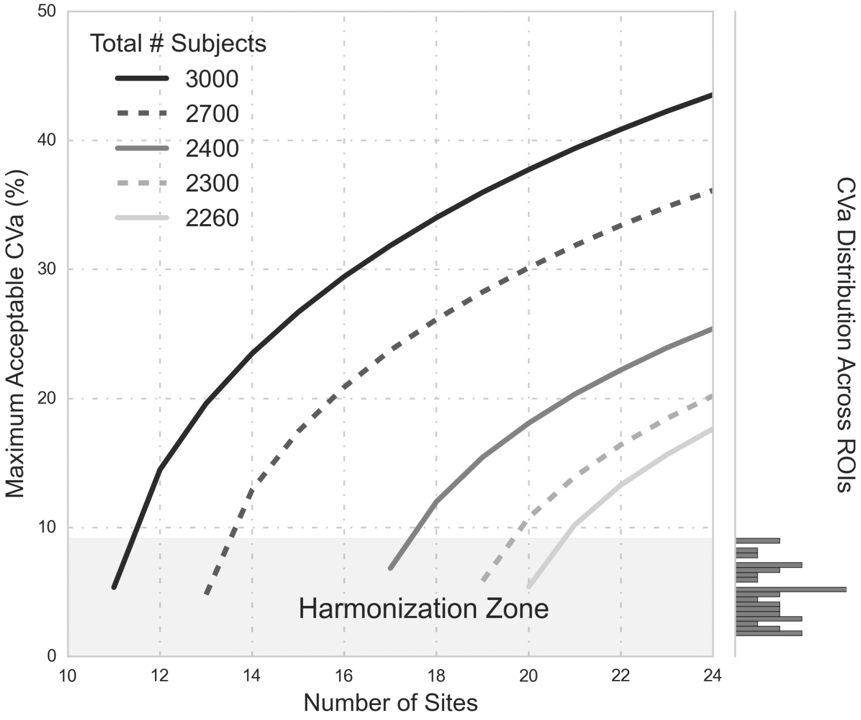


Figure 3.7: Shows power curves for 80% power for 2260 - 3000 total subjects, where the false positive rate is 0.002, and the effect size is 0.2. The lowest point of each curve shows the minimum number of sites required for a given number of subjects on the x-axis and the y-axis corresponds to the maximum acceptable coefficient of variability (CV_a, defined in 3.14) for that case. The right-hand side of the chart shows the distribution of CV_a values across all sites and all Freesurfer ROIs. When minimizing the total number of sites for a set number of subjects, the maximum allowable CV_a is around 5%, meaning that if the CV_a is higher than 5% for a particular ROI, the power of the model will fall below 80%. The shaded section on the bottom of the chart called the "Harmonization Zone" which indicates the regions of the graph where the maximum acceptable CV_a is below the largest CV_a across all freesurfer ROIs (which is the right amygdala at 9%). If site- and subject- level sample sizes fall within the harmonization zone, efforts to harmonize between sites is required to guarantee power for all ROIs.

3.10 Appendix

3.10.1 Variance of a Product of Random Variables

The proof for this is found in Introduction to the Theory of Statistics (1974) by Mood, Graybill and Boes [99], section 2.3, Thoerem 3:

Let X and Y be two random variables where $var[XY]$ exists, then

$$\begin{aligned} var[XY] = \mu_Y^2 var[X] + \mu_X^2 var[Y] + 2\mu_X\mu_Y cov[X,Y] \\ -(cov[X,Y])^2 + E[(X-\mu_X)^2(Y-\mu_Y)^2] \\ +2\mu_Y E[(X-\mu_X)^2(Y-\mu_Y)] + 2\mu_X E[(X-\mu_X)(Y-\mu_Y)^2] \end{aligned} \tag{3.17}$$

which can be obtained by computing $E[XY]$ and $E[(XY)^2]$ when XY is expressed as

$$XY = \mu_X\mu_Y + (X-\mu_X)\mu_Y + (Y-\mu_X)\mu_X + (X-\mu_X)(Y-\mu_Y) \tag{3.18}$$

If X and Y are independent, then $E[XY] = \mu_X\mu_Y$, the covariance terms are 0, and

$$E[(X-\mu_X)^2(Y-\mu_Y)^2] = E[(X-\mu_X)^2]E[(Y-\mu_Y)^2] = var[X]var[Y] \tag{3.19}$$

and

$$\mu_Y E[(X-\mu_X)^2(Y-\mu_Y)] = E[(X-\mu_X)^2]E[(Y-\mu_Y)] = 0 \tag{3.20}$$

$$\mu_X E[(Y-\mu_Y)^2(X-\mu_X)] = E[(Y-\mu_Y)^2]E[(X-\mu_X)] = 0 \tag{3.21}$$

Which gives

$$var[XY] = \mu_X^2 var[Y] + \mu_Y^2 var[X] + var[X]var[Y] \tag{3.22}$$

3.10.2 Maximum Likelihood

Note that the estimator defined in 3.9 is a maximum likelihood estimator under the condition of equal unexplained variance at each site and an equal number of subjects at each site. In the case with different number of subjects at each site, the maximum likelihood estimator for the disease effect, $\hat{\beta}_{10}$, is not the average of the site-level coefficients, but instead is the average weighted by the inverse error variance. This is a common method to run meta-analyses, for example, see [98, 59]. To show this, we follow the procedure from [98], and define the likelihood of the alternate hypotheses as

$$L_1 = \prod_j \frac{1}{\sqrt{2\pi a_j^2 V_j}} exp\Big(\frac{-(\beta_{1j} - \mu)^2}{2a_j^2 V_j}\Big) \tag{3.23}$$

for a non-zero μ and V_j defined as the unscaled error variance on $\hat{\beta_{1,j}}$. The maximum likelihood estimator $\hat{\mu}$ is found by taking the derivative of the log of (3.23), setting it equal to 0, and solving for μ,

$$\frac{\partial}{\partial\mu}\Big(log(L_1)\Big) = \frac{\partial}{\partial\mu}\Big(\sum_j^J log(\frac{1}{\sqrt{2\pi a_j^2 V_j}}) + \sum_j^J \frac{(\beta_{1j} - \mu)^2}{2a_j^2 V_j}\Big) = 0 \implies \hat{\mu} = \frac{\sum_j^J a_j^{-2} V_j^{-1} \beta_{1j}}{\sum_j^J a_j^{-2} V_j^{-1}} \tag{3.24}$$

which shows that the inverse variance weighted average is the maximum likelihood estimator for the overall treatment effect. If we assume that the unexplained variance (σ_0) is the same across all sites, which is a valid assumption if subjects are from the same population, the estimate can be expressed as

$$\hat{\beta}_{10} = \frac{\sum_{j=1}^{J} n_j \hat{\beta}_{1j}}{N} = \frac{\beta_{10} \sum_{j=1}^{J} n_j a_j}{N} \tag{3.25}$$

where $N = \sum^{J} n_j$ is the total number of subjects in the study. The variance of the estimate is

$$var(\hat{\beta}_{10}) = \frac{\sigma_0^2 \alpha_0^2}{N^2} \sum_{j=1}^{J} 4n_j + CV_\alpha^2 (4n_j + \delta^2 n_j^2) \tag{3.26}$$

and it follows that the noncentrality parameter is

$$\lambda = \frac{\delta^2 \left(\sum_{j=1}^{J} n_j \frac{a_j}{\mu_a} \right)^2}{\sum_{j=1}^{J} 4n_j + CV_a^2 (4n_j + \delta^2 n_j^2)} \tag{3.27}$$

which should be used for a more accurate power analysis if the specific number of subjects per site and the site's scaling factors are known.

4 PBRain- A Framework for the Curation and Execution of Neuroimaging Analyses

4.1 Introduction

Neuroimaging has helped us probe important questions about the brain, like the structure-function relationship and how the brain is affected by neurological and psychiatric illnesses. Given the large variability in brain structure across the population, studies must sample large numbers of people, and this presents many challenges associated with "big data". First, studies take a long time to complete, with multiple people working on data at different times. Data organization structures vary from person to person, along with in-house scripts configured for particular, non-standardized file structures. Next, datasets consist of multiple modalities in complex formats, which may change throughout the course of the study. Analyses scripts are structured with different programming styles, often in different programming languages, and are saved in no particular standard structure within the file system. Commenting style varies from person to person; some analyses scripts are sparsely commented, making it difficult for new people to learn and adapt scripts for different use cases. Neuroimaging processing steps are quite complex and require multiple steps that are dependent on each other; as such, it is difficult to know which scripts depend on others being run first, without proper consistent documentation. Finally, the provenance of files written to the filesystem is unclear when there is no standardized folder structure or naming convention for neuroimaging outputs.

Lack of a standardized organization scheme for 1) raw data, 2) processed data, 3) the location of analyses scripts, and 4) the internal organization and comment level of analyses scripts leads to many inefficiencies that hinder scientific progress in the neuroimaging field. Time is wasted re-writing scripts and reorganizing data. There is a high learning curve to editing scripts, which is particularly problematic in multi-disciplinary labs with members from backgrounds that don't include formal computer science training (psychologists, radiologists, etc). This leads to difficulties in transferring knowledge to newer lab members, which is important given that most research labs have a high turnover rate of research assistants and PhD students. It becomes difficult to collaborate and transfer knowledge outside of the lab, which is problematic for multisite neuroimaging studies that need to share data and analyses scripts. As the open-data movement continues to gain popularity, it is difficult to incorporate this data into local analyses within the lab when file organization structures are so different. This lack of consistency in file and analysis organization structure increases the likelihood of errors and data loss. An ideal solution would 1) minimize data curation time 2) minimize the likelihood of errors and data loss 3) minimize the learning curve for running basic analyses 4) separate data input/output (IO) from analyses scripts, so that analyses can be run on data that is organized differently, 5) standardize code and comment structure to make it easy for people to collaborate on analyses, and 6) enforce analysis script dependencies in order to reduce time spend debugging errors.

Neuroinformatics researchers have begun to address these "big data" problems, mostly in the functional MRI domain. In 2012, I wrote BrainImagingPipelines [100, 101], which was an open-source framework that provided curated and configurable data analysis pipelines for structural, functional, and diffusion imaging analyses. This framework was used by other researchers to design sparse-sampling methods for fMRI analyses [102], and to study reading [103], the default mode functional network [104], memory perception [105], and working memory [106] in adolescents. Soon after, the Configurable Pipeline for the Analysis of Connectomes (CPAC) was released [107], which consists of a set of workflows to analyze the brain connectome. In 2016, a standardized data organization structure to enable easier data sharing was proposed, as a more generalized extension of the OpenfMRI project[108] called Brain Imaging Data Structure [109]. As a follow up to BIDS, a set of pipelines was developed to analyze BIDS-compliant data in an August 2016 code-sprint that I participated in, which resulted in the collection of BIDS-Apps [110].

My work is focused on the clinical, multiple sclerosis use-case rather than fMRI. I present PBRain, which is a framework for the curation and execution of neuroimaging analyses. PBRain requires defining and enforcing the organization of 1) raw data 2) processed data, and 3) analyses code. The structure of the analyses code is standardized through the Nipype python library [53], and dependencies between workflows is defined as attributes to each PBRain workflow. Finally, version control is accomplished using git and GitHub, so that changes are always traced. This has led to a multidisciplinary collaboration between clinicians, research assistants, and software engineers in the radiology, neurology, and neurosurgery domains. For example, the pipeline developed by Kesshi Jordan to evaluate the functional consequences of brain tumor resection was integrated into the PBRain platform for easy reproducibility. Finally, the software is easy to install and configure in different computing environments, and has been run on the UCSD compute cluster. The following sections outline the design decisions, the implementation details, current pipeline implementations, and a discussion on the implications and future directions of the PBrain software package.

4.2 Design Requirements and Solutions

4.2.1 Input and Output (IO)

IO is often a major bottleneck of large neuroimaging analyses, because different scripts assume a certain file organization structure, and it can be difficult to separate the IO logic from the analysis logic. The first design requirement was to write code in a modular way such that the file IO handling was separated from the actual pipeline mechanics. This modularization enables us to switch data organization structure without interfering with the analysis, and results in code that is much more organized and easier to understand. Next, dependencies between analyses must be explicitly defined in the structure of the code, so that the software can automatically run any prerequisite analyses. Finally, workflow output structure needs to be explicitly defined and documented in the class, so that collaborators easily understand the nature of the files and how to locate them. Additionally, worflow outputs should be named in a way that is easy to trace back to the original file. This provenance information is saved in 2 ways: 1) by propagating the original filename throughout all outputs, and 2) by using Nipype's [53] built-in provenance tracking.

Figure 4.1A shows the PBrain interface class consists of 4 sections, the inputs, connections, run method, and outputs. Workflow autoassembly(figure 4.1B) is a natural result of this class specification. The IO structure of inputs/outputs is shown in figure 4.1AC and 4.1AD.

4.2.2 Collaboration

Keeping track of analysis code and outputs is essential for collaboration, and error checking. PBR is version controlled using Git, and hosted privately on Github. Git enables users to revert to previous versions of code, which is useful in tracking errors. Github allows users to post questions and have discussions on the "issues" page, and perform "code reviews". Documentation is crucial for efficient collaboration, and is implemented via the Sphinx python library, which can automatically generate documentation from class and function docstrings. Consistent code structure is needed to collaborate, as this reduces the difficulty in reading others' scripts. PBrain is based on the Nipype software package [53], where each node of the workflow is specified as an object, and are connected together to form workflows. The Nipype structures enables users to automatically generate graphs of their workflow logic (see figure 4.2).

4.3 Implementation Details

The PBRain interface model builds on the Nipype Interface model, such that PBRain interfaces inherit properties from the Nipype interface, such as 1) caching 2) workflow node connections 3) workflow execution engine, and 4) input/output specifications. Additionally, the PBRain interface requires specification of the connections between interfaces. This is an explicit definition of dependencies, and allows the workflows to automatically assemble. Each analysis is implemented as a Nipype Workflow, which is nested within the PBRain interface. Each input must be given a unique name, data type, and description. Documentation is automatically generated from this information. The output specification is where the standardized file structure is enforced, either the cross-sectional or longitudinal structure. The output specification is similar in structure to the input spec, where input file types and descriptions must be specified. Within the run method, a Nipype workflow is imported and run with the files in the PBR input specification. Because of this nested structure, parallelization can occur on two levels: the outer PBRinterface level or the inner Nipype interface level. A command line tool is exposed, such that users can call PBR via the following syntax:

```
pbr <exam ids> -w <workflow name> -ps <PACS database password> -p <parallelization arg
```

where the <exam id> argument can be a list of unique exam ids, specified in the command line or in a text file. The PACS database password is used to connect to the PACS archival dicom system to pull dicoms for exam ids that do not have raw data on the file system. This step is optional and can be bypassed with the

```
--no_dicoms
```

flag, but when using the

```
--no_dicoms
```

flag, the nifti data must be organized according to the PBRain specification. Other software "layers" include a config.py file, where directories and paths to standard atlases are set for a specific file system, and a heuristic file that maps series description names to modality types, like "T1", "T2", etc.

4.3.1 Example code structure

The following code structure, which is similar to the Nipype structure, shows how to define a PBrain interface, and is used as a template to write other PBrain workflows:

```
#imports
from nipype.interfaces.base import (BaseInterface, TraitedSpec, traits, File,
                                    OutputMultiPath, BaseInterfaceInputSpec,
                                    isdefined, InputMultiPath)


from ...config import config
from glob import glob
import os
from ...base import register_workflow, PBRBaseInputSpec, PBRBaseInterface

# explicit definition of inputs
class InterfacenameInputSpec(PBRBaseInputSpec):
    input_file = InputMultiPath(File(exists=True), description="useful description of
        the file")

# explicit definition of ouptuts
class IntefacenameOutputSpec(TraitedSpec):
    output_file = traits.List(File(exists=True), minlen=1, description="useful
        description of the output file")

class Interfacename(PBRBaseInterface):
    """
```

```
Docstring goes here
"""
input_spec = InterfacenameInputSpec
output_spec = InterfacenameOutputSpec
flag = "interfacename" #this is for pbr mse# −w interfacename
connections = [(" nifti ", " t1_files ", " input_file ")] #define how interfaces relate
    to each other

def _run_interface_pbr ( self , runtime):
    # how to run the interface
    wf = get_wf(config ) #function that returns a nipype workflow

    inputspec = wf.get_node("inputspec")
    inputspec . inputs . t1_files = self . inputs . input_file #connect inputs to the
        workflow
    wf. config = {"execution": {"crashdump_dir": os . path . join ( config ["
        crash_directory " ], self . inputs .mseID,           self . flag )}}
    wf.run(plugin=self . inputs . plugin , plugin_args=self . inputs . plugin_args ) #run
        the workflow
    return runtime



def _get_output_folder ( self ):
    return "output_folder_name"

def _list_outputs ( self ):
    #define how to locate the output files
```

```
        outputs = self._outputs().get()

        outputs[" output_file "] = glob(os.path.join(config[" output_directory "],

                                                    self.inputs.mseID, "output_folder_name

                                                       ","*.nii.gz"))


        return outputs


register_workflow(Interfacename) # register the class so it can be called from the

   command line
```

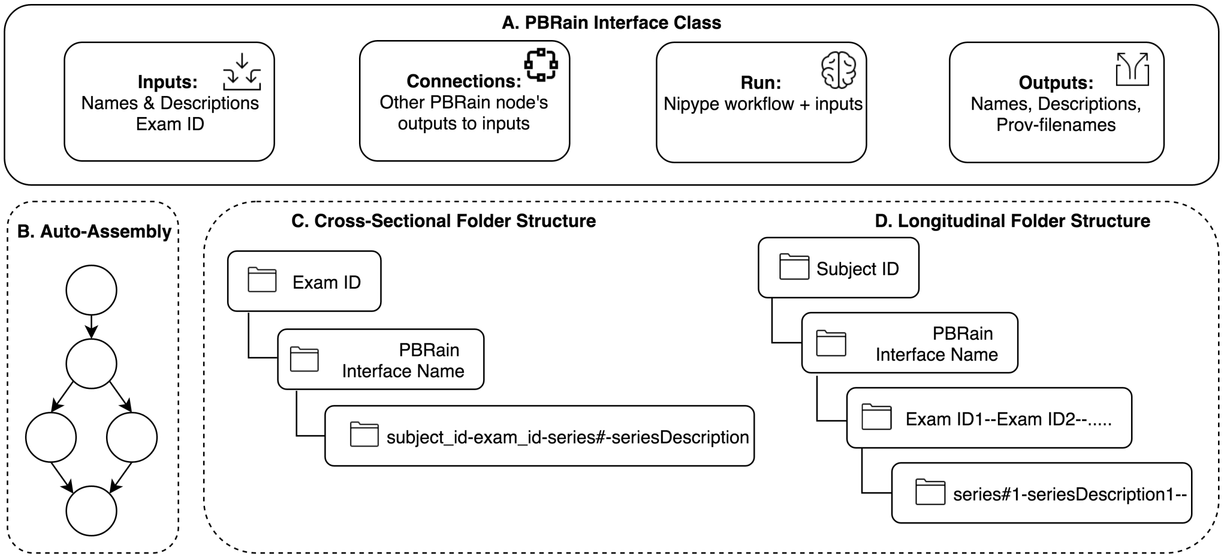


Figure 4.1: The PBRain Diagram. A) The base PBRain interface class consists of 1) the Inputs attribute, which specifies the input names, data types, and a description 2) the Connections attribute, which specifies which outputs are connected to the specified inputs 3) the Run method, where a generalized Nipype workflow is imported, the inputs are connected to the workflow, and the workflow is run, and 4) the outputs attribute, which specifies output names, data types, descriptions, and explicit paths to output files. B) Based on the Connections attribute, a PBRAIN workflow auto-assembles, and runs all dependencies before executing a particular node. C) Shows the cross-sectional folder structure naming convention for PBRain interfaces that run at a single timepoint. D) Shows the longitudinal folder structure when a PBRain interface is called on multiple exam IDs for the same subject.

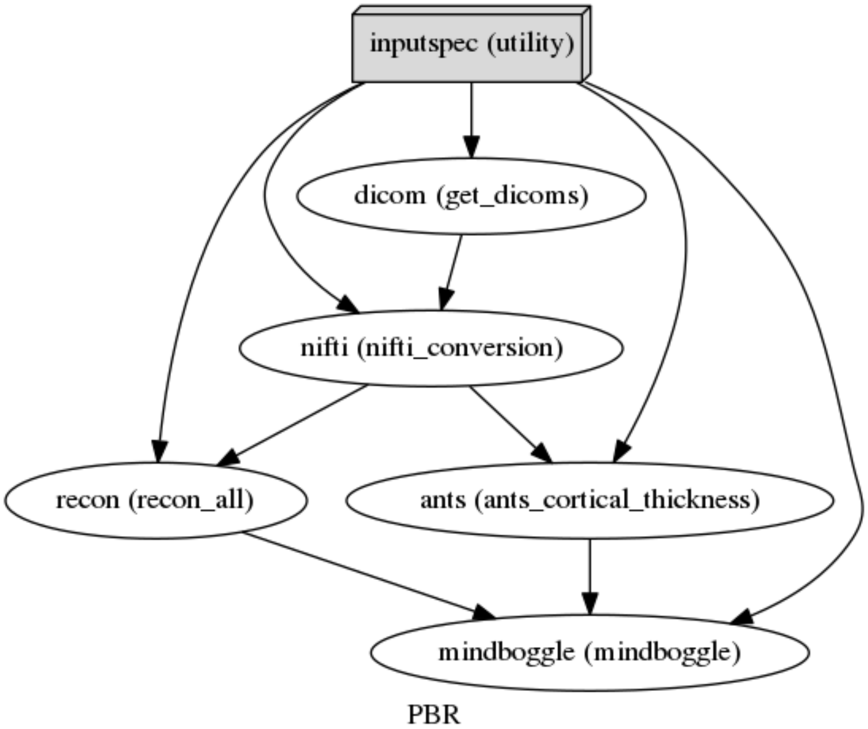


Figure 4.2: Workflow auto-assembly example for the Mindboggle workflow. When the user asks to run Mindboggle on a particular exam, the workflow recursively checks and connects dependencies. In this example, the steps are: 1) pull dicoms from our PACS server 2) Convert dicoms to NifTI format and organize according to the PBRain structure, 3) run Freesurfer's recon-all and ANTS Cortical Thickness pipeline, in parallel if specified, and finally, 4) to run Mindboggle.

4.4 Discussion

The PBrain software framework for the collaborative execution and curation of brain imaging data analyses has greatly improved the efficiency of data processing in our lab. Anyone in the lab can pull data from the PACS database and minimally process data with no knowledge of any programming language. This has reduced the startup time needed for new lab members to aquaint themselves with our computing environment. Regardless of the project, we are aware of the source of the processing outputs, which helps us more efficiently manage data storage in the lab. Version control of the different script versions that were used enabled more advanced lab members to add new processing scripts for others to run. Some examples of this include diffusion processing pipelines for brain tumor pre-surgical mapping from PhD student Kesshi Jordan, a spinal cord segmentation pipeline

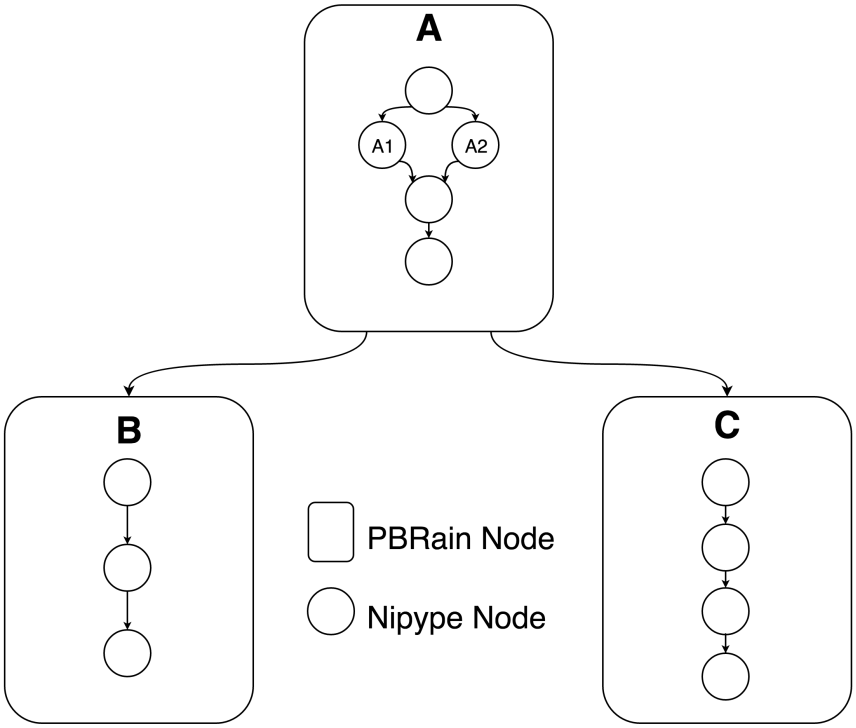


Figure 4.3: PBRain nested workflow structure. PBRain workflows are built on top of the Nipype worklfow architecture. Nipype workflows are nested inside PBRain nodes A, B and C. This modularity enables the user to specify parallelization mode with more detail. For example, the user can specify that the PBRain nodes run serially, but that the Nipype workflow within it runs in parallel. In this case, PBRain node A will run first, with Nipype nodes A1 and A2 running in parallel, followed by PBRain nodes B, and C. Conversely. the user can specify that Nipype nodes run serially, but PBRain nodes run in parallel. In this case, PBRain node A runs first, with Nipype node A1 running before A2, and then PBRain nodes B and C will run in parallel. The user can also run workflows fully in parallel, where both Nipype nodes and PBRain nodes run in parallel (when possible) or completely serially

from PhD student Esha Datta, a longitudinal atrophy pipeline from research assistant James Zhang, and a processing pipeline for a clinical MR visualization tool by research scientist Jason Crane.

The main limitation to PBrain is that there is a steep learning curve to the contribution of new pipelines. This requires an intimate knowledge of the Nipype framework, which involves mastery of object oriented programming techniques in Python. The dependence of PBrain on the Nipype framework is limiting, because if the Nipype structure changes drastically, all PBRain interfaces will need to be rewritten. PBRain is not open-source, so outside contributors will not be able to help debug problems; all debugging falls to any lab members with access to the code base. In the future, changing the file structure to the BIDS data structure will help foster collaboration between other neuroimaging labs, even if our PBRain codebase is private. Debugging PBRain workflows is also challenging, which becomes especially problematic when users run bulky batch processing. It is difficult to determine which nodes failed and for what reason. In the future, an interactive browser-based dashboard will be developed for PBRain to help beginner users debug failures.

4.5 Conclusion

As the number of neuroimaging data samples grow, efficient computing solutions will become crucial to the success of large scale projects. The framework presented here aids in the efficient storage and computation of large datasets, through standardized file structures, thorough documentation, and collaboration via GitHub. Finally, reducing the learning curve needed to batch process data is vital in the context of a research lab, given the high turnover rate of research assistants, graduate students and postdoctoral researchers; the framework presented here empowers new users to run complex data analyses to answer important clinical research questions.

5 Mindcontrol- A Web Application for Brain Segmentation Quality Control

Abstract

Tissue classification plays a crucial role in the investigation of normal neural development, brain-behavior relationships, and the disease mechanisms of many psychiatric and neurological illnesses. Ensuring the accuracy of tissue classification is important for quality research and, in particular, the translation of imaging biomarkers to clinical practice. Assessment with the human eye is vital to correct various errors inherent to all currently available segmentation algorithms. Manual quality assurance becomes methodologically difficult at a large scale - a problem of increasing importance as the number of data sets is on the rise. To make this process more efficient, we have developed Mindcontrol, an open-source web application for the collaborative quality control of neuroimaging processing outputs. The Mindcontrol platform consists of a dashboard to organize data, descriptive visualizations to explore the data, an imaging viewer, and an in-browser annotation and editing toolbox for data curation and quality control. Mindcontrol is flexible and can be configured for the outputs of any software package in any data organization structure. Example configurations for three large, open-source datasets are presented: the 1000 Functional Connectomes Project (FCP), the Consortium for Reliability and Reproducibility (CoRR), and the Autism Brain Imaging Data Exchange (ABIDE) Collection. These demo applications link descriptive quality control metrics, regional brain volumes, and thickness scalars to a 3D imaging viewer and editing module, resulting in an easy-to-implement quality control protocol that can be scaled for any size and complexity of study.

5.1 Background

Imaging biomarkers derived from MRI play a crucial role in the fields of neuroscience, neurology, and psychiatry. Estimates of regional brain volumes and shape features can track the disease progression of neurological and psychiatric diseases such as Alzheimer's disease [111, 112], Parkinson's disease [113], schizophrenia [114], depression [115], autism [116], and multiple sclerosis [117]. Given recent increases in data collection to accommodate modern precision-medicine approaches, assuring the quality of these biomarkers is vital as we scale their production.

Various semi-automated programs have been developed to estimate MRI biomarkers. While these applications are efficient, errors in regional segmentation are inevitable, given several methodological challenges inherent to both technological and clinical implementation limitations. First, the quality of the MRI scan itself due to motion artifacts or scanner instabilities could blur and distort anatomical boundaries [118, 119, 120, 121]. Differences in MRI hardware, software, and acquisition sequences also contribute to contrast differences and gradient distortions that affect tissue classification, which makes combining datasets across sites challenging [122]. An additional source of error comes from parameter selection for segmentation algorithms; different parameter choices can translate to widely varying results [123]. Furthermore, many MR segmentation algorithms were developed and tested on healthy adult brains; applying these algorithms to brain images of children, the elderly, or those with pathology may violate certain assumptions of the algorithm, resulting in drastically different results.

Several quality assurance strategies exist to address segmentation errors. In one approach, researchers flag low-quality scans prior to analysis by viewing the data before input to tissue classification algorithms. However, identifying "bad" datasets using the raw data is not always straightforward, and can be prohibitively time consuming for large datasets. Pre-processing protocols have been developed to extract metrics that can be viewed as a cohort-level summary from which outliers are selected for manual quality-assurance. For example, by running the Preprocessed-Connectomes Project's Quality Assurance Protocol (PCP-QAP) [124], researchers can view summary statistics that describe the quality of the raw data going into the algorithm and automatically remove subpar images. However, these metrics are limited because segmentation may still fail even if the quality of the scan is good. Another quality assurance strategy is to plot distributions of the segmentation output metrics themselves and remove any outlier volumes. However, without manual inspection, normal brains that naturally have very small or large estimates of brain size or pathological brains with valid segmentations may be inappropriately removed. Ideally, a link would exist between scalar summary statistics and 3D/4D volumes. Such a link would enable researchers to prioritize images for labor-intensive quality control (QC) procedures; to collaborate and organize QC procedures; and to understand how scalar quality metrics, such as signal to noise ratio, relate to the actual image and segmentation. In this report, we present a collaborative and efficient MRI QC platform that links group-level descriptive statistics to individual volume views of MRI images.

We propose an open source web-based brain quality control application called Mindcontrol: a dashboard to organize, QC, annotate, edit, and collaborate on neuroimaging processing. Mindcontrol provides an intuitive interface for examining distributions of descriptive measures from neuroimaging pipelines (e.g., surface area of right insula), and viewing the results of segmentation analyses using the Papaya.js volume viewer (https://github.com/rii-mango/Papaya). Users are able to annotate points and curves on the volume, edit voxels, and assign tasks to other users (e.g., to manually correct the segmentation of a particular image). The platform is pipeline agnostic, meaning that it can be configured to quality control any set of 3D volumes regardless of what neuroimaging software package produced it. In the following sections, we describe the implementation details of Mindcontrol, as well as its configuration for three open-source datasets, each with a different type of neuroimaging pipeline output.

5.2 Software Design and Implementation

5.2.1 Design Principles

Mindcontrol was developed with several design requirements. Mindcontrol must be easily accessible from any device, such as a Mac, Windows or even a tablet. Therefore, the best option was to develop a web application. Most tablets have limited storage capacity, so space-minimizing specifications were established. A dependence on cloud-based data storage was specified to accommodate large neuroimaging datasets without needing local storage. To efficiently store annotations and edited voxels, Mindcontrol only stores the changes to files, rather than whole-file information, on its database. Researchers must be able to QC outputs from any type of neuroimaging software package, so Mindcontrol was specified to flexibly accommodate any file organization structure, with configurable "modules" that can contain any type of descriptive statistics and 3D images. Mindcontrol configuration and database updates must require minimal Javascript knowledge, since Matlab/Octave, Python, R, and C are primarily used in the neuroimaging community for data analysis. Finally, changes to the database(like the addition of new images), changes in descriptive measures, and new edits/annotations, should be reflected in the application in real-time to foster collaboration.

5.2.2 Server Back-End Framework

Mindcontrol is built with Meteor (http://www.meteor.com), a full-stack javascript web-development platform. Meteor features a build tool, a package manager, the convenience of a single language (javascript) to develop both the front- and back-end of the application, and an abstracted implementation of full-stack reactivity. Data is transferred "over the wire" and rendered by the client (as opposed to the server sending HTML), which means that changes to the database automatically trigger changes to the application view. For example, as soon as a user finishes implementing QC procedures on an image and clicks "save", all other users can see the changes. A diagram of this process is provided in Figure 5.1.

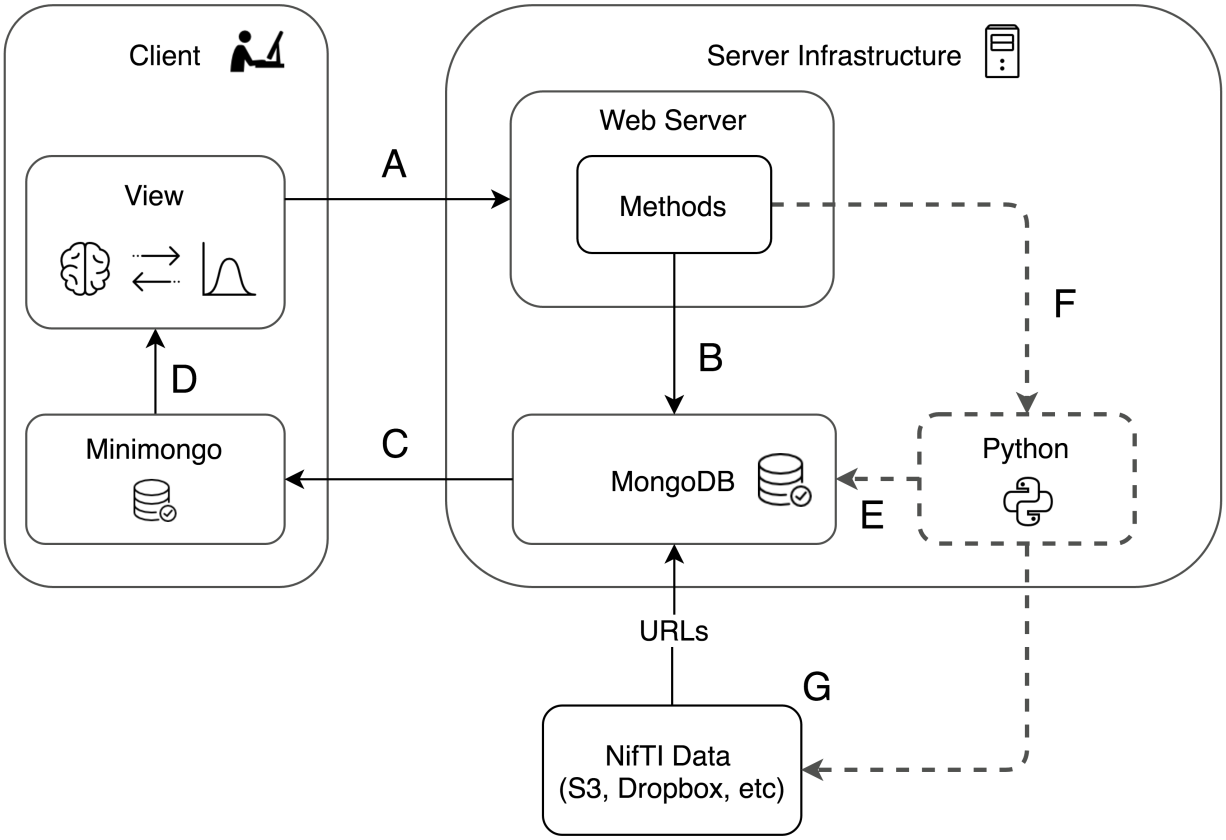


Figure 5.1: This diagram shows the different components of the Mindcontrol application. A) The client sends information, such as annotations and edits, to the server. B) The server calls a method that updates the mongoDB backend. C) When the back-end MongoDB database changes, these changes are automatically pushed to the minimongo database on the client. D) A change to the minimongo database automatically re-renders the view on the client. E) Users can optionally push changes to the client view via the MongoDB with Python MongoDB drivers. Drivers for C, C++, Scala, Java, PHP, Ruby, Perl, and Node.js are also available through MongoDB. F) Developers can optionally write server methods to launch Python or command-line processes that, in turn, use user annotations and edits to re-process images and update the MongoDB with new results. G) Imaging data (in NifTI format) is stored on an external server, such as Amazon S3 or Dropbox, and URLs to the images are stored in the MongoDB.

5.2.3 Client-Side Features

The user interface consists of a dashboard view and an imaging view, as shown in Figures 5.2 and 5.3, respectively. The primary dashboard view consists of processing module sections, a query controller, data tables, and descriptive statistic visualizations. Each entry in the table is a link that, when clicked, filters all tables on the page. The filters or queries can be saved, edited, and loaded in the query controller section, as shown in Figure 5.4.

Descriptive statistics are visualized using the D3 library (https://d3js.org/). Currently, two visualizations are provided: a calendar view of a heatmap that shows a histogram of the number of exams collected on a given day and 1D histograms of scalar metrics with dimensions that are swappable using a dropdown menu, as shown in Figure 5.2. Both histogram plots interactively filter the data tables below. Clicking on a particular date on the date-histogram plot filters all tables by the exams collected on that particular date. Users are able to "brush" sections of the 1D histogram to filter all tables with exams that meet requirements of values within that range (see Figure 5.5).

The imaging view is shown in Figure 5.3. The left-side column includes a section to label an image as "Pass", "Fail", "Edited", or "Needs Edits" and to provide notes. The status bar at the top-left portion updates instantaneously with information on which user checked the image, the quality status of the image, and when it was last checked. Users are also able to assign edits to be performed by other users on the system; for example, a research assistant can perform a general QC and assign difficult cases to a neuroradiologist. On the right-hand side, the Papaya.js viewer (http://rii-mango.github.io/Papaya/) is used to display the NifTI volumes of the original data and FreeSurfer segmentations.

Annotations of points and curves are shown in Figure 5.6. Using the *shift* key, users can click on the image to annotate points or select the "Logged Curves" toolbar. By *shift+click and dragging*, users can draw curves. Keyboard and mouse shortcuts provided by the Papaya.js viewer, along with Mindcontrol, include toggling overlays (*zz*) and undoing annotations (*dd*). Figure 5.7 shows the editing ("Painter") panel of the imaging view. Users set paintbrush values and *shift+click and drag* to change voxel values. For point and curve annotations and voxel editing, the images themselves are not changed, but world x,y,z coordinates, along with annotation text or paintbrush values, are saved to the mongo database when the user clicks “save”. Custom offline functions may be written to apply editing to images: for example, to implement pial surface edits from FreeSurfer.

5.2.4 Configuration Details

Mindcontrol can be configured for a study’s specific needs be specifying a configuration JSON file. The configuration file describes processing modules by module names, the columns to display in the data table below, and the type of graph to display (Date histogram or 1D histogram). Images must be hosted on a separate server or a content delivery network (CDN) and the Mindcontrol database populated with URLs to these images. An initial JSON file can be specified to populate the Mindcontrol MongoDB with entries on startup if the database is empty. Instructions and example JSON schema for the configuration file and the database entries can be found at `https://github.com/akeshavan/mindcontrol/wiki` along with a Python function to access the MongoDB, which can be used to write custom editing scripts and externally update the database.

5.3 Examples/Applications

Mindcontrol configurations were developed for selected data from the 1000 Functional Connectomes project (FCP), the consortium for reliability and reproducibility (CoRR), and the Autism Brain Imaging Data Exchange (ABIDE) Collection I.

The FCP consists of 1414 resting state fMRI and corresponding structural datasets collected from 35 sites around the world [125], which have been openly shared with the public. The purpose of the FCP collaboration is to comprehensively map the functional human connectome, to understand genetic influences on brain’s structure and function, and to understand how brain structure and function relate to human behavior [125]. Segmentation of 200 selected FCP anatomical images from Baltimore, Bangor, Berlin, ICBM, and Milwaukee was performed with Freesurfer (recon-all) version 5.3.0 [34] using the RedHat 7 operating system on IEEE 754 compliant hardware at UCSF. Regional volumes of subcortical and cerebellar regions were computed. Cortical volumes, areas, and thicknesses were also computed and averaged across hemispheres. Scan dates were simulated in order to demonstrate the date histogram shown in Figure 5.1B. The original anonymized T1-weighted images, along with the aparc+aseg output from Freesurfer, were converted to the compressed NifTI (.nii.gz) format and uploaded to Dropbox for the purpose of visualization within Mindcontrol. The Mindcontrol database was populated with URLs to these images, along with their corresponding FreeSurfer segmentation metrics. The demo of the FCP data is located at `http://mindcontrol.herokuapp.com`.

More recently, researchers have developed the Preprocessed-Connectomes Project's Quality Assurance Protocol (PCP-QAP) software, to provide anatomical and functional data quality measures in order to detect low-quality images before data processing and analysis [124]. Some metrics include contrast-to-noise ratio, signal-to-noise ratio, voxel smoothness, percentage of artifact voxels, foreground-to-background energy ratio, and entropy focus criterion [124]. The PCP-QAP protocol has been run on the Consortium for Reliability and Reproducibility (CoRR), and the Autism Brain Imaging Data Exchange (ABIDE) datasets and the results have been posted online.

The purpose of CoRR is to provide an open-science dataset to assess the reliability of functional and structural connectomics by defining test-retest reliability of commonly used MR metrics; to understand the variability of these metrics across sites; and to establish a standard benchmark dataset on which to evaluate new imaging metrics [126]. PCP-QAP normative data for the CoRR study was downloaded from https://github.com/preprocessed-connectomes-project/quality-assessment-protocol. The Mindcontrol database was populated with pointers to 2,963 CoRR structural images residing on an Amazon S3 bucket along with their corresponding PCP-QAP metrics. The demo of the CoRR dataset with PCP-QAP metrics is hosted at `http://mindcontrol-corr.herokuapp.com`.

The overarching goal of the ABIDE initiative is to expedite the discovery of the neural basis of autism by providing open access to a large, heterogeneous collection of structural and functional neuroimaging datasets collected from over 23 institutions [127]. The Preprocessed Connectomes Project provides cortical thickness measures of the ABIDE dataset output by the ANTs software package [128], along with summary statistics across regions of interests (ROIs) defined by the Desikan-Killiany-Tourville (DKT) protocol [129]. The Mindcontrol database was populated with pointer URLs to S3-hosted cortical thickness images and their corresponding ROI summary measures, along with PCP-QAP metrics. The demo of the ABIDE dataset with ANTS cortical thickness and PCP-QAP metrics is located at `http://mindcontrol-abide.herokuapp.com`.

5.4 Discussion

Mindcontrol is a configurable neuroinformatics dashboard that links study information and descriptive statistics with scientific data visualization, MR images, and their overlays (segmentation or otherwise). The three configurations demonstrated in this report show the link between MRI quality metrics and raw data, the link between Freesurfer regional volumes and segmentation quality, and the link between ANTS cortical thickness summary statistics and segmentation/thickness estimates of the volume. The platform is configurable, open-source, and software/pipeline agnostic, enabling researchers to configure it to their particular analyses. The dashboard allows researchers to assign editing tasks to others, who can then perform edits on the application itself.

The Mindcontrol platform streamlines and standardizes QC procedures. The traditional method of collaborative QC within a lab assembles disparate software components into a procedure that is vulnerable to clerical errors. The QC operators use existing viewers (such as FSLview or Freeview) to view and edit the images, making notes on a collaborative spreadsheet application (such as google-docs). They must carefully adhere to a common folder structure and naming convention so that other lab members and any automated processing scripts can locate these images. Distributions of scalar metrics are then plotted to identify outliers using a separate data analysis software program. The results of that analysis must then be reviewed using the imaging software to ensure that outliers are appropriately screened. This method is inherently inefficient because images must be loaded multiple times and attention split between the imaging platform and annotation software. Additionally, results of the QC must be maintained consistently across several software packages. Clerical errors are common and time-consuming to resolve because naming convention is not explicitly enforced, and manual edits could be lost within the filesystem without thorough documentation by research assistants. Google-spreadsheets are collaborative, but ideally this pass/fail/edited QC information would be directly linked to the data. Mindcontrol stores all notes, annotations, and QC results, and in-browser edits internally (Mongo database backend). User edits can be extracted automatically and written to the filesystem, eliminating the potential for clerical errors. An example python script to do this can be found at https://github.com/akeshavan/mindcontrol/wiki/Applying-Painter-Edits. Scalar metrics are linked to 3D images, enabling a user to inspect an outlier image with the click of a button. Mind-control is web-based, so it can be used on any device; QC operators can even use a tablet with stylus to edit, which is more natural than using a mouse.

There have been considerable efforts in this field to ensure data quality on a large scale. The human connectome project's extensive informatics pipeline, which includes a database service, QC procedures, and a data visualization platform, has been key to the project's success in collecting a large, high-quality dataset [130]. The Allen Brain Atlas offers a comprehensive genetic, neuroanatomical, and connectivity web-based data exploration portal, linking an MRI viewer to data tables [131]. The open-source LORIS web-based data management system integrates an image viewer with extensive QC modules [132]. Mindcontrol supplements these efforts by providing a lightweight and extensible data management and visualization system with the added capability to perform edits and curate annotations within the application.

Table 5.4 shows examples of subjects from the FCP, CoRR and ABIDE datasets with low-quality scans or segmentations, identified using Mindcontrol. The tails of various PCP-QAP metric distributions for both the ABIDE and CoRR datasets could be filtered interactively to isolate images with motion artifacts, extensive blurring, and noise. In the ABIDE dataset, filtering by the entropy focus criterion (EFC) greater than 17 identified images with extreme motion artifacts. The range of the EFC for the CoRR dataset was much smaller (less than 2) and the image with the highest EFC had no motion artifacts, but failed QC due to excessive defacing. In the ABIDE dataset, filtering for the high FWHM extremes identified images with motion artifacts, grainy/noisy images, and one extremely blurry image. On the other hand, in the CoRR dataset, the image with high FWHM had an extreme bias field. In the CoRR dataset, the images with very low contrast-to-noise (CNR) had motion artifacts, while the ABIDE images did not. Overall, examining the extremes of the PCP-QAP metrics with Mindcontrol identified outliers, but the relationship between artifacts and metrics varied by study.

Exploring the link between ANTS Cortical Thickness and the PCP-QAP metrics in the ABIDE dataset, we found that selecting the higher tail of average left- and right-postcentral gyrus thickness corresponded to datasets at the higher range of PCP-QAP FWHM. Conversely, selecting the lower tail of the precentral and postcentral gyrus thicknesses related to the lower range of the FWHM. It was difficult to pinpoint errors in the ANTS Cortical Thickness segmentation images because the data was normalized to MNI space. In the future, it would be better to QC each step of the ANTS pipeline to ensure that 1) segmentation in native space was accurate and 2) normalization to MNI space was reasonable.

Mindcontrol is particularly useful to investigate where errors occurred in segmentation algorithms. In the FCP dataset, the most common errors in segmentation with Freesurfer were that 1) parts of the temporal lobes were excluded from the segmentation and 2) the gray matter segmentation entered the dura. Scans with low-quality temporal lobe segmentation were found by selecting the lower tail of the amygdala or temporal pole volume distributions. Often, these images exhibited poor gray/white contrast in the temporal pole region, and low signal to noise. Initially, we observed that dura missclassification occurred most frequently in the precentral and postcentral gyri. We then selected data points with high precentral/postcentral volumes to locate these errors. However, scans in the middle of these metric distributions also exhibited dura misclassification, suggesting that this particular problem may be consistent across the entire dataset. In this example, it is necessary to QC every scan, regardless of where its summary metrics lie on the distribution.

In cases where every scan must be quality controlled, Mindcontrol's summary statistic distributions and annotation features serve as a triage tool, sorting cases that are likely to require more time or expertise. When training new editors, Mindcontrol's annotation and notes features enable users to ask questions, mark the image with the point or curve annotation feature, and assign images to more senior editors to review and provide feedback. An initial Mindcontrol quality check can be used to estimate total editing time and expertise needed for a study, enabling a strategic allocation of resources. Leveraging Mindcontrol as an integrated quality control tool can make processing methods more efficient, organized, and collaborative.

5.5 Future Directions

Mindcontrol is being actively developed to incorporate new features that will improve outlier detection, efficiency, and collaboration. New information visualizations to detect outliers include: scatter plots to compare two metrics against each other, and a longitudinal view of a single-subject trajectory for a given metric to detect uncharacteristic temporal changes. New scientific data visualizations are planned using the BrainBrowser library [133] to display cortical surfaces. A beta version of real-time collaborative annotations, where two users can annotate the same image and see the edits of the other user as they occur, is in the testing phase.

Currently, configuring Mindcontrol involves creating one JSON file to describe the different modules and another JSON file to populate the Mongo database with pointers to images and their scalar metrics. In the future, this process could be streamlined by creating a Mindcontrol configuration for datasets with a standardized folder structure, like the Brain Imaging Data Structure (BIDS) [3], and their BIDS-compliant derivatives [110]. Additionally, implementing the server-side application within a container, like Docker, will make it easier to deploy a Mindcontrol server. Further

Dataset	**Filter**	**Algorithm**	**Example Subject IDs**	**Observation**
ABIDE	Higher end of FWHM	PCP-QAP	50528, 0050511, 0050519	Motion artifact
ABIDE	Higher end of FWHM	PCP-QAP	50496	Very grainy
ABIDE	Higher end of FWHM	PCP-QAP	50611	Extremely blurry
ABIDE	High EFC	PCP-QAP	51160, 0051191, 0051166, 0051174, 0051192, 0051165, 0051186	Motion artifact
ABIDE	High QI1	PCP-QAP	50197, 50017	Very noisy, motion artifact
CoRR	Lower end of CNR	PCP-QAP	25073, 25085,	Motion artifact
CoRR	High EFC (range much lower than ABIDE)	PCP-QAP	25567	No motion artifact, but frontal lobe cut off (excessive defacing)
CoRR	High FWHM	PCP-QAP	27040	Needs major bias field correction
FCP	Lower end of Left-Amygdala, temporal-pole, Left-Amygdala distribution	Freesurfer	sub48830, sub93262, sub55176, sub75919	Temporal lobes not correctly segmented; gray white delineation difficult to see
FCP	Higher end of SuperiorFrontal, Precentral, Postcentral thickness	Freesurfer	sub98317, sub27536, sub28795, sub10582, sub93975	Gray matter segmentation enters dura

Table 5.1: A table of bad quality data or segmentations found on Mindcontrol

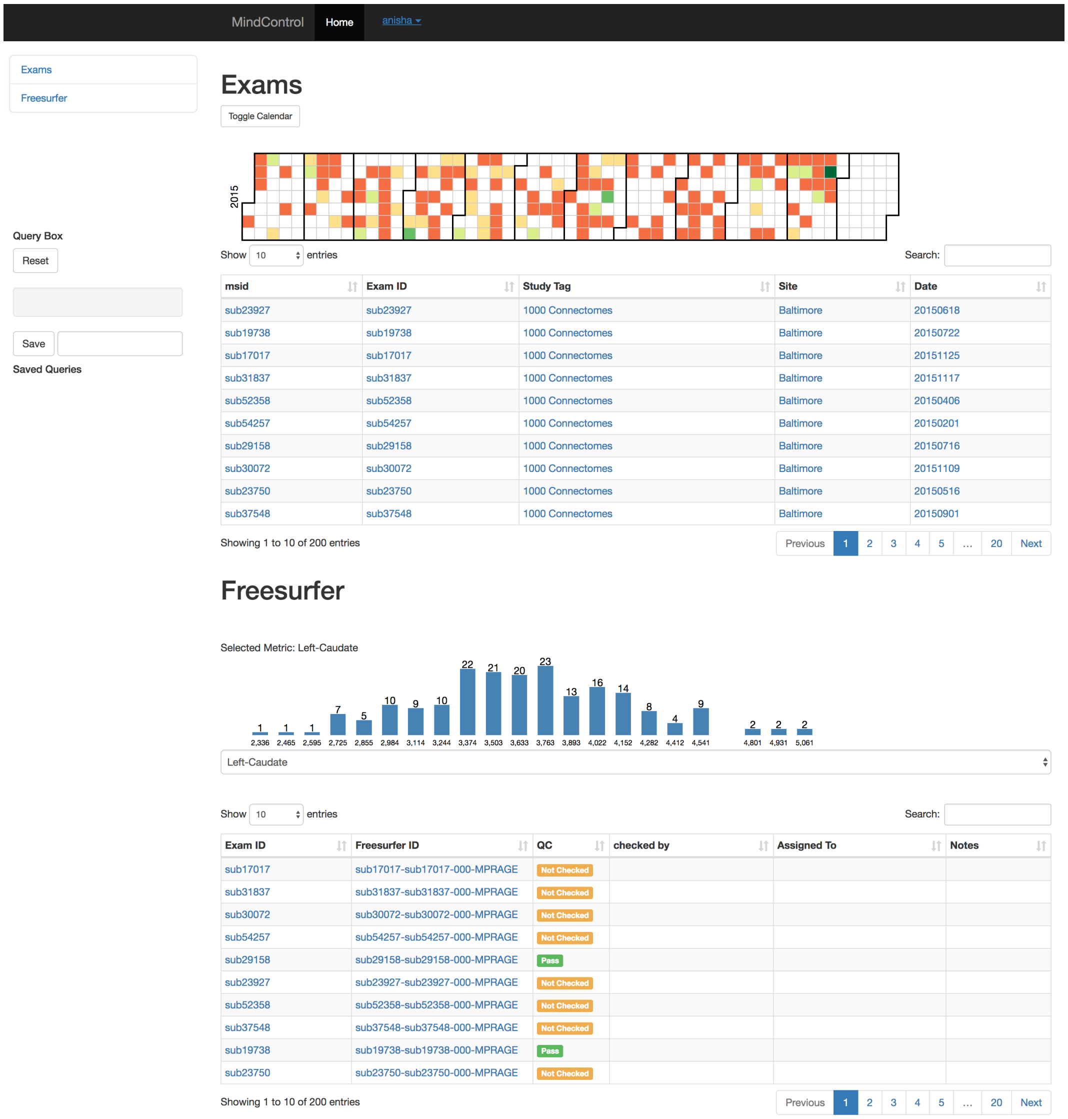
development of Mindcontrol will include the flexible importing of additional scalar metrics, such as measures of structural complexity, calculated by third-party toolboxes developed to complement standard analysis pipelines [134, 135]. This will enable researchers to collaborate on the same dataset by uploading metrics from their newly developed algorithms, and will enable them to easily explore their results in the context of metrics contributed by others. Finally, Mindcontrol has the potential to be a large-scale crowd-sourcing platform for segmentation editing and quality control. We hope the functionality, ease-of-use, and modularity offered by Mindcontrol will help to improve the standards used by studies relying on brain segmentation.

5.6 Software Availability

The Mindcontrol codebase is hosted on GitHub at `http://github.com/akeshavan/mindcontrol`, along with installation instructions. The Mindcontrol configuration of the FCP data is located on the master branch of the GitHub repository, and the configurations for CoRR and ABIDE are located at `http://github.com/akeshavan/mindcontrol_configs`, along with configuration documentation. Mindcontrol is licensed under the Apache License, Version 2.0.

5.7 Acknowledgements

AK would like to acknowledge Satrajit Ghosh and Arno Klein for their mentorship, the Papaya.js developers for their support, and Cameron Craddock for his assistance with the CoRR and ABIDE datasets. AK, ED, IM, KJ, and CM would like to thank the organizers of Neurohackweek and the Brainhack events for fostering a collaborative environment in which to develop new ideas and software. CM was supported by a fellowship from the Canadian Institutes of Health Research (FRN-146793). ED and KJ were supported by the National Defense Science and Engineering Graduate Fellowship. Icons from Figure 5.1 were sourced from https://icons8.com.



Exams

Show 10 entries Search:

msid	Exam ID	Study Tag	Site	Date
sub23927	sub23927	1000 Connectomes	Baltimore	20150618
sub19738	sub19738	1000 Connectomes	Baltimore	20150722
sub17017	sub17017	1000 Connectomes	Baltimore	20151125
sub31837	sub31837	1000 Connectomes	Baltimore	20151117
sub52358	sub52358	1000 Connectomes	Baltimore	20150406
sub54257	sub54257	1000 Connectomes	Baltimore	20150201
sub29158	sub29158	1000 Connectomes	Baltimore	20150716
sub30072	sub30072	1000 Connectomes	Baltimore	20151109
sub23750	sub23750	1000 Connectomes	Baltimore	20150516
sub37548	sub37548	1000 Connectomes	Baltimore	20150901

Showing 1 to 10 of 200 entries

Previous 1 2 3 4 5 … 20 Next

Freesurfer

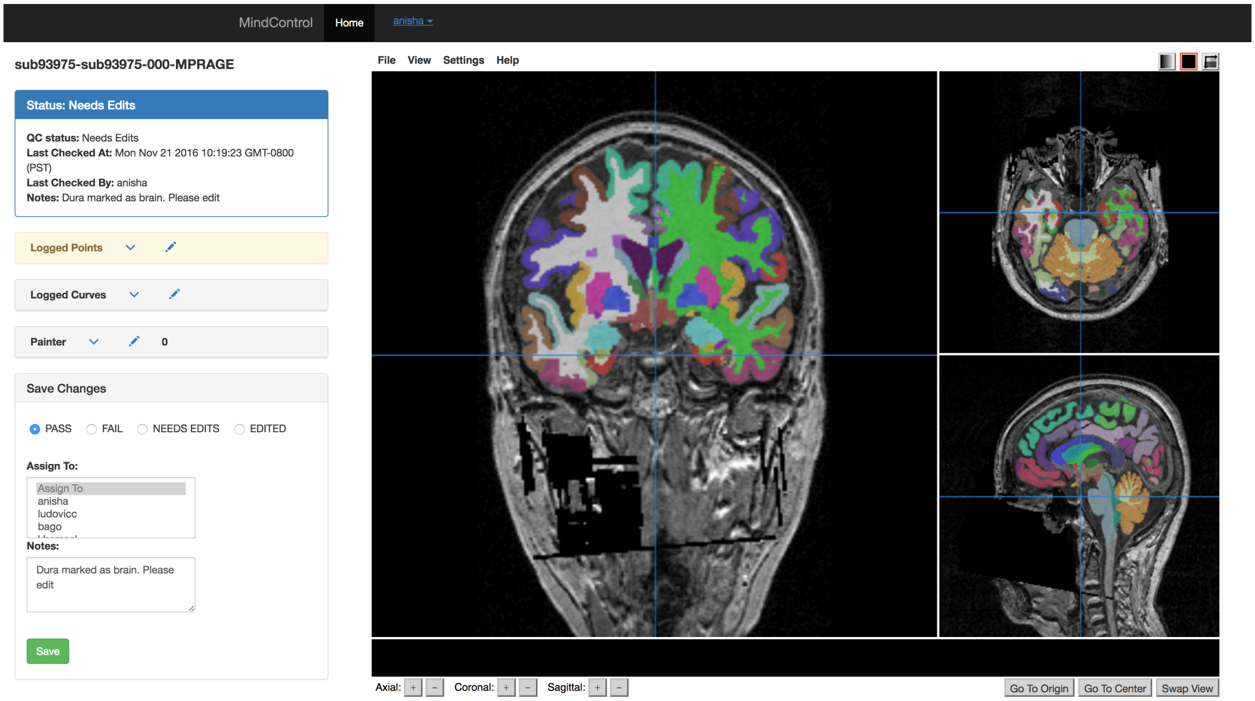
Show 10 entries Search:

Exam ID	Freesurfer ID	QC	checked by	Assigned To	Notes
sub17017	sub17017-sub17017-000-MPRAGE	Not Checked			
sub31837	sub31837-sub31837-000-MPRAGE	Not Checked			
sub30072	sub30072-sub30072-000-MPRAGE	Not Checked			
sub54257	sub54257-sub54257-000-MPRAGE	Not Checked			
sub29158	sub29158-sub29158-000-MPRAGE	Pass			
sub23927	sub23927-sub23927-000-MPRAGE	Not Checked			
sub52358	sub52358-sub52358-000-MPRAGE	Not Checked			
sub37548	sub37548-sub37548-000-MPRAGE	Not Checked			
sub19738	sub19738-sub19738-000-MPRAGE	Pass			
sub23750	sub23750-sub23750-000-MPRAGE	Not Checked			

Showing 1 to 10 of 200 entries

Previous 1 2 3 4 5 … 20 Next

Figure 5.2: This figure shows the Mindcontrol layout configured to quality check Freesurfer outputs from the 1000 Functional Connectomes Project (FCP). Part A shows the module navigator, which links to the different processing modules on the dashboard. Part B shows the different exams and the dates they were acquired as a heatmap, where green is more and orange is less scans collected on a given day. (For demonstration purposes, the dates depicted here do not reflect the actual dates the data were collected for the FCP, since this information was not provided at the time.) Clicking on data in any column of the exam table filters the data by that column. For example, clicking the site "Milwaukee" reduces both the "Exams" and the "FreeSurfer" tables to only show subjects from Milwaukee. Part C shows the Freesurfer table and regional volume distribution of the left caudate. A drop-down menu allows users to switch the descriptive metric. Clicking on a value in FreeSurfer ID column brings the user to the imaging view, as shown in Figure 5.3, where users can evaluate and annotate the quality status of the image. The value of the label in the "QC" column changes instantaneously due to Meteor's built in full-stack reactivity.



sub93975-sub93975-000-MPRAGE

Status: Needs Edits

QC status: Needs Edits
Last Checked At: Mon Nov 21 2016 10:19:23 GMT-0800 (PST)
Last Checked By: anisha
Notes: Dura marked as brain. Please edit

Painter 0

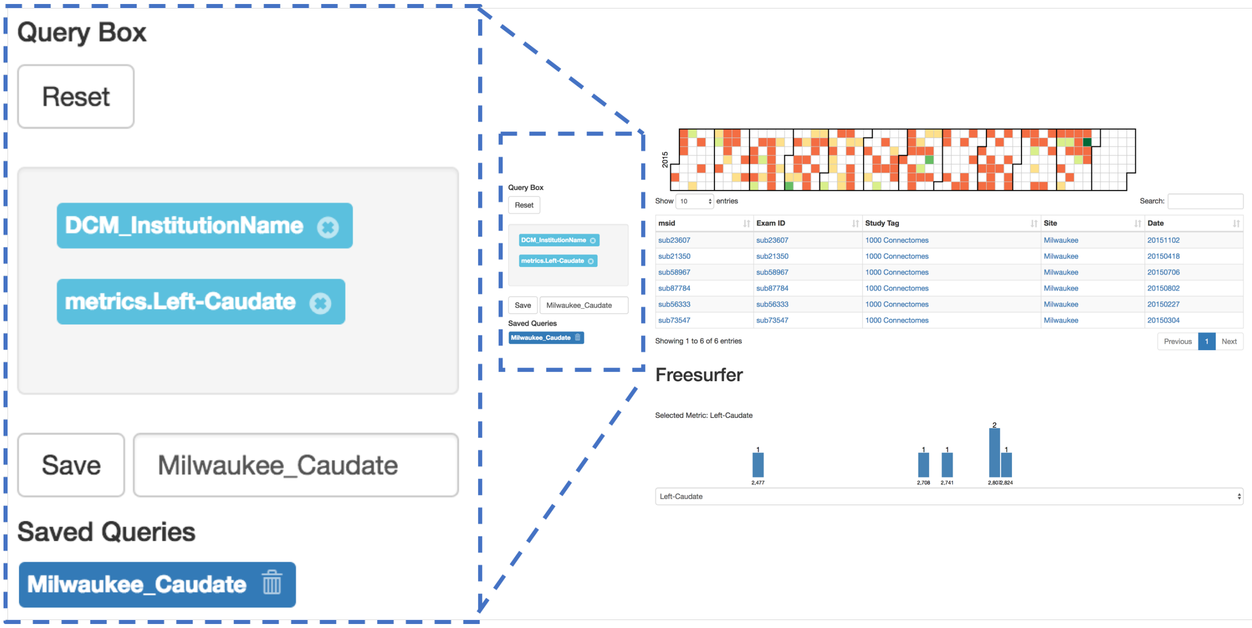
Save Changes

● PASS ○ FAIL ○ NEEDS EDITS ○ EDITED

Assign To:
Assign To
anisha
ludovicc
bago

Notes:
Dura marked as brain. Please edit

Figure 5.3: The imaging view of mindcontrol consists of a panel on the left-hand side that contains the QC status; a point annotation menu; a curve annotation menu; a voxel editing menu; and an editable sub-panel for QC status, notes, and editor assignment. On the right-hand side, the base MRI anatomical MPRAGE image is displayed with an overlay of the Freesurfer segmentation outputs using the Papaya.js viewer.

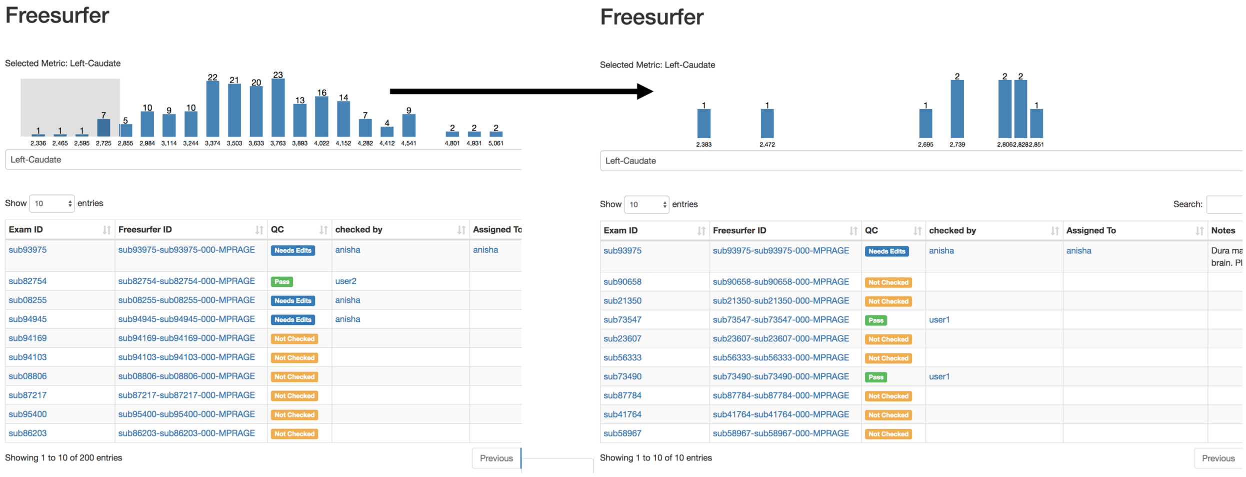


msid	Exam ID	Study Tag	Site	Date
sub23607	sub23607	1000 Connectomes	Milwaukee	20151102
sub21350	sub21350	1000 Connectomes	Milwaukee	20150418
sub58967	sub58967	1000 Connectomes	Milwaukee	20150706
sub87784	sub87784	1000 Connectomes	Milwaukee	20150802
sub56333	sub56333	1000 Connectomes	Milwaukee	20150227
sub73547	sub73547	1000 Connectomes	Milwaukee	20150304

Showing 1 to 6 of 6 entries

Freesurfer

Figure 5.4: The query controller shows the different filters that have been applied to this dataset. In this example, the exams have been filtered by institution (“Milwaukee”) and by a range of left caudate volumes (brushed from the histogram). Clicking the “x” next to the filter removes it, and the view updates. Queries can be saved and reloaded by providing a name in the text-entry box and clicking "Save". “Reset” removes all filters to show the whole dataset.



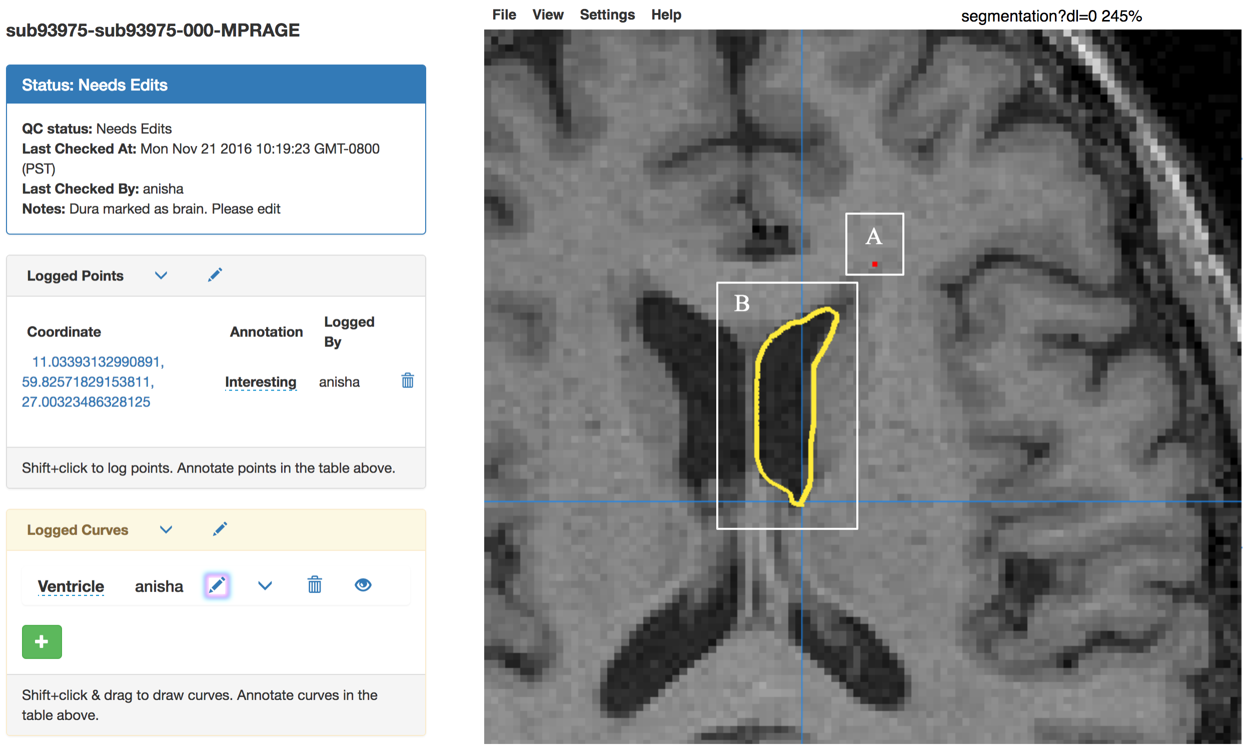
Exam ID	Freesurfer ID	QC	checked by	Assigned To
sub93975	sub93975-sub93975-000-MPRAGE	Needs Edits	anisha	anisha
sub82754	sub82754-sub82754-000-MPRAGE	Pass	user2	
sub08255	sub08255-sub08255-000-MPRAGE	Needs Edits	anisha	
sub94945	sub94945-sub94945-000-MPRAGE	Needs Edits	anisha	
sub94169	sub94169-sub94169-000-MPRAGE	Not Checked		
sub94103	sub94103-sub94103-000-MPRAGE	Not Checked		
sub08806	sub08806-sub08806-000-MPRAGE	Not Checked		
sub87217	sub87217-sub87217-000-MPRAGE	Not Checked		
sub95400	sub95400-sub95400-000-MPRAGE	Not Checked		
sub86203	sub86203-sub86203-000-MPRAGE	Not Checked		

Showing 1 to 10 of 200 entries

Exam ID	Freesurfer ID	QC	checked by	Assigned To	Notes
sub93975	sub93975-sub93975-000-MPRAGE	Needs Edits	anisha	anisha	Dura ma brain. Pl
sub90658	sub90658-sub90658-000-MPRAGE	Not Checked			
sub21350	sub21350-sub21350-000-MPRAGE	Not Checked			
sub73547	sub73547-sub73547-000-MPRAGE	Pass	user1		
sub23607	sub23607-sub23607-000-MPRAGE	Not Checked			
sub56333	sub56333-sub56333-000-MPRAGE	Not Checked			
sub73490	sub73490-sub73490-000-MPRAGE	Pass	user1		
sub87784	sub87784-sub87784-000-MPRAGE	Not Checked			
sub41764	sub41764-sub41764-000-MPRAGE	Not Checked			
sub58967	sub58967-sub58967-000-MPRAGE	Not Checked			

Showing 1 to 10 of 10 entries

Figure 5.5: This demonstrates the interactive brushing feature of Mindcontrol histograms. On the left, the user has brushed the tail end of the left caudate volume distribution from Freesurfer. On the right, the histogram has been redrawn with data from the brushed range, and the table beneath filtered from 200 entries to 14 entries based on the brushed caudate volumes.



sub93975-sub93975-000-MPRAGE

Status: Needs Edits

QC status: Needs Edits
Last Checked At: Mon Nov 21 2016 10:19:23 GMT-0800 (PST)
Last Checked By: anisha
Notes: Dura marked as brain. Please edit

Logged Points

Coordinate	Annotation	Logged By
11.03393132990891, 59.82571829153811, 27.00323486328125	Interesting	anisha

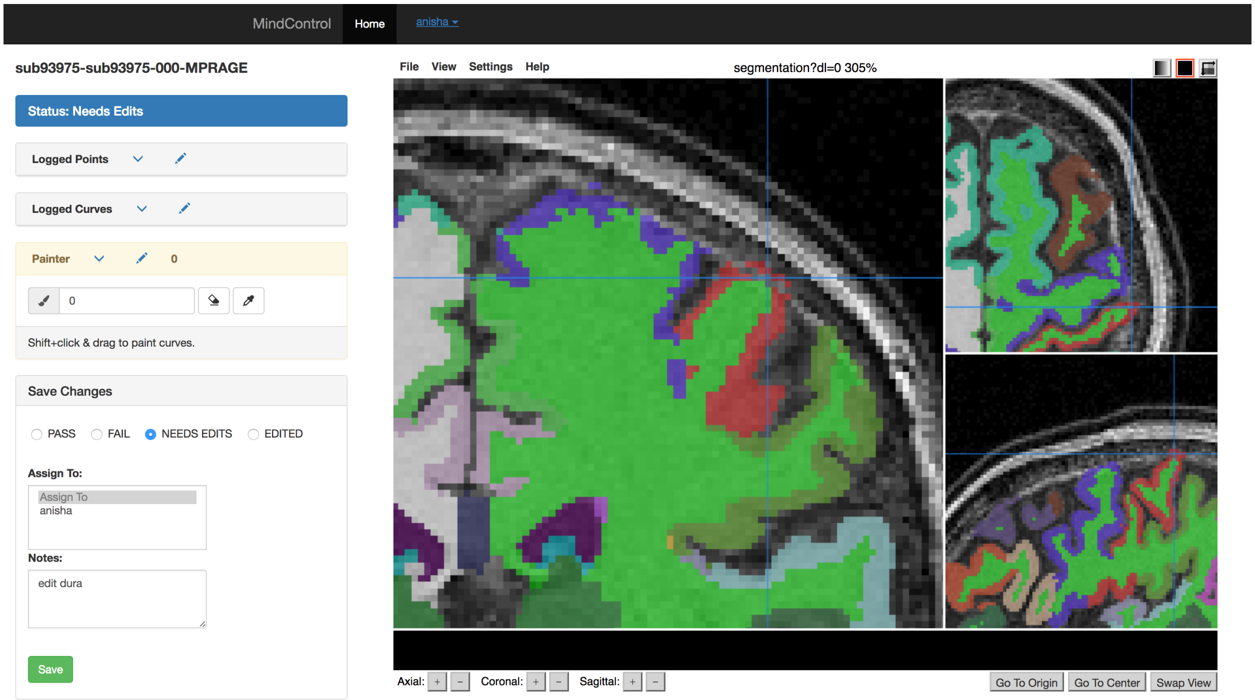
Shift+click to log points. Annotate points in the table above.

Logged Curves

Ventricle	anisha

Shift+click & drag to draw curves. Annotate curves in the table above.

Figure 5.6: The annotations panel can be used to annotate a single point (shown in red, part A) and curves (shown in B). When annotating points, the user is shown the selected x,y,z world coordinates and is able to name the annotation. In the curve annotation panel on the left sidebar, the user is able to name the curve and add/remove curves. Keyboard shortcuts: “dd” removes the previous annotation and “zz” toggles the segmentation overlay.



sub93975-sub93975-000-MPRAGE

Status: Needs Edits

0

Shift+click & drag to paint curves.

Save Changes

○ PASS ○ FAIL ● NEEDS EDITS ○ EDITED

Assign To:

Assign To
anisha

Notes:

edit dura

segmentation?dl=0 305%

Figure 5.7: The editing panel on the left shows the "Painter" toolbox in yellow, where users can input brush values or use the eyedropper tool to set the value to that of a clicked label. The eraser icon sets the brush value to 0, to delete or erase voxels. In the image above, the Freesurfer segmentation is being edited by erasing the voxels missclassified as dura.

6 Interactive Brain Data Visualization Platforms

6.1 Background

One main goal in the field of neuroimaging is to quantify information-rich 3D/4D images to informative scalar metrics that can accurately capture the variance associated with normal brain development or pathology. One example involves segmenting the tissues of anatomical MR images to measure the gray matter volume of an individual; abnormalities in this measurement can be related to a number of pathologies, including multiple sclerosis. A vital, and often overlooked step that must occur before this dimensionality reduction process is the visualization of the various features extracted from these images. When we quantify images as scalar metrics, we lose important spatial information. For example, in MS, the pattern of lesion distribution is unique, compared to other diseases with seemingly similar imaging phenotypes, such as vasculitis or neuromyelitis optica; the lesion volume scalar metric does not capture this important spatial lesion distribution feature. Ideally, there would be a more intuitive link between the visualization of scalar summary metrics, called information visualization, and MRI images, or more generally, "scientific data visualization" [136]. Such a visualization would be able to condense information-rich imaging biomarkers while simultaneously maintaining intuitive spatial relationships within the data.

Many software packages for scientific data visualization of MR images have been built; these include Freeview, part of the Freesurfer package [31], FSLVIEW[137], part of the FSL package [138], Trackvis, part of the Diffusion Toolkit[139]. Independent standalone packages include MRICron[140], 3D slicer[141], DataViewer3D [142] and the web-based BrainBrowser [133]. While these viewers have many useful interactive features, none are configured to integrate scalar information simultaneously with 3D images. In this chapter, I present two projects that address this gap: the ROYGBIV open-source brain viewer, and the MindMeld interactive dashboard.

6.2 Interactive online brain shape visualization (ROYGBIV)

The following is a reprint of [143], which was peer-reviewed and published in the Research Ideas and Outcomes journal. ROYGBIV visualizes outputs from the Mindboggle software package [144]. This work was begun at the 2015 Organization for Human Brain Mapping (OHBM) conference with collaborators Dr. Arno Klein and Dr. Ben Cipollini.

6.2.1 Introduction

Our goal for the hackathon was to create an interactive Web browser application to visualize human brain image data processed by the Mindboggle software package [145].

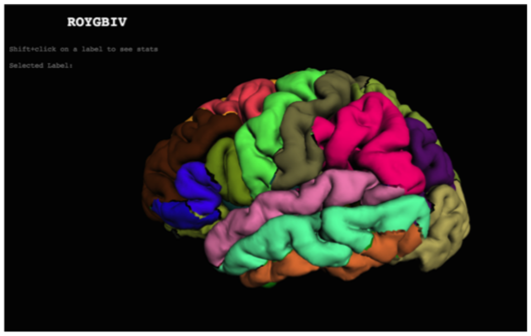
The Mindboggle project was initiated to improve the labeling as well as morphometry of brain imaging data, and to promote open science by making all data, software, and documentation freely and openly available. An interface for interactive visualization is essential for assessing issues in brain image processing and analysis, including surface reconstruction, labeling, and morphometry. Mindboggle processes human brain cortical surface meshes in the VTK format, and generates label and shape information for each anatomical region, where labels follow the Desikan-Killiany-Tourville protocol [146].

6.2.2 Approach

Over the course of two afternoons at the Human Brain Mapping 2015 conference's hackathon, we evaluated several JavaScript libraries for creating browser-based WebGL visualizations of brain surfaces, including three.js, XTK, and BrainBrowser. Three.js was chosen for ease of use and degree of active development and community support. To accompany these surface visualizations with graphical plots, we chose the d3 JavaScript library for its flexibility and widespread use.

6.2.3 Results

We completed an initial version of our browser-based interactive visualization tool; a left hemisphere of a human brain is available at `http://roygbiv.mindboggle.info`. Click and drag to rotate this brain, scroll to zoom in and out, and click on any region of the brain while pressing the shift key to produce an accompanying plot of shape measures for that region (fig. 6.2.3). This will render all other regions transparent. Figure 6.2.3 shows the distributions of travel depth, geodesic depth, mean curvature, freesurfer curvature, and freesurfer cortical thickness for the selected region. Shift-click outside the brain to return opacity to all regions.



ROYGBIV

Shift+click on a label to see stats

Selected Label:

Figure 6.1: Example visualization.

After the hackathon, we refactored the code to use an object-based approach. This allows multiple brains to be shown simultaneously. This approach was used to create a master-slave interaction: selection of a ROI in one hemisphere loads data for display on a second hemisphere. This approach

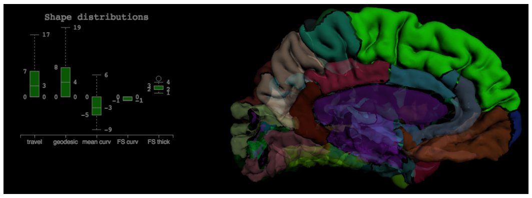


Figure 6.2: Example of a selected region and the accompanying boxplot.

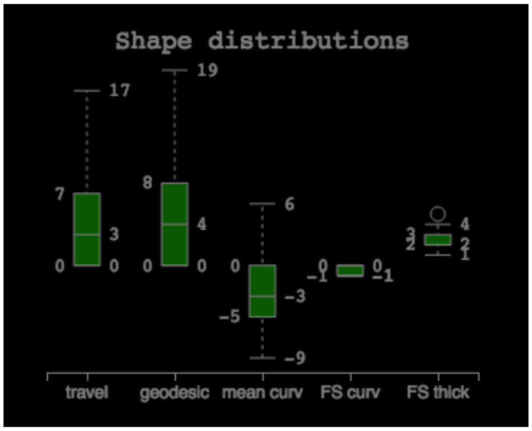


Figure 6.3: Example boxplot of a selected region that shows the distributions of shape features.

was used in a dynamic poster presented at Society for Neuroscience in 2015 [147].

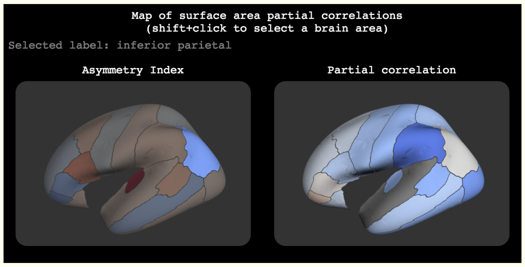

Map of surface area partial correlations
(shift+click to select a brain area)
Selected label: inferior parietal
Asymmetry Index
Partial correlation

Figure 6.4: Example master/slave visualization.

6.2.4 Conclusions

We have received very positive feedback for our efforts at the hackathon, and have since received several requests and encouragement to build this visualization out to accommodate other data besides shape information and to enable the visual evaluation of thousands of brains. We hope to continue this work with the help of others! To contribute to this project, please send pull requests to https://github.com/akeshavan/roygbiv.

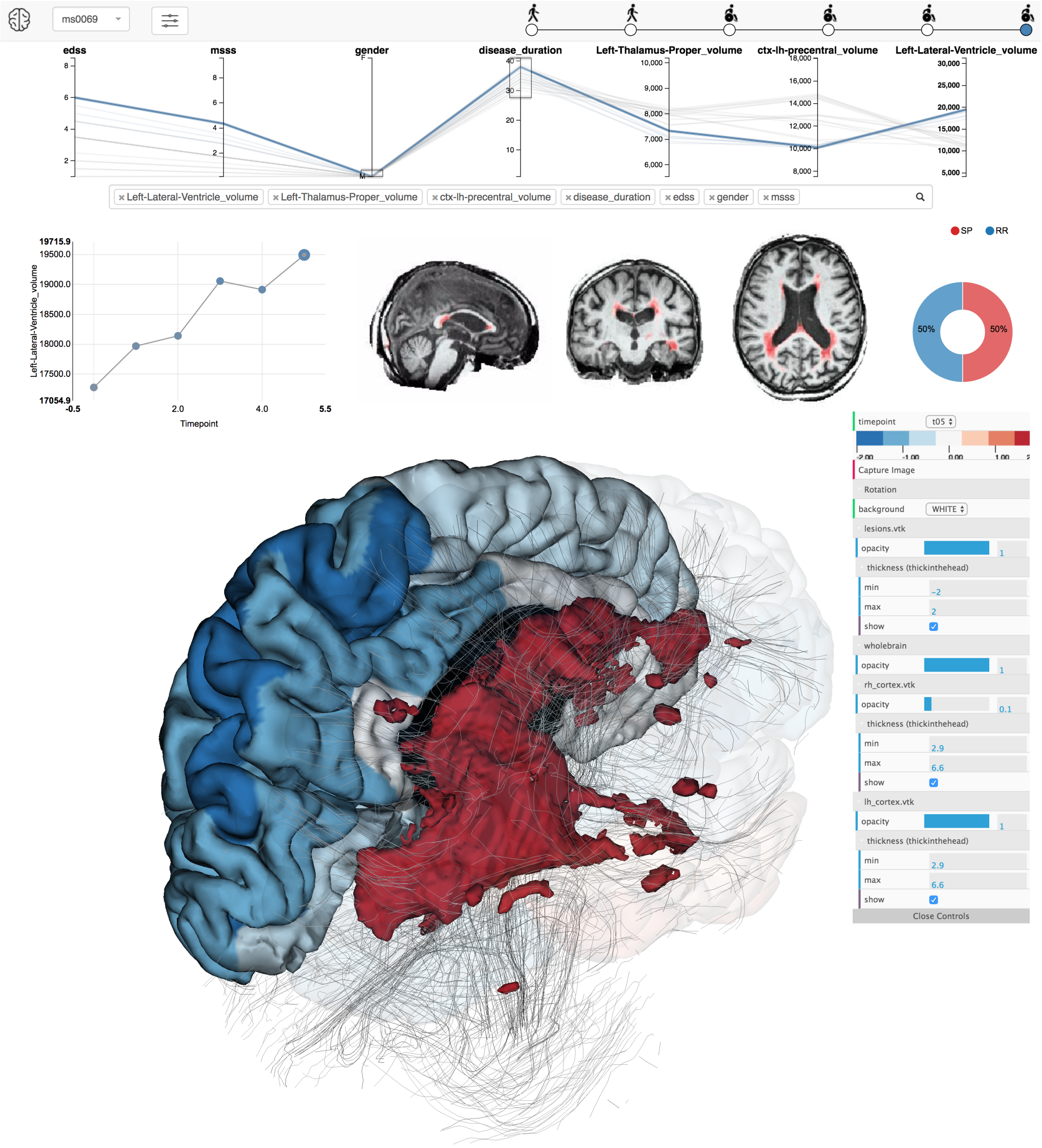
6.3 MindMeld

The goal of the MindMeld interactive visualization was to integrate longitudinal scalar metrics of MS patients with an interactive 3D brain visualization of cortical thickness, white matter tracts, and MS lesions. The key feature of this application is contextualization, which means that any metric, clinical or imaging based, should be displayed relative to a user-defined MS subpopulation. For example, the cortical thickness of a female MS patient relative to all female patients within the cohort, or the EDSS of a 50 year old male MS patient relative to all male MS patients between the ages of 55 and 65.

6.3.1 Design Requirements and Implementation

The application consists of 5 elements; 1) Contextualization filter to select a sub-population, 2) Longitudinal line plot to view any metric over time of a given patient, 3) Pie chart, to show the ratio of relapse-remitting to secondary progressive disease subtypes of the sub-population 4) Traditional slice-by-slice view of the MRI scan, and 5) 3D visualization of metrics extracted from the MPRAGE, FLAIR, and DWI scans. The main design requirement is that all five components must fit on the same page.

The main controller of the app is the contextualization filter, where the user defines ranges of values to select a reference sub-population that resembles the patient. Traditionally, this could be accomplished with text-input boxes; however, these components take up valuable screen space. Instead, a parallel coordinates plot [148] can be used to visualize a multidimensional distribution of metrics, with an added interactive filtering component on each axis. The parallel coordinates plot library was imported from https://github.com/syntagmatic/parallel-coordinates, and modified to enable brushing along a highlighted line and to add/remove data dimensions with a



timepoint t05

-2.00 -1.00 0.00 1.00 2

Capture Image

Rotation

background WHITE

lesions.vtk

opacity 1

thickness (thickinthehead)

min −2

max 2

show ☑

wholebrain

opacity 1

rh_cortex.vtk

opacity 0.1

thickness (thickinthehead)

min 2.9

max 6.6

show ☑

lh_cortex.vtk

opacity 1

thickness (thickinthehead)

min 2.9

max 6.6

show ☑

Close Controls

Figure 6.5: The MindMeld dashboard consists of the following components. Top Row: the parallel-coordinates contextualization filter, which can be brushed on each axis to filter the data. Second row, left: the longitudinal line plot, with markers colored by z-score compared to the reference population. Second row, center: the interactive volume view with lesions highlighted in red. Second row, right: a pie chart showing the ratio of subjects that converted to secondary progressive versus those who stayed in the relapse-remitting disease course. Bottom: the interactive 3D brain visualization with the cortical surface, with cortical thickness encoded in color by anatomical label, the grey streamlines, and FLAIR-derived lesions in red. A configuration side panel to the right enables users to change the opacity of different structures, apply an animation, and take a screenshot.

multiselect search bar (https://github.com/select2/select2). The style template used was Twitter Bootstrap (http://getbootstrap.com/) for a minimalist, responsive design.

The longitudinal line plot must display the evolution of any metric over time for a given patient. The context of each point in the line plot can be encoded by color, and linked to the contextualization filter by clicking an axis. The pie chart is needed to show how many patients in the sub-population transitioned to secondary progressive disease. This chart must also be linked to the contextualization filter. The two components should not take up too much screen space, since one does not need to closely inspect either chart to understand the overall trend. The line plot and pie chart were imported from the NVD3 library http://nvd3.org/, and the line plot was modified to encode contextualization in marker color using the d3 library (http://d3js.org). The line plot's y axis updates when an axis of the parallel coordinates chart is clicked, and the pie chart and line contextualization update after brushing of the parallel coordinates chart.

The standard volumetric MRI view is useful because it shows the raw intensity values of the gray matter, the white matter with its diffuse and focal MS lesions, and the ventricles, which are larger in brains with more atrophy. This view must be interactive, meaning that users should be able to click on any x,y,z coordinate and see the corresponding x,y, and z slices. The volumetric view does not need to be large, because close inspection of each slice would take too much time; instead, the 3D surface visualization should show the features extracted from the volumetric images. The volumetric MRI view was implemented using brainsprite.js (https://github.com/SIMEXP/brainsprite.js), which loads a standard image file (.jpg, .png) as a mosaic of brain slices, and reslices according to a user's mouse click.

The 3D surface visualization is capable of showing multiple modalities in one image, and therefore should take the most screen space. The full distribution of lesions, cortical folding patterns, cortical thickness by anatomical region, and fiber tracts must be shown simultaneously. The Brainbrowser library [133] was used to accomplish this. Custom additions to the Brainbrowser library were implemented such that the interaction of the sub-population selected by the parallel coordinates contextualizes the cortical thickness values on the surface.

6.3.2 Data Pipeline

A PBrain pipeline interface was built to calculate cortical shape metrics and create the appropriate data files for display within the browser. The PBrain interface is shown in figure 6.3.2. The PBrain interfaces run lesion segmentation, structural segmentation, and fiber tracking on FLAIR, MPRAGE, and high angular resolution diffusion imaging (HARDI), respectively. First, rigid body transformations were calculated from the FLAIR, HARDI, and T1+Gadolinium images across all timepoints to the anatomical MPRAGE image of the first timepoint using the FLIRT program from the FSL toolbox [138]. Lesion segmentation was calculated by the lesion segmentation toolbox [149] using the LPA algorithm, and streamline tractography was processed by the dipy toolbox [150]. The marching cubes algorithm was run on the lesions to create lesion surfaces. The cortical surface reconstruction of the anatomical image was calculated using Freesurfer [151], which also provided regional volumes and cortical thicknesses. The ANTS cortical thickness pipeline[128] was run in parallel to the Freesurfer pipeline. Shape features were extracted by the Mindboggle software package [152], which performs a more accurate labelling by combining the outputs of Freesurfer and ANTS, and also outputs the cortical surface mesh in the VTK-format. Lesion surfaces, cortical and subcortical surfaces, and streamlines were mapped to the anatomical image by applying the rigid body transformations so that all images are in the same coordinate space. To ensure processed data quality, processing outputs were quality controlled on the Mindcontrol application. The LST lesion probability map was displayed in Mindcontrol and quality controlled by a neuroradiologist, who annotated the false positive and false negative lesions using the point annotation feature. The false positive regions were removed and the false negative lesions were grown using a local threshold and size restriction ($< 1000\ mm^2$) around the click location. All surface-based data were saved in the .VTK format, except for the fiber tractography data, which was saved in the .OBJ model specified by BrainBrowser.

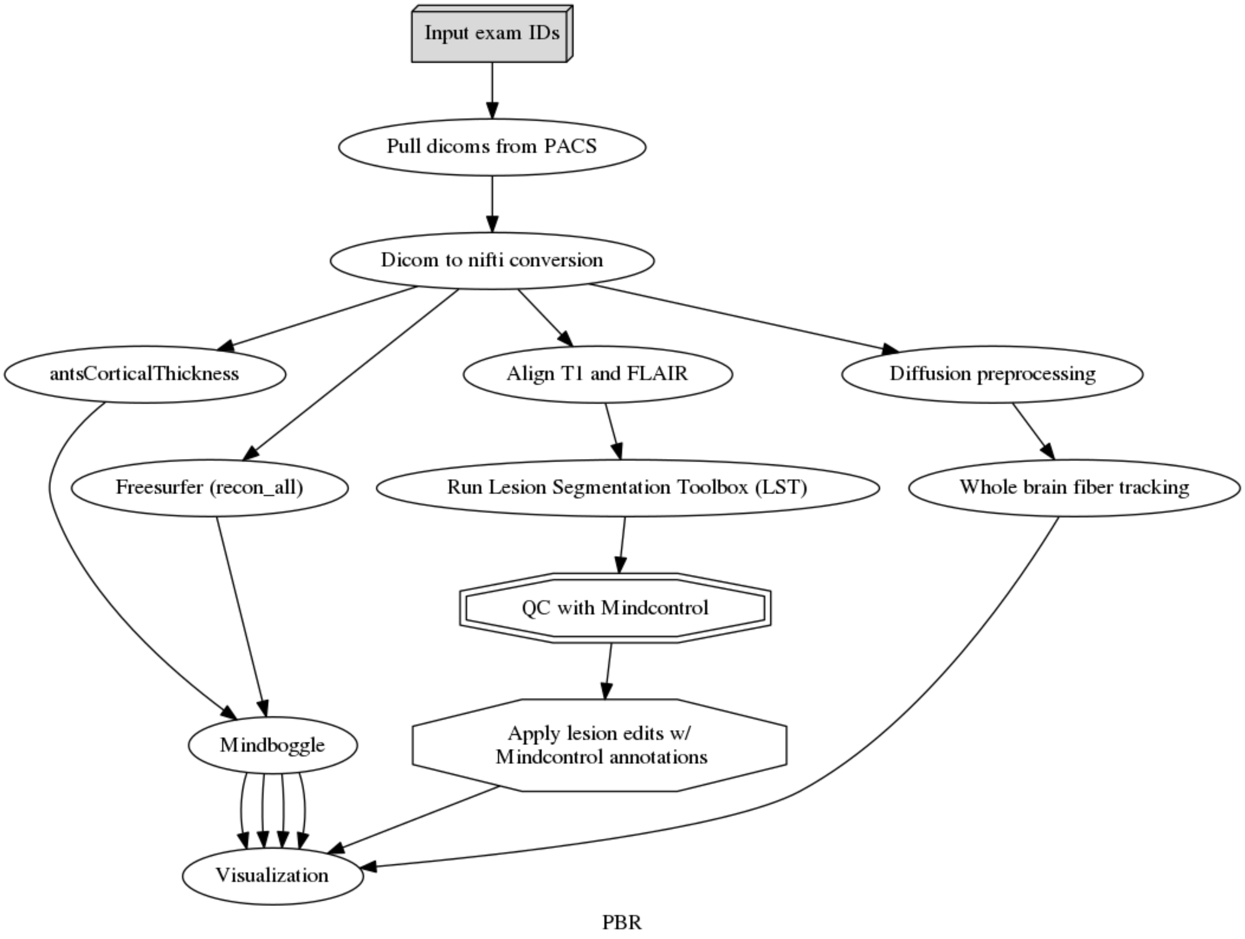


Figure 6.6: The PBRain workflow for the MindMeld visualization consists of 1) Pulling from the PACS database at UCSF 2) converteing all dicoms to the NifTI format, 3) Running antsCorticalThickness.sh, Freesurfer's recon-all, FLIRT alignment, and diffusion preprocessing to the the T1, FLAIR, and DW images. LST is run in the FLAIR image and quality controlled with Mindcontrol. The mindboggle software is run to improve the precision of labelling by combining the outputs of Freesurfer and ANTS. Whole brain fiber tracking is run on the preprocessed diffusion image to generate streamlines. All outputs feed into the visualization node, which prepares the data to be presented by the MindMeld application.

6.4 Discussion

The idiom, "An image is worth a thousand words", is particularly relevant in neuroimaging, where one image can produce thousands of features, based on shape, intensity, and image texture. The streamlined production of these features has been proposed through the PBrain interface, and their quality ensured by the Mindcontrol application, but the interpretation of such large amounts of data is still difficult. If we want to begin to understand these high-dimensional features, their relationship to each other, their normative distributions across populations, and how they can relate to disease progression, we must begin with an effective visualization. The information extracted from 4D imaging data are abstractions that highlight the important features of our data, in the same way a subway map highlights connections and line routes, and hides unnecessary information like the terrain. For example, for a given patient who presents a motor disability, a clinician might like to see the MS white-matter lesions over time, the corticospinal white matter tract, the cortical atrophy of the motor cortex relative to a reference population, and the walking speed of the patient over time, compared to disease trajectories of similar patients.

The MindMeld application presents all the clinical, genetic, and imaging data of a subject onto a one page interactive interface, that can be filtered to contextualize data based on a user selected sub-population. The visualization framework can present data from any number of MR modalities, any features extracted from them, and any relevant clinical, genetic, and epidemiological markers. The most striking feature in the 3D visualization from figure 6.5, is where the lesions are displayed on top of the white matter tracks, along with the 3D brain that shows cortical thickness by anatomical label. The spatial distribution of these lesions follow alongside the ventricles, which is a typical distribution pattern in MS [153, 154, 155]. Examining the pattern of lesions can be very informative in distinguishing between MS and other MS-like diseases, such as neuromyelitis optica. Another interesting feature seen from this application is the longitudinal atrophy pattern over time. Researchers have found different cortical atrophy patterns related to different disease subtypes and disabilities [156, 157]; the MindMeld visualization tool can help translate this group-level finding to the individual patient at the MS clinic. Finally, the link between cortical atrophy and lesion location [158] can be probed at the patient level, which could enable researchers to explore proposed mechanisms of cortical atrophy in MS, such as Wallerian degeneration [159].

6.5 Conclusion

The MindMeld application could help us gain a better understanding of the needs and preferences of clinicians, and could aid in medical decision making by providing an exploratory interactive data visualization platform. Some examples of exploratory analyses are: How does a patient's regional brain atrophy compare to a reference population of similar age, gender, and disease duration? How does a patient's brain atrophy pattern compare to a reference population of patients that present with similar symptoms? Is a patient's lesion volume at this time larger or smaller than a reference population of patients who are on a similar treatment? How does the disease trajectory of patients with similar cortical atrophy patterns compare to this patient, for the type of treatment received? The answers to these questions could be within our reach, through continued development of the MindMeld application.

We are moving into a world of big data, with an unprecedented amount of information available on each individual patient. This explosion of information availability has the capacity to revolutionize research and treatment of complex diseases such as MS. To harness the potential of these big data initiatives, platforms such as the MindMeld multiparametric visualization tool is crucial as we move forward into this new phase of interdisciplinary translational research.

7 Future Directions and Conclusions

7.1 MindGames- A Crowd-Sourcing Game Platform for Brain MRI Segmentation

7.1.1 Rationale

Advances in MRI technology and image segmentation algorithms have enabled researchers to begin to understand the mechanisms of healthy brain development [160] and neurological disorders such as multiple sclerosis [161]. Due to the wide variability of brain morphology coupled with pathological processes in the case of neurological disorders, increasingly large sample sizes are necessary to confidently answer the progressively complex biomedical questions of the research community. Automated algorithms have been developed to reduce information-rich 3D MRI images to 1-dimensional summary measures that describe tissue properties and are easy to interpret, such as total gray matter volume. Automated segmentation algorithms save considerable time compared to manual human inspection, but lack the advanced visual system of humans. As a result, these algorithms often make systematic errors, especially when analyzing brains with pathology or those in the early stages of development. Data science is poised to facilitate complex neuroscience research by fusing a crowdsourcing strategy with machine learning methods; automatic quantification can perform the bulk of the work efficiently and errors can be resolved by non-expert "citizen-scientists" with the advantage of the human visual system.

Crowdsourcing has been successful in many other disciplines [162], including mathematics [163], astronomy [164], and biochemistry [165] . Recently, over 200,000 "citizen-neuroscientists" from over 147 countries helped identify neuronal connections in a mouse retina through the Eyewire game [166]. This crowdsourced game led to a new understanding of how mammalian retinal cells detect motion. I propose to implement three key features of the EyeWire paradigm and adapt them for the segmentation of MRI data. First, by breaking up the problem into smaller "micro-tasks", Eyewire scientists were able to access a much larger user-pool of non-experts. In a similar vein, 3D MRI data can be divided into 2D slices to be segmented by users. Second, machine learning algorithms were trained to help with the task, which improved the speed of manual neuronal tracing and validated non-expert input in the Eyewire game. Deep learning methods have already shown to be successful at segmenting MRI data, and similar models could be built to support manual segmentation. Lastly, EyeWire transformed a dull, monotonous task for experts into a fun, competitive game that trained non-experts, and acquired valuable scientific data. The University of Washington is an ideal place to develop a similar game platform for MRI segmentation, using the resources at the Center for Game Science, led by Zoran Popovic. I propose to create an open-source platform for efficiently crowdsourcing brain tissue classification problems in order to answer neuroscience research questions with more precision.

Specific Aims

1. **Scaleable and Secure Micro-Tasks**: A scaleable database system and server backend that keeps data private by dividing it into small "micro-tasks"

2. **Learning by Example**: Machine learning algorithm that learns from human curation to improve efficiency of manual tasks

3. **Training through Gamification**: User interface that trains users to solve a specific problem, and keeps them engaged through a reward system

7.1.2 Specific Aim 1: Scaleable and Secure Micro-Tasks

This Aim will address two key challenges: 1) Partitioning 3D data into micro-tasks that keep data private, 2) Serving micro-tasks at scale. While there are many large-scale open-source data collection efforts, many datasets are kept private within research institutions due to IRB restrictions, so presenting a full 3D MRI volume to the public would be a violation. Serving smaller "chunks" of data serves two purposes: it allows us to keep data private (because you cannot see the whole brain), and it reduces the fatigue of non-experts (because you only need to edit a small section), which enables us to engage a larger user base. A scalable server will be implemented on a commercial cloud computing platform, with an API that allows researchers to upload MRI micro-tasks to the server database, and serves micro-tasks to users. Researchers will be asked to provide the following to the API: 1) an initial segmentation file from an automated algorithm 2) any original images (T1, T2, PD) that users need to properly edit the segmentation 3) directions on how the images should be sliced into micro-tasks (including the slicing plane and the number of slices). Additionally, researchers must provide a validation dataset, which includes "correctly" segmented images, which will be used to train non-experts in Aim 3. The resources and faculty at the eScience Institute will help me implement state-of-the-art database and cloud computing technologies in order to increase the delivery of micro-tasks to "citizen-scientists."

7.1.3 Specific Aim 2: Learning by Example

This Aim will address three challenges: 1) Resolve user input to create a final 3D volume, 2) Prioritize serving micro-tasks based on user consensus and 3) Predict the user-edited segmentation image. To reconstruct the micro-tasks back into a 3D image, a weighted consensus map will be computed, based on how accurately each user performed edits on training data. Micro-tasks with lower consensus scores will be served more frequently to users, until the consensus is high. Participants will also be scored based on how well their segmentations match with other users on the same image, and this will be used to reward users in Aim 3. Finally, improving automated segmentation algorithms based on human input will save time and reduce the number of editors assigned to each micro-task. For example, a dataset of 100 3D volumes could be broken into 20,000 patches, each of which would need to be manually edited. Alternatively, convolutional neural networks (CNNs) have been very successful at pattern recognition when trained on similarly large sample sizes, and could reduce the time spent editing each patch. I propose to build a CNN using existing architecture, such as Tensorflow or Theanet, to predict segmentation results, under the guidance of the machine learning experts at the eScience institute.

7.1.4 Specific Aim 3: Training and Gamification

For individuals with minimal neuroamatomy knowledge, the difficulty of manual neuroimaging segmentation will depend on the contrast of the image as well as the location/complexity of the target structure. An example of an easy task would be the segmentation of brain tissue from non-brain tissue, whereas a more difficult task would be the segmentation of multiple sclerosis lesions. This Aim will address simple as well as challenging problems through varying levels of training and rewards. A web application will be developed that hooks into the server developed in Aim 1. The app will include an in-browser brain editor (similar to the Mindcontrol application [5]), a reward structure and a scoreboard for the top users, and an optional link to the Amazon Turk engine, where users can be paid (in micro-payments) for completing micro-tasks. Initially, the user will only be presented with training tasks until they reach an adequate accuracy score. Next, the training tasks will be interspersed with new tasks, in order to detect performnce drift. The frequency of training tasks will increase based on the researcher's specification of task difficulty. The reward structure will be based on 1) how well the user edits training data, 2) how well the user segmentations match those of other users, and 3) how many voxels are edited by the user. The time spent on the task along with the number of edited voxels will also be used to validate whether or not the user completed the task with some thought. For example, a user's score would be penalized if a large number of voxels were edited too quickly for a difficult task. I plan to collaborate with the data scientists at the eScience Institute to build an intuitive and engaging crowd-sourcing user interface on the Amazon Turk Platform.

7.1.5 Broader Impacts

To summarize, I propose to develop an open-source platform for the crowd-sourced image segmentation of brain MRI data, under the guidance of Ariel Rokem and Jason Yeatman at the eScience Institute and the University of Washington Institute for Neuroengineering. Through gamification, piece-wise exposure, and machine learning, I plan to engage a large user base across a variety of image segmentation tasks. Example applications include parcellating gray and white matter in a low contrast image where traditional segmentation algorithms fail, and delineating multiple sclerosis lesions which usually requires trained neuroradiologists. For a particular application, the Yeatman Lab at UW is collecting a large, longitudinal MRI dataset on children undergoing an intensive learning program, with the goal of determining how experience shapes brain development. The segmentation data from the MindGames platform can be used to 1) define the typical time course of cortical changes by examining gray/white matter volumes from segmentation, 2) construct normative developmental curves in order to detect abnormalities, and 3) study how learning shapes brain development by analyzing quantitative MR intensities within the gray and white matter. The MindGames platform will help researchers by improving the precision of segmentation measures without advanced computer science expertise, but will also engage, educate and excite the public, and help advance cutting edge neuroscience research.

7.2 Overall Conclusion: Open Collaboration is Key

The overarching result of this work was the production of **open and collaborative** informatics tools to help scientists answer difficult and complex biomedical questions. The aim of the first part of this dissertation was to help multiple institutions pool data by lowering the requirements for data harmonization. The results of this study encourage researchers to collaborate by pooling both prospective and retrospectively acquired data, in order to define normative morphometry on a diverse population, study how deviations from normalcy are associated to pathology, probe genotype-phenotype relationships, and investigate rare diseases, to name a few possible applications. Next, the PBRain framework was proposed as a collaborative software code base, to streamline knowledge transfer in academic environments with generally high turnover rates of specialized researchers. This collaborative software empowers scientists to reproduce results quickly and easily, and lowers the learning curve to perofrm advanced neuroimaging analyses. The third project was Mindcontrol, an open-source, collaborative web-based platform for brain quality control. In the past, the sample size of a typical neuroimaging study was less than 100 subjects, and ensuring data quality through visual inspection was time consuming, but manageable. Advances in compute power and MR sequences led to an increased rate of data collection and processing, but visual inspection remained a major bottleneck. Mindcontrol, and its future iteration, MindGames, leverage web technology to bring data to and solicit contributions from remote researchers and citizen scientists. In the future, deep learning networks could be trained to quality check data; but to accomplish this, a large, collaboratively annotated data set is needed. Finally, a future application of the MindMeld platform is to be a collaborative, translational tool between researchers and clinicians. While rapid data collection and computing power accelerate the discovery of biomarkers in research, their utility in a real world setting remains unknown. The MindMeld visualization platform could be used as a staging ground for new biomarkers; their utility could be tested in the real world clinic by tracking use patterns and seeking feedback from clinicians. In the world of big data, open collaboration is a key strategy to tackle new challenges, to foster new ideas, and to solve the most pressing biomedical problems.

Project	URL	Contribution
Nipype [53]	`http://github.com/nipy/nipype`	Wrote new interfaces and submitted bug fixes
BIPS [100]	`http://github.com/INCF/BrainImagingPipelines`	Wrote reusable and configurable nipype pipelines
nbpapaya	`http://github.com/akeshavan/nbpapaya`	Wrote ipython notebook interface for 3D volumes
mindboggle [144]	`http://github.com/nipy/mindboggle`	Submitted bug fixes
roygbiv [167]	`http://github.com/akeshavan/roygbiv`	Wrote visualization for mindboggle outputs
BIDS-Apps [110]	`http://github.com/BIDS-Apps/mindboggle`	Wrote docker container for mindboggle
Mindcontrol [5]	`http://github.com/akeshavan/mindcontrol`	Wrote application for brain QC
Brainspell	`http://github.com/openneuro/brainspell-neo`	Wrote github integration for collaborative annotation of papers for meta analyses

Table 7.1: Table of my open source contributions and collaborations in the neuroimaging field.

Bibliography

[1] Eric D. Green. “Opening plenary speaker: Human genomics precision medicine, and advancing human health”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Institute of Electrical and Electronics Engineers (IEEE), Aug. 2016. DOI: `10.1109/embc.2016.7590619`. URL: `https://doi.org/10.1109%2Fembc.2016.7590619`.

[2] A Keshavan et al. “Power estimation for non-standardized multisite studies.” In: *Neuroimage* 134 (July 2016), pp. 281–94.

[3] Krzysztof J. Gorgolewski et al. “The brain imaging data structure a format for organizing and describing outputs of neuroimaging experiments”. In: *Scientific Data* 3 (June 2016), p. 160044. DOI: `10.1038/sdata.2016.44`. URL: `https://doi.org/10.1038%2Fsdata.2016.44`.

[4] Anisha Keshavan et al. “Mindcontrol: Organize quality control, annotate, edit, and collaborate on neuroimaging processing results”. In: *Research Ideas and Outcomes* 3 (Feb. 2017), e12276. DOI: `10.3897/rio.3.e12276`. URL: `https://doi.org/10.3897%2Frio.3.e12276`.

[5] Anisha Keshavan et al. “Mindcontrol: A Web Application for Brain Segmentation Quality Control”. In: *bioRxiv* (2016), p. 090431.

[6] J F Kurtzke. “Epidemiologic evidence for multiple sclerosis as an infection.” In: *Clinical Microbiology Reviews* 6.4 (Oct. 1993), pp. 382–427. DOI: `10.1128/cmr.6.4.382`. URL: `https://doi.org/10.1128%2Fcmr.6.4.382`.

[7] Casper Jersild et al. “HISTOCOMPATIBILITY DETERMINANTS IN MULTIPLE SCLEROSIS WITH SPECIAL REFERENCE TO CLINICAL COURSE”. In: *The Lancet* 302.7840 (Dec. 1973), pp. 1221–1225. DOI: 10.1016/s0140-6736(73)90970-7. URL: https://doi.org/10.1016%2Fs0140-6736%2873%2990970-7.

[8] M. R. Lincoln et al. “Epistasis among HLA-DRB1 HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility”. In: *Proceedings of the National Academy of Sciences* 106.18 (Apr. 2009), pp. 7542–7547. DOI: 10.1073/pnas.0812664106. URL: https://doi.org/10.1073%2Fpnas.0812664106.

[9] David A Dyment, George C Ebers, and A Dessa Sadovnick. “Genetics of multiple sclerosis”. In: *The Lancet Neurology* 3.2 (2004), pp. 104–110.

[10] C Luzzio and F Dangond. *Multiple Sclerosis Clinical Presentation*. 2017. URL: http://emedicine.medscape.com/article/1146199-clinical (visited on 05/12/2017).

[11] Nils Koch-Henriksen and Per Soelberg Sørensen. “The changing demographic pattern of multiple sclerosis epidemiology”. In: *The Lancet Neurology* 9.5 (May 2010), pp. 520–532. DOI: 10.1016/s1474-4422(10)70064-8. URL: https://doi.org/10.1016%2Fs1474-4422%2810%2970064-8.

[12] Douglas S Goodin et al. “Disease modifying therapies in multiple sclerosis”. In: *Neurology* 58.2 (2002), pp. 169–178.

[13] IR Young et al. “Nuclear magnetic resonance imaging of the brain in multiple sclerosis”. In: *The Lancet* 318.8255 (1981), pp. 1063–1066.

[14] W Ian McDonald et al. “Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis”. In: *Annals of neurology* 50.1 (2001), pp. 121–127.

[15] Barbara S Giesser. *Primer on multiple sclerosis*. Oxford University Press, 2015.

[16] Chris H Polman et al. “Diagnostic criteria for multiple sclerosis: 2005 revisions to the “McDonald Criteria””. In: *Annals of neurology* 58.6 (2005), pp. 840–846.

[17] JH Van Waesberghe et al. “Patterns of lesion development in multiple sclerosis: longitudinal observations with T1-weighted spin-echo and magnetization transfer MR.” In: *American journal of neuroradiology* 19.4 (1998), pp. 675–683.

[18] MAA Van Walderveen et al. “Histopathologic correlate of hypointense lesions on T1-weighted spin-echo MRI in multiple sclerosis”. In: *Neurology* 50.5 (1998), pp. 1282–1288.

[19] Belinda SY Li et al. “Brain metabolite profiles of T1-hypointense lesions in relapsing-remitting multiple sclerosis”. In: *American journal of neuroradiology* 24.1 (2003), pp. 68–74.

[20] M Filippi et al. “Correlations between changes in disability and T2-weighted brain MRI activity in multiple sclerosis A follow-up study”. In: *Neurology* 45.2 (1995), pp. 255–260.

[21] J Furby et al. “A longitudinal study of MRI-detected atrophy in secondary progressive multiple sclerosis”. In: *Journal of neurology* 257.9 (2010), pp. 1508–1516.

[22] Ludwig Kappos et al. “Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis”. In: *The Lancet* 353.9157 (1999), pp. 964–969.

[23] Frederik Barkhof. “The clinico-radiological paradox in multiple sclerosis revisited”. In: *Current opinion in neurology* 15.3 (2002), pp. 239–245.

[24] Elizabeth Fisher et al. “Gray matter atrophy in multiple sclerosis: a longitudinal study”. In: *Annals of neurology* 64.3 (2008), pp. 255–265.

[25] Leonora K Fisniku et al. “Gray matter atrophy is related to long-term disability in multiple sclerosis”. In: *Annals of neurology* 64.3 (2008), pp. 247–254.

[26] Catherine M Dalton et al. “Early development of multiple sclerosis is associated with progressive grey matter atrophy in patients presenting with clinically isolated syndromes”. In: *Brain* 127.5 (2004), pp. 1101–1107.

[27] Antonio Giorgio et al. “Brain atrophy assessment in multiple sclerosis: importance and limitations”. In: *Neuroimaging clinics of North America* 18.4 (2008), pp. 675–686.

[28] Mark W Woolrich et al. “Bayesian analysis of neuroimaging data in FSL”. In: *Neuroimage* 45.1 (2009), S173–S186.

[29] Allison C Nugent et al. “Automated subcortical segmentation using FIRST: test–retest reliability, interscanner reliability, and comparison to manual segmentation”. In: *Human brain mapping* 34.9 (2013), pp. 2313–2329.

[30] Brian B Avants et al. “An open source multivariate framework for n-tissue segmentation with evaluation on public data”. In: *Neuroinformatics* 9.4 (2011), pp. 381–400.

[31] Bruce Fischl. “FreeSurfer”. In: *Neuroimage* 62.2 (2012), pp. 774–781.

[32] STUART GEMAN and DONALD GEMAN. “Stochastic Relaxation Gibbs Distributions, and the Bayesian Restoration of Images”. In: *Readings in Computer Vision*. Elsevier BV, 1987, pp. 564–584. DOI: 10.1016/b978-0-08-051581-6.50057-x. URL: https://doi.org/10.1016%2Fb978-0-08-051581-6.50057-x.

[33] Ivana Despotović, Bart Goossens, and Wilfried Philips. “MRI Segmentation of the Human Brain: Challenges Methods, and Applications”. In: *Computational and Mathematical Methods in Medicine* 2015 (2015), pp. 1–23. DOI: 10.1155/2015/450341. URL: https://doi.org/10.1155%2F2015%2F450341.

[34] Bruce Fischl et al. “Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain”. In: *Neuron* 33.3 (2002), pp. 341–355.

[35] Jeroen JG Geurts and Frederik Barkhof. “Grey matter pathology in multiple sclerosis”. In: *The Lancet Neurology* 7.9 (2008), pp. 841–851.

[36] Lars Bø et al. “Subpial demyelination in the cerebral cortex of multiple sclerosis patients”. In: *Journal of Neuropathology & Experimental Neurology* 62.7 (2003), pp. 723–732.

[37] Ponnada A Narayana et al. “Regional cortical thickness in relapsing remitting multiple sclerosis: a multi-center study”. In: *NeuroImage: Clinical* 2 (2013), pp. 120–131.

[38] Michael Sailer et al. “Focal thinning of the cerebral cortex in multiple sclerosis”. In: *Brain* 126.8 (2003), pp. 1734–1744.

[39] Gro O Nygaard et al. “Cortical thickness and surface area relate to specific symptoms in early relapsing–remitting multiple sclerosis”. In: *Multiple Sclerosis Journal* (2014), p. 1352458514543811.

[40] JT Chen et al. “Relating neocortical pathology to disability progression in multiple sclerosis using MRI”. In: *Neuroimage* 23.3 (2004), pp. 1168–1175.

[41] Deepa Preeti Ramasamy et al. “Extent of cerebellum, subcortical and cortical atrophy in patients with MS: a case-control study”. In: *Journal of the neurological sciences* 282.1 (2009), pp. 47–54.

[42] Massimiliano Calabrese et al. “Widespread cortical thinning characterizes patients with MS with mild cognitive impairment”. In: *Neurology* 74.4 (2010), pp. 321–328.

[43] MK Houtchens et al. “Thalamic atrophy and cognition in multiple sclerosis”. In: *Neurology* 69.12 (2007), pp. 1213–1223.

[44] Sonia Batista et al. “Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis”. In: *Journal of neurology* 259.1 (2012), pp. 139–146.

[45] Robert Zivadinov et al. “Thalamic atrophy is associated with development of clinically definite multiple sclerosis”. In: *Radiology* 268.3 (2013), pp. 831–841.

[46] Alberto Cifelli et al. “Thalamic neurodegeneration in multiple sclerosis”. In: *Annals of neurology* 52.5 (2002), pp. 650–653.

[47] Gianna Riccitelli et al. “Cognitive impairment in multiple sclerosis is associated to different patterns of gray matter atrophy according to clinical phenotype”. In: *Human brain mapping* 32.10 (2011), pp. 1535–1543.

[48] Arnaud Charil et al. “Focal cortical atrophy in multiple sclerosis: relation to lesion load and disability”. In: *Neuroimage* 34.2 (2007), pp. 509–517.

[49] NL Sicotte et al. “Regional hippocampal atrophy in multiple sclerosis”. In: *Brain* 131.4 (2008), pp. 1134–1141.

[50] Carmen Tur et al. "HLA-DRB1* 15 influences the development of brain tissue damage in early PPMS". In: *Neurology* 83.19 (2014), pp. 1712–1718.

[51] DT Okuda et al. "Genotype–phenotype correlations in multiple sclerosis: HLA genes influence disease severity inferred by 1HMR spectroscopy and MRI measures". In: *Brain* 132.1 (2009), pp. 250–259.

[52] Noriko Isobe et al. "Association of HLA Genetic Risk Burden With Disease Phenotypes in Multiple Sclerosis". In: *JAMA neurology* 73.7 (2016), pp. 795–802.

[53] Krzysztof Gorgolewski et al. "Nipype: A Flexible Lightweight and Extensible Neuroimaging Data Processing Framework in Python". In: *Frontiers in Neuroinformatics* 5 (2011). DOI: 10.3389/fninf.2011.00013. URL: https://doi.org/10.3389%2Ffninf.2011.00013.

[54] Tyrone D. Cannon et al. "Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis". In: *Human Brain Mapping* 35.5 (2014), pp. 2424–2434. ISSN: 1097-0193. DOI: 10.1002/hbm.22338. URL: http://dx.doi.org/10.1002/hbm.22338.

[55] M Ewers et al. "Multicenter assessment of reliability of cranial MRI". In: *Neurobiology of Aging* 27.8 (2006), pp. 1051–1059.

[56] Jorge Jovicich et al. "Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data". In: *NeuroImage* 30.2 (2006), pp. 436–443. ISSN: 1053-8119. DOI: DOI:10.1016/j.neuroimage.2005.09.046. URL: http://www.sciencedirect.com/science/article/B6WNP-4HM7S0B-2/2/4fa5ff26cad90ba3c9ed12b7e12ce3b6.

[57] M Boccardi et al. "EADC-ADNI Working Group on The Harmonized Protocol for Hippocampal Volumetry and for the Alzheimer's Disease Neuroimaging Initiative: Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation". In: *Alzheimers Dement* (2013).

[58] Michael W Weiner et al. "The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception". In: *Alzheimer's & Dementia* 8.1 (2012), S1–S68.

[59] Paul M Thompson et al. “The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data”. In: *Brain imaging and behavior* 8.2 (2014), pp. 153–182.

[60] Vladimir S Fonov et al. “Improved precision in the measurement of longitudinal global and regional volumetric changes via a novel MRI gradient distortion characterization and correction technique”. In: *Medical Imaging and Augmented Reality*. Springer, 2010, pp. 324–333.

[61] Jeffrey L Gunter et al. “Measurement of MRI scanner performance with the ADNI phantom”. In: *Medical physics* 36.6 (2009), pp. 2193–2205.

[62] Zografos Caramanos et al. “Gradient distortions in MRI: characterizing and correcting for their effects on SIENA-generated measures of brain volume change”. In: *NeuroImage* 49.2 (2010), pp. 1601–1611.

[63] Amgad Droby et al. “A human post-mortem brain model for the standardization of multi-centre MRI studies”. In: *NeuroImage* 110 (2015), pp. 11–21.

[64] Christine L Tardif, D Louis Collins, and G Bruce Pike. “Regional impact of field strength on voxel-based morphometry results”. In: *Human brain mapping* 31.7 (2010), pp. 943–957.

[65] Christine L Tardif, D Louis Collins, and G Bruce Pike. “Sensitivity of voxel-based morphometry analysis to choice of imaging protocol at 3 T”. In: *Neuroimage* 44.3 (2009), pp. 827–838.

[66] Simon Brunton et al. “A voxel-based morphometry comparison of the 3.0T ADNI-1 and ADNI-2 MPRAGE protocols”. In: *Alzheimer’s & Dementia* 9.4 (July 2013), P581. DOI: `10.1016/j.jalz.2013.05.1154`. URL: `http://dx.doi.org/10.1016/j.jalz.2013.05.1154`.

[67] Robin Wolz et al. “Robustness of automated hippocampal volumetry across magnetic resonance field strengths and repeat images”. In: *Alzheimer’s & Dementia* 10.4 (July 2014), 430–438.e2. DOI: `10.1016/j.jalz.2013.09.014`. URL: `http://dx.doi.org/10.1016/j.jalz.2013.09.014`.

[68] Dana Horakova et al. “Clinical correlates of grey matter pathology in multiple sclerosis”. In: *BMC neurology* 12.1 (2012), p. 10.

[69] Michael P Sanfilipo et al. “Gray and white matter brain atrophy and neuropsychological impairment in multiple sclerosis”. In: *Neurology* 66.5 (2006), pp. 685–692.

[70] Alberto Cifelli et al. “Thalamic neurodegeneration in multiple sclerosis”. In: *Annals of neurology* 52.5 (2002), pp. 650–653.

[71] R Zivadinov et al. “Evolution of cortical and thalamus atrophy and disability progression in early relapsing-remitting MS during 5 years”. In: *American Journal of Neuroradiology* 34.10 (2013), pp. 1931–1939.

[72] MK Houtchens et al. “Thalamic atrophy and cognition in multiple sclerosis”. In: *Neurology* 69.12 (2007), pp. 1213–1223.

[73] M Wylezinska et al. “Thalamic neurodegeneration in relapsing-remitting multiple sclerosis”. In: *Neurology* 60.12 (2003), pp. 1949–1954.

[74] Robert A Bermel et al. “Selective caudate atrophy in multiple sclerosis: a 3D MRI parcellation study”. In: *Neuroreport* 14.3 (2003), pp. 335–339.

[75] Guozhi Tao et al. “Deep gray matter atrophy in multiple sclerosis: a tensor based morphometry”. In: *Journal of the neurological sciences* 282.1 (2009), pp. 39–46.

[76] G David Garson. “Fundamentals of hierarchical linear and multilevel modeling”. In: *Hierarchical linear modeling: guide and applications. Sage Publications Inc* (2013), pp. 3–25.

[77] Christine Fennema-Notestine et al. “Feasibility of multi-site clinical structural neuroimaging studies of aging using legacy data”. In: *Neuroinformatics* 5.4 (2007), pp. 235–245.

[78] S. W. Raudenbush and X. Liu. “Statistical power and optimal design for multisite randomized trials.” In: *Psychological methods* 5.2 (June 2000), pp. 199–213. ISSN: 1082-989X. URL: http://view.ncbi.nlm.nih.gov/pubmed/10937329.

[79] B. Fischl et al. “Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain”. In: *Neuron* 33 (2002), pp. 341–355.

[80] Randy L Buckner et al. “A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume”. In: *Neuroimage* 23.2 (2004), pp. 724–738.

[81] Lee Friedman et al. “Test–retest and between-site reliability in a multicenter fMRI study”. In: *Human brain mapping* 29.8 (2008), pp. 958–972.

[82] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.5.8. Northwestern University. Evanston, Illinois, 2015. URL: %7Bhttp://CRAN.R-project.org/package=psych%7D.

[83] J. F. Kurtzke. “Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)”. In: *Neurology* 33.11 (Nov. 1983), pp. 1444–1444. DOI: 10.1212/wnl.33.11.1444. URL: http://dx.doi.org/10.1212/wnl.33.11.1444.

[84] Daniel-Paolo Streitbürger et al. “Impact of image acquisition on voxel-based-morphometry investigations of age-related structural brain changes”. In: *Neuroimage* 87 (2014), pp. 170–182.

[85] Jorge Jovicich et al. “Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations”. In: *Neuroimage* 83 (2013), pp. 472–484.

[86] Hugo G Schnack et al. “Mapping reliability in multicenter MRI: Voxel-based morphometry and cortical thickness”. In: *Human Brain Mapping* 31.12 (2010), pp. 1967–1982.

[87] Brian Patenaude et al. “A Bayesian model of shape and appearance for subcortical brain segmentation”. In: *Neuroimage* 56.3 (2011), pp. 907–922.

[88] Allison C Nugent et al. “Automated subcortical segmentation using FIRST: Test–retest reliability, interscanner reliability, and comparison to manual segmentation”. In: *Human brain mapping* 34.9 (2013), pp. 2313–2329.

[89] Robin Wolz et al. “LEAP: Learning embeddings for atlas propagation”. In: *NeuroImage* 49.2 (Jan. 2010), pp. 1316–1325. DOI: 10.1016/j.neuroimage.2009.09.069. URL: http://dx.doi.org/10.1016/j.neuroimage.2009.09.069.

[90] Alexis Roche and Florence Forbes. “Partial Volume Estimation in Brain MRI Revisited”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Springer Science Business Media, 2014, pp. 771–778. DOI: 10.1007/978-3-319-10404-1_96. URL: http://dx.doi.org/10.1007/978-3-319-10404-1_96.

[91] Eun Young Kim and Hans J Johnson. “Robust Multi-site MR Data Processing: Iterative Optimization of Bias Correction, Tissue Classification, and Registration”. In: *Frontiers in Neuroinformatics* 7.29 (2013). ISSN: 1662-5196. DOI: 10.3389/fninf.2013.00029. URL: http://www.frontiersin.org/neuroinformatics/10.3389/fninf.2013.00029/abstract.

[92] Hongzhi Wang et al. “A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation”. In: *NeuroImage* 55.3 (2011), pp. 968–985. ISSN: 1053-8119. DOI: http://dx.doi.org/10.1016/j.neuroimage.2011.01.006. URL: http://www.sciencedirect.com/science/article/pii/S1053811911000243.

[93] Paul M Thompson et al. “Detecting Sisease-Specific Patterns of Brain Structure Using Cortical Pattern Matching and a Population-Based Probabilistic Brain Atlas”. In: *Information Processing in Medical Imaging*. Springer. 2001, pp. 488–501.

[94] Hugo G. Schnack et al. “Reliability of brain volumes from multicenter MRI acquisition: A calibration study”. In: *Human Brain Mapping* 22.4 (2004), pp. 312–320. DOI: 10.1002/hbm.20040. URL: http://dx.doi.org/10.1002/hbm.20040.

[95] Blake C Jones et al. “Quantification of multiple-sclerosis-related brain atrophy in two heterogeneous MRI datasets using mixed-effects modeling”. In: *NeuroImage: Clinical* 3 (2013), pp. 171–179.

[96] Jorge Jovicich et al. "MRI-derived measurements of human subcortical ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths". In: *NeuroImage* 46.1 (May 2009), pp. 177–192. DOI: 10.1016/j.neuroimage.2009.02.010. URL: http://dx.doi.org/10.1016/j.neuroimage.2009.02.010.

[97] Jennifer L. Whitwell. "Comparison of Imaging Biomarkers in the Alzheimer Disease Neuroimaging Initiative and the Mayo Clinic Study of Aging". In: *Arch Neurol* 69.5 (May 2012), p. 614. DOI: 10.1001/archneurol.2011.3029. URL: http://dx.doi.org/10.1001/archneurol.2011.3029.

[98] Buhm Han and Eleazar Eskin. "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies". In: *The American Journal of Human Genetics* 88.5 (2011), pp. 586–598.

[99] Duane C Boes, FA Graybill, and AM Mood. "Introduction to the Theory of Statistics". In: *Series in probabili* (1974).

[100] Satrajit Ghosh et al. "BIPS: A Framework for Curating and Executing Brain Imaging Pipelines". In: *Frontiers in Neuroinformatics* 53 (2014). ISSN: 1662-5196. DOI: 10.3389/conf.fninf.2014.08.00053. URL: http://www.frontiersin.org/neuroinformatics/10.3389/conf.fninf.2014.08.00053/full.

[101] Anisha Keshavan et al. *akeshavan/BrainImagingPipelines: Zenodo*. Mar. 2017. DOI: 10.5281/zenodo.346009. URL: https://doi.org/10.5281/zenodo.346009.

[102] Tyler K Perrachione and Satrajit S Ghosh. "Optimized design and analysis of sparse-sampling FMRI experiments". In: *Frontiers in neuroscience* 7 (2013), p. 55.

[103] Jonathan Smallwood et al. "The default modes of reading: modulation of posterior cingulate and medial prefrontal cortex connectivity associated with comprehension and task focus while reading". In: *Frontiers in human neuroscience* 7 (2013), p. 734.

[104] Alexander Schaefer et al. "Dynamic network participation of functional connectivity hubs assessed by resting-state fMRI". In: *Frontiers in human neuroscience* 8 (2014), p. 195.

[105] Benjamin Baird et al. "Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception". In: *Journal of Neuroscience* 33.42 (2013), pp. 16657–16665.

[106] Amy S Finn et al. "Functional brain organization of working memory in adolescents varies in relation to family income and academic achievement". In: *Developmental Science* (2016).

[107] C Craddock et al. "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac)". In: *Front Neuroinform* 42 (2013).

[108] Russell A Poldrack et al. "Toward open sharing of task-based fMRI data: the OpenfMRI project". In: *Frontiers in neuroinformatics* 7 (2013), p. 12.

[109] KJ Gorgolewski et al. "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments." In: *Sci Data* 3 (June 2016), p. 160044.

[110] Krzysztof J Gorgolewski et al. "BIDS Apps: Improving ease of use, accessibility and reproducibility of neuroimaging data analysis methods". In: *bioRxiv* (2016), p. 079145.

[111] BC Dickerson et al. "The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals." In: *Cerebral Cortex* 19 (Mar. 2009), pp. 497–510. DOI: 10.1093/cercor/bhn113.

[112] Prashanthi Vemuri and Clifford R Jack. "Role of structural MRI in Alzheimer's disease". In: *Alzheimers Research and Therapy* 2.4 (2010), p. 23. DOI: 10.1186/alzrt47. URL: http://dx.doi.org/10.1186/alzrt47.

[113] Paul J Tuite Silvia Mangia. "Magnetic Resonance Imaging (MRI) in Parkinson's Disease". In: *Journal of Alzheimer's Disease & Parkinsonism* 03.03 (2013). DOI: 10.4172/2161-0460.s1-001. URL: http://dx.doi.org/10.4172/2161-0460.s1-001.

[114] Martha E Shenton et al. "A review of MRI findings in schizophrenia". In: *Schizophrenia research* 49.1 (2001), pp. 1–52. DOI: 10.1016/S0920-9964(01)00163-3. URL: http://www.schres-journal.com/article/S0920-9964(01)00163-3/abstract.

[115] Francesco Amico et al. “Structural MRI correlates for vulnerability and resilience to major depressive disorder”. In: *Journal of psychiatry & neuroscience: JPN* 36.1 (2011), p. 15. DOI: 10.1503/jpn.090186.

[116] Paolo Brambilla et al. “Brain anatomy and development in autism: review of structural MRI studies”. In: *Brain research bulletin* 61.6 (2003), pp. 557–569. DOI: 10.1016/j.brainresbull.2003.06.001.

[117] M. Filippi et al. “Quantitative assessment of MRI lesion load in monitoring the evolution of multiple sclerosis”. In: *Brain* 118.6 (1995), pp. 1601–1612. DOI: 10.1093/brain/118.6.1601. URL: http://dx.doi.org/10.1093/brain/118.6.1601.

[118] Jonathan D. Blumenthal et al. “Motion Artifact in Magnetic Resonance Imaging: Implications for Automated Analysis”. In: *NeuroImage* 16.1 (May 2002), pp. 89–92. DOI: 10.1006/nimg.2002.1076. URL: http://dx.doi.org/10.1006/nimg.2002.1076.

[119] Heath R. Pardoe, Rebecca Kucharsky Hiess, and Ruben Kuzniecky. “Motion and morphometry in clinical and nonclinical populations”. In: *NeuroImage* 135 (2016), pp. 177–185. DOI: 10.1016/j.neuroimage.2016.05.005. URL: http://dx.doi.org/10.1016/j.neuroimage.2016.05.005.

[120] Martin Reuter et al. “Head motion during MRI acquisition reduces gray matter volume and thickness estimates”. In: *NeuroImage* 107 (Feb. 2015), pp. 107–115. DOI: 10.1016/j.neuroimage.2014.12.006. URL: http://dx.doi.org/10.1016/j.neuroimage.2014.12.006.

[121] Neil K. Savalia et al. “Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion”. In: *Human Brain Mapping* (Sept. 2016). DOI: 10.1002/hbm.23397. URL: http://dx.doi.org/10.1002/hbm.23397.

[122] Anisha Keshavan et al. “Power estimation for non-standardized multisite studies”. In: *NeuroImage* 134 (2016), pp. 281–294. DOI: 10.1016/j.neuroimage.2016.03.051.

[123] Xiao Han et al. “Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength scanner upgrade and manufacturer”. In: *NeuroImage*

32.1 (Aug. 2006), pp. 180–194. DOI: 10.1016/j.neuroimage.2006.02.051. URL: http://dx.doi.org/10.1016/j.neuroimage.2006.02.051.

[124] Zarrar Shehzad et al. "The Preprocessed Connectomes Project Quality Assessment Protocol: A resource for measuring the quality of MRI data". In: *Frontiers in Neuroscience* 47 (2015). ISSN: 1662-453X. DOI: 10.3389/conf.fnins.2015.91.00047. URL: http://www.frontiersin.org/10.3389/conf.fnins.2015.91.00047/event_abstract.

[125] Bharat B Biswal et al. "Toward discovery science of human brain function". In: *Proceedings of the National Academy of Sciences* 107.10 (2010), pp. 4734–4739. DOI: 10.1073/pnas.0911855107.

[126] Xi-Nian Zuo et al. "An open science resource for establishing reliability and reproducibility in functional connectomics". In: *Scientific Data* 1 (Dec. 2014), p. 140049. DOI: 10.1038/sdata.2014.49. URL: http://dx.doi.org/10.1038/sdata.2014.49.

[127] A Di Martino et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". In: *Molecular Psychiatry* 19.6 (June 2013), pp. 659–667. DOI: 10.1038/mp.2013.78. URL: http://dx.doi.org/10.1038/mp.2013.78.

[128] Brian B Avants, Nick Tustison, and Gang Song. "Advanced normalization tools (ANTS)". In: *Insight Journal* 2 (2009), pp. 1–35.

[129] Arno Klein and Jason Tourville. "101 Labeled Brain Images and a Consistent Human Cortical Labeling Protocol". In: *Frontiers in Neuroscience* 6 (2012). DOI: 10.3389/fnins.2012.00171. URL: http://dx.doi.org/10.3389/fnins.2012.00171.

[130] Daniel S. Marcus et al. "Human Connectome Project informatics: Quality control database services, and data visualization". In: *NeuroImage* 80 (Oct. 2013), pp. 202–219. DOI: 10.1016/j.neuroimage.2013.05.077. URL: http://dx.doi.org/10.1016/j.neuroimage.2013.05.077.

[131] S. M. Sunkin et al. “Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system”. In: *Nucleic Acids Research* 41.D1 (Nov. 2012), pp. D996–D1008. DOI: 10.1093/nar/gks1042. URL: http://dx.doi.org/10.1093/nar/gks1042.

[132] Samir Das et al. “LORIS: a web-based data management system for multi-center studies”. In: *Frontiers in Neuroinformatics* 5 (2012). DOI: 10.3389/fninf.2011.00037. URL: http://dx.doi.org/10.3389/fninf.2011.00037.

[133] Tarek Sherif et al. “BrainBrowser: distributed web-based neurological data visualization”. In: *Frontiers in Neuroinformatics* 8 (Jan. 2015), p. 89. DOI: 10.3389/fninf.2014.00089. URL: http://dx.doi.org/10.3389/fninf.2014.00089.

[134] Christopher R. Madan and Elizabeth A. Kensinger. “Cortical complexity as a measure of age-related brain atrophy”. In: *NeuroImage* 134 (2016), pp. 617–629. DOI: 10.1016/j.neuroimage.2016.04.029. URL: http://dx.doi.org/10.1016/j.neuroimage.2016.04.029.

[135] Christopher R. Madan and Elizabeth A. Kensinger. “Age-related differences in the structural complexity of subcortical and ventricular structures”. In: *Neurobiology of Aging* 50 (2017), pp. 87–95. DOI: 10.1016/j.neurobiolaging.2016.10.023. URL: http://dx.doi.org/10.1016/j.neurobiolaging.2016.10.023.

[136] Henrik R Nagel. “Scientific visualization versus information visualization”. In: *Workshop on state-of-the-art in scientific and parallel computing, Sweden*. Citeseer. 2006.

[137] D Flitney et al. “Anatomical brain atlases and their application in the FSLView visualisation tool”. In: *Thirteenth annual meeting of the Organization for Human Brain Mapping*. 2007.

[138] Mark Jenkinson et al. “Fsl”. In: *Neuroimage* 62.2 (2012), pp. 782–790.

[139] Ruopeng Wang et al. “Diffusion toolkit: a software package for diffusion imaging data processing and tractography”. In: *Proc Intl Soc Mag Reson Med*. Vol. 15. 3720. Berlin. 2007.

[140] C Rorden. *MRICron [computer software]*. 2007.

[141] Andriy Fedorov et al. “3D Slicer as an image computing platform for the Quantitative Imaging Network”. In: *Magnetic resonance imaging* 30.9 (2012), pp. 1323–1341.

[142] AD Gouws et al. “DataViewer3D: An open-source cross-platform multi-modal neuroimaging data visualization tool”. In: *NeuroImage* 47 (July 2009), S80. DOI: 10.1016/s1053-8119(09)70569-5. URL: https://doi.org/10.1016%2Fs1053-8119%2809%2970569-5.

[143] Anisha Keshavan, Arno Klein, and Ben Cipollini. “Interactive online brain shape visualization”. In: *Research Ideas and Outcomes* 3 (Feb. 2017), e12358. DOI: 10.3897/rio.3.e12358. URL: https://doi.org/10.3897/rio.3.e12358.

[144] Arno Klein et al. “Mindboggling morphometry of human brains”. In: *PLOS Computational Biology* 13.2 (Feb. 2017), pp. 1–40. DOI: 10.1371/journal.pcbi.1005350. URL: http://dx.doi.org/10.1371/journal.pcbi.1005350.

[145] A. Klein and J. Hirsch. “Mindboggle: a scatterbrained approach to automate brain labeling”. In: *Neuroimage* 24.2 (Jan. 2005), pp. 261–280.

[146] Arno Klein and Jason Tourville. “101 labeled brain images and a consistent human cortical labeling protocol”. In: *Frontiers in Neuroscience* 6.171 (2012). ISSN: 1662-453X. DOI: 10.3389/fnins.2012.00171. URL: http://www.frontiersin.org/brain_imaging_methods/10.3389/fnins.2012.00171/abstract.

[147] B Cipollini, H Bartsch, and G Cottrell. “Exploring the anatomy and genetics of cortical asymmetries in surface area and thickness.” In: *45th Annual Meeting of the Society for Neuroscience*. Chicago, 2015.

[148] Alfred Inselberg and Bernard Dimsdale. “Parallel coordinates: a tool for visualizing multidimensional geometry”. In: *Proceedings of the 1st conference on Visualization’90*. IEEE Computer Society Press. 1990, pp. 361–378.

[149] Paul Schmidt et al. “An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis”. In: *Neuroimage* 59.4 (2012), pp. 3774–3783.

[150] Eleftherios Garyfallidis et al. "Dipy a library for the analysis of diffusion MRI data". In: *Front. Neuroinform.* 8 (Feb. 2014). DOI: 10.3389/fninf.2014.00008. URL: http://dx.doi.org/10.3389/fninf.2014.00008.

[151] Anders M Dale, Bruce Fischl, and Martin I Sereno. "Cortical surface-based analysis: I. Segmentation and surface reconstruction". In: *Neuroimage* 9.2 (1999), pp. 179–194.

[152] Klein Arno et al. "Mindboggle 2: Automated human brain MRI feature extraction identification, shape analysis, and labeling". In: *Frontiers in Neuroinformatics* 8 (2014). DOI: 10.3389/conf.fninf.2014.08.00087. URL: http://dx.doi.org/10.3389/conf.fninf.2014.08.00087.

[153] Francesca Rossi et al. "Relevance of brain lesion location to cognition in relapsing multiple sclerosis". In: *PloS one* 7.11 (2012), e44826.

[154] Tian Ge et al. "Analysis of multiple sclerosis lesions via spatially varying coefficients". In: *The annals of applied statistics* 8.2 (2014), p. 1095.

[155] ZT Kincses et al. "Lesion probability mapping to explain clinical deficits and cognitive performance in multiple sclerosis". In: *Multiple Sclerosis Journal* 17.6 (2011), pp. 681–689.

[156] Z Khaleeli et al. "Localized grey matter damage in early primary progressive multiple sclerosis contributes to disability". In: *Neuroimage* 37.1 (2007), pp. 253–261.

[157] Katrin Morgen et al. "Evidence for a direct association between cortical atrophy and cognitive impairment in relapsing–remitting MS". In: *Neuroimage* 30.3 (2006), pp. 891–898.

[158] Kerstin Bendfeldt et al. "Spatiotemporal distribution pattern of white matter lesion volumes and their association with regional grey matter volume reductions in relapsing-remitting multiple sclerosis". In: *Human brain mapping* 31.10 (2010), pp. 1542–1555.

[159] JH Simon et al. "A Wallerian degeneration pattern in patients at risk for MS". In: *Neurology* 54.5 (2000), pp. 1155–1160.

[160] Jay N Giedd et al. "Brain development during childhood and adolescence: a longitudinal MRI study". In: *Nature neuroscience* 2.10 (1999), pp. 861–863.

[161] Rohit Bakshi et al. “MRI in multiple sclerosis: current status and future prospects”. In: *The Lancet Neurology* 7.7 (2008), pp. 615–625.

[162] Andrea Wiggins and Kevin Crowston. “From conservation to crowdsourcing: A typology of citizen science”. In: *System Sciences (HICSS), 2011 44th Hawaii international conference on*. IEEE. 2011, pp. 1–10.

[163] Justin Cranshaw and Aniket Kittur. “The polymath project: lessons from a successful online collaboration in mathematics”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 1865–1874.

[164] Chris J Lintott et al. “Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey”. In: *Monthly Notices of the Royal Astronomical Society* 389.3 (2008), pp. 1179–1189.

[165] Christopher B Eiben et al. “Increased Diels-Alderase activity through backbone remodeling guided by Foldit players”. In: *Nature biotechnology* 30.2 (2012), pp. 190–192.

[166] Jinseop S Kim et al. “Space-time wiring specificity supports direction selectivity in the retina”. In: *Nature* 509.7500 (2014), p. 331.

[167] Anisha Keshavan, Arno Klein, and Benjamin Cipollini. “Interactive online brain shape visualization”. In: *bioRxiv* (2016), p. 067678.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature .... Date ......6/6/17......................