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

Riddhik Basu[1] and Mentor: Arkaprava Roy[2]

1. North Carolina State University, Raleigh, North Carolina 27695, USA
2. Department of Biostatistics, University of Florida, Gainesville, Florida 32603, USA

November 2024

## Contents

# 1. Introduction

### 1) What is Alzheimer's disease?

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 (Reference). 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 (Reference). Currently, there is no known cure, but early diagnosis and interventions can help manage symptoms and improve the quality of life for affected individuals.

### 2) Different types of cognitive impairments: CN, MCI, AD

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 (Reference). 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 into an individual's cognitive abilities. They are particularly useful 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), Montreal Cognitive Assessment (MoCA) 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 (Reference).

### 3) Structural changes in the brain due to disease

Dementia is associated with notable structural changes in the brain, including atrophy of the hippocampus and cortical regions responsible for memory and cognition (Reference). 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 (Reference). These structural abnormalities correspond to cognitive decline, making FA an important biomarker for studying disease progression.

**4) Importance of genetic studies for AD**

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 (Reference). Exploring these genetic markers helps identify individuals at higher risk, informs personalized treatments, and advances the understanding of disease mechanisms.

**5) Overview of the analyses in this paper**

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), followed by further exploration of the most significant biomarkers. 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.

**6) Organization of the rest of the paper**

In Section 2, we describe the methodology, including data collection, and preprocessing steps, followed by Section 3, introducing our statistical framework. Section 4 discusses the findings of our overall, genetic and imaging analyses, as well as their implications, and potential limitations. Finally, in Section 5, we conclude with a summary of the key results and provide suggestions 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 Alzheimer’s Disease (AD). 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. 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. Initial data inspection and simple visualizations were performed using Microsoft Excel, allowing for easy manipulation. Preprocessing tasks were conducted using Python 3 within the Jupyter Notebook environment, employing packages such as *Numpy*, *Pandas, and Matplotlib* for data manipulation. For statistical analysis, we used R version 4.3.3 for its extensive libraries: *MASS, Hmisc and Foreign* for ordinal logistic regression, and *Forestplot* to generate confidence intervals. To handle high-dimensional data, we used Sure Independence Screening (*SIS*) for dimensionality reduction as well as *glmnet* 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:

- **Risk Factors**: Age, years of education, and APOE-4 status (genetic factor)
- **Cognitive Exams**: ADAS-Cog (ADAS11 and ADAS13), Clinical Dementia Rating 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
- **(DTI) Diffusion Tensor Imaging:** Fractional Anisotropy (FA) values of the Corpus Collasum (CC), Left and Right hemispheres of the brain.

*Table 1: Baseline Characteristics of Patients at Baseline across Diagnosis Group*

| ***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) |

*Figure 1. Age Distribution by APOE-4 Status across Disease Stage at Baseline.*

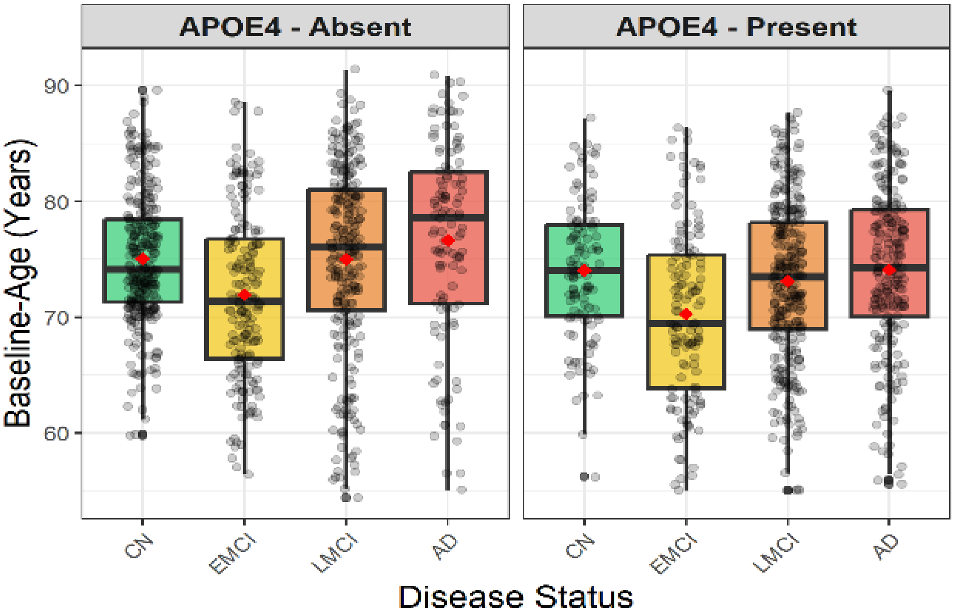


*Figure 1 visualizes baseline age across different disease states (CN, EMCI, LMCI, AD), stratified by APOE4 status (absent vs. present). Each boxplot represents the distribution of baseline ages for a given disease state, with the red diamonds indicating the average age within each group.*

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 Anisotropy (FA) in diffusion tensor imaging (DTI) data. FA analysis was insightful in identifying clusters of white matter associated with rapid neurodegeneration.

*The University of Southern California Neuroimaging and Informatics Institute* provided gene expression data, uniquely identified by probe-set, Locus Link (Gene-ID), and gene symbol. We filtered and merged the gene expression records with the imaging data and biomarkers, ultimately 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

Our primary aim is to first investigate biomarkers of significance, and then leverage them to identify gene expression levels with the greatest effect. To accomplish this, we break our methods into two groups: Low-Dimensional, and High-Dimensional.

The distinction between the two sections lies in their focus: the *Low-Dimensional* approach emphasizes identifying significant biomarkers using traditional statistical methods. In contrast, the vast amount of data for the *High-Dimensional* approach requires reduction. The latter will delve into more complex, data-intensive techniques.

## 3.1 Low-Dimensional Analysis:

### 3.1.1 Multiple Linear Regression

We use multiple linear regression as a tool to measure the relevance of predictors. When measuring multiple biomarkers against a single response, we need to filter out non-significant variables, as it would add additional noise and bias.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p \text{ ; \textit{with p predictors.}}$$

For the purposes of this analysis, we prioritize identifying and retaining only those 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.)

$$\log\left[\frac{P(Y \leq j)}{P(Y > j)}\right] = logit\big(P(Y \leq j)\big)$$

$$logit\big(P(Y \leq j)\big) = \beta_0 - n_1 x_1 - \cdots - n_p x_p \text{ ; with \textit{Y outcomes, j categories and p predictors.}}$$

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 Regression:

Due to the vast amount of genetic information, we need to reduce the dimensionality. To achieve this, we applied the Sure Independence Screening (SIS) method.

$$Y = X\beta + \varepsilon$$

*where* $Y \mid (Y_1, Y_2, \dots Y_n)^{T}$ *is an n-dimensional response vector,* $X \mid (x_1, x_2, \dots x_p)$ *is an n × p design matrix consisting of p covariates* $x_J$*'s,* $\beta \mid (\beta_1, \beta_2, \dots \beta_p)^{T}$ *is a p-dimensional regression coefficient vector, and* $\varepsilon \mid (\varepsilon_1, \varepsilon_2, \dots \varepsilon_n)^{T}$ *is an n-dimensional error vector.*

SIS effectively handles ultra-high dimensional data by ranking genetic features based on their marginal correlation with the outcome variable, allowing us to probabilistically assess the influence of each marker on disease progression. The method first screens the features, retaining only those with the strongest associations, reducing the complexity and noise in our data. By focusing on top-ranking features, SIS ensures that we do not lose important genetic markers, and we can then apply more refined variable selection techniques.

For both genetic and image analyses, we applied *Least Absolute Shrinkage and Selection Operator* (LASSO) as a penalty. LASSO is particularly beneficial in our case because it automates model selection through its shrinkage effect, as well as preventing overfitting. It effectively reduces the coefficients of less significant genes to zero, concentrating on relevant, non-zero predictors.

To further reduce dimensionality, we calculated the mean gene expression values and retained genes with a mean expression level exceeding the median, reducing the original set of 49,386 genes by half. These filtered gene expression data were then merged with DTI imaging data and the primary biomarkers.

### 3.2.1 Multi Gaussian

Another tool to help alleviate a vast number of responses is to collapse them into groups. The multi-response Gaussian family is particularly useful when there are several *(correlated)* responses, also known as the "multi-task learning" problem. In this case, a covariate is either included in the model for all the responses or excluded for all the responses.

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

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

*where $Y_{n\times m}$ is the matrix of response variables across n observations, $X_{n\times p}$ is the design matrix of predictors, $\beta_{p\times m}$ contains the regression coefficients linking predictors to responses, and $\varepsilon_{n\times m}$ represents the residual errors for each response.*

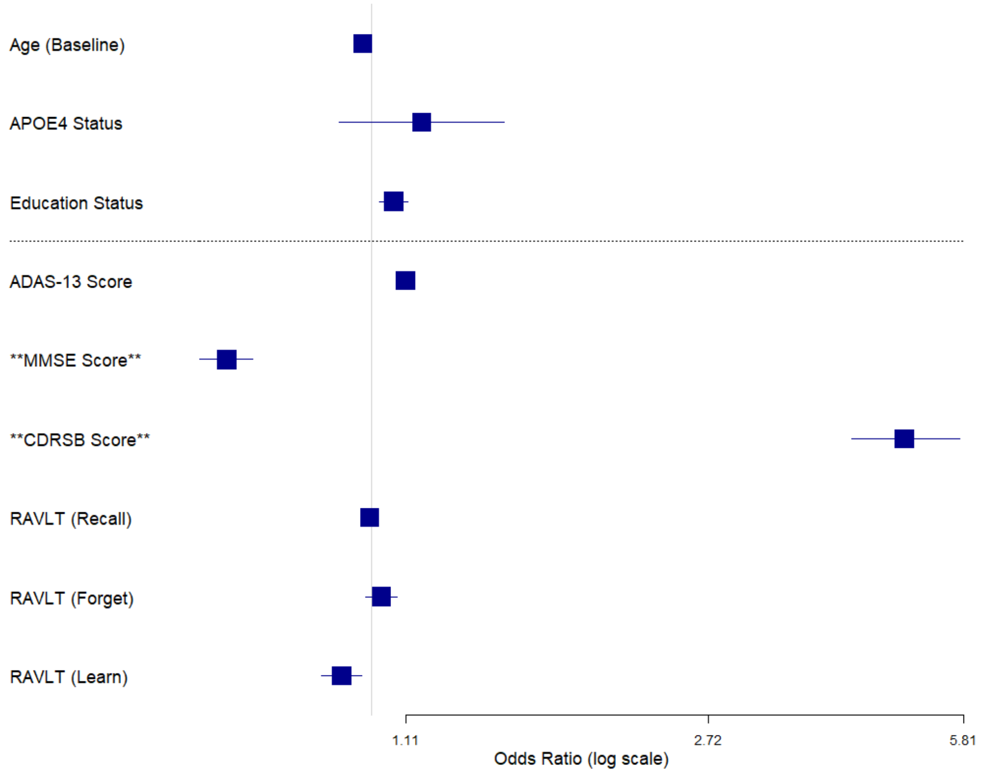
Unlike gaussian conditions with a single response, *Y* is not a vector but a matrix of quantitative responses. As a result, the coefficients at each value are also a matrix. This structure will be particularly helpful when addressing responses in brain regions.

# 4. Results

## 4.1 Low-Dimensional

We regress disease state using a combination of biomarkers as explanatory variables in our analyses. For our first model, we will threshold for the combined effect of risk factors and cognitive tests. Our results display the output from our ordinally regressed model compiling cognitive tests and risk factors. We will use the significance level of (α=0.05) to base the significance of our models on for comparison. Confidence intervals were constructed based on log-scale odds ratios

*Figure 2. Odds Ratio estimates for disease state on combined Risk factors and Cognitive Tests.*



Our ordinal regression analysis highlights significant associations between cognitive exams and disease progression, as visualized in *Figure 2*. Key predictors, including *MMSE* and *CDRSB*, show odds ratios greater than 1, indicating their contribution to higher disease states. *APOE4 genetic status* and *baseline age* also exhibit positive associations with disease progression, while *education status* and memory recall *(RAVLT)* tests show mixed effects.

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

## 4.1.1 Cognitive Testing 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. However, they have limitations; repeated exposure to questions or prior knowledge by the examiner can introduce bias, compromising test integrity.

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.

*Table 2: Multivariable Linear regression estimates for regressing the Clinical Dementia Rating scale Sum of Boxes (CDRSB) on the combined CSF measurements and Disease states. (N = 1,113)*

*Adjusted $R^2$ = 0.699, F-statistic: 370.6 on 7 and 1105 DF, Residual Standard Error: 0.9875 on 1105 DF*

| Biomarker | Estimate | Std. Error | T-value | P-Value |
|---|---|---|---|---|
| **Demographics** | | | | |
| *Baseline Age (years)* | $-4.919 * 10^{-4}$ | 0.004 | -0.119 | 0.905 |
| *APOE-4 Genetic Status* | $3.491 * 10^{-3}$ | 0.068 | 0.051 | 0.959 |
| **Disease State** | | | | |
| *Early Mild CI* | 2.909 | 0.072 | 40.512 | $< 2 * 10^{-16}$ |
| *Late Mild CI* | 0.764 | 0.061 | 12.601 | $< 2 * 10^{-16}$ |
| *Alzheimer's Disease* | 0.768 | 0.058 | 13.363 | $< 2 * 10^{-16}$ |
| **Cerebrospinal Fluid (CSF) Measurements** | | | | |
| *Phosphorylated Tau levels* | $5.108 * 10^{-3}$ | 0.002 | 2.265 | 0.024 |
| *Amyloid-Beta levels* | $-1.514 * 10^{-4}$ | $5.792 * 10^{-5}$ | -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 were around 98%, we only included PTAU.

Levels of *PTAU* emerged as significant, indicating that higher Ptau levels are associated with increased disease burden. In contrast, *ABETA* levels exhibited a significant inverse association with *CDRSB* scores, aligning with its established role in amyloid pathology and suggesting a potential mitigating effect on cognitive decline. Disease states served as essential contributors, with progression from *EMCI* to *LMCI* and Alzheimer's disease linked to significantly higher *CDRSB* scores, underscoring the pronounced cognitive decline across these stages.

Demographic variables, including *baseline age* and *APOE4* genetic status, were non-significant in this model. This suggests that while these factors may serve as background adjustors, they do not directly influence *CDRSB* scores in the presence of CSF biomarkers and disease states.

*Table 3: Multivariable Linear regression estimates for regressing the Mini Mental State (MMSE) the combined CSF measurements and Disease states. (N = 1,113)*

*Adjusted $R^2$ = 0.618, F-statistic: 257.8 on 7 and 1105 DF, Residual Standard Error: 1.632 on 1105 DF*

| **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 | $< 2 * 10^{-16}$ |
| *Late Mild CI* | -1.500 | 0.100 | -14.960 | $< 2 * 10^{-16}$ |
| *Alzheimer's Disease* | -0.594 | 0.095 | -6.245 | $6.04 * 10^{-10}$ |
| ***Cerebrospinal Fluid (CSF) Measurements*** | | | | |
| *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.

Disease states served as significant adjustors, showing progressive cognitive decline across stages. Baseline age was significantly associated with lower MMSE scores, reflecting a small effect of age on cognitive performance. APOE4 genetic status, however, was non-significant, underscoring its limited direct influence on MMSE in this model.

These findings highlight the critical role of CSF biomarkers, particularly Ptau and Amyloid-Beta, in understanding cognitive decline, and emphasize the increasing cognitive burden associated with progression through disease stages.

## 4.2 High-Dimensional

Recognizing the importance of *MMSE* and *CDRSB* as key markers in cognitive decline, we explore their associations from a genetic perspective. We implemented marginal-screening techniques to reduce the high dimensionality. To do so, we utilized the *SIS* method which screened out uninformative predictors (genes.) The SIS procedure ranks all predictors by using a marginal-utility measure between the response and each predictor and then retains the top variables for further investigation. With our subsample (n=468) we use the default of 10 partitions (nfolds=10.)

Lastly, we fit the linear model using a gaussian family, fitting our significant genes to our biomarkers selected from the ordinal regression analysis. Only genes which result as non-zero coefficients in this final model will be considered.

*Figure 3. Top 75 Genes identified from MMSE and CDRSB respectively using LASSO Regression*

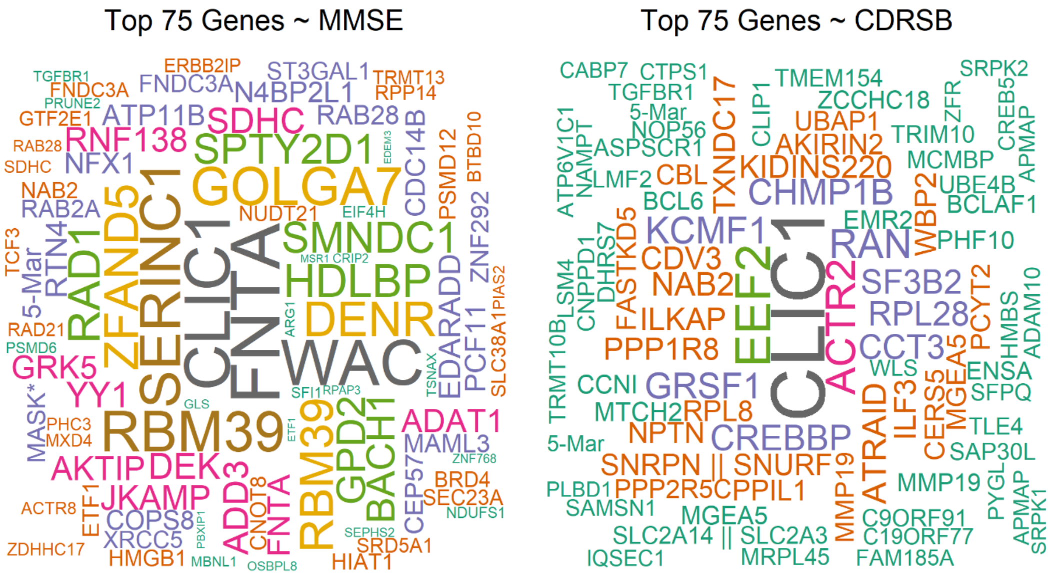
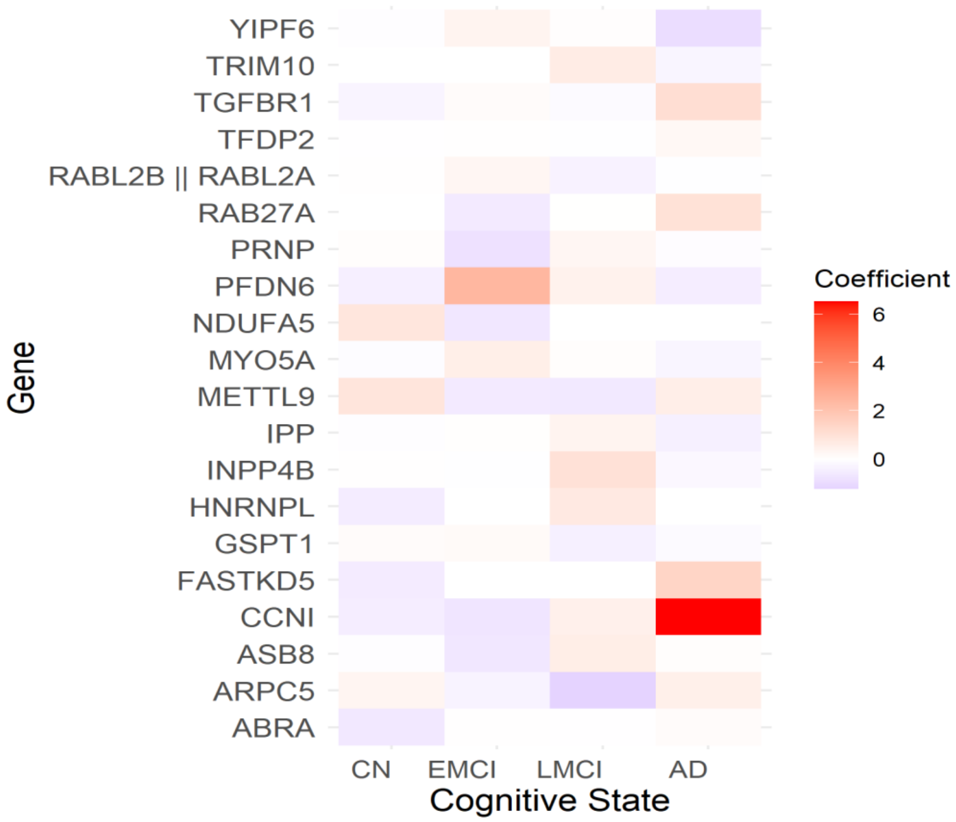


*Figure 3* represents the top 75 genes identified through this rigorous fitting process, with the size of each gene name reflecting its significance in relation to cognitive function. Modeled using MMSE and CDRSB as responses to capture cognitive decline across disease stages, four key genes-*NAB2, 5-MAR,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

### 4.2.1 Genetic + Disease State Analysis

We repeated the analysis to identify important genes across disease states. Since SIS does not support multinomial options, we processed each progressive state as independent binomial cases {e.g., CN-EMCI, CN-LMCI, CN-AD}. After screening, we used cross-validation with ungrouped multinomial options to capture varying genetic effects across response levels. This approach allows us to identify genes significant in some responses but not others, with the final output showing magnitudes and coefficients for each level. We then compiled a union and intersection set of non-zero coefficient genes from each category.

*Figure 4.*



The heatmap illustrates the association between the intersection set of genes and cognitive states, with the color gradient representing the strength of the magnitudes. Notably, *CCNI* shows a strong positive association with Alzheimer's Disease, while *FASTKD5* has moderate associations in early to late cognitive decline stages. In contrast, genes like *YIPF6* and *RAB27A* show negligible associations. These findings suggest that certain genes may serve as critical markers for cognitive decline, particularly in AD progression, warranting further validation and biological investigation.

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. That is, a gene which may show high effect at later stages of impairment may display a null effect at earlier stages. As such, finding intersections of the predictors reinforce our notion that a significant effect occurs in underlying mechanisms.

### 4.2.2 Imaging Genetic Analysis

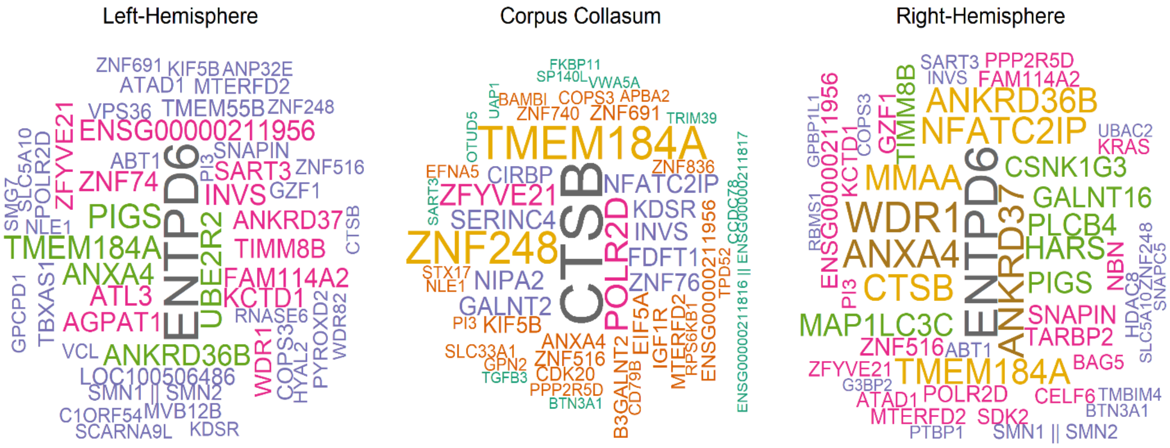
For our image analysis, we categorized the brain sections into three groups: the Corpus Callosum (CC), Right Brain, and Left Brain. The CC connects the cerebral hemispheres and is crucial for higher cognitive functions, often the first to show signs of neurodegeneration. The left hemisphere is responsible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

We have 57 FA (Fractional Anisotropy) image responses, with values ranging from 0 to 1. To handle this, we applied a logit transformation to the data. For fitting genetic predictors to these image responses, we applied SIS and the LASSO penalty. To account for heavy collinearity, the FA values were then split 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 5. Top 50 Genes identified from each Brain Region respectively using LASSO Regression.*



The figure 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, twelve predictors emerged as statistically significant in all three brain regions: TMEM184A, ZNF248, POLR2D, ZFYVE21, INVS, ANXA4, ZNF516, ENSG00000211956, MTERFD2, COPS3, PPP2R5D, and SART3.

Our results indicate that there exist more commonalities between the top genes from the two hemispheres versus the CC. For example, *"ENTPD6"* was found significant in measuring white matter in the hemispheres, but was not significant for the CC. In contrast, the gene "TMEM184A" was found significant in all 3 regions, to lesser effect. Genetic predictors found across all three brain sections may reveal stronger indications of white matter degradation and thus, act as influential predictors.

## 5. Concluding Remarks and Future Directions

This research explores the intricate associations between cognitive decline, genetic markers, and other biomarkers in Alzheimer's Disease progression. By employing low-dimensional and high-dimensional regression techniques, we sought to uncover the contributions of biological and genetic factors to cognitive impairment, assessed through Clinical Dementia Rating Sum of Boxes and Mini-Mental State Examination scores.

Our analyses demonstrated that CDRSB and MMSE scores are crucial in distinguishing disease states, with increasing CDRSB scores correlating with greater severity and MMSE improvements acting protectively. The importance of these biomarkers allowed us to utilize them from a genetic perspective.

Our genetic analysis utilized the Sure Independence Screening method to simplify data complexity, identifying significant markers that influence cognitive decline. The application of LASSO regression pinpointed impactful genes, including NAB2, 5-MAR, TGFBR1, and notably CLIC1, underscoring genetic contributions to neurodegeneration. Visual representations top predictors and emphasize the importance of integrating genetic data. Moreover, the differential effects of genes across disease stages suggest possible gene-environment interactions and stage-specific genetic susceptibilities, warranting further exploration into the molecular mechanisms of AD.

Our findings have several important implications for understanding the biological and genetic drivers of Alzheimer's Disease.

Interestingly, our finding of *CLIC1* aligns with growing evidence of its role in chronic Central Nervous System (CNS) inflammation. This suggests it could serve both as a marker for AD and a target for prevention. Similarly, *TGFBR1* is a confirmed receptor in the pathogenesis of Alzheimer's.

Future research should incorporate more advanced neuroimaging techniques to complement our current analyses, such as white matter integrity measures, to better capture neurochemical changes. Additionally, longitudinal studies could provide insights into the temporal evolution of cognitive and genetic markers, offering a more comprehensive understanding of disease progression. Exploring the interplay between genetic markers and environmental or lifestyle factors, such as physical activity and diet, may also shed light on modifiable risk factors for Alzheimer's Disease.

Our approach highlights the value of integrating statistical modeling, genetic analysis, and biomarker data to unravel the complexities of neurodegenerative diseases. While our models explain a significant portion of the variability in cognitive outcomes, they also underscore the multifactorial nature of Alzheimer's Disease, suggesting that a holistic approach is necessary to fully understand and design methods for preventative measures.

## 6. References