2022

Statistical Methods For The Analysis And Development Of Quantitative Imaging Biomarkers

Carolyn Lou
University of Pennsylvania

Statistical Methods For The Analysis And Development Of Quantitative Imaging Biomarkers

Abstract

The field of neuroimaging statistics is concerned with elucidating meaningful conclusions from high-dimensional imaging objects, often in the form of single-dimensioned summary statistics. Ideally, these summaries should provide interpretable biomarker measurements that can guide patient diagnoses or treatment decisions while minimizing information loss associated with dimension reduction. This dissertation is focused on (1) exploring methods for analyzing previously developed imaging biomarkers and (2) developing new imaging biomarkers using both well-established and novel imaging analysis techniques. We approach this problem in three ways: in our first project, we assess how previously developed imaging biomarkers can best be incorporated into downstream analyses in the context of a clinical trial. This work conceptualizes imaging biomarkers as measurements which intrinsically contain historical information on a patient and examines the effect of incorporating these predictors on the statistical power in a clinical trial analysis. For our second project, we develop a radiomic predictor that automatically identifies an important prognostic biomarker in multiple sclerosis, relying on quantification of imaging patterns potentially associated with brain atrophy and more severe disease courses. In our third project, we construct a coordinate system and framework for multiple sclerosis lesions analyses for more sensitive and specific biomarker development. We use dimension reduction and flexible nonparametric modelling to assess the diagnostic value of this method. These methods lay the groundwork for improving future work developing and utilizing imaging biomarkers with imaging statistics.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Russell T. Shinohara

Keywords

Machine Learning, Neuroimaging, Statistics

Subject Categories

Biostatistics

STATISTICAL METHODS FOR THE ANALYSIS AND DEVELOPMENT OF QUANTITATIVE IMAGING BIOMARKERS

Carolyn E. Lou

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Russell T. Shinohara

Associate Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra

Professor of Biostatistics

Dissertation Committee

Nandita Mitra, Professor of Biostatistics

Kristin A. Linn, Assistant Professor of Biostatistics

Pascal Sati, Associate Professor of Neurology

STATISTICAL METHODS FOR THE ANALYSIS AND DEVELOPMENT OF QUANTITATIVE IMAGING BIOMARKERS

ACKNOWLEDGMENT

This dissertation would not be complete without the guidance and support from many, many people. I’ve cried a lot, I’ve laughed a lot, I’ve learned a lot, and I’m so grateful for the community at Penn and the people I’ve met along the way.

First and foremost, I have to thank Taki Shinohara, my dissertation mentor. Taki is a brilliant statistician from whom I’ve had the honor of learning not only as a part of research but also through three distinct courses, covering topics ranging from key modern statistical techniques to professionalism as a modern statistician. Taki is a formidable leader of the PennSIVE center, and he has been instrumental in molding PennSIVE into the supportive and fun environment that it is. I’m so grateful that I got to be a part of this lab.

To my committee members Nandita, Kristin, and Pascal: Thank you for sharing your time and guidance. You have been able to provide guidance and ask interesting questions no matter the project, and I’m grateful for your insights.

To Eli Elliott and Cathy Vallejo: You have been my first point of contact for all of my biggest moments in graduate school, from booking my first flight out to Philadelphia when I interviewed to helping me with registering for my first conference to helping me stay on track with graduation deadlines. Thank you for your support.

To PennSIVE members, both past and present: Thank you for sharing your time and company, sitting through countless practice talks that have made each presentation thereafter stronger. You’re great company at conferences and your work is consistently inspiring.

To my cohort: Thanks for being some of my best friends these past 5 years. I’m so excited that half of us will be in New York after this and we can continue to get together and have fun.

To my family, who I usually only get to see over FaceTime: Thank you for your unconditional love and support, even from very far away.

And finally, to Nick, my fiancé: Thank you for being my best friend. I’m so lucky to get to spend every day with you. This dissertation would not have been completed without you.

ABSTRACT

STATISTICAL METHODS FOR THE ANALYSIS AND DEVELOPMENT OF QUANTITATIVE IMAGING BIOMARKERS

Carolyn E. Lou

Russell T. Shinohara

The field of neuroimaging statistics is concerned with elucidating meaningful conclusions from high-dimensional imaging objects, often in the form of single-dimensioned summary statistics. Ideally, these summaries should provide interpretable biomarker measurements that can guide patient diagnoses or treatment decisions while minimizing information loss associated with dimension reduction. This dissertation is focused on (1) exploring methods for analyzing previously developed imaging biomarkers and (2) developing new imaging biomarkers using both well-established and novel imaging analysis techniques. We approach this problem in three ways: in our first project, we assess how previously developed imaging biomarkers can best be incorporated into downstream analyses in the context of a clinical trial. This work conceptualizes imaging biomarkers as measurements which intrinsically contain historical information on a patient and examines the effect of incorporating these predictors on the statistical power in a clinical trial analysis. For our second project, we develop a radiomic predictor that automatically identifies an important prognostic biomarker in multiple sclerosis, relying on quantification of imaging patterns potentially associated with brain atrophy and more severe disease courses. In our third project, we construct a coordinate system and framework for multiple sclerosis lesions analyses for more sensitive and specific biomarker development. We use dimension reduction and flexible nonparametric modelling to assess the diagnostic value of this method. These methods lay the groundwork for improving future work developing and utilizing imaging biomarkers with imaging statistics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1:

INTRODUCTION

Neuroimaging data are more accessible and easily collected than ever, with heavy financial support from federal governments leading to an abundance of data in the form of brain images to be analyzed (Sporns, Tononi and Kötter, 2005; Insel, Landis and Collins, 2013; Jiang, 2013; Amunts, 2014). Despite growing access to computational resources, the complex nature of these data makes it difficult to draw conclusions from unprocessed images. Occasionally, large studies have resulted in spurious connections that have not held up under further scrutiny (Eklund, Nichols and Knutsson, 2016; Cremers, Wager and Yarkoni, 2017; Vandekar and Stephens, 2021). Statistical analyses using these data must be conducted carefully to avoid such results. Tools which can summarize large imaging data into a single-dimensioned, easily-interpretable metrics can aid in improving our ability to ascertain clinical insights and understand neurological phenomena.

In this dissertation, we aim to assess the utility and development of these single-dimensioned metrics, which we call radiomic predictors or imaging biomarkers. We first explore how incorporating imaging biomarkers into current and future clinical trial studies may be used to enhance their statistical power. Additionally, we develop novel imaging biomarkers that can be used to summarize complex imaging patterns present in magnetic resonance images (MRI) of multiple sclerosis (MS) lesions. These resulting biomarkers could potentially be used as a clinical tool in future clinical trials.

Traditional methods for estimating sample size requirements in a clinical trial rely on group mean data and population average treatment effects (Donner, 1984; Kirby, Gebski and Keech). Many historical control methods rely on group-level analyses such as pooling or Bayesian modelling (Viele et al., 2014). In Chapter 2, we introduce the concept of using individualized evaluation of treatment effects with neuroimaging biomarkers and provide a framework for practically incorporating this approach into future clinical trials of neurologic disease when baseline imaging is available (Lou et al., 2021a). We show that machine learning tools can provide individualized predictions for patients under study, which in turn can be used to inform sample size calculations with individualized estimates of clinical outcome in a trial. This methodology can substantially improve statistical power for detecting a treatment effect.

In Chapters 3 and 4, we shift our focus to development of imaging biomarkers that quantify potentially important prognostic and diagnostic imaging signals in MS lesions. As a demyelinating and inflammatory disorder, MS often manifests with lesions in the brain and spinal cord that can be detected *in vivo* with MRI (Sahraian and Radü, 2007). Imaging biomarkers such as total lesion volume and lesion count are commonly used both for diagnosis and for tracking disease progression, though recent studies suggest that misdiagnosis of MS is not uncommon, potentially due to the non-specificity of white-matter lesion presentation (Solomon et al., 2016; Thompson et al., 2018; Gaitán and Correale, 2019). New imaging signals such as the central vein sign and the paramagnetic rim signal provide possible avenues down which to more specifically characterize MS lesions while potentially assessing prognosis as well (Sati et al., 2016; Absinta et al., 2019; Maggi et al., 2020b).

In Chapter 3, we develop an automated tool for quantification of the paramagnetic rim signal, most notably shown be associated with greater disease burden and more severe tissue damage (Absinta et al., 2016; Tozlu et al., 2021). Manual inspection of MS lesion for the presence of a paramagnetic rim is time consuming and prone to inter- and intra-rater variability. We propose an automated method for identifying PRLs that would improve efficiency of study and facilitate translation of this biomarker into larger research studies and clinical practice. We use high-dimensional radiomic feature extraction along with a random forest classification model, which can flexibly model high dimensional data, to identify PRLs.

In Chapter 4, we develop our own method for quantifying imaging patterns. Quantitative radiomic analysis is a powerful tool for the analysis of focal MRI lesions but does not typically incorporate the spatial location of nor spatial patterns within a lesion. We leverage detection of known imaging biomarkers and estimation of sublesions with multimodal imaging, particularly relevant in the context of confluent clusters of lesions, to define a common coordinate system for all white matter lesions. We use dimension reduction to then assess the added value of our method by examining association with clinical outcomes.

CHAPTER 2:

LEVERAGING MACHINE LEARNING IMAGING BIOMARKERS TO AUGMENT STATISTICAL POWER IN CLINICAL TRIALS

2.1. Introduction

The power of a clinical trial is the probability of detecting a statistically significant difference between treatment groups under a set of assumptions. Power increases as the magnitude of the true difference in outcomes between treatment groups increases, as the accuracy of measurement for the outcome measure increases, and as sample size increases (Faul et al., 2007). When a treatment effect exists, failure to detect a statistically significant difference between treatment groups can occur as the result of a myriad of reasons, including small treatment effect, poor measurement of the primary outcome, inadequate sample size (underpowered studies), or treatment effect heterogeneity (Wittes, 2002; Anderson et al., 2017; Kent et al., 2019; Rekkas et al., 2020). Failure is more likely in studies of relatively rare neurologic diseases including glioblastoma multiforme (GBM) because enrolling large samples is difficult, but failure also occurs in more common diseases with substantial biological heterogeneity and unstable outcome measures, such as Alzheimer's disease (Cummings, 2018; McGranahan et al., 2019; Oxford, Stewart and Rohn, 2020).

Despite many long and expensive trials, no disease modifying drug for Alzheimer's disease has been approved (Petersen et al., 2010). Phase III trials for GBM have been more successful, but treatment efficacy has been modest, with an improvement in median survival of only 7 months (8–15 months) for patients in the treatment arms in 44 different trials (Anderson et al., 2008; Menze et al., 2015). Similar explanations have been proposed for the failure of trials of these two diseases, including biological heterogeneity, selection of ineffective treatments based on incomplete understanding of disease biology, starting treatment too late in disease development, incorrect drug doses, and unreliability of the primary outcome measurements (Davatzikos et al., 2011; Shaffer et al., 2013). All of these explanations may contribute to a reduction in the magnitude of the treatment effect. If the expected treatment benefit is overestimated, the study will be underpowered.

Traditionally, trials rely on empirical data from previously conducted studies (often phase II trials, if available) to estimate sample size requirements to achieve a particular level of power (i.e., 80%, 90%). These traditional methods for estimating sample size requirements rely on group mean data and calculating sample size requirements based on population average treatment effects. When historical control data are used, statisticians use methods such as pooling or Bayesian modeling, which also rely on group-level analyses (Pocock, 1976; Viele et al., 2014). Newer high-dimensional predictors such as neuroimaging or genomic data offer the opportunity to include individualized predictions. This allows for a more precise evaluation of the treatment effect for each person through comparison of their observed outcome with their predicted outcome, rather than relying on a group-level effect that determines average outcome. Because of this more precise evaluation, the residual variance decreases and thus the power to detect this treatment effect increases.

In this study, we introduce the concept of using individualized evaluation of treatment effects with neuroimaging biomarkers and provide a framework for practically incorporating this approach into future clinical trials of neurologic disease when baseline imaging is available. We show that machine learning tools can provide individualized predictions for patients with Alzheimer’s disease and GBM, which in turn can be used to inform sample size calculations with individualized estimates of clinical outcome in a trial. This methodology can substantially improve statistical power for detecting treatment effects, or alternatively, reduce the sample size needed to achieve the same power in a clinical trial.

2.2. Materials and methods

Our method relies on access to two sets of data: i) a current clinical trial designed to study an outcome of interest and ii) a previously observed cohort of similar subjects treated according to the current standard of care with data on the outcome of interest. We narrow our focus in this work to imaging biomarkers and associated studies, so we assume that imaging data has been gathered at study enrollment for both sets of trials. In both of our disease applications, imaging data are regularly obtained through standard course of care, either for exclusion of other pathologies or for diagnosis itself. The techniques proposed here are also directly applicable to other -omic modeling scenarios, and generally, to any predictive marker of standard of care outcome.

We aim to show that previously developed and validated radiomic prediction models, which summarize imaging patterns that predict future clinical outcomes of interest, can in some cases result in improved statistical power for detecting treatment effect (Fig. 1). These outcomes of interest can be endpoints such as response to treatment, patient survival, or progression-free survival. The model, which is built based on a historical cohort, can then be used in conjunction with data collected from the current trial to generate individualized values of the radiomic score for each of the current participants. These individualized scores represent predicted values for the outcomes of the treated individuals in the current trial had they instead been assigned to the control group. The incorporation of these predicted values as a covariate in the final analysis of the current trial lends power to the detection of the effect of a treatment by modeling the inter-subject variability in the outcome in terms of baseline heterogeneity represented in the baseline imaging.

In practice, this could be done by using a model developed for a previously validated radiomic predictor, applying it to data from a current trial of interest, and then incorporating the newly derived values of that radiomic predictor as a covariate in the study analysis. This would reduce uncertainty in the estimate of the overall treatment effect and therefore increase statistical power to detect a treatment effect.

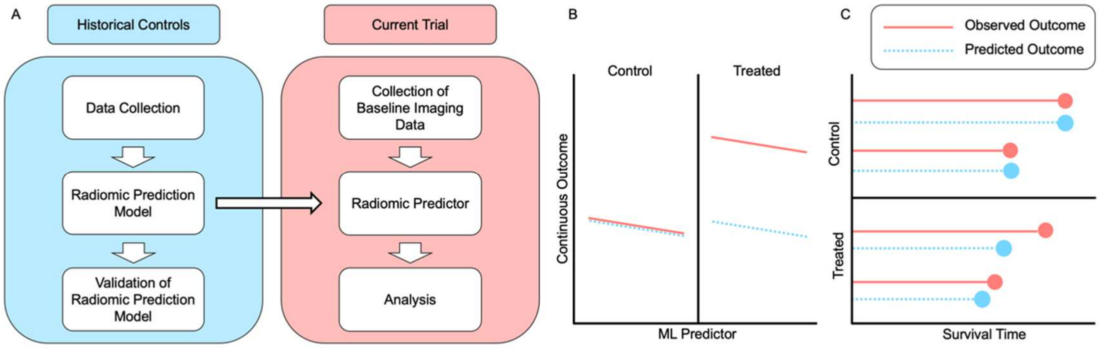


Figure 2.1: Method Visualization and Description. *A: Workflow for implementing the proposed method in a new clinical trial. B (continuous) and C (survival outcome): Schematic diagram for individualized predictions that are generated for each person in the current trial, where the solid red lines indicate observed outcome for the participants of the current trial and the dashed blue lines indicate predicted outcome for those participants had they not been treated. Fig. 1B illustrates the method for continuous outcomes, where the left side represents the outcomes of those randomized to the control arm and the right side represents the outcomes of treated participants. The predicted outcome values (dashed blue lines) for the control units had they not been treated would be exactly what they are observed to be (solid red lines), while the predicted outcome values (dashed blue lines) for the treated units had they not been treated are different from the observed outcome (solid red lines). Fig. 1C illustrates the analogous mechanism for survival outcomes, where the predicted survival times for the control units (dashed blue lines) are the same as the observed survival times (solid red lines), whereas the predicted survival for the treated individuals are lower than the observed survival times. Our method capitalizes on these differences to augment statistical power.*

To investigate the advantage of this approach, we implemented our models in two scenarios motivated by two different disease areas (GBM and AD). To better approximate real-life clinical trial performance, we use radiomic and outcome data from two observational studies to generate hypothetical study data, where the first focuses on the continuous outcomes of cognitive decline in prodromal Alzheimer's disease and the second on the survival after diagnosis with GBM. With these studies, we performed plasmode simulations, where we randomly split our observational data into a theoretical historical cohort and a theoretical trial cohort. We then simulate effects in a randomly selected subset, corresponding to one arm, of the trial cohort. We then compare the statistical power of our proposed approach with the classical modeling approach that does not include radiomic prediction-based modeling.

We also performed simulations with fully synthetic data, where the populations were generated to be homogeneous except for random error and treatment status. Code for both the synthetic data simulations and plasmode simulations are available on our GitHub (https://github.com/carolynlou/hcct).

2.2.1. Data

For our analyses, we relied only on observational data. These data were obtained from ADNI and the University of Pennsylvania for our AD and GBM studies respectively (Petersen et al., 2010; Macyszyn et al., 2016). There was no missingness in either of the two datasets, and no patients dropped out before baseline imaging data could be collected. Our studies of performance in a clinical trial setting were based on plasmode simulations, where we artificially generated hypothetical trial data from our observational data.

In our first case study, we focused on therapeutic trials for prevention of AD, in which the primary outcome is typically longitudinal cognitive change. Here, we simplified this outcome and quantified cognitive change as the difference between memory score measured 2 years from baseline and memory score measured at baseline. We used a predictive model, called the SPARE-AD score, which has been previously derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.usc.edu) on 283 subjects with mild cognitive impairment (MCI) who underwent serial MRIs at 1.5T (Davatzikos et al., 2009; Petersen et al., 2010). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.

SPARE-AD is derived from patterns of regional brain atrophy (volume loss) captured by atlas warping methods and high-dimensional pattern classification using support vector machines (SVM) aiming to differentiate cognitively normal (CN) and Alzheimer’s disease subjects (Fan et al., 2008; Davatzikos et al., 2009; Da et al., 2013). We used cognitive decline as our outcome here, as measured by 2-year change from baseline values of the ADNI composite memory score ADNI-MEM (Crane et al., 2012). Average 2-year change from baseline for ADNI-MEM in the current study was -0.17 (standard deviation 0.49). The average age of the participants was 74.8 years (sd = 7.32), and 99 (35%) of participants were female. A more detailed description of demographics and clinical characteristics of patients has been published previously (Da et al., 2013). We note that both the outcome and the disease status of the subjects studied here differ from the outcome and disease statuses that were used to build SPARE-AD. As a complementary analysis, we also examined conversion from mild cognitive impairment (MCI) to Alzheimer's disease as an outcome, employing time-to-event analysis methods, where we again used the SPARE-AD score as a radiomic predictor of interest. Approximately 60% of observations were censored.

As a second case study, we focused on trials for GBM therapies in which the primary outcome is overall survival time after diagnosis. We analyzed previously collected, anonymized data from 134 patients who were treated for newly diagnosed GBM at the Hospital of the University of Pennsylvania between 2006 and 2013. The median survival in this sample was 12 months, and survival data were assessed for all subjects with no loss to follow-up. The average age of patients in this study was 62.1 years (sd = 12.1), and 53 patients (40%) were female. Detailed demographics and a clinical description of these subjects have been previously published (Macyszyn et al., 2016). All people with access to the data were on an institutional IRB. For this second case study, we investigated the use of cross-validated predictions of survival time based on radiomic analyses of pre- and post-contrast T1-weighted, T2-weighted and T2- fluid attenuated inversion recovery (T2-FLAIR), diffusion, and perfusion MRI acquired pre-operatively at diagnosis. This GBM predictive model utilized an SVM model to differentiate short, medium, and long survival (Macyszyn et al., 2016).

2.2.2. Statistical Methods

All hypothesis testing was conducted assuming a 5% type 1 error rate and using two-sided alternatives. In lieu of data from a clinical trial, to explore the utility of our method, we employed plasmode simulation studies, in which all of our analyses were performed with datasets derived from ADNI and our GBM cohort, but we artificially generated treatment status and treatment effect. We also explored the method with synthetic data simulations (Fig. 3), in which we artificially generated a theoretical radiomic predictor, a binary treatment indicator, and an outcome.

Alzheimer's Disease Study. For our Alzheimer's disease plasmode simulation study, we used a continuous outcome and analyzed our data with linear regression. To compare our method to a more classical analysis, we fit the following two models, where Equation (1) represents our method and Equation (2) represents a classical analysis:

$$(1)\ Y_i = \alpha + \beta X_i + \gamma A_i + \epsilon_i$$

$$(2)\ Y_i = \alpha + \gamma A_i + \epsilon_i$$

Here, $\mathrm{Y_I}$ represents cognitive decline, defined as the difference between ADNI-MEM score observed at 2 years after baseline and ADNI-MEM score observed at baseline. X_i represents the radiomic predictor, A_i represents the treatment indicator, and ϵ_i represents random error. The parameters α and β are estimated from the data while γ is added in artificially, as described below. We note that these models are equivalent to ANOVA-CHANGE models as described in O'Connell et al (O'Connell et al., 2017).

To conduct the simulation, we randomly split our data into two equal portions, one representing the source of a treated population and one representing the source of a control population. We then generated a sample treated arm and a sample control arm that we used for downstream analysis by sampling with replacement from the respective source samples. For the first group, indexed by $i = 1, \ldots, \frac{n}{2}$, we set our treatment indicator $A_i = 0$ and record the observed outcome Y_i, as well as the value of the radiomic predictor X_i at baseline. For the second group, indexed by $i = \frac{n}{2} + 1, \ldots, n$, we introduced a treatment effect γ, set our treatment indicator $A_i = 1$, and again record outcome Y_i and baseline radiomic predictor measurement X_i.

We repeated this process 1000 times, recording the p-value corresponding to the test for treatment effect each time. We calculated type 1 error rate and power as the percentage of times the treatment effect was significant at the $\alpha = 0.05$ level, where γ is set to 0 to assess type 1 error and a non-zero value to assess power. In order to quantify the sample size benefits from using this method, we repeated the above procedure for a range of sample sizes n and recorded the smallest n for which power reached 80%. We explored this for a range of hypothetical effect sizes, which was defined as γ divided by the standard deviation of the outcome Y_i.

We also performed a similar analysis with a time-to-event outcome, studying the time to conversion from MCI to Alzheimer's disease measured in months. We analyzed this outcome with the following accelerated failure time models, assuming a log-logistic distribution:

$$(1)\ Y_i = \log(T_i) = \alpha + \beta X_i + \gamma A_i + \sigma\epsilon_i$$

$$(2)\ Y_i = \log(T_i) = \alpha + \gamma A_i + \sigma\epsilon_i$$

Here, X_i represents the radiomic predictor, and A_i represents the binary treatment indicator. We introduce a multiplicative treatment effect on observed survival or censoring time in the treatment group and refer to this multiplier as the effect size. In order to mimic a 3-year clinical trial, we introduce end-of-study censoring at 36 months. We conducted the simulation study as described previously, assessing sample size benefits as the minimum number of participants for the study.

Glioblastoma Multiforme Study. For our GBM plasmode simulation, we used survival outcomes and we assessed differences between treatment groups with and without adjustment for the radiomic prediction by assuming an accelerated failure time model. Specifically, we fit the following models:

$$(1)\ Y_i = \log(T_i) = \alpha + \beta X_i + \gamma A_i + \sigma\epsilon_i$$

$$(2)\ Y_i = \log(T_i) = \alpha + \gamma A_i + \sigma\epsilon_i$$

Here, Y_i is $\log(T_i)$, where T_i represents the time to event, X_i represents the radiomic predictor, A_i represents the treatment indicator, and ϵ_i represents random error. In this study, we modelled T_i using a log-logistic accelerated failure time model. We introduce a treatment effect and conduct the simulation study for this setting as described previously.

Synthetic Data Simulation Study. We start by discussing the continuous outcome case. We simulated data according to the following parametric form: $Y_i = 2 + 4X_i + \gamma A_i + \epsilon_i$, where X_i ~N(0, 0.25) denotes a continuous predictor, A_i is a binary treatment indicator simulated at random with $P(A_i = 1) = 0.5$, and $\epsilon_i \sim N(0,1)$ is a random error term for $i = 1, \ldots, n$. We then introduced a treatment effect of γ for these subjects and assessed the significance of the treatment effect with a Wald test via linear regression. We repeated this process 1000 times, recording the p-value corresponding to the test for treatment effect each time. We calculated type 1 error and power as the percentage of iterations in which the treatment effect was significant at the $\alpha = 0.05$ level, where we set γ to 0 to assess type 1 error and a non-zero value to assess power. In order to quantify the sample size benefits from using this method, we repeated the above procedure for a range of sample sizes n, corresponding to the total trial size, and recorded the smallest n for which power reaches 80%. We explored this for a range of hypothetical effect sizes, which we defined here as γ divided by the standard deviation of the outcome.

For the time-to-event outcome case, we followed a similar procedure. Here, the outcome Y_i was simulated according to a Weibull distribution such that $Y_i = \log(T_i) = 0.5 + 0.25X_i + 4\epsilon_i$, where T_i represents the time to event, $X_i \sim N(0, 1)$ denotes a continuous predictor, and ϵ_i is a random error term following an extreme value distribution with scale parameter of 4 for $i = 1, \ldots, n$, with total trial size n. Then, we introduced a treatment effect of γ with probability $P(A_i = 1) = 0.5$. We then used an accelerated failure time model to regress Y against A, testing for the treatment effect with a Wald test. We introduced end-of-study censoring at 36 months to mimic a 3-year clinical trial. We assessed power, type 1 error, and sample size benefits as in the continuous outcome case.

2.2.3. Data availability

The Alzheimer's disease data used for this study are publicly available and were obtained from the ADNI database (http://adni.loni.ucla.edu/). Data collected for this study was approved under institutional review board protocol #825722 sponsored by the National Institutes of Health. The GBM data have been uploaded to TCIA, and should be available to the public shortly. For the purposes of the review process, the data are available as Supporting Information. Data for this study was collected under institutional review board-approved protocol #706564 sponsored by the National Institutes of Health.

2.3 Results

For both continuous and time-to-event outcomes, the proposed method consistently reduced the minimum required sample size n for a given level of power in clinical trial analyses (Fig. 2). In the Alzheimer's disease plasmode simulations, where the outcome of interest is cognitive decline, with an effect size of 0.35, the total required sample size was 246 for the conventional analysis and 212 with the proposed historical control analysis. As the effect size increased, sample size requirements decreased for both approaches but decreased more rapidly for the conventional approach. In the GBM plasmode simulations, where the outcome of interest was survival time, at an effect size of 1.65, the total required sample size was 128 with the conventional analysis and 74 with the proposed historical control analysis.

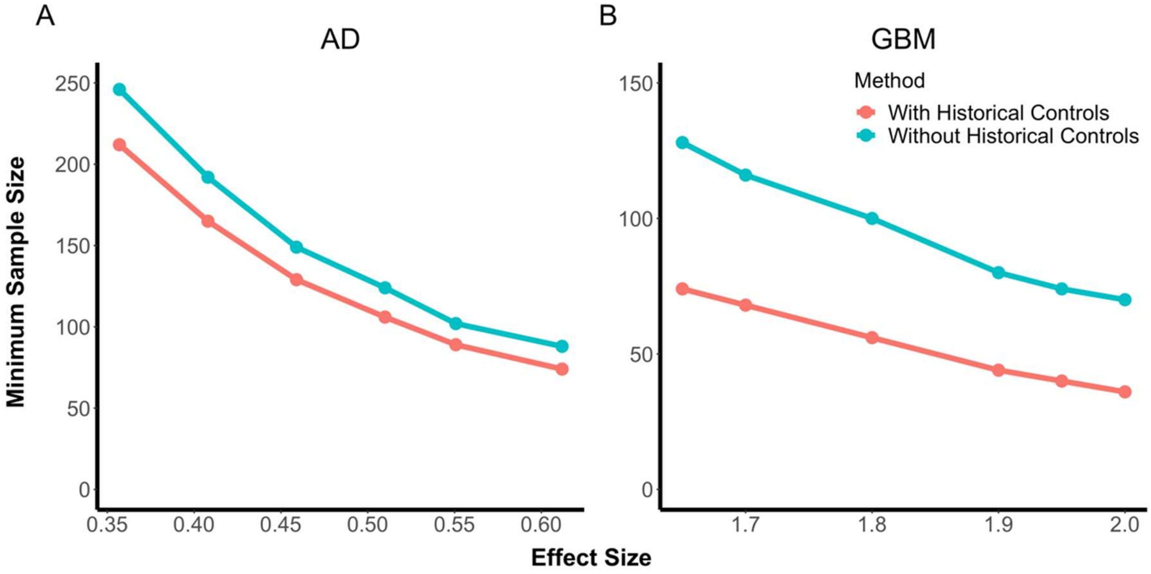


Figure 2.2: Plasmode simulation results. *Results from simulated studies under two scenarios. With the addition of historical controls, the minimum required sample size for 80% power is markedly lower than using classical two-sample clinical trial analysis. These figures show minimum sample size (vertical axes) required to achieve 80% power for a range of effect sizes (horizontal axes) based on observed outcome and radiomic predictions. Fig. 2A shows the results from simulations for continuous outcome measures of cognition in our Alzheimer's cohort from ADNI, analyzed using a linear regression model with and without incorporation of the radiomic predictor (left). Fig. 2B shows the results from simulations for survival in our glioblastoma cohort, comprised of 134 patients who were treated for newly diagnosed GBM at the Hospital of the University of Pennsylvania between 2006 and 2013 and analyzed with an accelerated failure time model with and without incorporation of the radiomic predictor. Note that the proposed method that leverages historical controls to build radiomic predictions (red) requires lower samples sizes than the classical approach (blue). Minimum required sample size was calculated as the smallest sample size that achieved 80% power as calculated by the percentage of Monte Carlo simulations with a non-zero treatment effect that were significant at the* α*=0.05 level.*

Reductions in sample size requirements were greater for smaller effect sizes, so the benefit of historical controls declined as effect size increased. Type 1 error remained controlled throughout all experiments conducted. In the Alzheimer's disease study with a continuous outcome, our proposed method resulted in a 14-16% decrease in the minimum required sample size. In the GBM study, our method reduced the required sample size by as much as 48%. Our simulations with fully synthetic data supported these findings.

Table 2.1 summarizes the effect of using our method on sample size across a range of power levels and effect sizes.

Table 2.1: Minimum Required Sample Size for Different Powers and Effect Sizes. *We provide the minimum required sample size for both 80% and 90% power across a range of effect sizes in both our ADNI*

cohort and our cohort of patients with glioblastoma multiforme (GBM). For our ADNI dataset, because the radiomic predictor of interest is known to be an accurate predictor of MCI to Alzheimer's disease conversion time, we also explored the utility of incorporating this method into a survival analysis.

Cohort-Outcome	**Power**	**Effect Size**	**Minimum Sample Size**	
			With Historical Controls	**Without Historical Controls**
ADNI-Continuous	0.8	0.40	174	200
		0.46	146	170
		0.61	74	88
	0.9	0.40	228	263
		0.46	172	204
		0.61	97	112
ADNI-Survival	0.8	1.7	242	287
		1.8	206	254
		1.9	180	219
	0.9	1.7	—	—
		1.8	271	332
		1.9	238	292
GBM-Survival	0.8	1.8	56	98
		1.9	44	82
		2	38	70
	0.9	1.8	74	130
		1.9	60	108
		2	50	90

We also explored our method with synthetic data simulations. Results for these simulations were similar to those from the plasmode simulation studies. With the synthetic data simulations, we noticed even greater sample size gains with the use of our proposed methodology. In settings with continuous outcomes, at the smallest effect size that we studied of 0.4, the total required sample size was 380 when using the conventional analysis and 230 when properly incorporating information from historical controls. As in the plasmode studies, this reduction became less pronounced as the effect size increased. This is partly due to a more rapid decrease in required sample size under the conventional approach than for the proposed method. In the time-to-event outcome simulations, at the smallest effect size studied of 1.1, the total required sample size was 540 with the conventional analysis and 310 with the proposed historical control analysis.

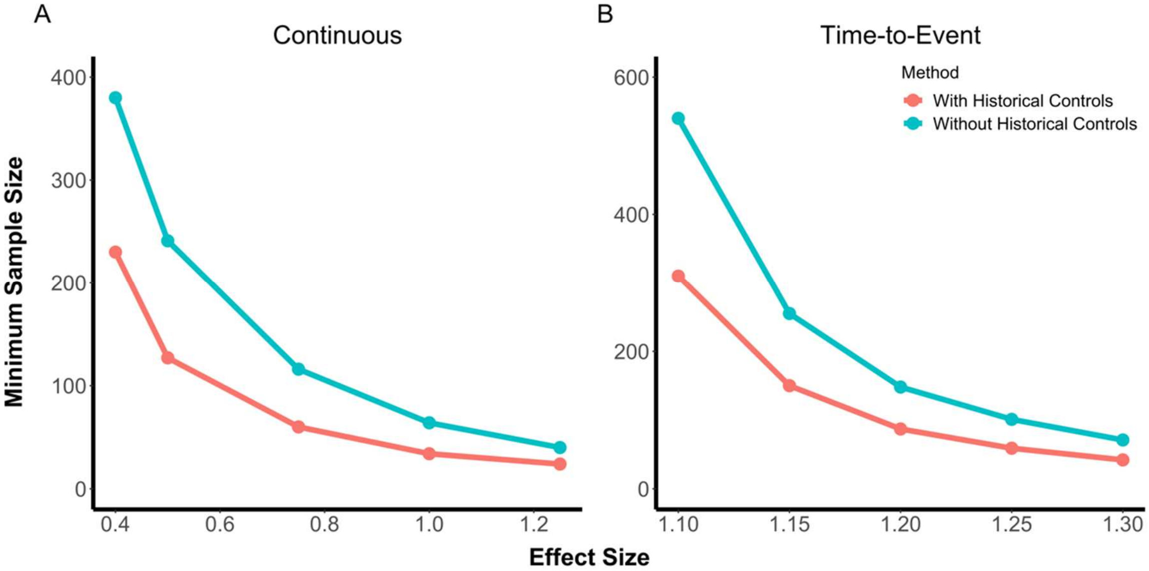


Figure 2.3: Synthetic data simulation results. *Results from simulated studies with synthetic data generated to be homogenous across all cohorts except for random error and treatment status. These figures show the minimum sample size required to achieve 80% power for a range of effect sizes, where minimum sample size was calculated as the smallest sample size for which at least 80% of Monte Carlo simulations with a non-zero treatment effect were significant at the α=0.05 level. Fig. 3A shows the results for simulations with a continuous outcome, analyzed using linear regression with and without incorporation of the radiomic predictor, and Fig. 3B shows the results for simulations with a survival outcome, analyzed using an accelerated failure time model with and without incorporation of the radiomic predictor. In both cases, the proposed method that leverages historical controls in the form of radiomic predictions (red) requires lower sample sizes than the classical approach (blue).*

2.4. Discussion

We have shown that individualized machine-learning-based imaging biomarkers can be useful tools in clinical trial analyses when the necessary information is available, offering decreased sample size requirements for a given effect size. The novelty of this method arises from the incorporation of individualized predictions based on powerful predictive algorithms which lend power to the detection of an average treatment effect due to targeting of the individuals in a given clinical trial. As robust neuroimaging biomarkers derived via machine learning models become more available, the historical datasets that can be analyzed with those models grow in size.

These changes are expected to strengthen the radiomic prediction models like the ones used in this study.

Our incorporation of the radiomic predictor relies on the existence of previously developed imaging biomarkers, which for the purposes of this paper, we theorize as having been trained on a historical cohort. Because the radiomic predictor has been previously trained on a historical cohort, the models that we fit to analyze a current trial inherently incorporate information from historical controls. When we do not include the radiomic predictor into the clinical trial analysis, we do not incorporate information from historical controls. We generate values of the radiomic predictor for the current sample using the model that was developed with a previous cohort. For our simulations, we assume that the current trial is being run on an experimental drug and the goal is to show superiority (O'Connell et al., 2017). Patients enrolled in a clinical trial or a cohort study may not be representative of patients in a population of interest. Differences between these populations are the result of the explicit inclusion/exclusion criteria of a clinical trial as well as the indirect differences between patients who are willing to volunteer for a clinical trial and those who are not (Jordan et al., 2013). Event rates can also be higher in a cohort study than in a clinical trial, potentially due to these same biases. However, randomization of the current trial participants ensures that even if the historical control population is different from the trial population in important ways we would not realize inflated type 1 error. The differences may however impact the predictive performance of the radiomic predictor for the outcome of interest, which could thus impact statistical power of the proposed methodology and attenuate the sample size benefits.

In the event that a primary analysis of an endpoint does not yield statistically significant results, this technique could potentially be used in a sensitivity analysis to aid interpretation. Performance of a secondary analysis looking to draw conclusions about the efficacy of a treatment could result in increases of type 1 error and thus spurious decisions, but incorporation of this technique into an exploratory analysis aimed at characterizing the impact of baseline heterogeneity on inference could illuminate important phenomena that would otherwise be missed.

We note that unexplained biological heterogeneity among the cohorts under study may have attenuated the power gains that were observed. To assess the potential gains of using this method in cases where the degree of biological heterogeneity explained was modifiable, we conducted simulation studies with data generated so as to be homogenous except for random error and treatment status (Fig. 3). In that setting, power gains and subsequent sample size reductions were much more dramatic.

Here, we used two previously developed biomarkers, one of which was trained to classify an outcome different from the target of the clinical trial analysis, and the other of which was trained to classify the same outcome as the clinical trial analysis. While both predictors offered gains in sample size reduction, the predictor built specifically for the outcome of interest in the clinical trial performed better and offered more substantial gains. We expect that the gains in power will likely be larger when the model is trained to predict the primary outcome of the clinical trial, though this needs to be empirically tested across a range of applications.

This approach is not limited to imaging biomarkers. It can be applied to a broad array of factors associated with the outcomes of interest in a clinical trial, such as clinical variables, blood or cerebrospinal fluid-based biomarkers, or genomic markers. The choice between use of biomarker-only prediction models as opposed to clinical and biomarker prediction models can be decided on a case-by-case basis, depending on the hypothesis of interest in a given study. In general, the more robust the associations among the predictors and the outcome of interest, the greater the anticipated gains in power or reduction in sample size required for a specified level of power.

The approach proposed in this paper has some limitations. First, the use of radiomic predictions can be hindered by the cost of collecting imaging data (Fleiss, 2011). Though many modern clinical trials of neurologic disease now incorporate baseline MR imaging into their protocol (Hammoud et al., 1996; Perry et al., 2017; Honig et al., 2018; Herrlinger et al., 2019; Egan et al., 2019; Wirsching et al., 2021; Mintun et al., 2021), especially in those of AD and GBM, which we use as examples in this manuscript, imaging remains expensive and potentially extraneous for the study of certain diseases. This method is only helpful when baseline imaging is available, and in the absence of baseline imaging for participants of a new study, incorporation of imaging biomarkers through prior scans from unknown length of time prior to the new study may result in attenuated benefits of statistical power and imperfect characterization of disease load. This can also impact the accuracy of the sample size calculation in that incorporation of data that does not reflect true baseline heterogeneity at the beginning of a new study can increase uncertainty in the analysis.

Furthermore, reductions in sample size requirement depend upon the strength of the prediction model. In the current study, both imaging biomarkers considered were built based on SVMs, but other machine learning techniques such as deep convolutional neural networks may provide more predictive power (Davatzikos, 2019). In addition, gains in power for the primary outcome will be associated with gains in power for secondary outcomes only to the extent that predictions from the prediction model are associated with the secondary outcomes. This will likely be determined by the degree of correlation between the primary outcomes and a set of secondary outcomes. In principle, within a single trial, separate prediction models could be developed for two or more co-primary endpoints. Incorporation of this method into randomized trials with more complex designs, such as one incorporating stratification by confounders or a one-arm trial, requires further statistical research.

Finally, if a radiomic predictor is trained on data sampled from a different population than that which is studied in the current trial, the improvements in statistical power may be less pronounced. However, due to the randomization in the study, the type 1 error rate is expected to be maintained and internal validation or calibration of the predictive model is possible using data from the control arm of a clinical trial.

The key conclusion arising from our study is that machine-learning-based predictive models can be used to effectively improve the statistical power of clinical trials by leveraging the wealth of information available in neuroimaging data to generate personalized predictions of outcome. Imaging biomarkers are seldom incorporated into clinical trial analyses, but we have demonstrated that when the necessary information is available, they can be a powerful tool, especially when evaluating therapies for rare diseases such as GBM or heterogenous diseases with long and slow progressions that require many years of patient follow-up such as AD.

CHAPTER 3:

AUTOMATED DETECTION OF PARAMAGNETIC RIM LESIONS IN MULTIPLE SCLEROSIS

3.1. Introduction

Multiple sclerosis is a demyelinating and inflammatory disorder whose hallmark is lesions in the brain and spinal cord (Sahraian and Radü, 2007). These lesions can be detected *in vivo* with MRI and are often quantified as total lesion volume and lesion count, both of which can be used as measures of disease burden and to track disease progression (Popescu et al., 2013). Imaging biomarkers such as these are commonly used in the clinic and as surrogate endpoints in clinical trials (Filippi and Agosta, 2010; Sormani and Bruzzi, 2013). However, other known biological processes of MS are left uncaptured.

Chronic active lesions, a subset of MS lesions that are more prevalent in patients with more severe disease (Frischer et al., 2015; Luchetti et al., 2018; Absinta et al., 2019), have imaging and histopathology findings suggestive of ongoing tissue damage (Absinta et al., 2016; Dal-Bianco et al., 2017; Kaunzner et al., 2019; Gillen et al., 2021) and have until recently only been detectable by histopathology. These lesions have also been termed as slowly expanding, or smoldering lesions. At an estimated prevalence of as low as 4% but up to 10-15% of all MS lesions, this type of lesion is sufficiently common and deleterious to warrant considerable efforts for biomarker development (Frischer et al., 2015; Absinta et al., 2016; Dal-Bianco et al., 2017; Chawla et al., 2018). On T2*-phase MRI contrast, they are identifiable by curvilinear hypointensity along the edge of the lesion that corresponds with iron laden phagocytic cells observed on histopathological specimens (Bagnato et al., 2011; Absinta et al., 2016; Dal-Bianco et al., 2017). Here, we refer to them as paramagnetic rim lesions (PRLs).

When first observed on MRI, the rim of a PRL was only visible on scans from ultra-high-field strength (7T) magnets (Hammond et al., 2008; Absinta et al., 2013; Bian et al., 2013; Mehta et al., 2013). Recently, PRLs have been shown to be identifiable on the more common high-field strength (3T) MRI scans as well, albeit with lower inter- and intra-rater reliability (Absinta et al., 2018). This development strengthens their viability as a target on clinical MRI protocols, particularly because the sequences studied can be acquired with high spatial resolution in less than 4 minutes (Sati et al., 2014a). Previous studies of PRLs have noted the geometric nature of the rim and worked to identify the rim on the quantitative susceptibility mapping contrast as well (Eskreis□Winkler et al., 2015; Wisnieff et al., 2015; Stüber, Pitt and Wang, 2016).

Manual inspection of MS lesion for the presence of a paramagnetic rim is difficult, time consuming, and prone to inter- and intra-rater variability. We propose an automated method for identifying PRLs that would improve efficiency of study and facilitate translation of this biomarker into larger research studies and clinical practice. One way to identify PRLs is through the quantification of visual patterns that characterize these data. Radiomics is an emerging field of research that encompasses the extraction of quantitative features from biomedical images that may reflect underlying pathophysiology (Rizzo et al., 2018). Studies have shown that radiomic features are often useful predictors of known hallmarks of disease (Coroller et al., 2016; Liu et al., 2016; Bakas et al., 2017; Sweeney et al., 2021), although they have not been used extensively in the MS literature. We use radiomic features along with a random forest classification model, which can flexibly model high dimensional data, to identify PRLs. Our method is fully automated and uses a T2*-phase volume with isometric voxels and high spatial resolution that is acquired in a clinically feasible acquisition time at 3T (Sati et al., 2014a).

3.2. Materials and Methods

3.2.1. Study population:

We studied 20 subjects with MS who were scanned under an institutional review board–approved natural history protocol at the National Institutes of Health (NIH), who were included in this study due to the presence of visible PRLs in MR scans. Subjects’ age at the time of scanning ranged from 20 to 66 years, with a mean age of 45 years (sd = 12) (Table 3.1). Written informed consent was obtained from all participants. Data from this study can be shared upon reasonable request and completion of a Data Transfer Agreement with the National Institutes of Health.

Table 3.1: Demographics of Study Sample

Demographics	
N	20
Age (mean (SD))	45.5 (12.4)
Male (%)	8 (40)
Phenotype (%)	
Primary progressive MS	3 (15)
Relapsing-remitting MS	12 (60)
Secondary progressive MS	5 (25)
Disease Duration (mean (SD))	15.1 (9.0)
EDSS (median (range))	2.5 (1.0–7.0)
Treatments (%)	
Untreated	6 (30)
Glatiramer acetate	1 (5)
Interferon beta-1a	4 (20)
Dimethyl fumarate	6 (30)
Fingolimod	1 (5)
Natalizumab	1 (5)
Rituximab	1 (5)

3.2.2. MR Imaging acquisition:

All subjects were imaged on a Siemens Magnetom Skyra (Siemens, Erlangen, Germany) 3T scanner, using a body transmit coil and a 32-channel receive array coil, at the National Institutes of Health in Bethesda, Maryland. Imaging acquisition included the following sequences:

- a whole-brain 3D T2-weighted fluid-attenuated inversion recovery (FLAIR) sequence (repetition time, TR = 4800 ms; echo time, TE = 354 ms; inversion time, TI = 1800 ms; flip angle, FA = 120°; acquisition time, TA = 6 minutes 30 seconds; 256 axial slices; 1mm isometric voxel resolution),
- a whole-brain 3D T1-weighted magnetization-prepared rapid gradient echo (T1) sequence (TR = 7.8 ms; TE = 3 ms; FA = 18°; TA = 3 minutes 35 seconds; 256 sagittal slices; 1mm isometric voxel resolution), and
- a 3D segmented echo-planar imaging (EPI) sequence with whole-brain coverage providing T2* magnitude and phase contrasts (TR = 64 ms; TE = 35 ms; flip angle, FA = 10°; TA = 5 minutes 46 seconds; 251 sagittal slices; 0.65mm isometric voxel resolution).

Additional standard MRI sequences, including a postcontrast 3D T1-weighted MPRAGE sequence for the identification of gadolinium-enhancing lesions, were also acquired but not incorporated into the automated assessment of PRLs.

3.2.3. Manual paramagnetic rim lesion assessment:

Supratentorial non-gadolinium enhancing MS lesions were visually inspected for the presence of a paramagnetic rim on T2* magnitude and unwrapped phase images by a neurologist with 14 years of experience in neuroimaging science (Absinta et al., 2013, 2018, 2019). Gadolinium enhancing lesions were excluded from the analysis because the main focus of this paper was to study chronic rim lesions. In Absinta et al. (2016) (Absinta et al., 2016), PRLs were found in 22 out of 40 gadolinium enhancing lesions. Of these 22, 45% of the rims disappeared within 3 months after enhancement resolved. As previously described (Yao et al., 2012), we identify a PRL when a hypointense signal on phase images is observed surrounding the periphery of the lesion, while being either hyper- or isointense in its inner portion. PRLs were delineated on the phase with a line through the center of the lesion along its longest axis on an axial slice.

3.2.4. Image preprocessing:

Phase images were unwrapped and filtered as previously described (Absinta et al., 2013). T1, FLAIR, and phase images were preprocessed using the *fslr* R package (Muschelli et al., 2015), an R wrapper for the FSL software(Smith et al., 2004; Jenkinson et al., 2012), further described below. Images were visualized with ITK-SNAP (Yushkevich et al., 2006). The T2*-magnitude contrast was not used in this method.

To preprocess our images, we first applied the N4 inhomogeneity correction algorithm to the T1, FLAIR, and phase images (Tustison et al., 2010). We then rigidly registered both the T1 and the FLAIR images to the T2*-phase image space, resampling to 0.65 mm isometric resolution and using a mutual information cost function and sinc interpolation. When deciding on registration parameters, we also considered using 9-parameter and 12-parameter registration but found that registration with those degrees of freedom resulted in some failed cases with warped images. We used multi-atlas skull stripping (MASS) to identify cerebral tissue in the images in T1 space (Doshi et al., 2013). In two cases, MASS yielded poorly skull-stripped images based on visual inspection. For those two cases, we instead used the FSL brain extraction tool for skull-stripping (Jenkinson et al., 2012). As a final step, we performed WhiteStripe intensity normalization on the otherwise preprocessed T1, FLAIR, and phase images (Shinohara et al., 2014).

3.2.4. Lesion labelling:

Our lesion labelling method relies on access to maps that represent voxel-wise probabilities of being a lesion. We use the automatic lesion segmentation method MIMoSA for its ability to integrate multimodal information and to provide voxel-wise probability maps (Valcarcel et al., 2018b). Manual lesion segmentation was conducted by a research assistant with 1 year of experience, who was trained by a board-certified neurologist with extensive expertise in neuroimmunology and MRI.

We trained the MIMoSA algorithm with manual segmentations as a gold standard and T1 and FLAIR images as input. We implemented a leave-one-out cross-validation approach, where data from all but one subject was used to train a MIMoSA model, and that model was subsequently applied to the remaining subject. We repeated this for every subject in our cohort.

From each k-fold model, we extracted probability maps that contained voxel-wise probabilities of being a white matter lesion. We then binarized these probability maps into lesion segmentation maps via a subject-specific estimated optimal threshold that was identified out of a user-provided range of possible thresholds and then chosen based on amount of overlap with a gold-standard lesion segmentation as measured by a Sørensen-Dice coefficient (Valcarcel et al., 2020).

Because our lesion segmentation masks did not always cover the entire area of a lesion, we then dilated the masks by one voxel in each direction to increase the likelihood of detecting the paramagnetic rim signal, which occurs on the boundary of lesions. In order to then mitigate the possibility that our dilation inadvertently resulted in the added inclusion of CSF or gray matter, we used FSL FAST to segment CSF and gray matter and masked those out of the voxels newly included through dilation (Zhang, Brady and Smith, 2000).

After lesion segmentation masks were obtained, we used the lesion probability maps as input to a center detection method (Dworkin et al., 2018b) to identify distinct lesions based on the texture of the lesion tissue. We then used a nearest-neighbor approach to classify the remainder of the lesion segmentation map into those identified lesions (Figure 1). At this point, we assigned PRL status to the identified lesions based on the presence of any overlap with the manual PRL labels described previously.

Due to failures in the lesion labelling process, a subset of abnormalities automatically identified by our method might, to a manual rater, be considered clusters of confluent lesions. Because we did not have access to manual segmentations of distinct lesions, we instead relied on a combination of our lesion labelling method and connected components analysis to label lesions as confluent. Specifically, if connected components identified one cluster where our lesion labelling method identified more than one lesion, we labelled the constituent lesions as confluent.

3.2.5. Feature extraction:

With the lesions identified by our automatic pipeline, we conducted a radiomic image analysis to characterize each lesion with intensity-based statistics on the phase contrast (Kolossváry et al., 2017). These include 44 features that summarize the intensities in an individual lesion in 3 general ways: by describing the average and spread of the intensities, by describing the shape of the distribution of intensities, and by describing the diversity of intensities(Kolossváry et al., 2017). For example, features like the mean, defined as $\frac{1}{n}\sum_{i=1}^{n} x_i$, and interquartile range, defined as $abs(x_{75\%} - x_{25\%})$, are included in the first group, where x_i represents intensity value at voxel i. Features like variance, defined as $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - mean(x)\right)^2$, and skew, defined as $\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - mean(x)\right)^3}{sd(x)^3}$, are included in the second group, and features like energy, defined as $\sum_{i=1}^{n} x_i^2$, uniformity, defined as $\sum_{i=1}^{n} p(x_i)^2$, and entropy, defined as $\sum_{i=1}^{n} -p(x_i)\log_2 p(x_i)$, are included in the third group. A full list and detailed equations for each of the first-order radiomic features can be found in the supplemental material of Kolossváry et al. (2017) (Kolossváry et al., 2017).

3.2.6. Prediction model:

The radiomic features were used as candidate predictors in our subsequent prediction modelling for classification of PRL lesions. Class labels for each lesion were previously assigned during the lesion labelling step. We split our dataset into a training set and test set by subject, randomly assigning lesions from 16 subjects into the training set and lesions from the remaining 4 subjects into the test set. Both sets were examined to ensure that at least 100 lesions were present in each group.

Because PRLs were of a minority class with a prevalence of approximately 12%, we used Synthetic Minority Oversampling TEchnique (SMOTE) to balance our data (Chawla et al., 2002). With SMOTE, we oversampled the PRLs by the reciprocal of the percentage of PRLs present in the dataset and we did not undersample the majority class. We then trained a random forest classifier with 10-fold cross-validation using the R package *caret* (Kuhn, 2008; Wright and Ziegler, 2017). We summarized performance results using an optimal threshold calculated based on Youden's J statistic, which maximizes the sum of sensitivity and specificity (Youden, 1950). We also derived empirical confidence intervals for those measurements by randomly reassigning the training and test set and repeating the above process 1000 times. We assessed variable importance in the random forest as the percent increase in mean-squared error for a model with the variable over a model with a permuted version of that variable, scaled for comparability across variables.

3.2.7. Post-hoc analyses:

An additional board-certified neurologist (MS) with extensive expertise in neuroimmunology and MRI, who was not involved in the generation of the manual PRL labels, examined each misclassified lesion. We rated lesions on a 5-point scale, where 1 indicated definitely not a PRL, 2 indicated probably not a PRL, 3 indicated uncertain, 4 indicated probably a PRL, and 5 indicated definitely a PRL. Some lesions were automatically labelled as one lesion but were actually a confluence of lesions (Figure 1). We assigned manual ratings to these confluent clusters based on the presence of at least one PRL. We additionally assessed APRL's performance only for lesions that were not part of a confluent cluster.

Because it is known that the sizes of PRLs tend to be larger than non-PRLs, we extracted lesion size for use as a potential feature in our prediction model, measured as the number of voxels in a given lesion.

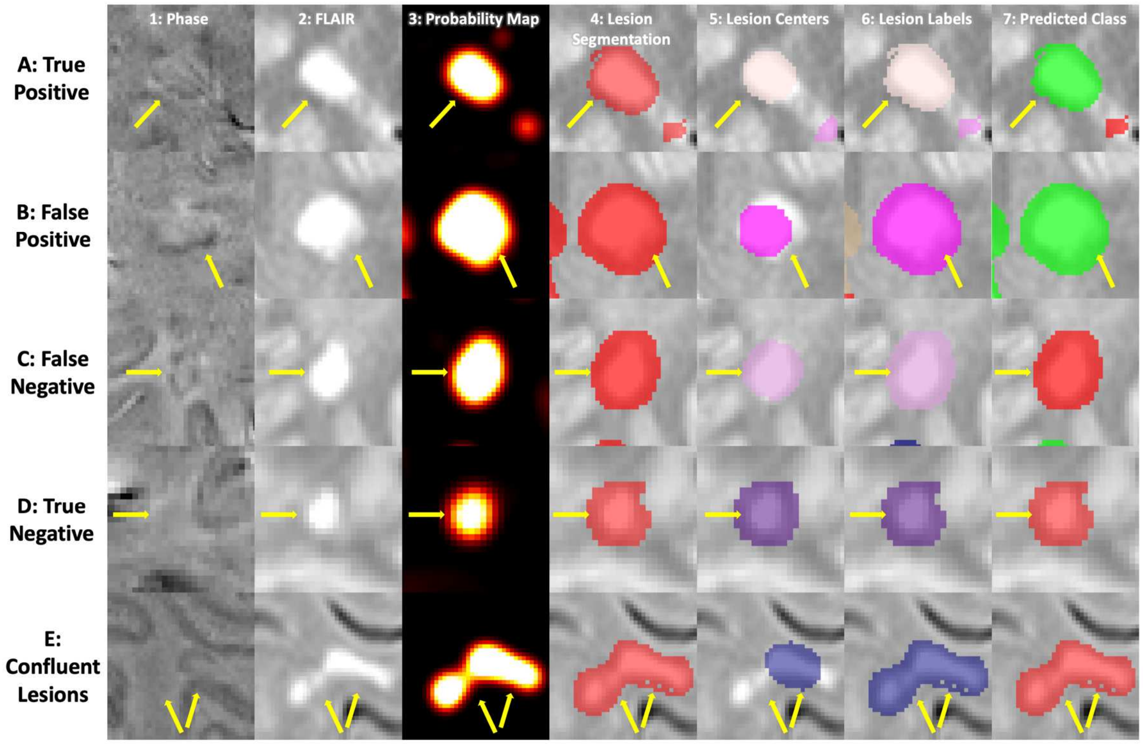


Figure 3.1: *A visualization of the steps of the method for five different lesions. Each column corresponds to a different part of the method, and each row corresponds to a different lesion of interest. In columns 5 and 6, the different colors represent different lesions, where the colors are arbitrarily assigned. In the last column, lesions classified as PRLs are visualized as green, and lesions classified as not PRLs are visualized as red. Subfigure A shows a lesion that was both manually identified as a PRL and classified as a PRL, i.e. a true positive. Subfigure B shows a lesion that was manually identified as not a PRL but classified as a PRL, i.e. a false positive. Correspondingly, subfigure C shows a false negative lesion, and subfigure D shows a true negative lesion. Subfigure E shows a lesion that was automatically labelled as a single lesion but is actually a confluence of lesions.*

3.3. Results

The final dataset included a total of 951 lesions in 20 subjects identified by our automated lesion labelling method, 113 (12%) of which we found to be PRLs by overlap with the manual annotation. The average number of lesions per subject was 47.6 (sd = 15.9), and the average number of automatically identified PRLs per subject was 5.7 (sd = 2.9). The number of identified

PRLs by our method was highly correlated with the gold standard count of PRLs, r = 0.86 (95% CI [0.68, 0.94]) (Figure 3.2).

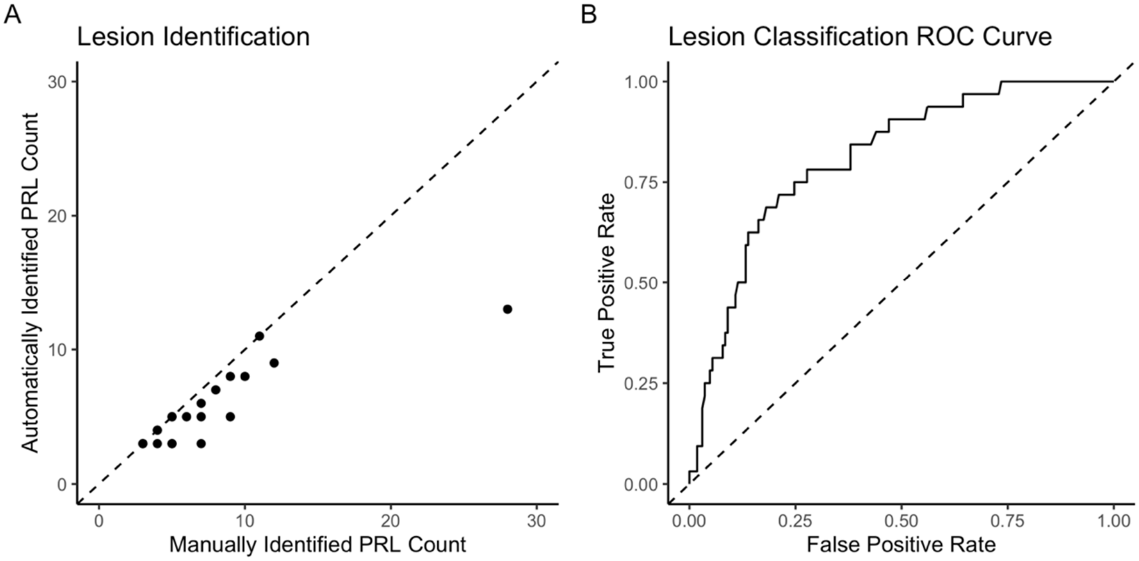


***Figure 3.2:** Subfigure A shows the manually identified count of PRLs against the number of PRLs estimated via our lesion identification method, r = 0.86 (0.68, 0.94). Subfigure B shows the ROC curve after classification, AUC = 0.82 (0.74, 0.92).*

We trained a random forest classification model using PRL status from the lesion labelling method as the label. In the iteration that we used to derive performance measures, there were 753 lesions in the training set, 81 of which were PRLs, and 198 lesions in the testing set, 47 of which were PRLs. Using only the undiscretized radiomic features, we were able to classify lesions with an AUC of 0.82 (95% CI [0.74, 0.92]). Using 0.502 as a probability threshold, the optimal threshold as determined by Youden's J, 135 lesions were accurately classified as not PRL, 31 lesions were false positives, 8 were false negatives, and 24 were classified correctly as PRL (Table 3.2). A breakdown of the classification results for the test set lesions by subject is provided in Table 3.2, in which we see that the distribution of classification results is not very different between patients.

***Table 3.2: Summary of Classification Performance Measures.** The table summarizes the performance measures we observed for the classification of PRLs, where counts in parentheses are counts excluding confluent lesions. 95% confidence intervals are provided for performance measures where available.*

Contingency Table (Excluding Confluent Lesions)		
Prediction	Reference	
	Rim Negative	Rim Positive
Rim Negative	135 (47)	8 (0)
Rim Positive	31 (19)	24 (6)

Testing Set Lesion Classification Count by Subject (Excluding Confluent Lesions)				
Subject	True Negative	False Negative	False Positive	True Positive
1	65 (24)	4 (0)	10 (4)	4 (1)
5	13 (4)	1 (0)	8 (2)	7 (0)
8	25 (7)	0 (0)	4 (3)	5 (0)
16	32 (12)	3 (0)	9 (10)	8 (5)

Performance Measures	With Confluent Lesions (95% CI)	Without Confluent Lesions
AUC	0.82 (0.74, 0.92)	0.88
Accuracy	0.8 (0.59, 0.91)	0.74
Positive Predictive Value	0.44 (0.17, 0.55)	0.24
Negative Predictive Value	0.94 (0.93, 1)	1
False Positive Rate	0.19 (0.07, 0.46)	0.29
False Negative Rate	0.25 (0, 0.37)	0
Sensitivity	0.75 (0.63, 1)	1
Specificity	0.81 (0.54, 0.93)	0.71

We also examined the results of the method for lesions that were not part of a confluent cluster. A total of 72 lesions in the test set were not confluent, and were able to be classified with an AUC of 0.88. Using 0.086 as the probability threshold, the optimal threshold for this subset of lesions as determined by Youden's J statistic, 47 lesions were accurately classified as not PRL, 19 were false positive, 0 were false negative, and 6 were accurately classified as PRL (Table 3.2). Additional performance measures are provided in Table 3.2. Because we examined confluent lesions as part of a post-hoc analysis, we did not derive confidence intervals for these performance measures.

A visualization of lesions that were true positive, false positive, false negative, and true negative respectively is provided in Figure 1. From subfigure B, where we see the method illustrated for a lesion that was falsely identified as a PRL, we can see that hypointensities can manifest around a lesion even when they cannot be rated as a rim. Conversely, from subfigure C, which shows a lesion that was falsely identified as not a PRL, we see that despite the presence of hypointensities that are visible to the eye, certain PRLs may not display a signal strong enough to be captured by radiomic features.

The random forest identified uniformity, entropy, and energy as the most important radiomic features for classifying lesions, which are all features that aim to describe the diversity of the data points (Figure 3.3). Other radiomic features that were important were mode, kurtosis, and skew. Entropy and uniformity were both higher in lesions that were not PRLs, and energy was higher in PRLs. In a model including lesion size as an additional predictor, the random forest identified lesion size as the most important feature, with PRLs expressing larger sizes than lesions that were not PRLs, predicting PRL status with an AUC of 0.81. A model using textural features classified lesions with an AUC of 0.72.

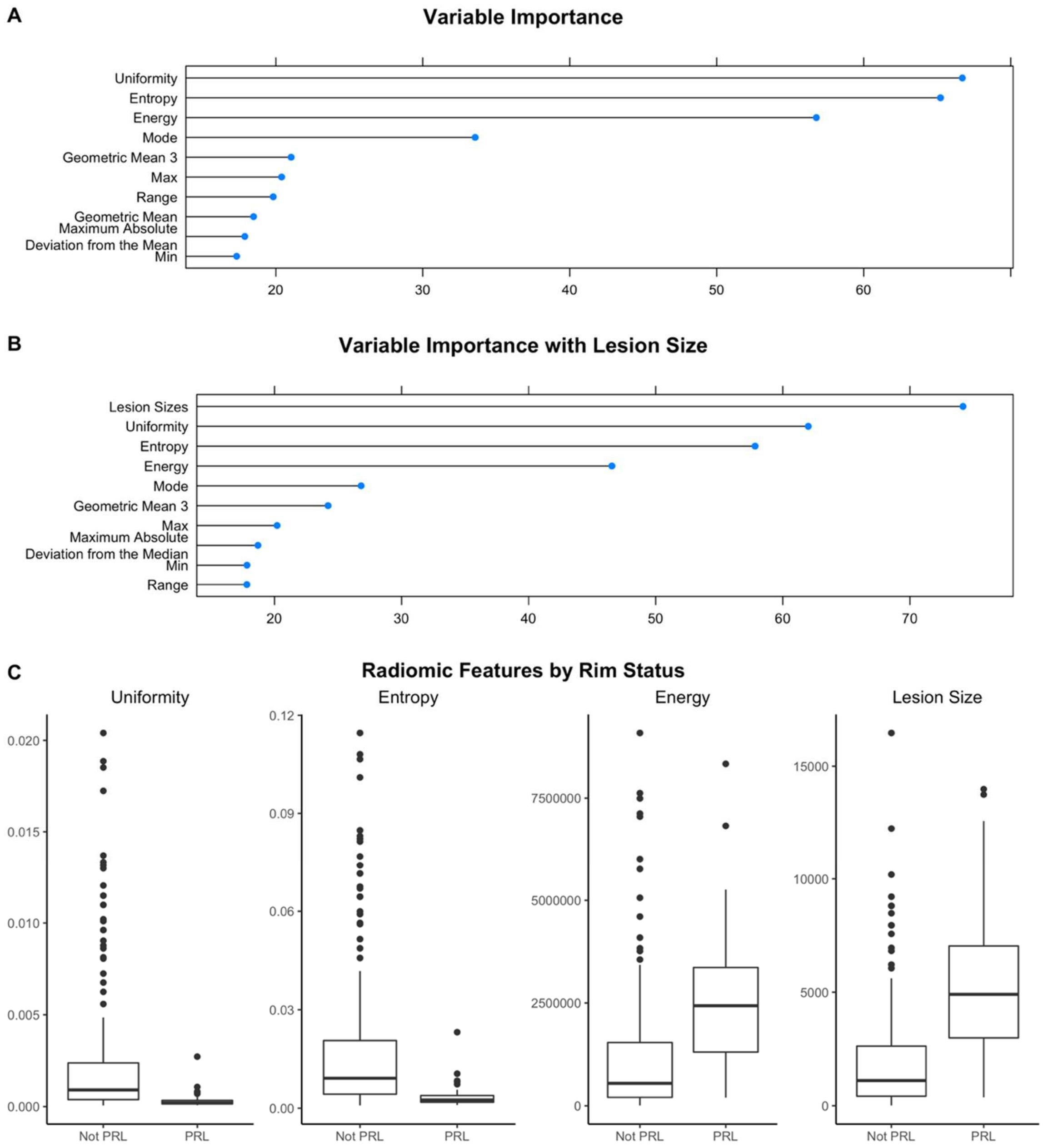


Figure 3.3: *The variables identified as the most important by APRL for determining the presence of PRLs were uniformity, entropy, and energy. Here, we measure variable importance as the percent increase in mean squared error for the model with the variable over the model with a permuted version of that variable, scaled for comparability across variables. Boxplots of uniformity, entropy, energy, and lesion size on the lesions from the test set show that PRLs and non-PRLs seem to differ on those measures, supporting the theory that they are important for distinguishing the two kinds of lesions.*

A second expert manually rated the 39 lesions that were misclassified by the model. The rater deemed that 1 lesion included too much artifact to assess PRL status, and 25 lesions were confluent. Of the lesions not part of a confluent cluster, 9 were false positive and 5 were false negative. Of those 9 false positive lesions, 4 were rated as definitely a PRL, 2 were rated as uncertain, 2 were rated as probably not a PRL, and 1 was rated as definitely not a PRL. For the 5 false negative lesions that were not confluent, 1 was rated as definitely a PRL, 2 were rated as probably a PRL, 1 was rated as uncertain, and 1 was rated as probably not a PRL.

As for confluent clusters, 22 were false positives and 3 were false negatives. These were rated according to the presence of at least one PRL in each confluent cluster. Of the 22 false positive lesions, 11 were rated as definitely a PRL, 5 were rated as probably a PRL, 1 was rated as uncertain, 3 were rated as probably not a PRL, and 2 were rated as definitely not a PRL. All 3 of the false negative lesions were rated as definitely a PRL. We note that the confluence defined here was a judgement made by the manual rater. This differs from but complements the confluence definition employed for the primary test set analysis, which was the definition based on the automated analysis used to derive the performance measures reported in Table 3.2.

3.4. Discussion

Preliminary studies have shown that the existence of a paramagnetic rim around an MS lesion is an important biomarker with potential clinical implications: indicative of chronic inflammation, associated with heightened disability, and resistant to current disease-modifying treatments (Absinta et al., 2019). However, paramagnetic rims are time-consuming to identify manually, even by highly trained experts (Absinta et al., 2018). In this paper, we developed APRL, a fully automatic method for detecting paramagnetic rim lesions on a 3T MRI using a submillimeter isometric, clinically feasible, segmented-EPI sequence (Sati et al., 2014a; Absinta et al., 2018). Automation of PRL identification that relies on objective assessment would aid larger scaled studies assessing this promising imaging biomarker in MS. Other automated approaches have also been explored (Barquero et al., 2020).

APRL relies on radiomics for automated PRL identification and classification. Radiomic features have not previously been used to classify PRLs. The radiomic features that were the most important in this context aimed to measure the variability of intensity within a lesion (entropy and uniformity) or quantify the magnitudes of the intensities themselves (energy).

Energy measures the magnitude of intensities within a lesion. On the phase image used in this study, PRLs manifested with higher energy because hypointensities represented more extreme negative values instead of values closer to 0, with more extreme hypointensities resulting in more extreme energy values.

Both entropy and uniformity are measures based on the probability of observing a particular intensity within a lesion. Because we did not bin the voxel intensities, the number of distinct intensities observed was large, so the probability of observing a particular intensity was fairly low. This was reflected in the observed range of uniformity in this study. Uniformity is a direct measure of homogeneity of the intensities within a lesion. We expected uniformity to be lower for PRLs due to the presence of both intensities representing normal appearing tissue and hypointensities from the paramagnetic rim. Lesions that were not PRLs did not appear with any distinct signature on a phase image, leading to a higher uniformity. In addition, the impact of the size of a region of interest on radiomic features in MS lesions has not been well studied and warrants further investigation.

Entropy takes the probability of observing a particular intensity within a lesion and transforms it to reflect the amount of observed variation. Because of the aforementioned lack of binning, here, entropy more accurately reflected lesion size in that given our more homogenous set of probabilities, a smaller probability of observing a given intensity resulted in a smaller measure of entropy. Larger lesions yielded a smaller probability of observing a given intensity. In this dataset, PRLs tended to have smaller values of entropy, possibly reflecting a larger size, which has been noted in previous studies of PRLs as well (Dal-Bianco et al., 2017). When we included lesion size in our classification model, we found that lesion size was an important predictor of PRL status in addition to uniformity, entropy, and energy, suggesting that these four measures provide potentially similar but nevertheless complementary forms of information for classifying PRLs. Many of the lesions that the model misclassified were confluent lesions that were labelled as a single lesion. According to our automated assessment of confluence, the percentage of confluent lesions among correctly classified lesions was 33%, while the percentage of confluent lesions among incorrectly classified lesions was 49%, suggesting that confluence negatively influences the model's ability to classify PRLs. According to our expert rater's visual assessment of confluence, nearly 65% of misclassified lesions were confluent. Of these, 88% were false positives, potentially reflective of heterogeneity in intensity that is more present for confluent lesions but also in lesions with a rim signal. Confluent lesions also tend to be larger, similar to PRLs, which may have also contributed to the misclassification.

We provide an example of one of these confluent lesions in Figure 1, Subfigure E. In this lesion, although one of the encompassed lesions contains a clear rim signal, the larger of the two does not. Because the majority of the voxels included in the confluent lesion belong to the encompassed one without a rim signal, the first-order radiomic features extracted from this confluent lesion reflected that signal.

We dilated our lesion segmentation map to increase the likelihood that a rim signal would be included in a lesion label. In order to mitigate the impact of inclusion of ventricular and cortical phase-hypointensities, we masked out cerebrospinal spinal fluid and gray matter from the dilations, but this dilation could have nevertheless resulted in the inclusion of non-lesional tissue that may have affected the calculation of radiomic features.

These issues could be addressed by taking a more nuanced approach to modelling the probability of having a rim. Here, we treated the identification of PRLs as a binary classification problem, invoking a random forest to predict if a given lesion was a PRL. However, the identification of PRLs can be difficult because of the myriad of factors that drive the clarity and strength of a rim signature, some of which are technical and some of which reflect biological processes. As noted in Figure 1, while some lesions exhibit a rim unequivocally, other lesions exhibit a more equivocal signature. This renders the task of identifying PRL lesions difficult, both for manual raters and automated classifiers. In fact, previous research has shown that intra- and interrater reliability for paramagnetic rim evaluation are substantial but not perfect, with a Cohen κ of 0.77 and 0.71 respectively (Absinta et al., 2018). A future approach could treat the presence of a rim as a continuous measure instead of a binary classification, where middling levels of this theoretical measure could represent both uncertainty about a lesion's classification and different stages of PRL progression. This would likely more accurately reflect underlying biological processes as well, as the amount of iron-containing phagocytes at the edge of a lesion can vary across lesions (Dal-Bianco et al., 2017).

3.4.1. Limitations:

A major limitation to current assessments of paramagnetic rims is that no international consensus exists on criteria for determining this imaging signature. This limitation may hinder the application of the proposed methodology to new studies in which differing definitions of paramagnetic rims may be desired based on local practices. While signal-to-noise ratio is higher on a 7T MR image, allowing for higher inter- and intra-rater reliability, they remain low across contrast types on 3T (Absinta et al., 2018). However, APRL relies on techniques that perform well on 3T images, so extensions to 7T would require additional validation.

This study may be improved by the collection of additional data containing delineations of rim signal locations. Increasing the sample size may allow for a more accurate reflection of the imaging signature associated with PRLs within the feature space, and a more specific delineation of the rim signal may improve APRL's ability to differentiate between hypointensity due to the presence of a rim and hypointensity due to noise or features like the central vein sign. In the current study, we did not explicitly assess for the presence of a central vein sign in each of the automatically identified lesions. Because the central vein sign also presents as hypointensity within a lesion on T2*-phase, a central vein sign might impact the calculation of first-order radiomic features. Textural features, which quantify the spatial relationship between voxel intensities, characterized PRLs less accurately than first-order features. Future studies may explore more direct methods for quantifying the central vein to disentangle the rim signal and the central vein sign. In addition, all the patients for this analysis had at least one PRL. Given recent histology work (Gillen et al., 2021), we do not suspect that patient without PRL lesions would have different radiomic signatures in their non-PRL lesions from patients with PRL lesions, but further work is warranted to investigate this.

Additionally, in the current study, we did not explicitly consider gadolinium-enhancement in our automated identification of PRLs. Gadolinium-enhancing lesions were specifically left out of the manual assessment in an effort to specifically study chronic rim lesions, whose presence has previously been shown to be associated with poor prognostic factors (Absinta et al., 2019). Paramagnetic rims in gadolinium-enhancing lesions fade within 3 months in a high percentage of cases (Absinta et al., 2016) and may exhibit features different from chronic rim lesions on imaging due to edema and tissue architecture, though this was not explicitly studied in this analysis.

3.5. Conclusion

This study introduces a fully automated method, APRL, for the identification and classification of paramagnetic rim lesions relying solely on 3T MR images, which are commonly available in a clinical setting. Automation of this process is important for the continued development of the scientific community's knowledge around these lesions and their implications for disease burden.

CHAPTER 4:

A NOVEL COORDINATE SYSTEM FOR MULTIPLE SCLEROSIS LESION EVALUATION ON MAGNETIC RESONANCE IMAGING

4.1. Introduction

Multiple sclerosis (MS) is a chronic, inflammatory demyelinating, and neurodegenerative disease of the central nervous system with no known cure. The symptoms and clinical trajectories of MS can vary widely, making diagnosis and prognosis difficult to determine. Around 20 to 30 percent of patients diagnosed with MS are actually misdiagnosed, and around 30 percent of these misdiagnosed patients incurred unnecessary morbidity due to their misdiagnosis (Solomon et al., 2016; Yamout et al., 2017; Kaisey et al., 2019). The McDonald criteria are the primary method of diagnosing MS, incorporating both clinical and imaging criteria, though misapplication of the McDonald criteria through overinterpretation of imaging results has heavily contributed to current misdiagnosis rates (Solomon, Klein and Bourdette, 2012; Thompson et al., 2018). The prevalence of misdiagnosis has motivated the study of imaging biomarkers that may hold improved diagnostic value.

Some biomarkers that have demonstrated potential in this space are specific to susceptibility-weighted imaging: in particular, the central vein sign, prominent on FLAIR* images, and the paramagnetic rim signal, visible on derivatives of susceptibility-weighted imaging, have been under study in recent years as potential diagnostic biomarkers (Sati et al., 2016; Solomon et al., 2018; Absinta et al., 2019; Sinnecker et al., 2019; Maggi et al., 2020b). Automated methods for both have been developed, and clinically relevant thresholds have been analyzed (Dworkin et al., 2018c; Maggi et al., 2020a; Barquero et al., 2020; Lou et al., 2021b; Ontaneda et al., 2021). The diagnostic value of both combined has also been shown to be high (Clarke et al., 2020). MS lesions are also thought to be shaped differently from lesions in patients with MS imaging mimics, in part also due to the perivenular nature of MS lesions, which leads to inflammation that is thought to gradually stem outwards from the source vein (Sinnecker et al., 2012; Wuerfel et al., 2012; Kilsdonk et al., 2014). Recent work has shown that 3-dimensional phenotyping of MS lesions that allow for assessment of both shape and surface characteristics also reveals differences in complexity of surface morphology and symmetry (Newton et al., 2017; Okuda et al., 2020).

These lesional features all aim to quantify ways that MS lesions are different from lesions from patients with an MS imaging mimic. In particular, the defining feature of the central vein sign is the existence of a hypointense tubular structure that appears at or near the center of a lesion. The defining feature of a paramagnetic rim signal is that it appears as a distinct signal at the lesion border. Lesion shapes rely on well-defined lesional boundaries that accurately reflect measures such as complex surface morphology and elongation. For each of these novel imaging biomarkers, a key characteristic is the specificity of both intensity and spatial distribution in the lesion. We propose a novel coordinate system for MS lesions that incorporates both spatial and voxel-wise intensity information in each lesion which can aid in discerning known diagnostic and prognostic MRI biomarkers, such as the central vein sign and the paramagnetic rim signal.

We borrow ideas from the human brain mapping space and functional data analysis to motivate the idea that building a common template for MS lesions through iterative registration can be a powerful framework for analyzing lesions (Avants et al., 2010, 2011). We propose a novel coordinate system that incorporates both spatial and voxel-wise intensity information in each lesion which could aid in discerning known diagnostic and prognostic MRI biomarkers, such as the central vein sign and the paramagnetic rim signal. Leveraging the central vein sign and estimated lesion boundaries, particularly relevant in the context of confluent clusters of lesions, we represent a given lesion in spherical coordinate space and then assess the added value of these features by examining association with clinical outcomes.

4.2. Materials and Methods

4.2.1. Data

The data used for this study comes from 68 people presenting for suspicion of multiple sclerosis who were imaged at the University of Vermont on a Phillips scanner. 23 of these patients were eventually diagnosed with MS while the remaining 45 were diagnosed with alternate disorders. 3D 3-Tesla magnetic resonance images with T1, T2-FLAIR, and T2* echo planar images were collected for all patients. Whole-brain 3D T_2-FLAIR, T_1, and T_2*-EPI (Sati et al., 2014b) volumes were acquired in a 3T Philips (dStream) MRI scanner. FLAIR and T_1 volumes were obtained with 1-mm resolution, and T_2*-EPI volumes were obtained with 0.55-mm isotropic resolution.

4.2.2. Image Processing

All images were bias corrected using N4 bias correction. T1 and T2*-magnitude images were skull-stripped using FSL BET (Jenkinson et al., 2012), the T1 was rigidly registered to the FLAIR image with windowed sinc interpolation, and these were both then also rigidly registered to T2*-magnitude space. We also derived a vesselness map using a Frangi filter (Frangi et al., 1998; Dworkin et al., 2018c). In order to get our centroids, we first intensity-normalize our T1 and FLAIR images using WhiteStripe (Shinohara et al., 2014) and then feed those into a pre-trained MIMoSA model with a 0.2 probability threshold (Valcarcel et al., 2018a). We then find the lesion centers using a technique proposed by Dworkin et al (Dworkin et al., 2018a) that relies on the texture of the lesion probability map to quantify areas with high probability of being a lesion. Periventricular centroids were then removed, and subsequently, all maps (i.e. T1, FLAIR, mimosa probability map, lesion segmentation map, and lesion centroid map) were registered to T2*-magnitude space.

4.2.3. Lesion Patch Generation

To create our template lesion, we first create image patches for each of our lesions on each of our subjects by masking out the lesion itself and then padding the image with 5 voxels on each side of each axis. For every lesion, we create patches of the FLAIR, EPI, and vesselness maps. We then create distance-to-boundary maps for each lesion and multiply that by the Frangi map for that lesion to create what we call a "coherence" map. We then run connected components on the map in order to identify a vein and determine each lesion to have the central vein sign if the largest component has more than 16 voxels. We then take the map of the connected components of the coherence map, now a binary mask, and rigidly register those together in hopes of aligning the most clear vein-like signal within a lesion. We apply these registrations to the magnitude and flair images and then start the template construction process.

4.2.4. Template Creation

For template construction, we utilize a two-stage process for constructing an all-subject template: first, we construct subject-specific templates using a multimodal template construction pipeline as described by Avants (Avants et al., 2011), and then we linearly register those subject-specific templates to one of the subject-specific templates, chosen manually for the presence of a distinct vein.

For the construction of subject-specific lesions, we use only the lesions with a strong vein-like signal. We first create coherence maps that enhance the strongest central tube-like signal in each lesion by binarizing the largest connected component on a given lesion's coherence map. We then rigidly register these enhanced coherence maps together and apply estimated registrations to the FLAIR and EPI lesions. We then take the rigidly registered FLAIR and EPI lesions and use multimodal template construction in the ANTs environment (Avants et al., 2011). We use a gradient step of 0.15, 10 iterations, and both a linear and a nonlinear registration, summarized using the mean of normalized intensities, with equal weighting on the FLAIR and EPI images.

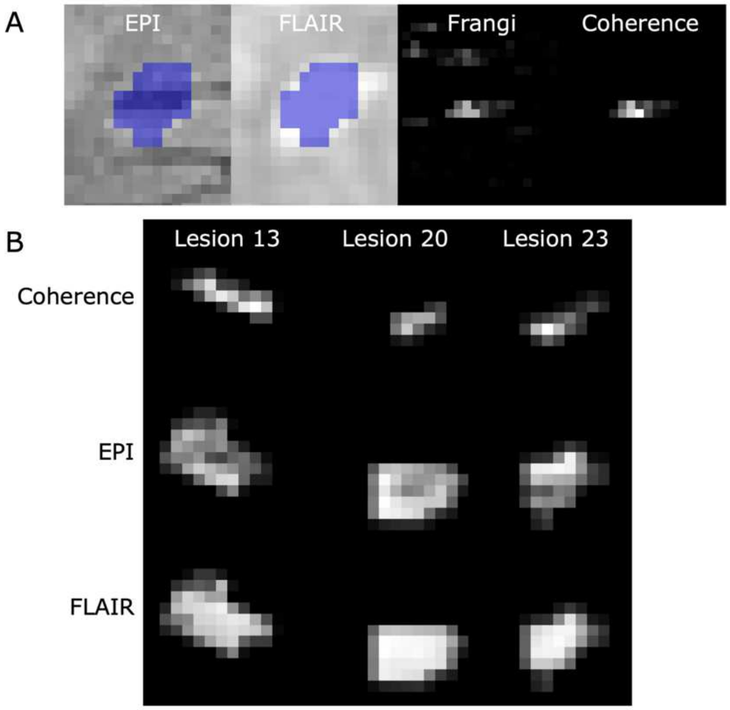


Figure 4.1: *Example images of MS lesions, processed and unprocessed. In subfigure A, we see an example of an MS lesion in native space. The blue segmentation indicates the automatically derived centroid mask. The Frangi vesselness filter and coherence map highlight the central vein in the center of the lesion. In subfigure B, we see 3 example lesions from the same subject. Our pre-template construction registration pipeline consists of initial within-subject rigid registration of coherence maps, followed by application of the estimated registration to EPI and FLAIR images. This process allows for template construction to align the vein signal as best as possible.*

Once the subject-specific templates are constructed, we then take the templates from patients who were diagnosed with MS and linearly register those subject-specific templates to a chosen subject template, smooth the registered subject-specific templates and then average them to form our all-subject template. For patients who were diagnosed with an alternate disorder, we use “similarity” registration (i.e. rotation, translation, and scaling) in order to register their subject-specific templates to the all-subject template and save that registration.

To then register all the lesions into all-subject template space, we nonlinearly (SyN) register them to their subject-specific template, and then use that and the transformation estimated by the registration of the subject-specific template to the all-subject template to register the EPI to all-subject template space (Avants et al., 2008). We use those transformations estimated by EPI-to-subject-specific template registration and subject-specific-to-all-subject-template space to also register our coherence maps, estimated in native space, to all-subject-template space.

4.2.5. Statistical Modelling

To identify axes of variation among lesions, we vectorize all lesion patches registered to template space and then perform PCA on our matrix of vectorized lesions, for which the intensities at each voxel location across lesions are centered but not scaled. We then extract the scores of each of the principal components (PCs), and analyze those with respect to the difference between lesions from MS patients and lesions from patients with an alternate disorder. We measure the degree to which that difference exists by performing a two-sided t-test that we Bonferroni-correct by the number of lesions in our study.

For the PCs that we identify as containing information that significantly differentiates MS and non-MS lesions, we then model the association between those PC scores and diagnosis using the following linear mixed effects model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{zu} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is the PC score under study, $\boldsymbol{\beta}$ is a vector containing an intercept term and a term for MS diagnosis, and $\boldsymbol{u}$ represents a random intercept on subject. We limit our study of PCs to the first 10, as we believe that interpretability is difficult for PCs that capture a small amount of variability within the original dataset.

We evaluate the significance of the association between diagnosis and PC score by using Satterthwaite’s degrees of freedom method (Satterthwaite, 1946; Kuznetsova, Brockhoff and Christensen, 2017). We also evaluate the diagnostic accuracy of those PCs by extracting a number of summary statistics and using them in a random forest model to model diagnosis. The summary statistics we calculate are as follows: maximum, minimum, median, mean, variance, interquartile range, skewness, and kurtosis. We cross-validate these results using leave-one-out cross validation.

In order to mitigate overfitting, we perform 5-fold cross validation with folds based on subject. For each fold, we first take our patients in the training set of the fold, perform PCA on the lesions from those patients, as outlined above, find the PC with scores that most significantly differentiate MS and non-MS lesions, and use statistics of that PC summarized by subject to model diagnosis. As an alternative feature set, we also extract the medians of all PCs that significantly differentiated MS and non-MS lesions and use those features in a random forest to model diagnosis. Diagnosis was modeled using, separately, a generalized linear model and a random forest. For the patients in the test set of the fold, we map the lesions from those patients into the PC space estimated by the training set, extract the PC identified by the training set as most significantly differentiating MS and non-MS patients, and predict diagnosis using summary statistics for that PC.

4.3. Results

After preprocessing, we had 3114 lesions from 63 patients under analysis, 22 of whom were diagnosed with MS and 41 of whom were diagnosed with an alternate disorder. There was an average of 49 lesions from each patient (Range: 8—200); among MS patients, this was an average of 68 lesions per patient (Range: 14—200), and among non-MS patients, this was an average of 39 lesions per patient (Range: 8—180). For subject-specific template construction, for which we only utilized a subset of our original set of lesions, we used an average of 7 lesions per patient.

For our study-wide template, we used a total of 252 lesions from 22 patients with MS, with an average of 11 lesions per patient, all of which were selected because of the existence of a strong vein signal in the center of the lesion. Our final study-wide template lesion was 23 voxels by 46 voxels by 29 voxels large, for a total of 30682 voxels.

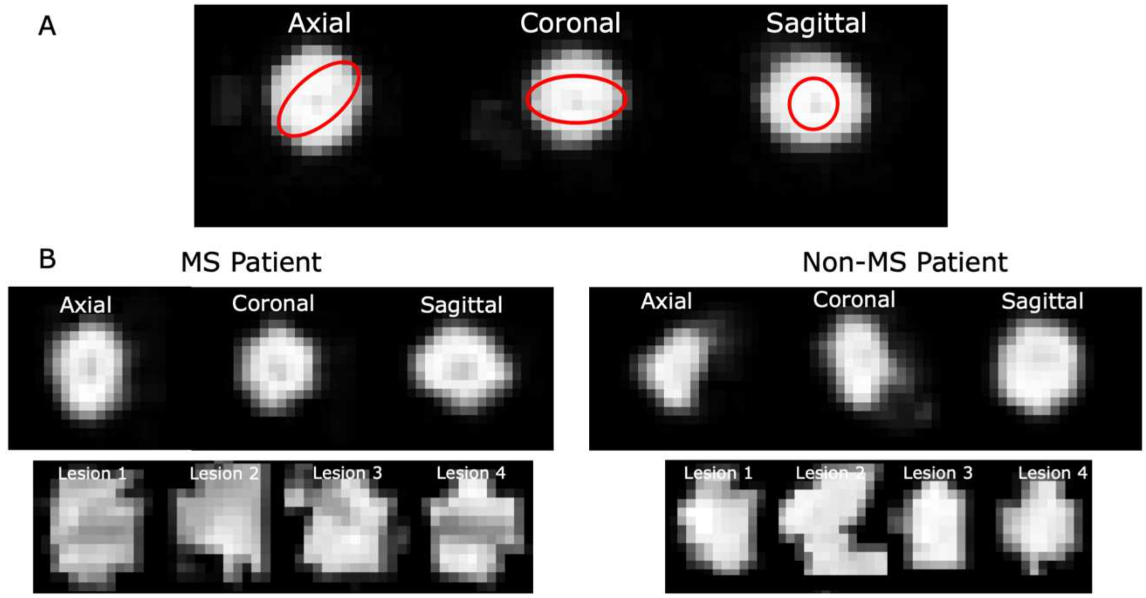


Figure 4.2: *Subfigure A: The all-subject template contains hypointensity in the center of the lesion that reflects hypointense pattern expected for a vein-like signal, circled in red. Subfigure B: The subject-specific template from an MS patient has a more prominent hypointensity in the center of the lesion than the subject-specific template from a non-MS patient. Lesions from MS patient registered to template space maintain visible vein-like pattern but lesions from non-MS patient do not show this pattern.*

To analyze our resultant lesions registered to template space, both MS and non-MS, we perform PCA on a vectorized version of the images in the form of a matrix of 3114 lesions by 30682 voxels. When we then analyze which principal component scores significantly differentiate MS and non-MS lesions after Bonferroni correction, we see that 7 of our principal components significantly differentiated MS and non-MS lesions. P-values assessing the significance of the difference between MS and non-MS lesions by PC score are reported in Table 4.1 below alongside inference derived from a mixed-effects model examining the association between PC score and diagnosis while accounting for subject-level variation.

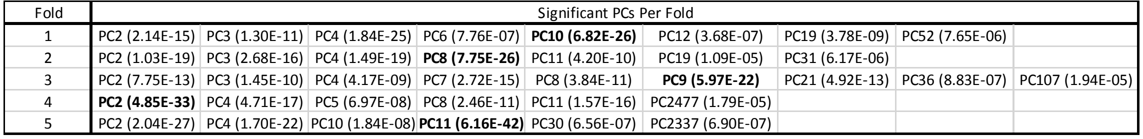
Table 4.1: Inference for PC Scores and MS Diagnosis. *Principal components with scores that significantly differentiated lesions from MS patients and lesions from non-MS patients are listed alongside the p-value for a test of significant differences. We also report results for a test of significance for the association between MS diagnosis and PC score within a mixed-effects model accounting for random variation on the subject level.*

Principal Component	Association between MS Diagnosis and PC Score: P-Value	Association between MS Diagnosis and PC Score accounting for subject-level variation: P-value
PC2	2.29E-25	0.0558
PC3	1.35E-06	0.719
PC4	9.63E-23	0.026
PC8	5.43E-06	0.0452
PC9	8.57E-08	0.345
PC10	2.14E-31	0.00306

The summary statistics of just PCs that are significantly associated with diagnosis after accounting for subject-level variation, when modeled univariately in a generalized linear model, yield AUCs of 0.83 and 0.76 for PC4 and PC10 respectively. If we cross-validate the predictions using a leave-one-out-cross-validation scheme across patients, we see AUCs of 0.66, 0.58, and 0.60. When we take the medians of each PC that significantly differentiated MS and non-MS lesions across subject and use those as features in a generalized linear model, we see an in-sample AUC of 0.87 and a leave-one-out-cross-validated AUC of 0.78.

After cross-validating the whole analysis pipeline, if we do an analogous analysis by taking the medians of all significant (Bonferroni-corrected) PCs and modeling diagnosis, we see AUCs of 0.54 and 0.57 for our GLM and random forest respectively. If we model diagnosis with summary statistics of just the PC that most significantly differentiated MS and non-MS lesions within that fold, we see AUCs of 0.64 and 0.59 for our GLM and random forest respectively. The most significant PC per fold and a list of the significant PCs per fold are listed in Table 4.2.

Table 4.2: List of Significant Principal Components per Cross-validation Fold. *Most significant PCs for each fold are highlighted.*



Fold	Significant PCs Per Fold								
1	PC2 (2.14E-15)	PC3 (1.30E-11)	PC4 (1.84E-25)	PC6 (7.76E-07)	**PC10 (6.82E-26)**	PC12 (3.68E-07)	PC19 (3.78E-09)	PC52 (7.65E-06)	
2	PC2 (1.03E-19)	PC3 (2.68E-16)	PC4 (1.49E-19)	**PC8 (7.75E-26)**	PC11 (4.20E-10)	PC19 (1.09E-05)	PC31 (6.17E-06)		
3	PC2 (7.75E-13)	PC3 (1.45E-10)	PC4 (4.17E-09)	PC7 (2.72E-15)	PC8 (3.84E-11)	**PC9 (5.97E-22)**	PC21 (4.92E-13)	PC36 (8.83E-07)	PC107 (1.94E-05)
4	**PC2 (4.85E-33)**	PC4 (4.71E-17)	PC5 (6.97E-08)	PC8 (2.46E-11)	PC11 (1.57E-16)	PC2477 (1.79E-05)			
5	PC2 (2.04E-27)	PC4 (1.70E-22)	PC10 (1.84E-08)	**PC11 (6.16E-42)**	PC30 (6.56E-07)	PC2337 (6.90E-07)			

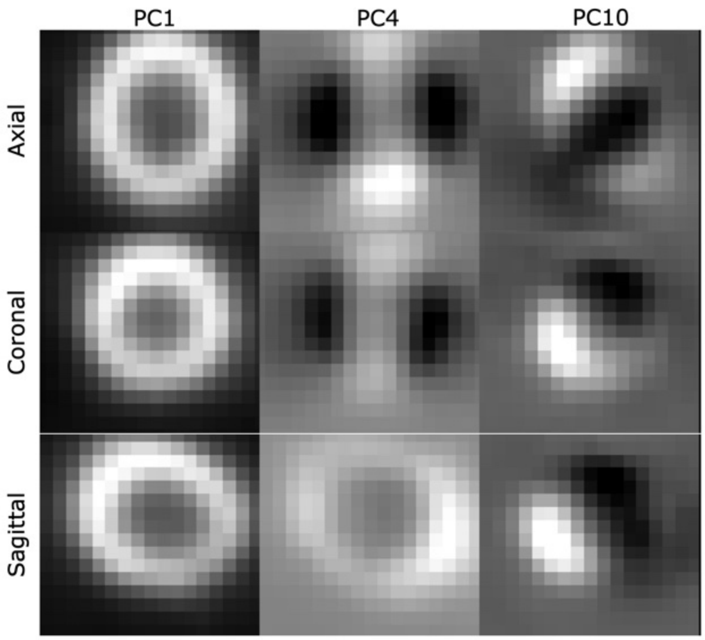


Figure 4.3: Voxel-wise loadings for PC1, PC4, and PC10. *The hyperintense voxels are voxels with high loadings for a given PC, and hypointense voxels are voxels with low loadings for a given PC. The loadings for these PCs seem to suggest that the in fact the largest amount of variation, as encapsulated by the first PC, occurs in a ring around the center of a lesion. This may be capturing residual differences in lesion size. PC4 and PC10 are each PCs that significantly differentiated MS and non-MS lesions, both with and without accounting for subject-level variability. PC4 seems to capture variability primarily occurring in one tubular direction and PC10 seems to capture variability occurring around the edge of lesions.*

4.4. Discussion

We develop a novel morphological coordinate system for quantifying lesion damage and repair that allows for simultaneous assessment of spatiality and intensity to address between-lesion heterogeneity. Downstream analysis of MS and non-MS lesions registered to the coordinate system shows the ability of our proposed method to highlight differences between lesions from patients with different diagnoses.

Subject-specific template lesions created for MS patients maintain hypointense signal in the center of the lesion while subject-specific template lesions created for non-MS patients do not show that pattern. Lesions from MS patients registered to template space maintain the central vein signal and appear closely aligned with each other while lesions from non-MS patients do not.

Principal components analysis reveals that the scores of 7 out of 3114 estimated principal components significantly differentiate MS and non-MS lesions. Three of these principal components additionally significantly differentiate MS and non-MS lesions after accounting for subject-level variation in lesion presentation. The loadings of these three principal components each highlight different spatial patterns in a lesion, one highlighting variation occurring primarily in a line through the center of the lesion, one highlighting variation primarily occurring on the outer edges of the lesion, and one highlighting variation occurring directly around a hypointense tubular form in the center of the lesion. Cross-validation shows that PC scores are a reliable way to identify directions of variation between lesions and that analysis of the PC that most strongly differentiates MS and non-MS lesions can yield diagnostic value.

The primary limitation of this implementation of this method, comes from the fact that we rely on automatic segmentations of lesion labels for this study. A template lesion is naturally going to be highly reliant on accurate delineation of lesion boundaries, but manual segmentation of lesional tissue and identification of distinct lesions with lesional tissue can be very time-consuming, as many MS patients have multiple lesions that are often confluent. For this reason, we rely on automatic methods both for identifying lesional tissue as well as identifying distinct lesions, but automatic methods are prone to mislabeling and may lead to inaccurate delineations of lesional boundaries. Mislabeling of lesional boundaries can lead to inconsistencies in our ability to measure the paramagnetic rim signal, which is defined by its presence specifically on the edge of an MS lesion, as well as morphological features such as elongation and complexity of surface morphology, which rely on lesion boundaries to define shape and size.

To avoid mishaps stemming from inaccurate delineation of lesion boundaries, in this iteration of this work, we focus our study to lesion centroids, defined as lesional tissue with the highest probabilities of being a lesion. By focusing the scope of our work to lesion centroids, we eschew the ability to incorporate the paramagnetic rim signal and lesion morphology into our assessment of diagnosis, but we are able to refine our template creation pipeline to highlight the central vein sign and amplify signals in the centers of lesions that may contain diagnostic value. Future iterations of this work may benefit from manual segmentation of distinct lesions that allow for more accurate measurements of the paramagnetic rim signal and lesion morphology alongside the central vein sign for a template lesion that truly encapsulates all of the imaging signals that are known to contain diagnostic value.

As an adjacent point, we primarily focus our downstream analyses here to just the template lesion created for the T2*-EPI contrast, but supplemental imaging contrasts may be incorporated for identification of the paramagnetic rim signal, most easily detectable on T2*-phase or quantitative susceptibility mapping contrast (QSM). Other imaging contrasts not incorporated in this work have also recently been shown to potentially contain important, complementary information for diagnosis (Fazekas et al., 1999; Thaler et al., 2015; Eskreis-Winkler et al., 2017; Zhang et al., 2018; Jang et al., 2020). Our current work only incorporates the T2*-EPI and the FLAIR contrast for template construction.

Further work may explore a more robust initial registration pipeline: even after highlighting just the most central tube-like structure within a lesion, rigid registration does not completely align the CVS within lesions. An additional registration step simply using rotation may improve alignment of vein signals. Also, this study incorporated images from patients with a range of alternative disorders, which may reduce the ability of diagnostic evaluation to accurately identify MS-specific traits. Comparison to a single alternative disorder may improve measurements of diagnostic value.

CHAPTER 5:
DISCUSSION

The growth in accessibility of neuroimaging data has prompted an increased focus on the study of imaging biomarkers. The development and evaluation of these novel indices of brain phenomena are critical to maximizing the utility of these data. Improper usage and development can lead to spurious conclusions and false promises.

5.1. Imaging biomarkers augment statistical power

In Chapter 2, we show that individualized machine-learning-based imaging biomarkers can be useful tools for augmenting the power of clinical trial analysis when the necessary information is available. This improvement can offer decreased sample size requirements for detecting a given effect size, providing greater flexibility in resource utilization. The novelty of this method arises from the incorporation of individualized predictions based on powerful predictive algorithms, which can lend power to the detection of an average treatment effect due to targeting of the individuals in a given clinical trial. As robust neuroimaging biomarkers derived via machine learning models become more available, the historical datasets that can be analyzed with those models grow in size. These changes are expected to strengthen the radiomic prediction models like the ones used in this study.

Our incorporation of the radiomic predictor relies on the existence of previously developed imaging biomarkers. For the purposes of this paper, we theorize that these models have been trained on a historical cohort. Because the radiomic predictor has been trained on a historical cohort, the models that we fit to analyze a current trial inherently incorporate information from historical controls. We generate values of the radiomic predictor for the current sample using the model that was developed with a previous cohort.

Reductions in sample size requirement depend upon the strength of the prediction model. In Chapter 2, both imaging biomarkers considered were built using SVMs. Other machine learning techniques such as deep convolutional neural networks may provide greater predictive power. In addition, gains in power for the primary outcome will be associated with gains in power for secondary outcomes only to the extent that predictions from the prediction model are associated with the secondary outcomes. We demonstrate that when the necessary information is available, imaging biomarkers can be a powerful tool, especially when evaluating therapies for rare diseases such as GBM or heterogenous diseases with long and slow progressions that require many years of patient follow-up such as AD.

5.2. Towards automated paramagnetic rim assessment

Preliminary studies have shown that the existence of a paramagnetic rim around an MS lesion is an important biomarker with potential clinical implications. Previous studies have found that paramagnetic rim lesions are (1) indicative of chronic inflammation, (2) associated with heightened disability, and (3) resistant to current disease-modifying treatments. However, paramagnetic rims are time-consuming to identify manually, even by highly trained experts. In Chapter 3, we develop APRL, a fully automatic method for detecting paramagnetic rim lesions on a 3T MRI using a submillimeter isometric, clinically feasible, segmented-EPI sequence. Automation of PRL identification that relies on objective assessment can aid larger scaled studies assessing this promising imaging biomarker in MS.

A major limitation to current assessments of paramagnetic rims is that no international consensus exists on criteria for determining this imaging signature. This limitation may hinder the application of the proposed methodology to new studies in which differing definitions of paramagnetic rims may be desired based on local practices. The identification of PRLs can also be difficult because of the myriad of factors that drive the clarity and strength of a rim signature, some of which are technical and some of which reflect biological processes. This renders the task of identifying PRL lesions difficult, both for manual raters and automated classifiers. A future approach could treat the presence of a rim as a continuous measure instead of a binary classification, where middling levels of this theoretical measure could represent both uncertainty about a lesion's classification and different stages of PRL progression. This would likely more accurately reflect underlying biological processes as well, as the number of iron-containing phagocytes at the edge of a lesion can vary across lesions.

5.3. The template multiple sclerosis lesion

In Chapter 4, we develop a novel morphological coordinate system for quantifying lesion damage and repair that allows for simultaneous assessment of spatiality and intensity to address between-lesion heterogeneity. Downstream analysis of MS and non-MS lesions registered to the coordinate system shows the ability of our proposed method to highlight differences in lesion presentation between patients with different diagnoses. Lesions from MS patients registered to template space maintain the central vein signal and appear closely aligned with each other while lesions from non-MS patients do not. Principal components analysis of lesions in template space reveals that the scores of 7 out of 3114 estimated principal components significantly differentiate MS and non-MS lesions. Three of these principal components additionally significantly differentiate MS and non-MS lesions after accounting for subject-level variation in lesion presentation. The loadings of these three principal components each highlight different spatial patterns in a lesion, one highlighting variation occurring primarily in a line through the center of the lesion, one highlighting variation primarily occurring on the outer edges of the lesion, and one highlighting variation occurring directly around a hypointense tubular form in the center of the lesion.

To avoid misrepresentation of lesional signals stemming from inaccurate delineation of lesion boundaries, we focus our study to lesion centroids, defined as lesional tissue with the highest probabilities of being a lesion. By focusing the scope of our work to lesion centroids, we eschew the ability to incorporate the paramagnetic rim signal and lesion morphology into our assessment of diagnosis, but we are able to refine our template creation pipeline to highlight the central vein sign and amplify signals in the centers of lesions that may contain diagnostic value. Future iterations of this work may benefit from manual segmentation of distinct lesions or statistical methods for defining lesional boundaries that allow for more accurate measurements of the paramagnetic rim signal and lesion morphology alongside the central vein sign for a template lesion that truly encapsulates all of the imaging signals that are known to contain diagnostic value.

5.4. Summary

In this dissertation, we propose novel methodology for quantifying signals in neuroimaging data and developing methodologies for incorporating these signals into statistical analyses. Each of our three projects centers a different "level" on the neuroimaging analysis pipeline. First, we think about how best to take advantage of previously built, well-validated imaging biomarkers that may have been built to predict one particular outcome and how best to utilize those models in a setting that may not always directly study what the radiomic predictor was built to analyze. Next, we use high-dimensional feature extraction and machine learning classification models to develop a novel radiomic predictor that quantifies an imaging signal that has shown clinical promise. Finally, we develop new methodology for measuring disease severity by focusing in on the relationships between intensity and spatiality on a voxel-level scale. We show that orientation within a common coordinate space may help to improve detection of prognostic and diagnostic lesional imaging signals.

BIBLIOGRAPHY

Absinta, M., Sati, P., Fechner, A., Schindler, M.K., Nair, G. and Reich, D.S. (2018) Identification of Chronic Active Multiple Sclerosis Lesions on 3T MRI. *American Journal of Neuroradiology*, **39**, 1233–1238.

Absinta, M., Sati, P., Gaitán, M.I., Maggi, P., Cortese, I.C.M., Filippi, M., et al. (2013) Seven-tesla phase imaging of acute multiple sclerosis lesions: A new window into the inflammatory process. *Annals of Neurology*, **74**, 669–678.

Absinta, M., Sati, P., Masuzzo, F., Nair, G., Sethi, V., Kolb, H., et al. (2019) Association of Chronic Active Multiple Sclerosis Lesions With Disability In Vivo. *JAMA Neurology*, **76**, 1474–1483.

Absinta, M., Sati, P., Schindler, M., Leibovitch, E.C., Ohayon, J., Wu, T., et al. (2016) Persistent 7-tesla phase rim predicts poor outcome in new multiple sclerosis patient lesions. *The Journal of Clinical Investigation*, **126**, 2597–2609.

Amunts, K. (2014) The human brain project: neuroscience perspectives and German contributions. *e-Neuroforum*, **20**, 43–50.

Anderson, E., Grant, R., Lewis, S.C. and Whittle, I.R. (2008) Randomized Phase III controlled trials of therapy in malignant glioma: where are we after 40 years? *British Journal of Neurosurgery*, **22**, 339–349.

Anderson, R.M., Hadjichrysanthou, C., Evans, S. and Wong, M.M. (2017) Why do so many clinical trials of therapies for Alzheimer's disease fail? *The Lancet*, **390**, 2327–2329.

Avants, B.B., Epstein, C.L., Grossman, M. and Gee, J.C. (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, **12**, 26–41.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A. and Gee, J.C. (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, **54**, 2033–2044.

Avants, B.B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., et al. (2010) The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, **49**, 2457–2466.

Bagnato, F., Hametner, S., Yao, B., van Gelderen, P., Merkle, H., Cantor, F.K., et al. (2011) Tracking iron in multiple sclerosis: a combined imaging and histopathological study at 7 Tesla. *Brain: A Journal of Neurology*, **134**, 3602–3615.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al. (2017) Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, **4**, 170117.

Barquero, G., La Rosa, F., Kebiri, H., Lu, P.-J., Rahmanzadeh, R., Weigel, M., et al. (2020) RimNet: A deep 3D multimodal MRI architecture for paramagnetic rim lesion assessment in multiple sclerosis. *NeuroImage: Clinical*, **28**, 102412.

Bian, W., Harter, K., Hammond-Rosenbluth, K.E., Lupo, J.M., Xu, D., Kelley, D.A., et al. (2013) A serial in vivo 7T magnetic resonance phase imaging study of white matter lesions in multiple sclerosis. *Multiple Sclerosis (Houndmills, Basingstoke, England)*, **19**, 69–75.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

Chawla, S., Kister, I., Sinnecker, T., Wuerfel, J., Brisset, J.-C., Paul, F., et al. (2018) Longitudinal study of multiple sclerosis lesions using ultra-high field (7T) multiparametric MR imaging. *PLOS ONE*, **13**, e0202918.

Clarke, M.A., Pareto, D., Pessini-Ferreira, L., Arrambide, G., Alberich, M., Crescenzo, F., et al. (2020) Value of 3T Susceptibility-Weighted Imaging in the Diagnosis of Multiple Sclerosis. *American Journal of Neuroradiology*, **41**, 1001–1008.

Coroller, T.P., Agrawal, V., Narayan, V., Hou, Y., Grossmann, P., Lee, S.W., et al. (2016) Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiotherapy and Oncology*, **119**, 480–486.

Crane, P.K., Carle, A., Gibbons, L.E., Insel, P., Mackin, R.S., Gross, A., et al. (2012) Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, **6**, 502–516.

Cremers, H.R., Wager, T.D. and Yarkoni, T. (2017) The relation between statistical power and inference in fMRI. *PLOS ONE*, **12**, e0184923.

Cummings, J. (2018) Lessons Learned from Alzheimer Disease: Clinical Trials with Negative Outcomes. *Clinical and Translational Science*, **11**, 147–152.

Da, X., Toledo, J.B., Zee, J., Wolk, D.A., Xie, S.X., Ou, Y., et al. (2013) Integration and relative value of biomarkers for prediction of MCI to AD progression: Spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *NeuroImage : Clinical*, **4**, 164–173.

Dal-Bianco, A., Grabner, G., Kronnerwetter, C., Weber, M., Höftberger, R., Berger, T., et al. (2017) Slow expansion of multiple sclerosis iron rim lesions: pathology and 7 T magnetic resonance imaging. *Acta Neuropathologica*, **133**, 25–42.

Davatzikos, C. (2019) Machine learning in neuroimaging: Progress and challenges. *NeuroImage*, **197**, 652–656.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N. and Trojanowski, J.Q. (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, **32**, 2322.e19–27.

Davatzikos, C., Xu, F., An, Y., Fan, Y. and Resnick, S.M. (2009) Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain*, **132**, 2026–2035.

Donner, A. (1984) Approaches to sample size estimation in the design of clinical trials—a review. *Statistics in Medicine*, **3**, 199–214.

Doshi, J., Erus, G., Ou, Y., Gaonkar, B. and Davatzikos, C. (2013) Multi-Atlas Skull-Stripping. *Academic Radiology*, **20**, 1566–1576.

Dworkin, J.D., Linn, K.A., Oguz, I., Fleishman, G.M., Bakshi, R., Nair, G., et al. (2018a) An Automated Statistical Technique for Counting Distinct Multiple Sclerosis Lesions. *American Journal of Neuroradiology*.

Dworkin, J.D., Linn, K.A., Oguz, I., Fleishman, G.M., Bakshi, R., Nair, G., et al. (2018b) An Automated Statistical Technique for Counting Distinct Multiple Sclerosis Lesions. *American Journal of Neuroradiology*, **39**, 626–633.

Dworkin, J.D., Sati, P., Solomon, A., Pham, D.L., Watts, R., Martin, M.L., et al. (2018c) Automated Integration of Multimodal MRI for the Probabilistic Detection of the Central Vein Sign in White Matter Lesions. *American Journal of Neuroradiology*, **39**, 1806–1813.

Egan, M.F., Kost, J., Voss, T., Mukai, Y., Aisen, P.S., Cummings, J.L., et al. (2019) Randomized Trial of Verubecestat for Prodromal Alzheimer's Disease. *New England Journal of Medicine*, **380**, 1408–1420.

Eklund, A., Nichols, T.E. and Knutsson, H. (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, **113**, 7900–7905.

Eskreis-Winkler, S., Deh, K., Gupta, A., Liu, T., Wisnieff, C., Jin, M., et al. (2015) Multiple sclerosis lesion geometry in quantitative susceptibility mapping (QSM) and phase imaging. *Journal of Magnetic Resonance Imaging*, **42**, 224–229.

Eskreis-Winkler, S., Zhang, Y., Zhang, J., Liu, Z., Dimov, A., Gupta, A., et al. (2017) The clinical utility of QSM: disease diagnosis, medical management, and surgical planning. *NMR in Biomedicine*, **30**, e3668.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., and Alzheimer's Disease Neuroimaging Initiative. (2008) Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, **39**, 1731–1743.

Faul, F., Erdfelder, E., Lang, A.-G. and Buchner, A. (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, **39**, 175–191.

Fazekas, F., Barkhof, F., Filippi, M., Grossman, R.I., Li, D.K.B., McDonald, W.I., et al. (1999) The contribution of magnetic resonance imaging to the diagnosis of multiple sclerosis. *Neurology*, **53**, 448–448.

Filippi, M. and Agosta, F. (2010) Imaging biomarkers in multiple sclerosis. *Journal of Magnetic Resonance Imaging*, **31**, 770–788.

Fleiss, J.L. (2011) *Design and Analysis of Clinical Experiments*. John Wiley & Sons.

Frangi, A.F., Niessen, W.J., Vincken, K.L. and Viergever, M.A. (1998) Multiscale vessel enhancement filtering. *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98* Lecture Notes in Computer Science. (eds W.M. Wells),, A. Colchester), & S. Delp), pp. 130–137. Springer Berlin Heidelberg, Berlin, Heidelberg.

Frischer, J.M., Weigand, S.D., Guo, Y., Kale, N., Parisi, J.E., Pirko, I., et al. (2015) Clinical and pathological insights into the dynamic nature of the white matter multiple sclerosis plaque. *Annals of Neurology*, **78**, 710–721.

Gaitán, M.I. and Correale, J. (2019) Multiple Sclerosis Misdiagnosis: A Persistent Problem to Solve. *Frontiers in Neurology*, **10**.

Gillen, K.M., Mubarak, M., Park, C., Ponath, G., Zhang, S., Dimov, A., et al. (2021) QSM is an imaging biomarker for chronic glial activation in multiple sclerosis lesions. *Annals of Clinical and Translational Neurology*, **8**, 877–886.

Hammond, K.E., Metcalf, M., Carvajal, L., Okuda, D.T., Srinivasan, R., Vigneron, D., et al. (2008) Quantitative in vivo magnetic resonance imaging of multiple sclerosis at 7 Tesla with sensitivity to iron. *Annals of Neurology*, **64**, 707–713.

Hammoud, M.A., Sawaya, R., Shi, W., Thall, P.F. and Leeds, N.E. (1996) Prognostic significance of preoperative MRI scans in glioblastoma multiforme. *Journal of Neuro-Oncology*, **27**, 65–73.

Herrlinger, U., Tzaridis, T., Mack, F., Steinbach, J.P., Schlegel, U., Sabel, M., et al. (2019) Lomustine-temozolomide combination therapy versus standard temozolomide therapy in patients with newly diagnosed glioblastoma with methylated MGMT promoter (CeTeG/NOA–09): a randomised, open-label, phase 3 trial. *The Lancet*, **393**, 678–688.

Honig, L.S., Vellas, B., Woodward, M., Boada, M., Bullock, R., Borrie, M., et al. (2018) Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. *New England Journal of Medicine*, **378**, 321–330.

Insel, T.R., Landis, S.C. and Collins, F.S. (2013) The NIH BRAIN Initiative. *Science*, **340**, 687–688.

Jang, J., Nam, Y., Choi, Y., Shin, N.-Y., An, J.Y., Ahn, K.-J., et al. (2020) Paramagnetic Rims in Multiple Sclerosis and Neuromyelitis Optica Spectrum Disorder: A Quantitative Susceptibility Mapping Study with 3-T MRI. *Journal of Clinical Neurology (Seoul, Korea)*, **16**, 562–572.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W. and Smith, S.M. (2012) FSL. *NeuroImage*, **62**, 782–790.

Jiang, T. (2013) Brainnetome: A new -ome to understand the brain and its disorders. *NeuroImage*, **80**, 263–272.

Jordan, S., Watkins, A., Storey, M., Allen, S.J., Brooks, C.J., Garaiova, I., et al. (2013) Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis. *PLoS ONE*, **8**, e67912.

Kaisey, M., Solomon, A.J., Luu, M., Giesser, B.S. and Sicotte, N.L. (2019) Incidence of multiple sclerosis misdiagnosis in referrals to two academic centers. *Multiple Sclerosis and Related Disorders*, **30**, 51–56.

Kaunzner, U.W., Kang, Y., Zhang, S., Morris, E., Yao, Y., Pandya, S., et al. (2019) Quantitative susceptibility mapping identifies inflammation in a subset of chronic multiple sclerosis lesions. *Brain: A Journal of Neurology*, **142**, 133–145.

Kent, D.M., Paulus, J.K., van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., et al. (2019) The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Annals of Internal Medicine*, **172**, 35–45.

Kilsdonk, I.D., Lopez-Soriano, A., Kuijer, J.P.A., de Graaf, W.L., Castelijns, J.A., Polman, C.H., et al. (2014) Morphological features of MS lesions on FLAIR* at 7 T and their relation to patient characteristics. *Journal of Neurology*, **261**, 1356–1364.

Kirby, A., Gebski, V. and Keech, A.C. Determining the sample size in a clinical trial. , 2.

Kolossváry, M., Karády, J., Szilveszter, B., Kitslaar, P., Hoffmann, U., Merkely, B., et al. (2017) Radiomic Features Are Superior to Conventional Quantitative Computed Tomographic Metrics to Identify Coronary Plaques With Napkin-Ring Sign. *Circulation. Cardiovascular Imaging*, **10**.

Kuhn, M. (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28**, 1–26.

Kuznetsova, A., Brockhoff, P.B. and Christensen, R.H.B. (2017) lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, **82**, 1–26.

Liu, Y., Kim, J., Balagurunathan, Y., Li, Q., Garcia, A.L., Stringfield, O., et al. (2016) Radiomic Features Are Associated With EGFR Mutation Status in Lung Adenocarcinomas. *Clinical Lung Cancer*, **17**, 441-448.e6.

Lou, C., Habes, M., Illenberger, N.A., Ezzati, A., Lipton, R.B., Shaw, P.A., et al. (2021a) Leveraging machine learning predictive biomarkers to augment the statistical power of clinical trials with baseline magnetic resonance imaging. *Brain Communications*, **3**, fcab264.

Lou, C., Sati, P., Absinta, M., Clark, K., Dworkin, J.D., Valcarcel, A.M., et al. (2021b) Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3T susceptibility-based MR imaging. *NeuroImage: Clinical*, **32**, 102796.

Luchetti, S., Fransen, N.L., van Eden, C.G., Ramaglia, V., Mason, M. and Huitinga, I. (2018) Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: a retrospective autopsy cohort analysis. *Acta Neuropathologica*, **135**, 511–528.

Macyszyn, L., Akbari, H., Pisapia, J.M., Da, X., Attiah, M., Pigrish, V., et al. (2016) Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-Oncology*, **18**, 417–425.

Maggi, P., Fartaria, M.J., Jorge, J., La Rosa, F., Absinta, M., Sati, P., et al. (2020a) CVSnet: A machine learning approach for automated central vein sign assessment in multiple sclerosis. *NMR in Biomedicine*, **33**, e4283.

Maggi, P., Sati, P., Nair, G., Cortese, I.C.M., Jacobson, S., Smith, B.R., et al. (2020b) Paramagnetic Rim Lesions are Specific to Multiple Sclerosis: An International Multicenter 3T MRI Study. *Annals of Neurology*, **88**, 1034–1042.

McGranahan, T., Therkelsen, K.E., Ahmad, S. and Nagpal, S. (2019) Current State of Immunotherapy for Treatment of Glioblastoma. *Current Treatment Options in Oncology*, **20**, 24.

Mehta, V., Pei, W., Yang, G., Li, S., Swamy, E., Boster, A., et al. (2013) Iron Is a Sensitive Biomarker for Inflammation in Multiple Sclerosis Lesions. *PLOS ONE*, **8**, e57573.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015) The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE transactions on medical imaging*, **34**, 1993–2024.

Mintun, M.A., Lo, A.C., Duggan Evans, C., Wessels, A.M., Ardayfio, P.A., Andersen, S.W., et al. (2021) Donanemab in Early Alzheimer’s Disease. *New England Journal of Medicine*, **384**, 1691–1704.

Muschelli, J., Sweeney, E., Lindquist, M. and Crainiceanu, C. (2015) fslr: Connecting the FSL Software with R. *The R journal*, **7**, 163–175.

Newton, B.D., Wright, K., Winkler, M.D., Bovis, F., Takahashi, M., Dimitrov, I.E., et al. (2017) Three-Dimensional Shape and Surface Features Distinguish Multiple Sclerosis Lesions from Nonspecific White Matter Disease. *Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging*, **27**, 613–619.

O’Connell, N.S., Dai, L., Jiang, Y., Speiser, J.L., Ward, R., Wei, W., et al. (2017) Methods for Analysis of Pre-Post Data in Clinical Research: A Comparison of Five Common Methods. *Journal of Biometrics & Biostatistics*, **8**, 1–8.

Okuda, D.T., Moog, T.M., McCreary, M., Bachand, J.N., Wilson, A., Wright, K., et al. (2020) Utility of shape evolution and displacement in the classification of chronic multiple sclerosis lesions. *Scientific Reports*, **10**, 19560.

Ontaneda, D., Sati, P., Raza, P., Kilbane, M., Gombos, E., Alvarez, E., et al. (2021) Central vein sign: A diagnostic biomarker in multiple sclerosis (CAVS-MS) study protocol for a prospective multicenter trial. *NeuroImage: Clinical*, **32**, 102834.

Oxford, A.E., Stewart, E.S. and Rohn, T.T. (2020) Clinical Trials in Alzheimer’s Disease: A Hurdle in the Path of Remedy. *International Journal of Alzheimer’s Disease*, **2020**.

Perry, J.R., Laperriere, N., O'Callaghan, C.J., Brandes, A.A., Menten, J., Phillips, C., et al. (2017) Short-Course Radiation plus Temozolomide in Elderly Patients with Glioblastoma. *New England Journal of Medicine*, **376**, 1027–1037.

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., et al. (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology*, **74**, 201–209.

Pocock, S.J. (1976) The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, **29**, 175–188.

Popescu, V., Agosta, F., Hulst, H.E., Sluimer, I.C., Knol, D.L., Sormani, M.P., et al. (2013) Brain atrophy and lesion load predict long term disability in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, **84**, 1082–1091.

Rekkas, A., Paulus, J.K., Raman, G., Wong, J.B., Steyerberg, E.W., Rijnbeek, P.R., et al. (2020) Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Medical Research Methodology*, **20**, 264.

Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A.G., et al. (2018) Radiomics: the facts and the challenges of image analysis. *European Radiology Experimental*, **2**.

Sahraian, M.A. and Radü, E.-W. (2007) *MRI Atlas of MS Lesions*. Springer Science & Business Media.

Sati, P., Oh, J., Constable, R.T., Evangelou, N., Guttmann, C.R.G., Henry, R.G., et al. (2016) The central vein sign and its clinical evaluation for the diagnosis of multiple sclerosis: a consensus statement from the North American Imaging in Multiple Sclerosis Cooperative. *Nature Reviews Neurology*, **12**, 714–722.

Sati, P., Thomasson, D.M., Li, N., Pham, D.L., Biassou, N.M., Reich, D.S., et al. (2014a) Rapid, high-resolution, whole-brain, susceptibility-based MRI of multiple sclerosis. *Multiple Sclerosis (Houndmills, Basingstoke, England)*, **20**, 1464–1470.

Sati, P., Thomasson, D., Li, N., Pham, D., Biassou, N., Reich, D., et al. (2014b) Rapid, high-resolution, whole-brain, susceptibility-based MRI of multiple sclerosis. *Multiple Sclerosis Journal*, **20**, 1464–1470.

Satterthwaite, F.E. (1946) An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, **2**, 110–114.

Shaffer, J.L., Petrella, J.R., Sheldon, F.C., Choudhury, K.R., Calhoun, V.D., Coleman, R.E., et al. (2013) Predicting Cognitive Decline in Subjects at Risk for Alzheimer Disease by Using Combined Cerebrospinal Fluid, MR Imaging, and PET Biomarkers. *Radiology*, **266**, 583–591.

Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., et al. (2014) Statistical normalization techniques for magnetic resonance imaging. *NeuroImage : Clinical*, **6**, 9–19.

Sinnecker, T., Clarke, M.A., Meier, D., Enzinger, C., Calabrese, M., De Stefano, N., et al. (2019) Evaluation of the Central Vein Sign as a Diagnostic Imaging Biomarker in Multiple Sclerosis. *JAMA Neurology*, **76**, 1446–1456.

Sinnecker, T., Dörr, J., Pfueller, C.F., Harms, L., Ruprecht, K., Jarius, S., et al. (2012) Distinct lesion morphology at 7-T MRI differentiates neuromyelitis optica from multiple sclerosis. *Neurology*, **79**, 708–714.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., et al. (2004) Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, **23**, S208–S219.

Solomon, A.J., Bourdette, D.N., Cross, A.H., Applebee, A., Skidd, P.M., Howard, D.B., et al. (2016) The contemporary spectrum of multiple sclerosis misdiagnosis. *Neurology*, **87**, 1393–1399.

Solomon, A.J., Klein, E.P. and Bourdette, D. (2012) “Undiagnosing” multiple sclerosis. *Neurology*, **78**, 1986–1991.

Solomon, A.J., Watts, R., Ontaneda, D., Absinta, M., Sati, P. and Reich, D.S. (2018) Diagnostic performance of central vein sign for multiple sclerosis with a simplified three-lesion algorithm. *Multiple Sclerosis Journal*, **24**, 750–757.

Sormani, M.P. and Bruzzi, P. (2013) MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. *The Lancet Neurology*, **12**, 669–676.

Sporns, O., Tononi, G. and Kötter, R. (2005) The Human Connectome: A Structural Description of the Human Brain. *PLOS Computational Biology*, **1**, e42.

Stüber, C., Pitt, D. and Wang, Y. (2016) Iron in Multiple Sclerosis and Its Noninvasive Imaging with Quantitative Susceptibility Mapping. *International Journal of Molecular Sciences*, **17**.

Sweeney, E.M., Nguyen, T.D., Kuceyeski, A., Ryan, S.M., Zhang, S., Zexter, L., et al. (2021) Estimation of Multiple Sclerosis lesion age on magnetic resonance imaging. *NeuroImage*, **225**, 117451.

Thaler, C., Faizy, T., Sedlacik, J., Holst, B., Stellmann, J.-P., Young, K.L., et al. (2015) T1-Thresholds in Black Holes Increase Clinical-Radiological Correlation in Multiple Sclerosis Patients. *PLOS ONE*, **10**, e0144693.

Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., et al. (2018) Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, **17**, 162–173.

Tozlu, C., Jamison, K., Nguyen, T., Zinger, N., Kaunzner, U., Pandya, S., et al. (2021) Structural disconnectivity from paramagnetic rim lesions is related to disability in multiple sclerosis. *Brain and Behavior*, **11**, e2353.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., et al. (2010) N4ITK: Improved N3 Bias Correction. *IEEE transactions on medical imaging*, **29**, 1310–1320.

Valcarcel, A.M., Linn, K.A., Khalid, F., Vandekar, S.N., Tauhid, S., Satterthwaite, T.D., et al. (2018a) A dual modeling approach to automatic segmentation of cerebral T2 hyperintensities and T1 black holes in multiple sclerosis. *NeuroImage: Clinical*, **20**, 1211–1221.

Valcarcel, A.M., Linn, K.A., Vandekar, S.N., Satterthwaite, T.D., Muschelli, J., Calabresi, P.A., et al. (2018b) MIMoSA: An Automated Method for Inter-Modal Segmentation Analysis of Multiple Sclerosis Brain Lesions. *Journal of neuroimaging : official journal of the American Society of Neuroimaging*, **28**, 389–398.

Valcarcel, A.M., Muschelli, J., Pham, D.L., Martin, M.L., Yushkevich, P., Brandstadter, R., et al. (2020) TAPAS: A Thresholding Approach for Probability Map Automatic Segmentation in Multiple Sclerosis. *NeuroImage: Clinical*, **27**, 102256.

Vandekar, S.N. and Stephens, J. (2021) Improving the replicability of neuroimaging findings by thresholding effect sizes instead of p-values. *Human Brain Mapping*, **42**, 2393–2398.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., et al. (2014) Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, **13**, 41–54.

Wirsching, H.-G., Roelcke, U., Weller, J., Hundsberger, T., Hottinger, A.F., Moos, R. von, et al. (2021) MRI and 18FET-PET Predict Survival Benefit from Bevacizumab Plus Radiotherapy in Patients with Isocitrate Dehydrogenase Wild-type Glioblastoma: Results from the Randomized ARTE Trial. *Clinical Cancer Research*, **27**, 179–188.

Wisnieff, C., Ramanan, S., Olesik, J., Gauthier, S., Wang, Y. and Pitt, D. (2015) Quantitative susceptibility mapping (QSM) of white matter multiple sclerosis lesions: Interpreting positive susceptibility and the presence of iron. *Magnetic Resonance in Medicine*, **74**, 564–570.

Wittes, J. (2002) Sample size calculations for randomized controlled trials. *Epidemiologic Reviews*, **24**, 39–53.

Wright, M.N. and Ziegler, A. (2017) ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**, 1–17.

Wuerfel, J., Sinnecker, T., Ringelstein, E.B., Jarius, S., Schwindt, W., Niendorf, T., et al. (2012) Lesion morphology at 7 Tesla MRI differentiates Susac syndrome from multiple sclerosis. *Multiple Sclerosis Journal*, **18**, 1592–1599.

Yamout, B.I., Khoury, S.J., Ayyoubi, N., Doumiati, H., Fakhreddine, M., Ahmed, S.F., et al. (2017) Alternative diagnoses in patients referred to specialized centers for suspected MS. *Multiple Sclerosis and Related Disorders*, **18**, 85–89.

Yao, B., Bagnato, F., Matsuura, E., Merkle, H., van Gelderen, P., Cantor, F.K., et al. (2012) Chronic multiple sclerosis lesions: characterization with high-field-strength MR imaging. *Radiology*, **262**, 206–215.

Youden, W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., et al. (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*, **31**, 1116–1128.

Zhang, Y., Brady, J.M. and Smith, S. (2000) Hidden Markov random field model for segmentation of brain MR image. *Medical Imaging 2000: Image Processing* pp. 1126–1137. International Society for Optics and Photonics.

Zhang, S., Nguyen, T.D., Zhao, Y., Gauthier, S.A., Wang, Y. and Gupta, A. (2018) Diagnostic accuracy of semiautomatic lesion detection plus quantitative susceptibility mapping in the identification of new and enhancing multiple sclerosis lesions. *NeuroImage: Clinical*, **18**, 143–148.