The Mechanistic Association between Cognition and Alzheimer's Disease using Machine Learning methods.

Riddhik Basu[1,*] and Arkaprava Roy[2,0]

[1] *Raleigh, North Carolina 27695, North Carolina State University, USA*
[2] *Gainesville, Florida 32603, University of Florida, USA*

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, structural brain changes, and genetic predispositions. This study leverages machine learning and statistical techniques to investigate the mechanistic relationships between cognitive function, genetic markers, and neuroimaging biomarkers in AD progression. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we perform both low-dimensional and high-dimensional analyses to identify key predictors of disease states, including cognitively normal (CN), mild cognitive impairment (MCI), and AD. Our low-dimensional approach utilizes multiple linear and ordinal logistic regression to examine the influence of cognitive scores, cerebrospinal fluid (CSF) biomarkers, and demographic factors on disease classification. The results highlight significant associations between Mini-Mental State Examination (MMSE), Clinical Dementia Rating Sum of Boxes (CDRSB), and phosphorylated tau levels in predicting cognitive decline. The high-dimensional analysis employs Sure Independence Screening (SIS) and LASSO regression to reduce dimensionality and identify genetic markers correlated with cognitive impairment and white matter integrity. Genes such as CLIC1, NAB2, and TGFBR1 emerge as significant predictors across multiple analyses, linking genetic expression to neurodegeneration. Additionally, imaging genetic analysis reveals shared genetic influences across brain hemispheres and the corpus callosum, suggesting distinct genetic contributions to white matter degradation. These findings enhance our understanding of AD pathology by integrating cognitive, genetic, and imaging data. Future research should explore longitudinal analyses and potential gene-environment interactions to further elucidate the biological mechanisms underlying AD progression.

Keywords
Alzheimer's Disease; Cognition; Gene Expression

[*] Corresponding author. Email: rbasu2@ncsu.edu or arkaprava.roy@ufl.edu.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that impairs memory, thinking, and behavior, leading to severe cognitive decline over time. It is the most common cause of dementia, accounting for 60–80% of all dementia cases (Irwin et al., 2018). AD typically begins with mild memory loss and gradually progresses to more severe impairments, ultimately impacting a person's ability to perform everyday tasks and maintain independence. The key pathological hallmarks of AD include the accumulation of amyloid-beta plaques and tau tangles in the brain, leading to neuronal damage and cognitive decline (DeTure and Dickson, 2019). Currently, there is no known cure, but early diagnosis and interventions can help manage symptoms and improve the quality of life for affected individuals.

Cognitive impairment presents itself in several stages, beginning with Cognitively Normal (CN), where individuals do not display any significant cognitive issues. This is followed by Mild Cognitive Impairment (MCI), a transitional stage between normal aging and Alzheimer's disease. MCI can be further categorized into early (EMCI) and late (LMCI) stages, based on the degree of impairment (Aisen et al., 2015). In its advanced form, cognitive decline progresses to AD, where patients experience significant memory loss, impaired reasoning, and behavioral disturbances. Neuropsychological exams are essential tools for assessing cognitive function and directly quantifying cognitive in diagnosed patients. These exams evaluate various domains such as memory, attention, executive function, and language, providing valuable insights for distinguishing normal aging, and differences between MCI versus AD. Examples of these exams include the Alzheimer's Disease Assessment Scores (ADAS), the Mini-Mental State Exam (MMSE), and more. As a commonly used assessment, the MMSE is scored on a 30-point scale, with higher scores indicating better cognitive function. Scores above 27 suggest normal cognition, scores between 24–27 indicate possible MCI, and scores below 24 typically indicate more severe impairments, such as in AD (Arevalo-Rodriguez, 2015).

Dementia is associated with notable structural changes in the brain, including atrophy of the hippocampus and cortical regions responsible for memory and cognition (Wiseman et al., 2004). White matter integrity is often disrupted, and diffusion tensor imaging (DTI) studies have identified changes in Fractional Anisotropy (FA), a measure of white matter microstructure. Lower FA values reflect impaired neuronal connectivity, which is commonly observed in patients with AD and MCI (Tae et al., 2018). These structural abnormalities correspond to cognitive decline, making FA an important biomarker for studying disease progression.

Genetic research plays a crucial role in understanding the development and progression of Alzheimer's disease. Variants in genes such as apolipoprotein-E (APOE) allele have been strongly associated with increased risk for AD. Beyond APOE, genome-wide association studies (GWAS) have identified other genes involved in amyloid processing, neuroinflammation, and lipid metabolism that may contribute to AD pathology (Alzheimer Disease Genetics Consortium , ADGC). Exploring these genetic markers helps identify individuals at higher risk, informs personalized treatments, and advances the understanding of disease mechanisms.

In this paper, we conduct multiple analyses to better understand the relationship between cognitive function, genetic markers, and structural brain changes. First, we apply statistical techniques to assess the influence of key biomarkers on disease states (CN, MCI, AD.) Next, we perform an association study between genetic markers and disease states to identify relevant genetic factors. We then investigate the relationship between FA values and genetic markers to explore how brain structure is influenced by genetic predisposition. Finally, we analyze common genes across disease states to identify shared pathways contributing to cognitive decline.

In Section 2, we describe the methodology, including data collection, and preprocessing steps, followed by Section 3, introducing our statistical methods to be used in the analysis. Section 4 discusses the findings of our overall genetic and imaging analyses, as well as their implications. Finally, in Section 5, we conclude with a summary of the key results including potential limitations and provide directions for future research.

2. ADNI-Data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is an ongoing longitudinal cohort study designed to develop clinical, imaging, and genetic markers for early detection and tracking of AD (Petersen, 2010). Eligible participants, aged 55-90 and in generally good health, were enrolled with either memory concerns or normal cognition. Each participant underwent comprehensive assessments, including cognitive testing, imaging, genetic evaluations, and both invasive and non-invasive medical procedures. Follow-up visits were conducted approximately every six months. The TADPOLE Challenge provided original data conducted from ADNI.

2.1 Cohort Description and Pre-Processing

Our study cohort consists of baseline visits from three cohorts: ADNI-1, ADNIGO, and ADNI-2 (Add phase-specific citations). All patients included in the analysis were de-identified and exhibited varying degrees of cognitive impairment. Although ADNI provides follow-up data, we focused exclusively on baseline data to better understand the population-level characteristics of cognitively impaired subjects.

For data preprocessing, visualization, and statistical analysis, we utilized a combination of tools. Preprocessing tasks were conducted using Python 3, employing packages such as Numpy (Harris, 2020), Pandas (McKinney, 2010), and Matplotlib (Hunter, 2007) for data manipulation. For statistical analysis, we used R statistical software version 4.3.3. The packages include Mass (Venables and Ripley, 2002) for ordinal logistic regression, Sure Independence Screening (SIS) for dimensionality reduction and glmnet (Friedman et al., 2010) for regularization techniques.

The primary goal of this study was to identify biomarkers associated with the acceleration of AD progression. After selecting baseline-only patients and removing incomplete or censored data, our final cohort consisted of 1,631 patients (819-ADNI1, 129-ADNIGO, and 683-ADNI2). Certain biological factors were excluded from analysis due to incomplete data or patient refusal. The selected predictors included cerebrospinal fluid (CSF) measurements, demographic data, neuropsychological tests, genetic and imaging markers:

Category	**Description**
Risk Factors	Age, years of education, and APOE-4 status (genetic factor)
Cognitive Exams	ADAS-Cog (ADAS-11 and ADAS-13), Clinical Dementia Rating Sum of Boxes (CDRSB), Mini Mental State Exam (MMSE), and Rey Auditory Verbal Learning Test (immediate, learning, and forgetting)
CSF Measures	Amyloid-beta, tau, and phosphorylated tau levels
Genetic Expression	Locus-Links, Probe Sets, Genes
Diffusion Tensor Imaging	Fractional Anisotropy (FA) values of the Corpus Callosum (CC), Left and Right hemispheres of the brain

Table 1: Summary of Key Measures and Variables

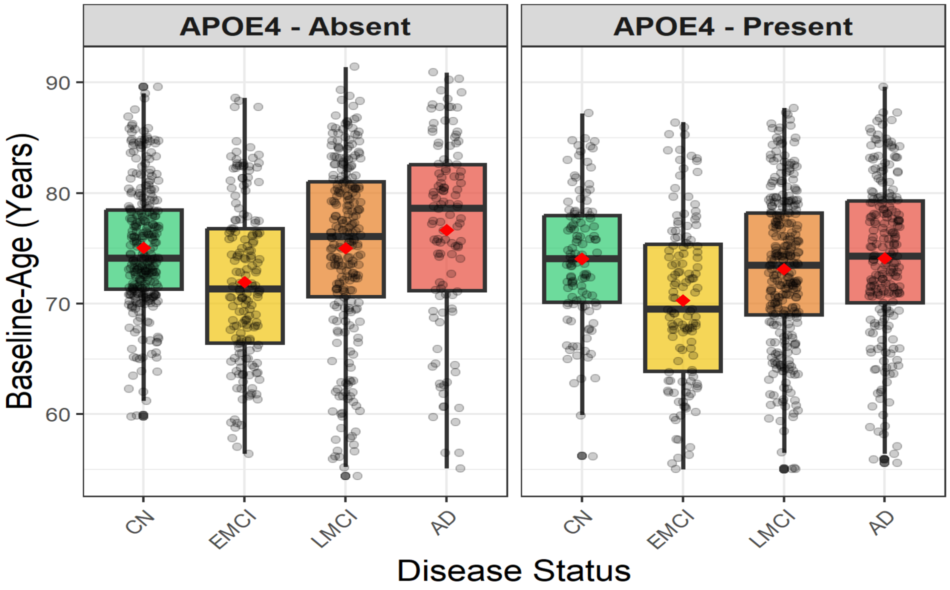

Figure 1: Add captions

Our secondary objective was to investigate genetic associations with disease progression. We analyzed gene expression data in relation to the biomarkers identified in our primary analysis and explored correlations between gene expression and brain structure changes by examining Fractional Anisotropic levels in diffusion tensor imaging (DTI) data. FA analysis was insightful in identifying clusters of white matter associated with rapid neurodegeneration (Poulakis, 2021).

The gene expression data include probe sets, Locus Link (Gene-ID), with a combined total of 49,386 genes. We filtered and merged the gene expression records with the rest of the data, yielding a sample of 468 patients with gene expression data and a sub-sample of 104 patients with both DTI imaging and genetic data.

3. Statistical Methods

We now review some of the low- and high-dimensional statistical methods that are considered in our analysis.

3.1 Low-Dimensional Analysis

3.1.1 Multiple Linear Regression

A multiple linear regression fits the following model for an outcome Y and an array of $p - 1$ predictors $X_1, X_2, ..., X_{p-1}$, assuming that the expectation of the error is 0, $E(e_i) = 0$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + e_i \tag{1}$$

For the purposes of this analysis, we prioritize identifying and retaining predictors that exhibit statistically significant associations with the outcome, ensuring the model's accuracy and interpretability while reducing the risk of overfitting. This careful selection of predictors is crucial for drawing valid inferences from the regression model. Later, we will augment this by implementing a penalty to shrink lesser-degree variables.

3.1.2 Ordinal Logistic Regression

The ordinal model is a statistical technique used to model the relationship between an ordinal dependent variable and independent variables. It utilizes logistic regression (like a link function) to measure the likelihood will fall into one of the categories. Our key assumption is that the relationship between predictors and log-odds of outcomes is proportional across all thresholds (proportional odds assumption.) (Yee, 2010)

$$log\left[\frac{P(Y \leqslant j)}{P(Y > j)}\right] = logit\left[P(Y \leqslant j)\right] \tag{2}$$

$$logit\big(P(Y \leqslant j)\big) = \beta_0 - n_1 x_1 - n_2 x_2 - ... - n_{p-1} x_{p-1}$$

Where: The function holds Y outcomes, j categories and $p-1$ predictors.

In our case, we can probabilistically assess how biomarkers, such as cognitive exams or genetic data, influence disease state progression. This allows us to understand and visualize how various biomarkers impact the likelihood of being in different stages of the disease, offering insights into the probabilistic nature of disease progression.

3.2 High-Dimensional Analysis

3.2.1 Screening

Due to the vast amount of genetic information, we reduce the dimensionality. To achieve this, we applied the Sure Independence Screening (SIS) method (Fan and Lv, 2008). SIS effectively selects a subset of the most important predictors by ranking them based on their marginal association with the outcome variable. By focusing on top-ranking features, SIS ensures that we do not lose important predictors, and we can then apply more refined variable selection techniques in a computationally efficient way.

We applied the Least Absolute Shrinkage and Selection Operator (LASSO) penalty-based regression methods. LASSO is particularly beneficial in our case because it automates model selection through its shrinkage effect, as well as preventing over-fitting (Tibshirani, 1996). It effectively reduces the coefficients of less significant genes to zero, concentrating on relevant, non-zero predictors.

We discuss the two models, considered in this paper. The first is the “multi-task learning” model, and the second is for multi-category outcomes.

3.2.2 Multivariate Response

The multi-response Gaussian family is particularly useful when there are several (correlated) responses, also known as the "multi-task learning" problem (Hastie et al., 2024). In this case, a covariate is either included in the model for all the responses or excluded for all the responses.

$$\begin{pmatrix} Y_{1,1} & \cdots & Y_{1,m} \\ \vdots & \ddots & \vdots \\ Y_{n,1} & \cdots & Y_{n,m} \end{pmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,m} \\ \vdots & \ddots & \vdots \\ \beta_{p,1} & \cdots & \beta_{p,m} \end{bmatrix} + \begin{pmatrix} \varepsilon_{1,1} & \cdots & \varepsilon_{1,m} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n,1} & \cdots & \varepsilon_{n,m} \end{pmatrix} \tag{3}$$

$$\to Y_{n\times m} = X_{n\times p}\, \beta_{p\times m} + \varepsilon_{n\times m}$$

The loss function consists of two main terms. The first term is the squared Frobenius norm of the residuals, measuring the difference between the observed and predicted responses across multiple outcomes. The second term is the LASSO penalty, which applies an ℓ_1 - norm regularization on the feature coefficients, encouraging sparsity by driving some coefficients to zero.

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{(p+1)\times K}} \quad \frac{1}{2N}\sum_{i=1}^{N} \|y_i - \beta_0 - \beta^T x_i\|_F^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_2 \tag{4}$$

where $\|\cdot\|_F$ and λ stand for

3.2.3 Multinomial Response

The multinomial model is particularly useful for classification problems where the response variable can take multiple categorical outcomes rather than just binary labels (Hastie et al., 2024). Unlike one-vs-all approaches, the multinomial model jointly models all class probabilities using the softmax function, ensuring that the probabilities sum to one. This allows the model to capture relationships between different categories rather than treating them independently.

Suppose the response variable has J levels, $G = 1, 2, \ldots, J$. Then our multinomial model is:

$$P(G = j \mid X = x) = \frac{e^{\beta_{0,j} + \beta_j^T x}}{\sum_{\ell=1}^{J} e^{\beta_{0,\ell} + \beta_\ell^T x}}. \tag{5}$$

Let Y be the $N \times K$ indicator response matrix, with elements $y_{i\ell} = I(g_i = \ell)$. Then the elastic net penalized negative log-likelihood function becomes:

$$\ell(\{\beta_{0k}, \beta_k\}_{k=1}^{K}) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{i,k}\log(\beta_{0k} + x_i^T\beta_k) - \log\left(\sum_{\ell=1}^{K} e^{\beta_{0\ell}+x_i^T\beta_\ell}\right)\right)\right] + \lambda\sum_{j=1}^{p}\|\beta_j\|_1. \tag{6}$$

The first term represents the negative log-likelihood for a multinomial classification model, which calculates the likelihood of observing the correct class labels, $y_{i,k}$, given the predicted probabilities modeled using the softmax function. The sum inside the logarithm computes the probability for each class, and the negative log-likelihood penalizes incorrect classifications. The second term is the LASSO penalty (ℓ_1-norm), which regularizes the model by enforcing sparsity in the coefficients. This term shrinks the coefficients, selecting only the most important features.

4. Results

With three distinct analytical objectives, we began with a full cohort of 1,631 patients, which was used for all low-dimensional analyses. From this cohort, we identified a subset of 468 patients with available genetic information. Among these, a further subset of 104 patients also had Fractional Anisotropy (FA) imaging data available. This sub-subset was used for imaging-informed kernel-based analyses. Across all levels of analysis, a p-value threshold of < 0.05 was used to define statistical significance.

List the R packages here with citations, saying 'We apply these R packages for our analysis.' Do not need to mention which is used where.

4.1 Low-dimensional Analysis

4.1.1 Regressing Disease State on Cognitive Scores

We regress disease states on a set of cognitive test scores and other risk factors using the ordinal regression model. Our selection of variables were taken from known demographic factors as well as the most common assessments given to dementia patients.

Figure 2 illustrates odds ratios in the log-scale from the ordinal regression model.

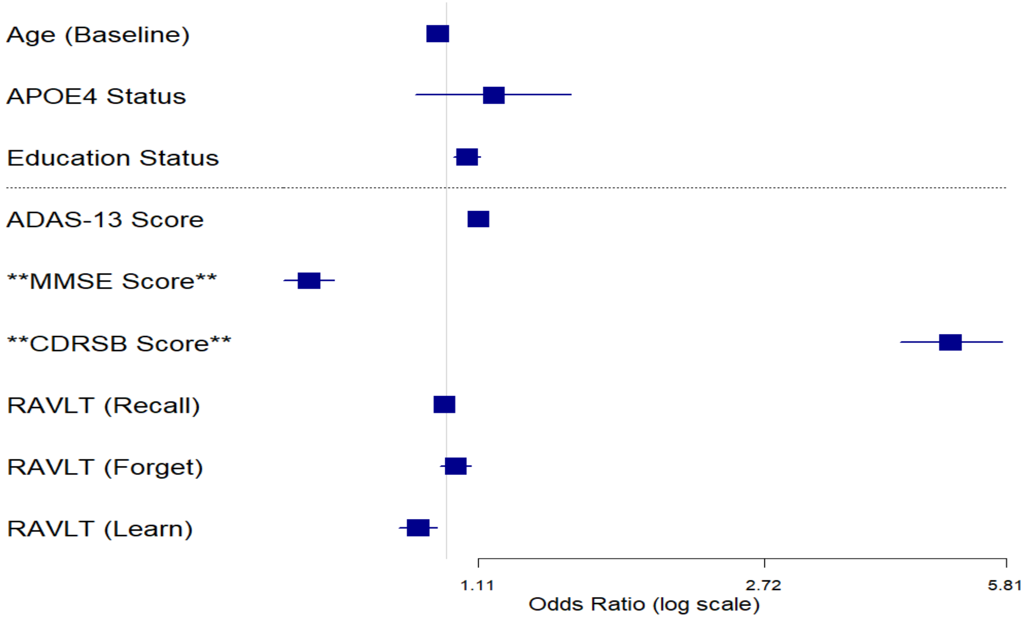

Figure 2: Add title

The results highlight statistically significant associations between several cognitive exam scores and disease progression, as visualized in Figure 2. Among the most influential predictors, the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating–Sum of Boxes (CDRSB) displayed the smallest and largest odds ratios, respectively. This is consistent with clinical expectations, as lower MMSE scores and higher CDRSB scores reflect worsening cognitive function and more advanced stages of Alzheimer's disease. In contrast, APOE4 carrier status showed a positive and statistically significant association with disease progression. However, variables such as education level, baseline age, and RAVLT (Rey Auditory Verbal Learning Test) scores did not reach statistical significance (p-values $>$ 0.05) in the multivariate model.

Notably, the MMSE and CDRSB scores demonstrate extreme magnitudes, indicating that each per-unit increase in MMSE results in a protective effect, while increments in CDRSB signal worsening cognitive impairment. Further analysis is required to investigate these biomarkers.

4.1.2 Cognitive Scores Analysis

Neuropsychological exams provide a direct measure of cognitive decline, assessing various skills such as memory, language, and visual processing. These tests help clinicians gauge a patient's overall cognitive awareness. In our analysis, we focus on the CDRSB and MMSE as response variables within a multivariable regression framework. We will include disease state as an explanatory variable to explore changes across individual conditions. Due to multicollinearity, no other neuropsychological tests will be incorporated into the model. We will first analyze the CDRSB to identify significant factors, followed by a regression of the MMSE.

Multivariable Regression Analysis of CDRSB

Table 2 presents the multivariable linear regression estimates for regressing the Clinical Dementia Rating scale Sum of Boxes (CDRSB) on the combined CSF measurements and disease states.

Table 2: Multivariable Linear Regression Estimates for CDRSB (N = 1,113)

Biomarker	**Estimate**	**Std. Error**	**T-value**	**P-Value**
Demographics				
Baseline Age (years)	-4.919 $\times 10^{-4}$	0.004	-0.119	0.905
APOE-4 Genetic Status	3.491 $\times 10^{-3}$	0.068	0.051	0.959
Disease State				
Early Mild CI	2.909	0.072	40.512	$< 2 \times 10^{-16}$
Late Mild CI	0.764	0.061	12.601	$< 2 \times 10^{-16}$
Alzheimer's Disease	0.768	0.058	13.363	$< 2 \times 10^{-16}$
Cerebrospinal Fluid (CSF)				
Phosphorylated Tau levels	5.108 $\times 10^{-3}$	0.002	2.265	0.024
Amyloid-Beta levels	-1.514 $\times 10^{-4}$	5.792 $\times 10^{-5}$	-2.614	0.009

The multivariable linear regression model assessing Clinical Dementia Rating Sum of Boxes (CDRSB) yielded an Adjusted R^2 of 0.699, indicating that approximately 70% of the variability in CDRSB is explained by the model. Central to this analysis were the cerebrospinal fluid (CSF) biomarkers, with Phosphorylated Tau levels (PTAU) and Amyloid-Beta levels (ABETA). Since the correlation between PTAU and TAU levels was around 98%, we only included PTAU.

Multivariable Regression Analysis of MMSE

Table 3 presents the multivariable linear regression estimates for regressing the Mini-Mental State Examination (MMSE) on the combined CSF measurements and disease states.

Table 3: Multivariable Linear Regression Estimates for MMSE (N = 1,113)

Biomarker	**Estimate**	**Std. Error**	**T-value**	**P-Value**
		Demographics		
Baseline Age (years)	-0.029	6.842×10^{-3}	-4.232	2.51×10^{-5}
APOE-4 Genetic Status	-0.072	0.113	-0.640	0.522
		Disease State		
Early Mild CI	-3.780	0.119	-31.843	$< 2 \times 10^{-16}$
Late Mild CI	-1.500	0.100	-14.960	$< 2 \times 10^{-16}$
Alzheimer's Disease	-0.594	0.095	-6.245	6.04×10^{-10}
		Cerebrospinal Fluid (CSF)		
Phosphorylated Tau levels	-0.012	0.004	-3.185	0.001
Amyloid-Beta levels	2.669×10^{-4}	9.575×10^{-5}	2.788	0.005

Similarly, the multivariable linear regression model for MMSE yielded an Adjusted R^2 of 0.618, indicating that 62% of the variability in MMSE is explained by the model. CSF biomarkers were again key, with PTAU showing a significant negative association with cognitive performance, highlighting its role in cognitive decline. Conversely, ABETA levels demonstrated a modest positive association with MMSE scores, consistent with their role in mitigating amyloid-related pathology.

4.2 High-Dimensional Analysis

4.2.1 Regressing CS on Gene Expression

Recognizing the importance of MMSE and CDRSB as key markers of cognitive decline, we examined their associations from a genetic perspective. To address the high dimensionality of the genetic data, we applied a marginal screening technique using the Sure Independence Screening (SIS) method, which ranks all predictors and retains only the top variables for further analysis. Using our genetic subsample (n = 468), we performed SIS with the default setting of 10 partitions (nfolds = 10). We then fit a linear model with a Gaussian family, regressing the selected cognitive biomarkers (MMSE and CDRSB) on the screened genetic features. Only genes with non-zero coefficients in this final model were considered as potential predictors of cognitive decline.

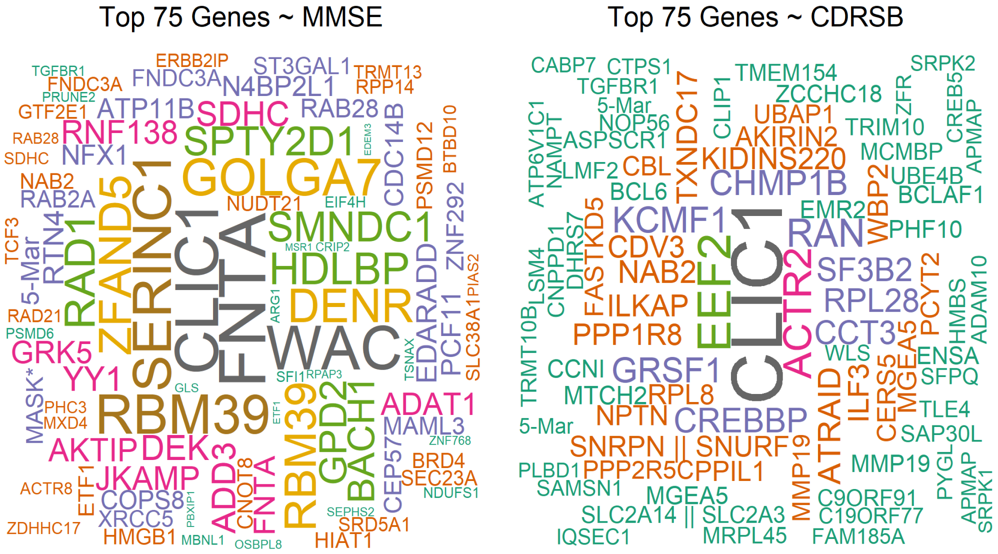

Figure 3: Word cloud of the top 75 genes selected via Sure Independence Screening with non-zero coefficients in a Gaussian model predicting cognitive decline.

Figure 3 represents the top 75 genes identified through this rigorous fitting process, with the size of each gene name reflecting its significance in relation to cognitive function. Modeled using MMSE and CDRSB as responses to capture cognitive decline across disease stages, four key genes-NAB2, 5-MAR, TGFBR1, and CLIC1-emerged as statistically significant in both models, with CLIC1 (Carlini, 2020) showing the greatest effect.

4.2.2 Regressing Disease State on Gene Expression

We repeated the analysis to identify key genes across disease progression stages. Since SIS does not support multinomial models, we treated each pairwise comparison between disease stages (CN, EMCI, LMCI, AD) as a separate binary classification. After SIS screening, we applied cross-validation using a multinomial model to capture genetic effects varying by disease stage. This approach allowed detection of genes significant in some transitions but not others. Finally, we compiled union and intersection sets of genes with non-zero coefficients across comparisons.

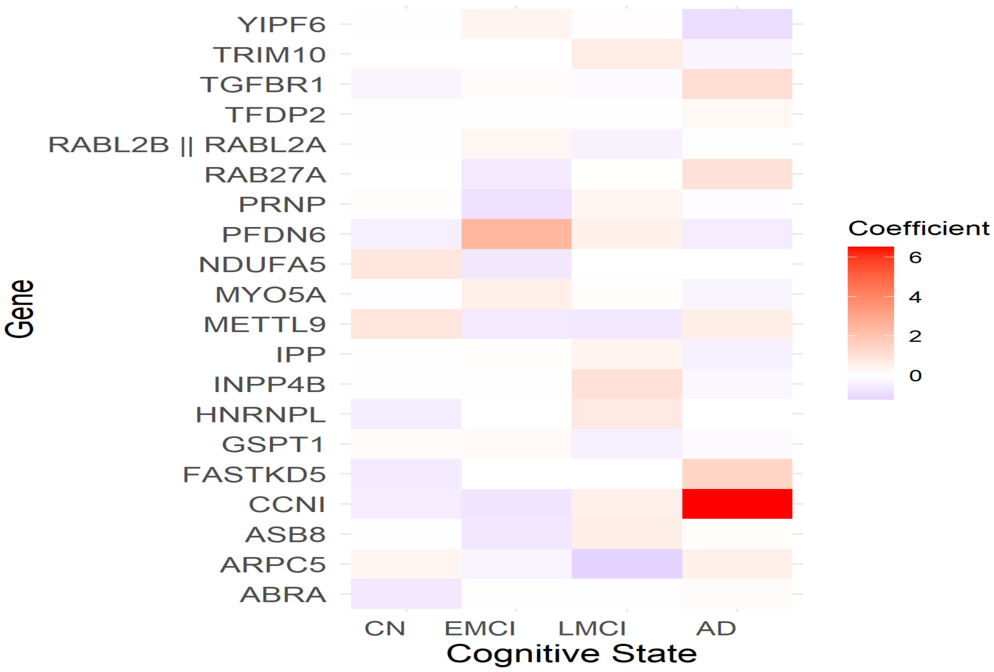

Figure 4

Figure 4 highlights the associations between intersecting genes and cognitive states, with color gradients indicating effect magnitude. Notably, CCNI shows a strong positive association with Alzheimer's Disease, while PRNP has moderate associations in early to late cognitive decline stages. In contrast, genes like YIPF6 and RAB27A show negligible associations. These findings suggest that a gene which may show high effect at earlier or later stages of impairment may display a null effect at different stages. As such, finding intersections of the predictors reinforce our notion that a significant effect occurs in underlying mechanisms.

In relation to the cognitive exam analyses from the previous step, we note that the following genes: TGFBR1, XRCC5 and CCNI were corelated for both disease progression and individual cognitive scorings. It is important to note that genetic effects may not reflect a constant effect as disease state progresses.

4.2.3 Imaging Genetic Analysis

For our image analysis, we categorized the brain sections into three groups: the Corpus Callosum (CC), Right Brain, and Left Brain. The CC connects the two cerebral hemispheres and is crucial for higher cognitive functions. It's often one of the first regions to show signs of neurodegeneration. The left hemisphere handles language processing, analytical thinking, and logic, while the right hemisphere is more involved in spatial awareness and emotional interpretation. Damage to these structures can lead to disruptions in the communication between the hemispheres, progressing neurodegenerative mechanisms.

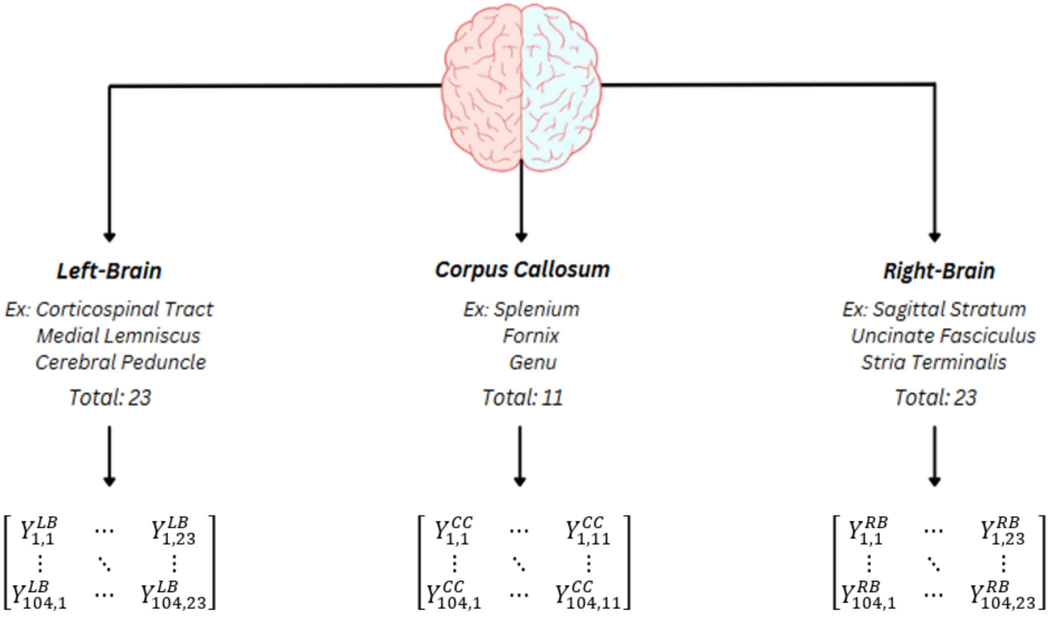
Left-Brain
Ex: Corticospinal Tract
Medial Lemniscus
Cerebral Peduncle
Total: 23
Corpus Callosum
Ex: Splenium
Fornix
Genu
Total: 11
Right-Brain
Ex: Sagittal Stratum
Uncinate Fasciculus
Stria Terminalis
Total: 23
$\begin{bmatrix} Y^{LB}_{1,1} & \cdots & Y^{LB}_{1,23} \\ \vdots & \ddots & \vdots \\ Y^{LB}_{104,1} & \cdots & Y^{LB}_{104,23} \end{bmatrix}$
$\begin{bmatrix} Y^{CC}_{1,1} & \cdots & Y^{CC}_{1,11} \\ \vdots & \ddots & \vdots \\ Y^{CC}_{104,1} & \cdots & Y^{CC}_{104,11} \end{bmatrix}$
$\begin{bmatrix} Y^{RB}_{1,1} & \cdots & Y^{RB}_{1,23} \\ \vdots & \ddots & \vdots \\ Y^{RB}_{104,1} & \cdots & Y^{RB}_{104,23} \end{bmatrix}$

Figure 5

We have 57 FA image responses, with values ranging from 0 to 1. For fitting genetic predictors to these image responses, we applied SIS and the LASSO penalty. To account for heavy collinearity, we first applied a logit transformation, and then split the responses into their respective brain categories and fitted using multi-gaussian. Due to the relatively small size of our subsample (n=104), we adjusted our cross-validation to 4 folds, (nfolds=4.) After the selection process, we fitted the significant genes across the 57 responses into Multivariate Gaussian models for each brain section. This resulted in three matrices, each corresponding to a set of non-zero gene predictors associated with a specific brain section.

In our final analysis, we set a threshold (e.g., the top 50 genes) for each brain section and identified predictors that intersected across multiple categories. This step involved examining the intersections with genetic predictors deemed significant from the initial genetic analysis. We observe many genetic indicators overlap across each of brain sections. For the Left and Right hemispheres, approximately 200 genes were associated with both, whereas 80 genes were strongly associated with both hemispheres and the Corpus Collasum.

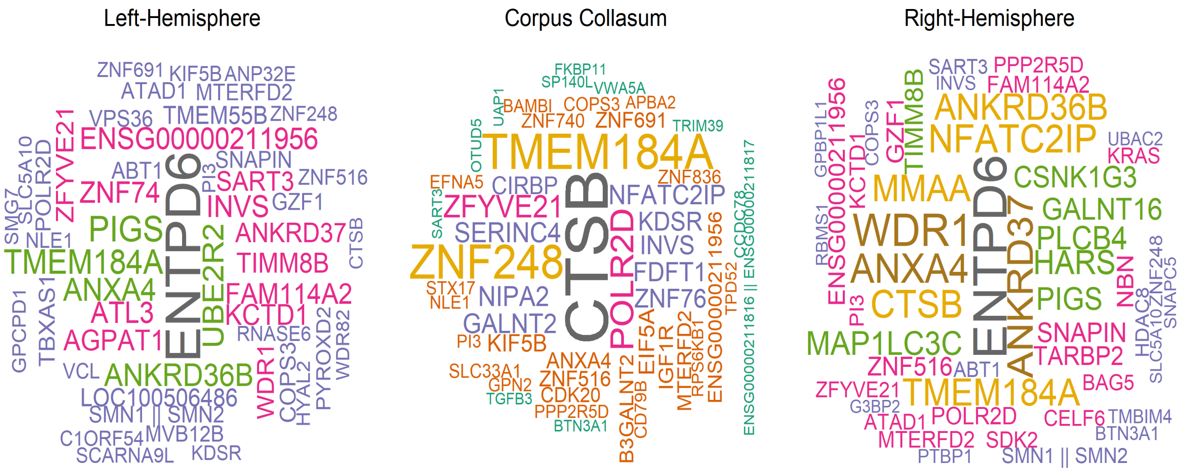

Figure 6

Figure 6 indicates the top 50 genes selected from the multi-gaussian fitting process, with the size of each gene name reflecting its significance in relation to white matter integrity. Using the FA measurements to capture genetic effect, many predictors emerged as statistically significant in all three brain regions. Some of the listed genes include: CTSB, TMEM184A, POLR2D, INVS, ANXA4, COPS3, PI3, and more.

Our results suggest that the top genes in the two hemispheres share more commonalities compared to those in the corpus callosum. For instance, ENTPD6 was identified as significant in measuring white matter in the hemispheres but showed no significance in the CC. This finding may indicate that the genetic mechanisms governing white matter in the hemispheres are more uniform or shared, reflecting their complementary roles in processing information and supporting localized brain functions. In contrast, the gene TMEM184A demonstrated significance across all three regions, to a lesser extent. Genes like TMEM184A, which are consistently significant across multiple brain regions, could serve as stronger indicators of white matter degradation and may hold greater potential as predictive markers.

5. Concluding Remarks and Future Directions

This study examines the complex relationships among cognitive decline, genetic markers, and other biomarkers in the progression of Alzheimer's Disease. By investigating biological and genetic contributions to cognitive impairment, our findings enhance the understanding of its pathology.

Our analysis identified the utility of CDRSB and MMSE scores in distinguishing disease states, with increasing CDRSB scores indicating greater severity and improvements in MMSE scores offering protective effects. These key biomarkers enabled us to adopt a genetic perspective in exploring AD progression. Using the Sure Independence Screening method, we simplified data complexity and identified significant genetic markers influencing cognitive decline. LASSO regression further highlighted key genes, including NAB2, 5-MAR, TGFBR1, and notably CLIC1, underscoring their role in neurodegeneration. Visual representations of top predictors emphasized the importance of integrating genetic data into AD research. Additionally, the differential effects of genetic markers across disease stages suggest potential gene-environment interactions and stage-specific susceptibilities, paving the way for further molecular studies.

Our findings carry significant implications for understanding the biological and genetic underpinnings of AD. For instance, the identification of CLIC1 aligns with growing evidence of its role in chronic central nervous system inflammation, highlighting its potential as both a biomarker and a therapeutic target. Similarly, the confirmation of TGFBR1 as a receptor in AD pathogenesis further supports its role in disease progression.

In our investigation of white matter imaging, we uncovered genetic commonalities across brain hemispheres and the corpus callosum. Genes such as CTSB and TMEM184A, found significant in all three brain regions, suggest a foundational role in maintaining white matter integrity throughout the brain, influencing intra- and inter-hemispheric communication. These genes hold promise as universal biomarkers or therapeutic targets for white matter-related diseases, offering insights into global brain function and pathology.

Future research should incorporate advanced neuroimaging techniques, such as refined measures of white matter integrity, to complement our analyses and capture neurochemical changes more effectively. Longitudinal studies would be invaluable for elucidating the temporal evolution of cognitive and genetic markers, providing a more comprehensive view of AD progression.

Additionally, exploring the interaction between genetic markers and environmental or lifestyle factors, such as physical activity and diet, may identify modifiable risk factors for AD.

Our approach demonstrates the value of integrating statistical modeling, genetic analysis, and biomarker data to unravel the complexities of neurodegenerative diseases. While our models explain a substantial portion of variability in cognitive outcomes, they underscore the multifactorial nature of AD. A holistic approach is essential to fully understand the mechanisms underlying the disease and to design effective preventive and therapeutic strategies.

References

Aisen PS, Petersen RC, Donohue M, Weiner MW, Initiative ADN (2015). Alzheimer's disease neuroimaging initiative 2 clinical core: Progress and plans. *Alzheimer's & Dementia*, 11(7): 734–739.

Alzheimer Disease Genetics Consortium (ADGC) ea (2019). Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates alpha-beta, tau and lipid processing. *Nature Genetics*, 51(3): 414–430.

Arevalo-Rodriguez Iea (2015). Mini-mental state examination (MMSE) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, 2015(3): CD010783.

Carlini Vea (2020). Clic1 protein accumulates in circulating monocyte membrane during neurodegeneration. *International Journal of Molecular Sciences*, 21(4): 1484.

DeTure MA, Dickson DW (2019). The neuropathological diagnosis of alzheimer's disease. *Molecular Neurodegeneration*, 14(1): 32.

Fan J, Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5): 849–911.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Harris CRea (2020). Array programming with numpy. *Nature*, 585(7825): 357–362.

Hastie T, Qian J, Tay K (2024). An introduction to glmnet. Online. Accessed: Nov. 24, 2024.

Hunter JD (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95.

Irwin K, Sexton C, Daniel T, Lawlor B, Naci L (2018). Healthy aging and dementia: Two roads diverging in midlife? *Frontiers in Aging Neuroscience*, 10: 275.

McKinney W (2010). Data structures for statistical computing in python. In: *Python in Science Conference*, 56–61. Austin, Texas.

Petersen RCea (2010). Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology*, 74(3): 201–209.

Poulakis Kea (2021). Longitudinal deterioration of white-matter integrity: heterogeneity in the ageing population. *Brain Communications*, 3(1): fcaa238.

Tae WS, Ham BJ, Pyun SB, Kang SH, Kim BJ (2018). Current clinical applications of diffusion-tensor imaging in neurological disorders. *Journal of Clinical Neurology (Seoul, Korea)*, 14(2): 129–140.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1): 267–288.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S.* Statistics and Computing. Springer, New York, 4th edition.

Wiseman RM, Saxby BK, Burton EJ, Barber R, Ford GA, O'Brien JT (2004). Hippocampal atrophy, whole brain volume, and white matter lesions in older hypertensive subjects. *Neurology*, 63(10): 1892–1897.

Yee TW (2010). Cumulative link models for ordinal regression. In: *Data Mining and Data Visualization* (C Rao, E Wegman, J Solka, eds.), volume 24 of *Handbook of Statistics*. Springer, New York, NY.

A Appendix section

Table 4: Patient Characteristics by Cognitive Status

Characteristic	**Cognitively Normal (417)**	**Early Mild CI (310)**	**Late Mild CI (562)**	**Diagnosed AD (342)**
Demographics				
Sex of Patient, n (%)				
Male	209 (50.12%)	171 (55.16%)	344 (61.21%)	189 (55.26%)
Female	208 (49.88%)	139 (44.84%)	218 (38.79%)	153 (44.74%)
Race of Patient, n (%)				
White	376 (90.17%)	286 (92.26%)	526 (93.59%)	317 (92.69%)
Black	30 (7.19%)	8 (2.58%)	22 (3.91%)	14 (4.09%)
Other	11 (2.64%)	16 (5.16%)	14 (2.42%)	11 (3.22%)
Formal Years of Education, Mean (SD)	16.28 (2.73)	15.96 (2.66)	15.88 (2.94)	15.18 (2.99)
Risk Factors				
Baseline Age (years), Mean (SD)	74.76 (5.73)	71.19 (7.50)	73.99 (7.50)	75.03 (7.79)
(Apolipoprotein E4) Genetic Status, n (%)				
Absent	301 (72.53%)	175 (57.19%)	256 (45.71%)	113 (33.43%)
Present	114 (27.47%)	131 (42.81%)	304 (54.29%)	225 (66.57%)
Cognitive Exams				
ADAS (13), Mean (SD)	9.34 (4.32)	12.65 (5.42)	18.66 (6.52)	29.87 (8.05)
CDRSB, Mean (SD)	0.03 (0.13)	1.29 (0.76)	1.65 (0.92)	4.39 (1.67)
MMSE, Mean (SD)	29.07 (1.12)	28.34 (1.56)	27.18 (1.80)	23.22 (2.07)
Rey Auditory General Test, Mean (SD)	44.35 (9.84)	39.55 (10.71)	31.32 (9.51)	22.82 (7.55)
Cerebrospinal Fluid (CSF) Measurements				
Amyloid-Beta levels, Mean (SD)	1,327.68 (660.95)	1,178.19 (587.49)	889.14 (490.26)	691.44 (416.43)
Tau levels, Mean (SD)	238.45 (88.96)	256.41 (121.74)	308.95 (128.96)	367.84 (144.92)
Phosphorylated Tau levels, Mean (SD)	22.00 (9.08)	24.25 (13.69)	30.46 (14.66)	36.66 (15.76)