

© Copyright 2022

Harkirat Kaur Sohi

Understanding the differences in cognitively defined subgroups in Alzheimer's disease:

A data science approach

Harkirat Kaur Sohi

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

John H. Gennari, Chair

Paul K. Crane

Ellen M. Wijsman

Program Authorized to Offer Degree:

Biomedical & Health Informatics

University of Washington

Abstract

Understanding the differences in cognitively defined subgroups in Alzheimer's disease:

A data science approach

Harkirat Kaur Sohi

Chair of the Supervisory Committee:

Graduate Program Director & Associate Professor John H. Gennari

Biomedical Informatics and Medical Education

My work connects two types of data in Alzheimer's Disease (AD): structural MRI data from Alzheimer's Disease Neuroimaging Initiative (ADNI) and cognition data in the form of AD subgroups. The subgroups (AD-Executive, AD-Language, AD-Memory and AD-Visuospatial), defined by Crane et al. (2017), are based on cognitive test scores from the time of AD diagnosis, and each subgroup is characterized by marked impairment in the specified cognitive domain relative to the other domains. My dissertation's focus is on data science and mathematical methods to understand how volumes of 70 brain regions of interest (ROIs) might differ across pairs of AD subgroups in cross-sectional data in time, specifically data from the time of AD diagnosis (Aim 1) and in longitudinal data (Aim 2). My work demonstrates a careful assessment and implementation of methods to best utilize the data available that is currently small in sample size, with imbalanced AD subgroup sizes and noisy in nature. In both aims, the following pairs of AD subgroups were compared: a.) AD-Language vs. AD-Memory, b.) AD-Memory vs. AD-

Visuospatial and c.) AD-Language vs. AD-Visuospatial. The AD-Executive group was excluded from the current analyses due to its small sample size.

In Aim 1, I explored supervised machine learning classification methods that provide insight into variable importance for identifying the most important brain ROIs for distinguishing between pairs of AD subgroups. I determined random forest to be the most appropriate method for this task, given the characteristics of the data. Prior to building classification models, I addressed specific challenges in cross-sectional data: potential noise due to non-ROI variables and imbalanced AD subgroup sizes. A challenge in using classification models in the domain of AD subgroups is that there is no gold standard for knowing how separable the AD subgroups are based on ROI volumes. The work presented here may be the first to establish a starting benchmark for classification accuracies for distinguishing between pairs of AD subgroups based on ROI volumes, although these models are not intended to be used for prediction in a clinical setting but rather to understand which brain regions are most important to distinguish the AD subgroups.

In Aim 2, I used linear mixed effects (LME) modeling on longitudinal data to determine which of the 70 ROIs' volume trajectories differ the most across pairs of AD subgroups in terms of longitudinal volume and rate of change of volume with respect to time. First, I laid out criteria for using data from specific MRI scans in an effort to reduce noise in data, instead of using the default longitudinal dataset. Given the small sample size of the AD subgroups and irregular data, I implemented LME modeling for each ROI on the original dataset consisting of all time points and also on a series of subsets of data that were obtained by restricting each AD subgroup's data to time points with a specific minimum number of subjects available.

An important finding of my work is that there was some overlap in the top ROIs that were determined to be important based on cross-sectional and longitudinal data analyses, for distinguishing between pairs of AD subgroups. Results from my Ph.D. work have potential implications for decisions about which brain regions may be relevant for future neuropathological studies in studying AD subgroups.

Acknowledgements

I would like to convey my sincere gratitude to many individuals who supported me during my Ph.D. journey.

A very special thanks to my advisor and mentor **Dr. John Gennari** for guiding me on the road of an interdisciplinary and collaborative project. Thank you for being an excellent mentor and for always making a welcoming space for me to express my ideas or concerns and for a positive outlook and excitement towards my work. I appreciated your regular availability through weekly meetings, check-ins and communication. I consider myself very lucky to have you as a mentor as your mentorship style worked very well with my working style. You modeled professionalism, respectfulness, openness to listening and listening to a student like a colleague. I appreciate that it is easy to approach you and I could always count on having a positive outcome from a meeting or discussion with you. Thanks for encouraging me to carve out time to think about future career prospects during the Ph.D. and working closely with me as I prepared for specific job interviews. Thanks for your detailed feedback on my dissertation thesis writing. It has been a pleasure to have you as my research advisor!

Thank you, **Dr. Paul Crane**, for taking a leap of faith in me as a student in BHI as we established this collaboration to combine biomedical informatics and Alzheimer's disease (AD) research. Thanks for your mentorship, your excitement towards our collaboration and for all the support in so many ways. Thank you for also providing a balance of perspectives and feedback on my work so that it fits in the context of the scientific implications and questions in Alzheimer's disease. I am especially thankful for the funding I received as a Research Assistant

for 3+ years from a grant headed by Dr. Crane. Thank you for the tips and advice for job interviews, as well as your detailed feedback on my dissertation thesis.

Thank you, **Dr. Ellen Wijsman**, for not only being the Graduate Student Representative on my Ph.D. committee, but also for funding me through a Research Assistantship in your lab during my first year at BHI. Thank you for being a long term mentor. I had the great opportunity to work in Dr. Wijsman's lab in a full time job before starting my Ph.D. Thank you for training me in so many necessary skills to be a good scientist before I embarked upon my Ph.D. journey. Wijsman lab's work environment introduced me to inspiring, intelligent and really nice folks who made me want to continue and grow further in my scientific career. Thanks also for the input on the methods and scientific insights for my dissertation work, along with your detailed feedback on my dissertation thesis.

Thank you, **Dr. Ali Shojaie**, for being my "to-go" person for all my machine learning and statistical methods questions. I cannot appreciate your help and input enough. I learned a lot in your 'BIOSTAT 546: Machine Learning for Biomedical and Public Health Big Data' class and it was one of the reasons why I felt inspired to have a machine learning component in my Ph.D. work.

Thank you, **Dr. David Crosslin**, for also helping me with statistical methods questions and for offering your time over zoom meetings despite the time zone differences. I appreciated all the encouragement and your willingness to help whenever you could.

Thank you **Dr. Shuai Huang** for providing feedback on some of my work and ideas in the exploratory stages of my dissertation work. The meetings provided me with some general excitement about ML methods in the field of AD research.

A special thanks to **Dr. Shannon Risacher** at Indiana University for providing the processed datasets that I used in my Ph.D. work. Thanks to **Dr. Rik Ossenkopele** and **Dr. Collin Groot** at Vrije Universiteit (VU) Amsterdam, **Dr. Shubhabrata Mukherjee**, **Dr. Christine MacDonald** and **Dr. Ariel Rokem** at the University of Washington (UW) for answering many questions and providing consultation as needed.

I would especially like to thank the **Alzheimer's Disease Neuroimaging Initiative (ADNI)** for providing access to the study's data which I used in my work. Thanks to all who have made the ADNI study possible and have kept it running at various levels: recruitment of individuals, data collection, data processing and database management. My Ph.D. work would not be possible without the individuals who have volunteered to be part of the ADNI study.

Thanks to many professors in my home department, Department of Biomedical & Health Informatics as well as other departments at UW for the great conversations and learning experiences! Thanks to **Heidi Kelm**, **Lora Brewsaugh**, **Jill Fulmore** and **Marni Levy** for answering so many questions and providing administrative support.

Thanks to friends who have especially checked-in with me in the last few months as I was getting closer to the finish line. A special shout-out to those who gave me feedback as I prepared for my job talk and Ph.D. defense talk.

Last, but not least, I want to thank my family for the encouragement and for cheering me on in this long journey. The list of gratitude and blessings is long but I will highlight my mother's home-cooked food and my father making many trips to bring the food to me in my most busy days of the Ph.D. Thanks to my brother who I could always count on (even on the phone) if I had any technical or software issues. I must mention I learned to write my first PERL script many years ago because of my brother's help who is four years younger than me. I want to

thank my grandparents for all their love, for believing in me and for inspiring me. To my aunt Sabi bhooa, who along with my parents, played an important role in my early childhood for laying a good foundation for my learning/educational experiences.

Thanks to so many wonderful teachers and mentors in my educational journey who encouraged my curiosity, met my questions with excitement and provided the right environment for me to continue to grow as a learner.

Table of Contents

Chapter 1: Overview, Significance, Impact & Background	7
1.1 Background: Cognitive subgroups, brain regions and ADNI	16
1.1.1. Clinical heterogeneity and AD Subgroups.....	16
1.1.2. Brain regions of interest.....	18
1.1.3. ADNI and FreeSurfer data	23
1.2 Organization of the dissertation work.....	28
Chapter 2: Aim 1 approach rationale, data and setting up the classification problem	31
2.1 Why I used machine learning methods and why I chose the ones I did?.....	33
2.2 Data for model building and the general classification problem in Aim 1	36
2.3 Cross-sectional data acquisition, pre-processing and cleaning	38
2.4 Training and Test data.....	41
2.5 Transformation of volume variables: Regressing out the effects of extraneous variables.....	43
2.6 SMOTE oversampling to balance AD-subgroup sample sizes	45
Chapter 3: Determining ROI importance in cross-sectional data using penalized logistic regression	
[Aim 1]	47
3.1 Why logistic regression based models as the first analysis method?	47
3.2 An introduction to logistic regression with regularization.....	48
3.3 Model building.....	55
3.3.1 Training and Test data.....	55
3.3.2 Cross-validation on training data to select tuning parameters α and λ	55
3.4 Model coefficients and model accuracy.....	64
3.4.1 AD-Language vs. AD-Memory	65
3.4.2. AD-Memory vs. AD-Visuospatial	70
3.4.3 AD-Language vs. AD-Visuospatial	74

3.5 Limitations of results of variable importance results based on penalized logistic regression.....	79
Chapter 4: ROI importance based on Random Forest binary classification models distinguishing between pairs of AD subgroups [Aim 1]	82
4.1 Why Random Forest?.....	82
4.2 Implementation of Random Forest.....	83
4.3 Variable importance results from Random Forest	85
4.4 Understanding how the ROIs identified as important differ across the AD subgroups	96
4.5 Limitations of the current approach, alternative solutions & next steps	103
4.6 Conclusions.....	107
Chapter 5: Overview of Longitudinal data analysis [Aim 2]: the scientific question, data challenges and workflow	112
5.1 Descriptive analysis of longitudinal MRI ADNI data: challenges of the data.....	113
5.2 Workflow for analysis of longitudinal data	132
Chapter 6: Assessing differences in ROI volume trajectories for pairs of AD subgroups using Linear Mixed Effects modeling [Aim 2]	134
6.1 Overview of Linear Mixed Effects modeling	135
6.2 Analysis design	141
6.3 Results and Discussion.....	152
6.3.1. AD-Language vs. AD-Memory	152
6.3.2 AD-Memory vs. AD-Visuospatial	172
6.3.3 AD-Language vs. AD-Visuospatial	185
6.4 Concluding remarks about longitudinal data analysis and future work	198
Chapter 7: Conclusions, Limitations & Future Work	204
Appendix A: Supplemental materials (Code and files)	221
Appendix B: Longitudinal data descriptive plots	221
References.....	223

Chapter 1: Overview, Significance, Impact & Background

Alzheimer's disease (AD) is a neurodegenerative condition that leads to cognitive decline and is the most common cause of dementia ("2021 Alzheimer's Disease Facts and Figures" 2021). AD affects many cognitive functions including memory, language skills, visual perception, problem solving, self-management, the ability to focus and pay attention and in some cases even emotion regulation ("What Is Dementia? Symptoms, Types, and Diagnosis" n.d.). Pathologically, AD is characterized by the deposition of the protein fragment beta-amyloid plaques outside neurons and the formation of neurofibrillary tangles inside neurons that are an accumulation of an abnormal form of the tau protein; both are believed to lead to neuronal cell damage and cell death. Over time, the above processes lead to a loss of connections between networks of neurons in the brain, leading to substantial loss of brain volume by the final stages of the disease ("What Happens to the Brain in Alzheimer's Disease?" n.d.). Worldwide, there were over 55 million people living with AD in 2021 and the number is estimated to be 78 million by 2030 (World Alzheimer Report 2021). As of 2021, an estimated 6.2 million Americans aged 65 years and older had AD dementia. AD is the sixth leading cause of death in the United States and the fifth leading cause of death for individuals ≥ 65 years of age (World Alzheimer Report 2021).

Recent work by Crane et al. has binned individuals with Alzheimer's disease (AD) into cognitively defined subgroups based on cognitive data from the earliest time point following the diagnosis of clinical AD dementia (Crane et al. 2017). These subgroups include AD-Executive, AD-Language, AD-Memory and AD-Visuospatial. Each group is characterized by marked impairment in the specified cognitive domain relative to other domains. This idea challenges the current nosology of AD and research and drug development strategies that implicitly assume

homogeneity in AD. Subgrouping AD into phenotypically distinct categories such as the above subgroups can be the basis of future targeted drug design and specific treatment options for the different disease variants.

My dissertation combined two types of AD data: (1) cognitively defined subgroups by Crane et al. representing different behavioral phenotypes within AD and (2) brain structural MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to understand which brain regions may be important in understanding AD subgroup differences; see figure 1. Specifically, I took a data science approach involving machine learning and other mathematical methods to connect the two knowledge domains (cognitive and imaging data) in AD, with the goal of shedding light on the underlying physiological differences in the brain across the different AD subgroups. If the cognitively defined AD subgroups also show biological differences in the brain, then these differences would be worth investigating in future precision medicine approaches in treating AD. One of the first steps in this investigation may be neuropathological studies. Through my work, I've hoped to provide insight into which brain regions to focus on for understanding AD subgroup differences in future neuropathological studies.

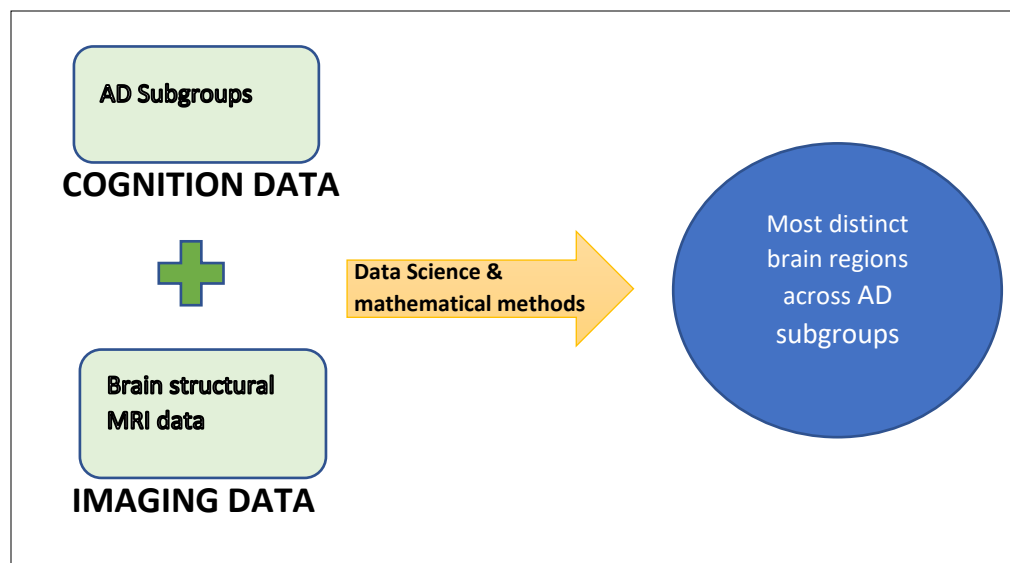


Figure 1: A graphical schematic of the overall goal of my dissertation.

There are two types of structural MRI data that I analyzed in my dissertation. The first is called cross-sectional data which refers to data at a single time point; in the context of the ADNI study, this translates to data from a single clinical MRI visit in the study. (See section 1.2.c for an overview of the ADNI study.) The second is longitudinal data which is data collected over a period of time. This corresponds to data from two or more clinical MRI visits for an individual during the study. Hence, my work is divided into two parts: first, an analysis of cross-sectional data, specifically data from the earliest time point following the diagnosis of clinical AD dementia, which I refer to as t_0 (Aim 1; Chapters 2-4) and second, an analysis of longitudinal data including t_0 (Aim 2; Chapters 6 and 7). In both cases, the AD subgroup assignments were based on cognitive data from t_0 . The common goal in both aims was to understand the differences across AD subgroups based on structural MRI data, specifically focusing on volumes of 70 brain regions of interest (ROI) that are discussed in 1.2.b. I compared the following AD subgroups: AD-Language, AD-Memory and AD-Visuospatial. The AD-Executive group was

excluded from the analysis due to its very small sample size; the details of sample sizes for all AD subgroups are discussed in 1.2.c.

My Ph.D. work focuses on applications of existing data science methods in a new domain (comparison of AD subgroups) and emphasizes a careful selection of methods to deal with the limitations and challenges of biomedical data from an ongoing clinical study. These limitations in the data made the task of looking for differences in the AD subgroups a challenging one.

Although ADNI is a large study and perhaps one of the biggest studies in terms of MRI collection and following individuals with AD over time, the sample sizes of the AD subgroups are still small from the perspective of machine learning and statistical methods. This is one of the biggest challenges of the current cross-sectional and longitudinal data that I analyzed. ADNI consists of cognitively normal, Mild Cognitive Impairment (MCI) and AD subjects. The division of the AD subjects into subgroups results in a relatively small sample size for each AD subgroup of interest. The small sample size is not something that can be resolved quickly or easily.

Although acquiring additional data for analysis can be challenging in any field, it is especially hard for MRI data and for a demographic of individuals who are older with physical and cognitive limitations. First, participants must be recruited into a study. Second, the cost of even a single MRI is substantial – both monetary costs and participant and staffing cost. Other common challenges of cross-sectional and longitudinal datasets were imbalance in AD subgroup sizes and several extraneous variables such as changing scanner properties over time and different testing sites, individuals' age, sex, etc. contributing to variation in ROI volumes observed across individuals, possibly diluting the signal (differences in brain regions) associated with AD subgroup differences. Additionally, in the current data, there is evidence of non-constant and unpredictable variation in the reported volumes for a given ROI from multiple MRI scans from

the same clinical visit for any given individual. This could be due to MRI measurement error as well as errors accumulated in the pre-processing pipeline for obtaining the volumes of ROIs from the MRI images. Together, these aspects of data made the problem of detecting differences in brain regions across AD subgroups a challenging one. While acknowledging the challenges and limitations of the current cross-sectional and longitudinal datasets, the goal of my Ph.D. work was to make the best of the current data at hand. A consistent theme of my Ph.D. work has been carefully selecting methods that are best suited for the characteristics of the data and methods that allow one to discover meaningful relationships in data despite the limitations of the data. Below, I briefly describe the scientific questions, analysis methods and findings of the two parts (Aim 1 and Aim 2) of my work. These questions rest upon the assumption that the AD subgroups defined by Crate et al. are indeed distinct enough groups.

In cross-sectional data analysis (Aim 1), I compared pairs of AD subgroups by constructing binary classification models using structural MRI data for the following: a.) AD-Language vs. AD-Memory, b.) AD-Memory vs. AD-Visuospatial and c.) AD-Language vs. AD-Visuospatial. The intent in developing classification models was to use the models to understand which brain ROI volume variables are important for distinguishing between the subgroups rather than using the models in a clinical setting to make classification decisions. To accomplish this, I explored two machine learning methods (logistic regression with elastic net penalty and random forest) that qualified as good candidates for providing some insight into variable importance. I determined random forest models to be a better fit for the data of interest due to theoretical reasons as well as data driven reasons which are discussed in Chapter 4. Using random forest, I obtained, for each classification, the topmost regions that have the greatest contribution in reducing the classification error. Through this analysis, I provided a ranking of brain regions that are most important in two

specific subgroup comparisons: a) AD-Language vs. AD-Memory: **right entorhinal cortex, right hippocampus, right lingual gyrus and left hippocampus** and b) AD-Memory vs. AD-Visuospatial: **left entorhinal cortex, right entorhinal cortex, right supramarginal gyrus and left postcentral gyrus**. I then determined the directionality of association between AD subgroup and each identified candidate brain region of interest by visually comparing distributions (violin plots) of the AD subgroups of interest. Results from the third comparison, AD-Language vs. AD-Visuospatial were not as informative; I discuss this further in Chapter 4. Overall, Aim 1 work demonstrates a careful selection of a machine learning method to best suit the characteristics of the data at hand; it is an application of random forest in a new domain: understanding AD subgroup differences. Since classification models have not been used to understand AD subgroup differences before, it is important to note there is no existing gold standard for how good the classification accuracies must be for distinguishing between pairs of AD subgroups.

The driving question for longitudinal data analysis (Aim 2) in my dissertation was to see whether the AD subgroups defined at t_0 , differ in their brain ROI volume trajectories over time, and if so, which ROIs show the most salient differences. In preliminary work for this analysis, I first simulated longitudinal data for two hypothetical groups based on average annual rates of change of volume of ROIs in AD obtained from literature, allowing tweaking of sample size and the factor by which two groups differ in ROI volumes. I intended the simulated data to serve as a test bed for analysis methods and applied linear mixed effects (LME) modeling on the data to determine the sample sizes and difference factors needed to detect engineered differences. For real longitudinal data, before applying LME modeling, I set criteria of working with data from specific MRI scans to reduce the effects of noise in data due to a.) moving MRI scanner technologies over time and b.) unaccountable MRI measurement error and ROI segmentation

error. Details of these criteria are discussed in Chapter 5. Using this data, I modeled each of the 70 ROI volumes over time using LME modeling to determine which (if any) ROIs' volume trajectories across two AD subgroups differ in overall volumes over time and the rates of change over time. I carried out this analysis on a series of subsets of data where each subset was generated by restricting each AD subgroup's data to time points with a specific minimum number of subjects available. I let this number range from 1 to 5, resulting in five different subsets of data for each AD subgroup. This was important to do since the longitudinal data is not only imbalanced in AD subgroup sizes but it is also irregular, i.e. not all individuals have data at all time points. Simply using all the data available at each timepoint for each AD subgroup may lead to biased results if the number of data points for an AD subgroup from a given time point is not large enough. For each dataset analyzed, I obtained two lists of top five ROIs, one based on the magnitude of difference in rates of change of ROI volume between the two AD subgroups being compared and the other based on the magnitude of difference in longitudinal volumes between the AD subgroups. All differences were normalized by ROI size for ranking ROIs for importance. Detailed results can be found in Chapter 6. A high-level result from this analysis is that for each of the three binary comparisons (AD-Language vs. AD-Memory, AD-Memory vs. AD-Visuospatial and AD-Language vs. AD-Visuospatial), there is some overlap between the lists of ROIs based on rates of change of ROI volume between the two groups being compared and average differences in ROI volumes between the groups at $t=0$. Additionally, some ROIs from these lists were also determined to be important for distinguishing between pairs of AD subgroups in the cross-sectional data analysis using random forest. These ROIs are: **right entorhinal cortex** (important in rate of change differences between groups and ROI volume differences between groups at $t=0$ from longitudinal data analysis) and **right hippocampus**

(important in ROI volume differences between groups at $t=0$ from longitudinal data analysis) **for AD-Language vs. AD-Memory. For AD-Memory vs. AD-Visuospatial**, the ROIs are **left entorhinal cortex, right entorhinal cortex and left parahippocampal gyrus**, all three of which are important in terms of rate of change differences between the groups as well as the ROI volume differences at $t=0$ for the two groups in longitudinal data. These findings preliminarily suggest that at least some of the ROIs that differ between AD subgroups at the time of diagnosis (or the first visit for AD individuals in the study) continue to show differences longitudinally.

The biomedical informatics contribution of my work is in dealing with the challenges and intricacies of real biological data, specifically structural MRI data collected in the ongoing ADNI study, for a meaningful analysis in the comparison of AD subgroups. As mentioned above, the biggest challenge of both cross-sectional and longitudinal data was the small sample sizes of the AD subgroups of interest (see Table 1 under 1.2.c), making it harder to find distinguishing features between subgroups. Accepting the limitation of small samples sizes, I focused my attention on addressing other challenges of the data that might help improve how well the subgroups can be distinguished from each other: imbalanced class sizes, accounting for correlation among brain ROI volume variables and potential noise in volume data due to variables other than ROI volumes. In addition to these challenges in cross-sectional data, in analyzing longitudinal data, I faced additional challenges of irregularly spaced and sparse data and changing technologies over time such as scanner MRI field strength. MRI volumes over time measured with even the same field strength were noisy, with large fluctuations in volume trajectories for individuals for many ROIs, instead of showing a steady decrease or mild fluctuations. In a scenario where the sample size is large enough, the noise in the ROI volume trajectories should not be of huge concern in LME modeling. Similarly, when the sample sizes

are large enough, the LME models should not be sensitive to which time points should or should not be included in the analysis for each AD subgroup. However, given the small sample sizes of the longitudinal data at hand, my analysis carefully took the above two points in consideration, rather than a direct application of LME models to the default dataset. Instead of an off-the-shelf and black box application of data science methods, my dissertation work demonstrated careful assessment of the chosen methods to suit the needs of the cross-sectional and longitudinal datasets.

My Ph.D. work is an application of existing methods (machine learning classification models and linear mixed effects modeling) in a new research domain: AD subgroups. From an AD researcher's perspective, results from my work may serve as a guide for which brain regions should be examined closely in future neuropathological studies for learning more about differences in the brains of individuals from different AD subgroups. In the case of classification models used for cross-sectional data, my work may be one of the first in establishing a documentation of classification accuracies for distinguishing between pairs of AD subgroups. Currently, there is no known gold standard or standards to be surpassed based on previous work for these classification accuracies.

Previous work focused on voxel-by-voxel comparison of AD subgroups based on cross-sectional structural MRI data (Crane and Group 2020). My Ph.D. work added to this line of research in a novel way, looking at a different entity of brain volume: brain regions of interest (ROI). I focused on 70 brain ROIs: 68 cortical regions defined by the Desikan-Killiany atlas (Desikan et al. 2006) and bilateral hippocampi. In addition, the methods I used for analysis are different from what has been done in previous analysis of cross-sectional data for comparing AD subgroups. I focused on variable importance through a carefully selected machine learning method (random

forest) and provided sets of ROIs that best distinguish a.) AD-Language vs. AD-Memory and b.) AD-Memory vs. AD-Visuospatial. In addition to extending previous work of analyzing structural MRI differences across AD subgroups to atlas defined brain ROIs, my analysis of cross-sectional data allowed for the possibility of non-linear relationships between ROIs and AD subgroups, and interactions between the ROI volumes, which is novel work in AD subgroups comparison research.

1.1 Background: Cognitive subgroups, brain regions and ADNI

To provide context for the analyses described in subsequent chapters, I begin with background information about AD subgroups, the brain regions of interest I chose to study, and the ADNI study which is the source of the data I used and an overview of the cross-sectional and longitudinal structural MRI data.

1.1.1. Clinical heterogeneity and AD Subgroups

Many diseases have been studied by focusing on subtypes within the disease, leading to more effective treatments for the subtypes. A successful example is breast cancer where molecular subtypes of breast cancer are used for predicting prognosis and planning treatments (Yang 2019). Other benefits of disease subtyping include being able to provide more precise estimates of the expected costs of care for different subtypes as well as better planning of clinical trials so that the subtypes can be represented in equal proportions through targeted recruitment (Saria and Goldenberg 2015). These serve as motivating factors for assessing the clinical variation within AD and defining subtypes.

Clinical heterogeneity observed in AD has suggested that AD is not a single uniform disease and that there are distinct variants within AD (Lam et al. 2013). Researchers have taken several approaches (Vardy et al. 2013), (Scheltens et al. 2016) and (Ferreira et al. 2017) to sub-

categorize people with AD. Other researchers have recognized three clinical variants of atypical AD, i.e. AD with relative preservation of memory but impairments in other cognitive functions: 1. posterior cortical atrophy (PCA) which presents with a predominant impairment in visual identification or in visuospatial function, 2. primary progressive aphasia (PPA) which is characterized by language impairment, semantic, syntactic and motor speech abilities, and 3. a behavioral variant of frontotemporal dementia characterized by progressive apathy, behavioral disinhibition or predominant executive dysfunction (Dubois et al. 2014). In other work, Dickerson & Wolk (Dickerson and Wolk 2011) focused on memory and executive functioning and characterized the primary impairments in individuals with AD by looking at the difference between memory and executive functioning in an individual based on neuropsychological tests. A recent development in this line of work has been the cognitively defined AD subgroups by Crane et al (Crane et al. 2017) which are the subgroups I used in my dissertation work. Crane et al. extended Dickerson and Wolk's idea of looking at contrasts between two domains to a difference-based approach for four domains (Executive & Attention, Language, Memory and Visuospatial) and in doing so, Crane et al. also did more extensive work in assessing the cognitive test items that would best define each domain.

In Crane et al.'s methodology, an expert panel of neuropsychologists assigned specific items from the Cognitive Abilities Screening Instrument (CASI) (Teng et al. 1994) and a neuropsychological batteryⁱ to one of the following cognitive domains: Executive & Attention, Language, Memory, and Visuospatial. Based on this assignment, scores for each domain were calculated for each individual. Then for each domain, raw scores for all subjects with incident

ⁱ The neuropsychological battery consisted of "clock drawing, verbal fluency, Mattis Dementia Rating Scale, Boston naming, verbal paired association and recall, logical memory and recall, Word List memory, Constructional Praxis and recall, Trails A and B, and Information and Comprehension subtest items.(McCarty et al. 2011)"

AD who had scores for all four domains (n=825) were scaled to have mean 0 and standard deviation (SD) 1. For each individual, if the difference between the average score of all four domains and the domain specific score was below or equal to an empirically selected threshold value (-0.75), then that domain was assigned to be the primary impairment domain for that individual. These procedures resulted in six AD subgroups: AD-Executive, AD-Language, AD-Memory, AD-Visuospatial, AD-No Domains, and AD-Multiple Domains. My dissertation focused on individuals in the first four single domain groups. Since the AD subgroup assignment is based on comparing each cognitive domain's score for an individual with the individual's average score, Crane et al.'s method controls for individual effects such as overall cognitive functionality in an individual and level of disease progression.

1.1.2. Brain regions of interest

The smallest unit of brain that can be analyzed using structural MRI corresponds to a voxel which is the smallest unit of measurement in a 3-D image. A typical voxel size for structural MRI is 1 mm³ ("Preliminaries — Introduction to MRI" n.d.) and it represents roughly 10⁵ neurons ("Lecture_2_-_mri_as_a_black_box.Pdf" n.d.). In looking for structural differences in the brains of individuals from different AD subgroups, one can analyze data ranging from the level of single voxels to brain regions comprised of one or more voxel. Previous work by Crane & Risacher (Crane and Group 2020) has shown some brain level differences across pairs of AD subgroups using voxel level analysis. To provide a different perspective on the question of brain regional differences in pairs of AD subgroups, I chose to use brain regions instead of individual voxels as the feature of interest for my analysis.

There are at least 46 known human brain atlases (Lawrence et al. 2021) that are region based, with varying numbers of brain regions ranging from 2 (in the Hemispheric atlas) to 1105 (in the

Talairach atlas). For my dissertation, I focused the analysis on cortical regions for ease of interpretation and also the hippocampus which is a subcortical structure known to be important in Alzheimer's disease due to its role in memory function (Parkin 1996). I also wanted to ensure that the number of regions (input features p) I use in my analysis are not too large, given the small sample size (n), as that combination could lead to model overfitting [cite]. In selecting a brain atlas for my analysis, I needed to make sure that cortical regions were included and that the number of regions was not too large. With the goal of presenting results that are reproducible in a software that is easily accessible as well as trusted in the neuroimaging analysis community, I chose to obtain atlas-based data from structural MRI scans using FreeSurfer ("FreeSurfer" n.d.). FreeSurfer is a widely used open source software in the field of neuroimaging analysis, and it is the structural MRI analysis software of choice for the Human Connectome Project ("Connectome Programs | Blueprint" n.d.), an initiative to map the anatomical and functional networks in the human brain. Among the cortical atlases, the FreeSurfer pipeline comes with three atlases: the Desikan Killiany (DK) atlas (68 regions), the Desikan-Killiany-Tourville (DKT) atlas (62 cortical regions) and the Destrieux atlas (148 cortical regions) (Destrieux et al. 2010). The Desikan-Killiany atlas seemed to be a more attractive choice because it provided a good balance of having enough regions of interest but not too many (so that potential problems in model building from the $p \geq n$ case could be avoided). I relied on expertise of Dr. Shannon Risacher at Indiana University for the processing of structural MRI data from ADNI using FreeSurfer to obtain data in the form of volume, surface area and thickness for brain regions based on an atlas. This processing in FreeSurfer consists of parcellation and segmentation which yields cortical and subcortical regions. The final dataset that I used in my dissertation consisted of 70 brain regions discussed next, courtesy of the FreeSurfer processed data from Dr. Risacher

using the Desikan-Killiany atlas and subcortical regions. In both cross-sectional and longitudinal data analysis, I focused on volumes of these selected 70 brain regions (68 cortical regions and two regions corresponding to the left and right hippocampus). In the next subsection (1.2.c), I describe the source of this data and the processing done by collaborators (Dr. Risacher) to prepare the data. Here, I discuss the 70 brain ROIs, including the Desikan-Killiany atlas.

The FreeSurfer processed MRI data from Dr. Risacher consisted of 120 volume, 68 cortical thickness and 70 surface area variables for each individual. This included cortical ROIs from the Desikan-Killiany atlas and subcortical ROIs. Volume, thickness and surface area are correlated measures, and although each measure captures slightly different aspects of the data, I chose to work with only volume variables to avoid multiple representations of the same region in the data for my analysis; I obtained the 70 volume variables corresponding to the 70 ROIs described in the previous paragraph. I excluded the remaining 50 volume variables in FreeSurfer data including the whole brain level or hemisphere level volumes, ventricular volumes and white matter variables. Since the goal is to understand regional brain differences across the AD subgroups, global volume measures at the whole brain or hemisphere level would not be useful. Ventricles are a series of interconnected cavities in the brain filled with cerebrospinal fluid (CSF), located in the core of the forebrain and the brainstem (Purves et al. 2001). There is no known unique function of ventricles. Ventricular volumes do not add much to the information that is already being provided by the surrounding brain regions. They represent “negative” spaces in the brain which are harder to interpret. White matter lies deep in the brain, beneath the grey matter cortex. It consists of millions of bundles of axons that connect neurons in different regions of the brain (Fields 2010). The white matter variables present in the FreeSurfer processed data provide a measure of white matter in left and right cerebellum, left and right cerebral cortex

and hypointensities. These global measures that are comprised of volumes of individual white matter tracts, are difficult to interpret and not useful for the analysis of looking at brain region differences across the AD subgroups. Hence, these variables were excluded and the final list of ROIs that I included in my analysis for both Aims 1 and 2 were volumes of the 68 cortical regions from the Desikan-Killiany atlas and the left and right hippocampus from the subcortical regions.

The Desikan-Killiany atlas is a human brain atlas of 34 cortical regions per hemisphere. It subdivides the human cerebral cortex on MRI into gyral based ROIs. The regions are organized into six lobes. Figure 2 provides a spatial context of the 68 cortical regions defined in the Desikan-Killiany atlas. The left and right hippocampi are shown on a different map (Figure 3) representing subcortical regions.

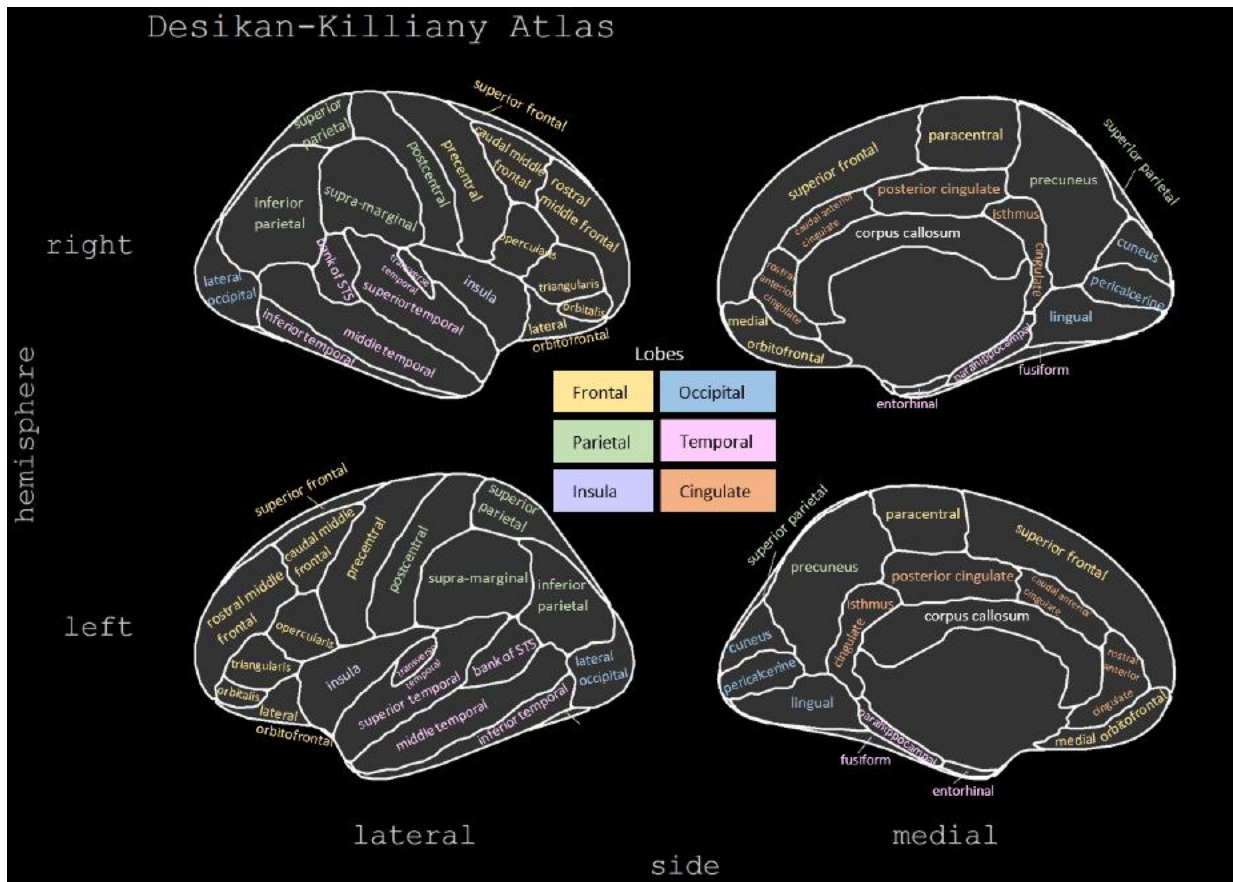


Figure 2: Labeled regions of the Desikan-Killiany atlas, in the lateral and medial views for both hemispheres of the brain. The different colors used for the region names represent different lobes. Base image created using the ggseg (Mowinckel and Vidal-Piñeiro 2020) package in R.

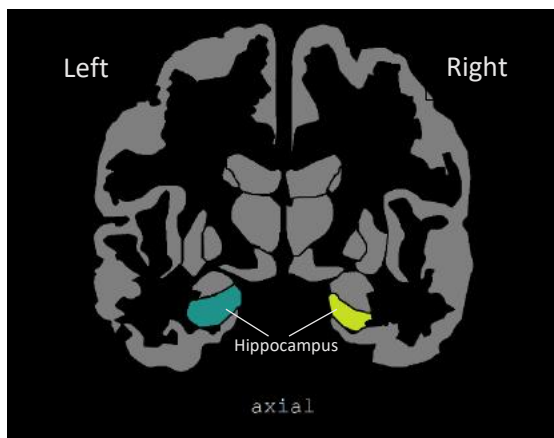


Figure 3: Subcortical regions shown in an axial view of the brain. The left and right hippocampi are among the 70 ROIs that were used in the analysis in my dissertation work. Base image created using the ggseg (Mowinckel and Vidal-Piñeiro 2020) package in R.

1.1.3. ADNI and FreeSurfer data

ADNI is a longitudinal multicenter study started in 2004 by Dr. Michael W. Weiner through funding from 20 companiesⁱⁱ, the National Institute of Aging (NIA), with the objective of developing clinical, imaging, genetic, and biochemical biomarkers for early detection and tracking of AD. The study has continued in different phases over the last 18 years. Figure 4 illustrates the intended subject additions in each phase and how they were carried through the study. ADNI-1 (2004-2009), the earliest phase of ADNI, consisted of 200 elderly controls (cognitively normal), 400 Mild Cognitive Impairment (MCI) and 200 AD individuals. The next phase, ADNI-Grand Opportunities, also known as ADNI-GO (2009-2011) comprised of all non-deceased individuals from ADNI-1 and an additional 200 individuals with early Mild Cognitive Impairment (MCI). ADNI-2 (2011-2016) continued to follow with existing ADNI-1 and ADNI-GO individuals and included new individuals: 150 individuals with normal cognition (cognitively normal controls), 100 individuals with early MCI, 150 individuals with late MCI and 150 individuals with AD. The latest phase of the study ADNI-3 (2016-2021) consists of all individuals from ADNI-1, ADNI-GO, ADNI-2 and new individuals: 133 controls, 151 MCI and 87 AD **Error! Bookmark not defined.** Protocols used in the ADNI-1 phase and the other phases were not identical due to technological advances over time. For example, ADNI-1 subjects' brains were scanned in MRI machines with a magnetic field strength of 1.5 Tesla (T) but in the subsequent phases of ADNI, the subjects were scanned in 3 T MRI machines. Additionally, the scanner type (manufacturer and model) varied across the different ADNI sites, but all the scanner types used were pre-approved by the ADNI study requirements. This variation in the ADNI

ⁱⁱ <https://adni.loni.usc.edu/about/#fund-container>

protocol across subjects and over time is an important point to note and one challenge of the ADNI data that I addressed in my data analysis.

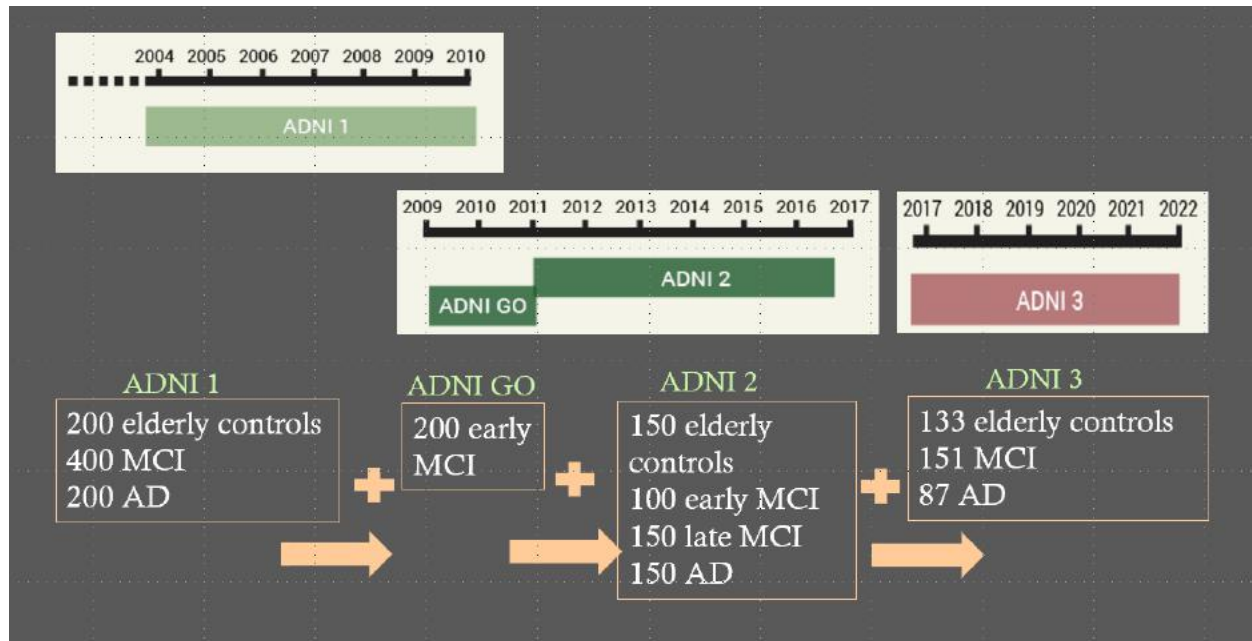


Figure 4: A chronological visualization of the phases in the ADNI study, and each phase's subject make-up. A total of 437 (and counting) AD subjects have been initiated into the study over the course of the various phases of the study. The timeline portions of this figure have been borrowed from the ADNI website.

The data available in the ADNI database includes raw MRI scans to FreeSurferⁱⁱⁱ processed MRI data in the form of volume, thickness and surface area values for brain regions defined by specific brain atlases used by FreeSurfer. The ADNI database has files corresponding to the FreeSurfer processed ROI data which includes volumes of 70 brain ROIs that I was interested in analyzing for individuals that have been assigned into AD subgroups by Crane et al. However,

ⁱⁱⁱ FreeSurfer is "an open source software suite for processing and analyzing (human) brain MRI images." [<https://surfer.nmr.mgh.harvard.edu/>] software that uses brain segmentation and parcellation according to the Desikan-Killiany (DK) atlas and calculates volumes and cortical thickness of the DK atlas brain regions.

the data for these individuals in the ADNI database is found under different files corresponding to different phases of the study (ADNI1, 2, GO, 3) which were processed using different FreeSurfer versions. In general, it is not recommended that data from different FreeSurfer versions be combined. Due to these standardization and interoperability issues in directly using FreeSurfer processed data from the ADNI database, Dr. Risacher at Indiana University processed all ADNI MRI scans using a single version and the latest versions of FreeSurfer at the time of analysis, version 6.0 (v6.0). I am grateful to Dr. Shannon Risacher at Indiana University who provided the FreeSurfer processed, and Quality Control (QC) checked data for my dissertation work.

In addition to imaging data, ADNI also contains clinical data for each individual. This includes the cognitive test scores that were used by Crane et al. to assign AD subgroups to individuals. Additionally, for each individual, many other variables are available including the following that I used in my analysis: “ptgender” (patient sex), “pteduc” (patient education), “MRI_Age” (age at the time of the MRI visit), “apoe4” (*APOE* genotype characterized as ≥ 1 $\epsilon 4$ alleles vs. 0 $\epsilon 4$ alleles), “ICV” (intracranial volume), “Field_Strength” (field strength of the MRI scanner) and “Scanner.Model.” Another variable of importance that I used in my analysis derived from ADNI data, involves w-scores,^{iv} which were calculated by Dr. Rik Ossenkoppele and Dr. Collin Groot at VU Amsterdam, using structural MRI data for each individual (Ossenkoppele et al. 2015). The w-score, for each voxel in an individual, represents a measure of the level of atrophy relative to cognitively normal controls. Ossenkoppele et al. used “a neuroimaging approach that sums the

^{iv} “W-scores (mean = 0, SD=1 in the control group, similar to z-scores) show, at each voxel, where a [participant’s] patient’s gray matter probability would fall on the normal distribution of gray matter probabilities in healthy controls, after taking nuisance factors into account [Jack et al., 1997; La Joie et al., 2012]. Our last steps were to binarize the W-score map for each [participant] patient at $W < -1.5$, and to sum the total number of suprathreshold voxels for every patient.(Ossenkoppele et al. 2015)”

number of brain voxels showing significantly lower gray matter volume than cognitively normal controls.” This approach of counting voxels below a w-score threshold gives a measure of the level of overall atrophy in the brain; Ossenkopele & Groot specified this threshold to be -1.5. The w-score voxel count can be thought of as an indicator of disease severity. All of the above variables could be potential contributors of variation in ROI volumes across individuals, and needed to be accounted for, while comparing AD subgroups; this is addressed in Chapter 2 for cross-sectional data and in Chapter 6 for longitudinal data. These variables are referred to as “non-ROI variables” in the rest of the document.

Although ADNI is a relatively large study with large sample sizes, the intersection of individuals in ADNI and those that have been assigned into cognitively defined AD subgroups is only a modest size from the perspective of machine learning and statistical analyses. Table 1 shows the final sample sizes after FreeSurfer QC (by Dr. Risacher) and additional data cleaning steps in which I removed individuals with missing w-scores. Given the small sample sizes, it was necessary to use data from both field strengths in the analysis. As detailed in subsequent chapters, for both cross-sectional and longitudinal data, my analysis accounted for the effects of field strength on volume measurements. For longitudinal data analysis, although I used data from both field strengths, for any given individual, I only used data from a single field strength over time.

The focus of my dissertation is on the four AD subgroups introduced earlier. There are two additional subgroups, AD-Multiple domains and AD-No domain, whose data I used in addition to the four groups’ in a preliminary data analysis step in Aim 1 work. For cross-sectional data, the data available (including both magnetic field strengths) in the six AD subgroups have the following sample sizes: AD-Executive (n=12), AD-Language (n=42), AD-Visuospatial (n=61),

AD-Memory (n=177), AD-Multiple domains (n=12) and AD-No domain (n=147). For longitudinal data, individuals who had data at t0 and at least one other time point from the same field strength had the following sample sizes: AD-Executive (n=19), AD-Language (n=42), AD-Visuospatial (n=64), AD-Memory (n = 190), AD-Multiple domains (n=20) and AD-No domain (n=235). Of these, over the span of the first four single domain subgroups, there were 13 individuals who had data for both 1.5T and 3T scans while in the AD-No domain group, there was 14 such individuals. For these individuals, I used data from the field strength that had data for more time points. If the number of time points were the same for the two field strengths for an individual, then I chose the 3T data. I used data from all six AD subgroups for cross-sectional and longitudinal data in a preliminary step for accounting for the effects of non-ROI variables on the volume data. For the main question of my dissertation, which is to understand how the single domain AD subgroups differ, I only analyzed differences across the following subgroups: AD-Language, AD-Memory and AD-Visuospatial, excluding AD-Executive due to its small sample size.

Sample size AD Subgroup	Cross-Sectional			Longitudinal		
	1.5 T	3 T	Total	1.5 T	3T	Total
AD-Executive	7	5	12	7	12	19
AD-Language	26	16	42	21	21	42
AD-Visuospatial	35	26	61	26	38	64
AD-Memory	115	62	177	97	93	190
AD-Multiple domains	12	10	22	10	10	20
AD-No domain	147	82	229	113	122	235
Total	342	201	543	274	296	570

Table 1: Sample sizes of AD subgroups. The first four subgroups were the focus of my dissertation work, although AD-Executive was excluded from comparison of AD subgroups due to the small sample size.

The data processed by Dr. Risacher that I used in my analysis originates from ADNI1, ADNI2 and ADNIGO for cross-sectional data and from all phases for longitudinal data. Raw MRI scans for these data can be accessed via the ADNI database with approval from the ADNI committee. All variables available for cross-sectional data are also available for each clinical visit in longitudinal data. Figure 5 shows the intended frequency of MRI visits for each phase of the ADNI study. Since I set the time axis relative to the time of AD diagnosis and analyzed data across all phases of ADNI in longitudinal data, the longitudinal dataset is irregular, i.e. not all individuals have data at all time points which is a common challenge of real world longitudinal datasets. Additionally, many individuals may have missing data because of missed visits.

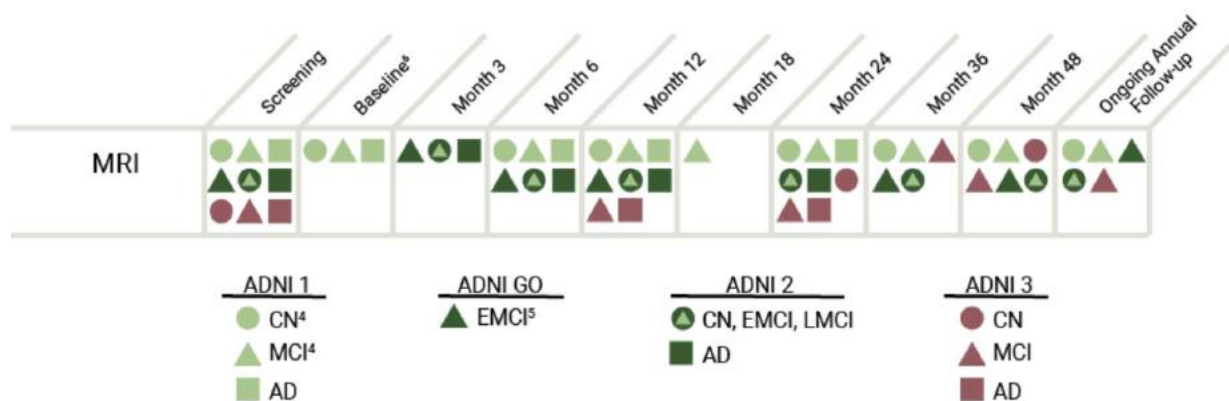


Figure 5: An overview of the MRI data collected throughout the ADNI study. Source: <http://adni.loni.usc.edu/data-samples/data-types/> AD= Alzheimer's disease, MCI = Mild Cognitive Impairment, EMCI = Early Mild Cognitive Impairment and LMCI = Late Mild Cognitive Impairment. A description of the superscripts on CN, MCI and EMCI could not be found at the source of this figure.

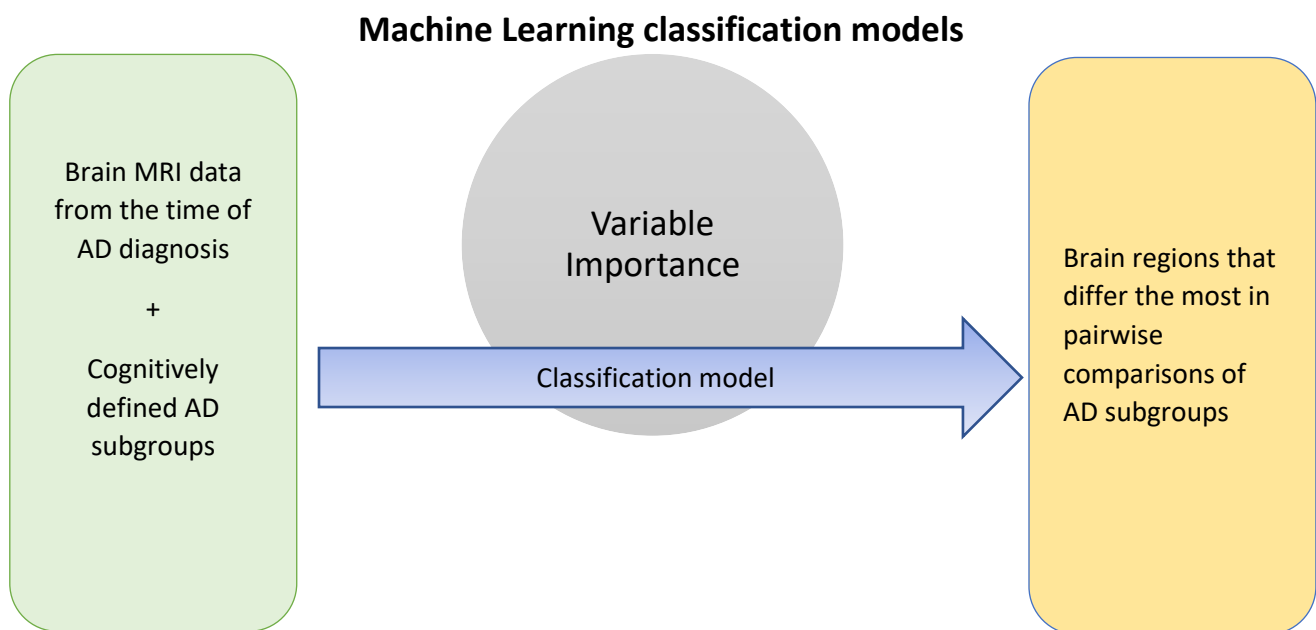
1.2 Organization of the dissertation work

The chapters of this dissertation are organized as follows: Chapters 2, 3 and 4 are about Aim 1: cross-sectional data analysis; Chapters 5 and 6 are about Aim 2: longitudinal data analysis, finally, Chapter 7 presents concluding thoughts, limitations and future research ideas. In Chapter 2, I describe the question tackled in Aim 1 (Cross-sectional data analysis), an overview of the

methodology and set up the mathematical problem of interest. I also provide a descriptive analysis of cross-sectional data in Chapter 2, including plots comparing some ROI volumes for pairs of AD subgroups. Descriptive plots for all cross-sectional ROI volumes can be found in the supplementary files (all files starting with “violin_BoxPlots” in their names). Chapters 3 and 4 are focused on the details of the analysis of cross-sectional data: determining ROI importance using classification models for distinguishing between pairs of AD subgroups using penalized logistic regression (Chapter 3) and random forest (Chapter 4). In Chapter 5, I provide an overview of longitudinal data, define the question of interest and discuss the characteristics and challenges of the current longitudinal data. In Chapter 6, I present an LME modeling analysis comparing rates of change and ROI volumes at $t=0$ in pairs of AD subgroups in longitudinal data originating from ADNI that were processed by Dr. Risacher using FreeSurfer. Through a series of LME models for each ROI for different subsets of data determined by excluding specific time points, Chapter 6 also attempts to check the robustness of LME models keeping some of the challenges of longitudinal data in mind such as irregularly spaced data over time and small sample sizes. Lastly, Chapter 7 concludes the dissertation with a summary of main findings, limitations of the current work and ideas for future work.

Aim 1: Cross-sectional Data Analysis

Machine Learning based approaches to identify brain regions from cross-sectional MRI data that are most important for distinguishing between AD-subgroups



Chapter 2: Aim 1 approach rationale, data and setting up the classification problem

The scientific goal of Aim 1 was to understand which brain regions are most important in distinguishing between pairs of AD subgroups, using FreeSurfer processed volumes of the 70 brain regions described in 1.2.b. from t0, the first MRI visit at or after AD diagnosis. In search of a method that would allow me to gain maximum insight into the differences across the AD subgroups, I explored two supervised machine learning based classification methods. My main criteria for choosing the machine learning methods were that they must be able to provide some insight into variable importance, and they must have ways of preventing overfitting the data. Based on these criteria, I implemented logistic regression with regularization (using the ridge, lasso and elastic net penalties) and random forest, the details of which are provided in the next two chapters. Although both of these models are interpretable for variable importance, logistic regression models are more interpretable than random forest. To aim for a highly interpretable model, I started the analysis of cross-sectional data using the logistic regression with regularization approach. Eventually, I selected random forest as the better method for the given data and question of interest (as discussed in Chapter 4) and used it to conclude which regions are important for distinguishing between pairs of AD subgroups.

Going into Aim 1 work, I hypothesized that there would be some differences across AD subgroups at the level of brain regions. I based this hypothesis on previous work (Crane and Group 2020) that had identified differences in grey matter density at the voxel level (in structural MRI data) for pairwise comparisons of AD subgroups, using independent models predicting a single voxel's volume as a function of AD-subgroup and confounding variables. A voxel represents the smallest unit of volume in an MRI. Each of the 70 regions considered in my

analysis consists of multiple voxels. As preliminary work of Aim 1, when I looked at independent models predicting each brain ROI's volume as a function of AD subgroup type (considering two subgroups at a time), I found the coefficient of the AD subgroup variable in the model to be statistically significant (different than 0) for some brain regions. This finding is consistent with my hypothesis that there are some differences across AD subgroups at the level of brain regions, when each brain region is analyzed separately.

However, for the chosen machine learning approaches, penalized logistic regression and random forest, which are both joint models of ROIs, I noted that there was no guarantee that these models would be able to separate the data sufficiently well into AD-subgroups just because the preliminary results had shown strong univariable (ROI) associations with AD subgroup membership. A univariable model provides information about marginal association of a single ROI variable with AD subgroup whereas a model based on multiple ROIs provides information about the relationship of a ROI variable with the AD-subgroup, given other ROI variables in the model. Whereas previous work in studying cognitively defined subgroups has shown marginal associations of voxels and ROIs with the AD subgroup type, the novel contribution through my Aim 1 work is that I present ensembles of brain regions for distinguishing between the different pairs of AD subgroups based on joint models that account for ROIs' relationships with one another.

2.1 Why I used machine learning methods and why I chose the ones I did?

A multivariable or a joint model of ROIs can be fit to the data to predict AD subgroup membership without using machine learning based methods specifically. There are specific reasons why I chose to use machine learning methods to do so. Before I explain the reasons, I first clarify the usage of the term “machine learning” in the context of my dissertation.

It is important to note that the term “machine learning” is not defined in a strict way across disciplines, and there is more than one interpretation of what may qualify as a machine learning method. For example, one of the methods that I implemented, logistic regression with regularization, is a method that has existed in the statistics field before the term “machine learning” came into wide use. In my dissertation work, I refer to it as a machine learning method because of how I implemented this method. The implementation is aligned with one of the standard problems in machine learning (classification) and some of the rule-of-thumb practices in the field: model building/learning and cross-validation on a subset of the data (training set) to choose optimal parameters and reporting test error by assessing model performance on unseen data (test set). Although having a training and a test set and using cross-validation are not unique to the field of machine learning, they are not necessarily a standard practice in traditional implementation of statistics methods. Hence, for my dissertation, I have used the term “machine learning” for any methods (including statistical methods) that have all three of the following criteria: methods that a.) are based on learning from a training dataset, b.) make the distinction between a training set and a test set, while making sure the test data is not used during the model building process and c.) follow the protocol of performing cross-validation on the training data in order to pick more reliable parameters for the model being constructed.

Note, that the test set to be used for the validation of the final model can be an unseen future dataset and it doesn't have to be a subset of the current dataset that we have. In my dissertation, I took this route because of the small sample sizes of the AD subgroups, and in order to build a reliable model, I wanted to maximize the training dataset size. Additionally, given the small sample size of the current dataset, if I split the data into training and test sets, the test set might not have enough samples to reliably test a model and report the associated test error. So, I chose to use all of the current dataset as the training data. In this approach, I note that the validation of the model using an independent dataset remains to be done in future work and I reported cross-validated training set errors as rough estimates of the test errors for the model.

Now that I have laid out a clear definition of what is included in the term “machine learning” in the context of my dissertation work, next, I explain why I used machine learning methods to understand the differences across AD subgroups. Firstly, there are techniques in some machine learning methods that allow one to minimize overfitting of the model. As noted in the opening paragraph of this chapter, this is one of the criteria I used to decide which machine learning methods I wanted to implement. Minimizing overfitting is important to obtain a reliable model that is more generalizable and more robust to changes in data. Generally, the more overfit or complex a model is, it will have a lower bias^v. Having a low bias is a good attribute of a model. However, highly overfit models generally have a high variance^{vi}. To avoid having a high variance, less overfit and less complex models are usually preferred over more complex ones, even if that comes at the expense at a slight increase of bias. Hence, machine learning methods

^v “...Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.” Pg. 34, Introduction to Statistical Learning (James et al. 2013)

^{vi} “Variance refers to the amount by which \hat{f} (estimate of true model) would change if we estimated it using a different training data.” Pg. 34, Introduction to Statistical Learning (James et al. 2013)

that allow one to minimize overfitting are useful in providing more robust classification models with low variance. The bias-variance trade-off is, however, important and having a high bias is also not good. I kept this point in mind while working with the machine learning methods that I chose; this is discussed later.

Probably one of the most attractive reasons for using machine learning methods for Aim 1 work is that for some methods, depending on the criteria used for controlling overfitting, the criteria can also be used as a way of feature selection (a selection of a subset of the variables to go into the model), which can be used to identify the important ROIs for distinguishing between pairs of AD subgroups. As noted earlier, feature selection is also one of the criteria I used to select which machine learning methods to implement. Feature selection can also be performed using techniques traditionally labeled as statistical methods, and as discussed above, there is an overlap of methods in statistics and machine learning.

Lastly, another advantage of using machine learning is that as a standard rule of practice in the machine learning field, tuning parameters of the machine learning methods are chosen through cross-validation of the training data. This procedure also contributes to the robustness of the final chosen model as the tuning parameters are selected based on an average measure over k subsets of the data instead of a single set of data. This makes the machine learning methods a better choice than traditional implementation of statistical methods where cross-validation may not be performed. Choosing tuning parameters based on cross-validation should theoretically result in a model with a low variance.

The key in statistical learning models is the bias-variance trade-off where both the bias and variance should be kept as low as possible. A model that minimizes the expected test error on unseen data simultaneously achieves low variance and low bias (James et al. 2013, 35). Model

complexity depends on attributes like the type of machine learning method chosen (example, linear vs. non-linear), the number of features represented in the model and the magnitudes of weights for the features in the model. For example, logistic regression is a more complex model than logistic regression with regularization since logistic regression without regularization has higher coefficients and/or a greater number of non-zero coefficients. Random Forest, a non-linear method, is more complex than any kind of logistic regression which is a linear method. Often, in practice, there is a gain of some bias at the expense of keeping the variance small, as the complexity of the models decreases. In choosing the machine learning methods that I decided to work with, I took the bias-variance tradeoff into account at a conceptual level, by choosing to opt for the implementation of the less complex machine learning models before trying the more complex ones.

Generally, model complexity also has an inverse relationship with model interpretability, which is something I also kept in mind while choosing the methods since being able to interpret the chosen classification model is very important for the scientific question of Aim 1. Given the small size of our data, I also decided not to work with machine learning methods (for example: neural networks) that generally only work well for large sample sizes.

2.2 Data for model building and the general classification problem in Aim 1

The term “cross-sectional” refers to data at a single time point; this translates to data from a single clinical MRI visit in the ADNI study. In Aim 1, the data used for each individual was cross-sectional data at the earliest clinical visit after AD diagnosis, referred as t_0 in this document. As discussed in Chapter 1, the cognitively defined subgroup assignment for each individual is also from t_0 . Mathematically, the problem of Aim 1 can be visualized as shown in Figure 6, where the goal is to build a classification model for predicting AD subgroup labels to

understand which brain ROIs are important in doing so, using data from the volumes of 70 brain regions for individuals from two given AD subgroups (shown as matrix X) and known AD subgroup labels for those individuals (shown as vector y).

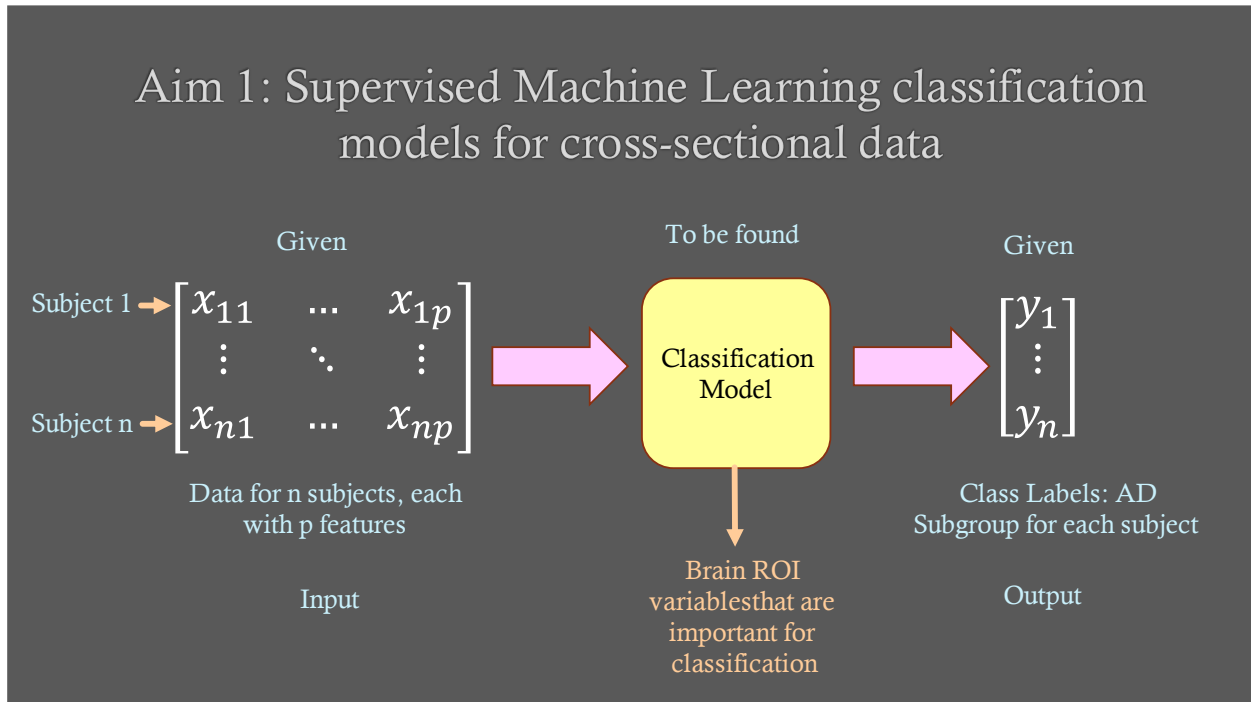


Figure 6: A visualization of the classification problem in Aim 1

The additional variables considered during model building are each individual's age, sex, years of education, intracranial volume, w-score^{vii} and apoE4 allele status (present or not). Intracranial volume provides an estimate of maximum premorbid brain size. Jenkins et al. (Jenkins et al. 2000) showed that unlike cerebral volume, intracranial volume is unaffected by atrophy due to neurodegeneration or ageing, and hence can be used as a parameter that allows one to account for variation in overall (pre-morbid) brain sizes when working with ROI data from different individuals. We considered all of the aforementioned variables as potential confounding

^{vii} See Chapter 1 for a definition of w-score.

variables. In penalized logistic regression's implementation in R, it is possible to specify these variables as "fixed variables" which means they will not be subjected to penalties for overfitting, and they will be included in the model as potential confounding variables to account for any potential association of these variables with the outcome (AD subgroup type) and with the ROI variables. In the other machine learning method that I used, random forest, it is harder to pre-include variables before tree building. Additionally, there are other variables in the dataset, nuisance variables such as scanner model type and the magnetic field strength of MRI scanners (1.5 T or 3 T); these potentially introduce noise in volume/thickness/surface area measurements in the dataset, but in theory, they should not have any association with the outcome to be predicted (AD subgroup type). I reasoned it would be incorrect to include these variables as covariates in the classification models. However, they still needed to be accounted for as contributors of potential noise in ROI volumes. As a consistent method to account for all non-ROI variables mentioned above, I regressed out the effects of these variables from each ROI volume using linear regression. This is discussed further in section 2.6.

In the current cross-sectional dataset analyzed in Aim 1, there are 177 individuals in the AD-Memory group, 61 individuals in the AD-Visuospatial group, 42 in AD-Language and 12 in AD-Executive. Given the small sample size of the AD-Executive group, I excluded it from the current analysis. Hence, I worked on the following classification problems: AD-Language vs. AD-Memory, AD-Language v. AD-Visuospatial, and AD-Memory vs. AD-Visuospatial.

2.3 Cross-sectional data acquisition, pre-processing and cleaning

I obtained the data for Aim 1 work from Dr. Shannon Risacher at Indiana University – Perdue University Indianapolis in the form of volumes, thickness and surface area values for brain regions defined by the Desikan-Killiany (DK) Atlas in FreeSurfer, along with several Quality

Control (QC) and meta-information variables. The volume, thickness and surface area values for the brain ROIs were obtained using FreeSurfer version 6.0. Dr. Risacher performed the necessary QC steps on the cross-sectional data that was originally collected across several institutions as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI). The QC'd dataset consisted of the above quantities for individuals without AD (controls) and individuals who have been classified by Paul Crane into the six subgroups of AD based on a method that uses results from cognitive tests described in Chapter 1.

Of the 790 individuals in the starting dataset that I received from Dr. Risacher, I dropped individuals whose FreeSurferQC value was listed as "FAILED" and individuals who were either controls (n=118) or did not have an AD subgroup classification available (n=0) from Crane et al.'s work. I also removed one individual (SubjID = 023_S_0604) because of discrepancy in QC from Voxel based morphometry (VBM) and FreeSurfer, as asked to do so by Dr. Risacher. Following the above steps, I checked for additional missingness in the data; all of the 624 subjects remaining had complete data for all FreeSurfer based volume, surface area and thickness variables. The sample sizes for the different groups were as follows: AD-Executive: 14, AD-Language: 46, AD-Memory: 189, AD-Visuospatial: 85, Multiple domains: 23 and No domain: 267. There were individuals in the dataset whose age was reported as ">89" (greater than 89); I changed these categorical values to a fixed numeric value of 90 as all other values for the age variable in the dataset are numeric values. Further, all individuals less than 65 years old were filtered out (n=68). As a last step of filtering data, I aligned the remaining individuals in this dataset with a dataset containing w-scores for individuals; I also acquired the w-score data file from Dr. Risacher. The w-score file only contains data for individuals who are 65 years or older and non-reverters. After dropping individuals from our dataset who didn't have w-score data, I

was left with the final cleaned dataset which is used in Aim 1 analysis, with sample sizes as follows: AD-Executive (n=12), AD-Language (n=42), AD-Memory (n=177), AD-Visuospatial (n=61), AD-No domain (n=229) and AD-Multiple domains (n=22).

Since the volume, thickness and surface area values for the different brain regions are derived from the MRI scans, it is important to take note of the source of the MRI data. Of the 624 individuals, the MRI data for 378 individuals came from scanners with a field strength of 1.5 Tesla (T) and 276 individuals' data came from 3 T MRI scanners. This divergence in the dataset is due to different phases of the ADNI study being represented here and the moving technological advancements over time. Additionally, the scanner type (manufacturer and model) varied across the different ADNI sites, but all the scanner types used were pre-approved by the ADNI study requirements.

FreeSurfer processed data contains 68 thickness variables, 70 surface area variables and 120 volume variables. As a starting point of Aim 1 analysis, we decided to work with volume variables as our variables of interest to find brain regions that are the most important for distinguishing between AD-subgroups. Although thickness, surface area and volume are correlated measures, each measure is still unique as it captures a different aspect of an ROI's geometry, and each of these measures may offer different insight into how the AD subgroups might differ from each other. Future work could include exploring thickness and surface area variables in separate classification problems. Furthermore, building classification models that take advantage of combining information from volume, thickness and surface area variables in a single model while taking correlation of the variables into account, is a very interesting and useful problem to be solved, and I propose that as future work beyond this dissertation.

Finally, I prepared a dataset extracting the following specific variables from the above pre-processed data: “rid” (unique subject ID), “sub_0_80” (AD subgroup label), ROI variables (volume, thickness and surface area variables from FreeSurfer), potential confounding variables: “ptgender” (sex), “pteduc” (patient education), “MRI_Age” (age at the time of the MRI), “apoe4” (apoE4 allele present or not), “wscore_wb” (whole brain average of w-score), “wscore_voxelcount_-1.5” (number of voxels below a w-score of -1.5), “ICV” (estimated total intracranial volume) and nuisance variables “Field_Strength” (field strength of the MRI scanner) and “Scanner.Model” (scanner model). Since wscore_wb and wscore_voxelcount_-1.5 are correlated, we chose to keep only one of these variables when we implemented the machine learning methods. I chose to keep the wscore_voxelcount_-1.5 variable.

Before diving into the analysis of cross-sectional data and building classification models for predicting AD subgroup, I carried out a descriptive analysis and visualization of the data to get a better sense of the overall data structure and to find any general patterns that might emerge. These include violin plots showing the distribution of volumes for each ROI for each AD subgroup; these plots can be found in the supplemental files provided with this document.

2.4 Training and Test data

After the necessary extraction and organization of data, the next step is to allocate training and test sets. When comparing multiple classification methods’ performance, it is best to have the same training and test sets across different methods, for consistency. I’ve described this step in this chapter as it is a common step for all machine learning methods of Aim 1. In section 4.2, I briefly talked about the general practice in the field of machine learning of having separate datasets for developing or training a model and for assessing the performance of the model. In

this section, I discuss the importance of doing so and I also provide reasons for why I used the entire sample size of the current cross-sectional data for training purposes.

The data used to develop the model is called the training set, and additional unseen data used to assess the performance of the model is called the test set. Testing the model on a dataset that was not used in the training process is important in order to provide a fair assessment of the model. If the model is tested on the same dataset that was used to develop the model, then the performance of the model would be spuriously biased towards being a good performance. In other words, training error (the error of the model based on the training data) is generally expected to be better than test error (the error of the model based on the test data), especially when the training set is small and when there is enough variation in the population. The test error allows one to know how generalizable the model is to new, unseen data. In an ideal situation where the sample size of the dataset is sufficiently large, splitting the data into training and test sets allows for each set to have enough samples for good training of the model and for reporting a reliable test error.

However, in dealing with a smaller sample size as is the case in the dataset available to me for cross-sectional analysis in Aim 1, I had to make the careful choice of choosing not to split the data into training and test sets, and to use all of the data as the training set. I made this decision in order to maximize the data available for training purposes and also because I realized that making a 80-20 split (80% data as training set and 20% as test set) would result in a test set with very small sample sizes which would not give a reliable test error anyway. The way to overcome the small sample size of the test set would be to reduce the amount of training data, which would come at a cost of a poorer classification model. Keeping this trade-off in mind, for all supervised machine learning methods used to develop classification models in Aim 1 work, I have chosen to use the entire current cross-sectional data as the training set. Regarding the test set, I envision it

to be a future dataset involving the same AD subgroup types, and hence, in my analysis for Aim 1, I do not report any test errors and leave this assessment of the models as future work as more data becomes available in the future.

For all methods used in Aim 1 work, I will report cross-validated training errors for each model while noting that these can be only interpreted as gross estimates of test errors but they are not the actual test errors for the models. Given the small sample size of the current dataset, working with the entire dataset as the training set seemed the most sensible choice for better model building.

2.5 Transformation of volume variables: Regressing out the effects of extraneous variables

A challenging aspect of the data is the variation across individuals in the volume of any given brain region that could be attributed to non-ROI variables. It is possible that some of these variables may be contributing to noise in volume ROI data and/or may be confounding variables that are associated with both the AD-subgroup and volume(s) of brain ROI(s). I allowed for the possibility of the following variable to be contributing to variation in volumes of brain regions across individuals: scanner field strength, scanner model, sex, age, years of education, *APOE* genotype, w-score and total intracranial volume (ICV). In Appendix A, I discuss some of these non-ROI variables in detail. To my knowledge and based on literature search, there isn't any data driven evidence or a theory that suggests that any of these non-ROI variables could be causally affecting the probability of being in a particular AD subgroup. Hence, although I acknowledge the possibility of sex, age, years of education, *APOE* genotype, w-score and total intracranial volume (ICV) being associated with AD subgroups, I do not model this potential phenomenon in our analysis.

To account for the effects of above-mentioned non-ROI variables, we transformed each of the 70 brain ROI volumes to a new variable space. We refer to this transformed variable as “residual volume” and it is defined in the following way. Using data of the 543 individuals in the cross-sectional dataset representing all individuals with AD (all six AD subgroups), I first modeled the volume of each brain region i as a linear function of the non-ROI variables, as in equation (1). I recoded some variables before using them in the model in (1). Specifically, I recoded the variable “Scanner.Model” representing 22 scanner models as 21 categorical variables (with values 0 or 1), “Field_Strength” as a categorical indicator variable for field strength of 3T and *APOE* genotype as a categorical indicator variable (1 for presence of the apoε4 allele and 0 for absence).

$$\begin{aligned}
 & model_vol_i \\
 & = \beta_0 + \beta_1 * EstimatedTotalIntraCranialVol + \beta_2 * Field_Strength + \beta_3 \\
 & * Wscore_voxelcount_ - 1.5 + \beta_4 * MRI_Age + \beta_5 * CSF + \beta_6 * ptgender + \beta_7 \\
 & * pteducat + \beta_8 * apoe4 + \beta_9 * Scanner.Model1 + \dots \beta_{29} \\
 & * Scanner.Model21
 \end{aligned} \tag{1}$$

Using the glm function under the Stats package in R (“R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>,” n.d.), I fit the above model and obtained a residual volume, $residual_vol_{ik}$, for each ROI i and individual k by subtracting the observed volume from predicted volume given by (1). Hence, from each brain ROI volume for an individual, I regressed out the effects of the non-ROI variables using a linear model. This procedure accomplishes two objectives: a.) removes the effects of non-ROI variables from ROI volumes and b.) allows one to only work with volume variables for feature selection (in step 4).

To have a large sample size and a more reliable model for regressing out the effects of extraneous variables, I used data from all six AD subgroups for the above procedure. For the

purposes of the scientific question of Aim 1, which is to understand which brain regions are most important in distinguishing pairs of AD subgroups, I focused only on the single domain specific subgroups. Since the AD-Multiple domains group consists of individuals who are roughly equally impaired in more than one cognitive domain, it would be difficult to discover differences between the AD-Multiple domains group and other single domain groups. Additionally, interpreting these differences is not straightforward or meaningful. For the comparison of the AD-No domain group (which can be thought of as a “generic” type of AD) with the single domain groups, one may not see as stark differences as comparing each of the single domain groups with one another. Hence, to focus efforts on an analysis which is more likely to reveal differences across the AD-subgroups, my dissertation is focused on the single domain AD subgroups. Further, my analysis had to be limited to the three single-domain subgroups with sufficient sample sizes (omitting AD-Executive).

2.6 SMOTE oversampling to balance AD-subgroup sample sizes

Another challenging aspect of our dataset, especially in the context of applying machine learning models for classifying individuals into AD subgroups, is the small and imbalanced sample sizes of the AD subgroups being compared: a.) AD-Language (42) vs. AD-Memory (177), b.) AD-Memory (177) vs. AD-Visuospatial (61) and c.) AD-Language (42) vs. AD-Visuospatial (61). Imbalanced class sizes can be problematic for many machine learning methods as the algorithm obtains a better overall classification accuracy by simply classifying most individuals into the majority class, regardless of the characteristics of the data. This results in a biased model with a good classification accuracy for the majority class but a poor classification accuracy for the minority class. I addressed the class imbalance by oversampling the minority class using SMOTE (Synthetic Minority Over-sampling Technique)(Chawla et al. 2002), using the SMOTE

package (Siriseriwan 2019) in R, to generate datasets that have roughly balanced class sizes for each of our three binary classification problems. The sample sizes after SMOTE oversampling for the three binary classification problems being considered: a.) AD-Language (n=168) vs. AD-Memory (n=177), b.) AD-Memory (n=177) vs. AD-Visuospatial (n=183) and c.) AD-Language (n=84) vs. AD-Visuospatial (n=61). Violin plots and boxplots for residual volume distributions for all 70 ROIs for each AD subgroup after SMOTE oversampling are provided in supplementary files.

In the upcoming chapters, I present details of the specific machine learning methods that I implemented to solve the classification problems defined in this chapter.

Chapter 3: Determining ROI importance in cross-sectional data using penalized logistic regression [Aim 1]

The first supervised machine learning method that I used for solving the classification problem outlined in Chapter 2, is logistic regression with regularization. In this chapter, I provide a brief mathematical overview of the method, followed by details of implementation, what I learned about separability of AD subgroups based on the 70 brain ROIs' volume data as well as a note on some limitations of the method.

3.1 Why logistic regression based models as the first analysis method?

A linear regression model is the simplest type of model one could use to model a supervised classification problem. It is also the easiest to interpret. Logistic regression models are slightly more complex and although not as easy to interpret as linear regression, they still rank well in model interpretability. Why would one want to use logistic regression over linear regression when linear regression is even simpler to interpret? The answer lies in the better suitability of logistic regression models for classification problems. The mathematical details of logistic regression are described in the next section, which will make the following points clear. Unlike a linear model, a logistic regression function has a range that is bounded $[0,1]$ and suitable for predicting probabilities of a given class. By the mathematical definition of a logistic function, its output values are guaranteed to be between 0 and 1. Additionally, a logistic regression based model can be extended to a multi-class classification problem in an interpretable way, unlike a linear model which would falsely introduce a ranking of classes if more than two classes are to be included in the output variable.

Keeping model interpretability in mind, I started the analysis in Aim 1 with a supervised machine learning method that is the least complex and a suitable method for classification. Additionally, keeping the bias-variance trade-off^{viii} in mind, working with a method that has a lower variance relative to other methods, seemed favorable. Depending on the true unknown decision boundary in the data for separating the AD subgroups, logistic regression may or may not have a higher bias than more complex supervised machine learning methods. It is generally better to explain the variation in the data with a simpler model before trying a more complex model because of lower variance and better model interpretability in a simpler model. Hence, I chose a logistic regression classification method as the first method for building a classification model for AD subgroups, for the purposes of understanding the relative importance of the brain ROIs in distinguishing between pairs of AD subgroups.

3.2 An introduction to logistic regression with regularization

In this section, I provide a mathematical overview of logistic regression with regularization. Logistic regression with regularization is a well-established technique (Zou and Hastie 2005) and this section may be skipped by readers who are familiar with the mathematical details of the method. The information laid in this section is important to understand the rationale behind some of the steps I carried out in the model building process (section 3.3) as well as the results (section 3.4).

Logistic regression with regularization is an extension of logistic regression. Regularization refers to constraints placed in the model fitting process. One reason to perform regularization is to control overfitting which is accomplished by imposing a penalty for large coefficients in the

^{viii} Bias-variance trade-off was first introduced and explained in Chapter 2.

model. This results in shrinkage of coefficients, and has the effect of reducing variance (James et al. 2013, Pg. 204). Below, I first briefly explain how a binomial classification problem can be modeled using logistic regression without regularization, the criteria used to find the coefficients. I then extend the discussion to how the criteria for model fitting changes when regularization is added.

Consider a response variable with two dummy values: $Y = 0$ and $Y=1$, representing two classes. A classification model can be built by predicting the probability vector $\mathbf{p}(\mathbf{X}) = \Pr(Y = 1 | \mathbf{X})$, where $Y=1$ is the class of interest and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p]$ is the input data matrix for p variables. Each vector in \mathbf{X} is N -dimensional (for example, $\mathbf{X}_1 = (x_{11}, x_{12}, \dots, x_{1N})^T$ where N is the number of measurements or samples. In the context of Aim 1, each measurement or sample corresponds to an individual denoted by $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{pk})$, for the k th individual. Hence, for a classification problem where there are N total individuals and p variables, \mathbf{X} is of size $N \times p$. For logistic regression based classification, the probability vector $\mathbf{p}(\mathbf{X})$ is represented by the following logistic function:

$$\mathbf{p}(\mathbf{X}) = \Pr(Y = 1 | \mathbf{X}) = \frac{e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{X}}}{1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{X}}} \quad (2)$$

Here, β_0 , the intercept and the vector $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]$ containing the coefficients for the p variables in the model, are estimated using the training data by maximizing the likelihood function, which is defined as follows:

$$\ell(\beta_0, \boldsymbol{\beta}) = \prod_{i: y_i=1} p(x_i) \prod_{i': y_{i'}=0} (1 - p(x_{i'})) \quad (3)$$

Note that the likelihood function $\ell(\beta_0, \boldsymbol{\beta})$ is a product of the probabilities of observations being in the correct class. Intuitively, one can think of the likelihood function as a measure of how

close the model is to predicting the correct class labels. For a model that is able to predict class labels really well, we would expect that all $p(x_i)$ s or $\Pr(Y = 1 | x_i)$ s would be close to 1 and $p(x_{i'})$ s or $\Pr(Y = 0 | x_{i'})$ s would be close to 0. Hence, in equation 3, having a better predicting model translates to having all $p(x_i)$ terms being closer to 1 and all $1 - p(x_{i'})$ terms to be close to 1 also. Since all probabilities must lie in the interval $[0,1]$, the best case scenario model should have $\prod_{i:y_i=1} p(x_i) = 1$ and $\prod_{i':y_{i'}=0}(1 - p(x_{i'})) = 1$, which would mean $\ell(\beta_0, \boldsymbol{\beta}) = 1$ for a perfectly predicting model. Note, the maximum value $\ell(\beta_0, \boldsymbol{\beta})$ can take is 1. Hence, to have the best performing model, the mathematical problem becomes of maximizing the likelihood function. With a few mathematical manipulations, one can show that maximizing the likelihood function is the same as minimizing the negative binomial log-likelihood function:

$$\min_{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}} \left[\frac{1}{N} \sum_{i=1}^N y_i (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) - \log \left(1 + e^{(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})} \right) \right] \quad (4)$$

where N is the number of observations, $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$ is the input data for the i th observation and y_i is the known output value (class label) for that sample. In the context of Aim 1, each sample \mathbf{x}_i corresponds to an individual. The criteria presented in Equation 4 is used to find β_0 and $\boldsymbol{\beta}$ for logistic regression without any regularization.

Using equation 3 as the criteria to find the coefficients of the model may result in coefficients that have been highly overfit to the data. This corresponds to coefficients being too large. Such a model fits the training data really well but it is expected to be not robust to data changes. To avoid this issue of overfitting which tends to be a problem when the sample size is small and when the number of variables p is close to N , additional constraints are added to the above

function (in equation 4) for finding the coefficients, controlled by two parameters, λ and α . The following equation is used to determine the coefficients for logistic regression with regularization:

$$\min_{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}} \left[\frac{1}{N} \sum_{i=1}^N y_i (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) - \log (1 + e^{(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})}) \right] + \lambda \left[\frac{1}{2} (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \quad (5)$$

The first term in equation 5 is the same as the negative binomial log-likelihood from equation 4.

The second term in the equation 5 contains the additional constraint of regularization: the penalties (determined by values of λ and α) to be imposed during model fitting that allow one to control the level of overfitting of coefficients in the model. λ can take on any positive real value, while α can take any value in the interval $[0,1]$. Intuitively, α controls the number of non-zero coefficients in the model while λ determines how much one wants to penalize having large coefficients in the model. Dissecting the different parts of the second term in equation 5 can help us understand these specific roles of λ and α . $\|\boldsymbol{\beta}\|_2$ is the L-2 norm or the Euclidean distance,

defined as $\sqrt{\beta_1^2 + \beta_2^2 + \dots + \beta_p^2}$. $\|\boldsymbol{\beta}\|_1$ is the L1 norm, defined as $\sum_{i=1}^p |\beta_i|$, also known as the

Taxicab norm or Manhattan norm, which refers to “the distance a taxi has to drive in a rectangular grid to get from the origin to the point”^{ix} $\boldsymbol{\beta}$.

When $\alpha = 0$, the logistic regression model is said to have a ridge penalty; note that the regularization penalty term in Equation 5 becomes $\frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2$ when $\alpha = 0$. In the ridge penalty, there is no constraint imposed on the number of non-zero coefficients in the model. A model with a ridge penalty has the maximum number of non-zero coefficients possible in the model, hence there are p non-zero coefficients in a model with $\alpha = 0$. The only penalty in a ridge model comes

^{ix} Cite source for definition. Currently, I used Wikipedia.

from the tuning parameter λ which is represented in the $\frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2$ term. λ and $\|\boldsymbol{\beta}\|_2^2$ are inversely related. As λ increases, $\|\boldsymbol{\beta}\|_2$ (and hence $\beta_1, \beta_2, \dots, \beta_p$) must decrease in order to keep the $\frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2$ term small enough. Thus, the larger the value of λ , the greater the penalization on large coefficients.

When $\alpha = 1$, the logistic regression model is said to have a lasso (Least Absolute Shrinkage and Selection Operator) penalty. In the case of the lasso penalty, the regularization penalty term in equation 5 simplifies to $\lambda \|\boldsymbol{\beta}\|_1$ and now λ controls how much one wants to penalize having a large $\|\boldsymbol{\beta}\|_1$. Given that $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^p |\beta_i|$, one can see that a higher value of λ places a higher penalty on the coefficients $\beta_1, \beta_2, \dots, \beta_p$ being large. A less obvious fact of the lasso penalty is that it results in some coefficients being exactly 0. To illustrate this phenomenon of the lasso penalty, consider a simple case where $\boldsymbol{\beta} = [\beta_1, \beta_2]$. In the case of the lasso penalty, the negative log likelihood minimization problem is subjected to the constraint of $\lambda \|\boldsymbol{\beta}\|_1 \leq m$, where m is a constant. The constraint is equivalent to $\|\boldsymbol{\beta}\|_1 \leq s$ where $s = \frac{m}{\lambda}$ is another constant. $\|\boldsymbol{\beta}\|_1 \leq s$ which is the same as $|\beta_1| + |\beta_2| \leq s$ can be visualized in the (β_1, β_2) coordinate system; this inequality is shown as the green region in Figure 3.1a. $\hat{\boldsymbol{\beta}}$ in figure 3.1a represents the solution of the minimization of the negative log-likelihood problem. The ellipses around $\hat{\boldsymbol{\beta}}$ represent regions of constant negative log-likelihood. The first point at which an ellipse intersects with the constrain region (the green region) is the solution to the logistic regression with the lasso penalty. Since the region given by $|\beta_1| + |\beta_2| \leq s$ has sharp edges, the intersection of a constant negative log-likelihood ellipse and this region often occurs at one of the corners of the region, which corresponds to either β_1 or β_2 being 0. The property of sharp edges of the constraint

region of the lasso penalty, also extends to higher dimensions where $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]$, resulting in some of the coefficients becoming exactly 0.

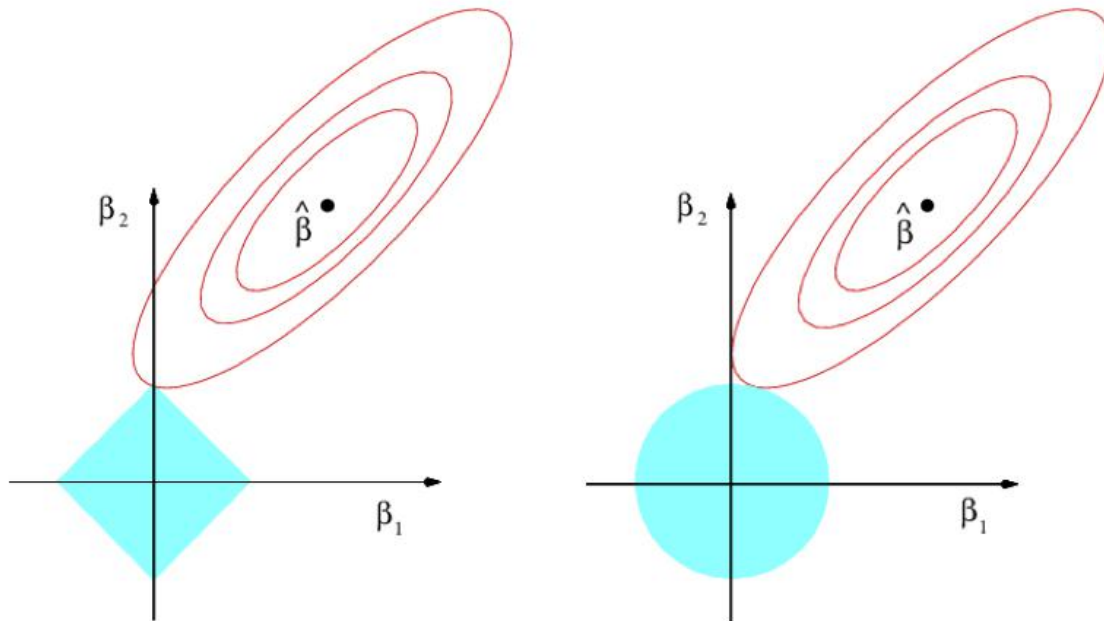


Figure 3.1: A visualization of contours of the error and constraint functions for the lasso (left; 3.1a) and ridge regression (right; 3.1b). Image source: Figure 6.7, *Introduction to Statistical Learning*, Pg. 222

Figure 3.1b illustrates why the ridge penalty, unlike lasso, does not result in some coefficients becoming exactly 0. In the case of the ridge penalty, the constraint region shown in green is $\|\boldsymbol{\beta}\|_2^2 \leq s$, which in the case of two dimensions is $\beta_1^2 + \beta_2^2 \leq s$ which is a circle. Due to the concavity of this region, the intersection with an ellipse and the region are less likely to happen at the points on the circle that are on the axes than all other points collectively on the circle. So, in the case of the ridge penalty, more often, the solution will include non-zero values for both coefficients β_1 and β_2 than the case of one of the coefficients being 0. In higher dimensions, the constraint region is not a circle but a hyper surface that still has the same properties that make the

intersection of the ellipses with non-zero set of values for $\beta_1, \beta_2, \dots, \beta_p$ more likely than some of them being exactly 0. Thus, the ridge penalty for logistic regression results in non-zero coefficients that are subjected to be small based on the value of λ , whereas the lasso penalty results in complete zeroing of some coefficients and some non-zero coefficients that are subjected to be small based on the value of λ .

Above, I explained the details of two special cases of the value of α : $\alpha = 0$ (Ridge) and $\alpha = 1$ (Lasso), When $0 < \alpha < 1$, the penalty is known as the elastic net penalty and the behavior of the models lies in between the extreme cases of the ridge penalty (all coefficients are non-zero) and the lasso penalty (the least number of non-zero coefficients possible are in the model). In summary, the parameter α controls the number of non-zero coefficients in the model while the parameter λ controls how much one wants to penalize the non-zero coefficients for being large. Both the number of non-zero coefficients and the magnitudes of non-zero coefficients play a role in overfitting. For every combination of α and λ , there is a set of coefficients $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]$ that will satisfy equation 5. Together, combinations of α and λ give different models representing different levels of overfitting of coefficients. Using logistic regression with regularization for Aim 1 work as a first method is appealing not only for its ease of model interpretability and controlling overfitting but also because it allows for varying levels of zeroing of coefficients depending on the value of α chosen, which results in varying levels of feature selection. As discussed in Chapter 2, feature selection is one of the criteria I kept in mind while choosing my method of analyses for Aim 1 work, as the scientific goal is to understand which brain ROI variables together are important for distinguishing between AD subgroups. So, a classification model such as logistic regression with regularization that has feature selection built into it, is useful for answering the question of Aim 1.

In section 3.3.2, I discuss how I chose values for α and λ while constructing the binary classification models for distinguishing between AD subgroups.

3.3 Model building

3.3.1 Training and Test data

As discussed in Chapter 2, due to the small sample size of the current cross-sectional dataset, I used the entirety of the current data sample as the training data and did not allocate a subset of the data as the test set. I assessed the performance of my models based on cross-validation error for the training data. I anticipate the test set based analyses to be future work as new data becomes available to form a reliable sized test dataset.

After SMOTE oversampling (see Section 2.8), below were the sample sizes of the training data for the different binary classification problems tackled using models based on logistic regression with regularization:

AD-Language (n=168) vs. AD-Memory (n=177)

AD-Memory (n=177) vs. AD-Visuospatial (n=183)

AD-Language (n=84) vs. AD-Visuospatial (n=61)

3.3.2 Cross-validation on training data to select tuning parameters α and λ

For each binary classification problem, I performed 5-fold cross-validation on the training set to find the tuning parameters α and λ that correspond to models with the smallest average deviance over the five folds. Deviance^x is defined as

$$D = 2 (\log(\ell(\text{sat})) - \log(\ell(\beta_0, \boldsymbol{\beta}))) \quad (6)$$

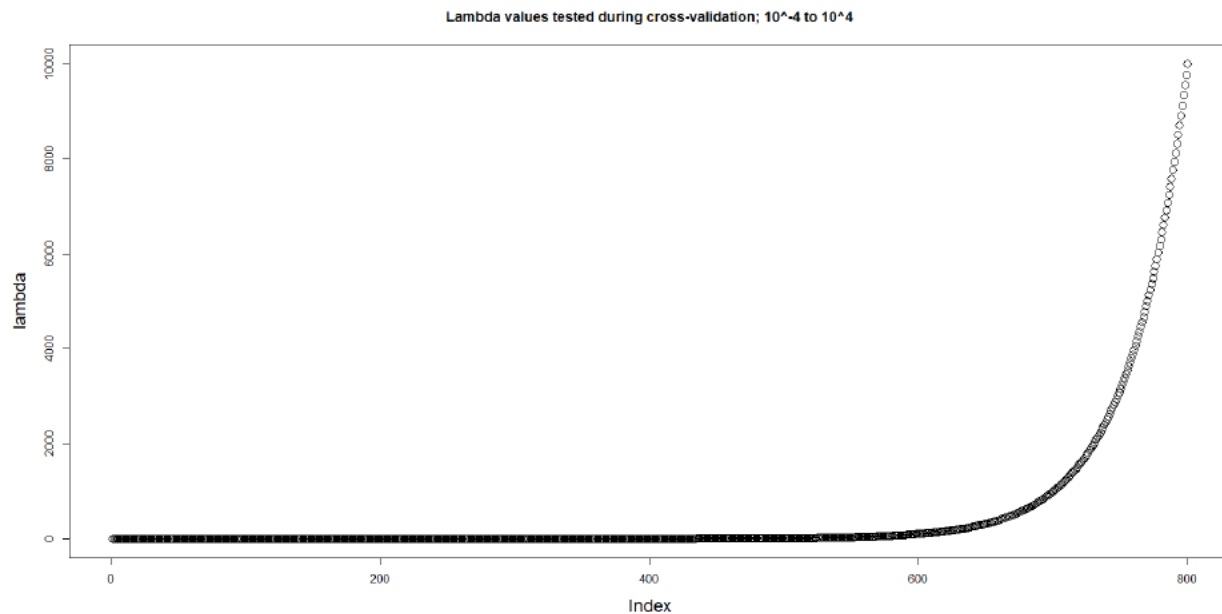
^x <https://www.rdocumentation.org/packages/glmnet/versions/4.0-2/topics/deviance.glmnet>

where $\log(\ell(\text{sat}))$ is the log-likelihood for the saturated model (a model with a free parameter per observation) and $\log(\ell(\beta_0, \boldsymbol{\beta}))$ is the log-likelihood of the model fit to the data. The deviance can be thought of as an indicator of how much of an improvement is the model of interest compared to the saturated model. $\ell(\beta_0, \boldsymbol{\beta})$ was defined in equation 3 and the expression for $-\log(\ell(\beta_0, \boldsymbol{\beta}))$ was shown in equation 5 for logistic regression with elastic net penalty. Since $\log(\ell(\text{sat}))$ is a constant term for the saturated model fit to the given data, minimizing deviance is related to the problem of minimizing the negative log-likelihood from equation 5. For determining the optimal set of values for α and λ for model fitting, I selected values that minimized average deviance calculated during 5-fold cross-validation.

For cross-validation, I divided the training data into five folds. Although SMOTE oversampling of the minority classes in each binary classification brought the ratio of the two classes close to 1, the ratio was not exactly equal to 1 due to limitations in the SMOTE package in R. Keeping the slight class imbalances in mind, I generated the folds so that the ratio of the two class sizes (after SMOTE oversampling) was maintained in each fold to ensure it reflects the ratio of the class sizes in the entire training data. Manually generating the folds for cross-validation instead of having an analysis package randomly assign individuals into folds also ensured reproducibility in analysis. Additionally, for model fitting, I specified weights for the penalty for misclassifying observations from the majority and minority classes as follows: $w = [w_{\text{minority}}, w_{\text{majority}}]$, where $w_{\text{minority}} = r * w_{\text{majority}}$, and r is the ratio of the sample size of the majority class to the sample size of the minority class in the training set; $r = \frac{n_{\text{majority}}}{n_{\text{minority}}}$. For each fold, models were fit using data from the other four folds using α values in the interval $[0,1]$ in increments of 0.05, and for each value of α , 800 values of λ ranging from 10^{-4} to 10^4 were used to construct 800 different models. The idea in choosing these specifics for the λ values were to cover a wide

range of lambda values and to test a large enough number of lambda values over the range. Figures 3.2a and 3.2b provide a visualization of the λ values considered for model building during cross-validation. For α , testing 20 equally spaced values between 0 and 1 seemed sufficient as the minimum average deviance values differed gradually between consecutive α values; this point will be more clear in Figures 3.3-3.5 that are discussed later. I used the `cv.glmnet` function in R to implement the above specifics.

A. λ values tested during 5-fold cross-validation



B. Powers of 10 in λ values tested during 5-fold cross-validation

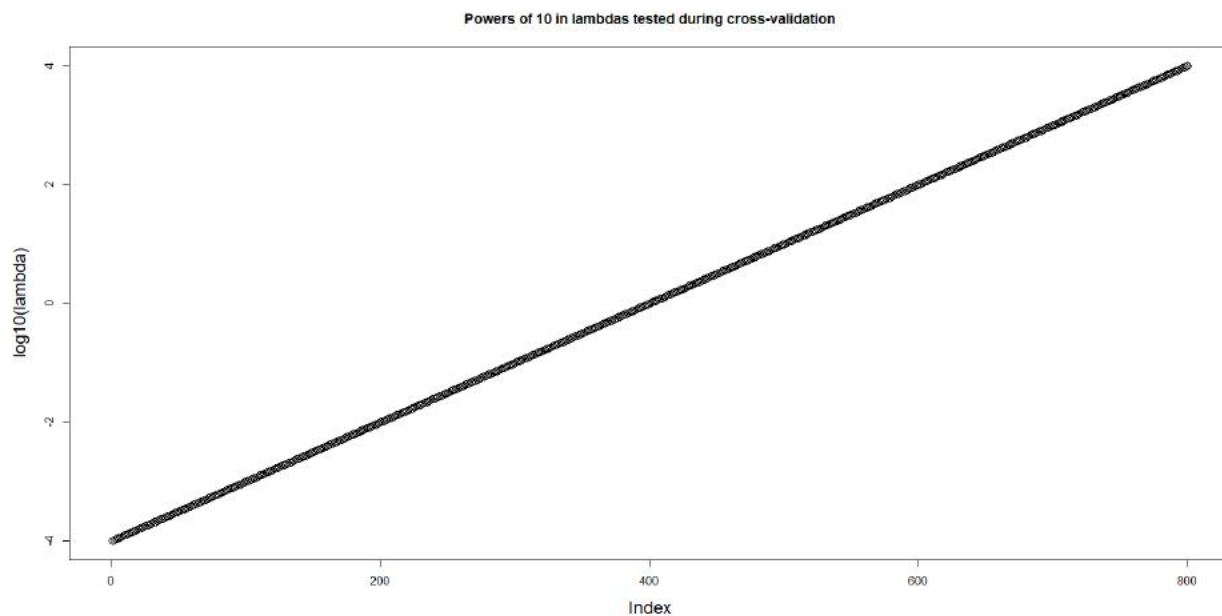


Figure 3.2: Both A and B show the λ values tested during cross-validation, but in different ways. The two figures highlight some important features of the set of λ values tested during cross-validation. A. 800 exponentially spaced values between 10^{-4} and 10^4 were tested during cross-validation. B. When expressed as powers of 10, the values in the exponent of the λ values are equally spaced between -4 and 4; there are 800 of these base 10 exponent values.

The vector of 800 λ values used is on an exponential scale; the intervals between consecutive values of λ increase for larger values of λ . Such a scale is used to cover a wide range of λ values that can be tested during cross-validation, without linearly increasing the number of λ values. I picked the λ range from 10^{-4} to 10^4 as it allowed me to cover the range of λ values over which the average cross-validated deviance curve reaches a minimum, as seen in the Results section. The range of λ values is often adjusted depending on whether the range is able to cover values over which the cross-validated deviance curve reaches a “valley” part of a curve. Figures 3.3, 3.5 and 3.7 show that the chosen λ range was sufficient to identify the minimum value from each average deviance curve from 5-fold CV for all 20 α values, for all three classification problems being considered.

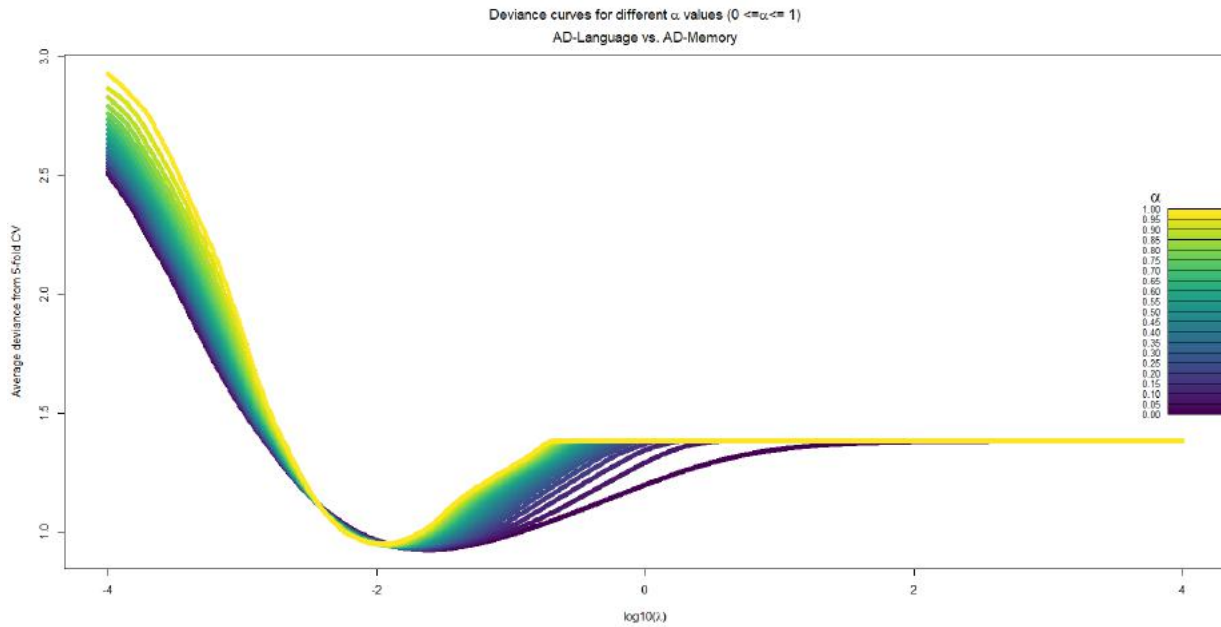


Figure 3.3 Average deviance from 5-fold CV as a function of $\log_{10}(\lambda)$ for each alpha value tested during CV, for AD-Language vs. AD-Memory regularized logistic regression classification models.

So, in each k th fold, for each value of α , there were 800 different models that were constructed using data from the other folds and validated using data from the current (k th) fold. In the validation procedure in each fold, the deviance measure defined in Equation 6 was used to assess the performance of each model corresponding to a given set of α and λ values. Deviances calculated through each fold of model validation were averaged over the five folds to obtain a cross-validated average deviance for each model corresponding to a particular α and λ set (see figures 3.3, 3.5 and 3.7). I used the `cv.glmnet` function in R to perform this for each α , using the “deviance” option for the `type.measure` parameter. For each α value, I obtained a λ value that corresponds to models with the lowest average deviance. Then from these twenty sets of α and λ values, I chose the set that corresponds to a model with the lowest average cross-validated deviance. In total, the above process involved 5-fold cross-validation of a total of $800 * 20 = 16,000$ models (corresponding to unique combinations of α and λ values) for each binary classification problem of the AD-subgroups. Figures 3.4, 3.6 and 3.8 below summarize the findings from the above process of cross-validation for finding the best suited set of α and λ values for model fitting for the three binary classifications problems.

For AD-Language vs. AD-Memory, $\alpha = 0$ and $\lambda = 0.0247$ resulted in the lowest average deviance during cross-validation. This corresponds to the ridge penalty, which means all 70 brain ROI volume variables got included in the model. Figure 3.4 shows the number of non-zero coefficients in the model corresponding to lowest average deviance at each α value.

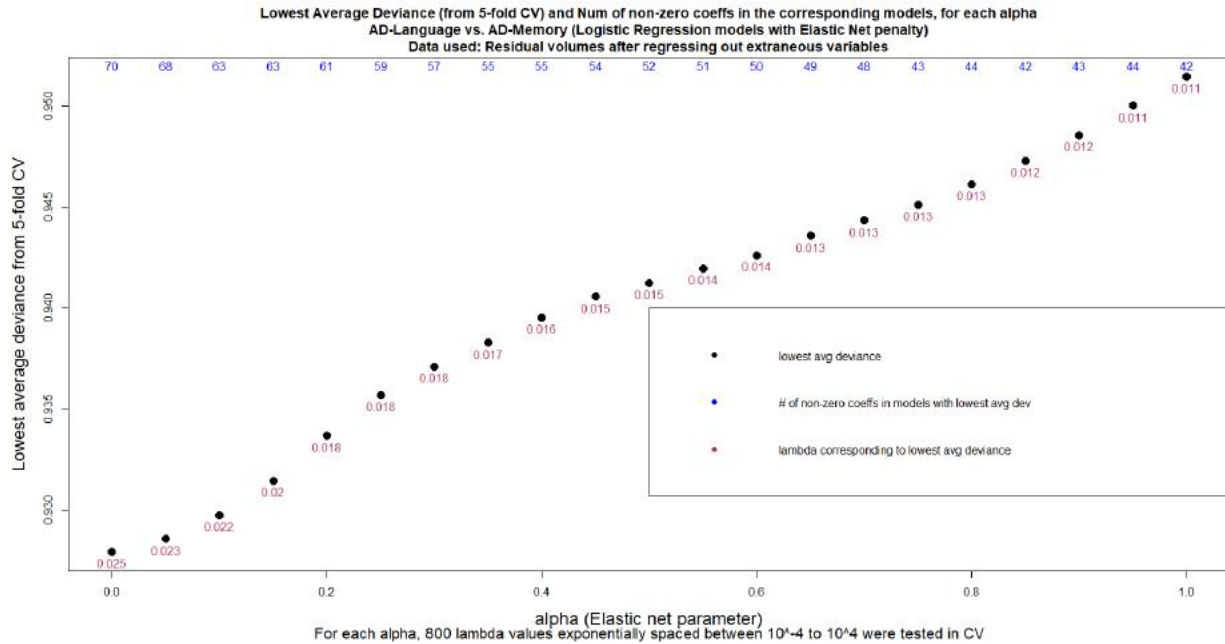


Figure 3.4 Lowest average deviance form 5-fold CV for models for AD-Language vs. AD-Memory for each alpha value and number of non-zero coefficients in the corresponding model.

For AD-Memory vs. AD-Visuospatial, the lowest average deviance from 5-fold cross-validation corresponded to the models with $\alpha = 0.25$ and $\lambda = 0.0430$ (See Figure 3.6). This resulted in 46 non-zero coefficients in the chosen model for AD-Memory vs. AD-Visuospatial.

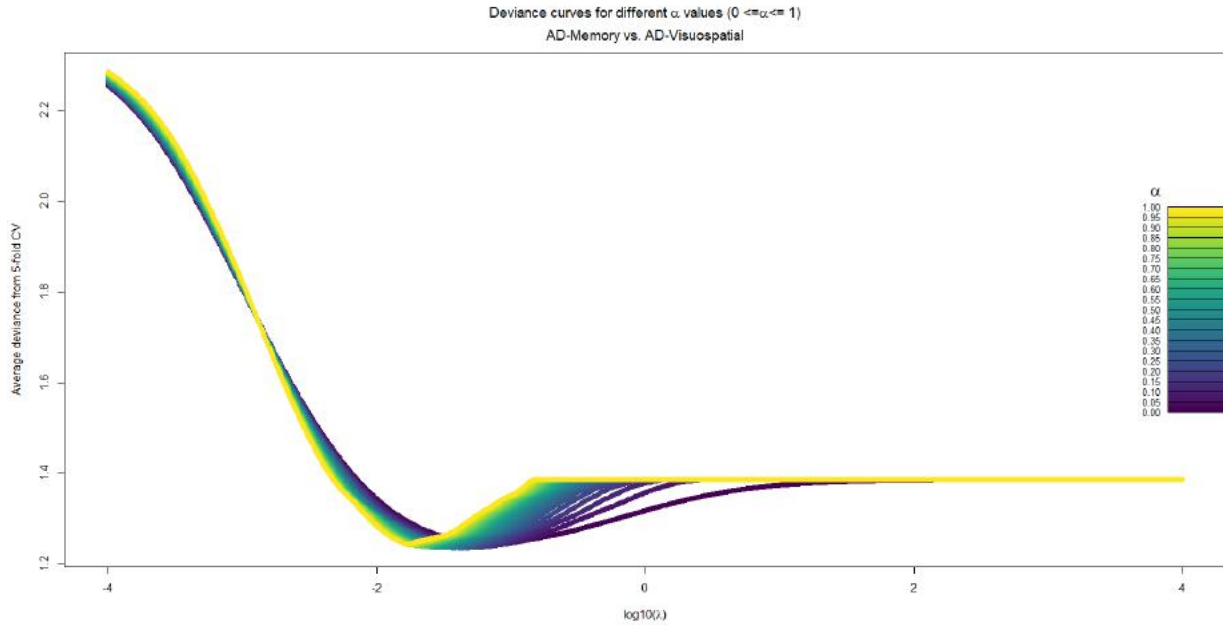


Figure 3.5 Average deviance from 5-fold CV as a function of $\log_{10}(\lambda)$ for each alpha value tested during CV, for AD-Memory vs. AD-Visuospatial regularized logistic regression classification models.

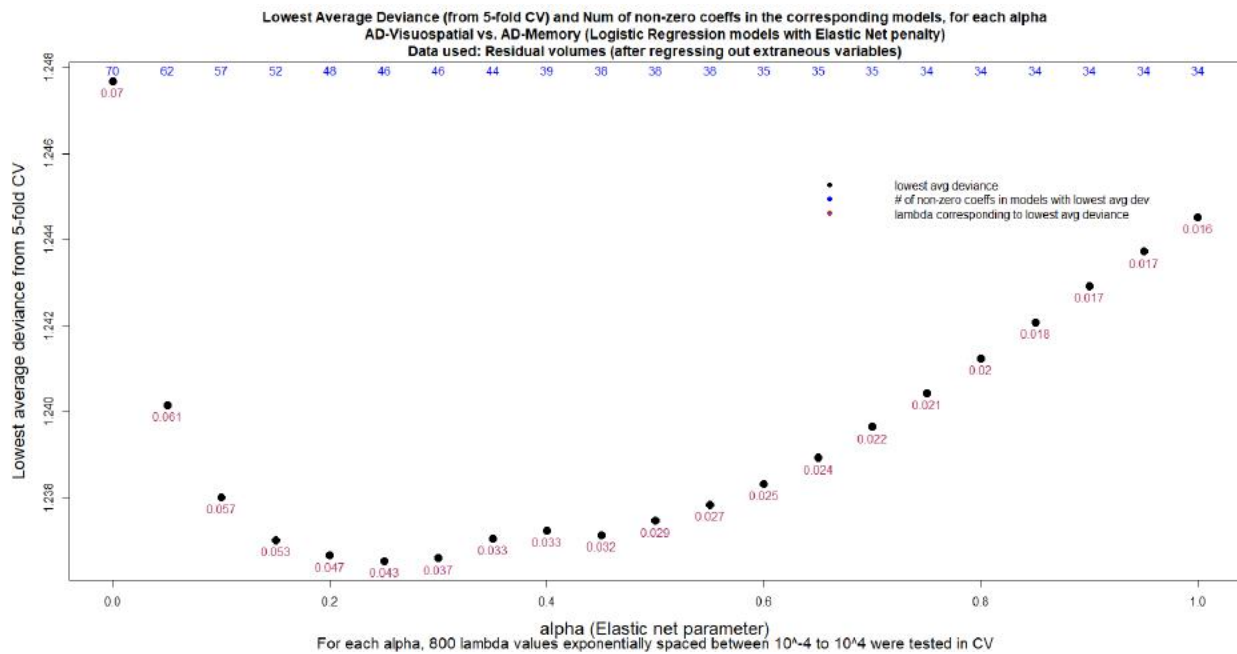


Figure 3.6 Lowest average deviance from 5-fold CV for models for AD-Memory vs. AD-Visuospatial for each alpha value and number of non-zero coefficients in the corresponding model.

Lastly, for AD-Language vs. AD-Visuospatial, models corresponding to $\alpha = 0$ and $\lambda = 0.2596$ yielded the lowest average deviance during 5-fold cross-validation (see Figure 3.8).

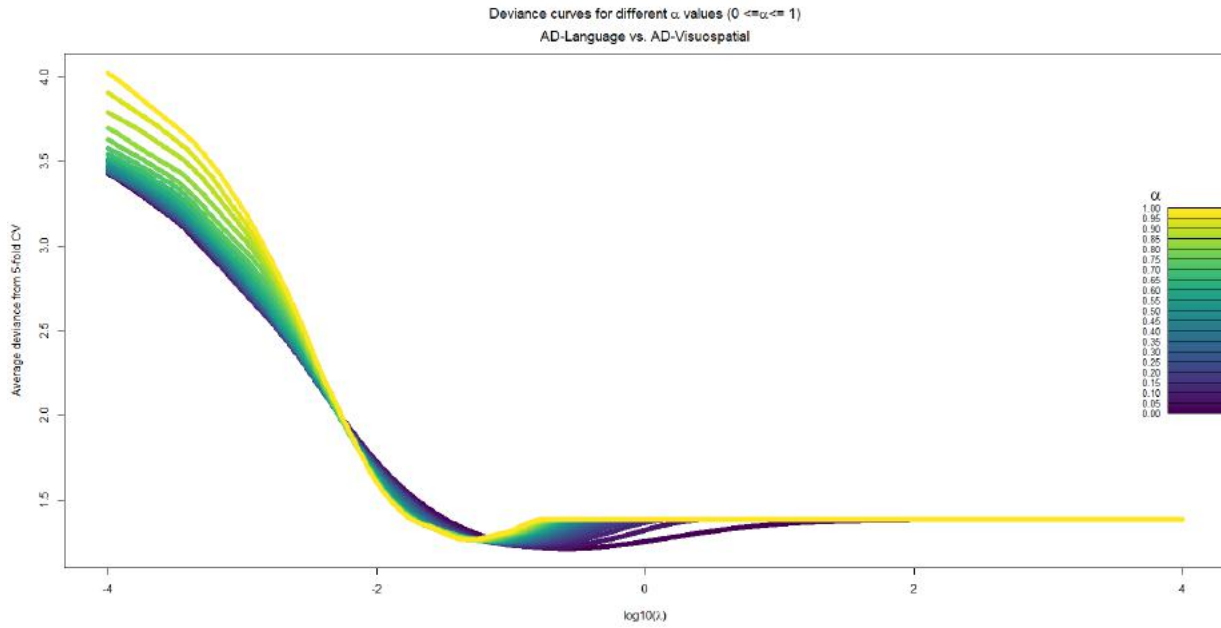


Figure 3.7 Average deviance from 5-fold CV as a function of $\log_{10}(\lambda)$ for each α value tested during CV, for AD-Language vs. AD-Visuospatial regularized logistic regression classification models.

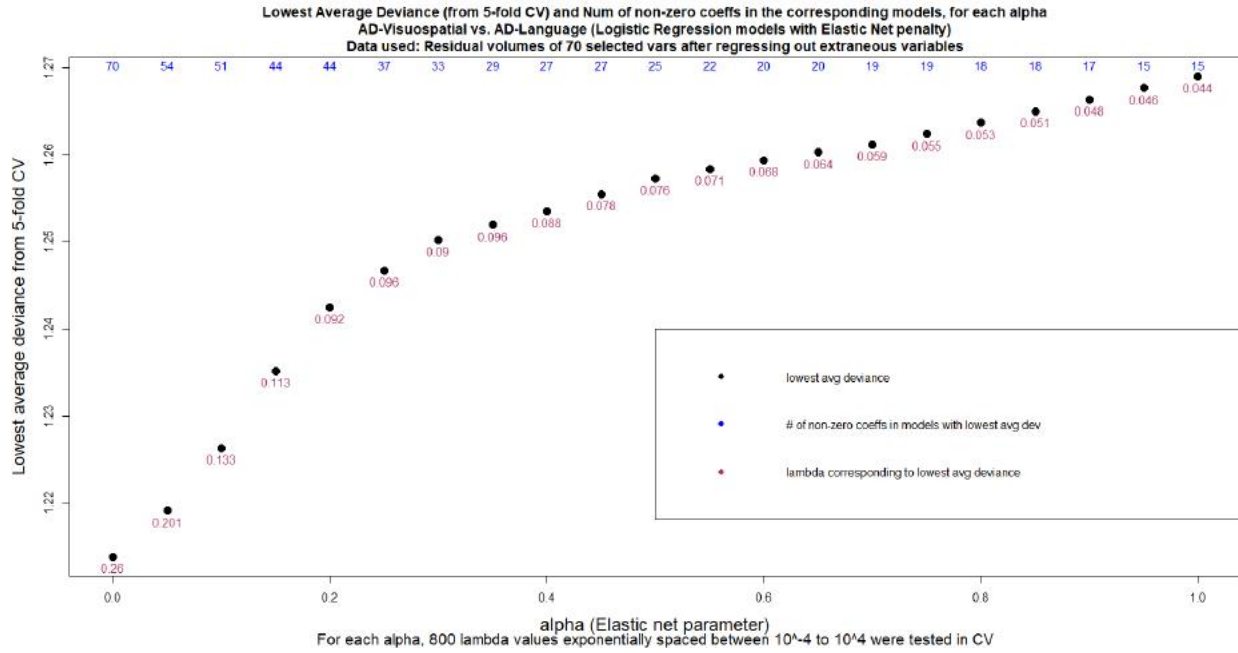


Figure 3.8 Lowest average deviance form 5-fold CV for models for AD-Language vs. AD-Visuospatial for each alpha value and number of non-zero coefficients in the corresponding model.

3.4 Model coefficients and model accuracy

Using the tuning parameters determined during cross-validation (in section 3.3), I fit a model to the entire training data for each of the following: AD-Language vs. AD-Memory, AD-Memory vs. AD-Visuospatial and AD-Language vs. AD-Visuospatial, to obtain the corresponding coefficients for each model. Since I did not have an allocated test data set to evaluate model performance, I instead, used the average misclassification error from cross-validation corresponding to the chosen α and λ values, as an estimate for the test error for the models. The misclassification errors (MCEs) based on 5-fold cross-validation were: 20.64% for AD-Language vs. AD-Memory, 33.12% for AD-Memory vs. AD-Visuospatial and 30.39% for AD-Language vs. AD-Visuospatial. The total sample size for the AD-Memory vs. AD-Visuospatial classification was slightly larger than AD-Language vs. AD-Memory. The relative average MCEs above suggest that simply having a larger sample size is not the reason for better

performance in logistic regression based classification. AD-Language vs. AD-Memory may actually be more separable than the other two pairwise comparisons.

Standardized coefficients of the models are provided in Tables 3.1-3.3; the β coefficients are sorted in increasing order. The magnitude of a given coefficient can be roughly interpreted as how important the corresponding variable is in distinguishing between the two AD subgroups, given all other ROI variables in the model. To focus on ROI volume variables that had relatively higher magnitudes of coefficients, I plotted the magnitudes of the coefficients in each model, as shown in figures 3.6-3.8. Below, I discuss the main findings from each model. The threshold of coefficient magnitudes to determine the most important variables for a given binary classification is subjective and arbitrary. As a stringent strategy, my approach gave a spotlight to variables with much higher magnitude coefficients than the rest of the variables in the model.

3.4.1 AD-Language vs. AD-Memory

The coefficients of the AD-Language vs. AD-Memory model are presented in Table 3.1. The AD-Language vs. AD-Memory classification model (as coded) predicted the probability of being in the AD-Memory group. Hence, variables with positive coefficients correspond to variables where greater volumes are associated with a higher probability of being in the AD-Memory group compared to the AD-Language group. Similarly, variables with negative coefficients correspond to variables where smaller volumes (more atrophy in an ROI volume) are associated with a higher probability of being in the AD-Memory group. Based on a visual evaluation of the plot of coefficient magnitudes in Figure 3.6, I found two variables to have much higher magnitude coefficients than the rest. Referring to Table 3.1, one can see that the ROI volumes corresponding to these high magnitude coefficients are right entorhinal volume (-0.876) and right superiorfrontal volume (-0.810). This indicates that lower volumes in the right entorhinal cortex

and the right superiorfrontal gyrus are associated with a higher probability of being in the AD-Memory group than AD-Language. It is important to note that the importance of ROIs in the model is on a continuous scale, and while my analysis of comparing the coefficients of ROIs showed that the above two ROIs had much higher magnitude coefficients, it does not mean that the variables with coefficients smaller in magnitude than above should be completely disregarded when looking for important ROIs.

Table 3.1: Classification model for
AD-Language vs. AD-Memory
Logistic Regression with regularization
($\alpha = 0$ and $\lambda = 0.0247$)

Number of non-zero coefficients in the model: 70

	Standardized coefficients
Intercept	0.185304
rh_entorhinal_volume	-0.87568
rh_superiorfrontal_volume	-0.81033
rh_rostralmiddlefrontal_volume	-0.54044
Right.Hippocampus	-0.4731
lh_posteriorcingulate_volume	-0.41306
lh_entorhinal_volume	-0.40187
lh_frontalpole_volume	-0.40057

rh_fusiform_volume	-0.35457
lh_lateraloccipital_volume	-0.33725
lh_parstriangularis_volume	-0.29299
Left.Hippocampus	-0.28475
lh_lingual_volume	-0.27985
lh_precentral_volume	-0.26351
rh_postcentral_volume	-0.23316
rh_parsopercularis_volume	-0.22724
rh_superiortemporal_volume	-0.20904
lh_parsorbitalis_volume	-0.20726
lh_pericalcarine_volume	-0.19891
rh_caudalanteriorcingulate_volume	-0.19713
rh_rostralanteriorcingulate_volume	-0.17951
rh_precuneus_volume	-0.16016
lh_caudalanteriorcingulate_volume	-0.15852
lh_paracentral_volume	-0.12817
lh_middletemporal_volume	-0.12251
rh_parsorbitalis_volume	-0.10337
lh_caudalmiddlefrontal_volume	-0.07616
lh_superiorparietal_volume	-0.07512
lh_bankssts_volume	-0.06332
rh_inferiorparietal_volume	-0.06278
lh_temporalpole_volume	-0.06081

lh_inferiorparietal_volume	-0.06014
lh_isthmuscingulate_volume	-0.02592
rh_lingual_volume	-0.02237
rh_caudalmiddlefrontal_volume	-0.01043
rh_isthmuscingulate_volume	-0.00762
rh_frontalpole_volume	0.00034
lh_supramarginal_volume	0.024568
rh_medialorbitofrontal_volume	0.027024
rh parahippocampal_volume	0.028052
rh_transversetemporal_volume	0.028712
lh_cuneus_volume	0.029381
rh_middletemporal_volume	0.066253
rh_bankssts_volume	0.067227
rh_supramarginal_volume	0.075562
rh_paracentral_volume	0.079796
lh_parsopercularis_volume	0.100012
rh_precentral_volume	0.136992
rh_posteriorcingulate_volume	0.138819
lh_precuneus_volume	0.13983
lh parahippocampal_volume	0.141146
rh_insula_volume	0.14311
lh_insula_volume	0.157009
lh_fusiform_volume	0.184948

lh_rostralanteriorcingulate_volume	0.185198
rh_superiorparietal_volume	0.194033
lh_lateralorbitofrontal_volume	0.208654
rh_inferiortemporal_volume	0.229043
rh_pericalcarine_volume	0.232077
rh_lateraloccipital_volume	0.248581
lh_superiortemporal_volume	0.276005
lh_postcentral_volume	0.296633
lh_transversetemporal_volume	0.298224
lh_medialorbitofrontal_volume	0.312501
rh_temporalpole_volume	0.396523
lh_inferiortemporal_volume	0.410534
rh_parstriangularis_volume	0.419537
rh_cuneus_volume	0.429712
lh_superiorfrontal_volume	0.438756
lh_rostralmiddlefrontal_volume	0.46696
rh_lateralorbitofrontal_volume	0.503051

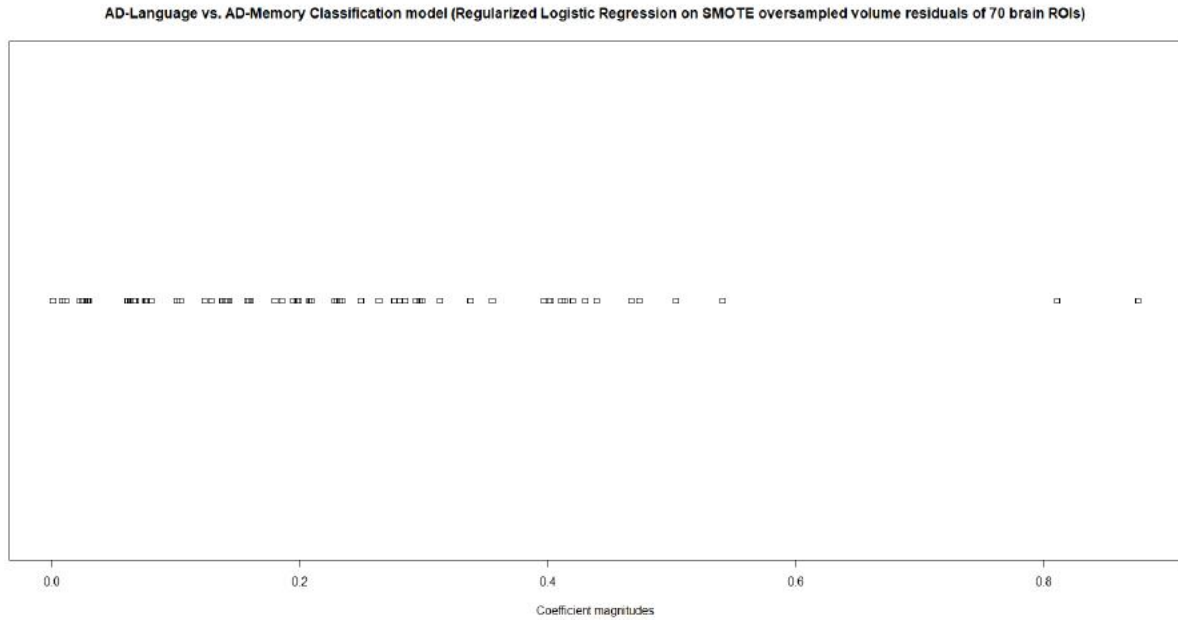


Figure 3.6 Magnitudes of coefficients in the AD-Language vs. AD-Memory model (Regularized logistic regression; $\alpha = 0$ and $\lambda = 0.0247$).

3.4.2. AD-Memory vs. AD-Visuospatial

The model with the lowest average deviance from 5-fold cross-validation for AD-Memory vs. AD-Visuospatial corresponded to $\alpha = 0.25$ and $\lambda = 0.0430$. Fitting a model to the entire training data with these parameters yielded a model with 46 non-zero coefficients; the model coefficients are presented in Table 3.2. Here, AD-Visuospatial is the target class for classification. Positive coefficients correspond to ROIs that have greater volume in the AD-Visuospatial group compared to the AD-Memory group. Alternatively, greater atrophy in these ROIs in an individual corresponds to a higher probability of the individual being in the AD-Memory group. Negative coefficients correspond to ROIs with greater atrophy in the AD-Visuospatial group on average compared to the AD-Memory group. Visually comparing the magnitudes of coefficients (Figure 3.7), I selected ROI volume variables corresponding to the top four magnitudes of coefficients as the more important ROIs for distinguishing between the AD-Memory and AD-Visuospatial groups. Again, this does not mean that any variables with coefficients smaller in

magnitude are not important. The top four important ROI volume variables based on coefficient magnitudes are left entorhinal cortex (0.434), right entorhinal cortex (0.283), right cuneus (-0.400) and right supramarginal gyrus (-0.359). Having greater atrophy in left entorhinal cortex and right entorhinal cortex is associated with a higher probability of being in the AD-Memory group compared to the AD-Visuospatial group. Having greater atrophy in the right cuneus and right supramarginal gyrus is associated with a higher probability of being in the AD-Visuospatial group.

Table 3.2: Classification model for
AD-Memory vs. AD-Visuospatial
Logistic Regression with regularization
($\alpha = 0.25$ and $\lambda = 0.0430$)

Number of non-zero coefficients in the model: 46

	standardized_coeff
intercept	0.143148
rh_cuneus_volume	-0.39815
rh_supramarginal_volume	-0.35909
lh_precentral_volume	-0.20961
lh_lateralorbitofrontal_volume	-0.17636
rh_temporalpole_volume	-0.15319
rh_postcentral_volume	-0.14691
rh_inferiortemporal_volume	-0.10449

lh_medialorbitofrontal_volume	-0.0992
rh_fusiform_volume	-0.09731
rh_superiorparietal_volume	-0.0755
rh_transversetemporal_volume	-0.0679
lh_pericalcarine_volume	-0.06635
rh_precentral_volume	-0.06276
lh_paracentral_volume	-0.06157
lh_bankssts_volume	-0.05649
rh_superiorfrontal_volume	-0.05054
rh_paracentral_volume	-0.03942
rh_inferiorparietal_volume	-0.03699
lh_parstriangularis_volume	-0.02837
rh_frontalpole_volume	-0.0057
lh_caudalmiddlefrontal_volume	-0.00456
lh_supramarginal_volume	0.009289
lh_parsopercularis_volume	0.013215
lh_superiortemporal_volume	0.014513
rh_pericalcarine_volume	0.018212
rh_precuneus_volume	0.022222
rh_parstriangularis_volume	0.029038
rh_lateraloccipital_volume	0.031696
rh_lingual_volume	0.032447
rh_medialorbitofrontal_volume	0.034615

rh_posteriorcingulate_volume	0.045634
lh_isthmuscingulate_volume	0.047394
lh_middletemporal_volume	0.070091
rh_isthmuscingulate_volume	0.076213
rh_lateralorbitofrontal_volume	0.105194
rh_superiortemporal_volume	0.113576
rh_parsopercularis_volume	0.119713
rh_caudalmiddlefrontal_volume	0.120178
lh_transversetemporal_volume	0.157949
lh_superiorfrontal_volume	0.186396
lh_lateraloccipital_volume	0.200335
lh_caudalanteriorcingulate_volume	0.214554
rh parahippocampal_volume	0.221048
rh_parsorbitalis_volume	0.241707
rh_entorhinal_volume	0.282579
lh_entorhinal_volume	0.433645

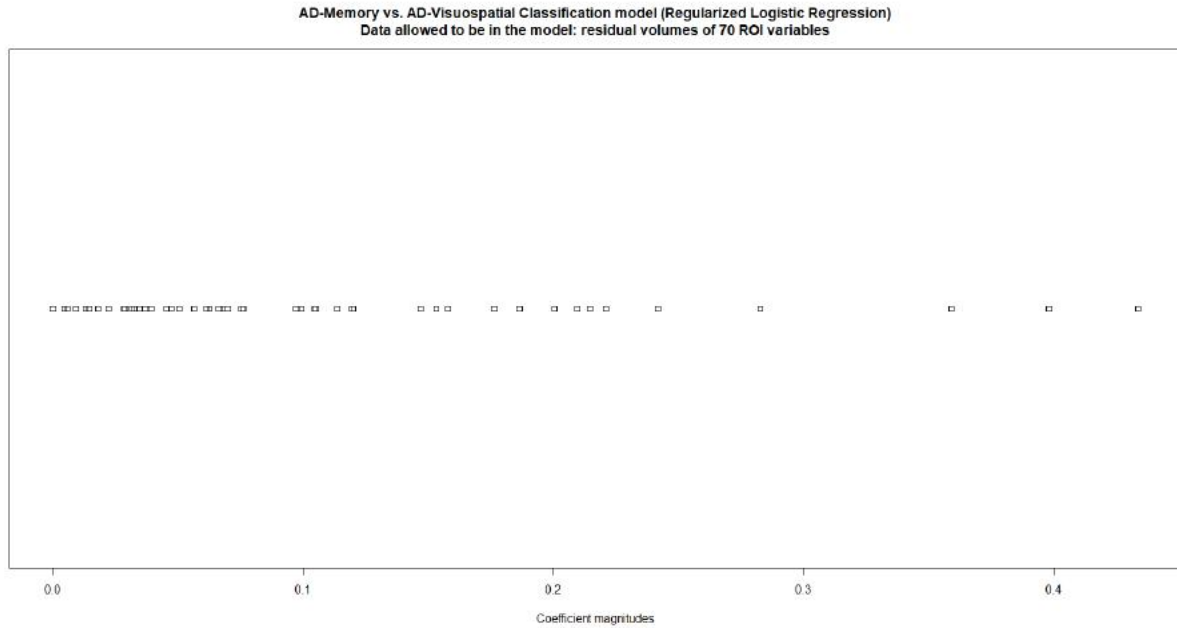


Figure 3.7: Magnitudes of coefficients in the AD-Memory vs. AD-Visuospatial model (Regularized logistic regression; $\alpha = 0.25$ and $\lambda = 0.0430$).

3.4.3 AD-Language vs. AD-Visuospatial

Lowest average deviance from 5-fold CV for AD-Language vs. AD-Visuospatial models was for the model corresponding to $\alpha = 0$ (ridge regression) and $\lambda = 0.2596$. The coefficients for all 70 ROIs are shown in Table 3.3; the target class for classification was AD-Visuospatial. Figure 3.7 shows the spread of the standardized coefficients for the model. Based on a visual inspection of this spread, I selected variables corresponding to the largest six coefficient magnitudes as the variables of greater importance as they seemed to be in a distinct cluster than the rest of the coefficient magnitudes. These coefficients correspond to the following ROI volumes: right hippocampus (-0.228), right entorhinal cortex (-0.210), right superior frontal gyrus (-0.201), left transverse temporal gyrus (0.172), left inferior temporal gyrus (0.162) and left superior temporal gyrus (0.161). The signs of these coefficients reveal that greater atrophy in the right hippocampus, right entorhinal cortex and right superior frontal gyrus is associated with

a higher probability of an individual being in the AD-Visuospatial group. Greater atrophy in the left transverse temporal gyrus, left inferior temporal gyrus and left superior temporal gyrus is associated with higher probability of being in the AD-Language group compared to the AD-Visuospatial group.

Table 3.3: Classification model for
AD-Language vs. AD-Visuospatial
Logistic Regression with regularization
($\alpha = 0$ and $\lambda = 0.2596$)

Number of non-zero coefficients in the model: 70

	standardized_coeff
intercept	-0.40902
Right.Hippocampus	-0.22802
rh_entorhinal_volume	-0.21035
rh_superiorfrontal_volume	-0.20074
rh_supramarginal_volume	-0.12458
rh_postcentral_volume	-0.12119
lh_posteriorcingulate_volume	-0.12043
rh_fusiform_volume	-0.11675
lh_precentral_volume	-0.10376
rh_precentral_volume	-0.09943
rh_cuneus_volume	-0.09667

lh_frontalpole_volume	-0.09513
lh_paracentral_volume	-0.09177
rh_superiorparietal_volume	-0.07976
rh_caudalanteriorcingulate_volume	-0.0703
lh_superiorparietal_volume	-0.06895
rh_transversetemporal_volume	-0.06492
rh_frontalpole_volume	-0.05611
rh_bankssts_volume	-0.05315
Left.Hippocampus	-0.05166
lh_pericalcarine_volume	-0.04433
rh_paracentral_volume	-0.04328
rh_inferiorparietal_volume	-0.04188
rh_rostralanteriorcingulate_volume	-0.04128
rh_precuneus_volume	-0.04067
lh_lingual_volume	-0.03585
lh_bankssts_volume	-0.02498
rh_rostralmiddlefrontal_volume	-0.02121
rh_temporalpole_volume	-0.01881
lh_caudalmiddlefrontal_volume	-0.0154
lh_cuneus_volume	-0.01442
lh_precuneus_volume	-0.01008
lh_parstriangularis_volume	-0.00889
lh_postcentral_volume	-0.00424

lh_caudalanteriorcingulate_volume	-0.00063
lh_isthmuscingulate_volume	0.000767
rh_middletemporal_volume	0.005603
lh_parsorbitalis_volume	0.007848
rh_insula_volume	0.012905
lh_entorhinal_volume	0.020631
lh_inferiorparietal_volume	0.024853
rh_superiortemporal_volume	0.026345
rh_parsopercularis_volume	0.029186
lh_lateralorbitofrontal_volume	0.032739
rh_inferiortemporal_volume	0.034878
lh_medialorbitofrontal_volume	0.034896
rh_lingual_volume	0.046676
rh_parstriangularis_volume	0.047961
rh parahippocampal_volume	0.048775
rh_lateraloccipital_volume	0.052137
rh_caudalmiddlefrontal_volume	0.054147
rh_parsorbitalis_volume	0.059011
rh_isthmuscingulate_volume	0.059672
lh_rostralanteriorcingulate_volume	0.060567
lh_lateraloccipital_volume	0.063596
lh_supramarginal_volume	0.067028
lh_temporalpole_volume	0.068201

rh_medialorbitofrontal_volume	0.075228
lh_fusiform_volume	0.079016
lh_rostralmiddlefrontal_volume	0.08368
lh_middletemporal_volume	0.090911
rh_pericalcarine_volume	0.105969
lh_superiorfrontal_volume	0.10821
lh_parsopercularis_volume	0.116273
lh parahippocampal_volume	0.119612
lh_insula_volume	0.123326
rh_posteriorcingulate_volume	0.123407
rh_lateralorbitofrontal_volume	0.131016
lh_superiortemporal_volume	0.160876
lh_inferiortemporal_volume	0.161968
lh_transversetemporal_volume	0.172061

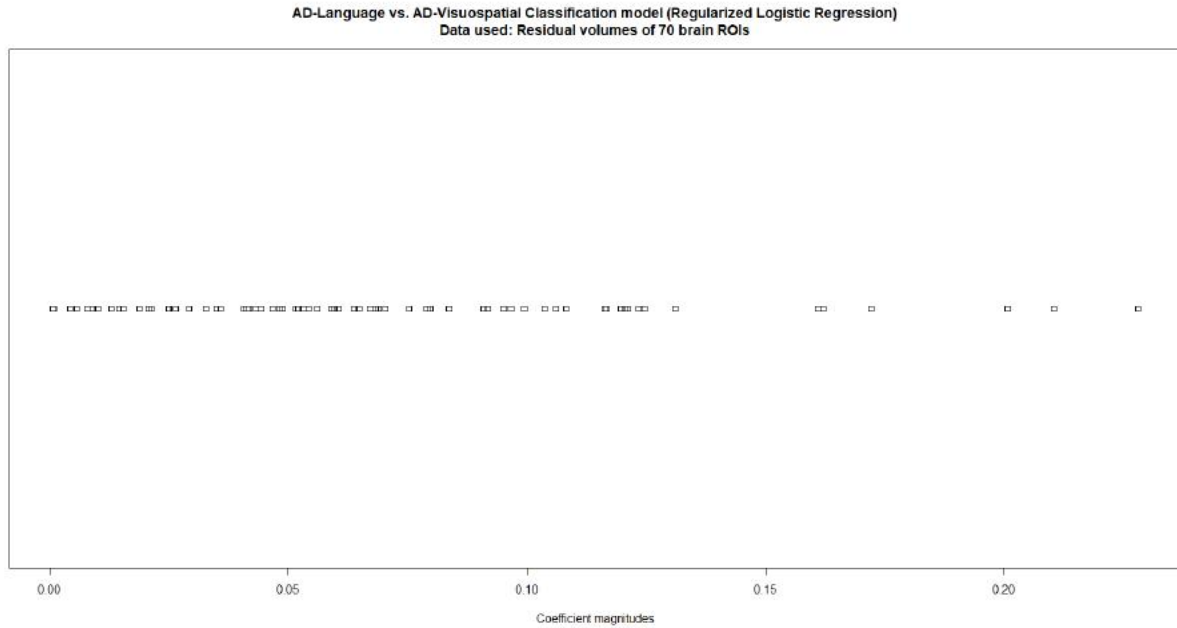


Figure 3.8: Magnitudes of coefficients in the AD-Language vs. AD-Visuospatial model (Regularized logistic regression; $\alpha = 0$ and $\lambda = 0.2596$).

3.5 Limitations of results of variable importance results based on penalized logistic regression

The goal of using classification models in my Aim 1 work was to interpret the models to shed light on which ROIs are most important in separating pairs of AD subgroups. There are a few suboptimal points to be noted in the results from classification models based on logistic regression with regularization. The first has to do with the classification accuracy of the models. The best classification accuracy was obtained for the AD-Language vs. AD-Memory group, with an average misclassification error calculated from 5-fold CV to be 20.64%. Since there is no known golden standard for the misclassification error for the binary classification models for the AD subgroups, it is difficult to assess whether a model with a classification accuracy of 79.36% is considered to be a “good model” or not. Nonetheless, intuitively, the greater the classification accuracy (alternatively, the smaller the misclassification error) for a model, the more reliable the model for interpreting variable importance. The classification accuracy for the other two

classifications were: 66.88% for AD-Memory vs. AD-Visuospatial and 69.61% for AD-Language vs. AD-Visuospatial. The current classification accuracies obtained for the models using regularized logistic regression may be considered modest. These accuracies suggested a need to turn to another type of classification model to see if that might yield better classification accuracies.

Another suboptimal aspect of regularized logistic regression is that it may not always result in feature selection if the goal is to find a model corresponding to smallest average deviance during cross-validation, as seen in Section 3.4. Secondly, when this method does result in feature selection, there is the possibility of missing out on variables that play a role in classification but are collinear with a variable included in the model. Below I discuss these points.

One of the main reasons for choosing regularized logistic regression as a first method of choice for building classification models was its feature selection capabilities when $0 < \alpha \leq 1$. Instead of arbitrarily choosing an alpha value (in the range $0 \leq \alpha \leq 1$) that results in zeroing of a specific number of coefficients, I chose the alpha value corresponding to the lowest average deviance during cross-validation, to obtain a model that would best separate the data from two classes. In two of the classifications, AD-Language vs. AD-Memory and AD-Language vs. AD-Visuospatial, the models with the lowest average deviance corresponded to $\alpha = 0$ (ridge regression) which results in all variables to be kept in the model. Hence, for these two classifications, having the model with the lowest average deviance meant having no feature selection. Although not ideal, this aspect (no feature selection) was not a major hurdle in determining variable importance as I used the relative magnitudes of coefficients of the variables in each model to determine the most important variables.

There is a greater care needed in the interpretation of a regularized logistic regression model, however, when the lowest average deviance model corresponds to a model that indeed results in feature selection (AD-Memory vs. AD-Visuospatial which had only 46 non-zero coefficients). Although feature selection was my original intent in using regularized logistic regression, upon reflecting more on the nature of brain ROI data, I realized that using regularized logistic regression for feature selection may not be optimal as the alpha based penalty results in zeroing of coefficient of a variable if it is correlated with another variable that is included in the model. In the case of brain ROI data, correlations among ROIs are expected due to physical proximity of ROIs in the brain. Hence, using feature selection based on regularized logistic regression, specifically for the case of $0 < \alpha \leq 1$, may not give a complete picture of which ROIs are important in distinguishing between a pair of AD subgroups.

The above points warranted analyzing the data using another classification method besides regularized logistic regression. I was also interested in seeing if the classification accuracies might improve if I used a non-linear method for classification. I chose random forest as the method to try next, which is the topic of the next chapter. Random forest is nowhere as interpretable as logistic regression but it does allow one to assess variable importance, so it satisfied the criteria laid in Chapter 2 for the methods I proposed to use for building classification models for distinguishing between pairs of AD subgroups.

Chapter 4: ROI importance based on Random Forest binary classification models distinguishing between pairs of AD subgroups [Aim 1]

In this chapter, I discuss the use of random forest classification models for determining variable importance among the 70 brain ROIs for distinguishing between the following pairs of AD subgroups: a.) AD-Language vs. AD-Memory, b.) AD-Memory vs. AD-Visuospatial and c.) AD-Language vs. AD-Visuospatial. As in the case of penalized logistic regression classification models discussed in the previous chapter, the random forest classification models here are not intended to be used for classification of new data in a clinical setting. Rather, the goal is to learn from the models based on current data about important features (volumes of brain regions) that best distinguish between pairs of subgroups. Here, I first provide a brief explanation for choosing random forest classification models to determine variable importance (Section 4.1), followed by details of the implementation of the models (Section 4.2), results (Section 4.3), discussion and interpretation of the findings (Section 4.4), some limitations of the current approach (Section 4.5) and conclusions and next steps to improve the analysis (Section 4.5). The final results of ROI importance for cross-sectional data (Aim 1) were deduced based on the random forest models discussed here.

4.1 Why Random Forest?

Although penalized logistic regression (PLR) was a good first method to explore for understanding variable importance through binary classification models of AD subgroups due to its ease of interpretability, the results in the previous chapter pointed at a few non-optimal aspects of the PLR classification models for answering the question of interest. Although there is no gold standard for classification accuracies for distinguishing between pairs of AD subgroups, PLR models yielded

classification accuracies that may be considered modest (66.88%, 69.1% and 79.36%), prompting for experimentation with another classification method that may lead to improvement in accuracies. Greater classification accuracies would in turn make the variable importance results from the models more reliable. Secondly, models with the elastic net penalty or the lasso penalty may not provide a complete story of which brain ROIs are important, as these penalties can result in zeroing of coefficients of correlated ROIs. In the data being analyzed, it is expected that there is some correlation between the ROI volumes due to physical proximity, hence the elastic net or the lasso penalty may not be best suited for the purposes of understanding relative ROI importance. Random forest qualifies as a good method to implement for the data at hand and for the question of interest for a few reasons. Variable importance can be obtained from random forest classification models even though the models themselves are not directly interpretable. Unlike logistic regression, random forest allows for non-linear decision boundaries for classes as well as interaction among variables in a single model. When working with complex biological data, such as brain ROI volumes, interactions and non-linear relationships are not uncommon. Additionally, random forest may better capture relationships among correlated variables, such as the brain ROIs. Random forest achieves these properties by decorrelating regression and decision trees to reduce variance and improve generalizability. The decorrelation is achieved by using a random subset of variables at each node to build each tree. This gives a fair chance for correlated variables to be considered for variable selection at the different nodes across multiple trees.

4.2 Implementation of Random Forest

Using SMOTE oversampled data of residual volumes of the 70 ROIs (described in chapter _) as input features, I generated models for the three classification problems used the RandomForest package (Liaw and Wiener 2002) in R: a.) AD-Language (n=168) vs. AD-Memory (n=177), b.)

AD-Memory (n=177) vs. AD-Visuospatial (n=183) and c.) AD-Language (n=84) vs. AD-Visuospatial (n=61). To ensure that the results are not seed dependent, for each of the three classification problems, I used 100 unique seeds to generate 100 different random forest models. In each random forest model, the number of trees was specified to be 500 which I reasoned to be large number for a dataset with 70 variables. In general, the larger the number of trees, the less overfit is the final model. For the number of randomly selected predictors considered at each split, I used the recommendation specified by James et. al (James et al. 2013, 319) of $m = \sqrt{p}$ for classification problems where $p=70$ is the total number of predictors.

By design of the Random Forest method, the data used for each tree is a bootstrapped subset of all observations (which is on average, two thirds of all observations), and the performance of each tree is tested on the remaining one third of the data, also known as the out-of-bag (OOB) observations. The class prediction for any observation i is based on averaging the prediction from all trees in which that observation was OOB, which yields around $B/3$ predictions for the i th observation (James et al. 2013, 318), where B is the total number of trees. The OOB error for the overall model is the misclassification error based on the averaged predictions from $B/3$ trees for each observation in the dataset. Whereas for penalized logistic regression models (discussed in Chapter 3), I reported average misclassification errors from cross-validation, in tree methods that are based on multiple trees and bootstrapped data, the OOB error is typically used to report misclassification error instead of formally carrying out cross-validation. The construct of using a bootstrapped subset of data to build a tree and using OOB samples to evaluate the performance of the tree is equivalent to cross-validation, though not in the controlled fashion of dividing data into folds. I preferred this choice of using the OOB error as a measure of model performance

rather than average misclassification error from a formal cross-validation procedure which would further reduce the sample size used for building trees during cross-validation after bootstrapping.

4.3 Variable importance results from Random Forest

Although a model based on a collection of trees, as in random forest, is more difficult to interpret than a single tree, one can obtain a summary of overall importance of the predictors in classification trees using a measure such as the mean decrease in accuracy. The mean decrease in accuracy indicates the decrease in accuracy in predictions on the OOB samples when a given predictor is excluded from the model (James et al. 2013). From the random forest models for each classification problem, I used the mean decrease in accuracy to identify the most important brain ROIs for distinguishing between two given AD subgroups. Specifically, for each of the three binary classification problems, I averaged the mean decrease in accuracy values from the 100 random forest models corresponding to the 100 different seeds and used the average values for comparing the ROIs for importance. It should be noted that there is no absolute threshold of the mean decrease in accuracy value for determining the most important variables. The variable importance scale is relative and the actual values of variable importance for the different variables have little meaning for interpretation. To provide a spatial context, I created visualizations on the brain surface using the `ggseg` (Mowinckel and Vidal-Piñeiro 2020) package in R, showing relative importance of all brain regions based on the averaged mean decrease in accuracy (from 100 random forest models) for each classification (See Figures 4.2 and 4.4 under Results).

Random forest variable importance can be used to determine the relative importance of the ROIs but it does not provide any information about how an ROI volume might differ across the AD subgroups being compared. This relationship is investigated in the next section (4.4). The focus

of this section is to bring attention to parts of the brain that seem to have greater importance in distinguishing between the pairs of AD subgroups based on the random forest classification models.

Results from all three classification problems are described below. It is not sufficient to simply obtain a ranking of importance for the ROI volumes based on the models for a given classification. It is important that the models from which the variable importance is deduced have a decent classification performance. We do not have a gold standard for what qualifies as a good classification performance for the AD subgroups, however, the better the classification performance, the more reliably we can state the results from variable importance from the given models. I used the OOB error as a measure of classification performance for the random forest models.

The OOB error estimates from the 100 random forest models for each classification were averaged to obtain the average OOB error. See Table 4.1 for a summary of the average OOB errors for each classification problem. The average OOB errors from random forest models for all three classification problems were lower than the average misclassification errors from cross-validation in the corresponding penalized logistic regression models (See chapter 3). Although the average OOB error and the average misclassification error are not the same exact measure, both measures represent an estimate of model performance. It seems reasonable to compare these measures for a rough and broad understanding of whether random forest models or penalized logistic regression models have better performance for the given data. Based on the model performance estimated by OOB classification errors, one can say that random forest classification models may be better suited for the given data compared to penalized logistic regression classification models. The better classification performance of the random forest

models consequently allows one to be more confident about reporting the relative importance of the ROIs deduced from the models.

a. AD-Language vs. AD-Memory

Avg overall OOB error for AD-Language vs. AD-Memory	Avg OOB error for AD-Language	Avg OOB error for AD-Memory
6.4%	5.8%	7.0%

b. AD-Memory vs. AD-Visuospatial

Avg overall OOB error for AD-Memory vs. AD-Visuospatial	Avg OOB error for AD-Memory	Avg OOB error for AD-Visuospatial
11.7%	14.5%	9.0%

c. AD-Language vs. AD-Visuospatial

Avg overall OOB error for AD-Language vs. AD-Visuospatial	Avg OOB error for AD-Language	Avg OOB error for AD-Visuospatial
21.1%	10.8%	35.2%

Table 4.1: Averaged Out of Bag (OOB) errors from 100 random forest models corresponding to 100 seeds for each of the classifications: a. AD-Language vs. AD-Memory, b. AD-Memory vs. AD-Visuospatial and c. AD-Language vs. AD-Visuospatial.

a. AD-Language vs. AD-Memory

Of the three classification problems, the distinction between the AD-Language and AD-Memory groups was the strongest, with an average OOB classification error estimate of $6.37\% \pm .06\%$ (standard error) over the 100 random forest models. The average OOB error for the AD-Language group was $5.75\% \pm 0.10\%$ and $6.70\% \pm 0.10\%$ for the AD-Memory group. The variable importance plot based on the ‘mean decrease in accuracy’ measure is shown in Figure 4.1. The topmost regions that stand out in importance (in decreasing order) based on this plot are the **right entorhinal cortex, right hippocampus, right lingual gyrus** and **left hippocampus**. Figure 4.2 provides a visualization of the ROI importance results on the brain surface for all 70 regions shown on the Desikan-Killiany cortical atlas and the subcortical atlas separately.

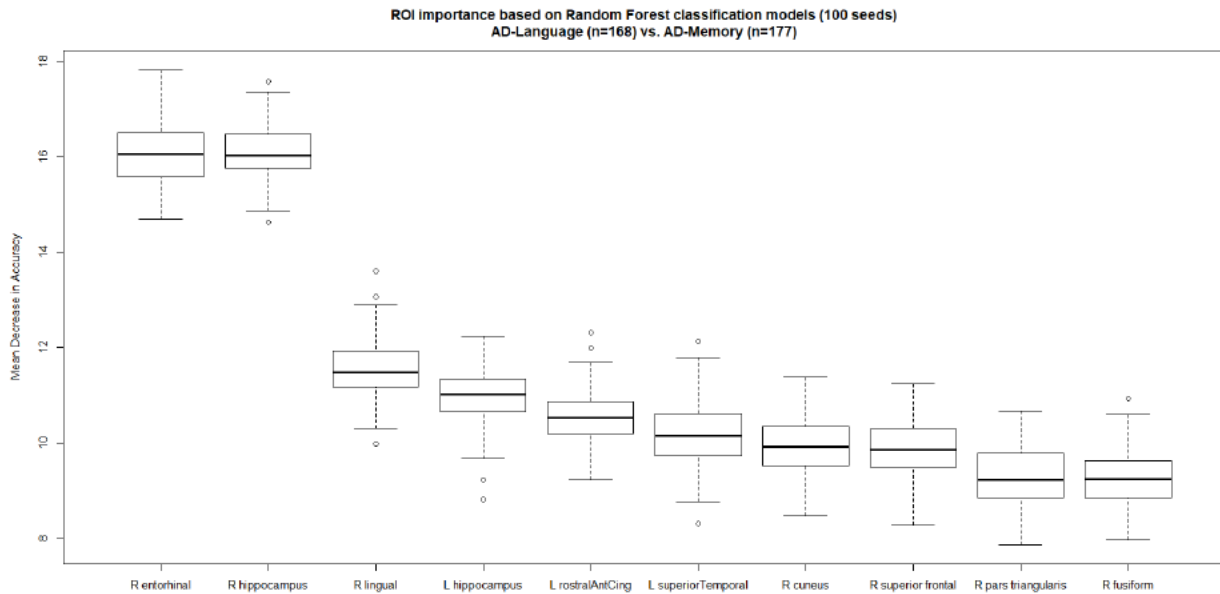


Figure 4.1: Random Forest variable importance from 100 models (100 seeds) for the AD-Language vs. AD-Memory classification. Each boxplot represents the distribution of the Mean Decrease in Accuracy from the 100 models. Top 10 variables are shown based on the Mean Decrease in Accuracy.

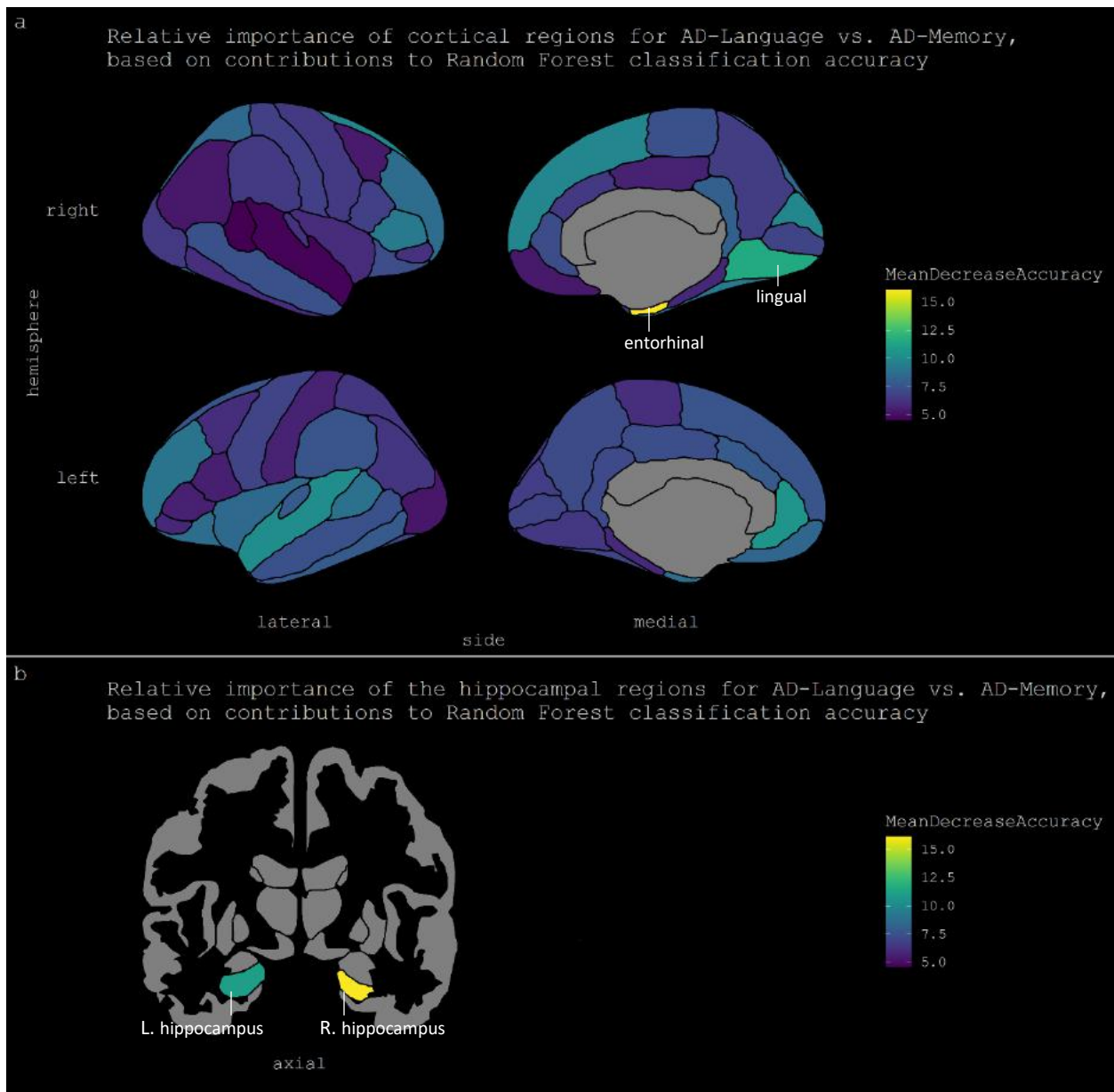


Figure 4.2: Relative importance of ROIs in the a.) Desikan-Killiany (DK) atlas and b.) the hippocampal regions, based on contributions to classification accuracy from the random forest models for AD-Language vs. AD-Memory. Mean Decrease in Accuracy values for each ROI were averaged over the 100 random forest models (corresponding to 100 seeds). The top four regions of importance are labeled.

b. AD-Memory vs. AD-Visuospatial

The AD-Memory vs. AD-Visuospatial classification had the next best distinction of the AD subgroups after the AD-Language vs. AD-Memory classification. The AD-Memory vs. AD-Visuospatial classification problem had an average OOB classification error estimate of $11.7\% \pm 0.08$ over the 100 random forest models. The individual class classification error averages were $14.5\% \pm 0.14\%$ for the AD-Memory group and $9.0\% \pm 0.07\%$ for the AD-Visuospatial group. Based on the contributions to classification accuracy in Random Forest (average value of mean decrease in accuracy for each ROI over the 100 random forest models), following were the top four ROIs (in descending order of importance): **left entorhinal cortex, right entorhinal cortex, right supramarginal gyrus and left postcentral gyrus**. As noted earlier, the scale of variable importance is continuous and the regions with lower mean decrease in accuracy measures than the above should not be completely disregarded when looking for ROI importance. The relative variable importance measures for all 70 ROIs and their locations on the brain are shown in Figure 4.4 based on the average of mean decrease in accuracy values from 100 random forest models.

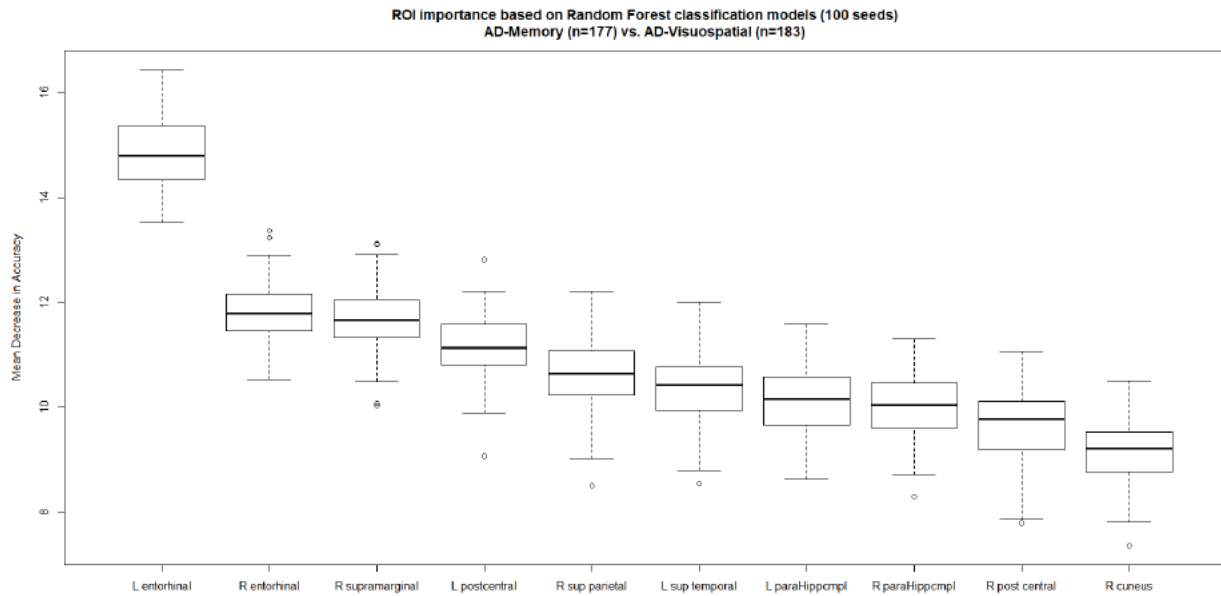


Figure 4.3: Random Forest variable importance from 100 models (100 seeds) for the AD-Memory vs. AD-Visuospatial classification. Each boxplot represents the distribution of the Mean Decrease in Accuracy from the 100 models. Top 10 variables are shown based on the Mean Decrease in Accuracy.

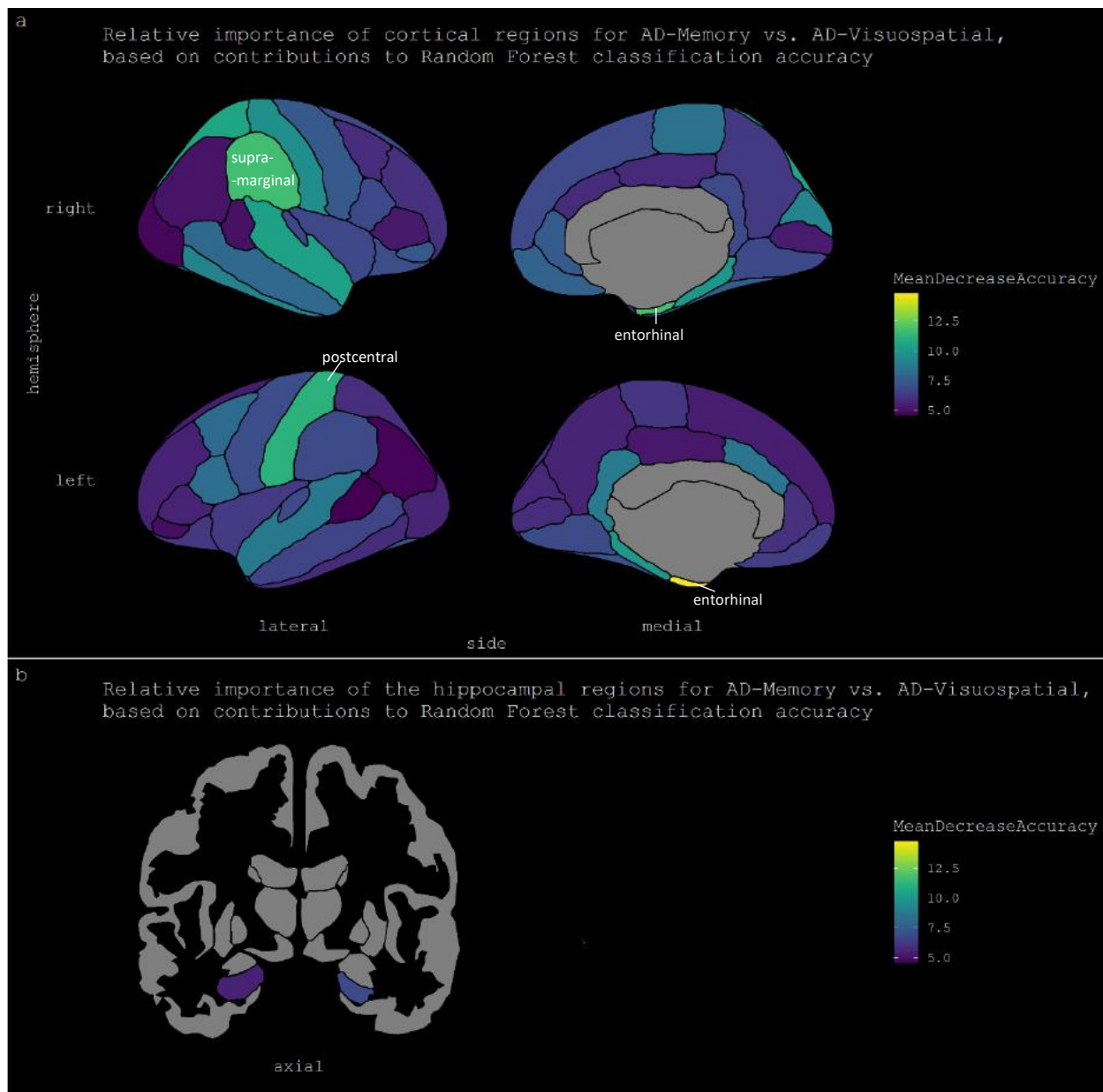


Figure 4.4: Relative importance of ROIs in the a.) Desikan-Killiany (DK) atlas and b.) the hippocampal regions, based on contributions to classification accuracy from the random forest models for AD-Memory vs. AD-Visuospatial. Mean Decrease in Accuracy values for each ROI were averaged over the 100 random forest models (corresponding to 100 seeds). The top four regions of importance are labeled.

c. AD-Language vs. AD-Visuospatial

The AD-Language and AD-Visuospatial groups were the most difficult to separate based on brain regions' volumes; the average OOB classification error from 100 random forest models was 21.1% \pm 0.16% and individual class classification OOB error averages were 10.8% \pm 0.14% for AD-Language and 35.3% \pm 0.37% for AD-Visuospatial. Note that the AD-Visuospatial group has a much higher average OOB error than the AD-Language group. As demonstrated by the disproportionate average OOB errors for the two classes here, the current models based on random forest do not separate the data from AD-Language and AD-Visuospatial groups well enough to be used for deducing ROI importance from them. For completeness, the variable importance rankings based on averaged mean decrease in accuracy values from 100 random forest models for AD-Language vs. AD-Visuospatial are shown in Figures 4.5 and 4.6 but these results should not be used for making remarks about which ROIs are important for distinguishing between the two AD subgroups. Future work is needed to understand whether these two groups are actually not that separable based on brain ROI volumes or if another method could yield a better and class-balanced classification performance. It is also worth probing in detail in the future why there is an imbalance in classification accuracies for AD-Language and AD-Visuospatial groups, and the scientific interpretation of this imbalance. A preliminary hypothesis could be that members of the AD-Language group have atrophy patterns that are more ROI specific whereas members of the AD-Visuospatial have much less regional specific atrophy patterns. Additional work is needed to corroborate this hypothesis.

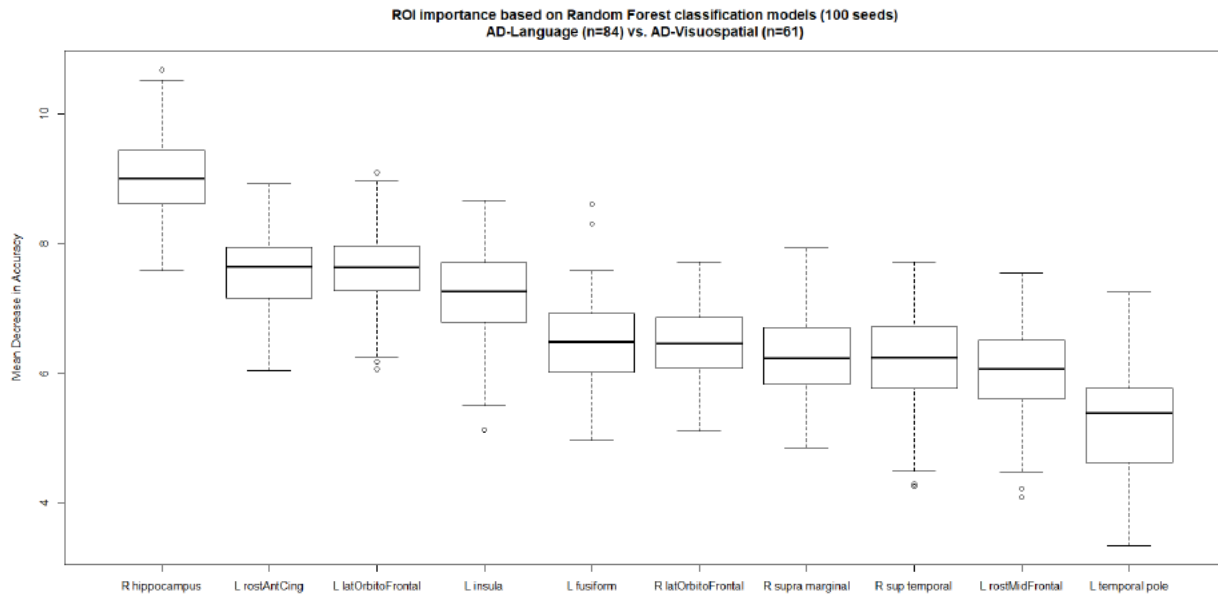


Figure 4.5 Random Forest variable importance from 100 models (100 seeds) for the AD-Language vs. AD-Visuospatial classification. Each boxplot represents the distribution of the Mean Decrease in Accuracy from the 100 models. Top 10 variables are shown based on the Mean Decrease in Accuracy. Due to the great imbalance in classification accuracies for the two subgroups from the current random forest models for this classification problem, the above results should not be used for a reliable reporting of ROI importance.

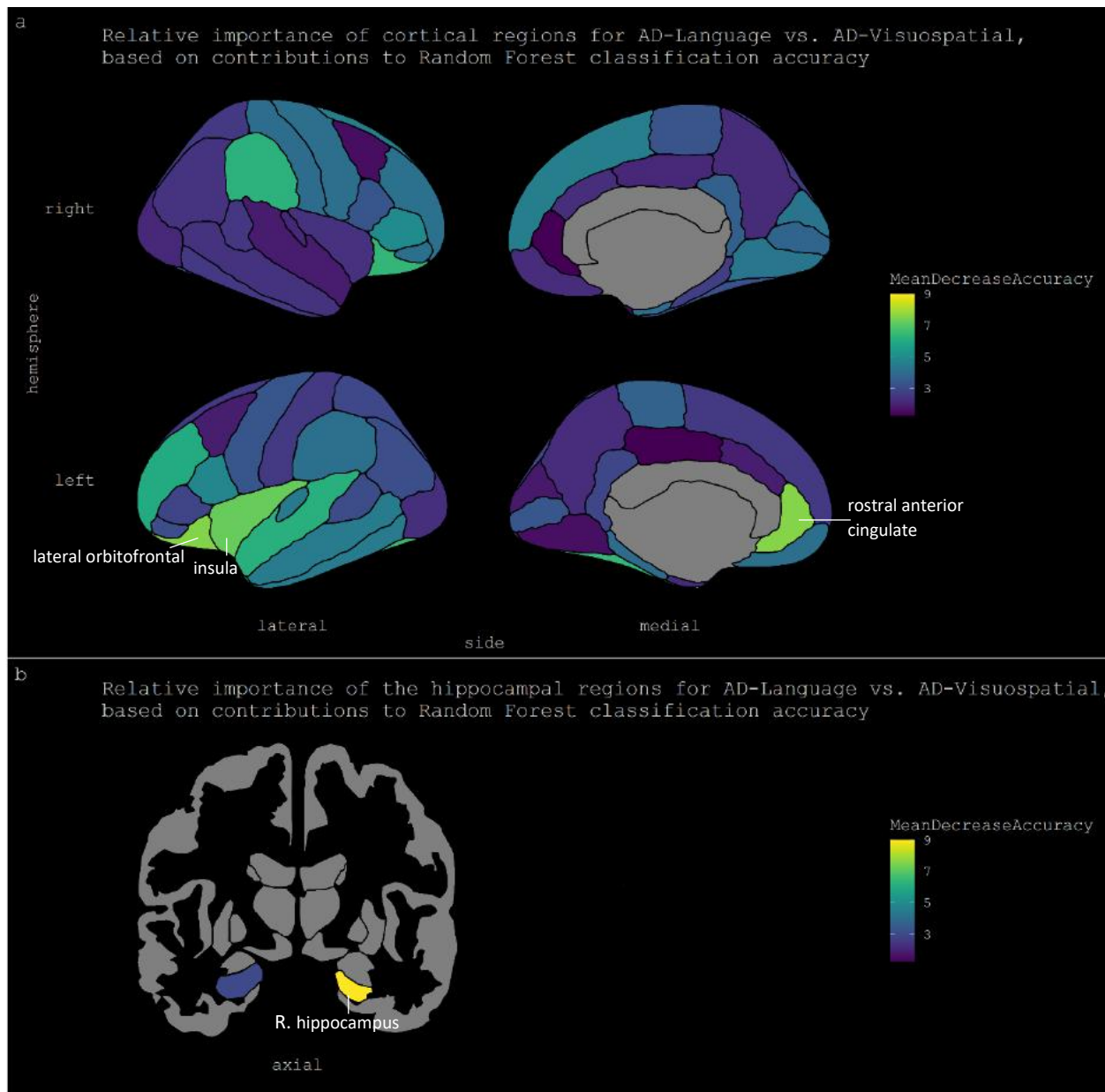


Figure 4.6 Relative importance of ROIs in the a.) Desikan-Killiany (DK) atlas and b.) the hippocampal regions, based on contributions to classification accuracy from the random forest models for AD-Language vs. AD-Visuospatial. Mean Decrease in Accuracy values for each ROI were averaged over the 100 random forest models (corresponding to 100 seeds). The top four regions of importance are labeled. However, the class imbalance in classification accuracies of the AD-Language compared to AD-Visuospatial do not make the current results reliable to assess ROI importance for distinguishing between these two groups.

4.4 Understanding how the ROIs identified as important differ across the AD subgroups

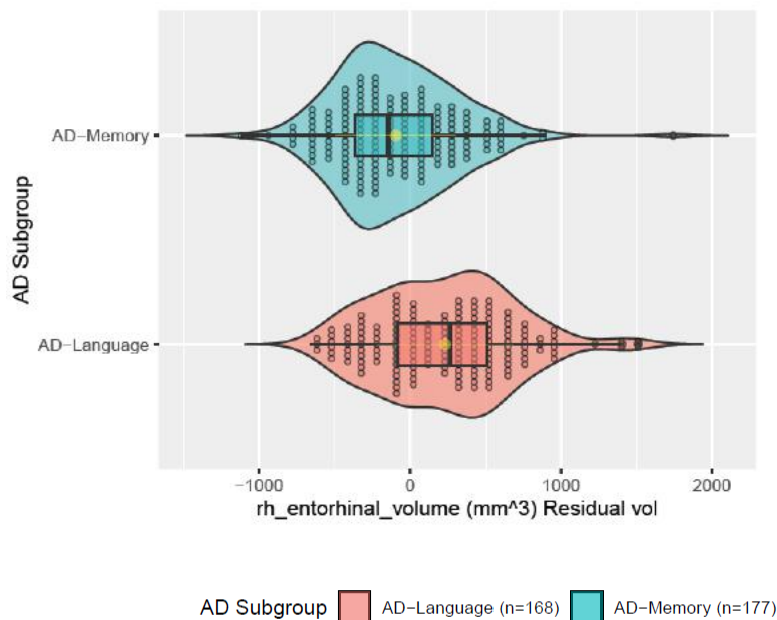
While random forest classification models can be used to identify the topmost important ROIs for distinguishing between the pairs of AD subgroups, they do not provide any information on how those ROIs differ across the AD subgroups being compared. To provide insight into this, I constructed violin plots for the AD subgroups of interest for the selected ROIs and visually compared the distributions and spread of the data in the two AD subgroups to understand which AD subgroup has smaller volumes in the selected ROIs. In addition to the violin plots in the visualizations in figures 4.7- 4.9, I've also shown the individual data points in each distribution as stacked points in each bin; a total of 30 bins were used. The relationships of the important ROIs with the AD subgroups in each of the binary classifications based on these plots are described below. I chose to focus on the top four ROIs for this discussion. The number of top ROIs to focus on is a subjective choice and would be best guided by recommendations from neuropathologists who may be interested in the results of this work. Violin plots comparison for all 70 ROIs are available in the supplementary files: [violin_boxPlots_L_M_volResiduals_SMOTE.pdf](#), [violin_boxPlots_M_VS_volResiduals_SMOTE.pdf](#) and [violin_boxPlots_L_VS_volResiduals_SMOTE.pdf](#).

a. AD-Language vs. AD-Memory

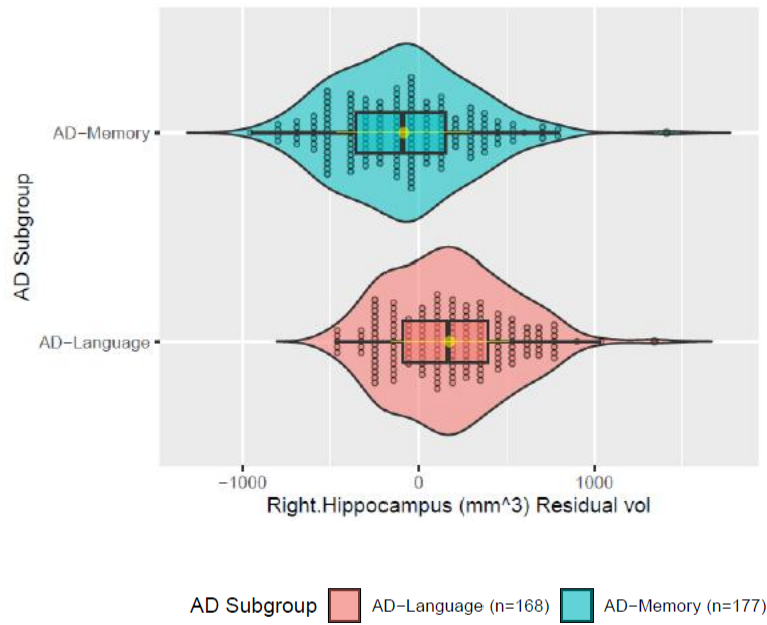
Variable importance results from random forest (described in the previous section) pointed at the following regions to be most important for distinguishing the AD-Language group from the AD-Memory group: **right entorhinal cortex, right hippocampus, right lingual gyrus, and left hippocampus**. Figure 4.7 shows violin plots for SMOTE oversampled residual volume data for the two subgroups for these selected ROIs. It can be seen from the plots that the AD-Memory group has lower volumes for right entorhinal cortex, right hippocampus and left hippocampus

than the AD-Language group. For right lingual gyrus, the distributions for the AD-Language and AD-Memory groups are very much overlapping, with boxplots centered roughly around the same mean value. It is unclear why the right lingual gyrus shows up as a variable with one of the higher variable importance values based on random forest classification, yet the distributions of the two AD subgroups do not show any visually noticeable differences. One possibility is that random forest is able to shed light on differences across the subgroups after taking non-linear relationships and interactions among the ROI variables into account, which may be not be apparent in simpler isolated comparisons of individual ROI volumes of subgroups. Another possibility is that there is bias in random forest variable importance results that is spuriously placing a higher importance on the right lingual gyrus volume. This possibility is further discussed in the next section (Section 4.5) and a plan is proposed to reduce potential bias in random forest variable importance results.

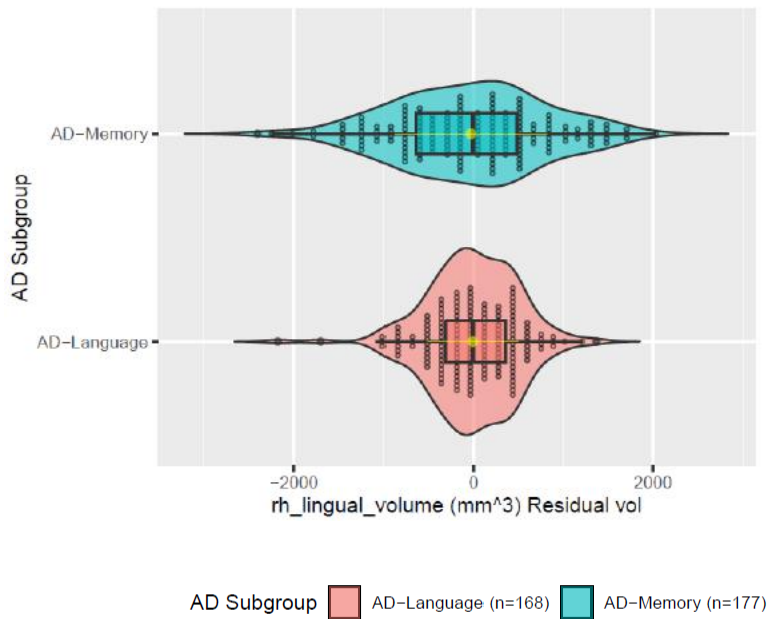
1. Right entorhinal cortex



2. Right hippocampus



3. Right lingual gyrus



4. Left hippocampus

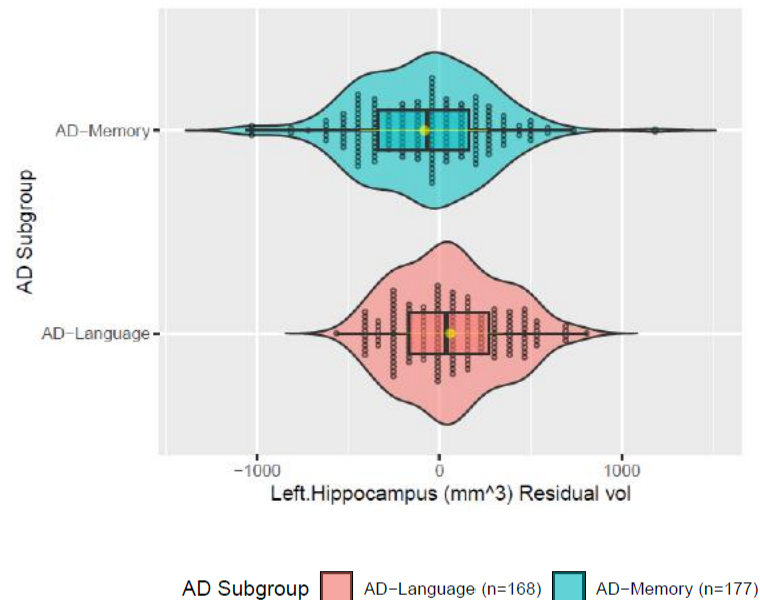


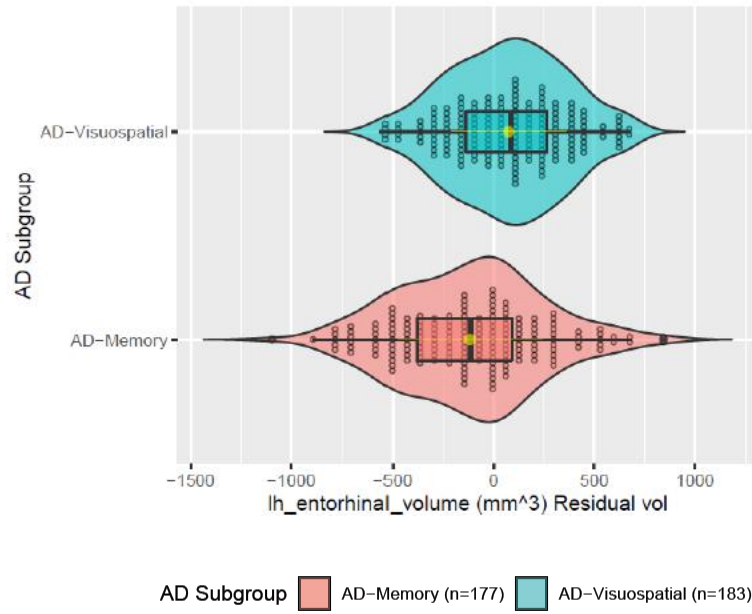
Figure 4.7: Violin distribution plots comparing AD-Language and AD-Memory groups for the top four brain ROIs identified through random forest variable importance results. The mean value for each distribution is shown by the yellow dot. Individual data points in the distribution are shown stacked at the corresponding volume bins. Number of bins = 30.

b. AD-Memory vs. AD-Visuospatial

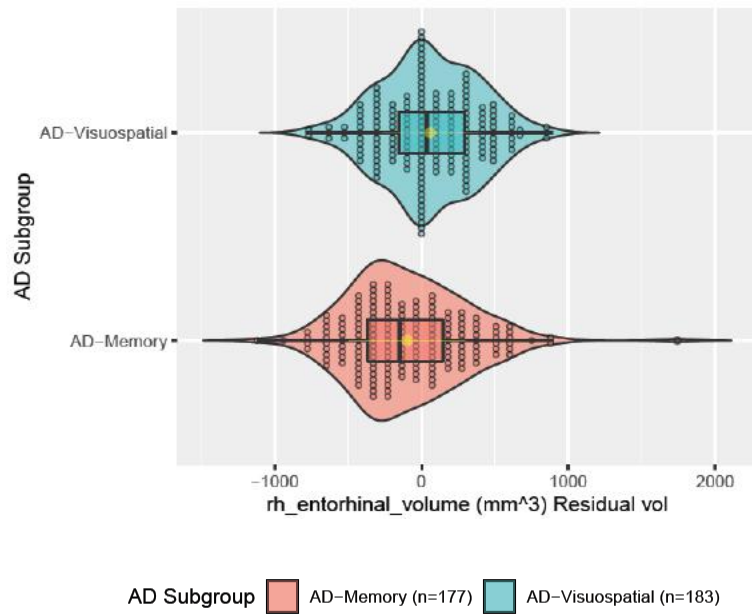
Comparing the violin distribution plots for AD-Memory and AD-Visuospatial groups for the top identified regions of importance (Figure 4.8): **left entorhinal cortex, right entorhinal cortex, right supramarginal gyrus and left postcentral gyrus**, one can see that in general, the distributions for the AD-Memory group are more spread out compared to the AD-Visuospatial group. This could be because of the smaller starting sample size of the AD-Visuospatial group (which may not be enough to represent the true population characteristics), which was then oversampled using SMOTE. It is worth noting that the AD-Language group had an even smaller sample size than the AD-Visuospatial group but the distribution spreads for AD-Language and AD-Memory were not as different in their variance as is seen in the case of AD-Memory vs. AD-

Visuospatial. Although the means of the distributions may not be that different in the two AD subgroups for some of the above noted regions, there are still some noticeable differences in the shapes of the distributions for the ROIs shown in Figure 4.8. For left entorhinal cortex, it can be seen that the AD-Memory group has a longer left tail than the AD-Language group, indicating a shift towards lower volumes for the AD-Memory group compared to the AD-Language group. Similarly, the right entorhinal cortex also shows a slight left shift for the AD-Memory group's distribution compared to the AD-Language group. The differences in the distributions across the two subgroups for the other two ROI volumes (right supramarginal gyrus and left postcentral gyrus) are less pronounced. For right supramarginal gyrus, the AD-Visuospatial group's distribution is slightly shifted to the left compared to the AD-Memory group's distribution, hence indicating a slightly greater proportion of smaller volumes in the AD-Visuospatial group. However, it is difficult to make a strong claim about this difference between the two groups. The ambiguity in AD subgroup differences in the volume of left postcentral gyrus based on violin plot comparisons of the AD-Memory and AD-Visuospatial groups is even greater and adds to the case of investigating potential biases in random forest variable importance results. This is a point of discussion in Section 4.5.

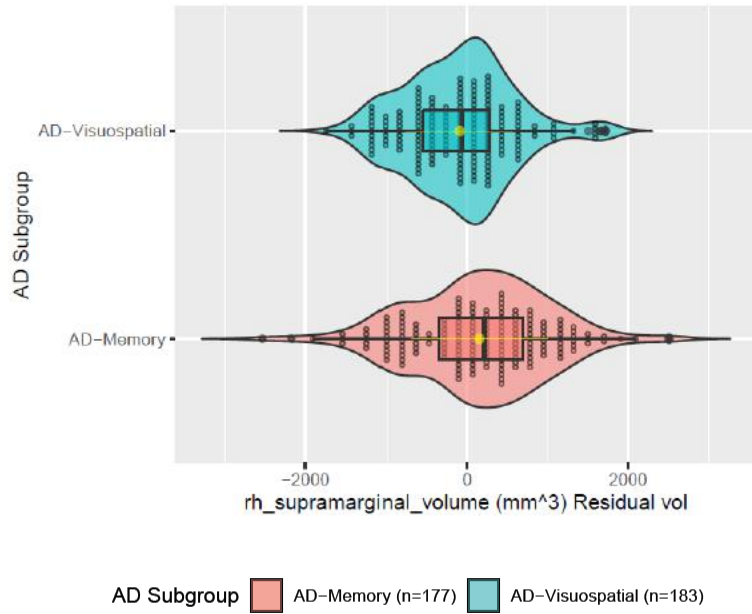
1. Left entorhinal cortex



2. Right entorhinal cortex



3. Right supramarginal gyrus



4. Left postcentral gyrus

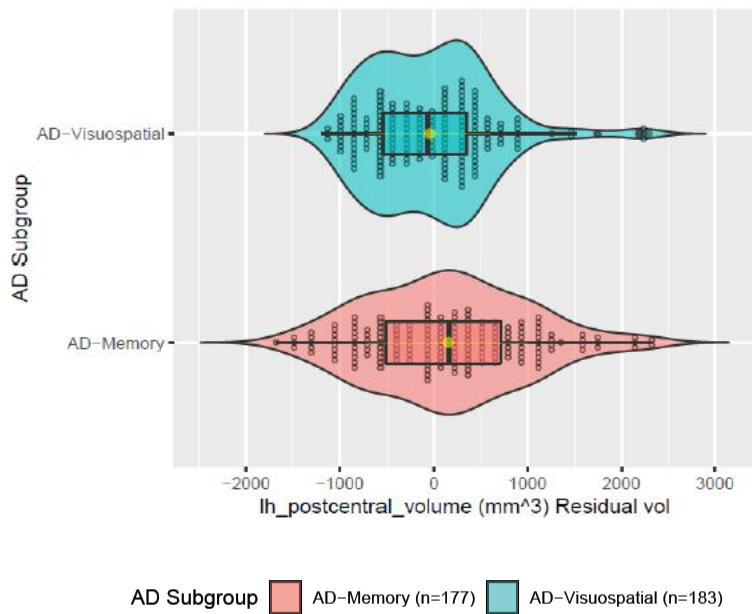


Figure 4.8: Violin distribution plots comparing AD-Memory and AD-Visuospatial groups for the top four brain ROIs identified through random forest variable importance results. The mean value for each distribution is shown by the yellow dot. Individual data points in the distribution are shown stacked at the corresponding volume bins. Number of bins = 30.

c. AD-Language vs. AD-Visuospatial

Given that the current classification model for AD-Language v. AD-Visuospatial should not be used for interpreting which ROIs are most important (due to reasons given in Section 4.3), it is currently unknown which ROIs should be examined for violin plot comparisons between the AD-Language and AD-Visuospatial groups. Hence, currently, the AD-Language vs. AD-Visuospatial analysis is not at the stage of comparing violin plots for understanding which of the two subgroups has more atrophy in the ROIs of interest.

4.5 Limitations of the current approach, alternative solutions & next steps

While majority of the results from variable importance based on random forest classification models make sense when violin plots for the AD subgroups of interest are compared for a given ROI (i.e. the violin plot distributions look different enough for the two AD subgroups being compared), some results need further investigation. As discussed in section 4.3, for the AD-Language vs. AD-Memory classification, right lingual gyrus is ranked to be the third most important brain region based on the mean decrease in accuracy measure. However, when the distributions of the AD-Language and AD-Memory groups for this region are compared using violin plots, the distributions are very much overlapping and don't show noticeable visual differences (see Figure 4.7). A similar ambiguity exists in AD subgroup differences among AD-Memory and AD-Visuospatial for the volumes of right supramarginal gyrus and left postcentral gyrus based on violin plot comparisons, yet these regions are in the top four important regions based on random forest models for AD-Memory vs. AD-Visuospatial. These results bring up two possibilities. One possibility is that random forest results for variable importance for these ROIs are indeed correct and random forest sheds light on differences in AD subgroups by accounting for non-linear relationships and interactions among the ROI variables, which may not be apparent

in isolated comparisons of individual ROI volume distributions for the subgroups. The other possibility is that there are potential biases in variable importance results from random forest which could be causing a spurious identification of certain ROIs as important.

Looking into the possibility of biases in random forest variable importance, I came across a study by Strobl et al. (*James et al. 2013, 312, 319, 330*) that investigated this using simulated data. The simulation studies by Strobl et al. illustrated that when variables with different data types are used, the variable importance results from random forest can be misleading. This was not the case in the data that I analyzed since all variables considered for random forest model building were ROI variables representing residual volumes. However, the phenomenon of “different data types” may also encompass the case where variables are of the same data type but with very different scales. Strobl et al. showed that random forest variable importance may not be reliable in situations where the variables being considered for importance vary in their scale of measurement or their number of categories.

In the case of the analysis that I did using random forest, I did not have any categorical variables as all variables were ROI volumes, so the issue of having greatly different number of categories for variables does not apply to the data that I analyzed. However, within the continuous variables, it is possible to have variables that have more “categories” or values than some other variables that may have a less continuous set of values. So, while I didn’t have any categorical variables for random forest model building, it is possible that the number of unique values (“categories”) in each continuous variable affects which variable is more likely to be chosen at a tree node. The idea behind this finding is that random forest will tend to favor variables that are more continuous as that provides more opportunities to partition the data.

Regarding the variation in the scale of measurement, although all variables in my analysis represented residual volumes for different ROIs, it is possible that the scales for different ROIs were different due to the differences in the original sizes (and hence volumes) of the ROIs, in which case this could be a potential point causing bias in random forest variable importance in my results. Unlike logistic regression, Random forest does not need a scaling or standardization of input variables as the method is not distance based [cite], i.e. the data partitioning at each node does not depend on the actual values of the observations for the given variable. Hence, in my implementation of random forest models, I had not scaled or standardized the input variables. However, Strobl et al.'s work suggests the importance of having input variables on very similar scales when it comes to obtaining variable importance without bias.

The reasons for bias in random forest variable importance, as noted by Strobl et al., are attributed to two factors: 1.) biased variable selection in the classification trees due to the points discussed above (input variables with different data types or different scales and different number of categories), and 2. bias induced by bootstrapping (random sampling with replacement) in random forest. Strobl et al. propose an alternative to random forest: a method known as conditional inference forest which generates random forests from unbiased classification trees based on a conditional inference framework (Strobl et al. 2008). The simulation design used by Strobl et al. represented a scenario of a binary variable to be predicted from a set of variables that vary in their scale of measurement and number of categories. The first variable X1 was continuous while the other variables X2,...X5 were categorical with number of categories ranging from 2 to 20. The sample size was set to 120 which is comparable to the sample size of the AD subgroups I used for in random forest model building. Although this simulation study considers an extreme scenario of the input variables representing different types of data with different scales of measurements and

categories which is unlike the ROI residual volume data that I analyzed using random forest, the conclusions drawn from the results of this simulation study are still worth paying attention to for designing future work for reducing bias in variable importance. Strope et al.'s simulation studies showed that for the randomForest function, all three variable importance measures are unreliable for the simulated data: the Gini index, selection frequency and permutation importance; the Gini index was the most unreliable. For conditional inference forest, the authors used the cforest function in R and found that reliable results of variable importance can be obtained both with the selection frequency and the permutation importance if cforest is used with subsampling without replacement. The cforest function does not have the option of Gini index as a variable importance measure. Using subsampling with replacement in cforest also gives biased results.

Hence, a solution to address potential bias in random forest variable importance in my current work is to use the cforest function in R, with subsampling without replacement. Before this implementation, however, a careful understanding of the conditional inference forest method is necessary to assess whether it is appropriate for the data at hand. Conditional inference trees like random forest trees are based on recursive partitioning. The additional restriction in conditional inference forests is that the variable to be used for splitting the data at each node and the split points are determined based on statistical significance tests. Mingers (*Mingers 1987*) notes that a recursive partitioning algorithm that chooses the split variable simply based on the contribution to the information measure such as the Gini index or contributions to prediction accuracy such as in random forest “has no concept of statistical significance, and so cannot distinguish between a significant and an insignificant improvement in the information measure.” Conditional inference trees are an attempt to provide “a statistical approach [to recursive partitioning] which takes into account the distributional properties of measures.” (*Hothorn, Hornik, and Zeileis, n.d.*)

An unrelated limitation based on the current random forest results is that it did not result in a satisfactory classification model performance for AD-Language vs. AD-Visuospatial as reflected by the disproportionate OOB errors for the two classes. The AD-Visuospatial group had a much higher average OOB error than the AD-Language group. Because of this, the models for this classification could not be used to say anything about which ROIs are important for this particular comparison of AD subgroups. The classification model from penalized logistic regression had a poorer overall classification performance (5-fold CV misclassification error of 30.37%) than the average overall from the current random forest models. Hence, current results do not warrant the use of either the penalized logistic regression model or the random forest models to reliably interpret ROI importance for the classification. For future work, it would be interesting to see if using conditional inference forests with subsampling without replacement may yield a better classification performance for AD-Language vs. AD-Visuospatial, with balanced OOB errors for the two classes. Other methods may also need to be explored to see if these two groups can be separated well using ROI volume data or if they are actually not that separable.

Within the random forest or cforest framework, for future work, parameters such as the number of trees, the number of variables to use at the nodes for splitting the data, and the sample size for subsampling (portion of the data to be used for each tree building) can be fine-tuned using cross-validation.

4.6 Conclusions

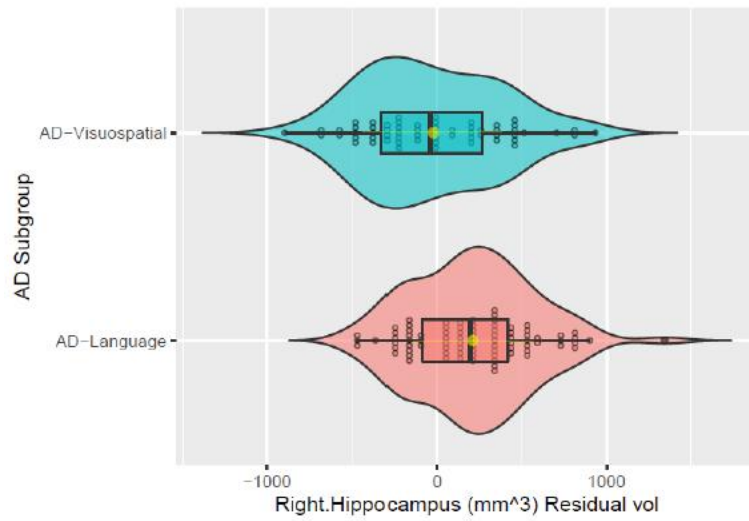
The work described in this chapter is an application of random forest classification models in the domain of AD subgroups. This is a novel way of shedding light on which brain ROIs are most important to distinguish between pairs of AD subgroups. This approach allows for the accounting of potential non-linear relationships between ROIs and AD subgroups and interactions of ROI

volumes, which to my knowledge have not been considered in previous model building for analyzing AD subgroup differences. By determining variable importance based on contributions to classification accuracy based on 100 random forest models (100 seeds) for each binary classification (AD-Language vs. AD-Memory, AD-Memory vs. AD-Visuospatial and AD-Language vs. AD-Visuospatial), I successfully identified some ROIs as candidates that could be explored in future neuropathological studies for understanding AD subgroup differences. Current results yielded decent classification performance and hence reliable variable importance assessment for AD-Language vs. AD-Memory and for AD-Memory vs. AD-Visuospatial. As noted in section 4.5, future work will involve developing better models to understand the ROI differences between AD-Language vs. AD-Visuospatial as well as implementing condition inference forest to address potential biases in random forest variable importance. By averaging the variable importance results and classification accuracies over 100 models based on 100 unique seed specifications in R, I hoped to have introduced more robustness in the results. Next, by making visual comparisons of ROI volume distributions of the subgroups being compared I was also able to provide a sense of which AD subgroup is expected to have a greater atrophy in the identified ROIs. The brain regions identified through the analysis here may be those that would be most fruitful for targeting with subsequent neuropathological evaluations to identify differential microscopic pathology underlying these AD subgroup differences in atrophy discernible with MRI scans.

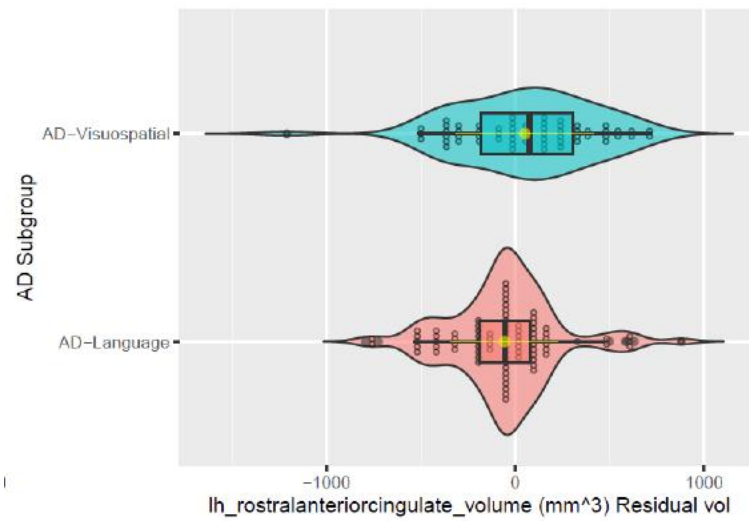
Appendix for Chapter 4

1. Violin plots for AD-Language vs. AD-Visuospatial

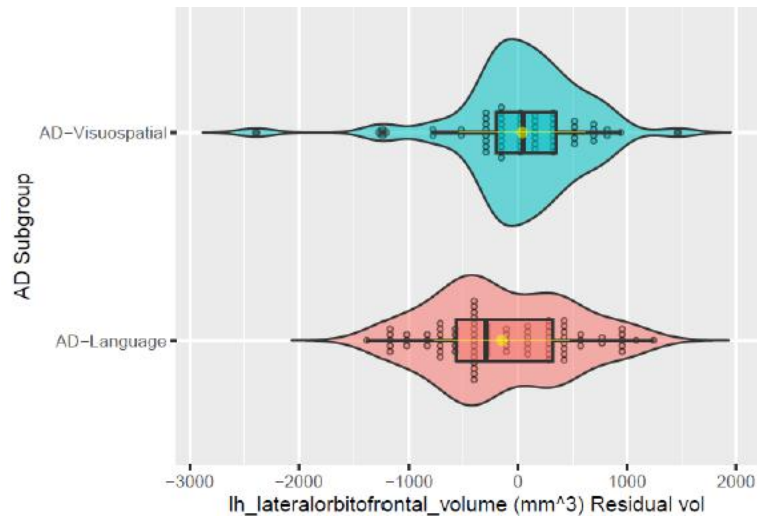
1. Right hippocampus



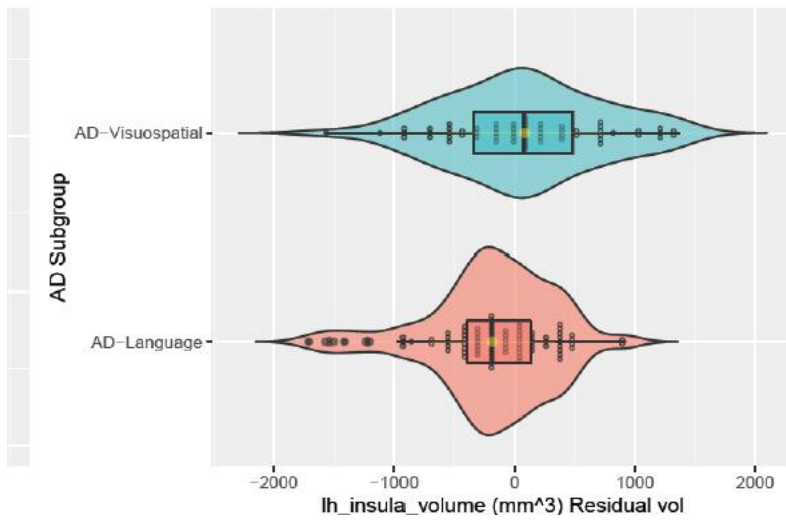
2. Left Rostral anterior cingulate gyrus



3. Left lateral orbitofrontal gyrus



4. Left insula



Aim 2: Longitudinal Data Analysis

Do cognitively defined AD subgroups from the time of AD diagnosis differ in brain ROI volume trajectories over time? If so, in which brain regions?

Methods for a meaningful analysis of noisy, small sample sized, class-imbalanced and irregular longitudinal brain ROI volume data for AD subgroups

Chapter 5: Overview of Longitudinal data analysis [Aim 2]: the scientific question, data challenges and workflow

In Aim 1 of my dissertation, the focus was on understanding which brain regions are most important for distinguishing between the AD subgroups, based on ROI volume data from 70 brain regions from the time of AD diagnosis. A natural extension of this question/investigation is to ask whether there are ROIs for which the AD subgroups differ the most in their volume trajectories over time. This led to the question for the second part of my dissertation [Aim 2]: Do cognitively defined AD subgroups from the time of diagnosis show differences in their ROI volume trajectories over time? If so, which ROIs show the most substantial differences between pairs of AD subgroups? Specifically, I focused on understanding how two aspects of each ROI's average volume trajectory differ between pairs AD subgroups: rate of change of ROI volume over time (slope) and the ROI volume at $t=0$ (intercept). I used linear mixed effects modeling to understand these differences, which is the topic of discussion of Chapter 6. Figure 5.1 provides an overarching overview of the problem tackled in longitudinal data analysis, including a description of the data, the analysis method used and the expected scientific finding from the analysis.

The workflow for Aim 2 work has two parts. The first part is focused on pre-processing of the data and the second part is the linear mixed effects modeling analysis. The latter is the topic of Chapter 6. In this chapter, in the next sections, I provide details of the longitudinal dataset, and the characteristics of data that led me to carry out specific pre-processing steps. I conclude the

chapter with a schematic of the full workflow for Aim 2 work, which connects the contents of this chapter and the next.

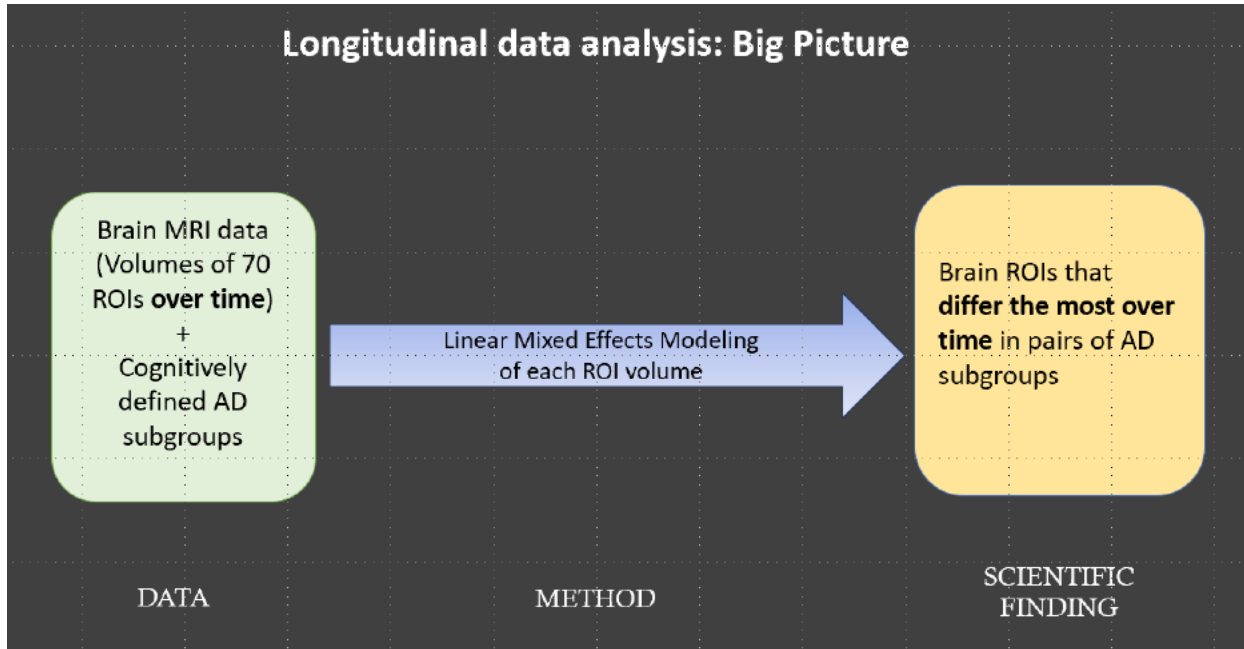


Figure 5.1: Overview of Longitudinal data analysis

5.1 Descriptive analysis of longitudinal MRI ADNI data: challenges of the data

Before diving into the details of analysis of longitudinal data using linear mixed effects modeling (Chapter 6), it is important to understand the characteristics of the current longitudinal dataset. Many of the pre-processing steps in the longitudinal data analysis workflow were driven by the nature of the current data in an attempt to reduce the effects of noise in the data.

I obtained the longitudinal dataset from Dr. Risacher at Indiana University, who processed the raw structural MRI data from ADNI using the same pipeline that she had used for the cross-sectional dataset. Note that the cross-sectional dataset that I used in Aim 1 work was not obtained by taking a subset of current longitudinal dataset corresponding to the AD diagnosis visit. Hence, the individuals in the longitudinal dataset (individuals with data at the time of AD

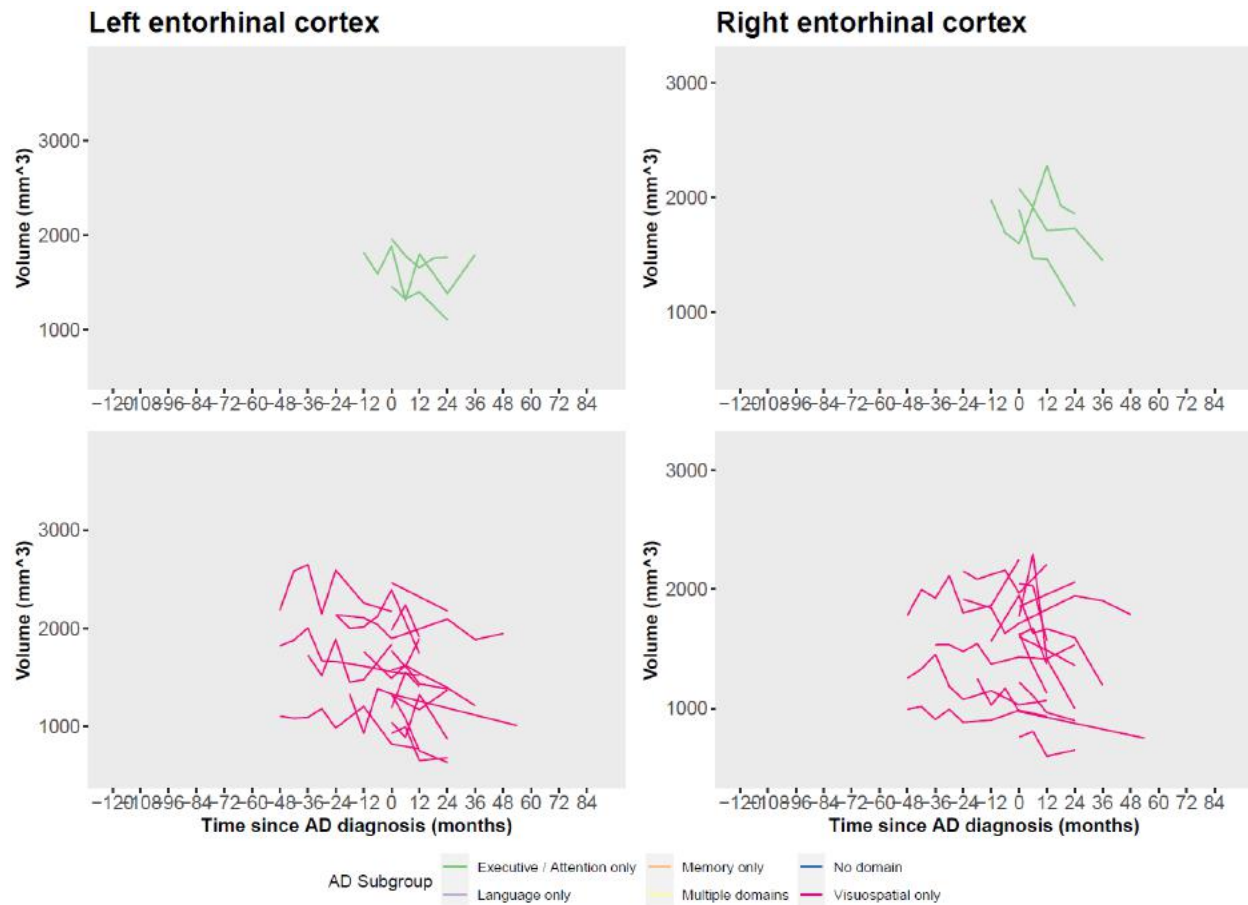
diagnosis visit and at least one other visit) are not simply a subset of the cross-sectional dataset (individuals with data at the time of AD diagnosis.) Although there is a large overlap of individuals in the cross-sectional and longitudinal datasets that I used, there are some additional individuals in the longitudinal dataset that were not present in the cross-sectional dataset, and vice versa. The variables in the longitudinal dataset were the same as cross-sectional dataset: volumes of 70 ROIs that are the focus of my dissertation work, other volume variables such as ICV and total gray volume, and non-ROI variables (magnetic field strength of the scanner, gender, age, years of education and *APOE* genotype status.) *W*-scores were not available for all individuals (about 50 individuals) in the longitudinal dataset. Hence, I used an alternative measure derived from the ICV and total gray volume to represent how progressed an individual is in terms of volume loss over the entire brain. The details of this measure are discussed in Chapter 6 when the linear mixed effects models are described. The longitudinal dataset from Dr. Risacher was re-formatted by Phoebe Scollard, a member of Dr. Crane's research group. In the original dataset from Dr. Risacher, data from ROI variables from every visit for an individual were in a single row. Phoebe Scollard re-formatted the data so that each row in the dataset represents an individual's data from a single scan from a single visit. This re-formatted dataset is what I used in my Aim 2 work.

By definition, longitudinal data includes data from one or more time points. Here, time points refer to visits in the ADNI study for an individual, which are typically every six months. The longitudinal dataset is more complex than the cross-sectional dataset due to a few reasons. First, due to changing technologies in different phases of the ADNI study over time, a given individual may have MRI scans conducted on scanners with different magnetic field strengths across different visits in the study. In some cases, an individual may have data from both 1.5T and 3T

scanners for a given visit. Further, from a given field strength scanner, most individuals had more than one scan per visit. Most likely, these multiple scans per visit exist to make sure there was at least one scan of reasonable quality available out of the few that are collected. These points present some choices to be made about which scans should be used for a given individual. Some of these choices contributed to the pre-processing steps described in Figure 5.4, which are discussed in more detail later. Additionally, in the current dataset, not all individuals had data available according to the visits' scheduled timeline of the ADNI study, i.e. the longitudinal dataset is irregular.

As a preliminary step in exploring the dataset, to describe each individual's disease timeline relative to a common point across all individuals, I anchored all individuals' data in time relative to each individual's time of AD diagnosis. The visit of AD diagnosis was defined to be $t=0$. All other visits were defined relative to the AD diagnosis visit. For example, a visit 6 months before the AD diagnosis visit corresponds to $t=-6$ months while a visit 2 years after the AD diagnosis visit corresponds to $t=24$ months in the transformed dataset. Below, I've presented volume trajectories for four ROIs using data corresponding to the 1st scan from each visit for individuals from all four AD subgroups, as examples, to illustrate some of the characteristics of the data (Figures 5.2 and 5.3). Generally, one would expect that ROI volumes decrease over time or more or less stay constant in an older sample of individuals such as the one represented in this dataset. For the first two ROIs (Figures 5.2a and 5.2b), left entorhinal cortex and right entorhinal cortex, one can see that the volumes in many trajectories fluctuate over time, although a general downward trend can be seen in most cases. Figures 5.3a and 5.3b show the volume trajectories for left frontal pole and right frontal pole. The trajectories for these ROIs seem to fluctuate even more compared to trajectories for left entorhinal cortex and right entorhinal cortex, and it is

difficult to see a general decreasing trend for a large number of trajectories. Figures 5.2 and 5.3 demonstrate that there is noise in ROI volume measurements over time in the current longitudinal dataset. For some ROIs, there may be more noise than others. The plots also demonstrate that the amount of noise is roughly the same for 1.5T and 3T data. Other characteristics of the data that can be seen from Figures 5.2 and 5.3 are imbalanced class (AD subgroup) sizes, small sample sizes and irregular data.



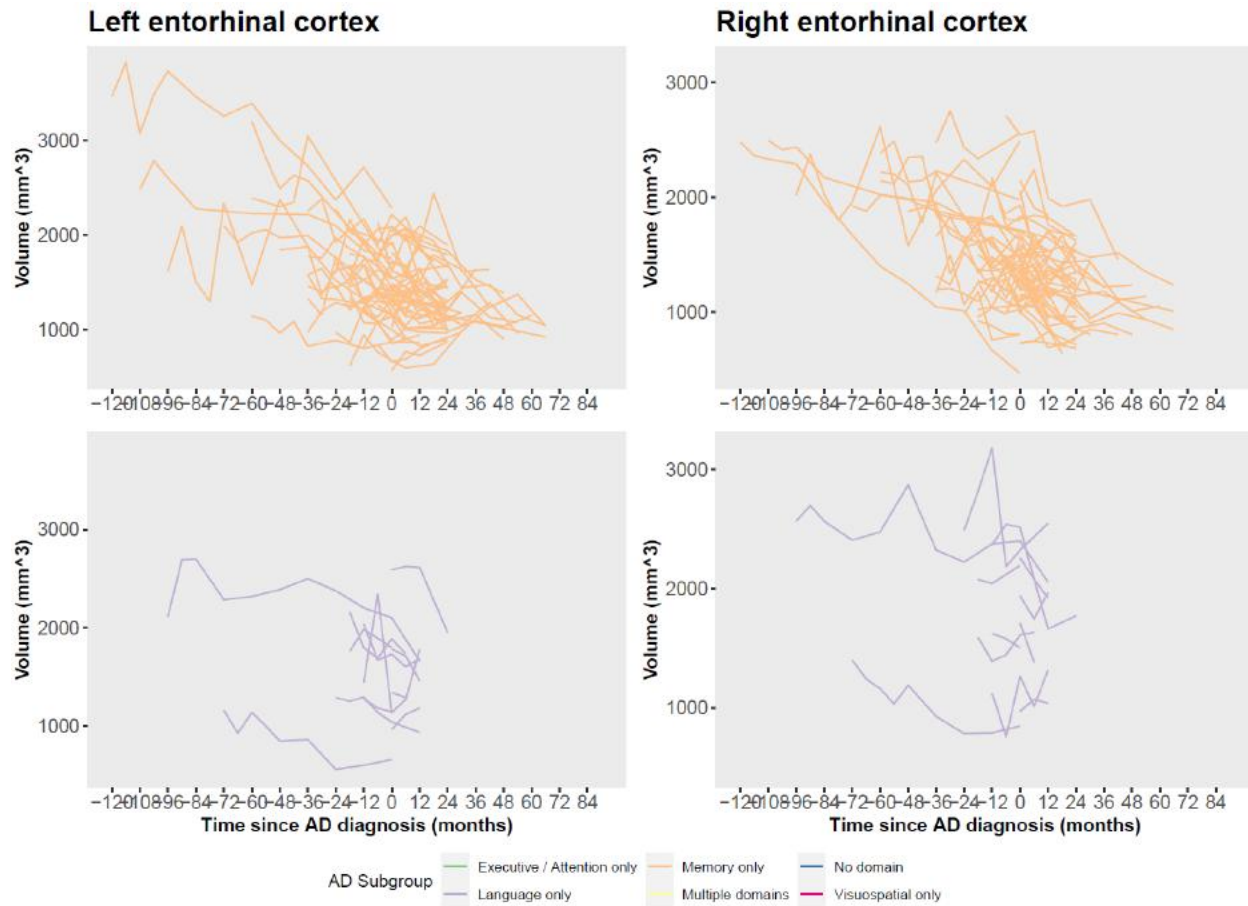
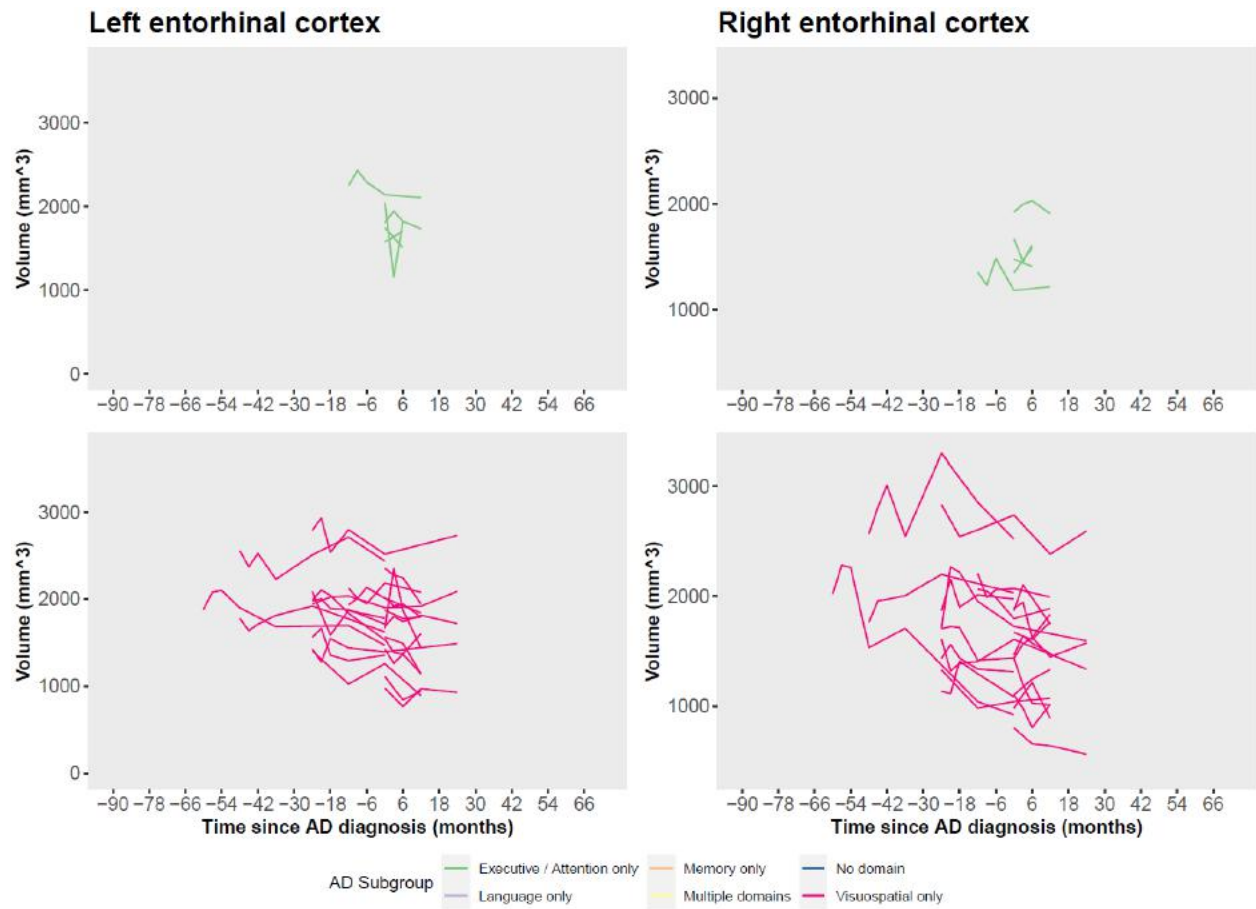


Figure 5.2a: Volume trajectories for left entorhinal cortex and right entorhinal cortex for individuals in all four AD subgroups. Data is from 1st scans from all 1.5T visits in the longitudinal dataset.



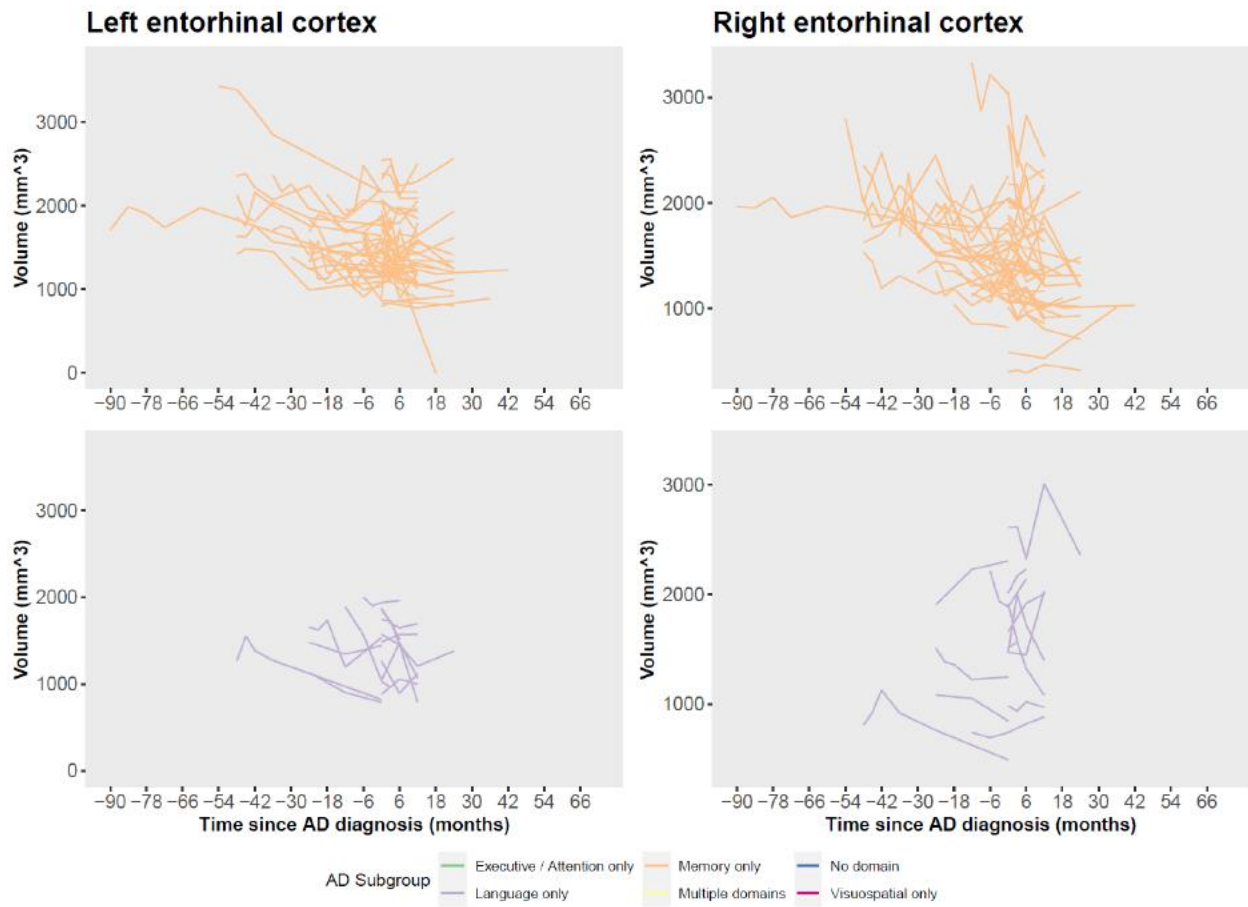
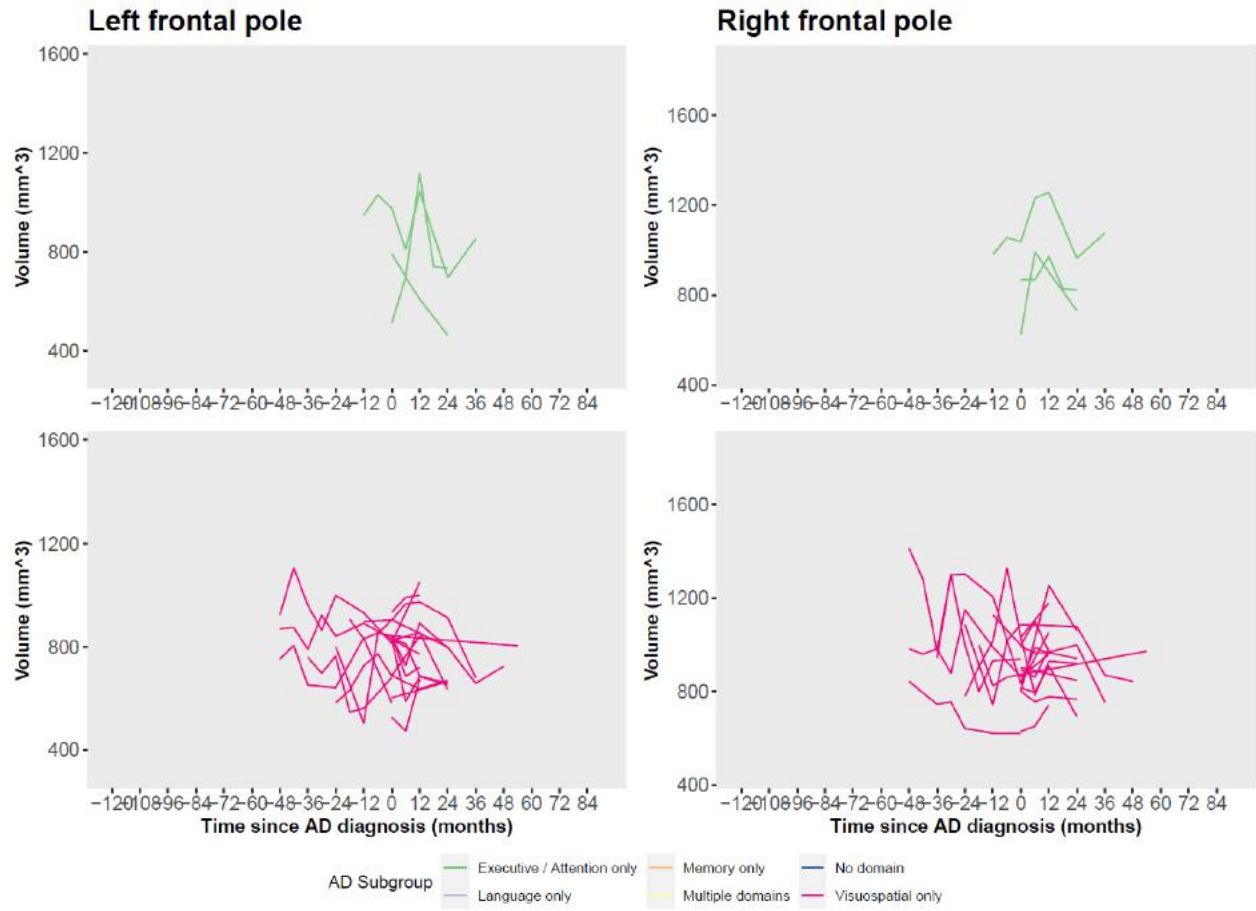


Figure 5.2b: Volume trajectories for left entorhinal cortex and right entorhinal cortex for individuals in all four AD subgroups. Data is from 1st scans from all 3T visits in the longitudinal dataset.



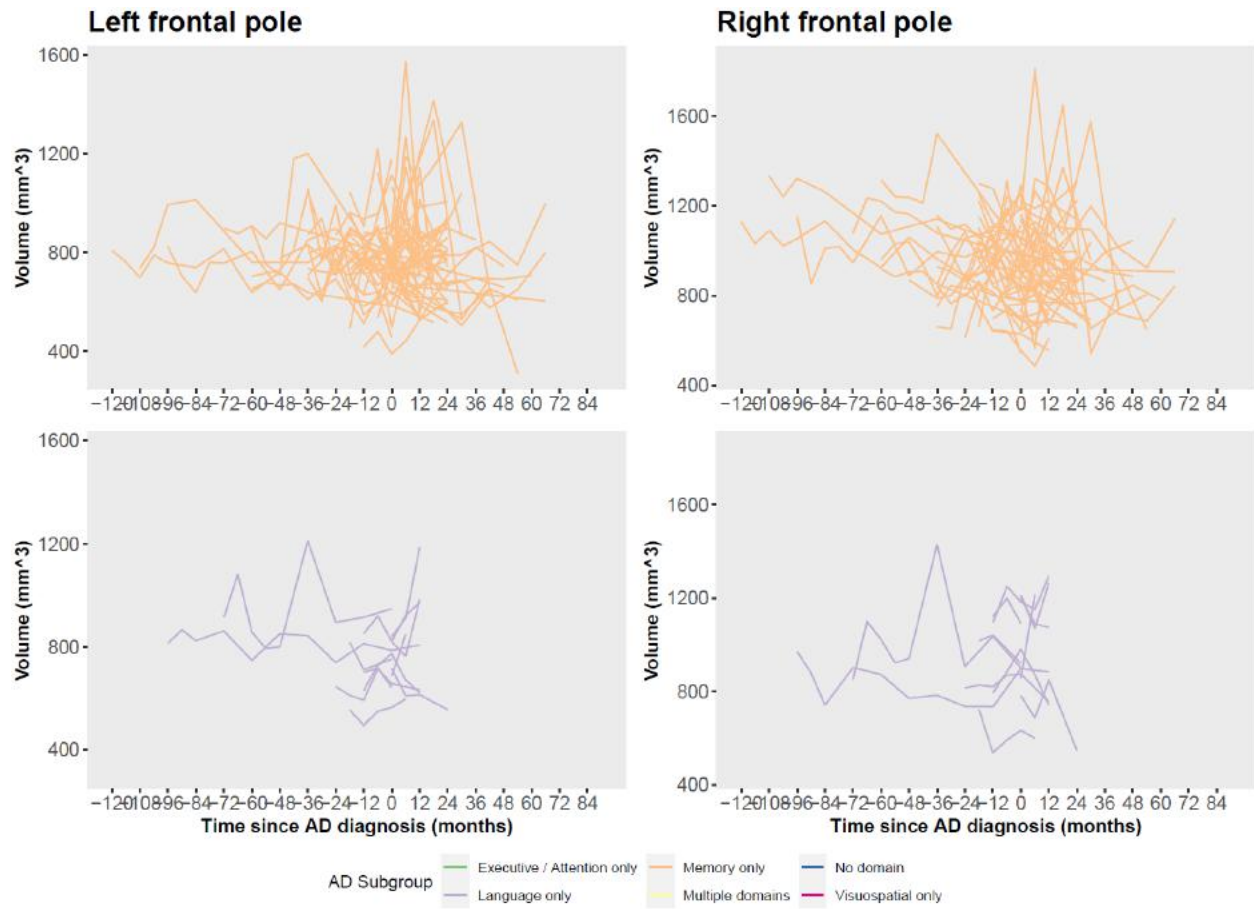
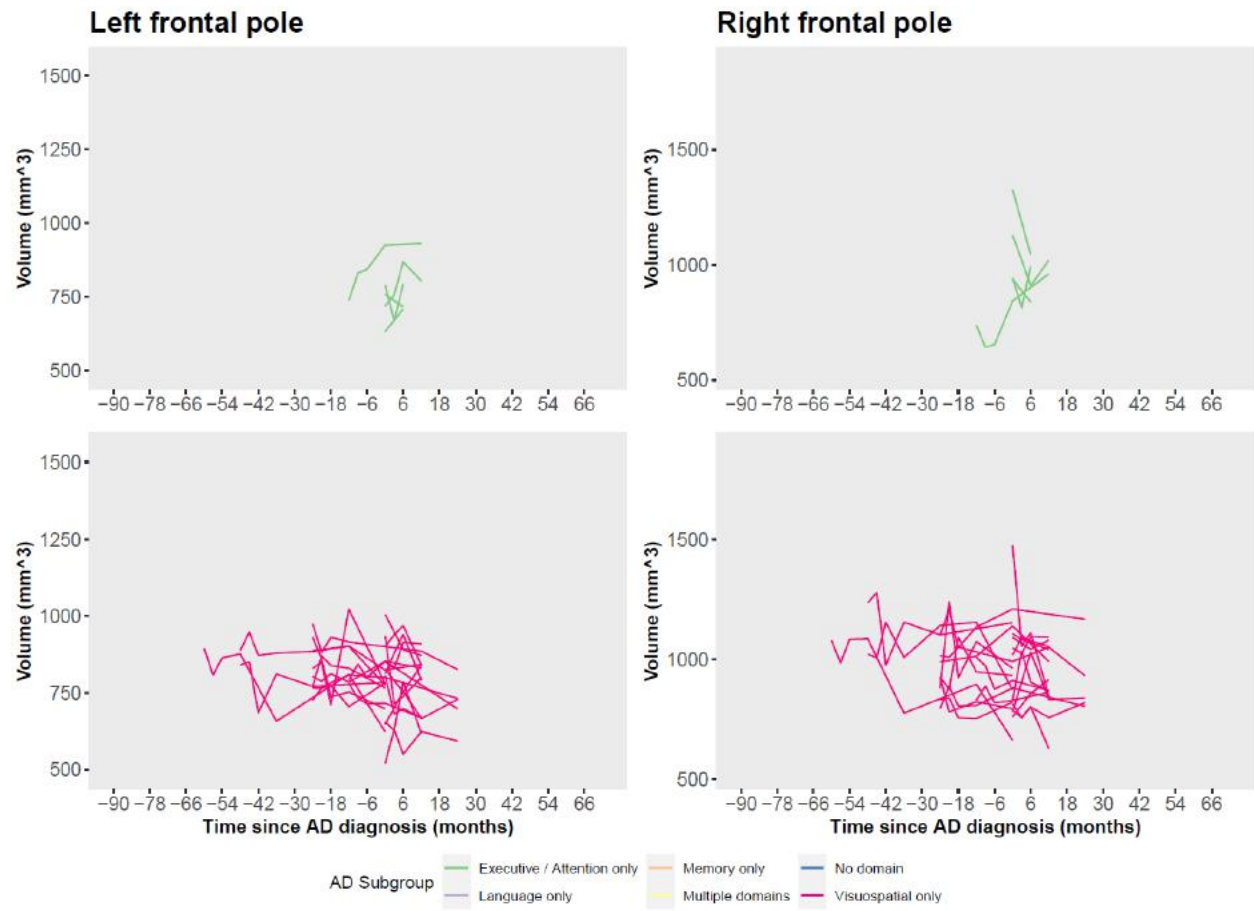


Figure 5.3a: Volume trajectories for left frontal pole and right frontal pole for individuals in all four AD subgroups. Data is from 1st scans from all 1.5T visits in the longitudinal dataset.



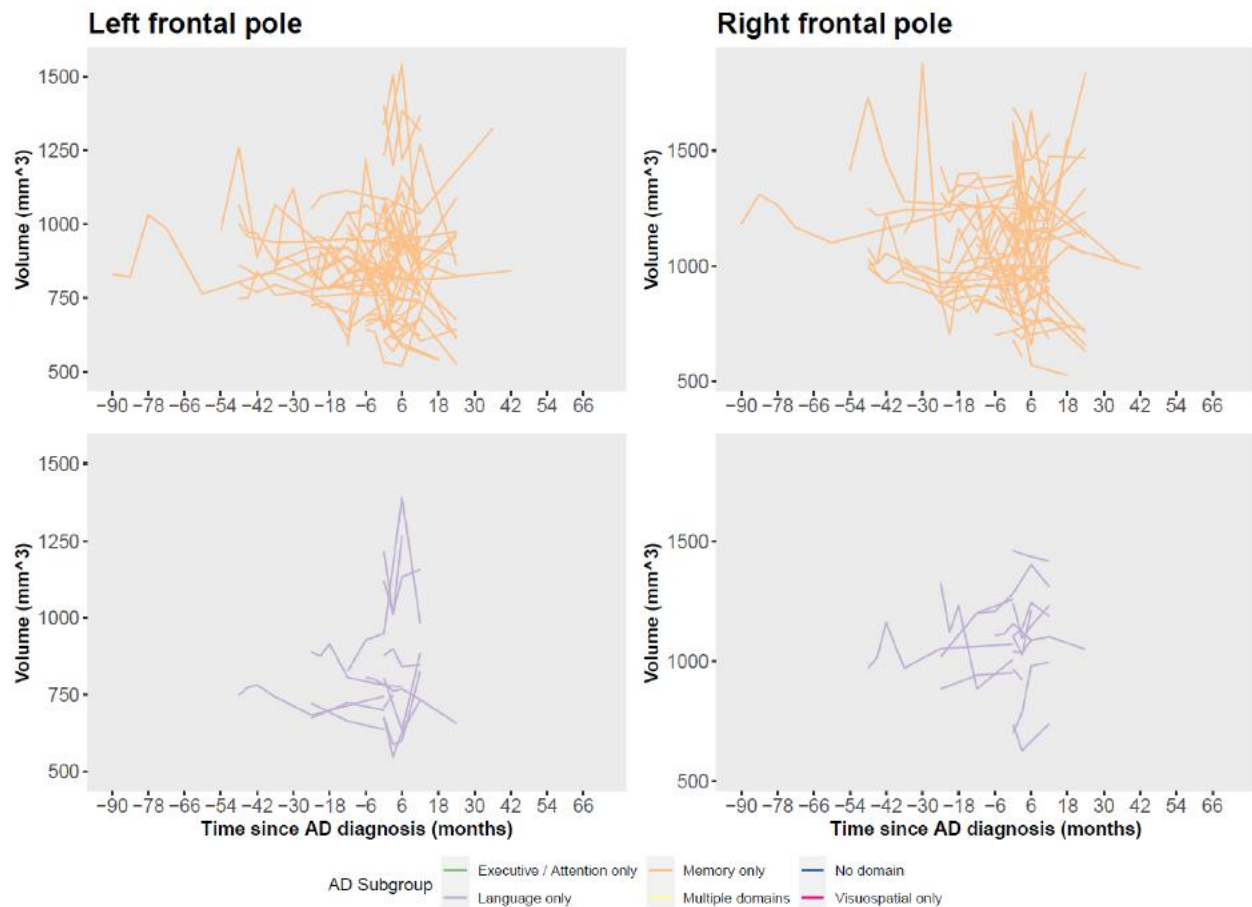


Figure 5.3b: Volume trajectories for left frontal pole and right frontal pole for individuals in all four AD subgroups. Data is from 1st scans from all 3T visits in the longitudinal dataset.

In the first part of workflow for longitudinal data analysis (outlined in Figure 5.4 below), some of the pre-processing steps that I carried out (Steps 2 and 3) were an attempt to reduce the effects of noise in the longitudinal dataset. The first step was a data filtering step where I made sure that I was only considering individuals who have data present at the time of AD diagnosis and at least one other visit. Next, I used the magnetic field strength from the visit of AD diagnosis for each individual as the field strength of choice for all visits for that individual. For individuals who had data for both field strengths at the visit of AD diagnosis, I used the field strength that would result in a larger number of visits' data to be used. By restricting to data from only one type of

field strength for all visits for an individual, the noise within a trajectory in ROI volumes due to different field strengths should be removed. In the last step of pre-processing workflow, I further sought to reduce the noise in ROI volume trajectories by choosing a higher quality scan for each visit for which multiple scans were available. Specifically, I chose data corresponding to the scan with the highest contrast to noise ratio (CNR) from FreeSurfer processing, instead of the first scan by default.

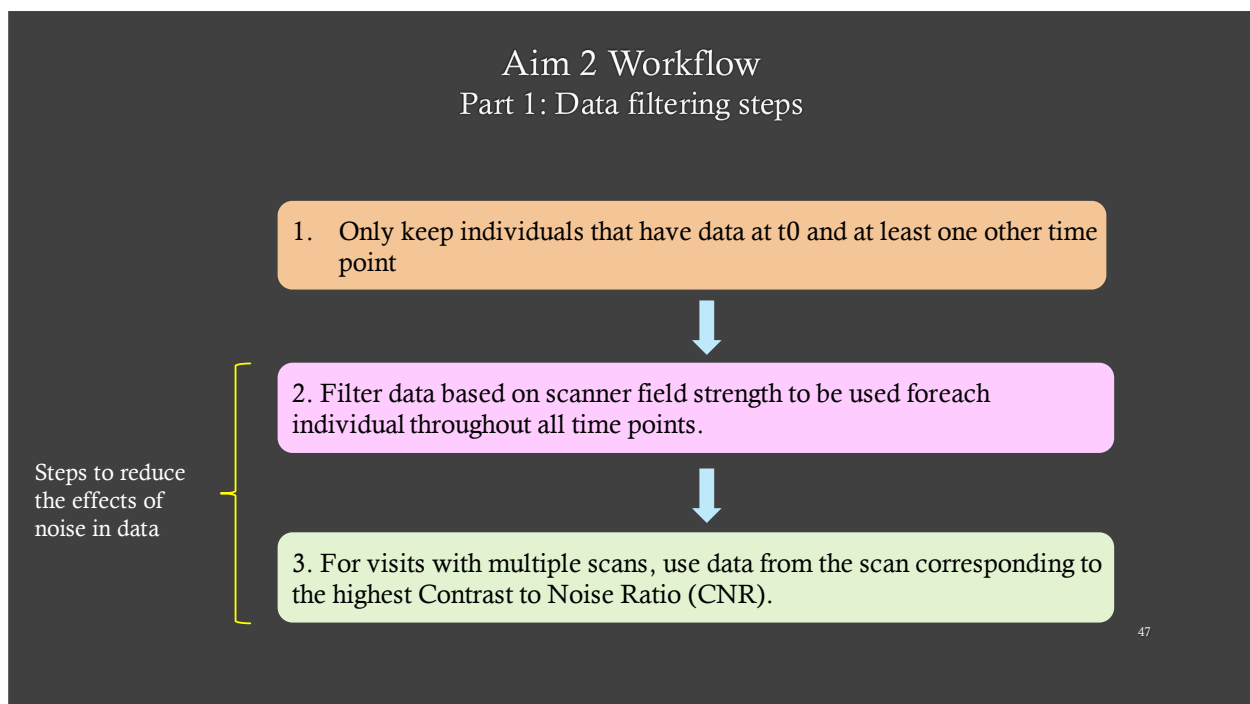


Figure 5.4: Part 1 of Longitudinal data analysis workflow: Data filtering and pre-processing

Figures 5.5a and 5.5b show what the data availability looked like for each time point after steps 1 and 2 of workflow. The final sample sizes from each AD subgroup are summarized in Table 5.1. In the analysis method used for longitudinal data, linear mixed effects modeling, which is the topic of next chapter, data from both field strengths were used together.

AD Subgroup	1.5 T	3T	Total
AD-Executive	7	12	19
AD-Language	21	21	42
AD-Visuospatial	26	38	64
AD-Memory	97	93	190

Table 5.1: Sample sizes for AD subgroups in the longitudinal dataset

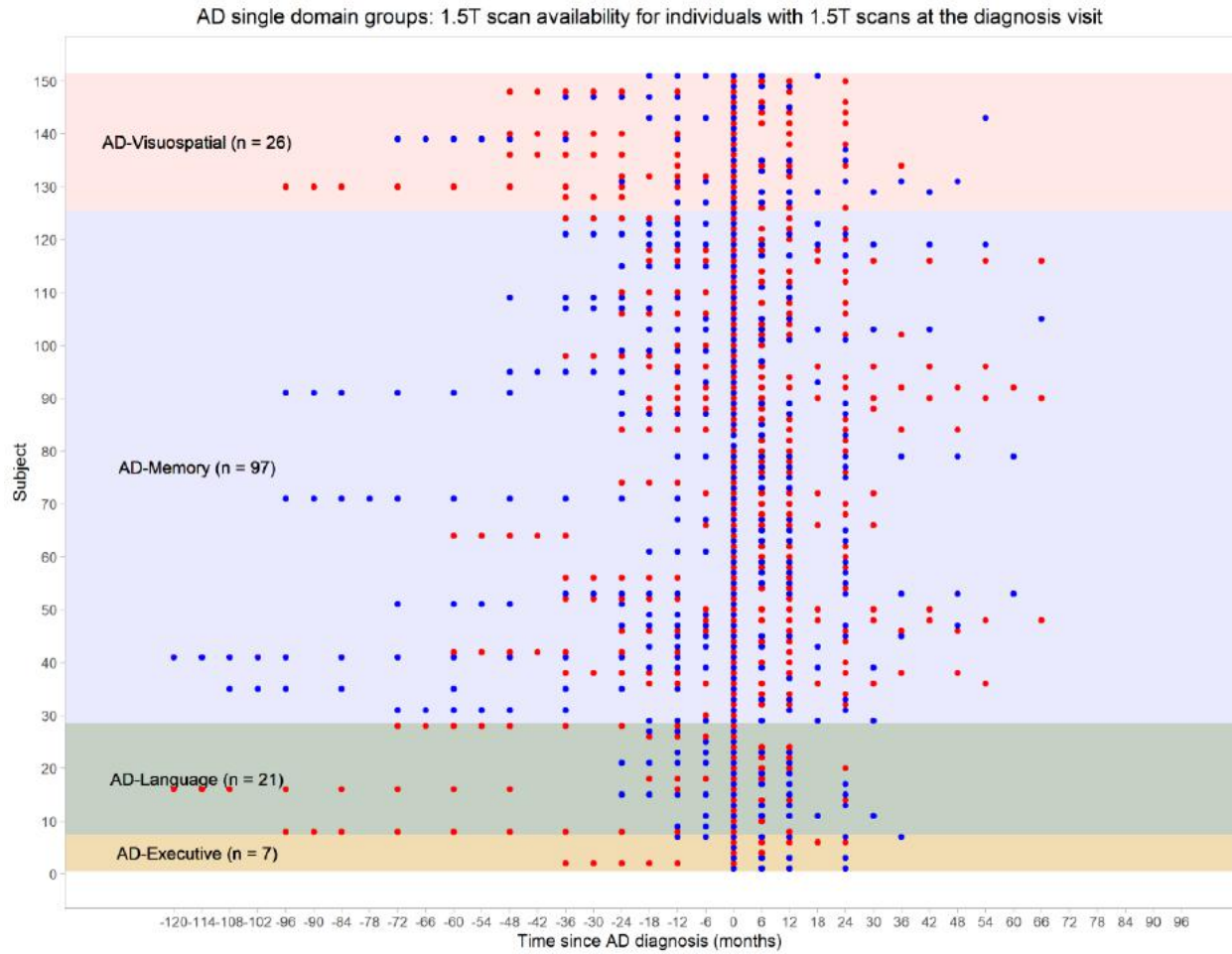


Figure 5.5a: A visualization of data availability at each time point for individuals with 1.5T scans at the time of AD diagnosis. Each row of dots represents an individual's data over time. Red and blue colors are used to distinguish between consecutive individuals.

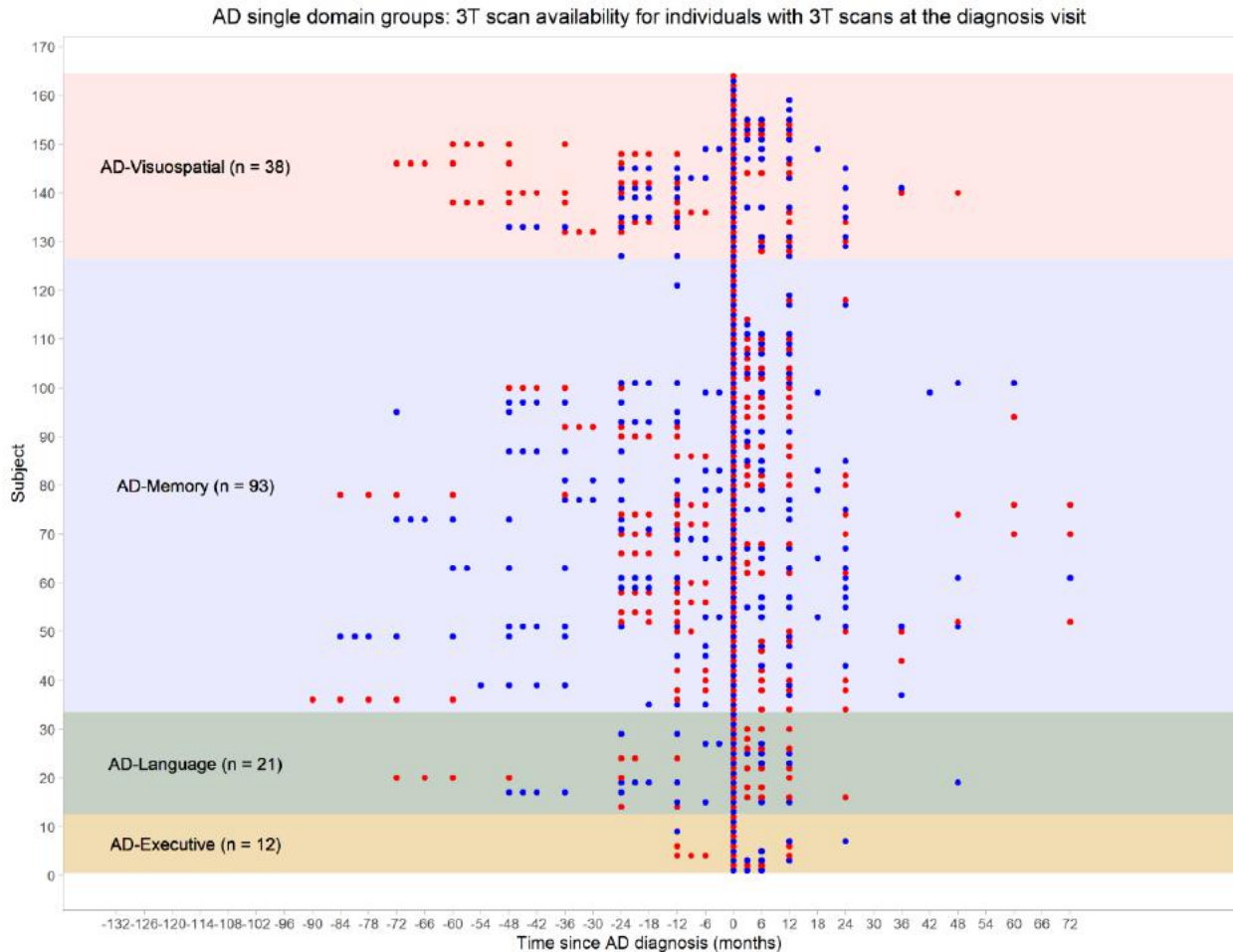
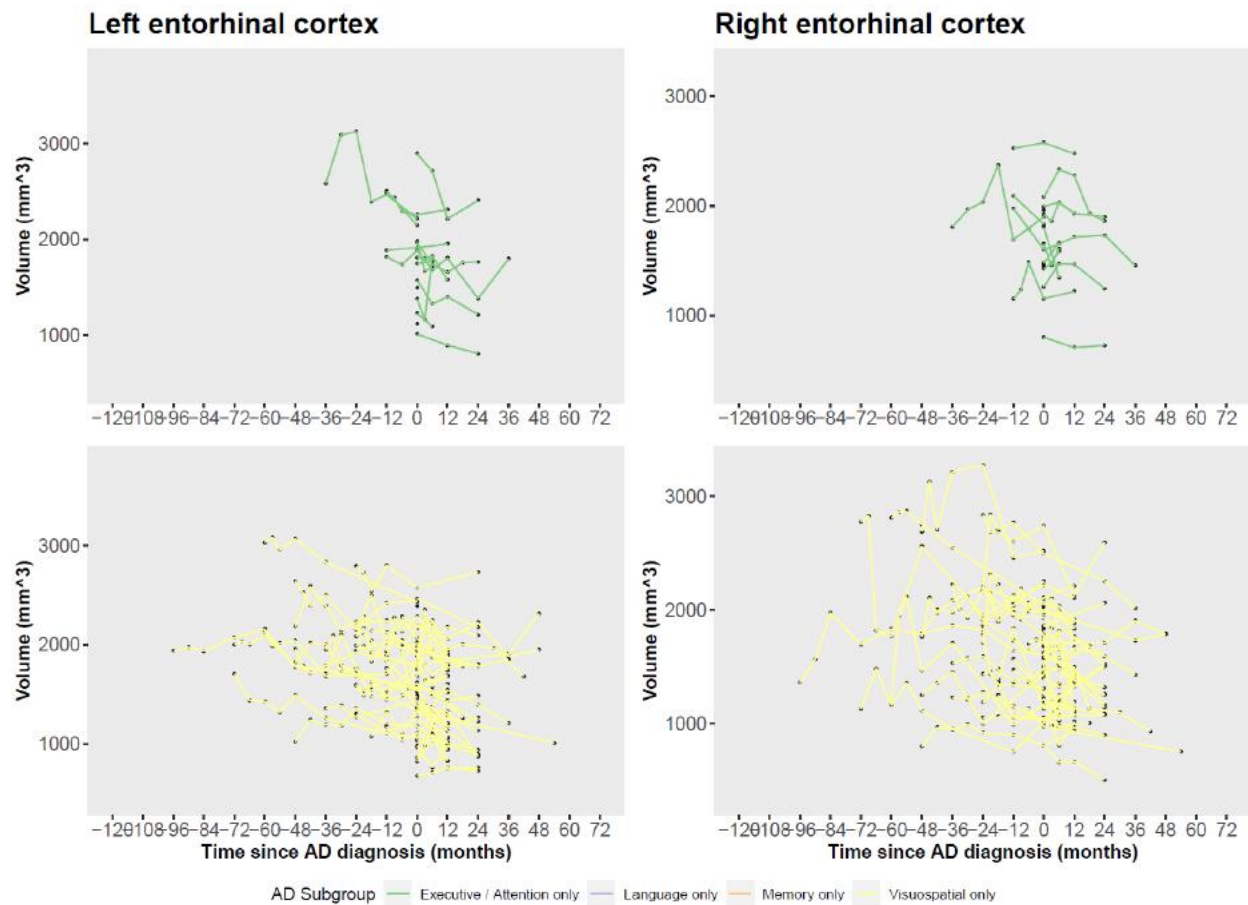


Figure 5.5b: A visualization of data availability at each time point for individuals with 3T scans at the time of AD diagnosis. Each row of dots represents an individual's data over time. Red and blue colors are used to distinguish between consecutive individuals.

ROI volume trajectories after the above pre-processing steps are shown below in Figures 5.6 and 5.7 for the same four ROIs whose trajectories were shown above (Figures 5.2 and 5.3).

Individuals with 1.5T and 3T scans are shown together in these plots. Although there seems to be a very slight reduction in the noise in the trajectories after the pre-processing step of using data corresponding to the best CNR score from each visit, there is still considerable amount of

fluctuation in most volume trajectories. Using the scans with the highest CNR score per visit was an attempt to reduce the noise in the data. However, it did not get rid of noise completely as can be seen from the trajectories in Figures 5.6 and 5.7, and noisy data still continue to remain a limitation of the current longitudinal dataset.



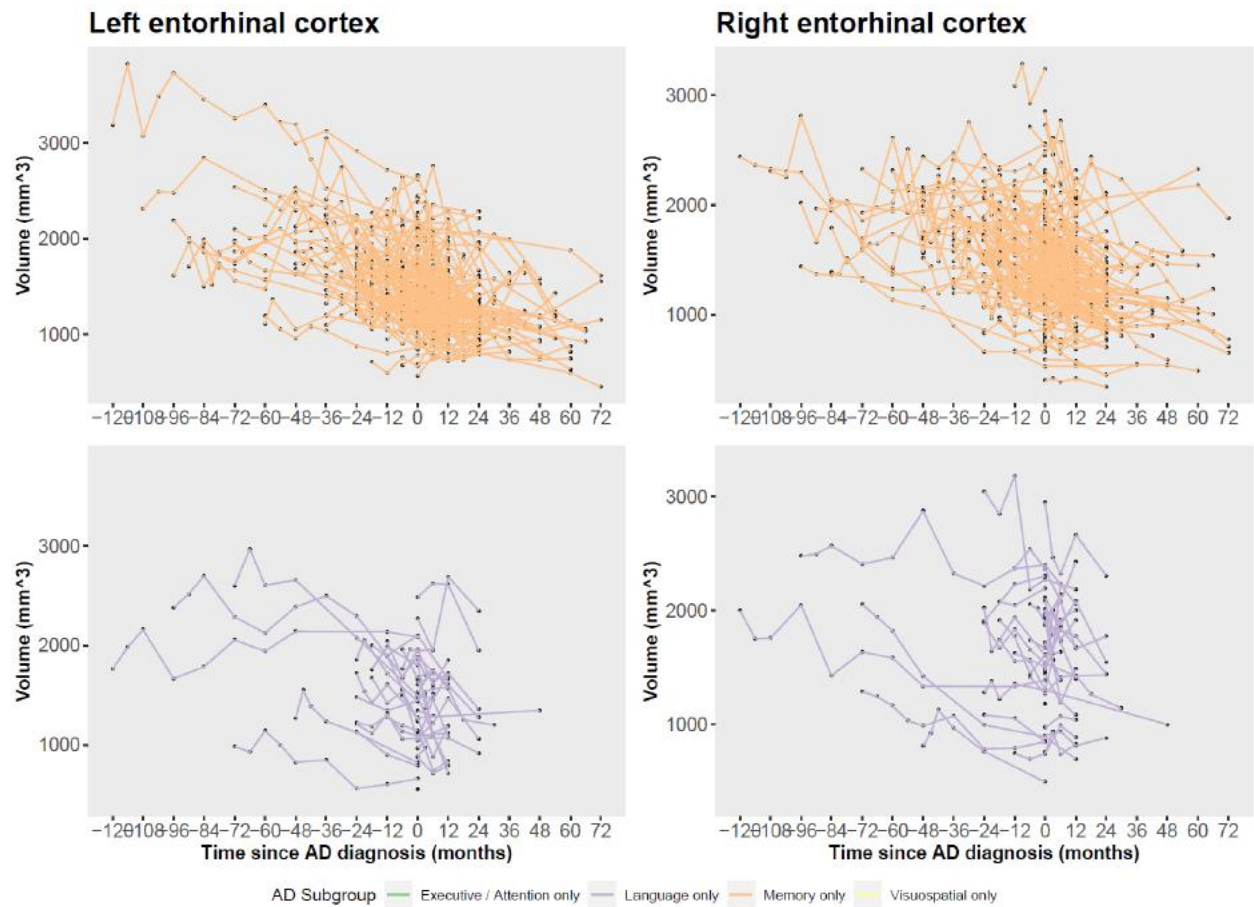
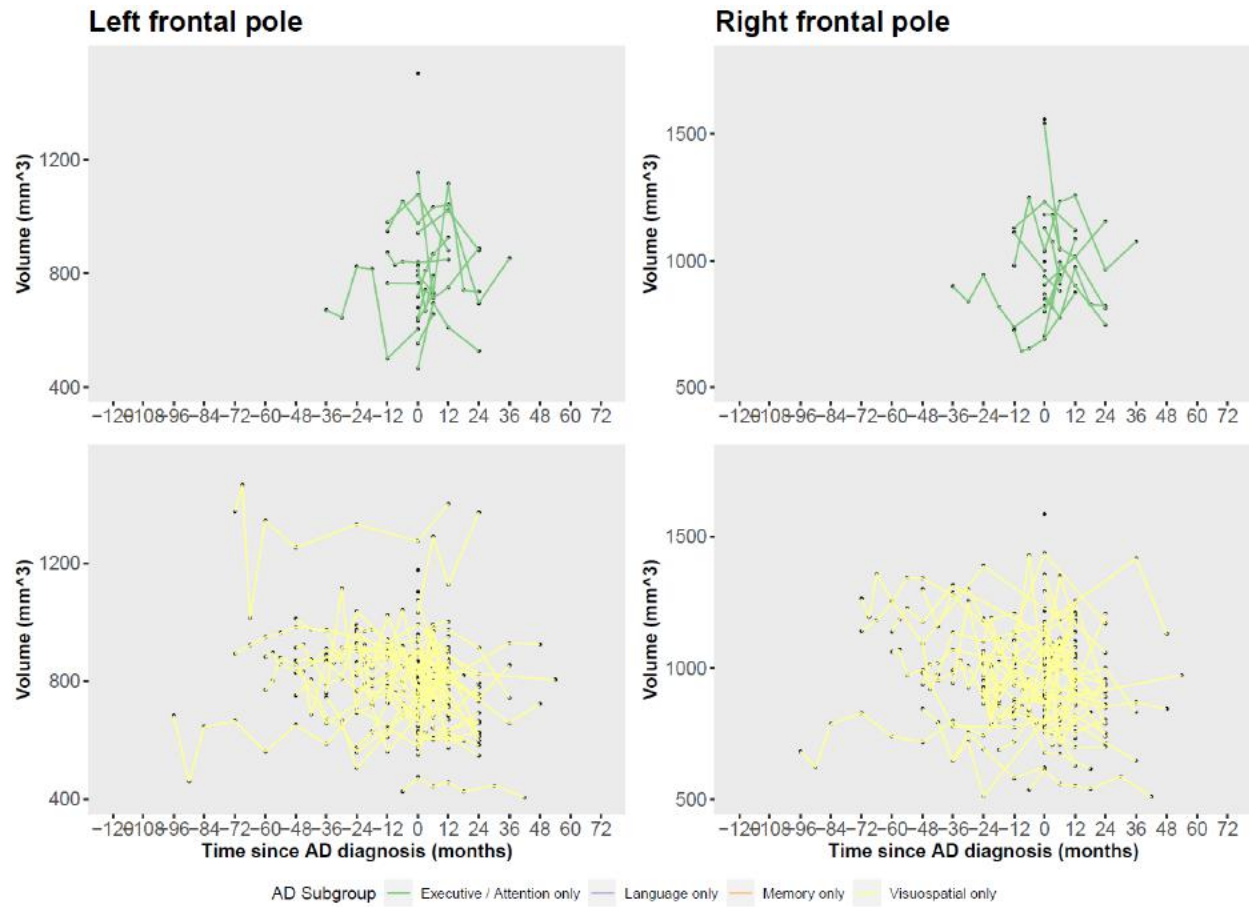


Figure 5.5: Volume trajectories for left entorhinal cortex and right entorhinal cortex for individuals in all four AD subgroups. Data is from the scans with the highest CNR score from FreeSurfer for all 3T visits in the longitudinal dataset.



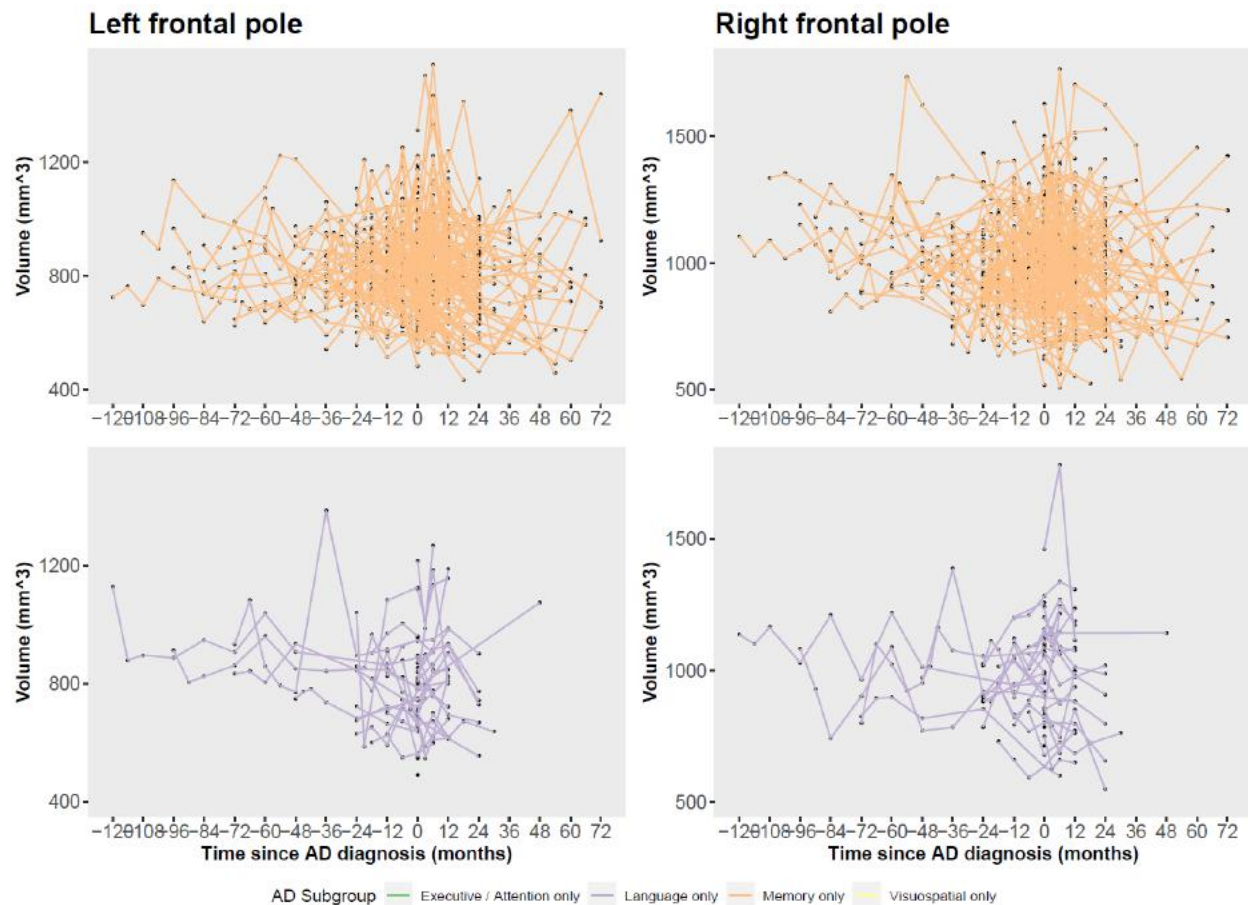


Figure 5.6 : Volume trajectories for left frontal pole and right frontal pole for individuals in all four AD subgroups. Data is from the scans with the highest CNR score from FreeSurfer for all 3T visits in the longitudinal dataset.

The characteristics of the longitudinal dataset described above highlight the complexities of the data: changing MRI field strength over time, noise in volume data, class imbalance, small sample size and irregular data. Through the pre-processing steps, I attempted to reduce some noise in data by ensuring each individual's data over time only consists of one type of field strength and by using data from the scan with the highest CNR score for each visit with multiple scans. However, trajectories with large fluctuations over time, potentially due to errors in MRI measurement and/or segmentation continue to be an issue. These factors make the current

longitudinal dataset a challenging one for finding differences in ROI volume trajectories across AD subgroups.

Volume trajectories for all 70 ROIs (based on data from best CNR scans) are provided in a supplementary file: Ch5_supplemental_VolTraject_70ROIs_bestCNR.pdf

5.2 Workflow for analysis of longitudinal data

I conclude this chapter by providing a complete workflow of the longitudinal data analysis (see next page), connecting the work discussed above (Part 1 of the workflow, also shown in Figure 5.4 above) and the analysis described in the next chapter (Part 2 of the workflow). The data exploration and preliminary analysis described in this chapter shed light on the challenges of the current dataset, which was crucial in helping me decide on linear mixed effects modeling as an appropriate method for analysis for the current longitudinal dataset. Mixed effects models are very suitable for dealing with some typical characteristics of longitudinal data such as correlation between serial measurements and irregular data. In the next chapter, I describe linear mixed effects modeling and its suitability for longitudinal data, details of implementation of the models and other steps described in Part 2 of Aim 2 workflow.

Longitudinal Data Analysis (Aim 2) Workflow

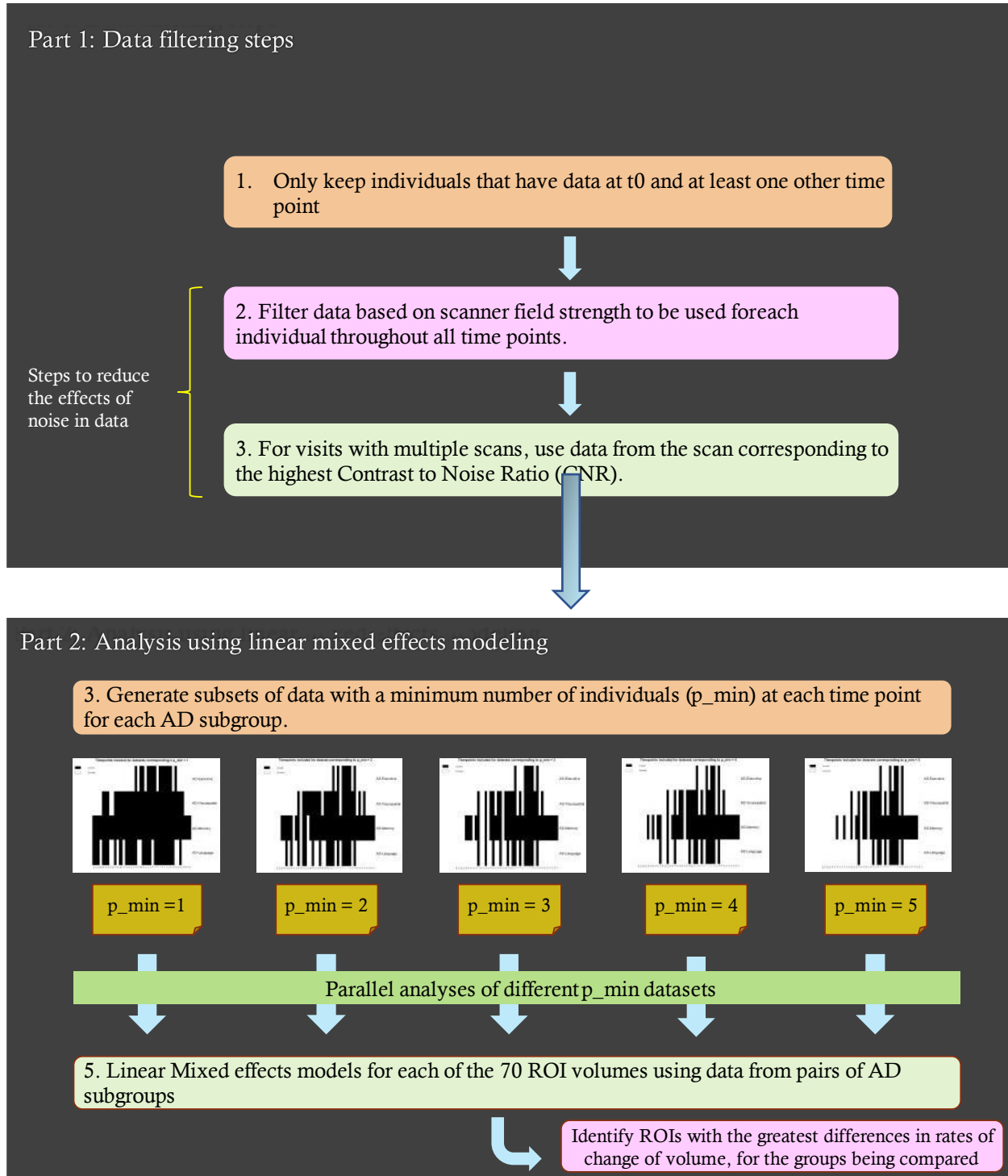


Figure 5.7: Workflow for longitudinal data analysis (Aim 2)

Chapter 6: Assessing differences in ROI volume trajectories for pairs of AD subgroups using Linear Mixed Effects modeling [Aim 2]

Of the commonly used methods for analyzing longitudinal data, two methods are 1. Repeated measures analysis of variance (or within-subject ANOVA) (Girden 1992) and 2. Cross-sectional (General Linear Model – GLM based) analysis of summary measurements such as percent annual change (Desikan et al. 2011); (Fjell et al. 2009). In my preliminary work for Aim 2, I considered the idea of using percent annual change in ROI volume as a single input data value for each individual and to analyze this transformed dataset for all ROIs and individuals using a machine learning based classification approach as I did for cross-sectional data in Aim 1. An issue with this approach is that the number of data points and the time range over which data are available for different individuals in the longitudinal dataset are not the same. So, firstly, using a summary measure is problematic in this scenario since it is not calculated using the same number of data points for different individuals. To address the different time ranges over which the summary measure is calculated, one could restrict the analysis to a specific common time period for all individuals. However, this would come at the expense of losing valuable information from data points that were excluded from the analysis due to the time window restriction. In addition to these two issues, a major concern of the approach involving analysis of summary measurements is that it does not account for the correlation structure in longitudinal data (described in section 6.1). This is also a weakness of the repeated measures analysis of variance approach. An approach that does take into account the correlation structure in longitudinal data is linear mixed effects (LME) modeling. It is also better suited for data from an ongoing research study like ADNI; characteristics of the longitudinal data in the context of such a study are highlighted below.

6.1 Overview of Linear Mixed Effects modeling

Linear mixed effects (LME) modeling has been advocated as a powerful statistical framework for analyzing longitudinal data (Bernal-Rusiel et al. 2013). Bernall et al. demonstrated the use of LME models on data from ADNI in a Matlab implementation they wrote, comparing ROI volume for different groups in ADNI: healthy controls (HC), other groups representing individuals in different phases between healthy and dementia, and individuals with AD. They showed that LME models are better suited for data from longitudinal studies and lead to an “improvement in statistical detection” compared to the other benchmark methods mentioned above. Although LME models were first introduced in 1950s and applied to a real world problem in 1959 (Henderson et al. 1959), and have been used by researchers since then, Bernall et al. believe they have been under-utilized in analyzing longitudinal brain data compared to the other two methods mentioned above due to two reasons: 1. a lack of software tools that can perform such analysis in existing pipelines for neuroimaging analysis such as FreeSurfer and 2. a lack of an understanding of the intricacies of longitudinal data in a study like the ADNI.

Here are some key characteristics of longitudinal data that must be considered for a good analysis of the data. Longitudinal data consists of measurements that are ordered in time. An important aspect of longitudinal data that needs to be accounted for in the modeling process is the correlation among serial measurements. Each individual’s measurements over different time points (repeated measures) are positively correlated and repeated measures from time points closer to each other are expected to be more correlated. Another characteristic of longitudinal data is that the between-subject variance may not be constant over time. Bernal-Rusiel et al. note three potential sources of variability affecting the correlation structure in longitudinal data: 1. Between-subject variation, 2. inherent within-subject biological change and 3. measurement

error. Lastly, missing data and irregular data are common features of longitudinal studies, i.e. not every time point has data available for every individual in the dataset. LME modeling allows for the modeling of the mean trajectory of the variable of interest over time while also taking into account the correlation structure among serial measurements. Additionally, the LME modeling framework does not require that all subjects in the study have a common set of measurement times.

Mixed effects models allow for two levels of dynamics in data to be described. These models use regression to describe the population level dynamics of a variable (the dependent variable of interest for my Aim 2 question is ROI volume); these dynamics are known as fixed effects and include all variables that we think affect the outcome variable. Individual effects, which are deviations from the population level dynamics are captured through specification of random effects, also in the form of regression coefficients. This means that certain aspects of a population level trajectory are allowed to vary “randomly” due to subject level variation that cannot be explained by fixed effects. See equations below. Linear mixed effects modeling uses linear regression specifically. Based on the ROI volume trajectories presented in the previous chapter, linear models seemed to be an appropriate choice for analyzing the current longitudinal data. Also, given the small sample size and noise in the ROI volume trajectories, I did not want to overfit the data by using a non-linear model.

Consider the equation below that describes an LME model

$$Y_i = X_i\beta + Z_ib_i + e_i \quad (6.1)$$

where Y_i is a vector of n measurements of an ROI volume (from n time points) for an individual i . X_i is a $n \times p$ design matrix for fixed effects. This includes data for all independent variables

that may have an association with ROI volume including gender, MRI scanner field strength, education, age, AD subgroup, *APOE* genotype and scan time. β is a $p \times 1$ vector of unknown fixed effects regression coefficients, to be found. Individual effects which may vary across individuals such as the intercept of ROI volume and the slope of the ROI volume with respect to time are captured through random effects. Z_i is the $n \times q$ design matrix for random effects which is a subset of X_i . b_i is a $q \times 1$ vector of q random effects. In the analysis that I carried out, I specified two random effects ($q=2$): a random effect for the intercept captured in the coefficient b_{0i} for an individual i and a random effect for the slope with respect to time captured by the coefficient b_{1i} for individual i . e_i is a $n \times 1$ vector of measurement errors for individual i . In my Aim 2 work, I used this framework to model each of the 70 ROI volumes using 70 independent LME models.

Each individual's volume trajectory for a given ROI can be described by

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2 ADSubgroup + \beta_3 field_strength + \beta_4 gender + \beta_5 age + \beta_6 yrs_Educate + \beta_7 APOEgroup + \beta_8 ICV_t0_Scaled + \beta_9 TotalGrayVol_ICV_ratio_t0 + \beta_{10} ADSubgroup * t_{ij} + e_{ij} \quad (6.2)$$

where Y_{ij} is the ROI volume for individual i at timepoint j , t_{ij} is the scan timepoint j for individual i and e_{ij} is the measurement error at timepoint j for individual i . ICV_t0_scaled represents the intracranial volume (ICV) for each individual at the time of AD diagnosis; I used this as a measure to control for overall head size of different individuals. Raw ICV values are very large compared to values of any of the other variables being used in the model. This results in issues in the LME model fitting process as the matrix to be inverted becomes close to singular. This happens because the relatively small values from the other variables in the matrix get

zeroed when being compared to the much larger values of the ICV variable. Hence, to avoid this issue, ICV values needed to be scaled so they are not on a drastically different scale than the other variables. I used the proportion of maximum scaling (POMS) method (Little 2013) to scale the ICV values which transforms the values to a 0 to 1 scale:

$$\text{POMS} = [(observed - minimum)/(maximum - minimum)] \quad (6.3)$$

Specifically,

$$\text{ICV}_{t0_scaled} = (\text{ICV}_{t0} - \min(\text{ICV}_{t0})) / (\max(\text{ICV}_{t0}) - \min(\text{ICV}_{t0})) \quad (6.4)$$

Unlike standardization, this maintains the proportion of distances between the observations. In my analysis of longitudinal data, I made a deliberate choice of not standardizing variables in the implementation of LME models to keep the interpretation of results as simple as possible and also because of the complexities of longitudinal data which makes it hard to standardize variables in a sensible way (Moeller 2015).

For controlling for effects of overall atrophy of the brain, instead of using w-scores as I did in the cross-sectional data analysis, here in longitudinal data analysis, I defined and used a measure called the TotalGrayVol_ICV_ratio which is defined as Total gray volume at t0 / ICV at t0 as a measure of how much gray matter is present in the brain relative to a person's head size at the first visit in the study when an individual is diagnosed with AD. The reason for not using w-scores in longitudinal data analysis is that w-scores were not available for about 50 individuals in the data. Using w-scores in the current analysis would have meant reducing the sample size by 50 individuals. Given the already small sample size, I opted not to include w-scores in the analysis and to use the TotalGrayVol_ICV measure as a way of approximating overall atrophy in

the brain. The TotalGrayVol_ICV_ratio does not need to be scaled as the ratio of total gray volume to ICV is still on a comparable scale as other variables in the model.

In the current model, the rate of change of ROI volume is allowed to vary by AD subgroup. Other covariates: gender, MRI scanner field strength, education, age, *APOE* genotype are allowed to be associated with ROI volume but these variables have not been modeled to contribute to differences in rate of change of ROI volume. MRI scanner field strength is a variable that is not related to brain processes, so it should not have an effect on the rate of change of ROI volume. The other covariates gender, education level, age and *APOE* genotype may have an impact on how fast brain volume decreases over time but it seems unlikely that these variables would affect rates of change of different ROIs in different ways. To start with a simpler model, I did not include interaction terms between time and each of the following: gender, education, age and *APOE* genotype. One advantage of this simpler model where the only time dependent terms are t and $ADSubgroup * t$ in the model is that an average rate of change can be obtained for all individuals in the reference group ($ADSubgroup = 0$), given by the $\beta_1 t$ term regardless of an individual's gender, education level, age or *APOE* genotype status. Knowing this average rate of change for the reference group is useful for comparing the magnitude of the difference in the rates of change between the two AD subgroups.

For the purpose of my analysis which is understanding how a given ROI differs longitudinally across pairs of AD subgroups, the coefficients of interest are β_2 , the coefficient for AD subgroup and β_{10} , the coefficient for the interaction between AD subgroup and time. β_2 represents the difference in ROI volumes between the AD subgroups at $t=0$ (time corresponding to the AD diagnosis visit) and β_{10} represents the difference in the rate of change of volume with respect to time for the two AD subgroups. Evaluating the LME model (Equation 6.2) at specific values for

$ADSubgroup$ and t allow one to see these interpretations of β_2 and β_{10} . Since I was interested in understanding how the ROI volumes differ across two given AD subgroups, the $ADSubgroup$ variable was modeled as a binary variable. $ADSubgroup$ was defined as 0 for the reference AD subgroup and 1 for the target AD subgroup. The choice of which AD subgroup is chosen to be the reference group is arbitrary. Evaluating the LME model (Equation 6.2) at $t=0$ for each of the AD subgroups, one can show that the ROI volumes for the two AD subgroups at $t=0$ differ by β_2 . Evaluating the LME model for $ADSubgroup = 0$, one can see that the rate of change of ROI volume is given by β_1 , while when $ADSubgroup = 1$, the rate of change of ROI volume is given by $\beta_1 + \beta_{10}$. Hence, the population level difference in rates of change of ROI volume for the two groups is given by β_{10} based on the model.

The following assumptions are made in the LME modeling framework for estimating β from the data. First, the following distributions are assumed: $b_i \sim N(0, D)$ and $e_i \sim N(0, \sigma^2 I_{n_i})$ where the notation $N(0, \Sigma)$ indicates a zero mean multivariate Gaussian with covariance matrix Σ . I_{n_i} is a $n_i \times n_i$ identity matrix, and all $b_1, \dots, b_m, e_1, \dots, e_m$ are independent where m is the number of subjects. There is an important distinction between the marginal and conditional means of Y_i in LME models. The marginal or population-averaged mean of Y_i is the expected value of equation 6.1, which yields

$$E(Y_i) = X_i\beta \tag{6.3}$$

The conditional or subject-specific mean of Y_i , given b_1 is

$$E(Y_i | b_i) = X_i\beta + Z_i b_i \tag{6.4}$$

The corresponding marginal and conditional covariances are given below in equations 6.5 and 6.6 respectively.

$$Cov(Y_i) = Cov(Z_i b_i) + Cov(e_i) = Z_i D Z_i^T + \sigma^2 I_{n_i} \quad (6.5)$$

$$Cov(Y_i | b_i) = Cov(e_i) = \sigma^2 I_{n_i} \quad (6.6)$$

Ultimately, fitting an LME model is equivalent to estimating the unknown β coefficients and the model parameters D and σ . The details of parameter estimation are not discussed here; they are clearly laid out in Bernal et al. (2013)'s work. I modeled each ROI volume over time using Bernal et al.'s Matlab implementation of univariate LME modeling which uses iterative solvers to estimate D and σ , which are then used to estimate β coefficients as a closed-form solution. Hence, the β coefficients in an LME model describe the average population level dynamics, while depending on D and σ , which collectively present information about subject level deviations from the population average and the correlation structure in data.

6.2 Analysis design

This section describes in detail, the steps outlined in Part 2 of the workflow diagram which was presented at the end of Chapter 5. Readers may wish to refer to Figure 5.7 for an overview of the workflow.

Due to the small sample sizes, irregular data and imbalanced class sizes in the current longitudinal dataset, some timepoints may have very few data points from a given AD subgroup. In such scenarios, it is possible that the LME results (model coefficients) are biased as the small number of data points for a given time point may not be a true representation of the overall data trends. It was therefore necessary to carry out an analysis that considers the effect of excluding different time points for particular AD subgroups from the analysis based on a restriction on the minimum number of data points required for the subgroup at a given time point.

I generated a series of subsets of data for each AD subgroup based on the following restrictions for the minimum number of data points (individuals) required at each time point. Let this number be called p_{\min} . In the first case, I allowed all timepoints to be included in the analysis for each AD subgroup being considered. This is the entire dataset and for this case, $p_{\min} = 0$, i.e. each time point must have at least 0 data points for each AD subgroup. Next, I generated a subset based on $p_{\min} = 1$. This represents datasets for each AD subgroup where at each time point, the given AD subgroup has at least one data point. In this fashion, I obtained additional subsets of data for each AD subgroup corresponding to $p_{\min} = 2, 3, 4$ and 5 . Here, a dataset with $p_{\min} = 5$ is the most restrictive dataset, where each time point kept in the data for an AD subgroup was required to have at least 5 data points for that subgroup. Note, that although the timepoints that get included in datasets corresponding to $p_{\min}=0$ (all time points) and $p_{\min} = 1$ (all time points with at least one data point) may be different, the results from LME modeling are expected to be the same for these two cases. This is because in the case of the $p_{\min} = 0$ dataset, it is expected that the timepoints with no datapoints will not have any contribution towards the model. Essentially, the datasets corresponding to $p_{\min} = 0$ and $p_{\min} = 1$ are the same, with the case of $p_{\min} = 0$ having no subjects' data for the extra time points that are part of this dataset. I confirmed that the results from LME modeling on the datasets with $p_{\min} = 0$ and $p_{\min} = 1$ were identical. Here, I present the analysis for the cases of $p_{\min} = 1, 2, 3, 4$ and 5 .

Below is a tabulation of time points that were included in each of the subsets for each AD subgroup; all time points are in months since AD diagnosis. The final datasets used in linear mixed effects modeling consisted of combined datasets from each AD subgroup for each p_{\min} . For example, for the case of comparing AD-Language and AD-Visuospatial corresponding to $p_{\min} = 5$, the dataset used in linear mixed effects modeling for each ROI consisted of data from

the timepoints -48, -24, -18, -12, -6, 0, 3, 6, 12, 24 for AD-Language and data from timepoints for -60, -48, -42, -36, -30, -24, -21, -18, -12, -6, 0, 3, 6, 12, 24 for AD-Memory. Tabulations of which timepoints got included in the datasets corresponding to different p_{\min} values are shown in Figures 6.1a-e.

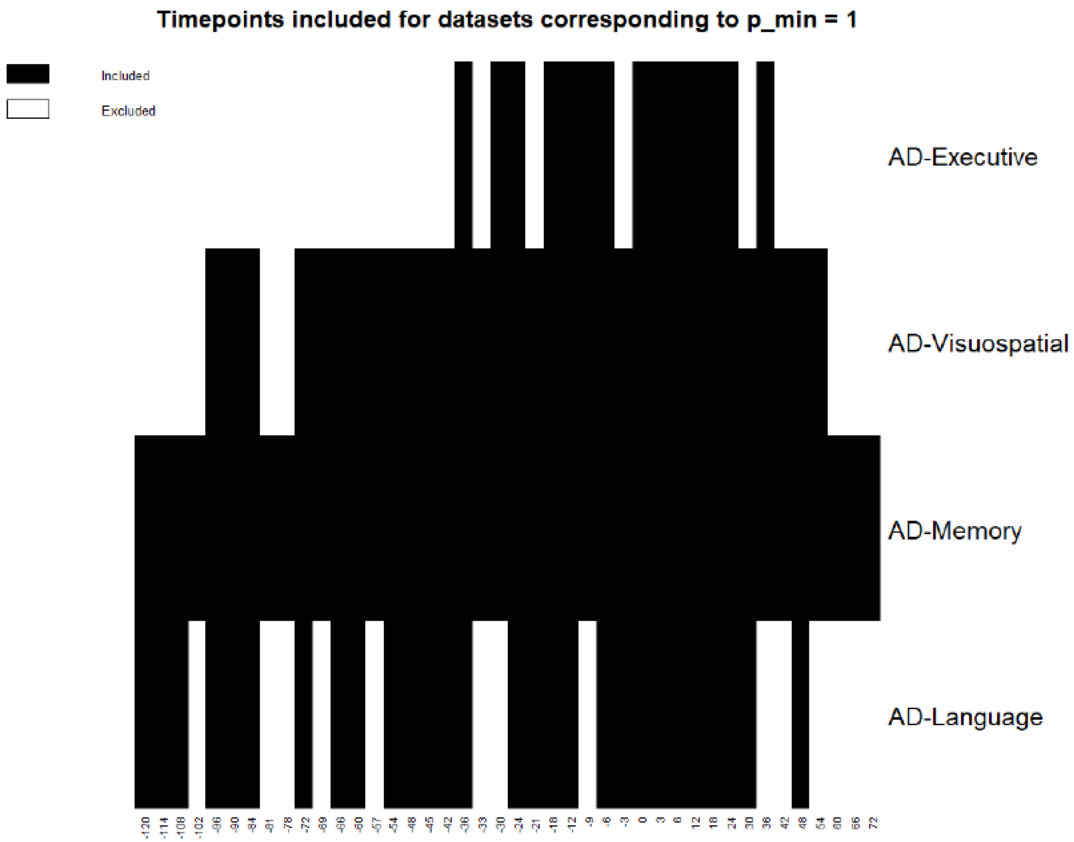


Figure 6.1a: A tabulation of timepoints that got included in datasets corresponding to $p_{\min} = 1$, i.e. all timepoints with at least one individual's data for a given AD subgroup.

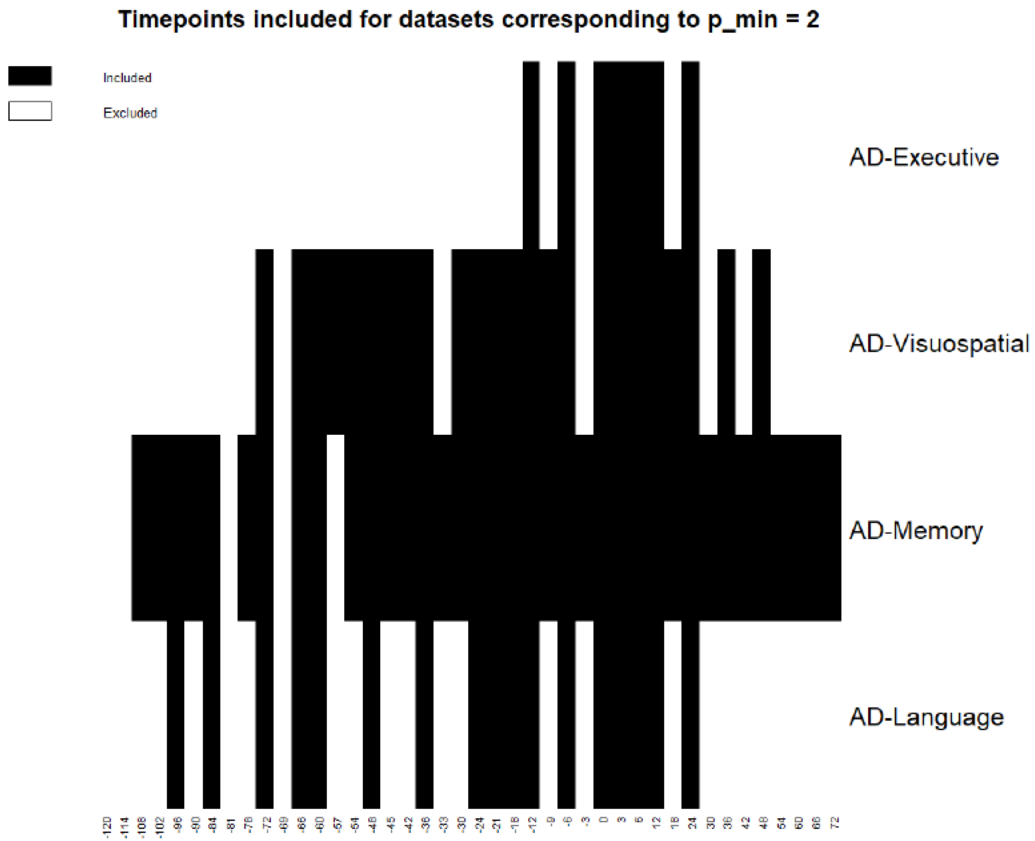


Figure 6.1b: A tabulation of timepoints that got included in datasets corresponding to $p_{min} = 2$, i.e. all timepoints with at least two individuals' data for a given AD subgroup.

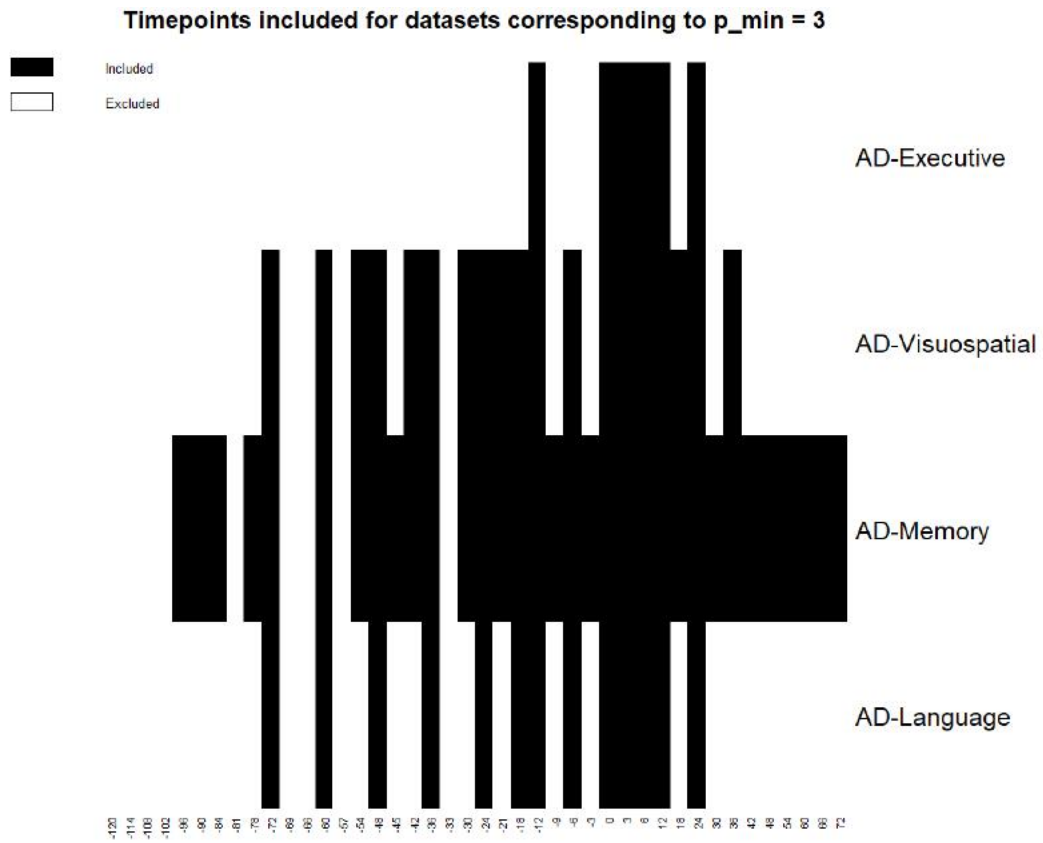


Figure 6.1c: A tabulation of timepoints that got included in datasets corresponding to $p_{\min} = 3$, i.e. all timepoints with at least three individuals' data for a given AD subgroup.

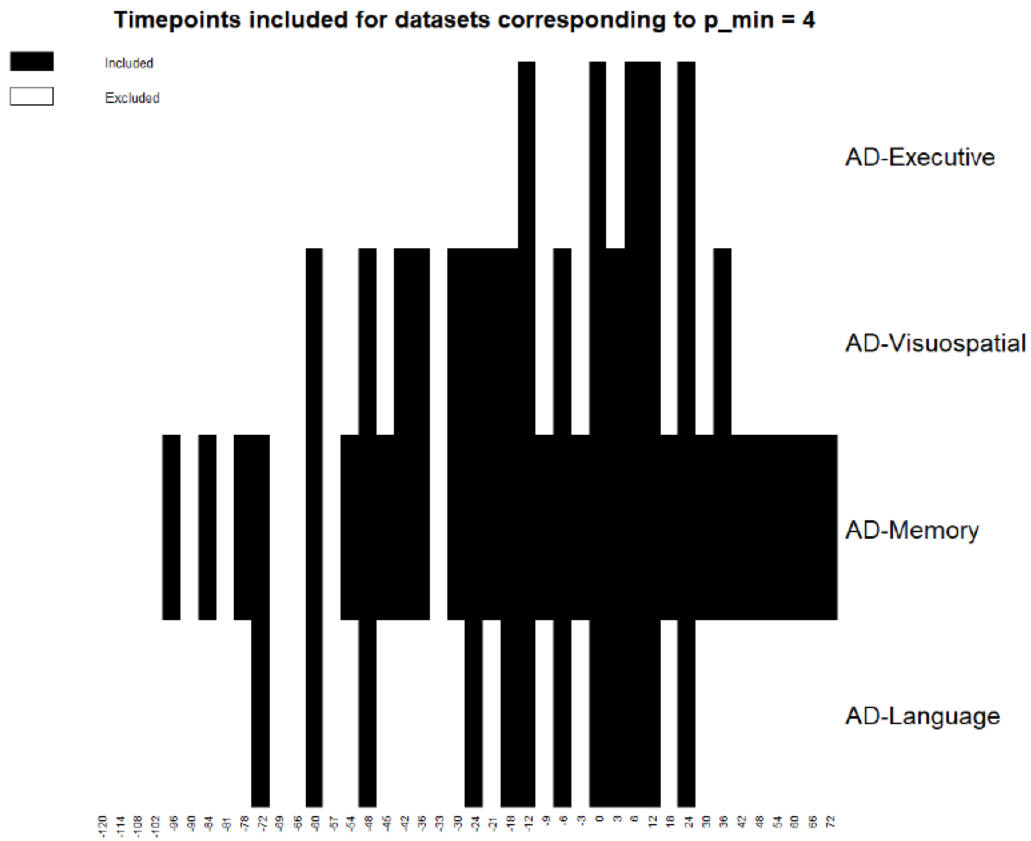


Figure 6.1d: A tabulation of timepoints that got included in datasets corresponding to $p_{\min} = 4$, i.e. all timepoints with at least four individuals' data for a given AD subgroup.

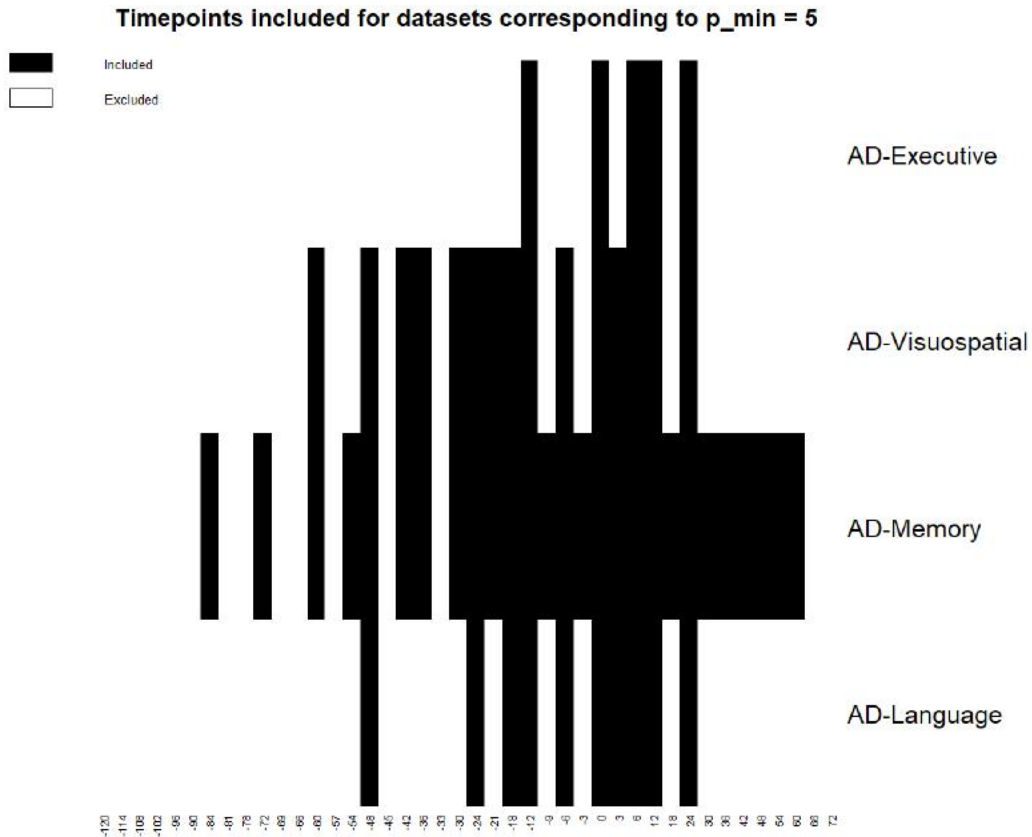


Figure 6.1e: A tabulation of timepoints that got included in datasets corresponding to $p_{\min} = 5$, i.e. all timepoints with at least five individuals' data for a given AD subgroup.

The goal in carrying out the LME analysis on the different subsets of data based on p_{\min} was to get a sense of the robustness of the models. For example, how much do the coefficients of interest (β_2 , difference in ROI volumes at $t=0$ for the two AD subgroups and β_{10} , which represents the difference in rate of change of ROI volume for the two AD subgroups) change as different datasets corresponding to different p_{\min} values are analyzed. On one hand, the greater the value of p_{\min} , we expect the LME results to be most reliable. However, increasing the value of p_{\min} comes at the expense of losing some data points. Hence, in increasing p_{\min} ,

there is a trade-off between improving the reliability of LME model results and potentially losing valuable information through exclusion of data points. The decision to limit my analyses to an upper bound of $p_min = 5$ is subjective; it is guided by the choice of not wanting to lose too many data points. This analysis does not establish which p_min dataset is the most favorable to analyze using LME models for the current dataset. Rather, it gives a sense of the sensitivity of LME model results for the current longitudinal dataset to the inclusion/exclusion of time points with a number of individuals below a threshold.

The following LME model was implemented for each ROI on each dataset from two given AD subgroups corresponding to different p_min values. Individual effects (each individual's deviations from the population dynamics) were represented in the model by specification of the intercept and rate of change of ROI volume with respect to time as random effects. In equation 6.2, β coefficients represent the fixed effects aspects of the variables, and they provide information about the population level dynamics. The b coefficients represent random effects for the chosen variables, based on data from each individual i and timepoint j . e_{ij} is the measurement error for individual i at time j which for the current data we specified to be 0. To make it easier to parse and bring focus to the coefficients of interest, equation 6.2 can be re-expressed as below where each individual i 's ROI volume at time j , Y_{ij} , is given by:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2 ADSubgroup + other\ fixed\ effects\ terms + \beta_{10} ADSubgroup * t_{ij} + e_{ij} \quad (6.7)$$

where $other\ fixed\ effects\ terms = \beta_3\ field_strength + \beta_4\ gender + \beta_5\ age + \beta_6\ yrs_Educate + \beta_7\ APOEgroup + \beta_8\ ICV_t0_Scaled + \beta_9\ TotalGrayVol_ICV_ratio_t0$ as discussed in the previous section.

The population level trajectory for an ROI volume based on the data from two AD subgroups is given by

$$ROI_vol(t) = \beta_0 + \beta_1 t + \beta_2 ADSubgroup + other\ fixed\ effects\ terms + \beta_{10} ADSubgroup * t \quad (6.8)$$

For each of the 70 ROIs, data from the following pairs of AD subgroups was considered separately: a. AD-Language and AD-Memory, b. AD-Memory and AD-Visuospatial and c. AD-Language and AD-Visuospatial. Within each of these, five variations of datasets given by different p_min values were analyzed separately using an LME model for a given ROI. As noted in the previous chapter, the AD-Executive group was excluded from the LME modeling analysis due to small sample size.

Hypothesis testing for the significance of the β_2 and β_{10} coefficients for each ROI was considered, with the null hypothesis was that the coefficient of interest is 0. If the p-value was below a threshold, the null hypothesis could be rejected. Determining this threshold required correction for multiple hypotheses testing since there are 70 ROI models for a given pair of AD subgroups being compared, and five variations of datasets used for a given comparison, which results in a total of 350 models whose p-values are being analyzed. A Holm-Bonferroni correction for a family wise error rate of 0.05 yields a threshold of 1.43×10^{-4} for the smallest p-value which means that the smallest p-value in the results should be less than this threshold for it to be significant. The Holm-Bonferroni corrected thresholds for the next four p-values in increasing order are 1.43×10^{-4} , 1.44×10^{-4} , 1.44×10^{-4} and 1.45×10^{-4} . Not only are these thresholds very restrictive (compared to observed p-values presented in the next section) for finding ROIs that might be important in AD subgroup differences, using p-values to assess how important an ROI is may not be suitable when some of the datasets might only have a few data

points for some time points and when the overall sample sizes of AD subgroups are not that large. In such cases, the p-values can differ quite a bit depending on which data points got included in the analysis or not.

In the analysis presented below, I did not focus on the significance of p-values but rather looked at the relative magnitudes of coefficients (normalized by ROI size) across the different ROI models to obtain a list of ROIs that might be most important to pay attention to for AD subgroup differences. The small sample sizes of the AD subgroups, imbalanced class sizes and the different number of data points at different time points could lead to p-values that are sensitive to individual data points as noted earlier. Hence, given the current dataset, I chose not to focus on p-values. Additional data may be needed to carry out a reliable analysis based on p-values, which is forecasted as future work. P-values are still presented in results but they are not the focus of prioritizing the ROIs for subgroup differences.

For each pair of AD subgroups analyzed using LME models, I analyzed the coefficient values for β_2 and β_{10} in datasets corresponding to each p_min value. For each of these coefficients, I made a list of the top 5 ROIs from the analysis for each p_min dataset based on the magnitude of coefficients. It is important to note that the ROIs in different LME models were of different sizes. Hence, for the purposes of comparing the coefficients from different ROI LME models, instead of directly using the magnitudes of coefficients (β_{10} units: mm³/year and β_2 units: mm³) to rank the ROIs, I normalized the coefficients by a measure that gives a rough sense of ROI size so that the coefficients are represented relative to each ROI's size. As a rough estimate of this measure for each ROI, I used the average ROI volume calculated from all four AD subgroups from the time of AD diagnosis; I used the cross-sectional dataset for this calculation. Hence, for each ROI model, I calculated from β_2 the difference in ROI volume at t=0 for two AD subgroups

as a percentage of the average ROI volume ($\frac{\beta_2}{Average\ ROI\ volume} * 100\%$) and from β_{10} the difference in rate of change of ROI volume between two AD subgroups as a percentage of the average ROI volume ($\frac{\beta_{10}}{Average\ ROI\ volume} * 100\%$). Based on these measures, I obtained two lists of ROIs that appear in the top 5 ROIs for different p_min datasets, one based on the normalized β_2 coefficient (difference in ROI volume at t=0 as % of average ROI volume) and the other based on the normalized β_{10} coefficient (difference in rates of change of ROI volume as % of average ROI volume). I then analyzed how the coefficients for each of these ROIs differ over analyses of different subsets of data corresponding to p_min values. To assess whether the differences in rate of change of ROI volume for two AD subgroups as given by the normalized β_{10} coefficient are substantial for a given ROI, I compared the differences in rate of change to the rate of change of each the two AD subgroups which are given by $\frac{\beta_1}{Average\ ROI\ volume} * 100\%$ for the reference group and $\frac{\beta_1 + \beta_{10}}{Average\ ROI\ volume} * 100\%$ for the target group. This process was carried out for each of the three comparisons: AD-Language vs. AD-Memory, AD-Memory vs. AD-Visuospatial and AD-Language vs. AD-Visuospatial. The decision to focus on the top 5 ROIs was subjective and there is no absolute threshold of how many top ROIs should be considered. The goal was to obtain a handful of ROIs that show greater differences across pairs of AD subgroups compared to other ROIs. The goal of my work is hypothesis generation for which ROIs might be important to look at in future neuropathological studies for understanding AD subgroup differences. As a starting point, I've chosen to focus on five ROIs for each p_min dataset and that is the focus of the results presented in my dissertation. However, in future, the current analysis can easily be extended to report a greater number of ROIs in the hypothesis generation. The appropriate number of ROIs to be reported as important for looking at AD

subgroup differences would be best determined by consulting neuropathologists who may be using the results from this hypothesis generation work.

6.3 Results and Discussion

The majority of the LME models for 70 ROIs across the different p_{\min} datasets converged, i.e. a solution could be found for the coefficients in the model based on the Bernal et al. implementation in Matlab. A few exceptions were the models for the following ROIs and datasets. For models based on data from AD-Language and AD-Memory individuals, the model for right caudal anterior cingulate gyrus volume did not converge for the $p_{\min} = 5$ case. For models based on data from the AD-Language and AD-Visuospatial groups, the following models did not converge: right insula (for $p_{\min} = 0, 1, 2, 3, 4$), left insula (for $p_{\min} = 2, 5$), right medial orbitofrontal gyrus (for $p_{\min} = 2, 4$) and left temporal pole (for $p_{\min} = 5$). Only LME models that converged were used for determining the top ROIs that show the most differences across pairs of AD subgroups in longitudinal data.

6.3.1. AD-Language vs. AD-Memory

For AD-Language vs. AD-Memory, the top 5 ROIs based on the magnitude of differences in rates of change of volume between the two AD subgroups, normalized by ROI size, for different subsets of data are shown in Table 6.1. Based on these rankings, I focused on the following ROIs for subsequent analysis of rate of change differences between the two subgroups: **left posterior cingulate gyrus, left entorhinal cortex, right entorhinal cortex, left caudal anterior cingulate gyrus, right caudal middle frontal gyrus, left parahippocampal gyrus, right parahippocampal gyrus, and right frontal pole**. Table 6.2 shows the top 5 ROIs from the LME analyses of the different p_{\min} datasets based on the magnitude of the differences in ROI volumes between the two AD subgroups at $t=0$, normalized by ROI size. These ROIs were: **right**

entorhinal cortex, left temporal pole gyrus, right hippocampus, left inferior temporal gyrus and left middle temporal gyrus. In the LME models for ROI volumes based on AD-Language and AD-Memory data, AD-Language was coded as 0 (the reference group) and AD-Memory was coded as 1 (the target group). Below, I discuss the results from LME models for the ROIs mentioned above, first for the rate of change differences between the subgroups and then for the differences in ROI volumes at $t=0$.

ROI rankings for AD-Language vs. AD-Memory based on differences in rate of change of volume	
Dataset (p_min)	Top 5 ROIs based on β_{10} coefficient (normalized by average ROI volume in AD)
1	1. Left posterior cingulate, 2. Left entorhinal, 3. Right entorhinal, 4. Left parahippocampal and right caudal middle frontal, 5. Left caudal anterior cingulate
2	1. Left posterior cingulate, 2. Left entorhinal, 3. Left caudal anterior cingulate, 4. Right entorhinal, 5. Right caudal middle frontal
3	1. Left posterior cingulate, 2. Left entorhinal, 3. Right entorhinal, 4. Left caudal anterior cingulate, 5. Right caudal middle frontal
4	1. Left posterior cingulate, 2. Left entorhinal, 3. Right entorhinal, 4. Left caudal anterior cingulate, 5. Right caudal middle frontal
5	1. Left entorhinal, 2. Left posterior cingulate, 3. Left caudal anterior cingulate, 4. Right frontal pole, 5. Right parahippocampal

Table 6.1: Top 5 ROIs based on differences in rates of change of ROI volume (given by β_{10}) between AD-Language and AD-Memory groups, from linear mixed effects models on different subsets of data.

ROI rankings for AD-Language vs. AD-Memory based on differences in longitudinal volume	
Dataset (p_min)	Top 5 ROIs based on β_2 coefficient (normalized by average ROI volume in AD)
1	1. Right entorhinal, 2. Right hippocampus, 3. Left temporal pole, 4. Left inferior temporal, 5. Left middle temporal
2, 3, 4, 5	1. Right entorhinal, 2. Left temporal pole, 3. Right hippocampus, 4. Left inferior temporal, 5. Left middle temporal

Table 6.2: Top 5 ROIs based on differences in rates of change of ROI volume (given by β_2) between AD-Language and AD-Memory groups, from linear mixed effects models on different subsets of data.

6.3.1.1 Differences in rates of change of ROI volume

Although the ranking order of the ROIs based on the normalized β_{10} coefficients differed slightly for datasets corresponding to different p_min values, there was a substantial overlap in the ROIs that appeared in the top five ROIs throughout all p_min datasets' analyses. For each of these regions, I looked at how the normalized β_{10} coefficient and corresponding p-values differed for different subsets of data corresponding to different values of p_min. As noted earlier, the p-values were not used to make any assessments about how important an ROI is for AD subgroup differences. They are presented here for completeness and to demonstrate that the p-values from the current analysis are not in the range of values needed for significance as was determined by Holm-Bonferroni correction in the previous section. The results are organized in the following way: first, I've presented Tables 6.3a-6.3e (one table for each p_min dataset) summarizing the differences in rate of change of ROI volume between the two AD subgroups as

a percentage of average ROI volume as well as the rate of change of ROI volume for each of the subgroups, also as a percentage of average ROI volume. Next, I've presented figures that summarize the results over all p_min datasets, including plots that show the average differences in rates of change of ROI volume across all p_min datasets and the ROIs of importance (Figure 6.1) as well as how substantial these differences between the AD subgroups are relative to the rates of change for each of the two subgroups and which subgroup has the greater rate of decline for the ROIs of importance (Figure 6.2). P-values associated with the β_{10} coefficient across the different p_min datasets for each of the selected ROIs are shown in Figure 6.3. All rates in the figures are shown as % of average ROI volume which was calculated from the cross-sectional dataset for all 4 AD subgroups. As discussed in section 6.2, the average ROI volume is used as a rough estimate for ROI size, and rates were normalized by the average ROI volume for comparing values across ROIs of different sizes.

As per the coding of the AD subgroups in the models, a positive β_{10} coefficient corresponds to a more positive rate of change in the AD-Memory group or equivalently a slower decline in ROI volume in the AD-Memory group compared to the AD-Language group. As can be seen in Figure 6.1 and Tables 6.3a-6.3e, the estimate of β_{10} from the LME models differs for the different p_min datasets, as one may expect. For some ROIs, the variation in β_{10} across the five p_min datasets is greater than the others. For example, the left entorhinal cortex stands out as having a large error bar around the mean difference in rates of change of ROI volume between the two AD subgroups, especially since the estimated difference for the p_min =5 dataset is substantially different than the estimated values for the other p_min datasets. However, qualitatively, it seems that for most ROIs, the β_{10} estimates do not differ in a huge way for different p_min datasets. A formal quantitative analysis is needed as future work to assess

whether the differences observed in the β_{10} estimates from models for a given ROI for different p_min datasets are small or large and to determine which p_min dataset is most reliable for reporting results. For the current analysis, the interpretations I make below for the chosen ROIs for AD-Language vs. AD-Memory differences in rates of change of ROI volume are based on the mean of the β_{10} estimates across the five p_min datasets.

As can be seen in Figure 6.2, of the ROIs chosen to be most important for rate of change differences between AD-Language and AD-Memory groups, **left posterior cingulate gyrus, left entorhinal cortex, right entorhinal cortex, and left caudal anterior cingulate gyrus** have **greater rates of decline of ROI volume in the AD-Language group**, while **right caudal middle frontal gyrus, left parahippocampal gyrus and right frontal pole** have **greater rates of decline in the AD-Memory group**. Figure 6.2 also illustrates the relative difference in rates between the groups for the different ROIs. Qualitatively speaking, all ROIs presented in Figure 6.2 show substantial differences in the rates between the two subgroups when compared to the rate for each subgroup. On average, based on the results from different p_min datasets, **left caudal anterior cingulate gyrus, left posterior cingulate gyrus, right frontal pole and right caudal middle frontal gyrus** have the most substantial differences relative to the rates of the two groups. Note that while the left entorhinal cortex has the fastest declining rates for both groups (-6.3% for AD-Language and -4.8% for the AD-Memory group) based on the LME models, the greatest difference in rates of change relative to the rates of each of the AD subgroups are for the left posterior cingulate gyrus (-2.7% for the AD-Language group and 1.2% for the AD-Memory group.) A difference of 1.5% in the rates of the two groups for the left posterior cingulate gyrus is more substantial than a difference of 1.5% in the two groups for the left entorhinal cortex.

Tables 6.3a – 6.3e: The ROIs shown here are a collective list based on the top 5 ROIs for AD-Language vs. AD-Memory from the LME analysis of each p_min dataset based on the magnitude of the β_{10} coefficient, normalized by average ROI volume ('Difference in rates as % of Avg ROI volume' column in the table). Individuals with AD* indicated in the second column of each table are individuals from the 4 single domain groups: AD-Executive, AD-Language, AD-Memory and AD-Visuospatial. Positive differences in rates correspond to a greater rate of decline of ROI volume in the AD-Language group compared to the AD-Memory group. Negative differences in rates of change correspond to a greater rate of decline of ROI volume in the AD-Memory group.

AD-Language vs. AD-Memory					
p_min = 1					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Language group, as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume
Left posterior cingulate gyrus	2532.1	34.6	1.4	-2.6	-1.2
Left entorhinal cortex	1521.9	17.7	1.2	-5.9	-4.7
Right entorhinal cortex	1508.8	17.2	1.1	-5.2	-4.0
Left caudal anterior cingulate gyrus	1491.9	12.1	0.8	-1.4	-0.6
Right caudal middle frontal gyrus	4886.0	-46.1	-0.9	-0.9	-1.8
Left parahippocampal gyrus	1677.4	-15.8	-0.9	-1.9	-2.9
Right parahippocampal gyrus	1596.6	9.4	0.6	-3.1	-2.5
Right frontal pole	987.4	-6.0	-0.6	-0.6	-1.2

Table 6.3a : Differences in rates of change of ROI volume between AD-Language and AD-Memory groups for selected ROIs of importance. ROIs in the top 5 rankings for p_min = 1 are highlighted in blue.

AD-Language vs. AD-Memory					
p_min = 2					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Language group, as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume
Left posterior cingulate gyrus	2532.1	39.9	1.6	-2.8	-1.2
Left entorhinal cortex	1521.9	16.9	1.1	-5.8	-4.7
Right entorhinal cortex	1508.8	15.9	1.1	-5.1	-4.0
Left caudal anterior cingulate gyrus	1491.9	16.2	1.1	-1.7	-0.6
Right caudal middle frontal gyrus	4886.0	-47.4	-1.0	-0.9	-1.9
Left parahippocampal gyrus	1677.4	-12.5	-0.7	-2.1	-2.9
Right parahippocampal gyrus	1596.6	11.0	0.7	-3.2	-2.5
Right frontal pole	987.4	-4.3	-0.4	-0.8	-1.2

Table 6.3b: Differences in rates of change of ROI volume between AD-Language and AD-Memory groups for selected ROIs of importance. The top 5 ROIs for p_min = 2 are highlighted in blue.

AD-Language vs. AD-Memory					
p_min = 3					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Language group, as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume
Left posterior cingulate gyrus	2532.1	38.9	1.5	-2.7	-1.2
Left entorhinal cortex	1521.9	21.4	1.4	-6.2	-4.8
Right entorhinal cortex	1508.8	17.2	1.1	-5.2	-4.0
Left caudal anterior cingulate gyrus	1491.9	15.3	1.0	-1.6	-0.6

Right caudal middle frontal gyrus	4886.0	-42.4	-0.9	-1.0	-1.9
Left parahippocampal gyrus	1677.4	-12.9	-0.8	-2.2	-2.9
Right parahippocampal gyrus	1596.6	12.1	0.8	-3.3	-2.5
Right frontal pole	987.4	-5.7	-0.6	-0.6	-1.2

Table 6.3c : Differences in rates of change of ROI volume between AD-Language and AD-Memory groups for selected ROIs of importance. The top 5 ROIs for $p_{\min} = 3$ are highlighted in blue.

AD-Language vs. AD-Memory					
$p_{\min} = 4$					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Language group, as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume
Left posterior cingulate gyrus	2532.1	37.9	1.5	-2.7	-1.2
Left entorhinal cortex	1521.9	20.5	1.3	-6.1	-4.8
Right entorhinal cortex	1508.8	16.3	1.1	-5.1	-4.0
Left caudal anterior cingulate gyrus	1491.9	14.7	1.0	-1.6	-0.6
Right caudal middle frontal gyrus	4886.0	-41.6	-0.9	-1.0	-1.9
Left parahippocampal gyrus	1677.4	-13.8	-0.8	-2.1	-2.9
Right parahippocampal gyrus	1596.6	11.5	0.7	-3.2	-2.5
Right frontal pole	987.4	-6.4	-0.6	-0.6	-1.2

Table 6.3d : Differences in rates of change of ROI volume between AD-Language and AD-Memory groups for selected ROIs of importance. The top 5 ROIs for $p_{\min} = 4$ are highlighted in blue.

AD-Language vs. AD-Memory					
p_min = 5					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Language group, as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume
Left posterior cingulate gyrus	2532.1	37.0	1.5	-2.7	-1.2
Left entorhinal cortex	1521.9	37.3	2.4	-7.3	-4.8
Right entorhinal cortex	1508.8	13.8	0.9	-4.9	-4.0
Left caudal anterior cingulate gyrus	1491.9	17.9	1.2	-1.7	-0.5
Right caudal middle frontal gyrus	4886.0	-30.9	-0.6	-1.2	-1.8
Left parahippocampal gyrus	1677.4	-15.5	-0.9	-2.1	-3.0
Right parahippocampal gyrus	1596.6	8.9	0.6	-3.0	-2.5
Right frontal pole	987.4	-9.5	-1.0	-0.3	-1.3

Table 6.3e : Differences in rates of change of ROI volume between AD-Language and AD-Memory groups for selected ROIs of importance. The top 5 ROIs for p_min = 5 are highlighted in blue.

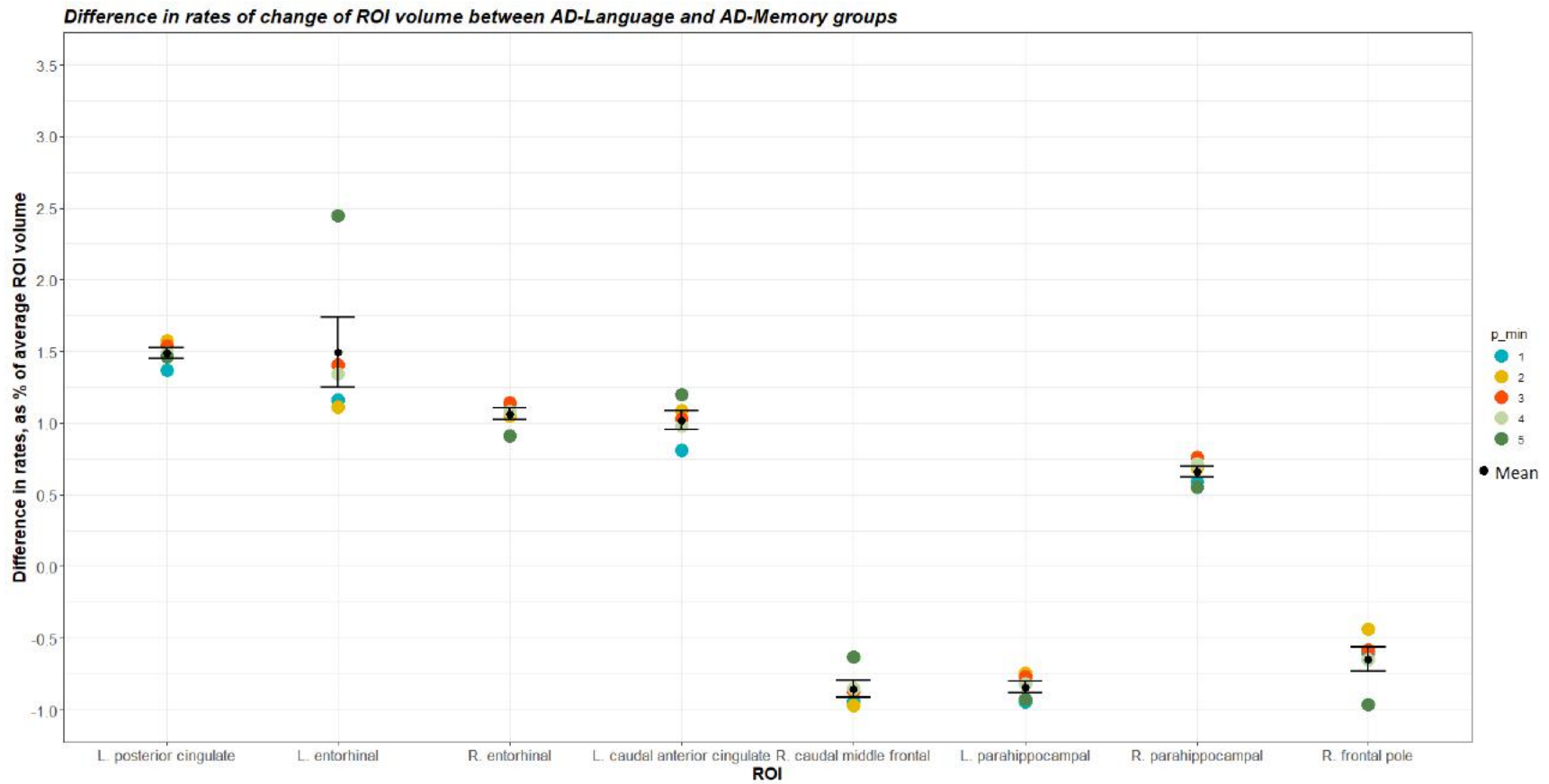


Figure 6.1: Results from β_{10} coefficients in linear mixed effects models for different p_min datasets: Difference in rates of change of ROI volume, as a % of average ROI volume. The mean of the differences across the different p_min datasets is shown in black, along with error bars (based on standard error of the mean).

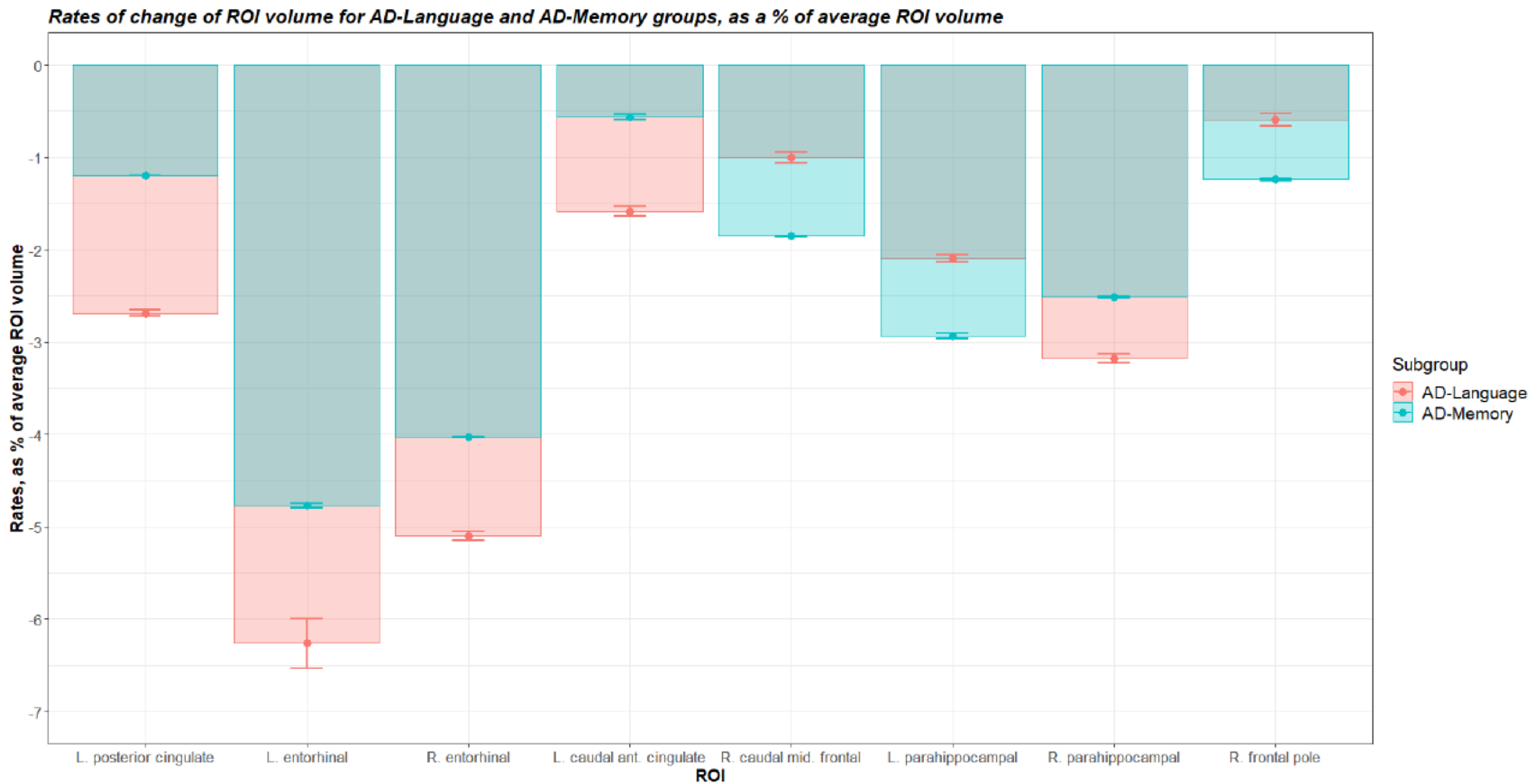


Figure 6.2: A comparison of the rates of change of ROI volume for the AD-Language and AD-Memory groups. The height of the bars represents the mean rate of change for a given group from the analyses of the different p_{min} (1, 2, 3, 4, 5) datasets. Mean values are also plotted as points with error bars based on standard error of the mean from the five p_{min} datasets. A more negative rate indicates a faster decline of ROI volume. This figure delivers the following key points: 1. Which AD subgroup has a faster rate of decline for the different ROIs, 2. The absolute difference in rates between the groups, indicated by the non-overlapping part of the bars for the two subgroups and 3. How much this difference is in proportion to the rate of change of each of the subgroups.

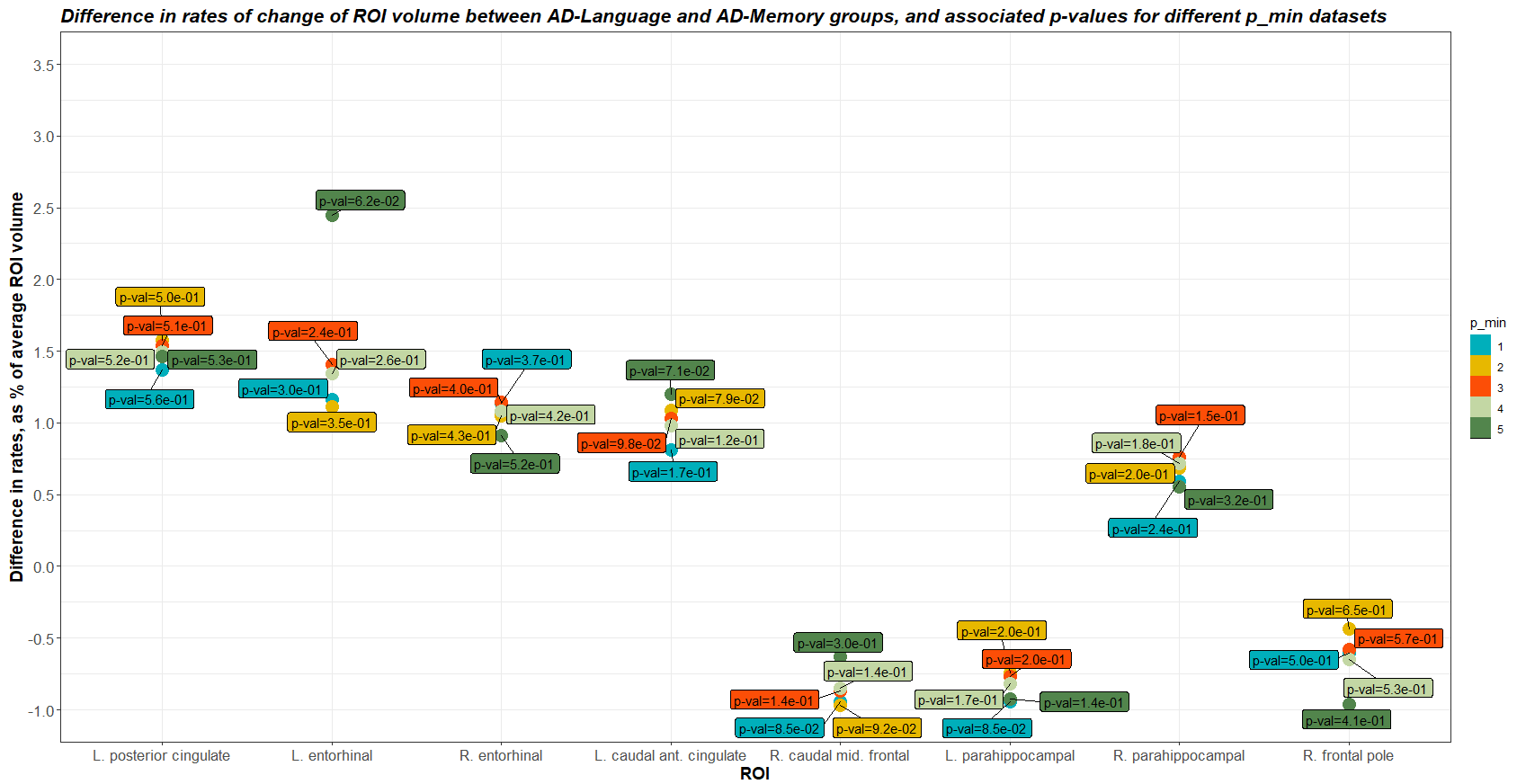


Figure 6.3: p-values associated with each estimate of the difference in rates between AD-Language and AD-Memory groups, from linear mixed effects modeling analysis of different p_min datasets.

6.3.1.2 Differences in average ROI volume at the time of AD diagnosis

Although the intent in modeling ROI volume using LME modeling on longitudinal data is to understand the differences in average dynamics over time for the respective AD subgroups being compared, it can also be used to learn about the AD subgroup differences in ROI volume at the time of AD diagnosis. As discussed in section 6.2, this information can be interpreted from the β_2 coefficient in each ROI volume model. Note that although the interpretation of β_2 is the average difference in ROI volumes at $t=0$ between the two AD subgroups, it is an estimate based on longitudinal data and not just data at $t=0$ (time of AD diagnosis). This is a difference between the results from cross-sectional data where the differences between AD subgroups were based strictly on data from the time of AD diagnosis and the longitudinal data analysis results referring to $t=0$.

Results based on the β_2 coefficient are summarized in Tables 6.4a-6.4e for the ROIs that were determined to have the largest β_2 values relative to ROI size (average ROI volume). Results from all p_min datasets were in close agreement in terms of which ROIs show the greatest differences in ROI volumes at $t=0$ between the AD-Language and AD-Memory. The greatest difference as estimated by the LME models was for the right entorhinal cortex in all p_min datasets. For the case of p_min = 1, right hippocampus showed the next greatest difference, while for p_min = 2, 3, 4 and 5, left temporal pole showed the next greatest difference. For all p_min datasets, left inferior temporal gyrus and left middle temporal gyrus were in 4th and 5th rank respectively for the differences in ROI volume at $t=0$ between the two AD subgroups.

LME models for all p-min datasets showed that **right entorhinal cortex** and **right hippocampus** have a negative β_2 coefficient, hence **smaller volume on average in the AD-Memory group** compared to the AD-Language group. **Left temporal pole, left inferior**

temporal gyrus and **left middle temporal gyrus** had a positive β_2 coefficient which indicates that these ROIs had a **smaller volume on average in the AD-Language group** compared to the AD-Memory group. It is worth noting that the right entorhinal cortex and the right hippocampus also appeared as the top two ROIs in variable importance results from cross-sectional data analysis for AD-Language vs. AD-Memory using random forest. Violin plot distributions of these two ROIs in cross-sectional data analysis had shown that the AD-Memory group has smaller volumes in these ROIs. This is consistent with the finding above based on the interpretation of the sign of the β_2 coefficients for these ROIs.

Note that p-values for the differences for the noted ROIs at $t=0$ are smaller than the p-values seen in the case of the analysis of differences in rates of change. But they are still above the threshold determined by the Holm-Bonferroni correction for multiple hypotheses testing. The right hippocampus, however, has p-values on the order of 10^{-4} across all p_{\min} datasets, which is on the same order as the Holm-Bonferroni corrected threshold for the top-ranking p-values. This could suggest that the right hippocampus is indeed a very important ROI to be considered for AD-Language vs. AD-Memory differences in ROI volume at the time of AD diagnosis, as it has a fairly small p-value associated with the differences even with the current small sample sizes.

Of the top ROIs for differences between AD-Language and AD-Memory in β_2 (this subsection) and β_{10} (section 6.3.1.1), the right entorhinal cortex is the only region that appeared in both lists. It is interesting to note that as per the LME models, while the right entorhinal cortex had a faster decline over time in the AD-Language group, it had a lower volume in the AD-Memory group at $t=0$.

The variation in differences of ROI volume at $t=0$ for the two subgroups across the different p_{\min} datasets are summarized in Table 6.5.

Tables 6.4a – 6.4e: The ROIs shown here are a collective list based on the top 5 ROIs for AD-Language vs. AD-Memory from the LME analysis of each p_min dataset based on the magnitude of the β_2 coefficient, normalized by average ROI volume ('Difference in rates as % of Avg ROI volume' column in the table). Individuals with AD* indicated in the second column of each table are individuals from the 4 single domain groups: AD-Executive, AD-Language, AD-Memory and AD-Visuospatial ROIs with a positive β_2 (greater volume in the AD-Memory group at t=0) are shown in pink while ROIs with a negative β_2 (smaller volume in the AD-Memory group at t=0) are shown in green.

AD-Language vs. AD-Memory				
p_min = 1				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right entorhinal cortex	1508.8	-129.0	-8.5	7.0E-02
Left temporal pole	2053.0	164.2	8.0	1.5E-02
Right hippocampus	2982.8	-244.7	-8.2	2.0E-04
Left inferior temporal gyrus	8517.7	575.5	6.8	7.5E-03
Left middle temporal gyrus	8130.4	547.0	6.7	3.0E-03

Table 6.4a: Differences in ROI volume at t=0 between AD-Language and AD-Memory groups for selected ROIs of importance, for the dataset with p_min = 1. See note above “Tables 6.4a-e” for color coding.

AD-Language vs. AD-Memory				
p_min = 2				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right entorhinal cortex	1508.8	-129.2	-8.6	7.0E-02
Left temporal pole	2053.0	170.9	8.3	1.1E-02
Right hippocampus	2982.8	-243.3	-8.2	2.1E-04
Left inferior temporal gyrus	8517.7	569.6	6.7	8.2E-03
Left middle temporal gyrus	8130.4	541.8	6.7	3.4E-03

Table 6.4b: Differences in ROI volume at t=0 between AD-Language and AD-Memory groups for selected ROIs of importance, for the dataset with p_min = 2. See note above “Tables 6.4a-e” for color coding.

AD-Language vs. AD-Memory				
p_min = 3				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm^3)	β_2 (Difference in ROI volumes at t=0) (mm^3)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right entorhinal cortex	1508.8	-130.1	-8.6	6.8E-02
Left temporal pole	2053.0	169.4	8.3	1.1E-02
Right hippocampus	2982.8	-241.1	-8.1	2.4E-04
Left inferior temporal gyrus	8517.7	569.2	6.7	8.2E-03
Left middle temporal gyrus	8130.4	543.2	6.7	3.3E-03

Table 6.4c: Differences in ROI volume at t=0 between AD-Language and AD-Memory groups for selected ROIs of importance, for the dataset with p_min = 3. See note above “Tables 6.4a-e” for color coding.

AD-Language vs. AD-Memory				
p_min = 4				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm^3)	β_2 (Difference in ROI volumes at t=0) (mm^3)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right entorhinal cortex	1508.8	-129.8	-8.6	6.9E-02
Left temporal pole	2053.0	169.5	8.3	1.1E-02
Right hippocampus	2982.8	-240.6	-8.1	2.5E-04
Left inferior temporal gyrus	8517.7	569.0	6.7	8.3E-03
Left middle temporal gyrus	8130.4	543.0	6.7	3.2E-03

Table 6.4d: Differences in ROI volume at t=0 between AD-Language and AD-Memory groups for selected ROIs of importance, for the dataset with p_min = 4. See note above “Tables 6.4a-e” for color coding.

AD-Language vs. AD-Memory				
p_min = 5				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right entorhinal cortex	1508.8	-129.4	-8.6	7.0E-02
Left temporal pole	2053.0	170.3	8.3	1.1E-02
Right hippocampus	2982.8	-241.2	-8.1	2.4E-04
Left inferior temporal gyrus	8517.7	569.0	6.7	8.2E-03
Left middle temporal gyrus	8130.4	540.9	6.7	3.4E-03

Table 6.4e: Differences in ROI volume at t=0 between AD-Language and AD-Memory groups for selected ROIs of importance, for the dataset with p_min = 5. See note above “Tables 6.4a-e” for color coding.

AD-Language vs. AD-Memory		
ROI	Mean difference in ROI vol at t=0 (across different p_min datasets), as % of average ROI volume	Std. Error of mean
Right entorhinal cortex	-8.58	0.01
Left temporal pole	8.23	0.06
Right hippocampus	-8.12	0.03
Left inferior temporal gyrus	6.70	0.01
Left middle temporal gyrus	6.68	0.01

Table 6.5: Variation in the difference in ROI volume at t=0 across different p_min datasets for AD-Language vs. AD-Memory analysis. Mean and standard error of the mean are shown for each ROI.

The p-values for β_2 seem to differ less across the different p_min datasets' LME models compared to the p-values for β_{10} for the ROIs that were determined to be important in each case. This may suggest that the β_2 estimates are more stable than the β_{10} estimates across different p_min cases. Additionally, a larger number of ROIs in the β_2 based ROI importance list have smaller p-values compared to the β_{10} ROI importance list. Although current sample size is not sufficient to make conclusive statements based on p-value comparisons, one may expect that the differences in ROI volumes between the AD subgroups at t=0 may be more significant than the differences in rates of change of ROI volume between the groups.

6.3.2 AD-Memory vs. AD-Visuospatial

In the next comparison considered through LME models, AD-Memory vs. AD-Visuospatial, the following ROIs were in the top five ROIs for different p_min datasets for the analysis of rate of change difference between the AD subgroups: **left temporal pole, right temporal pole, left entorhinal cortex, right entorhinal cortex, left rostral anterior cingulate gyrus, left frontal pole and left parahippocampal gyrus**. These rankings were based on magnitudes of normalized β_{10} coefficients (difference in rate of change of ROI volume as a percentage of average ROI volume) from all 70 ROI LME models. The rankings of the ROIs for each of the p_min datasets is shown in Table 6.6. The top five ROIs corresponding to the highest magnitudes of β_2 relative to the respective average ROI volumes were **left entorhinal cortex, right entorhinal cortex, left parahippocampal gyrus, left temporal pole and left hippocampus** across all p_min datasets. The details of the results based on the β_{10} and β_2 coefficients are discussed in the next two subsections respectively.

Dataset (p_min)	Top 5 ROIs based on β_{10} coefficient (normalized by average ROI volume in AD)
1	1. Left temporal pole, 2. Left entorhinal, 3. Right temporal pole, 4. Right entorhinal, 5. Left rostral anterior cingulate
2	1. Left temporal pole, 2. Right temporal pole, 3. Left entorhinal, 4. Right entorhinal, 5. Left Rostral anterior cingulate
3	1. Right temporal pole, 2. Left temporal pole, 3. Left entorhinal, 4. Right entorhinal, 5. Left frontal pole
4	1. Right entorhinal, 2. Left temporal pole, 3. Left entorhinal, 4. Right temporal, 5. Left parahippocampal
5	1. Left temporal, 2. Right entorhinal, 3. Right temporal, 4. Left entorhinal, 5. Left parahippocampal

Table 6.6: Top 5 ROIs based on differences in rates of change of ROI volume (given by β_{10}) between the AD-Memory and AD-Visuospatial groups.

ROI rankings for AD-Memory vs. AD-Visuospatial based on differences in longitudinal volumes	
Dataset (p_min)	Top 5 ROIs based on β_2 coefficient (normalized by average ROI volume in AD)
1, 2, 3, 4, 5	1. Left entorhinal, 2. Right entorhinal, 3. Left parahippocampal, 4. Left temporal pole, 5. Left hippocampus

Table 6.7: Top 5 ROIs based on differences in longitudinal volumes (given by β_2) between the AD-Memory and AD-Visuospatial groups.

6.3.2.1 Differences in rates of change of ROI volume

The datasets corresponding to $p_{\min} = 1, 2$ and 3 had the same ROIs in the list of top 5 ROIs based on β_{10} . For datasets corresponding to $p_{\min} = 4$ and 5 , the list of top 5 ROIs was different by one ROI compared to the list for $p_{\min} = 1, 2$ and 3 . Overall, there was a great overlap in the number of ROIs that appeared as the top five ROIs from the analyses based on different datasets corresponding to different p_{\min} values. Tables 6.8a-e and Figures 6.4-6.5 show how the differences in the rate of change of ROI volume differed across the analyses of the different datasets for AD-Memory vs. AD-Visuospatial for the seven chosen ROIs as per the list above. The p-values associated with the β_{10} coefficients are shown in Figure 6.6.

In the LME models based on the data from the AD-Memory and AD-Visuospatial groups, the ADSubgroup variable was coded as 0 for AD-Memory and 1 for AD-Visuospatial. Hence, a

negative β_{10} coefficient here indicates a more negative rate of change (faster decline) of ROI volume in the AD-Visuospatial group compared to the AD-Memory group. Based on this, it can be seen from Tables 6.8a-e and Figures 6.5 and 6.6 that **individuals in the AD-Visuospatial group, on average, decline faster in their volume for the left frontal pole** compared to the AD-Memory group while for the other ROIs (**right temporal pole, left entorhinal cortex, right entorhinal cortex, left rostral anterior cingulate gyrus, left frontal pole and left parahippocampal gyrus**), **individuals in the AD-Memory group decline faster** on average. Figure 6.5 illustrates that **left temporal pole, right temporal pole, left rostral anterior cingulate and left frontal pole** have the most substantial differences in rates between the two AD subgroups, relative to the rates of each of the subgroups.

From Figure 6.6, one can see that the p-values associated with the β_{10} coefficient from the LME models for a given ROI across the different p_min datasets are not always stable or small enough. This behavior of the p-values corroborates the decision to not rely on p-values for determining ROIs that show the most differences across the AD subgroups of interest. A more enriched dataset both in the number of individuals and the number of time points for each individual is needed to extend the current analysis to an inference based analysis using p-values.

Tables 6.8a – 6.8e: The ROIs shown here are a collective list based on the top 5 ROIs for AD-Memory vs. AD-Visuospatial from the LME analysis of each p_min dataset based on the magnitude of the β_{10} coefficient, normalized by average ROI volume ('Difference in rates as % of Avg ROI volume' column in the table). Individuals with AD* indicated in the second column of each table are individuals from the 4 single domain groups: AD-Executive, AD-Language, AD-Memory and AD-Visuospatial. Positive differences in rates correspond to a greater rate of decline of ROI volume in the AD-Memory group compared to the AD-Memory group. Negative differences in rates of change correspond to a greater rate of decline of ROI volume in the AD-Visuospatial group.

AD-Memory vs. AD-Visuospatial					
p_min = 1					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left temporal pole	2053.0	32.3	1.6	-2.7	-1.1
Right temporal pole	2098.4	25.4	1.2	-2.7	-1.5
Left entorhinal cortex	1521.9	20.0	1.3	-4.5	-3.2
Right entorhinal cortex	1508.8	14.6	1.0	-4.0	-3.1
Left rostral anterior cingulate gyrus	2249.1	18.1	0.8	-1.7	-0.9
Left frontal pole	809.2	-5.2	-0.6	-0.8	-1.4
Left parahippocampal gyrus	1677.4	12.8	0.8	-3.0	-2.2

Table 6.8a: Differences in rates of change of ROI volume between AD-Memory and AD-Visuospatial groups for selected ROIs of importance. The top 5 ROIs for p_min = 1 are highlighted in blue.

AD-Memory vs. AD-Visuospatial					
p_min =2					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left temporal pole	2053.0	32.2	1.6	-2.7	-1.2
Right temporal pole	2098.4	30.6	1.5	-2.7	-1.3
Left entorhinal cortex	1521.9	18.9	1.2	-4.5	-3.3
Right entorhinal cortex	1508.8	15.1	1.0	-4.0	-3.0
Left rostral anterior cingulate gyrus	2249.1	18.6	0.8	-1.7	-0.8
Left frontal pole	809.2	-5.2	-0.6	-0.8	-1.4
Left parahippocampal gyrus	1677.4	11.2	0.7	-3.0	-2.3

Table 6.8b: Differences in rates of change of ROI volume between AD-Memory and AD-Visuospatial groups for selected ROIs of importance. The top 5 ROIs for p_min = 2 are highlighted in blue.

AD-Memory vs. AD-Visuospatial					
p_min =3					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left temporal pole	2053.0	26.2	1.3	-2.7	-1.5
Right temporal pole	2098.4	29.6	1.4	-2.7	-1.3
Left entorhinal cortex	1521.9	17.0	1.1	-4.6	-3.5
Right entorhinal cortex	1508.8	16.4	1.1	-4.1	-3.0
Left rostral anterior cingulate gyrus	2249.1	16.9	0.8	-1.7	-0.9
Left frontal pole	809.2	-6.5	-0.8	-0.8	-1.6
Left parahippocampal gyrus	1677.4	12.6	0.8	-3.0	-2.2

Table 6.8c: Differences in rates of change of ROI volume between AD-Memory and AD-Visuospatial groups for selected ROIs of importance. The top 5 ROIs for p_min = 3 are highlighted in blue.

AD-Memory vs. AD-Visuospatial					
p_min = 4					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left temporal pole	2053.0	27.6	1.3	-2.8	-1.4
Right temporal pole	2098.4	26.0	1.2	-2.7	-1.5
Left entorhinal cortex	1521.9	19.5	1.3	-4.6	-3.4
Right entorhinal cortex	1508.8	20.6	1.4	-4.1	-2.7
Left rostral anterior cingulate gyrus	2249.1	16.5	0.7	-1.7	-0.9
Left frontal pole	809.2	-6.2	-0.8	-0.8	-1.6
Left parahippocampal gyrus	1677.4	14.9	0.9	-3.0	-2.1

Table 6.8d: Differences in rates of change of ROI volume between AD-Memory and AD-Visuospatial groups for selected ROIs of importance. The top 5 ROIs for p_min = 4 are highlighted in blue.

AD-Memory vs. AD-Visuospatial					
p_min = 5					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left temporal pole	2053.0	28.2	1.4	-2.9	-1.5
Right temporal pole	2098.4	25.7	1.2	-2.7	-1.5
Left entorhinal cortex	1521.9	18.3	1.2	-4.7	-3.5
Right entorhinal cortex	1508.8	20.6	1.4	-4.0	-2.7
Left rostral anterior cingulate gyrus	2249.1	13.1	0.6	-1.5	-0.9
Left frontal pole	809.2	-5.9	-0.7	-0.9	-1.6
Left parahippocampal gyrus	1677.4	14.4	0.9	-3.1	-2.2

Table 6.8e: Differences in rates of change of ROI volume between AD-Memory and AD-Visuospatial groups for selected ROIs of importance. The top 5 ROIs for p_min = 5 are highlighted in blue.

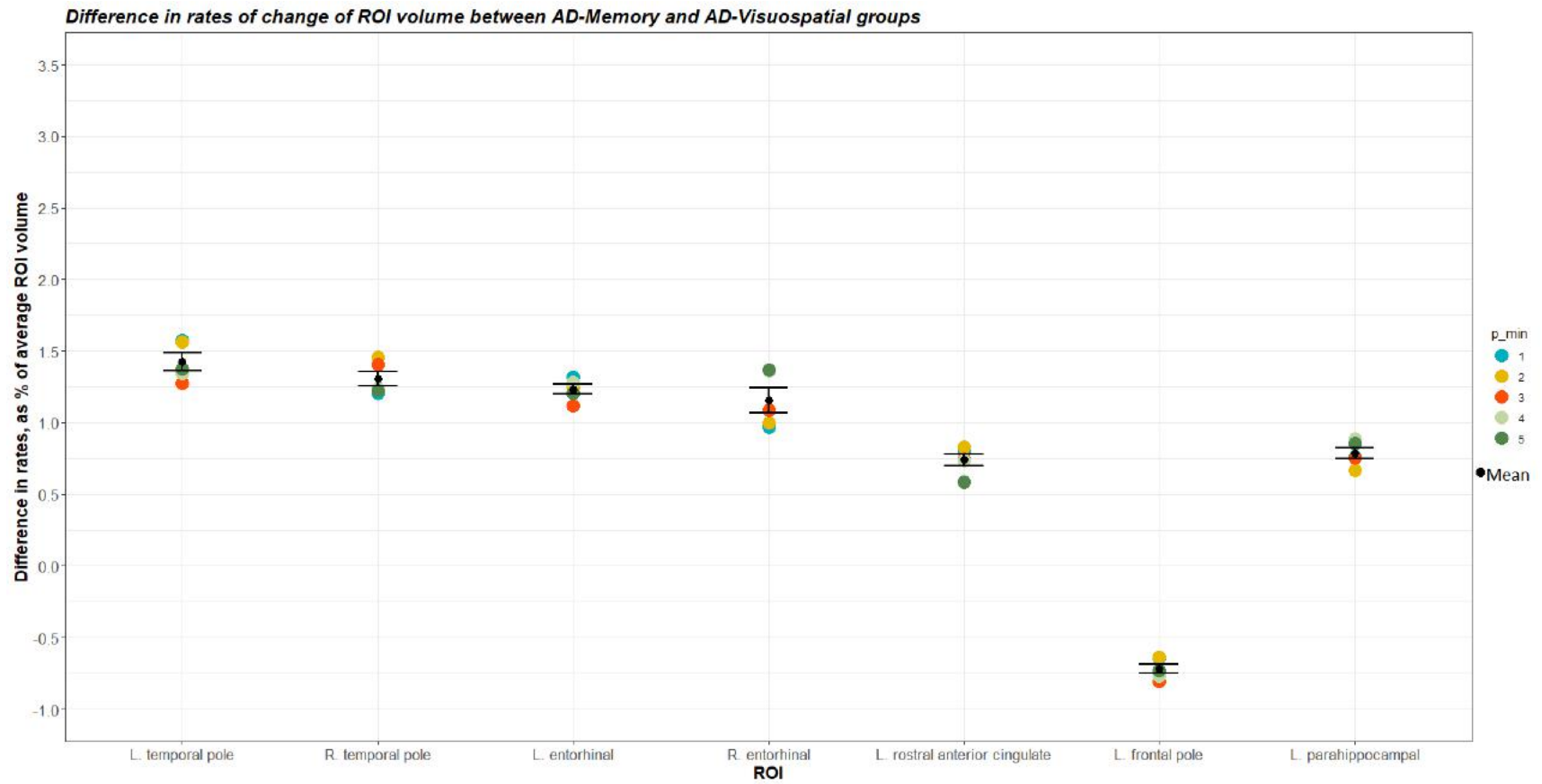


Figure 6.4: Results from β_{10} coefficients in linear mixed effects models for different p_min datasets: Difference in rates of change of ROI volume, as a % of average ROI volume. The mean of the differences across the different p_min datasets is shown in black, along with error bars (based on standard error of the mean).

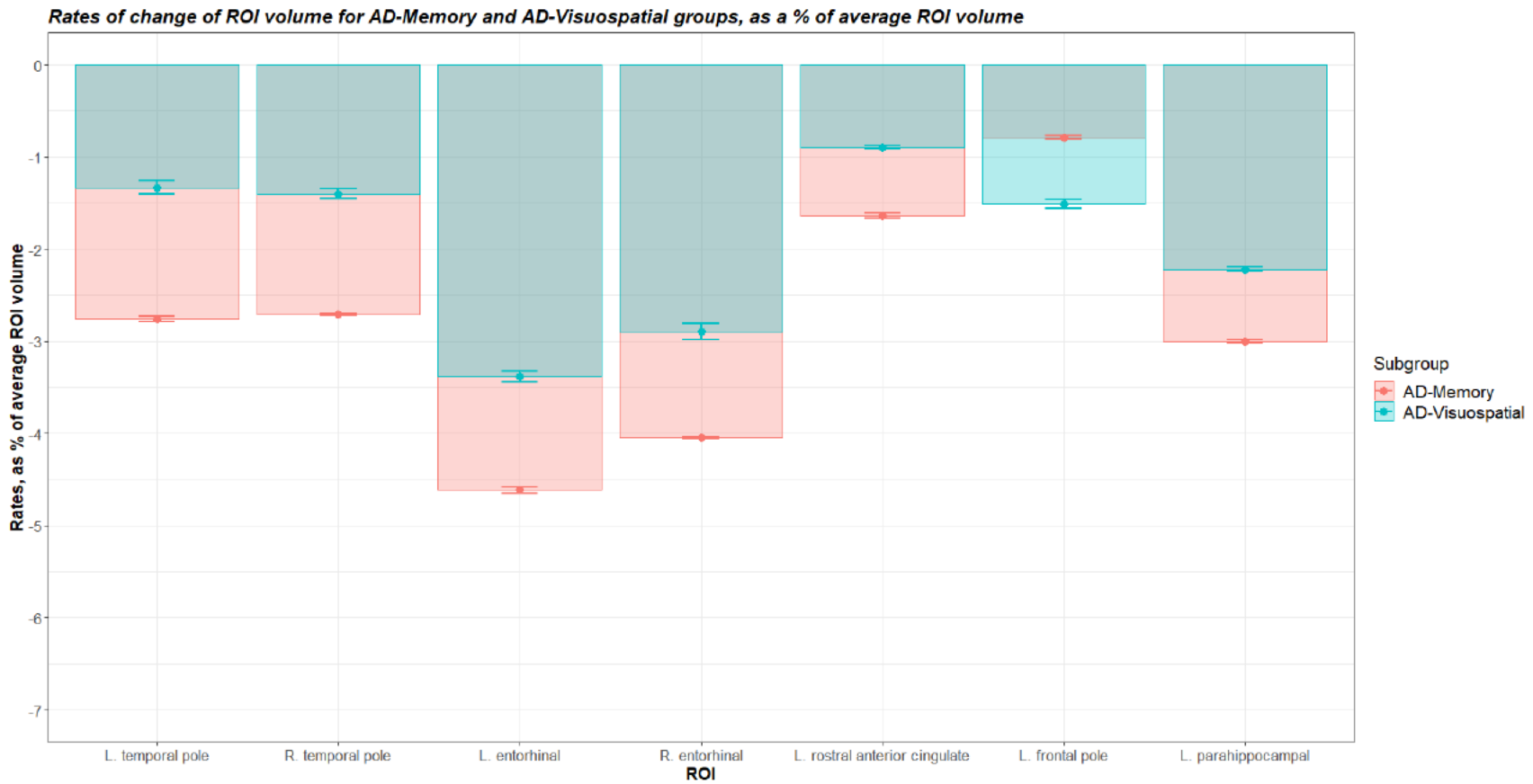


Figure 6.5: A comparison of the rates of change of ROI volume for the AD-Memory and AD-Visuospatial groups. The height of the bars represents the mean rate of change for a given group from the analyses of the different p_{min} (1, 2, 3, 4, 5) datasets. Mean values are also plotted as points with error bars based on standard error of the mean from the five p_{min} datasets. A more negative rate indicates a faster decline of ROI volume. This figure delivers the following key points: 1. Which AD subgroup has a faster rate of decline for the different ROIs, 2. The absolute difference in rates between the groups, indicated by the non-overlapping part of the bars for the two subgroups and 3. How much this difference is in proportion to the rate of change of each of the subgroups.

Difference in rates of change of ROI volume between AD-Memory and AD-Visuospatial groups, and associated p-values for different p_min datasets

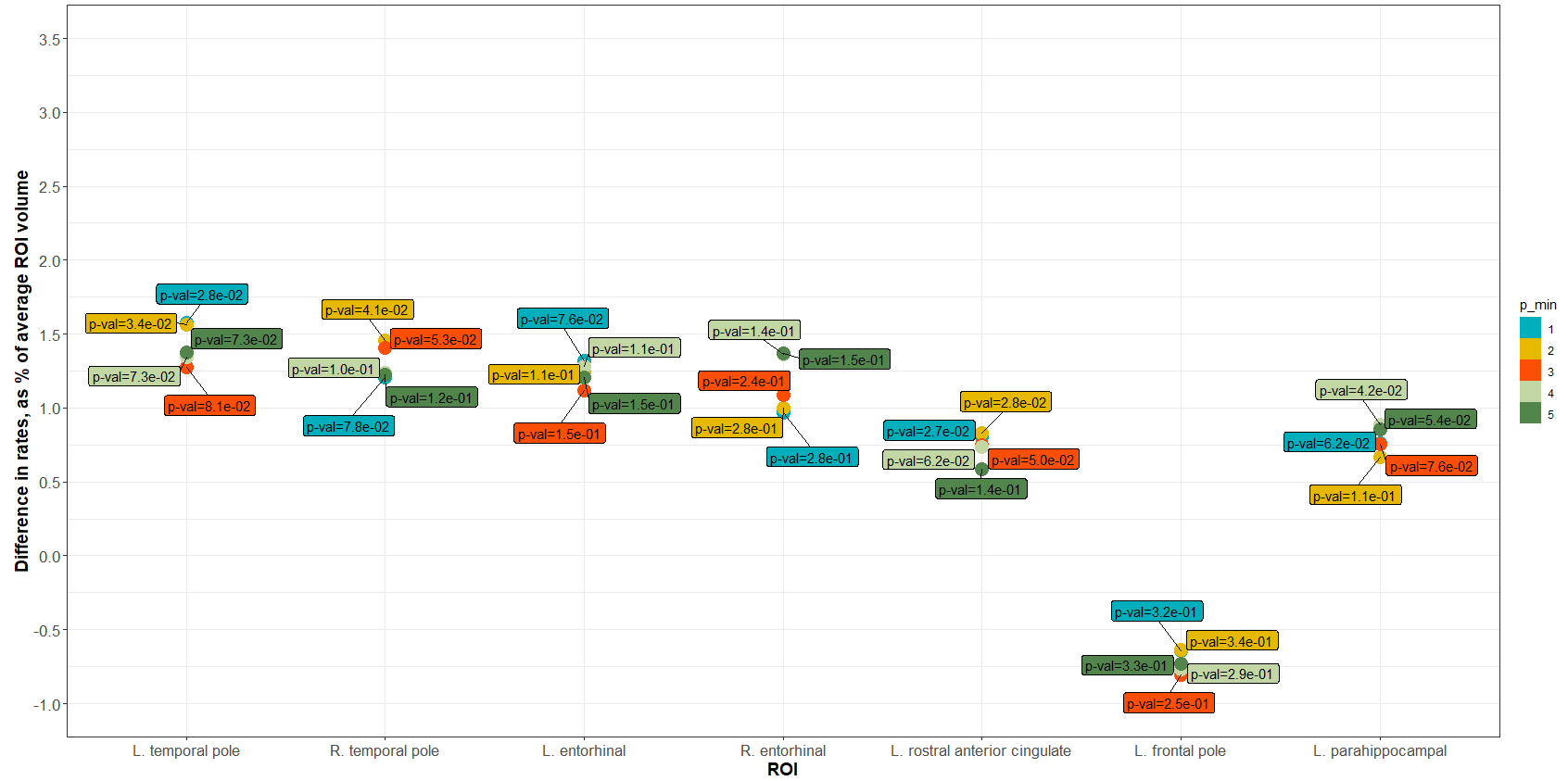


Figure 6.6: P-values associated with each estimate of the difference in rates between AD-Memory and AD-Visuospatial groups, from linear mixed effects modeling analysis of different p_min datasets.

6.3.2.2 Differences in average ROI volume at the time of AD diagnosis

In terms of the β_2 coefficient, which represents the average difference in ROI volumes between the two AD subgroups at $t=0$, the following ROIs appeared in the top 5 ROIs based on the results from datasets corresponding to different p_{\min} values for AD-Memory vs. AD-Visuospatial: **left entorhinal cortex, right entorhinal cortex, left parahippocampal gyrus, left temporal pole and left hippocampus**. All five p_{\min} datasets had the same order of ranking for these ROIs. The β_2 coefficients from these ROI models were all positive (See Tables 6.9a-e), which indicates that on average, the AD-Visuospatial group had a greater volume in these ROIs at the time of diagnosis compared to the AD-Memory group. Equivalently, the LME models suggest that **these ROIs had a smaller volume at $t=0$ (time of AD diagnosis) for the AD-Memory group compared to the AD-Visuospatial group**. These results seem to be consistent with the known role of the entorhinal cortex, the hippocampus and the parahippocampal gyrus in the memory function (Izquierdo and Medina 1993), (Takehara-Nishiuchi 2014), (Parkin 1996) and (Ward et al. 2013). As seen in section 6.3.2.1, left entorhinal cortex, right entorhinal cortex, left parahippocampal gyrus and left temporal pole are ROIs that were also determined to be important based on differences in rate of change of ROI volume between the two AD subgroups. In cross-sectional data analysis for AD-Memory vs. AD-Visuospatial, left entorhinal cortex and right entorhinal cortex were the top two regions of importance while the left parahippocampal gyrus was the seventh most important region. Once again, it is assuring to see some overlap between ROIs that were determined to be important in cross-sectional data analysis and longitudinal data analysis carried out using independent methods.

It is worth noting that the p-values associated with β_2 for the left hippocampus volume model are on the order of 10^{-4} for all p_{\min} datasets. The Holm-Bonferroni corrected threshold for

significance (with a familywise error rate of 0.05) was determined to be in the range of 1.43×10^{-4} to 1.45×10^{-4} for the smallest five p-values, for a correction of 350 hypotheses tests (see section 6.2). While the p-values associated with β_2 for the left hippocampus volume model are still above the Holm-Bonferroni corrected thresholds, they are close to the threshold. This indicates that the left hippocampus differences in ROI volume at t=0 across AD-Memory and AD-Visuospatial might be worth paying attention to as the β_2 coefficient for this ROI's model for each p_min dataset has a very small p-value despite the current small sample sizes.

Tables 6.9a – 6.9e: The ROIs shown here are a collective list based on the top 5 ROIs for AD-Memory vs. AD-Visuospatial from the LME analysis of each p_min dataset based on the magnitude of the β_2 coefficient, normalized by average ROI volume ('Difference in rates as % of Avg ROI volume' column in the table). Individuals with AD* indicated in the second column of each table are individuals from the 4 single domain groups: AD-Executive, AD-Language, AD-Memory and AD-Visuospatial. All ROIs shown in the tables had a positive β_2 (smaller volume in the AD-Memory group at t=0).

AD-Memory vs. AD-Visuospatial				
p_min = 1				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Left entorhinal cortex	1521.9	161.6	10.6	2.4E-03
Right entorhinal cortex	1508.8	120.3	8.0	3.0E-02
Left parahippocampal gyrus	1677.4	103.1	6.1	2.0E-02
Left temporal pole	2053.0	123.7	6.0	2.1E-02
Left hippocampus	2854.8	163.9	5.7	5.7E-04

Table 6.9a: Differences in ROI volume at t=0 between AD-Memory and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 1. See note "Tables 6.9a-e" above for interpretation of the sign of β_2 .

AD-Memory vs. AD-Visuospatial				
p_min = 2				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Left entorhinal cortex	1521.9	161.9	10.6	2.4E-03
Right entorhinal cortex	1508.8	119.5	7.9	3.1E-02
Left parahippocampal gyrus	1677.4	103.3	6.2	1.9E-02
Left temporal pole	2053.0	124.1	6.0	2.1E-02
Left hippocampus	2854.8	163.5	5.7	5.8E-04

Table 6.9b: Differences in ROI volume at t=0 between AD-Memory and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 2. See note “Tables 6.9a-e” above for interpretation of the sign of β_2 .

AD-Memory vs. AD-Visuospatial				
p_min = 3				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Left entorhinal cortex	1521.9	159.2	10.5	2.6E-03
Right entorhinal cortex	1508.8	119.4	7.9	3.1E-02
Left parahippocampal gyrus	1677.4	103.7	6.2	1.9E-02
Left temporal pole	2053.0	120.1	5.8	2.5E-02
Left hippocampus	2854.8	162.6	5.7	6.3E-04

Table 6.9c: Differences in ROI volume at t=0 between AD-Memory and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 3. See note “Tables 6.9a-e” above for interpretation of the sign of β_2 .

AD-Memory vs. AD-Visuospatial				
p_min = 4				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Left entorhinal cortex	1521.9	160.4	10.5	2.5E-03
Right entorhinal cortex	1508.8	118.8	7.9	3.2E-02
Left parahippocampal gyrus	1677.4	104.5	6.2	1.8E-02
Left temporal pole	2053.0	122.3	6.0	2.3E-02
Left hippocampus	2854.8	162.8	5.7	6.1E-04

Table 6.9d: Differences in ROI volume at t=0 between AD-Memory and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 4. See note “Tables 6.9a-e” above for interpretation of the sign of β_2 .

AD-Memory vs. AD-Visuospatial				
p_min = 5				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Left entorhinal cortex	1521.9	158.9	10.4	2.7E-03
Right entorhinal cortex	1508.8	119.7	7.9	3.1E-02
Left parahippocampal gyrus	1677.4	103.3	6.2	1.9E-02
Left temporal pole	2053.0	121.2	5.9	2.4E-02
Left hippocampus	2854.8	167.3	5.9	4.6E-04

Table 6.9e: Differences in ROI volume at t=0 between AD-Memory and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 5. See note “Tables 6.9a-e” above for interpretation of the sign of β_2 .

AD-Memory vs. AD-Visuospatial		
ROI	Mean difference in ROI vol at t=0 (across different p_min datasets), as % of average ROI volume	Std. Error of mean
Left entorhinal cortex	10.54	0.04
Right entorhinal cortex	7.92	0.02
Left parahippocampal gyrus	6.17	0.01
Left temporal pole	5.96	0.04
Left hippocampus	5.75	0.03

Table 6.10: Variation in the difference in ROI volume at t=0 across the five p_min datasets for AD-Language vs. AD-Memory analysis. Mean and standard error of the mean are shown for each ROI.

6.3.3 AD-Language vs. AD-Visuospatial

The following ROIs were in the top 5 ROIs for the different p_min datasets, based on the normalized magnitude of the β_{10} coefficient from the LME models: **left entorhinal cortex, right**

entorhinal cortex, left temporal pole, right temporal pole gyrus, left caudal anterior cingulate gyrus, right inferior temporal gyrus and left rostral anterior cingulate gyrus. The top 5 ROIs in each p_min dataset's analysis are shown in Table 6.11. Although the order of the rankings of ROIs changes slightly and the coefficient values also differ over the different p_min datasets, there is a great overlap of ROIs found in the top 5 ROIs across the different datasets, which provides more confidence in the collective list of ROIs above. In terms of volume differences at t=0 between the two AD subgroups, the following ROIs were determined to be important based on the β_2 coefficient: **right temporal pole, left entorhinal cortex, left inferior temporal gyrus, left middle temporal gyrus, and right inferior parietal lobule.** These ROIs were the top 5 ROIs from the analyses of all datasets corresponding to the five p_min values; see Table 6.12. As it can be seen, the **left entorhinal cortex** is the common ROI that shows salient differences in both the rate of change of ROI volume between the two AD subgroups and ROI volume at t=0 across the two groups, according to the LME models. Details of the results based on β_{10} and β_2 coefficients are discussed in the next two subsection respectively.

ROI rankings for AD-Language vs. AD-Visuospatial based on differences in rate of change of volume	
Dataset (p_min)	Top 5 ROIs based on β_{10} coefficient (normalized by average ROI volume in AD)
1	1. Left entorhinal, 2. Right entorhinal, 3. Left Temporal pole, 4. Left caudal anterior cingulate, 5. Right inferior temporal
2	1. Left entorhinal, 2. Left temporal pole, 3. Right entorhinal, 4. Right temporal, 5. Right inferior temporal
3	1. Left entorhinal, 2. Right entorhinal, 3. Left temporal pole, 4. Right temporal pole, 5. Right inferior temporal
4	1. Right entorhinal, 2. Left entorhinal, 3. Left temporal pole, 4. Right inferior temporal, 5. Right temporal pole
5	1. Left entorhinal, 2. Right entorhinal, 3. Right inferior temporal, 5. Left caudal anterior cingulate, 5. Left rostral anterior cingulate

Table 6.11: Top 5 ROIs based on differences in rates of change of ROI volume (given by β_{10}) between the AD-Language and AD-Visuospatial groups.

ROI rankings for AD-Language vs. AD-Visuospatial based on differences in longitudinal volumes	
Dataset (p_min)	Top 5 ROIs based on β_2 coefficient (normalized by average ROI volume in AD)
1, 2, 3, 4, 5	1. Right temporal pole, 2. Left entorhinal, 3. Left inferior temporal, 4. Left middle temporal, 5. Right inferior parietal

Table 6.12: Top 5 ROIs based on differences in longitudinal volumes (given by β_2) between the AD-Language and AD-Visuospatial groups.

6.3.3.1 Differences in rates of change of ROI volume

The coefficient values for β_{10} (mm^3/year) and β_{10} as a percentage of average ROI volume are shown in Tables 6.13a-e for different p_min datasets. Figures 6.7 and 6.8 present this information visually. The associated p-values for the coefficients are shown in Figure 6.9. In these models, the AD-Language group was coded as “0” for the ADSubgroup variable and the AD-Visuospatial group was coded as “1.” Hence, a positive β_{10} value corresponds to a more positive rate of change of ROI volume in the AD-Visuospatial group compared to the AD-Language group, or equivalently, a more negative rate of change of ROI volume in the AD-Language group compared to the AD-Visuospatial group. In each of the ROIs noted above for differences in rate of change of ROI volume between the two AD subgroups (**left entorhinal cortex, right entorhinal cortex, left temporal pole, right temporal pole gyrus, left caudal anterior cingulate gyrus, right inferior temporal gyrus and left rostral anterior cingulate gyrus**), the β_{10} coefficient is positive, which indicates that **the AD-Language group had a greater rate of decline of ROI volume compared to the AD-Visuospatial group for these ROIs.** This is visually illustrated in Figure 6.8. Figure 6.8 also shows that the **most substantial differences in rates of change of ROI volume between the two AD subgroups, relative to the rates of each of the subgroups are for left caudal anterior cingulate gyrus, left temporal pole and right rostral anterior cingulate.**

P-values in the AD-Language vs. AD-Visuospatial analysis also seem to be sensitive to which dataset was used for LME modeling. As in the case of the AD-Language vs. AD-Memory and AD-Memory vs. AD-Visuospatial analyses for differences in rates of change for the subgroups, it is worth noting that even the smaller p-values obtained from this analysis are much larger than the Holm-Bonferroni corrected thresholds for significance. No major interpretations in the current analysis were made using p-values. This is anticipated as future work as more longitudinal data becomes available for these AD subgroups.

Tables 6.13a – 6.13e: The ROIs shown here are a collective list based on the top 5 ROIs for AD-Language vs. AD-Visuospatial from the LME analysis of each p_min dataset based on the magnitude of the β_{10} coefficient, normalized by average ROI volume ('Difference in rates as % of Avg ROI volume' column in the table). Individuals with AD* indicated in the second column of each table are individuals from the 4 single domain groups: AD-Executive, AD-Language, AD-Memory and AD-Visuospatial. Positive differences in rates correspond to a greater rate of decline of ROI volume in the AD-Language group compared to the AD-Visuospatial group. Negative differences in rates of change correspond to a greater rate of decline of ROI volume in the AD-Visuospatial group.

AD-Language vs. AD-Visuospatial					
p_min = 1					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left entorhinal cortex	1521.9	41.4	2.7	-6.9	-4.2
Right entorhinal cortex	1508.8	34.3	2.3	-6.0	-3.7
Left temporal pole	2053.0	31.2	1.5	-2.5	-1.0
Right temporal pole gyrus	2098.4	22.1	1.1	-3.1	-2.1
Left caudal anterior cingulate gyrus	1491.9	18.7	1.3	-1.1	0.1
Right inferior temporal gyrus	8408.3	103.3	1.2	-3.4	-2.2
Left rostral anterior cingulate gyrus	2249.1	16.0	0.7	-1.6	-0.9

Table 6.13a: Differences in rates of change of ROI volume between AD-Language and AD-Visuospatial groups for selected ROIs of importance. ROIs in the top 5 rankings for p_min = 1 are highlighted in blue.

AD-Language vs. AD-Visuospatial					
p_min = 2					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left entorhinal cortex	1521.9	38.9	2.6	-6.8	-4.3
Right entorhinal cortex	1508.8	35.1	2.3	-6.0	-3.7
Left temporal pole	2053.0	50.5	2.5	-3.3	-0.9
Right temporal pole gyrus	2098.4	34.9	1.7	-3.5	-1.8
Left caudal anterior cingulate gyrus	1491.9	19.8	1.3	-1.3	0.0
Right inferior temporal gyrus	8408.3	121.8	1.4	-3.6	-2.2
Left rostral anterior cingulate gyrus	2249.1	20.0	0.9	-1.7	-0.9

Table 6.13b: Differences in rates of change of ROI volume between AD-Language and AD-Visuospatial groups for selected ROIs of importance. ROIs in the top 5 rankings for p_min = 2 are highlighted in blue.

AD-Language vs. AD-Visuospatial					
p_min = 3					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left entorhinal cortex	1521.9	40.0	2.6	-6.8	-4.2
Right entorhinal cortex	1508.8	38.7	2.6	-6.2	-3.7
Left temporal pole	2053.0	36.7	1.8	-3.0	-1.2
Right temporal pole gyrus	2098.4	31.0	1.5	-3.3	-1.8
Left caudal anterior cingulate gyrus	1491.9	18.2	1.2	-1.3	-0.1
Right inferior temporal gyrus	8408.3	120.5	1.4	-3.6	-2.2
Left rostral anterior cingulate gyrus	2249.1	20.3	0.9	-1.8	-0.9

Table 6.13c: Differences in rates of change of ROI volume between AD-Language and AD-Visuospatial groups for selected ROIs of importance. ROIs in the top 5 rankings for p_min = 3 are highlighted in blue.

AD-Language vs. AD-Visuospatial					
p_min = 4					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left entorhinal cortex	1521.9	41.2	2.7	-6.8	-4.1
Right entorhinal cortex	1508.8	42.3	2.8	-6.1	-3.3
Left temporal pole	2053.0	35.3	1.7	-2.9	-1.2
Right temporal pole gyrus	2098.4	27.0	1.3	-3.3	-2.0
Left caudal anterior cingulate gyrus	1491.9	17.9	1.2	-1.2	0.0
Right inferior temporal gyrus	8408.3	114.4	1.4	-3.6	-2.2
Left rostral anterior cingulate gyrus	2249.1	19.1	0.8	-1.8	-0.9

Table 6.13d: Differences in rates of change of ROI volume between AD-Language and AD-Visuospatial groups for selected ROIs of importance. ROIs in the top 5 rankings for p_min = 4 are highlighted in blue.

AD-Language vs. AD-Visuospatial					
p_min = 5					
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	Difference in rates (mm ³ /year)	Difference in rates as % of Avg ROI volume	Rate of change for the AD-Memory group, as % of Avg ROI volume	Rate of change for the AD-Visuospatial group, as % of Avg ROI volume
Left entorhinal cortex	1521.9	53.6	3.5	-7.7	-4.1
Right entorhinal cortex	1508.8	41.2	2.7	-6.1	-3.3
Left temporal pole	2053.0	Model did not converge	NA	NA	NA
Right temporal pole gyrus	2098.4	11.8	0.6	-2.6	-2.1
Left caudal anterior cingulate gyrus	1491.9	20.1	1.4	-1.4	0.0
Right inferior temporal gyrus	8408.3	122.2	1.5	-3.7	-2.3
Left rostral anterior cingulate gyrus	2249.1	25.8	1.1	-2.1	-1.0

Table 6.13e: Differences in rates of change of ROI volume between AD-Language and AD-Visuospatial groups for selected ROIs of importance. ROIs in the top 5 rankings for p_min = 5 are highlighted in blue.

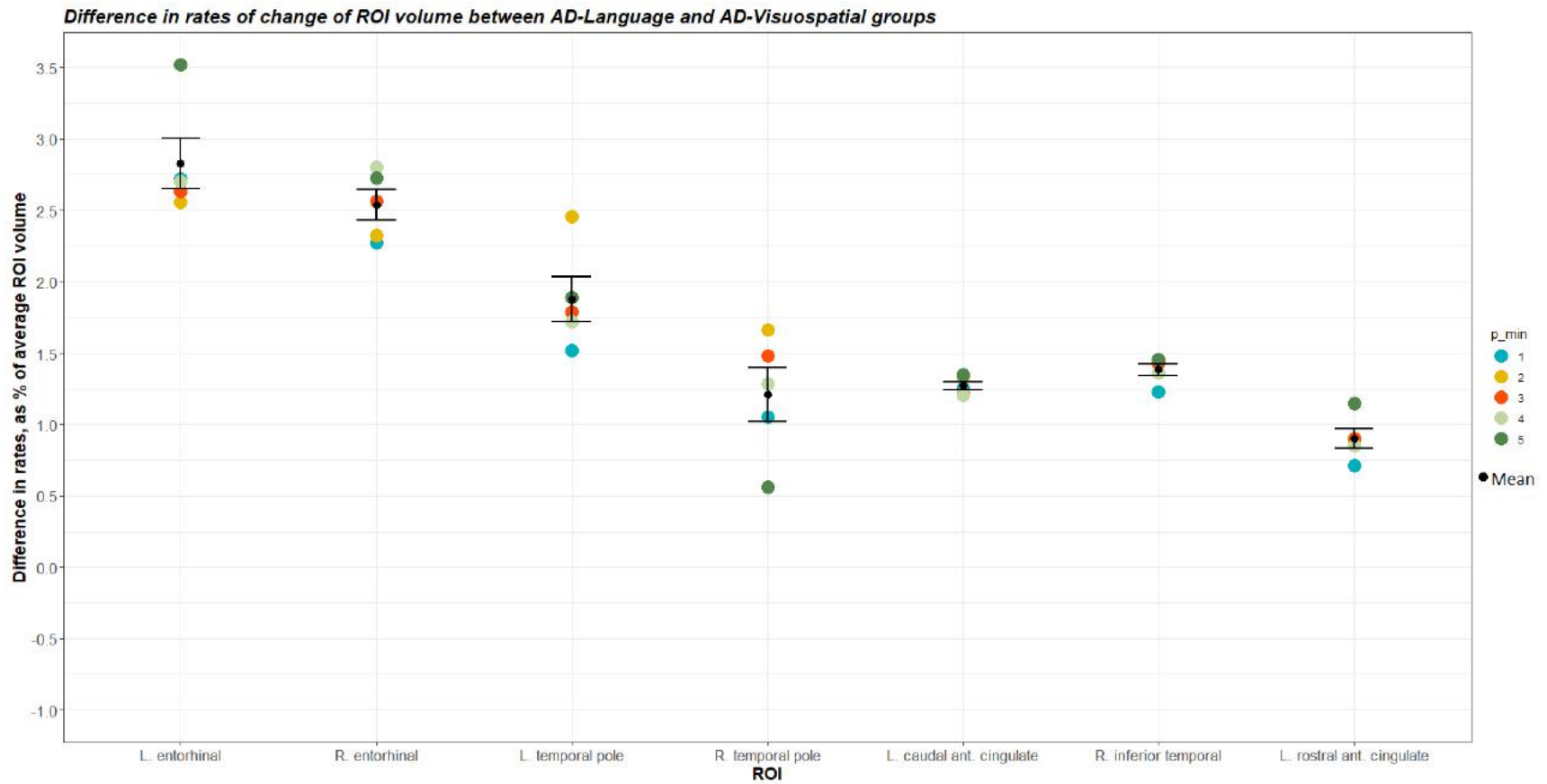


Figure 6.7: Results from β_{10} coefficients in linear mixed effects models for different p_min datasets: Difference in rates of change of ROI volume, as a % of average ROI volume. The mean of the differences across the different p_min datasets is shown in black, along with error bars (based on standard error of the mean).

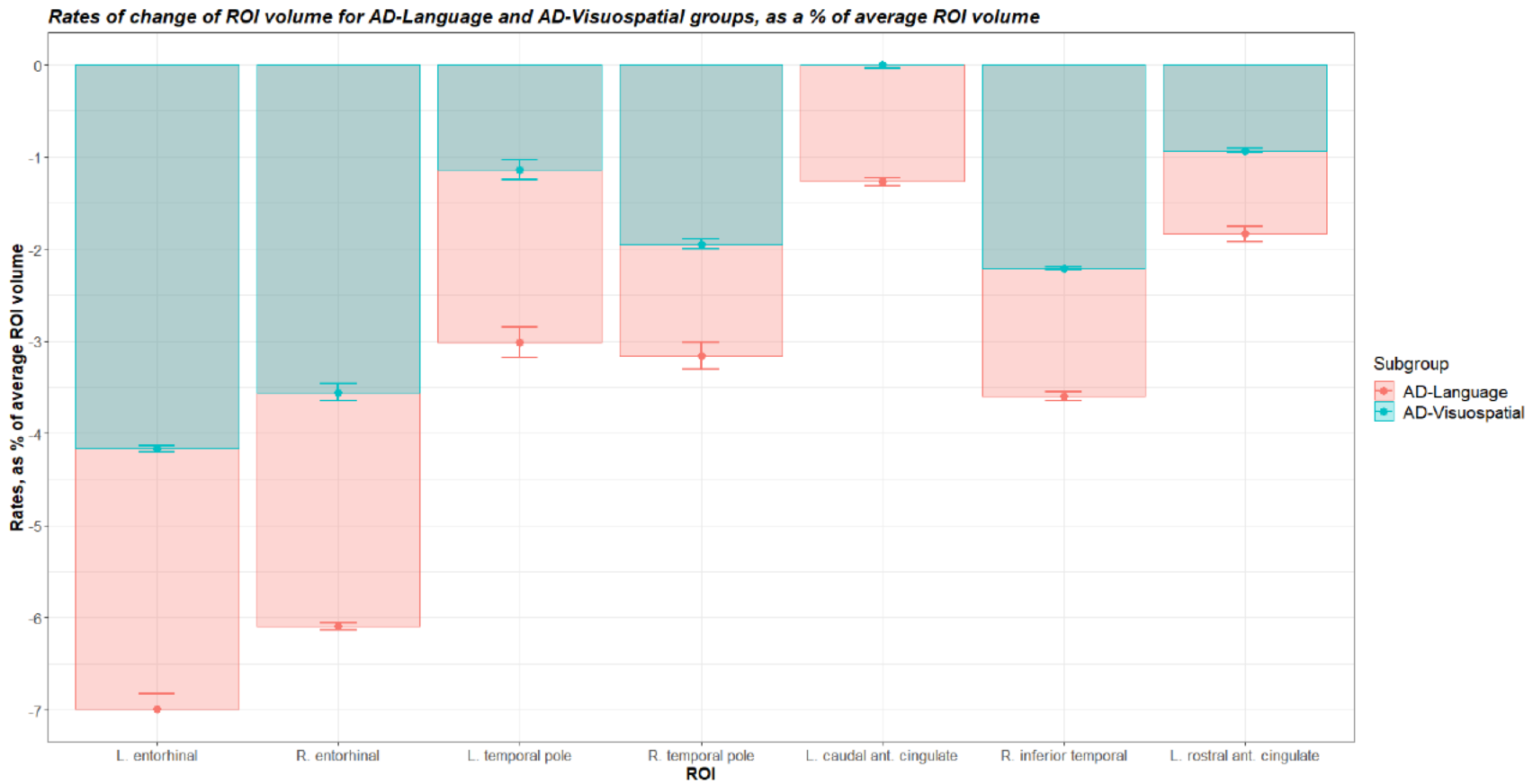


Figure 6.8: A comparison of the rates of change of ROI volume for the AD-Language and AD-Visuospatial groups. The height of the bars represents the mean rate of change for a given group from the analyses of the different p_{min} (1, 2, 3, 4, 5) datasets. Mean values are also plotted as points with error bars based on standard error of the mean from the five p_{min} datasets. A more negative rate indicates a faster decline of ROI volume. This figure delivers the following key points: 1. Which AD subgroup has a faster rate of decline for the different ROIs, 2. The absolute difference in rates between the groups, indicated by the non-overlapping part of the bars for the two subgroups and 3. How much this difference is in proportion to the rate of change of each of the subgroups.

Difference in rates of change of ROI volume between AD-Language and AD-Visuospatial groups, and associated p-values for different p_min datasets

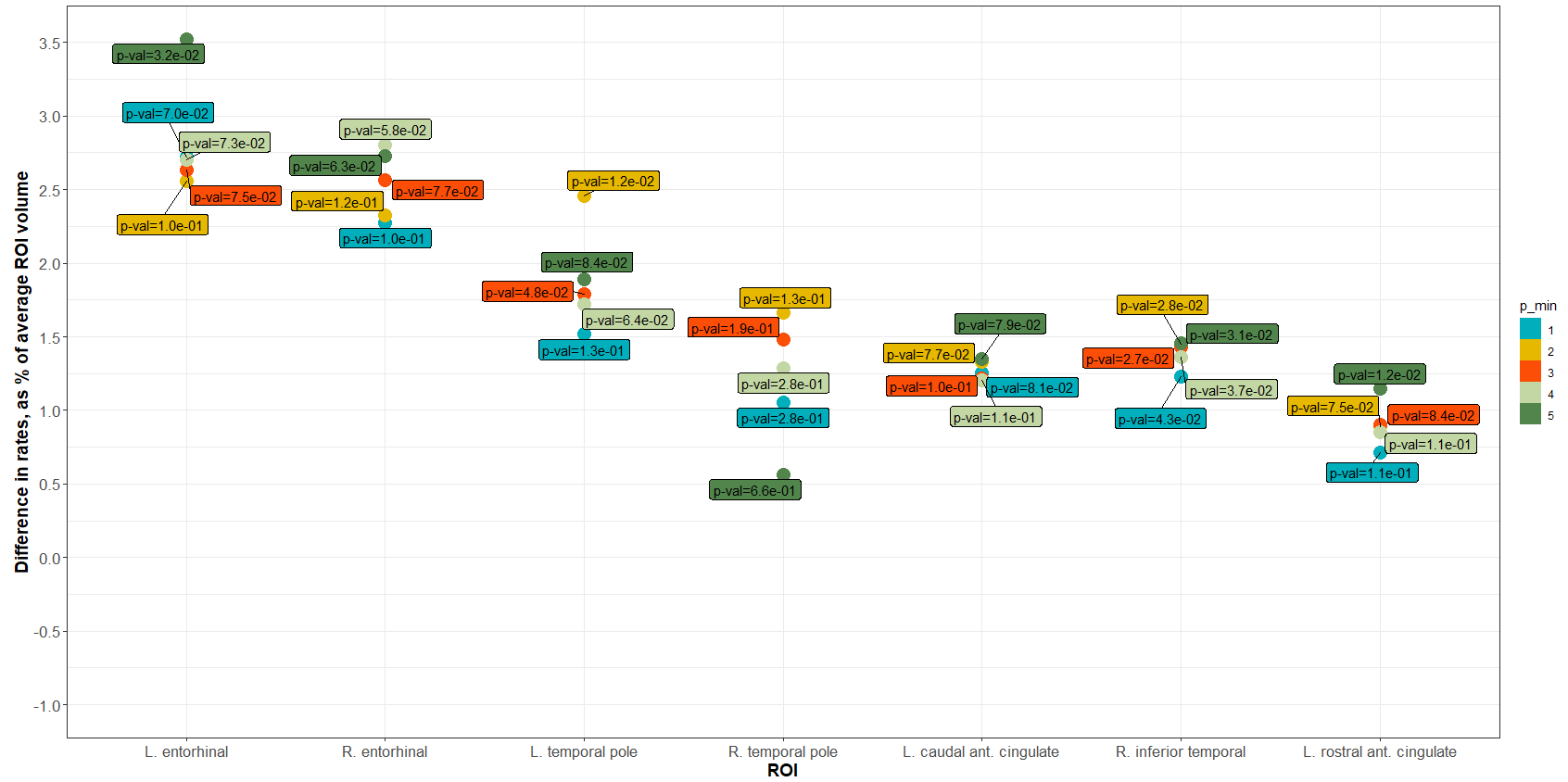


Figure 6.9: P-values associated with each estimate of the difference in rates between AD-Language and AD-Visuospatial groups, from linear mixed effects modeling analysis of different p_min datasets.

6.3.3.2 Differences in average ROI volume at t=0

In terms of the analysis of β_2 coefficients, a positive β_2 coefficient from the LME models for the AD-Language and AD-Visuospatial data corresponds to a greater average ROI volume at t=0 in the AD-Visuospatial group compared to the AD-Language group, while a negative β_2 indicates the opposite. It can be seen from Tables 6.13a-e that the **right inferior parietal lobule has a smaller volume at t=0 for the AD-Language group compared to the AD-Visuospatial group.** All other ROIs (**right temporal pole, left entorhinal cortex, left inferior temporal gyrus and left middle temporal gyrus**) have **smaller volumes at t=0 for the AD-Visuospatial group compared to the AD-Language group.** The associated p-values for each of these differences (last column in Tables 6.13a-e) are at least an order larger than the threshold needed for significance with the Holm-Bonferroni correction with a family wise error rate of 0.05.

Right temporal pole and **left entorhinal cortex** are the common ROIs found in the lists of importance based on average differences in ROI volumes at t=0 and average differences in rate of change of ROI volume between the AD-Language and AD-Visuospatial groups, as predicted by the LME models. Since the current cross-sectional data analysis for AD-Language vs. AD-Visuospatial did not yield a reliable classification model, a comparison of which ROI are important across cross-sectional and longitudinal data analysis cannot yet be performed.

Tables 6.13a – 6.13e: The ROIs shown here are a collective list based on the top 5 ROIs for AD-Language vs. AD-Visuospatial from the LME analysis of each p_min dataset based on the magnitude of the β_2 coefficient, normalized by average ROI volume ('Difference in rates as % of Avg ROI volume' column in the table). Individuals with AD* indicated in the second column of each table are individuals from the 4 single domain groups: AD-Executive, AD-Language, AD-Memory and AD-Visuospatial ROIs with a positive β_2 (greater volume in the AD-Visuospatial group at t=0) are shown in pink while ROIs with a negative β_2 (smaller volume in the AD-Visuospatial group at t=0) are shown in green.

AD-Language vs. AD-Visuospatial				
p_min = 1				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right temporal pole	2098.4	248.6	11.8	7.7E-03
Left entorhinal	1521.9	172.4	11.3	2.8E-02
Left inferior temporal	8517.7	676.3	7.9	1.2E-02
Left middle temporal	8130.4	484.5	6.0	3.5E-02
Right inferior parietal	11043.6	-525.4	-4.8	4.2E-02

Table 6.13a: Differences in ROI volume at t=0 between AD-Language and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 1. See note “Tables 6.13a-e” above for color coding.

AD-Language vs. AD-Visuospatial				
p_min = 2				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right temporal pole	2098.4	249.5	11.9	7.2E-03
Left entorhinal	1521.9	174.0	11.4	2.7E-02
Left inferior temporal	8517.7	683.0	8.0	1.1E-02
Left middle temporal	8130.4	483.1	5.9	3.7E-02
Right inferior parietal	11043.6	-524.8	-4.8	4.2E-02

Table 6.13b: Differences in ROI volume at t=0 between AD-Language and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 2. See note “Tables 6.13a-e” above for color coding.

AD-Language vs. AD-Visuospatial				
p_min = 3				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right temporal pole	2098.4	253.1	12.1	6.3E-03
Left entorhinal	1521.9	174.7	11.5	2.6E-02
Left inferior temporal	8517.7	689.3	8.1	1.0E-02
Left middle temporal	8130.4	490.9	6.0	3.5E-02
Right inferior parietal	11043.6	-524.9	-4.8	4.2E-02

Table 6.13c: Differences in ROI volume at t=0 between AD-Language and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 3. See note “Tables 6.13a-e” above for color coding.

AD-Language vs. AD-Visuospatial				
p_min = 4				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right temporal pole	2098.4	249.8	11.9	7.1E-03
Left entorhinal	1521.9	176.8	11.6	2.4E-02
Left inferior temporal	8517.7	689.4	8.1	1.0E-02
Left middle temporal	8130.4	490.1	6.0	3.5E-02
Right inferior parietal	11043.6	-529.6	-4.8	4.1E-02

Table 6.13d: Differences in ROI volume at t=0 between AD-Language and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 4. See note “Tables 6.13a-e” above for color coding.

AD-Language vs. AD-Visuospatial				
p_min = 5				
ROI	Estimated average ROI volume for individuals with AD* at time of diagnosis (mm ³)	β_2 (Difference in ROI volumes at t=0) (mm ³)	Difference in ROI volumes at t=0, as % of average ROI volume	p-value β_2
Right temporal pole	2098.4	249.7	11.9	7.2E-03
Left entorhinal	1521.9	176.8	11.6	2.4E-02
Left inferior temporal	8517.7	702.9	8.3	8.6E-03
Left middle temporal	8130.4	492.7	6.1	3.6E-02
Right inferior parietal	11043.6	-532.4	-4.8	4.0E-02

Table 6.13e: Differences in ROI volume at t=0 between AD-Language and AD-Visuospatial groups for selected ROIs of importance, for the dataset with p_min = 5. See note “Tables 6.13a-e” above for color coding.

AD-Language vs. AD-Visuospatial		
ROI	Mean difference in ROI vol at t=0 (across different p_min datasets), as % of average ROI volume	Std. Error of mean
Right temporal pole	11.92	0.04
Left entorhinal	11.50	0.05
Left inferior temporal	8.08	0.05
Left middle temporal	6.01	0.02
Right inferior parietal	-4.78	0.01

Table 6.14: Variation in the difference in ROI volume at t=0 across the five p_min datasets for AD-Language vs. AD-Visuospatial analysis. Mean and standard error of the mean are shown for each ROI.

6.4 Concluding remarks about longitudinal data analysis and future work

Overall, for all three comparisons: a.) AD-Language vs. AD-Memory, b.) AD-Memory vs. AD-Visuospatial and AD-Language vs. AD-Visuospatial, the top ranking ROIs in each comparison showed substantial differences in rates of change of ROI volumes relative to the rates for each of the subgroups being compared. Based on the current analysis using LME models of ROI volumes from longitudinal data, my work in hypothesis generation of which ROIs are most important to investigate for AD subgroup differences in future studies including neuropathological studies points at the ROIs listed in Table 6.15.

Comparison	Differences between subgroups	Top ROIs of importance
AD-Language vs. AD-Memory	Average differences in rates of change of ROI volume	Left posterior cingulate, left entorhinal cortex, right entorhinal cortex, left caudal anterior cingulate, right caudal middle frontal, left parahippocampal, right frontal pole.
	Average differences in ROI volumes at t=0	Right entorhinal, right hippocampus, left temporal pole, left inferior temporal and left middle temporal

AD-Memory vs. AD-Visuospatial	Average differences in rates of change of ROI volume	Right temporal pole, left entorhinal, left rostral anterior cingulate, left frontal pole, left parahippocampal, left frontal pole
	Average differences in ROI volumes at t=0	Left entorhinal cortex, right entorhinal cortex, left parahippocampal gyrus, left temporal pole and left hippocampus
AD-Language vs. AD-Visuospatial	Average differences in rates of change of ROI volume	Left entorhinal, right entorhinal, left temporal pole, right temporal pole, left caudal anterior cingulate, right inferior temporal, left rostral anterior cingulate
	Average differences in ROI volumes at t=0	Right inferior parietal, right temporal pole, left entorhinal cortex, left inferior temporal, left middle temporal

Table 6.15: Top ROIs of importance for differences between pairs of AD subgroups at the time of AD diagnosis, based on coefficient estimates from linear mixed effects modeling for each ROI volume. Color codings for ROIs indicate a greater rate of decline or smaller average volumes in the AD subgroup indicated by the color.

The LME models based on longitudinal data also provided insight into the population level differences between the AD subgroups at t=0, the time of AD diagnosis. There were some ROIs that showed differences between the pairs of AD subgroups based on both ROI volumes at t=0 and rates of change of volume with respect to time. For example, the right entorhinal cortex was an ROI determined to be important for AD-Language vs. AD-Memory differences based on both differences in rates of change of ROI volume for the two subgroups and differences in ROI volumes at t=0. Additionally, in the AD-Memory vs. AD-Visuospatial analysis, left entorhinal cortex, right entorhinal cortex and left parahippocampal gyrus were determined to be important based on differences in rates of change of ROI volume as well as differences in ROI volume at t=0. Right temporal pole and left entorhinal cortex are the common ROIs found in the lists of

importance based on average differences in ROI volumes at $t=0$ and average differences in rate of change of ROI volume between the AD-Language and AD-Visuospatial groups, as predicted by the LME models. These results indicate that some ROIs that might be most differentiating for the two given AD subgroups at the time of AD diagnosis continue to be important longitudinally as well.

For AD-Language vs. AD-Memory and AD-Memory vs. AD-Visuospatial comparisons, where ROI importance was determined successfully in cross-sectional data, there were some ROIs that were found in the top ROIs of importance based on both the cross-sectional and longitudinal data analysis. Further, among the two given AD subgroups being considered for difference, the longitudinal and cross-sectional analysis of these ROIs were consistent in which AD subgroup has less volume at $t=0$.

LME models are a powerful framework for learning about population level dynamics of an ROI volume while allowing for individual level deviations from the population level trajectories and account for the correlation structure in longitudinal data. Current analysis included 70 ROI models for each of the three binary comparisons of AD subgroups, using all of longitudinal data available and using four subsets of data based on a minimum number of data points needed at each time point for each AD subgroup. For each binary comparison of AD subgroups, this resulted in $70 \times 5 = 350$ models. As noted in the sections above, a correction for multiple hypotheses testing was needed to determine a threshold of significance for the p-values associated with the coefficients of interest (β_2 and β_{10}). These thresholds based on a family-wise error rate of 0.05 were calculated using the Holm-Bonferroni correction. According to this correction, the thresholds for the top 5 (smallest) p-values are on the 1.43×10^{-4} to 1.45×10^{-4} . For all three binary comparisons of AD subgroups, the β_{10} coefficients had p-values on the order

of 10^{-1} or 10^{-2} . The β_2 coefficients for most ROIs of importance were on the order of 10^{-2} or 10^{-3} . Of all the ROIs analyzed, only two ROIs (right hippocampus in the comparison of AD-Language vs. AD-Memory and left hippocampus in the comparison of AD-Memory vs. AD-Visuospatial) had p-values associated with the β_2 coefficient in the 10^{-4} range. Considering that these two ROIs are in the much lower p-value range despite the current small sample size, these ROIs may be recommended as the most important ROIs to look into for volume differences at $t=0$.

Given that the majority of p-values were much above the threshold needed for the chosen level of significance, the current analysis did not focus on p-values to determine how important an ROI is for AD subgroup differences. The current analysis was based on coefficient magnitudes, normalized by ROI size. Given the small sample sizes of the AD subgroups and the sparseness of the current longitudinal data, an analysis based on coefficients magnitudes rather than p-values seemed more reasonable and reliable. The results from analyses of different subsets of data (corresponding to different p_{\min} values) showed that in the current dataset, p-values are sensitive to the removal or addition of data points. As more data becomes available in the future, it is anticipated that one could move towards an inference analysis based on p-values. Not being able to use the p-values for inference in the current analysis does not mean that the current results based on magnitudes of coefficients are not valid. However, the analysis can be stronger in the future and p-values can add a layer of confidence to the current results. Another point to be tackled in future work is a method to assess how good the current LME models are for each ROI volume. In the current work, all models converged except for a handful cases mentioned in the Results section. Of the models that converged, in the current analysis, there was no established method of determining how much a given ROI model can be trusted. Having this assessment in

the future can refine the current results and can also provide more confidence about which ROIs are most important in longitudinal differences across pairs of AD subgroups.

The LME model itself can also be improved in future work. In the current model of an ROI volume, males and females were modeled to have the same rate of change of volume over time. Individuals across different ages, education levels and *APOE* genotype status were also assumed to have the same rate of change of ROI volume. The current model can be made more sophisticated in the future by allowing for the rate of change of ROI volume to vary with gender, age, education levels and *APOE* genotype status. Additionally, one can explore non-linear relationships of covariates in a mixed effects model for ROI volume in the future. The work can also be extended in the future to account for potential correlation among ROIs, which has not been taken into account in the current univariate model approach. Given the current small sample size and noise in the current data, a simpler model for each ROI volume as currently implemented seemed to be a good starting point for the analysis of this dataset.

The analysis can also be improved in the future by analyzing two subsets of data separately using separate LME models: data before the time of AD diagnosis and data after the time of diagnosis. The hypothesis for this analysis would be that the rates of change of ROI volume are different for these two datasets. Consequently, such an analysis may point at different ROIs as being important in the time leading up to AD diagnosis and the time after.

In the current analysis scheme of longitudinal data with LME models, I explored the sensitivity of LME models to removal of timepoints with a number of individuals below a threshold (p_{\min}) for a given AD subgroup. Results showed that there is some variability across different p_{\min} datasets in both β_{10} , the estimated average differences in rate of change of ROI volume between the AD subgroups being compared and β_2 , the estimated average difference in ROI

volume between the subgroups at $t=0$. P-values also vary for a given coefficient for models of a given ROI volume based on different p_{\min} datasets. Qualitatively, the variation in β_{10} and β_2 estimates from LME models based on different p_{\min} datasets did not seem large. However, a quantitative method is needed as part of future work to assess whether the variation observed in coefficient values and variation in p-values across different p_{\min} datasets is small or large. I hypothesized that using data corresponding to $p_{\min} = 5$ would be the most robust since each time point has at least five individuals' data for each of the AD subgroups being compared. However, the dataset corresponding to $p_{\min} = 5$ has less data overall compared to the lower p_{\min} cases. The current analysis did not answer the question of which p_{\min} dataset would be the most appropriate to report results from. For the results reported above, I used average values based on all five p_{\min} datasets for differences in rate of change of volume and differences in volume at $t=0$. In future, developing a metric for which p_{\min} dataset results are most reliable can be useful.

Chapter 7: Conclusions, Limitations & Future Work

My dissertation focused on hypothesis generation directed towards which brain ROIs could be important for future studies. This includes neuropathological studies for AD subgroup differences which may have implications in precision medicine approaches in treating AD. I analyzed cross-sectional data from the time of AD diagnosis using machine learning classification models (Aim 1) and longitudinal data using linear mixed effects modeling (Aim 2) from individuals in the ADNI study to arrive at lists of ROIs that are important in three binary comparisons of AD subgroups. My work was an application of data science methods in a new domain: AD subgroup analysis. In both cross-sectional and longitudinal data analyses, my focus was on a careful assessment and implementation of methods to best make use of the data available in its current form, which has small sample sizes, imbalanced AD subgroup sizes and is noisy in nature. Recommendations for which ROIs may be important to focus on in future neuropathological studies based on my analyses are summarized in Tables 7.1 and 7.2.

In cross-sectional data analysis, using variable importance measures from a random forest analysis, I determined relative importance of all 70 ROIs for distinguishing between pairs of AD subgroups (Chapter 4) and presented this information for all ROIs in a color gradient for importance on a brain diagram. To focus the discussion of important ROIs to a few ROIs, I presented a list of the top 10 ROIs that I determined to be important for each binary comparison (see Table 7.1), and focused on the top four ROIs for a more detailed analysis: **right entorhinal cortex, right hippocampus, right lingual gyrus and left hippocampus** (AD-Language vs. AD-Visuospatial) and **left entorhinal cortex, right entorhinal cortex, right supramarginal gyrus and left postcentral gyrus** (for AD-Memory vs. AD-Visuospatial).

Cross-sectional data analysis: Top ROIs of importance for differences between pairs of AD subgroups

Comparison	Top ROIs of importance
AD-Language vs. AD-Memory	Right entorhinal, right hippocampus, right lingual, left hippocampus, left rostral anterior cingulate, left superior temporal, right cuneus, right superior frontal, right pars triangularis, right fusiform
AD-Memory vs. AD-Visuospatial	Left entorhinal, right entorhinal, right supramarginal, left postcentral, right superior parietal, left superior temporal, left parahippocampal, right parahippocampal, right post central, right cuneus
AD-Language vs. AD-Visuospatial	Additional work is needed for a reliable classification model to report the ROIs of importance.

Table 7.1: Cross-sectional data analysis: Top ROIs of importance for distinguishing between pairs of AD subgroups at the time of AD diagnosis, based on an importance measure from random forest classification models. Color codings for ROIs indicate smaller average volumes in the AD subgroup indicated by the color. ROIs without the color coding are the ones for which the current results were inconclusive about which AD subgroup has the smaller volumes.

Violin plot distributions from cross-sectional data for the above ROIs of the respective AD subgroups showed that on average, individuals in the AD-Memory group had lower volumes in the right entorhinal cortex and right hippocampus compared to the AD-Language group. For the next two ROIs of importance for AD-Language vs. AD-Memory, right lingual gyrus and the left hippocampus, average differences between the two subgroups were not so distinguishable based on the violin plot distribution comparison. For the AD-Memory vs. AD-Visuospatial comparison, on average, the AD-Memory group had lower volumes in the left entorhinal cortex and right entorhinal cortex. The AD-Visuospatial group had slightly smaller volumes on average for the right superiormarginal gyrus compared to the AD-Memory group, although the difference was not very pronounced based on the distributions of the two groups. The distributions for the

two groups were even harder to spot a pronounced difference for in the left postcentral gyrus. This inconsistency in interpretations from the two methods, random forest variable importance and violin plot distributions, pointed at two possibilities. One possibility is that random forest which allows for interaction of ROI volume variables and non-linear relationships to be considered in the tree building process, was able to pick out differences between AD subgroups that may not be apparent in a simpler isolated analysis of an ROI. The other possibility is that there is potential bias in random forest importance that results in spurious identification of certain variables as important. This was one limitation of the current results based on random forest for cross-sectional data analysis. In future work, it will be necessary to resolve this ambiguity around specific ROIs from random forest results. One proposed solution is to use conditional inference forests (cforest) with subsampling without replacement instead of random forests to determine if the same questionable ROIs appear or not in the variable importance results from cforest. Cforest with subsampling without replacement has been shown to correctly identify the variables of importance in simulation studies, unlike random forest which may favor variables with a large number of categories.

Current cross-sectional data analysis for AD-Language vs. AD-Visuospatial (see row 3 of Table 7.2) did not yield reasonable models for interpreting ROI importance. In random forest models, the classification errors for the two groups were imbalanced (roughly 11% for one AD subgroup and 35% for the other AD subgroup). The reason for this imbalance in the classification errors needs to be explored in future work. The ROIs that were the most important for this classification are still listed in Chapter 4 for completeness, along with the relative importance of each of the 70 ROIs mapped in a color gradient on diagrams of the brain. In future, it would be worth seeing if conditional inference forest with subsampling without replacement also results in a similar class

imbalance for classification errors. In the current work, the penalized logistic regression model had a much higher classification error overall (~30%) for AD-Language vs. AD-Visuospatial, which was the reason for not reporting variable importance results from that model. If the class imbalance in accuracies in random forest for AD-Language vs. AD-Visuospatial cannot be resolved in future work, it may be necessary to resort to the current random forest or penalized logistic regression models as a starting point for getting some insight into which ROIs are important for differences in these AD subgroups at the time of AD diagnosis.

There is another approach that could be considered to gain some insight into the AD-Language vs. AD-Visuospatial regions of importance at the time of AD diagnosis. This approach is to utilize the results from linear mixed effects (LME) modeling of longitudinal data (Aim 2 work), focusing on the β_2 coefficient that represents the difference in ROI volumes at t=0 between the two subgroups being compared. The LME models pointed at the following ROIs as the most important for AD-Language and AD-Visuospatial average group differences at t=0 (See Table 7.2): right temporal pole, left entorhinal, left inferior temporal, left middle temporal and right inferior parietal. A limitation of this proposed approach is that there was not a metric used to assess how good the LME models are other than making sure the models converged. So, it is difficult to know whether the LME model results for t=0 based on longitudinal data can be trusted more than the random forest or penalized logistic regression models based on cross-sectional data. Additionally, the current LME models for each ROI are separate models which don't allow for correlation among ROIs and conditional probabilities of being in a particular group to be taken into account unlike random forest and penalized logistic regression models in Aim 1 work that are joint models of ROIs. For cross-sectional data which are more straightforward to model than longitudinal data in models involving multiple independent

variables together, joint models of ROIs are preferred for a more comprehensive story of the relative importance of ROIs. Nonetheless, in the absence of reliable joint models for cross-sectional data for the AD-Language vs. AD-Visuospatial classification, individual LME models for the ROIs from longitudinal data analysis could be a starting point for preliminary hypothesis generation for ROIs to be explored in future studies.

Some of the approaches that I implemented for dealing with the challenges of cross-sectional data addressed the issues but it is important to recognize the limitations of these approaches. First, oversampling the minority AD subgroups using SMOTE for creating balanced class sizes for binary classification models potentially introduced some bias into the data. Although SMOTE is considered to be a less biased oversampling technique, it still involves generating synthetic data which introduces some bias into the data to be used for downstream analysis. Second, using linear regression models to remove the effects of extraneous variables from ROI volumes also has its limitations. I made a simplistic assumption about the relationship of these extraneous variables with ROI volume to be linear. Preliminary analysis using plots of any given ROI volume vs. these extraneous variables did not show strong correlations except for age and sex. Using non-linear models for regressing out the effects of any of these extraneous variables did not seem to be an appropriate choice given the small sample size and noise in the relationships of continuous extraneous variables with ROI volumes. In the current work, the residual volumes used in downstream cross-sectional data analysis were obtained by subtracting the modeled effects of the extraneous variables from the raw ROI volumes, assuming that the effects got modeled as well as they could based on the data. In future work, it may be useful to assess how good the models were for removing the effects of the extraneous variables from ROI volumes and to fine-tune the models while also keeping the risk of overfitting in mind.

Overall, through cross-sectional data analysis (Aim 1), I demonstrated the use of supervised machine learning methods focusing on variable importance to understand which ROIs are most important for distinguishing between AD subgroups. I showed that random forest models had better classification accuracies than penalized logistic regression models for the data at hand. Through the use of random forest, the ROI volume variables were allowed to have non-linear relationships with AD subgroup as well as interactions among the ROI volume variables, which may be a good way to model the unknown complexities of biological data. My work is an application of machine learning methods in a new domain: AD subgroups. There is no known gold standard for how well the data can be separated into AD subgroups based on classification models. To my knowledge, there is currently no documentation of classification accuracies for distinguishing pairs of cognitively defined AD subgroups based on machine learning methods. Hence, the work I've done here may provide a starting benchmark for classification accuracies for distinguishing between pairs of AD subgroups based on ROI volumes from the time of AD diagnosis. Future models for distinguishing between AD subgroups could aim at improving upon these classification accuracies. It is important to note that these models are not intended to be used in a clinical setting for making predictions about which AD subgroup an individual belongs to. The classification accuracies are a way to assess how good the model is which is to be used for determining which ROIs are important for AD subgroup differences.

The highlights of longitudinal data analysis (Aim 2) from my dissertation work are that for each of the three binary comparisons evaluated, there are ROIs whose differences in rates of change of ROI volume between the two subgroups are substantial relative to the rates of each of the subgroups. The top ROIs of importance are presented in Table 7.2. There are two lists of ROIs for each binary comparison of AD subgroups. One is based on average differences in rates of

change of ROI volume between two AD subgroups and the other is based average differences in ROI volumes at t=0 between the groups. The ROI importance results were based on the magnitudes of the coefficients of interest from the LME models, relative to the average ROI volume from all single domain subgroups at the time of diagnosis.

Top ROIs of importance for differences between pairs of AD subgroups: Longitudinal data analysis

Comparison	Differences between subgroups	Top ROIs of importance
AD-Language vs. AD-Memory	Average differences in rates of change of ROI volume	Left posterior cingulate, left entorhinal cortex, right entorhinal cortex, left caudal anterior cingulate, right caudal middle frontal, left parahippocampal, right frontal pole.
	Average differences in ROI volumes at t=0	Right entorhinal, right hippocampus, left temporal pole, left inferior temporal and left middle temporal
AD-Memory vs. AD-Visuospatial	Average differences in rates of change of ROI volume	Right temporal pole, left entorhinal, left rostral anterior cingulate, left frontal pole, left parahippocampal, left frontal pole
	Average differences in ROI volumes at t=0	Left entorhinal cortex, right entorhinal cortex, left parahippocampal gyrus, left temporal pole and left hippocampus
AD-Language vs. AD-Visuospatial	Average differences in rates of change of ROI volume	Left entorhinal, right entorhinal, left temporal pole, right temporal pole, left caudal anterior cingulate, right inferior temporal, left rostral anterior cingulate
	Average differences in ROI volumes at t=0	Right inferior parietal, right temporal pole, left entorhinal cortex, left inferior temporal, left middle temporal

Table 7.2: Longitudinal data analysis: Top ROIs of importance for differences between pairs of AD subgroups at the time of AD diagnosis, based on coefficient estimates from linear mixed effects modeling for each ROI volume. Color codings for ROIs indicate a greater rate of decline or smaller average volumes in the AD subgroup indicated by the color.

An interesting question to be asked with results from both the cross-sectional and longitudinal data analyses is whether the ROIs determined to be important at the time of AD diagnosis continue to be important for AD subgroup differences over time. For AD-Language vs. AD-Visuospatial, such a comparison could not be performed due to the lack of reliable results from cross-sectional data analysis for comparing the two groups. However, for both AD-Language vs. AD-Memory and AD-Memory vs. AD-Visuospatial, in comparison of cross-sectional and longitudinal data analysis results, there were some ROIs that were found to be important for distinguishing between the two subgroups based on both cross-sectional data and longitudinal data. The details of these common ROIs are described in the next few paragraphs.

For AD-Language vs. AD-Memory, the right entorhinal cortex which ranked the highest in variable importance in cross-sectional data analysis. This was also one of the top regions that showed substantial differences between the two subgroups in longitudinal data analysis, both in terms of average difference in rate of change of volume and average difference in ROI volume at $t=0$ (as predicted by the LME models on longitudinal data). The result from longitudinal data analysis corresponding to the average difference in ROI volume at $t=0$ was consistent with which AD subgroup had lower volume in cross-sectional data (AD-Memory). The difference in rate of change of volume of right entorhinal cortex between the two groups, however, indicated that the AD-Language group had a faster decline in volume for this region compared to the AD-Memory group. A hypothesis to explain this could be that since the entorhinal cortex is known to play a vital role in the memory function, it may be the case that individuals in the AD-Memory group have already had a great rate of decline in the right entorhinal cortex compared to the AD-Language group by the time they get diagnosed with AD. The lower volumes for this ROI in the AD-Memory group at the time of AD diagnosis may also be a result of this hypothesized fast

decline before AD diagnosis. After AD diagnosis, individuals in the AD-Language group may start showing a faster decline in this ROI compared to individuals in the AD-Memory group as the fast phase of decline for this ROI in the AD-Memory group has already taken place. This is only one hypothesis (and not a conclusion) to explain the greater rate of decline in the right entorhinal cortex for the AD-Language group compared to the AD-Memory group.

Another ROI that was determined to be important for AD-Language vs. AD-Memory in cross-sectional data analysis and longitudinal data analysis is the right hippocampus. The right hippocampus was the second most important ROI for AD-Language vs. AD-Memory based on the random forest analysis of the cross-sectional dataset and had lower volumes in the AD-Memory group. In the longitudinal data analysis, the LME model for the right hippocampus volume estimated that at $t=0$, the AD-memory group has lower ROI volumes than the AD-Language group. This region was not in the list of ROIs determined to be important for differences in rate of change of ROI volume for the two groups.

For AD-Memory vs. AD-Visuospatial, there were several common ROIs of most importance in both cross-sectional data analysis and longitudinal data analysis. Based on differences between the AD subgroups in rate of change of ROI volume and differences in ROI volumes at $t=0$, these were left entorhinal cortex, right entorhinal cortex and left parahippocampal gyurs. In cross-sectional data analysis, left entorhinal cortex and right entorhinal cortex were the top two ROIs based on random forest variable importance results and left parahippocampal gyrus which was the 7th most important ROI, all three of which had on average lower volumes in the AD-Memory group compared to the AD-Language group. In longitudinal data analysis, all three of these ROIs were estimated to have a faster rate of decline in the AD-Memory group. The LME models also estimated that ROI volume at $t=0$ was smaller, on average, in the AD-Memory group than in the

AD-Visuospatial group, which is consistent with cross-sectional data analysis results. In this comparison of cross-sectional and longitudinal data results, all three regions discussed here (left entorhinal cortex, right entorhinal cortex and left parahippocampal gyrus) are known to play a role in the memory function. It is interesting to note that when the AD-Memory group is compared to AD-Language, it has a slower decline in the right entorhinal cortex whereas when the AD-Memory group is compared to the AD-Visuospatial group, it has a faster decline in the same ROI.

Hence, results based on cross-sectional data and longitudinal data show some common ROIs that are determined to be important for AD subgroup differences. It is important to keep in mind that the methods used to determine ROI importance in cross-sectional data and longitudinal data are different. To arrive at some overlapping results for which ROIs are most important at the time of AD diagnosis based on two different methods is reassuring and adds more confidence to the results. Further, it is insightful to know that some ROIs determined to be important at the time of AD diagnosis continue to be important in rate of change differences between the subgroups. These ROIs may be the most crucial to study further for AD subgroup differences.

It is equally interesting to note the ROIs that are in the topmost important ROIs for rate of change of volume differences between the subgroups but do not show up in the top most important ROIs for AD subgroup differences at the time of AD diagnosis. It is possible some of these ROIs have different levels or directions of differences between the two subgroups being compared in the time periods before and after AD diagnosis. This is an avenue to be explored in future work: to study the differences between AD subgroups from two subsets of data, one including data from all visits before and including AD diagnosis, and the other including data from all visits after and including AD diagnosis. The analysis can also be extended to study data

from different time windows around AD diagnosis. A hypothesis for this future analysis is that the AD subgroup differences would be the greatest and strongest in the time period leading up to AD diagnosis and also in perhaps some smaller time period after AD diagnosis. As AD progresses in each individual over time, the differences between AD subgroups may become more diluted as all of the brain tissue (hence all ROIs) undergoes atrophy. Hence, studying data from different time windows in the longitudinal dataset separately may be an insightful analysis in the future. This approach may come at the expense of having a sparser dataset to be modeled for a chosen time window, which with the current sample size may not be ideal.

A topic that needs further discussion is how to choose the number of ROIs that one should report as the most important for each binary comparison of AD subgroups. This is relevant to both cross-sectional and longitudinal data analyses. This is an important topic to tackle as one heads towards future work or one moves in the direction of providing recommendations for ROIs of interest to neuropathologists for understanding AD subgroup differences. In my dissertation work, I made a subjective call to provide comparisons of the importance measure in cross-sectional data analysis in plots showing the top 10 ROIs (ranked by the importance measure). I then focused on the top four ROIs for comparing distributions for the AD subgroups of interest to determine which of the two AD subgroups has smaller volumes for a given ROI. In longitudinal data analysis, I focused on the top 5 ROIs for each of the p_min datasets for the two AD subgroups being compared, and then used the union of the list of top 5 ROIs corresponding to each p_min dataset analysis as the most important ROIs for the given AD subgroups' comparison. At the end, I had two collective list of ROIs for each binary comparison of AD subgroups based on all p_min datasets' analyses: one based on average differences in the rate of change of ROI volume between the AD subgroups being compared and the other based on

average differences in ROI volumes between the subgroups at t=0. The decision to focus on the top 5 ROIs from the analysis of each p_min dataset for a given binary comparison of AD subgroups and to report a collective list of ROIs that is the union of the top 5 ROIs across the different p_min datasets' analyses is subjective. As noted in Chapter 6, the answer to the question of how many ROIs should be included in the top ROIs of importance can be best obtained by consulting neuropathologists who may be interested in using the results from the current work for designing neuropathological studies for understanding AD subgroup differences. In both cross-sectional and longitudinal data analyses, I applied the chosen ranking metrics for importance on all 70 ROIs. Hence, one can easily extend the current analysis to focus on a different number of top ROIs in the future, as per the needs of future studies that may be interested in the results of my hypothesis generation work for ROIs of importance in AD subgroup differences.

Given the current small sizes of AD subgroups, a p-value based analysis for determining statistical significance of the differences between subgroups in longitudinal data was not deemed to be appropriate or reliable. When the sample size is small, the p-values can be sensitive to the inclusion or exclusion of specific data points. A more enriched dataset both in the number of individuals and time points is needed to move towards an inference based analysis using p-values, which is anticipated as future work as more data becomes available. Although p-values were not used to assess ROI importance in the current analysis, I did provide them in the results in Chapter 6 for completeness. The majority of the p-values were much larger than the threshold for significance determined by the Holm-Bonferroni correction ($\sim 10^{-4}$) for multiple hypotheses testing with a family wise error rate of 0.05. Most p-values were on the order of 10^{-1} and 10^{-2} for differences in rates of change of ROI volume and on the order of 10^{-3} for differences in ROI

volume between the subgroups at $t=0$. The right hippocampus in the comparison of AD-Language vs. AD-Memory and left hippocampus in the comparison of AD-Memory vs. AD-Visuospatial were exceptions for the analysis of average differences in ROI volumes between the subgroups; the associated p-values were in the 10^{-4} range, although still higher than the exact threshold given by the Holm-Bonferroni correction.

Additionally, in my analysis scheme for longitudinal data analysis, I considered whether it was appropriate to include timepoints that may have a very small number of individuals from a given AD subgroup. Such timepoints may not represent the population dynamics appropriately. I carried out an analysis to explore the sensitivity of LME models to removal of timepoints with a number of individuals below a threshold (p_{\min}) for a given AD subgroup. For each ROI and for each binary comparison of AD subgroups, I carried out five parallel analyses using LME modeling on datasets corresponding to p_{\min} values of 1, 2, 3, 4 and 5. As expected, there was variation in the coefficients of interest and the associated p-values across the models from different p_{\min} datasets. Qualitatively, the variation did not seem to be large, and one could say that the LME models' results are not too sensitive to the exclusion/inclusion of specific timepoints in the current longitudinal dataset. However, a quantitative method is needed to assess whether this variation is large or small. Further, in future work, it could be useful to develop a metric to determine the value of p_{\min} at which the model results (coefficient estimates) begin to stabilize, i.e. the difference in results between two consecutive p_{\min} values is sufficiently small. The final results of ROI importance that I reported were based on a collective list of top 5 ROIs from each of the five p_{\min} dataset's analysis. In the final summarized results for the important ROIs for each binary comparison of AD subgroups, I used average values based on all five p_{\min} datasets for differences in rate of change of volume between the AD subgroups and

differences in volume at $t=0$. If one wishes to focus on a specific p_min dataset, my analysis also provided separate results for each of the p_min datasets for both the differences in AD subgroups for rates of change of ROI volume and differences in ROI volume at $t=0$.

In future work, the mixed effects models used in the longitudinal data analysis could be improved by allowing rates of change of ROI volume to vary between different groups in the dataset besides just AD subgroups. This would mean allowing rates of change to vary with gender, age, education levels and *APOE* genotype status. Although linear models seemed to be appropriate for the current dataset (which had small sample sizes and noise), non-linear models may be considered in the future as more data becomes available and if the data suggests non-linear relationships between the covariates and ROI volume. In the future, one can also consider joint models of ROIs in the mixed effects modeling framework to account for potential correlations among ROIs. Given the small sample size of the current dataset and the noise in the ROI volume trajectories, the current modeling approach of starting with a simpler ROI volume model and using a model for each ROI separately seemed to be a good starting point to learn about ROI importance from this dataset. I used model convergence as the criteria for using the results from the current LME models. However, in future work, it would be useful to have a metric to formally assess the goodness of fit for each of the models.

Overall, my dissertation work aimed to shed light on which brain regions are most important for distinguishing between pairs of cognitively defined AD subgroups. The major contribution of this work towards AD research is in a preliminary hypothesis generation for which ROIs may be important to explore in future neuropathological studies for precision medicine approaches in treating AD. The list of ROIs is summarized in Tables 7.1 and 7.2 above. My work is an application of data science methods in a new domain: AD subgroups' differences. The dataset

that I worked with was challenging to draw easy conclusions from as it consisted of small samples sizes of AD subgroups, imbalanced sizes for the AD subgroups and noise in ROI volumes. The biomedical informatics niche of my dissertation work was in implementing appropriate mathematical and data science methods to “find the needle in a haystack.” Instead of an off the shelf application of data science methods, my work demonstrated a careful assessment and implementation of the methods to learn as much as possible from the current data despite the challenges of the data, in both cross-sectional and longitudinal datasets. I used different methods for analyzing cross-sectional and longitudinal data. In cross-sectional data analysis, I interpreted variable importance based on random forest classification models while in longitudinal data analysis, I used linear mixed effects modeling of each ROI volume to understand which ROIs have the most substantial differences between pairs of AD subgroups. The choices of methods were carefully based on the characteristics of data. My work in cross-sectional data analysis may be one of the first in documenting classification accuracies for distinguishing between pairs of AD subgroups based on ROI volumes from the time of AD diagnosis. These classification accuracies could be used as a benchmark in developing better classification models in the future for interpreting ROI importance from the models with more confidence. Although there are many avenues of improvement in the current work as discussed in the above paragraphs, there is a level of confidence in the results I’ve presented. One demonstration of this confidence is that my analysis using random forest for cross-sectional data and linear mixed effects modeling for longitudinal data pointed at some common ROIs as the most important ROIs for a given comparison of AD subgroups at the time of AD diagnosis. My dissertation work is a starting point for hypothesis generation of which ROIs may be useful to explore in future neuropathological studies for the purposes of understanding AD subgroup differences. I expect

the results from this work to get more refined and consolidated as more data may become available in the future.

As a budding scientist who has had the opportunity to work at the intersection of data science methods and Alzheimer's disease data, through my Ph.D. dissertation, I have developed a sense of appreciation and excitement for data science and informatics tools to help answer questions in AD research. At the same time, I have developed a humble perspective of the limitations of many analysis methods in dealing with real-world challenges of biomedical data from an ongoing study of individuals with AD. As someone who was provided access to structural MRI data and other data from the ADNI study which I used in my dissertation work, I am grateful and appreciative of each and every individual who volunteered for their data to be collected in the ongoing ADNI study, especially given the cognitive challenges of AD. I am also thankful for the many collaborators because of whom I was able to work with the data that I did including members of the ADNI project and collaborators at the University of Washington, Indiana University and Vrije Universiteit (VU) Amsterdam. I look forward to keeping up with continued future work at the intersection of AD subgroups and data science & mathematical methods, and hope that my dissertation has served as a good example of this inter-disciplinary work.

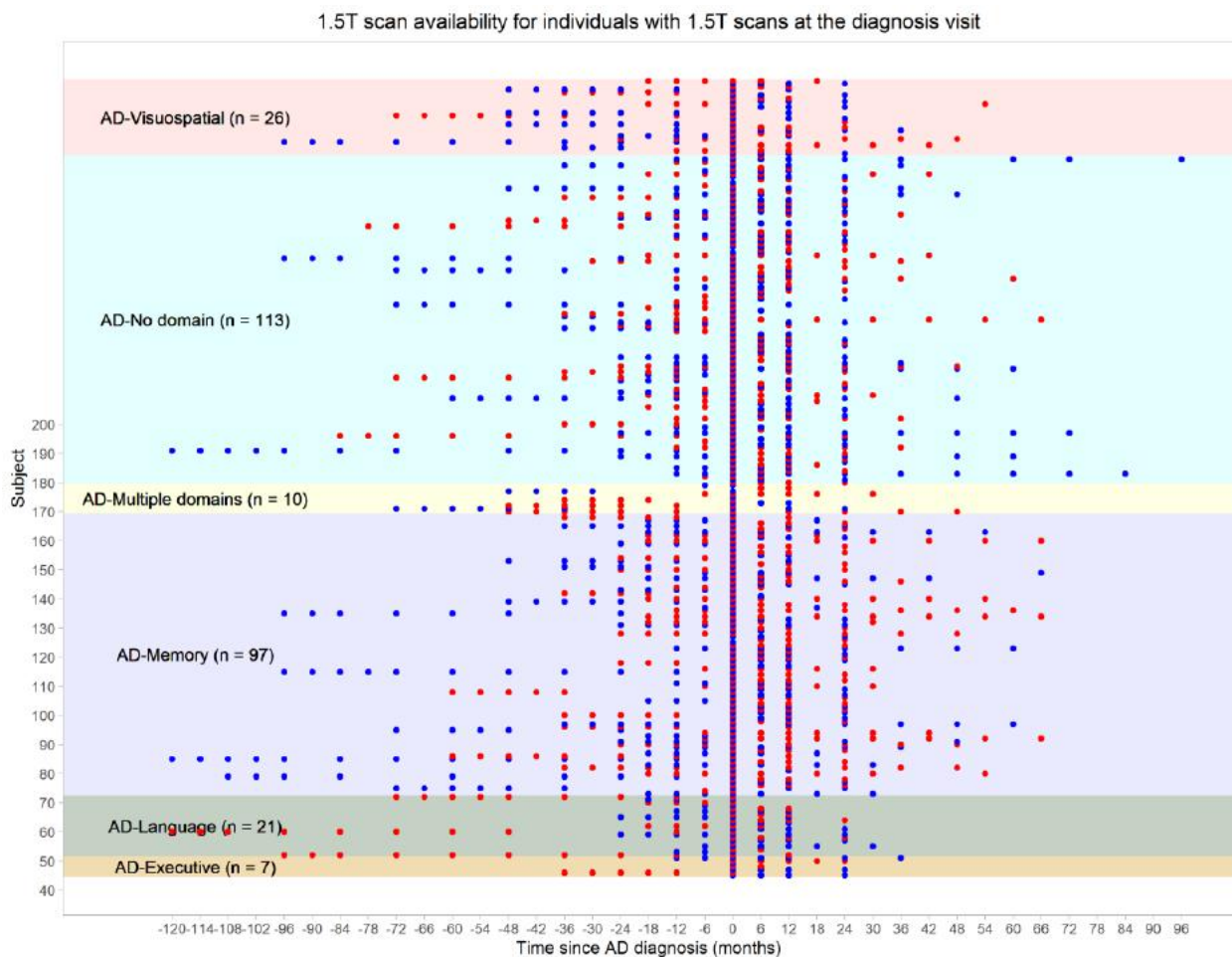
Appendix A: Supplemental materials (Code and files)

All code and scripts that I wrote and used in Aim 1 and Aim 2 as well as the supplementary files mentioned in this document will be uploaded to the following Github page: <https://github.com/sohih>

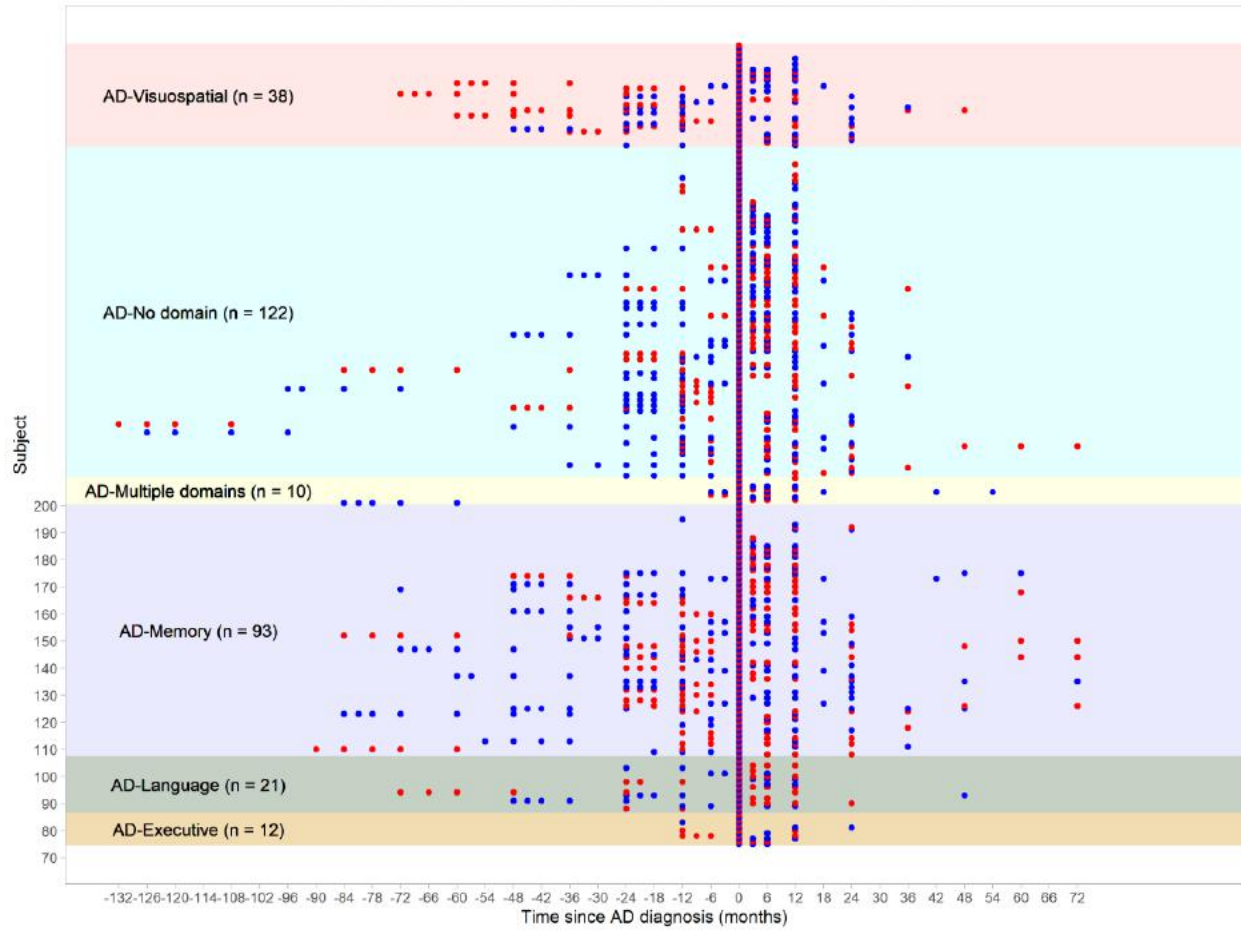
Supplementary files are also being submitted along with the submission of this document to the University of Washington Graduate School. These files include descriptive plots for cross-sectional data for all 70 ROIs (Chapter 2), cross-sectional data violin plot distribution comparison plots for pairs of AD subgroups for all 70 ROIs (Chapter 3), longitudinal ROI volume trajectories for all 70 ROIs for each of the AD subgroups after Part 1 of Aim 2 workflow (Chapter 5).

Appendix B: Longitudinal data descriptive plots

Plots showing data availability over time for all six AD subgroups are provided here.



3T scan availability for individuals with 3T scans at the diagnosis visit



References

- “2021 Alzheimer’s Disease Facts and Figures.” 2021. *Alzheimer’s & Dementia* 17 (3): 327–406. <https://doi.org/10.1002/alz.12328>.
- Bernal-Rusiel, Jorge L., Douglas N. Greve, Martin Reuter, Bruce Fischl, Mert R. Sabuncu, and Alzheimer’s Disease Neuroimaging Initiative. 2013. “Statistical Analysis of Longitudinal Neuroimage Data with Linear Mixed Effects Models.” *NeuroImage* 66 (February): 249–60. <https://doi.org/10.1016/j.neuroimage.2012.10.065>.
- Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16 (June): 321–57. <https://doi.org/10.1613/jair.953>.
- “Connectome Programs | Blueprint.” n.d. Accessed August 4, 2022. <https://neuroscienceblueprint.nih.gov/human-connectome/connectome-programs>.
- Crane, Paul K., and for the ADNI Psychometrics Study Group. 2020. “Cognitively Defined Subtypes of Alzheimer’s Disease.” *Alzheimer’s & Dementia* 16 (S6): e040621. <https://doi.org/10.1002/alz.040621>.
- Crane, Paul K., Emily Trittschuh, Shubhabrata Mukherjee, Andrew J. Saykin, R. Elizabeth Sanders, Eric B. Larson, Susan M. McCurry, et al. 2017. “Incidence of Cognitively-Defined Late-Onset Alzheimer’s Dementia Subgroups from a Prospective Cohort Study.” *Alzheimer’s & Dementia : The Journal of the Alzheimer’s Association* 13 (12): 1307–16. <https://doi.org/10.1016/j.jalz.2017.04.011>.
- Desikan, Rahul S., Linda K. McEvoy, Wes Thompson, Dominic Holland, J. Cooper Roddey, Kaj Blennow, Paul S. Aisen, James B. Brewer, Bradley T. Hyman, and Anders M. Dale. 2011. “Amyloid- β Associated Volume Loss Occurs Only in the Presence of Phospho-Tau.” *Annals of Neurology* 70 (4): 657–61. <https://doi.org/10.1002/ana.22509>.
- Desikan, Rahul S., Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, et al. 2006. “An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans into Gyral Based Regions of Interest.” *NeuroImage* 31 (3): 968–80. <https://doi.org/10.1016/j.neuroimage.2006.01.021>.
- Destrieux, Christophe, Bruce Fischl, Anders Dale, and Eric Halgren. 2010. “Automatic Parcellation of Human Cortical Gyri and Sulci Using Standard Anatomical Nomenclature.” *NeuroImage* 53 (1): 1–15. <https://doi.org/10.1016/j.neuroimage.2010.06.010>.
- Dickerson, Bradford C, and David A Wolk. 2011. “Dysexecutive versus Amnesic Phenotypes of Very Mild Alzheimer’s Disease Are Associated with Distinct Clinical, Genetic and Cortical Thinning Characteristics.” *Journal of Neurology, Neurosurgery, and Psychiatry* 82 (1): 45–51. <https://doi.org/10.1136/jnnp.2009.199505>.
- Dubois, Bruno, Howard H. Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T. DeKosky, et al. 2014. “Advancing Research Diagnostic Criteria for Alzheimer’s Disease: The IWG-2 Criteria.” *The Lancet. Neurology* 13 (6): 614–29. [https://doi.org/10.1016/S1474-4422\(14\)70090-0](https://doi.org/10.1016/S1474-4422(14)70090-0).
- Ferreira, Daniel, Chloë Verhagen, Juan Andrés Hernández-Cabrera, Lena Cavallin, Chun-Jie Guo, Urban Ekman, J.-Sebastian Muehlboeck, et al. 2017. “Distinct Subtypes of Alzheimer’s Disease Based on Patterns of Brain Atrophy: Longitudinal Trajectories and Clinical Applications.” *Scientific Reports* 7 (April): 46263. <https://doi.org/10.1038/srep46263>.

- Fields, R. Douglas. 2010. "Change in the Brain's White Matter." *Science (New York, N.Y.)* 330 (6005): 768–69. <https://doi.org/10.1126/science.1199139>.
- Fjell, Anders M., Kristine B. Walhovd, Christine Fennema-Notestine, Linda K. McEvoy, Donald J. Hagler, Dominic Holland, James B. Brewer, and Anders M. Dale. 2009. "One-Year Brain Atrophy Evident in Healthy Aging." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (48): 15223–31. <https://doi.org/10.1523/JNEUROSCI.3252-09.2009>.
- "FreeSurfer." n.d. FreeSurfer. Accessed August 4, 2022. <https://surfer.nmr.mgh.harvard.edu>.
- Girden, Ellen R. 1992. *ANOVA: Repeated Measures*. ANOVA: Repeated Measures. Thousand Oaks, CA, US: Sage Publications, Inc.
- Henderson, C. R., Oscar Kempthorne, S. R. Searle, and C. M. von Krosigk. 1959. "The Estimation of Environmental and Genetic Trends from Records Subject to Culling." *Biometrics* 15 (2): 192–218. <https://doi.org/10.2307/2527669>.
- Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. n.d. "Ctree: Conditional Inference Trees," 34. International, Alzheimer's Disease, and McGill University. 2021. "World Alzheimer Report 2021: Journey through the Diagnosis of Dementia," September. <https://www.alzint.org/resource/world-alzheimer-report-2021/>.
- Izquierdo, I, and J H Medina. 1993. "Role of the Amygdala, Hippocampus and Entorhinal Cortex in Memory Consolidation and Expression." *Brazilian Journal of Medical and Biological Research = Revista Brasileira de Pesquisas Medicas e Biologicas* 26 (6): 573–89.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, eds. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer.
- Jenkins, R., N. C. Fox, A. M. Rossor, R. J. Harvey, and M. N. Rossor. 2000. "Intracranial Volume and Alzheimer Disease: Evidence against the Cerebral Reserve Hypothesis." *Archives of Neurology* 57 (2): 220–24. <https://doi.org/10.1001/archneur.57.2.220>.
- Lam, Benjamin, Mario Masellis, Morris Freedman, Donald T. Stuss, and Sandra E. Black. 2013. "Clinical, Imaging, and Pathological Heterogeneity of the Alzheimer's Disease Syndrome." *Alzheimer's Research & Therapy* 5 (1): 1. <https://doi.org/10.1186/alzrt155>.
- Lawrence, Ross M., Eric W. Bridgeford, Patrick E. Myers, Ganesh C. Arvapalli, Sandhya C. Ramachandran, Derek A. Pisner, Paige F. Frank, Allison D. Lemmer, Aki Nikolaidis, and Joshua T. Vogelstein. 2021. "Standardizing Human Brain Parcellations." *Scientific Data* 8 (1): 78. <https://doi.org/10.1038/s41597-021-00849-3>.
- "Lecture_2_-_mri_as_a_black_box.Pdf." n.d. Accessed August 4, 2022. https://www.weizmann.ac.il/chembiophys/assaf_tal/sites/chemphys.assaf_tal/files/upload_s/lecture_2_-_mri_as_a_black_box.pdf.
- Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by RandomForest." *R News*, 2002.
- Little, Todd D. 2013. *Longitudinal Structural Equation Modeling*. Longitudinal Structural Equation Modeling. New York, NY, US: Guilford Press.
- McCarty, Catherine A., Rex L. Chisholm, Christopher G. Chute, Iftikhar J. Kullo, Gail P. Jarvik, Eric B. Larson, Rongling Li, et al. 2011. "The EMERGE Network: A Consortium of Biorepositories Linked to Electronic Medical Records Data for Conducting Genomic Studies." *BMC Medical Genomics* 4 (13). <https://doi.org/10.1186/1755-8794-4-13>.

- Mingers, John. 1987. "Expert Systems-Rule Induction with Statistical Data." *The Journal of the Operational Research Society* 38 (1): 39–47. <https://doi.org/10.2307/2582520>.
- Moeller, Julia. 2015. "A Word on Standardization in Longitudinal Studies: Don't." *Frontiers in Psychology* 6 (September): 1389. <https://doi.org/10.3389/fpsyg.2015.01389>.
- Mowinckel, Athanasia M., and Didac Vidal-Piñeiro. 2020. "Visualization of Brain Statistics With R Packages *Ggseg* and *Ggseg3d*." *Advances in Methods and Practices in Psychological Science* 3 (4): 466–83. <https://doi.org/10.1177/2515245920928009>.
- Ossenkoppele, Rik, Brendan I. Cohn-Sheehy, Renaud La Joie, Jacob W. Vogel, Christiane Möller, Manja Lehmann, Bart N. M. van Berckel, et al. 2015. "Atrophy Patterns in Early Clinical Stages across Distinct Phenotypes of Alzheimer's Disease." *Human Brain Mapping* 36 (11): 4421–37. <https://doi.org/10.1002/hbm.22927>.
- Parkin, Alan J. 1996. "Human Memory: The Hippocampus Is the Key." *Current Biology* 6 (12): 1583–85. [https://doi.org/10.1016/S0960-9822\(02\)70778-1](https://doi.org/10.1016/S0960-9822(02)70778-1).
- "Preliminaries — Introduction to MRI." n.d. Accessed August 4, 2022. <http://jpeelle.net/mri/general/preliminaries.html>.
- Purves, Dale, George J. Augustine, David Fitzpatrick, Lawrence C. Katz, Anthony-Samuel LaMantia, James O. McNamara, and S. Mark Williams. 2001. "The Ventricular System." *Neuroscience. 2nd Edition*. <https://www.ncbi.nlm.nih.gov/books/NBK11083/>.
- "R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>." n.d.
- Saria, Suchi, and Anna Goldenberg. 2015. "Subtyping: What It Is and Its Role in Precision Medicine." *IEEE Intelligent Systems* 30 (4): 70–75. <https://doi.org/10.1109/MIS.2015.60>.
- Scheltens, Nienke M. E., Francisca Galindo-Garre, Yolande A. L. Pijnenburg, Annelies E. van der Vlies, Lieke L. Smits, Teddy Koene, Charlotte E. Teunissen, et al. 2016. "The Identification of Cognitive Subtypes in Alzheimer's Disease Dementia Using Latent Class Analysis." *Journal of Neurology, Neurosurgery, and Psychiatry* 87 (3): 235–43. <https://doi.org/10.1136/jnnp-2014-309582>.
- Siriseriwan, Wacharasak. 2019. "Smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE." <https://CRAN.R-project.org/package=smotefamily>.
- Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9 (July): 307. <https://doi.org/10.1186/1471-2105-9-307>.
- Takehara-Nishiuchi, Kaori. 2014. "Entorhinal Cortex and Consolidated Memory." *Neuroscience Research* 84 (July): 27–33. <https://doi.org/10.1016/j.neures.2014.02.012>.
- Teng, E. L., K. Hasegawa, A. Homma, Y. Imai, E. Larson, A. Graves, K. Sugimoto, T. Yamaguchi, H. Sasaki, and D. Chiu. 1994. "The Cognitive Abilities Screening Instrument (CASI): A Practical Test for Cross-Cultural Epidemiological Studies of Dementia." *International Psychogeriatrics* 6 (1): 45–58; discussion 62. <https://doi.org/10.1017/s1041610294001602>.
- Vardy, Emma R. L. C., Andrew H. Ford, Peter Gallagher, Rosie Watson, Ian G. McKeith, Andrew Blamire, and John T. O'Brien. 2013. "Distinct Cognitive Phenotypes in Alzheimer's Disease in Older People." *International Psychogeriatrics* 25 (10): 1659–66. <https://doi.org/10.1017/S1041610213000914>.
- Ward, Andrew M., Aaron P. Schultz, Willem Huijbers, Koene R.A. Van Dijk, Trey Hedden, and Reisa A. Sperling. 2013. "The Parahippocampal Gyrus Links the Default-mode Cortical

- Network with the Medial Temporal Lobe Memory System.” *Human Brain Mapping* 35 (3): 1061–73. <https://doi.org/10.1002/hbm.22234>.
- “What Happens to the Brain in Alzheimer’s Disease?” n.d. National Institute on Aging. Accessed March 7, 2022. <http://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease>.
- “What Is Dementia? Symptoms, Types, and Diagnosis.” n.d. National Institute on Aging. Accessed March 7, 2022. <http://www.nia.nih.gov/health/what-is-dementia>.
- Yang, Sherry. 2019. “The Relationship between Cancer Subtype and Treatment: New Development Using Breast Cancer Subtypes to Predict Clinical Outcomes.” *Research Outreach*, no. 108 (July): 26–29. <https://doi.org/10.32907/RO-108-2629>.
- Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2): 301–20.