JUNE 13, 2024

EXPLORING PLASMA BIOMARKERS FOR ALZHEIMER'S DISEASE DETECTION

EVALUATING THE FEASIBILITY AND ACCURACY OF PREDICTING AMYLOID-β PLAQUES

B. CONTRERAS, H. GARCIA, C. PAJARILLO, L. SIMMONS & E. TOOBIAN
PORTLAND STATE UNIVERSITY – STAT 409
Professors: Dr. Ge Zhao & Dr. Bruno Jedynak
Instructors: Jacob Schultz & Adam MacBale

Table of Contents

1 Objective/Overview

1.1 Goals of the Project

This project focuses on Alzheimer's disease (AD), particularly examining the relationship between amyloid-β plaques and plasma biomarkers. Alzheimer's disease is characterized by the misfolding of amyloid-β and tau proteins, resulting in the formation of amyloid-β plaques and tau tangles. These aggregates disrupt neuronal function and impede synaptic communication, ultimately resulting in the degeneration of neural pathways.

Traditionally, amyloid-β plaques can be detected using Positron Emission Tomography (PET) imaging, specifically Amyloid PET. However, PET imaging is expensive and only available at research hospitals. Due to the high costs and limited accessibility of PET scans, this study turns to plasma biomarkers as an alternative in the prediction of these amyloid-β plaques.

The goals of this project are to assess the feasibility of predicting Amyloid PET using plasma biomarkers, to identify which biomarkers can most accurately predict the concentration of amyloid, and evaluate the influence of demographic factors on these predictions. To accomplish these goals, we will utilize patient medical test result data and demographic data to develop predictive models aiming to understand the relationship between the plasma biomarkers and Amyloid PET scores. We aim to determine if we can predict amyloid positivity, defined as Amyloid PET scores above an index of 1.17, and determine if we can ascertain the levels of amyloid using plasma biomarkers. We additionally aim to find which biomarkers are the most helpful for these predictions and the impact of age, gender, genetic factors, or other covariates on the accuracy of these predictions. Ultimately, the purpose of this study is to enable early detection of Alzheimer's Disease prior to the onset of clinical symptoms, thereby potentially enhancing intervention strategies.

1.2 Project Context

1.2.1 The Wisconsin Registry for Alzheimer's Prevention (WRAP) Study

The Wisconsin Registry for Alzheimer's Prevention (WRAP) is one of the world's largest and longest-running studies of individuals at risk for Alzheimer's disease. WRAP is a longitudinal study that follows a risk-enriched cohort from late midlife into old age. The study includes 1,729 participants (1,380 active) who enrolled in midlife (baseline mean age 54 years) and have been followed biannually for an average of 12 years. Data collected during hospital visits include cognitive assessments, lifestyle factors, medical history, structural and functional imaging, cerebrospinal fluid, and blood samples. The goals of the WRAP study are to detect the emergence of Alzheimer's disease (AD) proteinopathy and cognitive decline prior to overt clinical symptoms, gain a comprehensive picture of the effects of nonmodifiable genetics and modifiable health and lifestyle factors on cognitive and AD biomarker onsets and trajectories, and characterize the presence and impact of other diseases on cognitive decline, chiefly vascular disease (Jedynak 2024).

1.2.2 Collaborators and Sponsors of WRAP

- **Collaborators**
  - Sterling Johnson, Professor, University of Wisconsin and WRAP team
  - Dr. Bruno Jedynak, Professor, Fariborz Maseeh Dept. of Mathematics and Statistics, Portland State University
  - Adam Macbale, Manager of the Data-Science and Statistics Consulting Lab at PSU, de-identification of the data
- **Sponsors**
  - NIH: National Institute of Aging

1.2.3 Acknowledgements of Our Project

- **Project Advisors**
    - Adam Macbale, Lab Manager, Data Science and Statistics Consulting Lab, Portland State University, primary advisor
    - Jacob Schultz, Graduate Teaching Assistant, Fariborz Maseeh Dept. of Mathematics and Statistics, Portland State University, secondary advisor
    - Dr. Ge Zhao, Professor, Fariborz Maseeh Dept. of Mathematics and Statistics, Portland State University, additional guidance and support

1.3 Summary of Available Data

1.3.1 Pittsburgh Compound B (PIB) Distribution Volume Ratio (DVR) Automated Anatomical Labeling (AAL) Data

The PIB_DVR_AAL dataset includes 232 measurements from 480 patients across 1,010 PET scans. Pittsburgh Compound B (PIB) is a radiolabeled tracer specifically used in PET scans to visualize amyloid plaques. PIB binds directly to these plaques, aiding in their detection and analysis. The quantitative measure in this dataset, the Distribution Volume Ratio (DVR), represents the ratio of the tracer concentration in targeted brain regions relative to a reference region believed to be free from amyloid deposits (Klunk et a., 2004). Additionally, the dataset employs Automated Anatomical Labeling (AAL), which segments the brain into standardized regions based on a neurological atlas. This method ensures precise localization and consistent analysis of amyloid plaque distribution across different scanned individuals.

Additionally, the dataset includes each subject's ID number and age at time of scan. DVR values from these PET scans have been aggregated into a single weighted volumetric index through weighted sums of readings from each brain region. This index, along with the subject's ID number and age at time of the scan, is provided in a secondary dataset, "Amyloid_PET".

1.3.2 Amyloid Positron Emission Tomography (PET) Data

The Amyloid_PET dataset consists of a weighted index derived from the DVR values obtained from each subject's PET scans, as recorded in the PIB_DVR_AAL dataset. This index provides a consolidated view of amyloid deposition across various brain regions. Additionally, this dataset includes the subject's ID number and age at the time of the PET scan. It retains the same number of patients and tests as the initial PIB_DVR_AAL dataset, including 480 patients across 1,010 PET scans. This consistency ensures a one-to-one correspondence between the measurements and subjects across both datasets. Previous studies have established a threshold of 1.17 for this weighted index, which is used to determine amyloid positivity or negativity, indicating the presence or absence of significant amyloid deposition (Jack Jr et al., 2018).

1.3.3 Plasma Single Molecular Array (SIMOA) Gothenburg Data

The Plasma_Simoa_Gothenburg dataset consists of 1,253 observations from 422 subjects, utilizing Single Molecular Array (Simoa) technology for highly sensitive detection of plasma biomarkers. The primary biomarkers include Amyloid beta 40 (Ab40), Amyloid beta 42 (Ab42), Glial fibrillary acidic protein (GFAP), Neurofilament light-chain (NFL), and phosphorylated tau at position 181 (ptau181), position 231 (ptau231), and position 217 (pTau217). The 'Ab40', 'Ab42', 'GFAP', and 'NFL' biomarkers were measured using

Quanterix Simoa assays, while 'ptau181' was quantified using Quanterix V2 assay, 'ptau231' via an in-house Simoa assay, and 'pTau217' through the AlzPath method (Andreasson et al., 2021; Ashton et al., 2021).

The inclusion criteria for the dataset assumes that subjects must have undergone at least one amyloid PET scan or lumbar puncture. Each record also includes the subject's ID number, visit number, and age at the time of sample acquisition. The dataset specifies the Lower Limits of Quantification (LLOQ) for each biomarker as follows:

DETECTION LIMITS:

- Ab40: functional LLOQ = 4.08 pg/mL
- Ab42: functional LLOQ = 1.51 pg/mL
- GFAP: functional LLOQ = 11.6 pg/mL
- NfL: functional LLOQ = 1.6 pg/mL
- pTau181: functional LLOQ = 0.338 pg/mL
- pTau231: LLOQ = 2 pg/mL
- pTau217: Not yet available (README file, 2023)

To better understand the progression of Alzheimer's disease biomarkers over time we can consider the data longitudinally. The visualizations of the plasma biomarker data across multiple visits shown in Figure 1 demonstrates how, for most, biomarker levels evolve gradually over time.

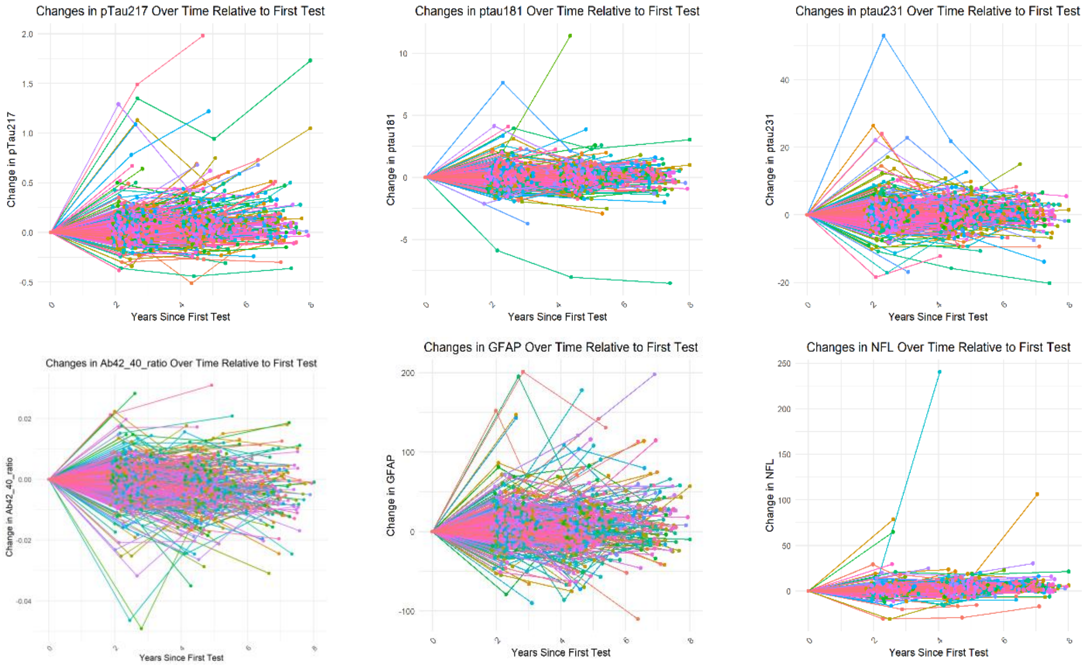

Figure 1: Plasma Biomarkers Over Time

1.3.4 Participants Data

The Participants dataset contains demographic and genetic information for 774 subjects. This includes demographic data such as ‘sex’, ‘race_primary’, and a variable ‘hispanic’. Additionally, the dataset contains genetic data represented by the variables ‘apoe_e1’ and ‘apoe_e2’, which specify the APOE genotype, a genetic marker used in assessing the inherited risk of developing Alzheimer’s disease. We will exclude columns 'race_secondary', 'race_tertiary', and 'hisp_orig' as they contain only N/A data, offering no valuable information. Upon consultation, it was advised to exclude ‘ed_years’ from our analysis as ‘ed_years’ was not deemed a critical factor in our study. Additionally, a sizable portion of ‘ed_years’ is composed of N/A data which would distort the analysis. Thus, the exclusion avoids significant data loss and maintains the integrity of the dataset, allowing for a more comprehensive and robust analysis.

1.3.5 Cognitive Status Data

The Cognitive Status dataset contains information on the cognitive status of 1758 subjects over a total of 7906 visits. The information in this dataset includes the subject’s ID number, visit number, and their cognitive status at the time of that visit. The cognitive status is described as one of six possible labels:

- **No Diagnosis Calculated**: Indicates that no cognitive status diagnosis was made during the visit.
- **Cognitive Unimpaired and Stable:** The subject shows no signs of cognitive impairment and their cognitive function is stable at the time of the visit.
- **Cognitive Unimpaired and Declining:** The subject shows no sign of cognitive impairment currently, but there are indications of a decline in cognitive function.
- **Impaired Not MCI:** The subject shows cognitive impairment, but it does not meet the criteria for Clinical Mild Cognitive Impairment (MCI).
- **Clinical MCI:** The subject meets the criteria for mild cognitive impairment.
- **Dementia:** The subject has significant cognitive decline that meets the criteria for dementia.

These labels represent the progression of cognitive decline and its relationship with Alzheimer’s disease. Early cognitive decline, such as in “Cognitive Unimpaired and Declining” could be an early sign of developing cognitive impairment or Alzheimer’s disease. However, even at the stage of “Impaired Not MCI” in this progression, this impairment could be due to other factors or conditions. “Clinical MCI is a condition often considered a precursor to Alzheimer’s disease. Individuals with MCI exhibit noticeable cognitive decline, but not severe enough to interfere significantly with daily life. Finally, subjects who meet the criteria of “Dementia” have significant cognitive decline that interferes with daily functioning. Alzheimer’s disease is categorized as dementia (Jack et al., 2018).

2 Data Analysis & Preprocessing

2.1 Initial Data Analysis:

In the initial stage of data analysis, several basic checks were conducted to ensure the quality and consistency of the datasets. This included examining each dataset for missing values and duplicate records, which is crucial for identifying potential data quality issues that could impact the analysis. Missing values were identified in several columns, and appropriate handling strategies for these gaps were planned.

Additionally, tables summarizing the qualitative variables were created. These tables provided insight into the meanings and distributions of the categorical levels, which informed subsequent data cleaning and preprocessing steps.

These initial analyses established a solid foundation for the data cleaning and preprocessing steps that followed, ensuring that the datasets were adequately prepared for further exploration and modeling.

2.2 Data Cleaning:

Data cleaning involved several essential steps to prepare the datasets for analysis. This process ensured that the data was accurate, consistent, and ready for subsequent modeling and analysis.

2.2.1 Handling Missing and Duplicate Data

Initial check revealed several columns with missing values. Columns with all N/A values or those deemed irrelevant, such as index variables or specific metrics from previous studies that did not provide useful information in our context, were removed. For columns with partial missing data, rows were removed as the imputation was not appropriate for this data. Duplicate rows were identified and removed to maintain the integrity of the data.

2.2.2 Addition of New Variables

At this stage, several new variables were introduced:

- *amyloid_binary:* A binary variable indicating amyloid positivity based on the established threshold of 1.17. This variable helps in distinguishing between subjects with and without significant amyloid deposition and is crucial for the classification model later used in this analysis.
- *apoe_ternary* and *apoe_binary:* These variables were created utilizing the existing variables of 'apoe_e1' and 'apoe_e2', which represent a genetic marker from each biological parent. APOE genotype data is used to assess the risk of developing Alzheimer's disease. They are categorized in the original variables as integer values from "2" to "4", with "4" indicating high risk and thus two "4"s even higher risk. The engineered variable 'apoe_binary' classifies each subject into one of two categories indicating whether they have a level "4" in either APOE variable. The engineered variable 'apoe_ternary' classifies each subject into one of three categories indicating a count of instances of "4" between both of these APOE variables.
- *Ab_ratio:* This variable represents the ratio of Amyloid-β 42 to Amyloid-β 40. The use of this ratio is supported by research indicating that the Aβ42/40 ratio is a more reliable biomarker for amyloid deposition compared to Aβ42 or Aβ40 alone. According to Amft et al. (2022), "the cerebrospinal fluid biomarker ratio Aβ42/40 identifies amyloid positron emission tomography positivity better than Aβ42 alone in a heterogeneous memory clinic cohort," making it a crucial predictor for our analysis.

2.2.3 Factorization and Numerization:

Categorical variables such as 'sex', 'race_primary', 'hispanic', and 'Calculated_Consensus_dx' were converted into factor variables to facilitate analysis. This transformation ensures that these variables are appropriately handled in statistic models. Continuous variables, like 'ptau181' were converted to numeric format where necessary to enable mathematical operations and statistical analysis.

2.2.4 Exclusion of Columns:

Upon consultation, the column 'ed_years' was excluded from the analysis. It was deemed not critical for the study, and a significant portion of the data in this column consisted of N/A values, which could distort the analysis. Thus, the exclusion of this column helped maintain the integrity of the dataset.

These steps in the data cleaning were critical in preparing the datasets for further exploration and modeling. Details of this process are shown in the "Data Cleaning & Integration" flowchart in Figure 2.

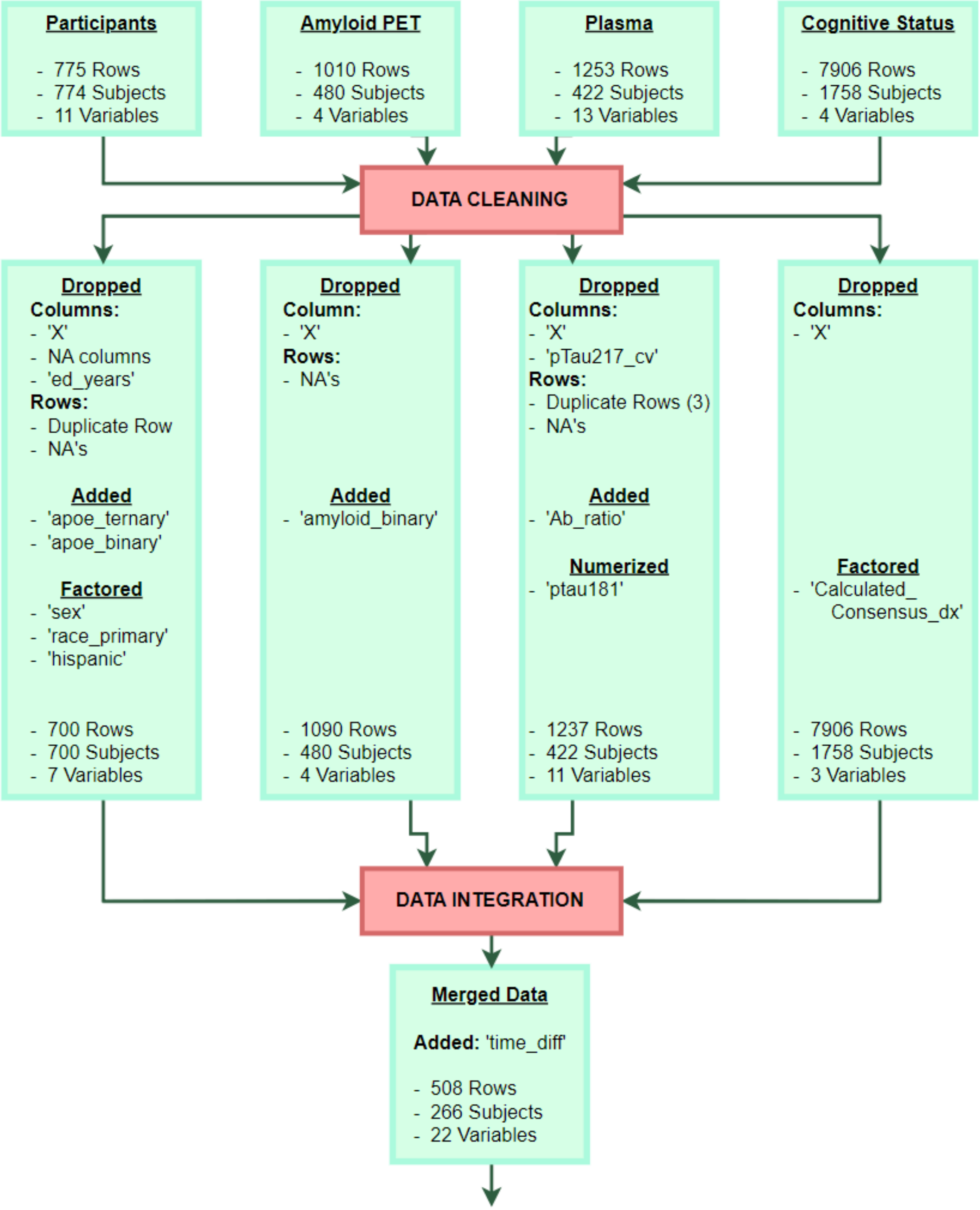

Figure 2: Data Cleaning and Integration Flowchart

2.3 Data Integration:

Data integration was a critical step to endure a comprehensive and cohesive analysis. This process involved combining various datasets into a single, unified dataset, which was essential for our analysis.

2.3.1 Merging Datasets

The merging process involved aligning datasets based on common identifiers such as subject ID and visit number. This alignment ensured that data from different sources corresponded accurately to each participant and their respective visits. The primary datasets included information on amyloid PET scans, plasma biomarkers, cognitive status, and participant genetic information and demographics.

2.3.2 Consistency Checks

To maintain consistency across the merged dataset, several checks were performed:

- Ensuring no data was lost or duplicated during the merging process.
- Verifying key variables, such as Subject ID and visit number, were consistent across all datasets.
- Checking for any discrepancies or mismatches in the data entries.

2.3.3 Addressing Timing Discrepancies

A significant challenge encountered during data integration was the timing discrepancy between amyloid PET tests and plasma tests. These tests were often conducted at different times, complicating the analysis. Considerable time was spent exploring various options before settling on a matching method based on the closest absolute time difference.

For each subject who had an amyloid PET test, the closest plasma test was identified by matching based on the smallest absolute value of the time difference, captured in an engineered variable called 'time_diff'. This variable quantifies the difference between the times of the two tests. Although the matching was done based on the smallest absolute time difference, the 'time_diff' variable retains its positive or negative value to indicate the sequence of the tests.

2.3.4 Cross-Sectional Study Constraints

Although the data is from a longitudinal study, a cross-sectional analysis was requested. A time window of 4 years was chosen, meaning that any plasma test conducted more than 2 years before or after an amyloid PET test (i.e., 'time_diff' > 2) would not be considered. This constraint was critical in ensuring the relevance of the matched tests and maintaining the focus of the analysis.

These steps ensured that the datasets were accurately combined, providing a reliable foundation for further analysis and modeling. For more detailed information on the variables maintained in the merged dataset, please refer to *Appendix B: Merged Dataset Variables*.

3 Exploratory Data Analysis

The primary objective of the Exploratory Data Analysis (EDA) is to provide a comprehensive understanding of the dataset, identify key patterns, spot anomalies, and inform subsequent modeling decisions. This involves examining the frequency distributions of categorical variables, understanding the distributions of quantitative variables, and investigating correlations between variables.

3.1 Overview of Qualitative Variables

The objective of this section is to examine the distribution of categorical variables and identify necessary transformations or groupings. This process is crucial for addressing data sparsity and ensuring robust model performance. We begin by examining the counts and frequencies of these variables. A comprehensive table showing these findings can be found in *Appendix C: Qualitative Variable Analysis*.

3.1.1 Amyloid_binary and Sex

The 'amyloid_binary' variable indicates the presence (1) or absence (0) of significant amyloid deposition, while the 'sex' variable categorizes subjects as either male or female. The distribution analysis shows that the majority of subjects are amyloid negative, with a higher number of female participants compared to males. These distributions are illustrated in the histograms in Figure 3:

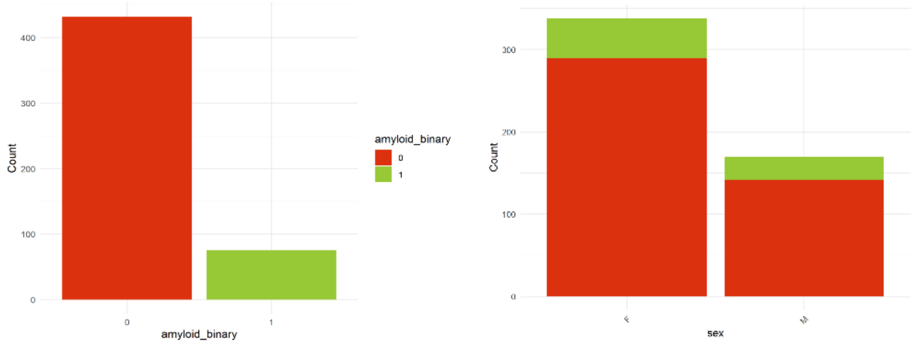

Figure 3: Distributions of 'amyloid_binary' and 'sex'

3.1.2 Race_primary and Hispanic

The initial categories for 'race_primary' included "American Indian or Alaska Native", "Asian", "Black or African American", "White", and "Other". The 'hispanic' variable had only 3 instances of "Yes," which matched the 3 instances in 'race_primary' labeled as "Other." This lack of additional information and potential for singularity issues when 'race_primary' was one-hot encoded led to the decision to drop the 'hispanic variable'. Additionally, the 'race_primary' variable was also dropped as the vast majority of subjects (95.28%) were "White", providing limited variability for analysis. The distributions for these variables can be found in Figure 4.

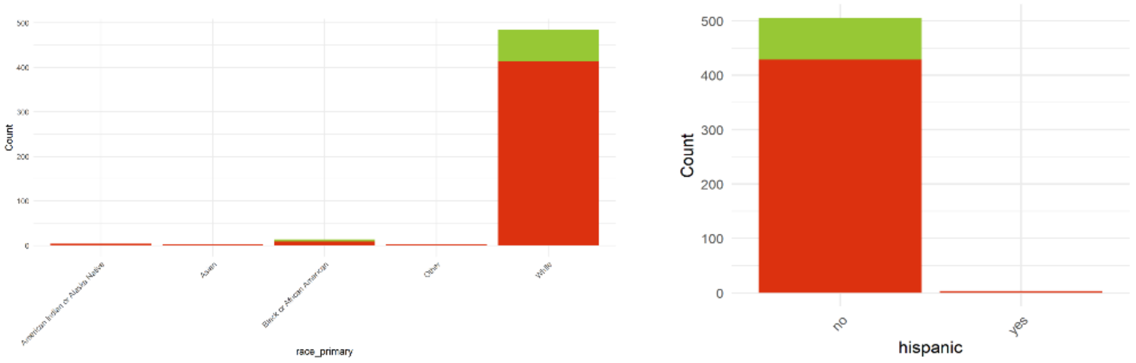

Figure 4: Distributions of 'race_primary' and 'hispanic'

3.1.3 Calculated_Consensus_dx

The 'Calculated_Consensus_dx' variable initially includes categories of "Cog_Unimpaired_Stable", "Cog_Unimpaired_Declining", "Impaired_Not_MCI", "Clinical_MCI", and "Dementia". These levels were very unevenly distributed, with most falling into "Cog_Unimpaired_Stable". To achieve slightly better distribution while maintaining meaning within this variable, the levels were merged into three levels representing a lack of impairment, clinical impairment, and an intermediate stage. "Cog_Unimpaired_Stable" remained the same. "Cog_Unimpaired_Declining" and "Impaired_Not_MCI" were merged to a single level of "Unimpaired_Declining_or_Impaired". "Clinical_MCI" and "Dementia" were merged into a single category of "MCI_or_Dementia". Although this process slightly improved the distribution, there was still extremely uneven distribution, which would need to be addressed in our modeling methods. The distributions of this variable before and after merging levels are shown in Figure 5.

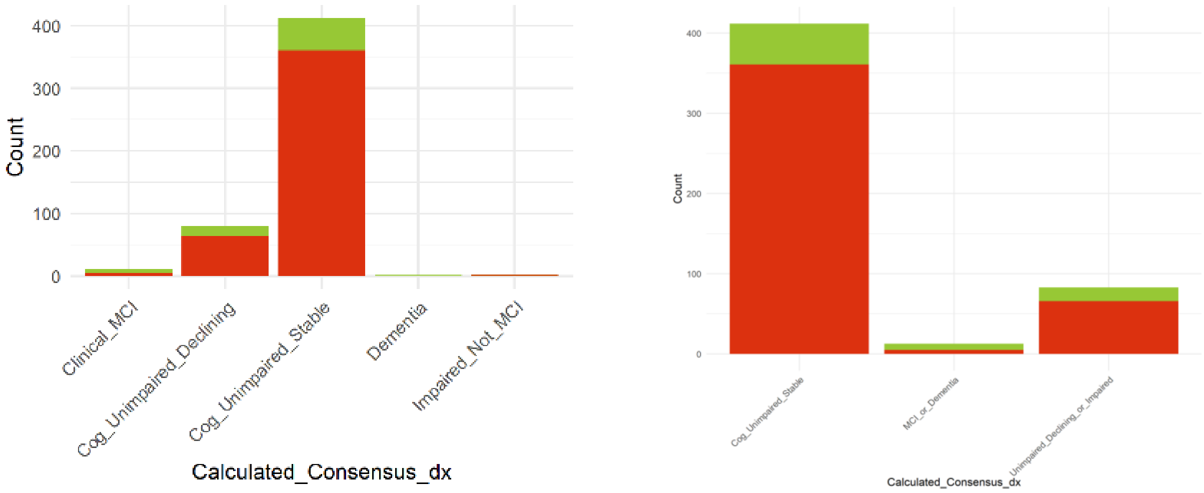

Figure 5: 'Calculated_Consensus_dx' Before & After Level Merge

3.1.4 Apoe_binary and Apoe_ternary

Both ‘apoe_binary’ and ‘apoe_ternary’ were analyzed to provide a detailed view of the genetic risk factors for Alzheimer's disease. The ‘apoe_binary’ variable indicates the presence of the APOE4 allele, a known genetic risk factor, while the ‘apoe_ternary’ variable provides a more nuanced view by indicating the count of APOE4 alleles. This dual analysis allows for a deeper understanding of genetic influences and potential stratification of risk. Although ‘apoe_ternary’ is unevenly distributed among the levels, ‘apoe_binary’ is relatively well balanced. The distributions of each are shown in Figure 6.

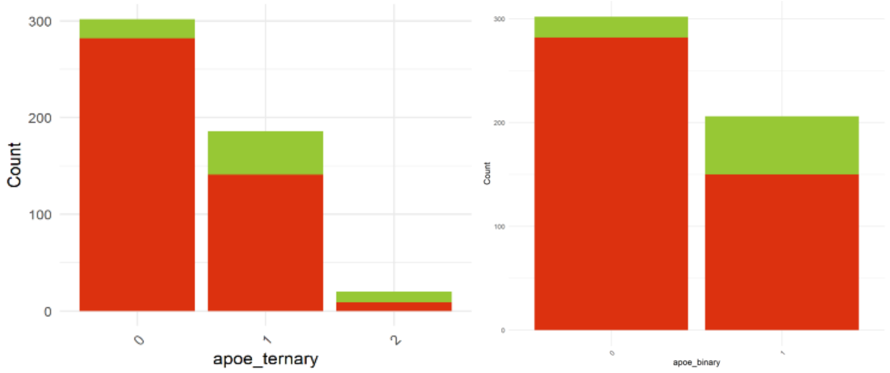

Figure 6: Distributions of 'apoe_binary' and 'apoe_ternary'

3.1.5 Results Table

The effect of the merging of levels and the removal of variables no longer being considered are reflected in Table 1 below.

VARIABLE	CATEGORIES	COUNTS			RELATIVE FREQUENCIES (%)		
		Amyloid Negative	Amyloid Positive	TOTAL Count	Amyloid Negative	Amyloid Positive	% of TOTAL
amyloid binary	Negative (0)	432	---	432	100.0	---	85.04
	Positive (1)	---	76	76	---	100.0	14.96
sex	Female	290	48	338	67.13	63.16	66.55
	Male	142	28	170	32.87	36.84	33.46
Calculated_Consensus_dx	Cog_Unimpaired_Stable	361	51	421	83.56	67.11	81.10
	Unimpaired_Declining_or_Impaired	66	17	83	15.28	22.37	16.33
	MCI_or_Dementia	5	8	13	1.16	10.53	2.56
apoe_ternary	0	282	20	302	65.28	26.32	59.45
	1	141	45	186	32.63	59.21	36.61
	2	9	11	20	2.08	14.47	3.94
apoe_binary	0	282	20	302	65.28	26.32	59.45
	1	150	56	206	34.71	73.68	40.55

Table 1: Results of Merging Levels and Removal of Considered Variables in Qualitative Data

3.2 Overview of Quantitative Variables

The objective of this section is to summarize and visualize the distributions of quantitative variables, identifying any necessary transformations. This is crucial for understanding the dataset's characteristics and preparing it for further analysis.

3.2.1 Distributions

Histograms and boxplots were used to visualize the distribution of each variable, as shown below in Figure 7. These visualizations help understand the spread and central tendencies of the data, allowing for the identification of potential outliers and non-normal distributions. A thorough analysis revealed that many variables were skewed, particularly the plasma biomarker data. Boxplots compared the distribution of continuous variables against the amyloid binary values, which classify amyloid PET results as positive or negative based on the threshold of 1.17. By dividing the data into distinct groups, the boxplots visually compare the distributions between amyloid positive and negative groups, enabling us to identify differences. The uneven distribution in some of the plasma biomarker data will be addressed further in our linear regression analysis.

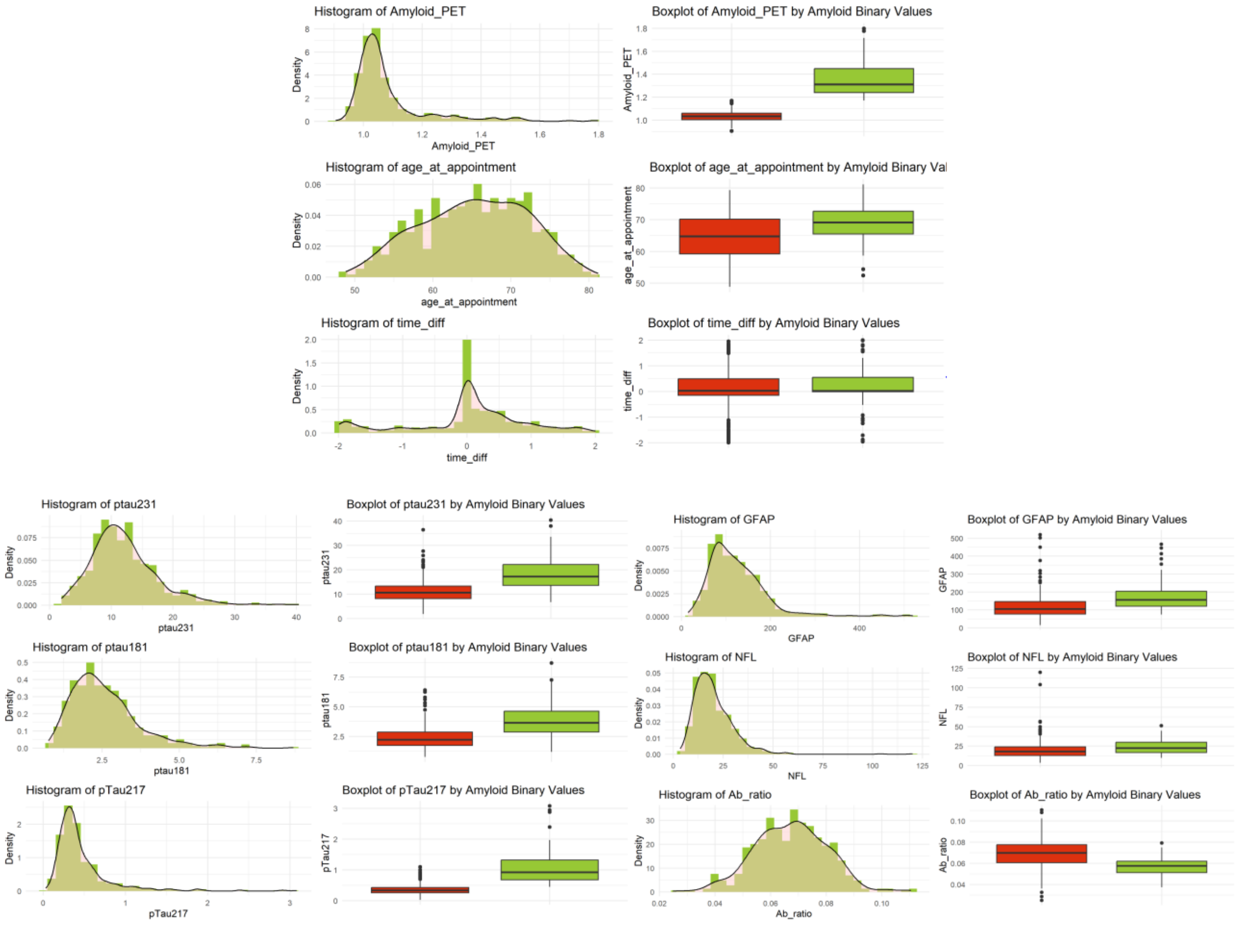

Figure 7: Distributions and Boxplots of Quantitative Variables

3.2.2 Standardizations

To address the varying scales of the plasma biomarkers, z-score normalization was applied using the scale() function. This standardization ensures that each variable contributes equally to the analysis, improving model performance and interpretability. After this standardization, summary statistics were generated for the primary continuous variables, providing an overview of their central tendencies and dispersion. Table 2 below summarizes the mean, median, standard deviation, minimum, and maximum values for each variable.

VARIABLE	MEAN	MEDIAN	STD. DEV.	MIN	MAX
Age at Appointment	65.28	65.59	8.44	48.83	81.16
Amyloid_PET	1.084	1.044	0.17	0.905	1.799
Ab40	112.93	110.00	24.91	27.70	214.00
Ab42	7.517	7.430	2.06	0.690	13.60
GFAP	0.00	-0.2077	1.44	-1.5641	5.4976
NFL	0.00	-0.1670	1.42	-1.5611	9.2957
ptau181	0.00	-0.2089	1.49	-1.6199	5.3606
ptau231	0.00	-0.1560	1.07	-1.8402	5.0579
pTau217	0.00	-0.2712	1.14	-1.1693	6.8871
Ab_ratio	0.00	0.02681	1.12	-3.25038	3.30770
Time Difference	0.02898	0.02000	0.57	-1.99000	2.00000

Table 2: Summarization of Qualitative Variables

3.2.3 Correlation

A correlation matrix, shown in the figure below, was constructed in order to analyze the relationship between our continuous variables. This matrix allowed us to identify several key insights. A strong positive correlation is shown between 'Amyloid_PET' and 'pTau217'. We also identified a small negative correlation between 'Amyloid_PET' and 'Ab_ratio'. We also identified early signs of multicollinearity as shown by the strong positive correlation between 'ptau181' and 'ptau231'. We will take this into consideration during our model selection process.

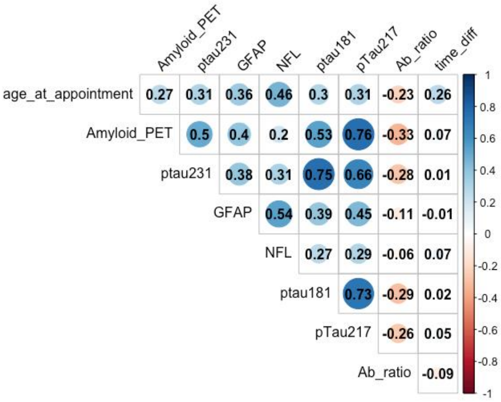

Figure 8: Correlation Matrix of Continuous Variables

4 Fitting Models

4.1 Initial Method

We initially explored using a mixed effects model to account for both fixed and random effects in our data. Mixed effects models (as described in *Appendix A: Summary of Statistical Terms*) are particularly useful in handling data with grouped structures, such as repeated measurements from the same patient. They allow us to model the individual variability by including random effects, which can capture the differences between patients while considering the fixed effects that apply to the entire population. However, in our dataset, the majority of patients had only 1 or tests, which is shown in Figure 9. This limited number of repeated measures per patient poses a challenge for the mixed effects model, which relies on sufficient within-group data to accurately estimate random effects. Without enough repeated measures, this model cannot adequately differentiate between the within-patient variability and the overall population trends leading to unreliable predictions.

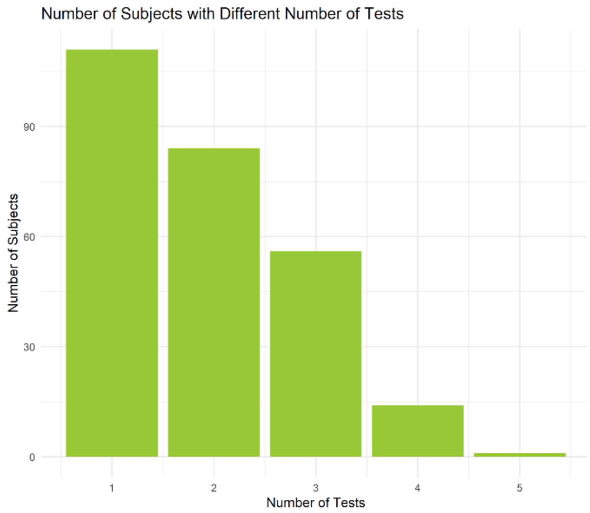

Figure 9: Number of Tests per Subjects

4.2 Key Methods for Model Evaluation and Selection

The categorical variables in our dataset exhibited uneven distributions, which posed a challenge for cross-validation methods that might not represent minority classes in each fold. To address this, we utilized Leave-One-Out-Cross-Validation (LOOCV), which ensures that every data point, including those from underrepresented categories, is used for both training and testing. This approach provides a more robust and unbiased evaluation of our model's performance across the entire dataset, ensuring that the variability within each category is accurately captured.

To refine our predictive models, we employed evaluation metrics such as Log Loss and PRESS. Where each evaluation metric is tailored to the nature of each model type. Log Loss, specifically employed for logistic regression, quantifies how close the probabilistic predictions are to the actual values, weighted by these predictive values, and penalizes deviations from the true outcomes. The more that the predicted probabilities deviate from the actual value, the higher the Log Loss. Lower Log Loss values signify better alignment between predicted and actual outcomes. Initially, AIC was employed to perform this task, however, while AIC is useful for feature selection, it's not specifically designed for evaluating the performance of classification models. Rather, it balances fit and complexity but does not provide a direct measure of prediction accuracy for classification.

Meanwhile, PRESS, utilized for linear regression, assesses predictive power by evaluating model performance on new data points by utilizing LOOCV, ensuring robust generalization beyond the training dataset. The sum of squared prediction errors for each left-out point constitutes the PRESS statistic. This statistic is employed to identify the features in the linear regression model that provide the strongest predictive power, with lower PRESS values indicating better model performance.

4.3 Methodology - Both Models

To address whether Amyloid can be predicted using a threshold of 1.17, and to determine if the level of Amyloid-β plaques can be predicted, we employed logistic and linear regression respectively. Our methodology also explores which subset of biomarkers is most effective for these predictions, and whether incorporating covariates such as demographic or genetic factors enhances the predictive accuracy.

In our analysis, we developed two predictive models: the "reduced" model and the "full" model. During model selection of the reduced model, we removed the variables 'Calculated_consensus_dx', 'apoe_ternary', and 'apoe_binary' and then compared the models against one another using all combinations of the remaining predictor variables. These variables represent information that is expensive to collect and often not available in a real-world scenario. By excluding these variables, this model represents more cost-effective and easier to gather information. Conversely, we included these variables during the full model selection. The choice between these models depends on balancing cost constraints and the desired level of predictive precision.

1) Find the most effective combination of plasma predictors and covariates for optimal prediction power (using either PRESS for lin. reg.; or for log reg. using Log Loss)
   a) Test all combinations of plasma biomarkers and covariates for both 'apoe_ternary' & 'apoe_binary'.
   b) Model is chosen based on the lowest Log Loss/PRESS Value.
   c) Full model will include all predictor variable options while reduced model will include combination of predictors with lowest PRESS or LogLoss, excluding any "expensive" variables.
   d) For linear regression model, any transformations necessary will be performed at this point.
   e) Run LOOCV on the models containing the strongest predictive set of predictors for both the full and reduced models.
      i) LINEAR REGRESSION (Per Results from PRESS):
         - **FULL: Amyloid_PET ~ ptau181 + pTau217 + GFAP + NFL + Ab_ratio + sex + Calculated_Consensus_dx + apoe_ternary**
         - **REDUCED: Amyloid_PET ~ ptau217 + GFAP + NFL + Ab_ratio + sex**
      ii) LOGISTIC REGRESSION (Per Results from LogLoss):
         - **FULL: amyloid_binary ~ ptau231 + ptau181 + pTau217 + GFAP + NFL + Ab_ratio + age_at_appointment + apoe_ternary**
         - **REDUCED: amyloid_binary ~ ptau181 + pTau217 + GFAP + NFL + Ab_ratio**

   f) Analyze and aggregate the results to compare the performance of both regression models.
   g) Additionally, evaluate the models' effectiveness in distinguishing between the full versus reduced models and in predicting outcomes above versus below the threshold.

The flowchart in Figure 10 shows the general steps involved in our methodology.

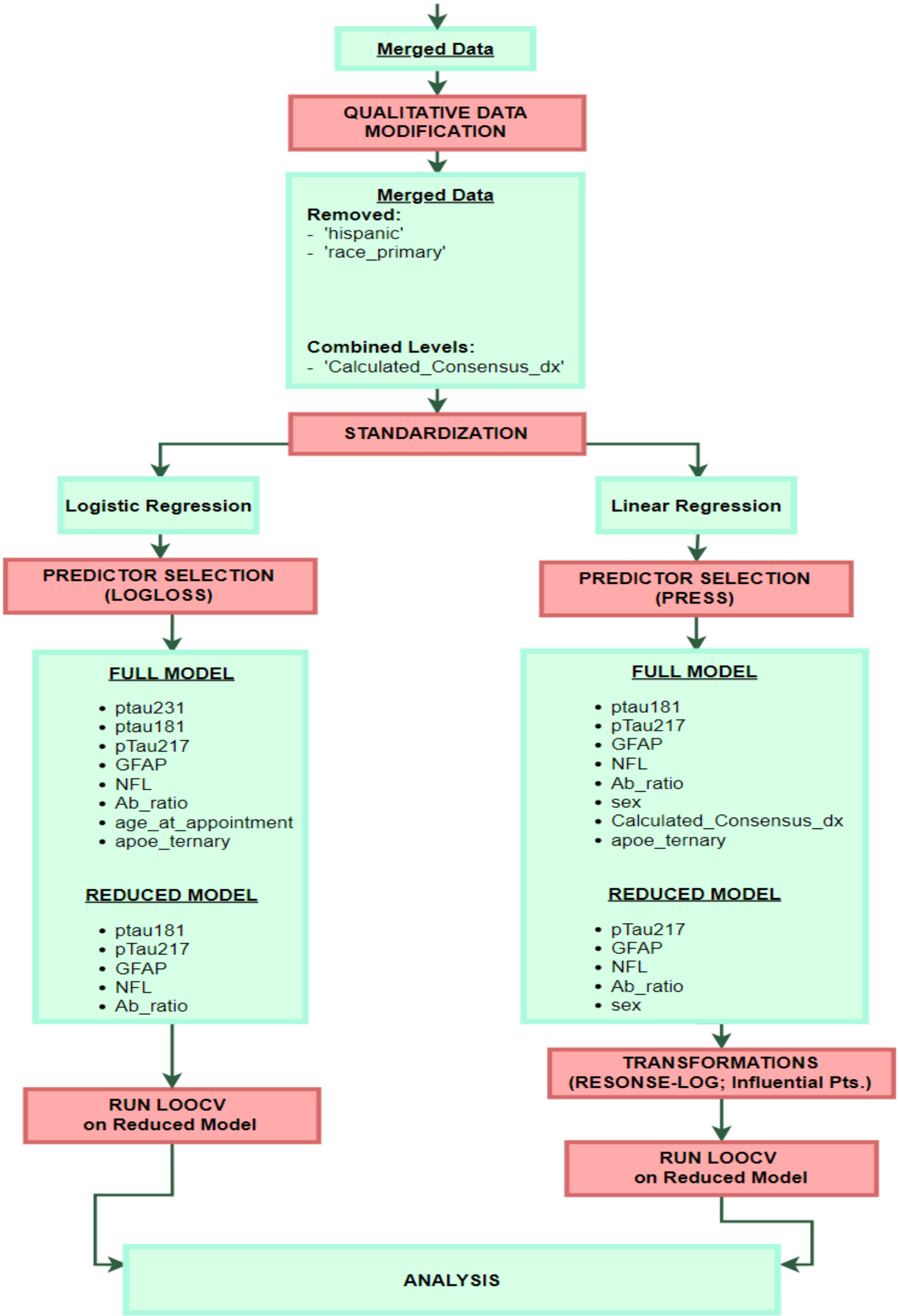

Figure 10: METHODOLOGY

4.4 Logistic Regression

In our process of addressing the question whether amyloid positivity can be predicted based on a threshold value of 1.17. We chose to employ logistic regression to address this inquiry. By modeling the relationship between the binary outcome variable (amyloid positivity) and one or more independent variables, such as biomarker levels and demographic factors, logistic regression can estimate the probability of amyloid positivity given certain predictor values. The analysis aims to evaluate the predictive capability of the chosen threshold of 1.17 for distinguishing amyloid-positive cases from negative ones, providing valuable insights into the effectiveness of this threshold in identifying individuals with amyloid positivity.

4.4.1 Validation and Feature Selection Method

To assess our model's performance in predicting Amyloid Positivity, we evaluated all combinations of predictors. Utilizing Leave One Out Cross Validation (LOOCV) at each step, we computed the Log Loss for every predictor combination. Ultimately, we generated a list of Log Loss values associated with distinct predictor sets, prioritizing those with the lowest Log Loss. This approach identifies both the full reduced model employed in our logistic regression analysis, emphasizing the significance of Log Loss as a key evaluation metric.

4.4.2 Performance Analysis:

Analysis of the logistic regression compares the performance of the two predictive models, labeled as the "full" and "reduced" models, in predicting amyloid positivity using the specified threshold of 1.17. The evaluation includes metrics such as accuracy, sensitivity (true positive rate), specificity (true negative rate), and the Area Under the Curve (AUC). The section summarizes the effectiveness of both models in correctly classifying amyloid-positive and amyloid-negative cases, providing insights into their predictive capabilities. For a detailed summary table of both models reference *Appendix D: Logistic Regression Model Summaries*.

<u>Full Model</u>:

	Predicted	
Actual	0	1
0	420	19
1	12	57

Table 3: Full Model Aggregated Confusion Matrix

- Log loss: 0.157
- Accuracy: 0.939
- Sensitivity (TPR): 0.75
- Specificity (TNR): 0.972
- AUC: 0.969

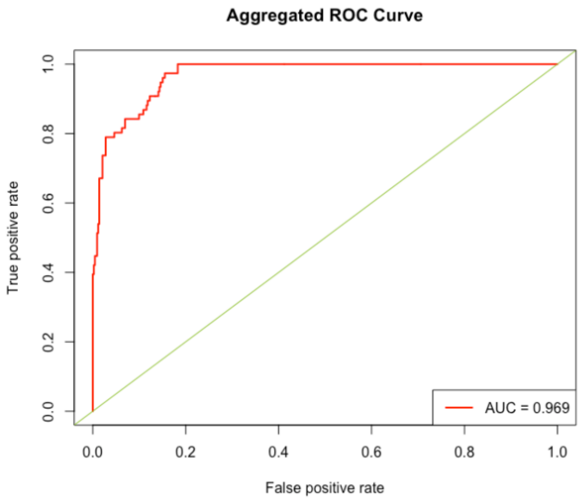

Figure 11: Aggregated ROC Curve - Full Logistic Model

<u>Reduced Model</u>:

	Predicted	
Actual	0	1
0	421	21
1	11	55

Table 4: Full Model Aggregated Confusion Matrix

- Log loss: 0.169
- Accuracy: 0.937
- Sensitivity (TPR): 0.724
- Specificity (TNR): 0.975
- AUC: 0.971

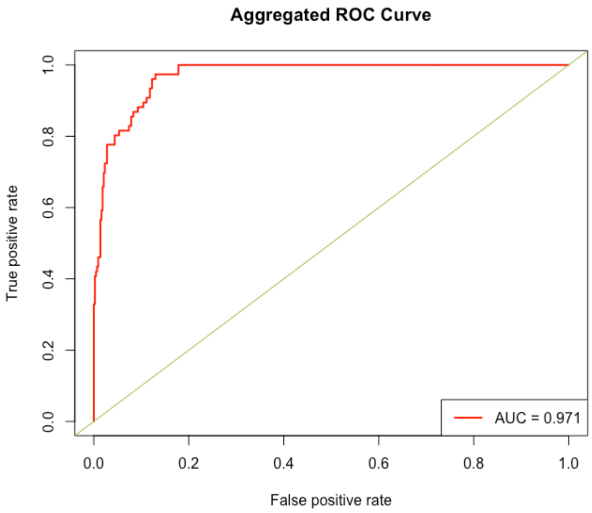

Figure 12: Aggregated ROC Curve - Reduced Logistic Model

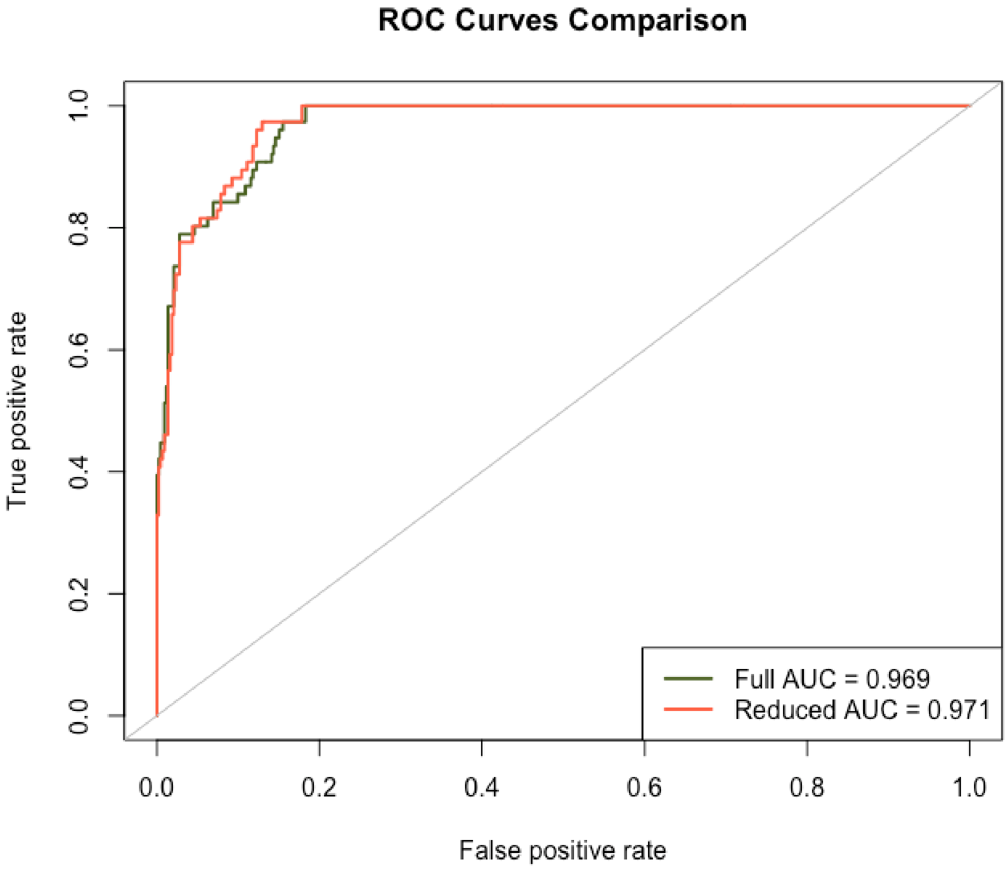

Figure 13: Comparison of Aggregated ROC Curves for Both Models

To evaluate the discriminative ability of the models, we generated aggregated Receiver Operating Characteristic (ROC) curves and computed the Area Under the Curve (AUC) values for each model, as shown in Figure 13. Definitions for AUC and ROC can be found in *Appendix A: Summary of Statistical Terms*. The full model achieved an AUC of 0.969, while the reduced model achieved an AUC of 0.971. These values indicate excellent performance in distinguishing between amyloid positive and negative cases.

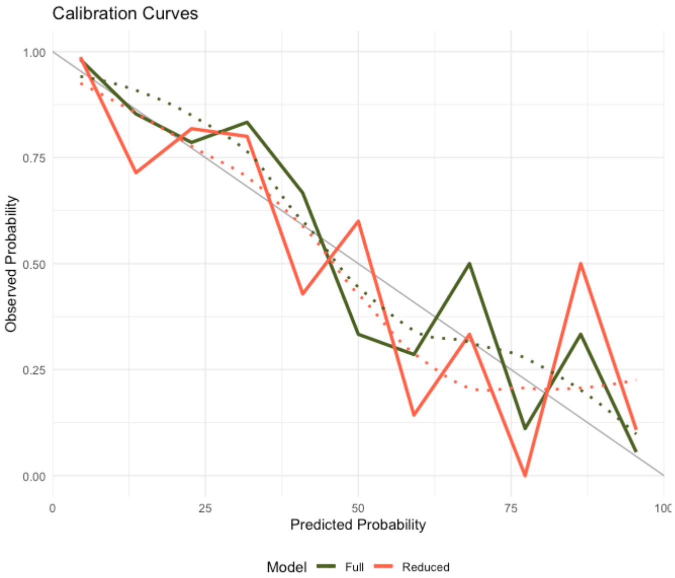

Figure 14: Calibration Curve of Both Models

Figure 14 illustrates the calibration curves for the full and reduced predictive models. The curves compare the predicted probabilities against the observed probabilities for amyloid positivity. The calibration curve assesses how closely the predicted probabilities align with actual outcomes, indicating the reliability and accuracy of each model's probability estimates. A perfectly calibrated model would have points lying on the diagonal line, demonstrating a one-to-one correspondence between predicted and observed probabilities.

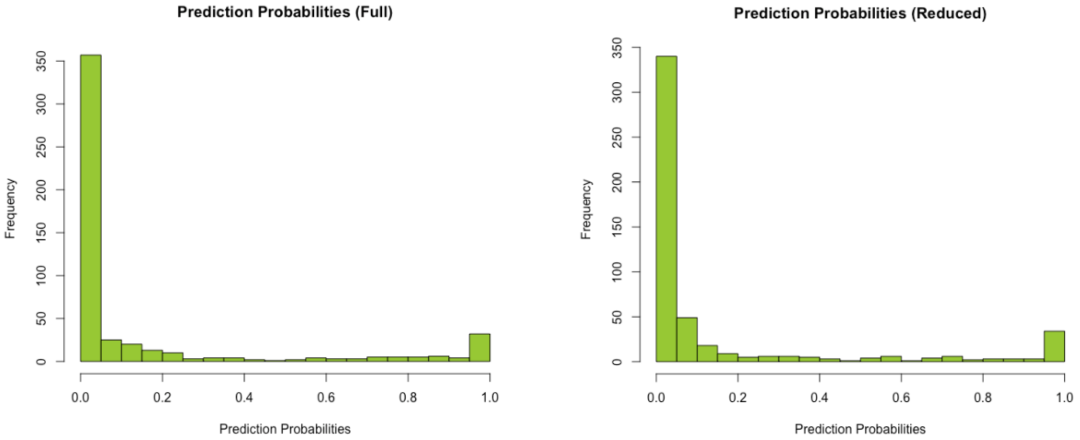

Figure 15: Distribution of Prediction Probabilities for Both Models

Figure 15 depicts the distribution of prediction probabilities for the full and reduced models. The histograms show the frequency of predicted probabilities for amyloid positivity across the dataset.

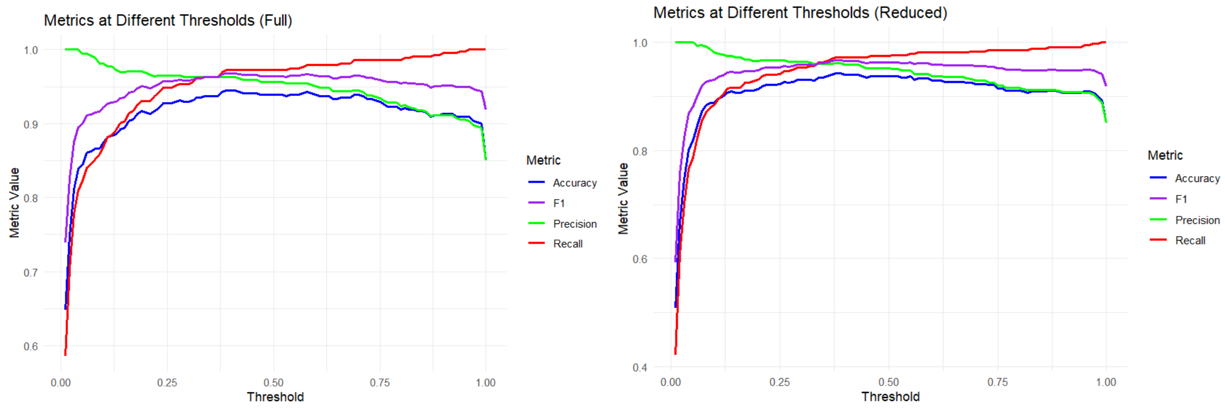

Figure 16: Metrics at Different Thresholds for Both Models

Figure 16 presents the performance metrics - accuracy, F1 score, precision, and recall - at varying thresholds for the "Full" and "Reduced" models. The graphs illustrate how these metrics evolve as the threshold for predicting amyloid positivity is adjusted, providing insight into the models' behavior and performance across different decision boundaries.

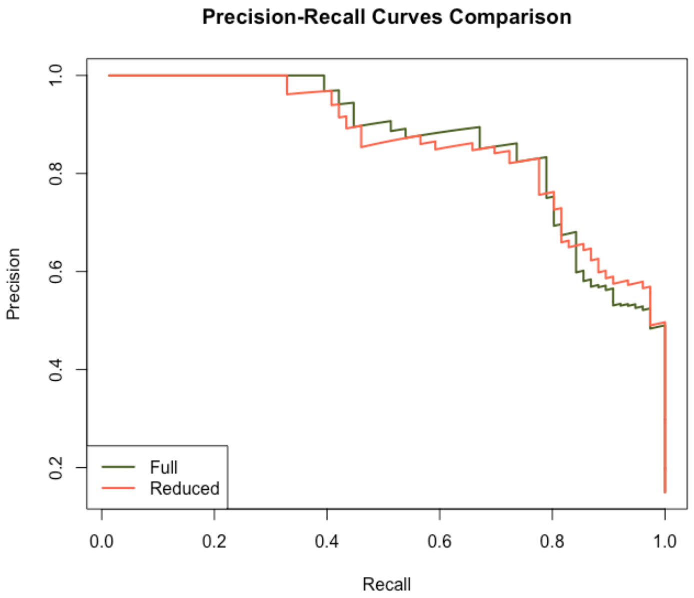

Figure 17: Precision-Recall Curve for Both Models

Figure 17 shows the precision-recall curves for the "Full" and "Reduced" models. The curves compare precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positives) across different thresholds. This analysis highlights the trade-off between precision and recall and helps evaluate the models' effectiveness in identifying true positives in the presence of imbalanced classes.

4.4.3 Summary

In comparing the full and reduced models' performance in predicting amyloid positivity, we find very similar results across key metrics. Both models demonstrate high accuracy, with the full model at 93.9% and the reduced model at 93.7%. Furthermore, their specificity (true negative rate) remains consistent, with the full model achieving 97.2% and the reduced model achieving 97.5%.

Moreover, while the full model exhibits a slightly higher sensitivity (true positive rate) of 75% compared to 72.4% for the reduced model, this difference is relatively small. Likewise, their AUC (Area Under the Curve) values are very close, with the full model at 0.969 and the "Cheap" model at 0.971, indicating nearly identical performance.

However, in our logistic regression model comparison, we evaluated the performance of the full and reduced models. The ANOVA test results revealed a significant difference in model fit between these models, $p = 0.00071$. This indicates that the additional predictor variables included in the full model significantly enhance the model's predictive power compared to the reduced model.

Based on the statistical findings, it appears that there is a limited discrepancy in the performance between the "full" and "reduced" models. This suggests that variables typically considered more "expensive”, such as cognitive status and genetic markers, may not notably influence the predictive capability of the model. Consequently, employing either model is likely to yield comparable outcomes, even in the absence of the "expensive" variables, thereby enhancing accessibility. However, it's important to note that these observations are based solely on numerical disparities, and interpretations may vary. Additionally, while we can discuss trade-offs and offer recommendations, definitive conclusions about what constitutes significant influence or the best approach should be approached with caution, given our expertise lies in statistical analysis rather than medical inference.

4.5 Linear Regression

To address the second objective of predicting amyloid PET levels, we employed linear regression models using plasma biomarkers and covariates. Initially, we tested all possible combinations of these predictors, including both ‘apoe_ternary’ and ‘apoe_binary’, to identify the most effective combinations. We also tested different models using test sets, different variable combinations, and other regularization techniques. Ultimately, the models that were selected were based on the lowest Predictive Residual Sum of Squares (PRESS) values, which helped us select the best predictors. This approach aided us in achieving objectives three and four as well.

4.5.1 Prediction Selection

In our linear regression analysis, we utilized the Predictive Residual Sum of Squares (PRESS) statistic to determine the most effective set of predictors for our models. For more information on PRESS and its use in our model refer to *Appendix A: Summary of Statistical Terms.*

The models resulting in the lowest PRESS values were considered to be most suitable for predicting Amyloid PET levels. Ultimately, our full model included the following predictors: ‘pTau181’, ‘pTau217’, ‘GFAP’, ‘NFL’, ‘Ab_ratio’, ‘sex’, ‘Calculated_Consensus_dx’, and ‘apoe_ternary’; yielding a PRESS statistic of 3.633477077. Conversely, the reduced linear model includes: ‘pTau217’, ‘GFAP’, ‘NFL’, ‘Ab_ratio’, and ‘sex’; resulting in a PRESS statistic of 3.696805781.

4.5.2 Initial Model Analysis:

We started tuning the models by plotting both models (Figure 18) and began to establish the four assumptions required for a linear model: linearity, independence, normality, and homoscedasticity.

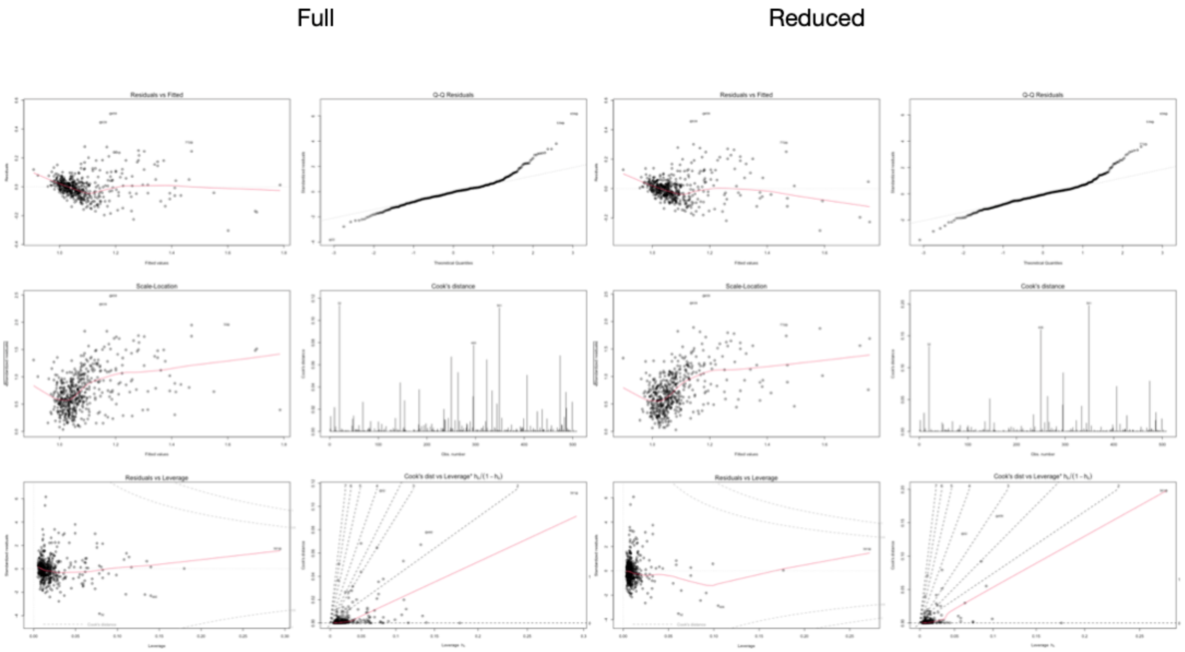

Figure 18: Initial Plots of Full and Reduced Linear Regression Models

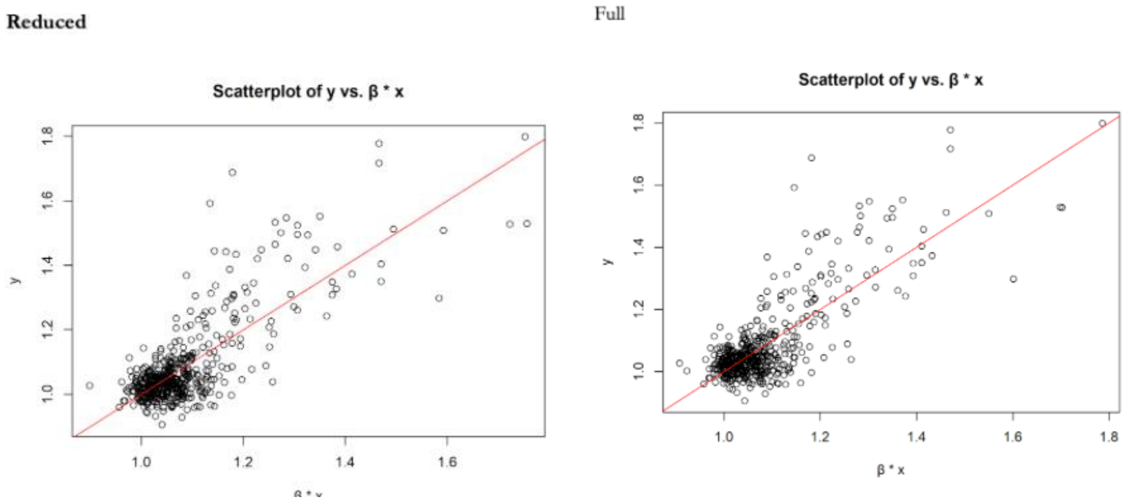

Figure 19: Linear Relationships Predictors v. Response

A linear relationship between our response and predictor variables was visibly confirmed with a plot of 'Amyloid_Pet' against each models' respective predictors. While not ideal, we found acceptable results, we then moved on to address the other assumptions.

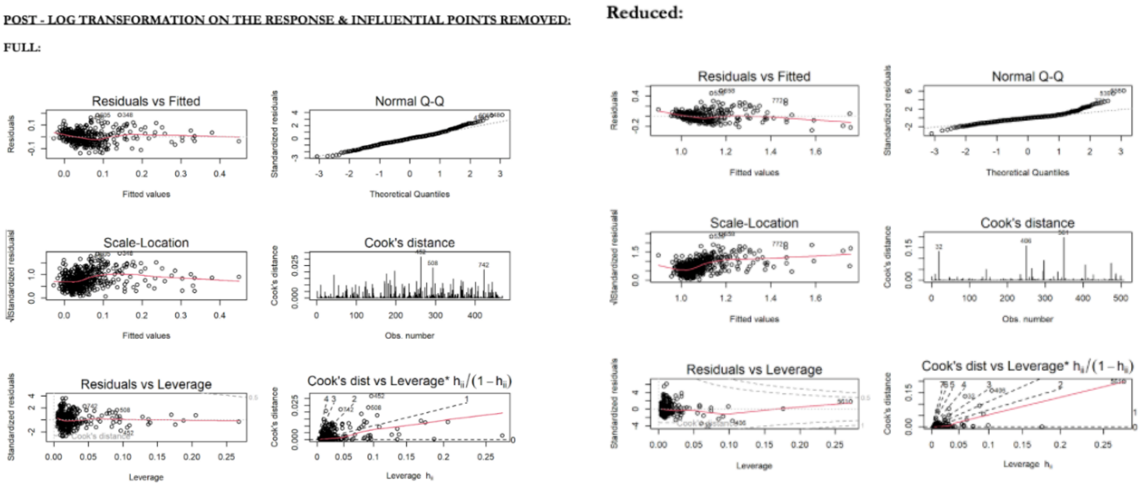

Figure 20: Plots of Model Fit Post-Response Log Transformation

Log transformation is widely used to address several major issues with response variables. Firstly, it stabilizes the variance of residuals, correcting their heteroscedasticity and ensuring homoscedasticity, which is vital for the validity of the linear regression assumptions. Secondly, it corrects skewed distributions, making them more symmetric and closer to a normal distribution. It also linearizes non-linear relationships between independent and dependent variables, thus simplifying the interpretation of the model and improving model fit. Lastly, it greatly reduces the impact of outliers by compressing the data range. The application of a log transformation to the Amyloid_PET data enhances the validity and interpretability of the regression models, resulting in more robust outcomes.

An ANOVA was done to compare the performance of linear models with and without log-transformed response variables in predicting Amyloid PET levels. The first one compared a full model with the non-transformed response variable against a log-transformed version of the same model. Both models had a residual degrees of freedom of 497 and a residual sum of squares of 3.406, indicating no difference between the models(Fd = 0, Sum of Sq = 0). A second ANOVA compared a reduced model with the non-transformed response variable, which included predictors. Both models had a residual degrees of freedom of 502 and a residual sum of squares of 3.5598, showing, again, that the models do not differ (Df = 0, Sum of Sq = 0). In both comparisons, the lack of change in the residual sum of squares and degrees of freedom indicates that log-transforming the response variable does not significantly alter the explanatory power of the models in predicting Amyloid PET levels but does help address the four assumptions.

4.5.3 Performance Analysis:

This section compares the performance of the two prediction models, referred to as Full and Reduced, in predicting the amyloid PET levels. Utilizing Leave One Out Cross Validation, our evaluation includes metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R squared (R2), and adjusted R squared. It summarizes the effectiveness of both models in accurately estimating amyloid PET levels and their prediction capabilities.

The full model includes a more comprehensive set of predictors: ‘pTau181’, ‘pTau217’, ‘GFAP’, ‘NFL’, ‘Ab_ratio’, ‘sex’, ‘Calculated_Consensus_dx’, and ‘apoe_ternary’. In contrast, the reduced model uses a more limited set of predictors: ‘pTau217’, ‘GFAP’, ‘NFL’, ‘Ab_ratio’, and ‘sex’. The difference in predictors is based on the ease of data collection. Predictors such as ‘Calculated_Consensus_dx’ and ‘apoe_ternary’ require more expensive and intrusive testing, whereas the data for the predictors for our reduced model are available from a plasma sample and more easily gathered information such as age or sex.

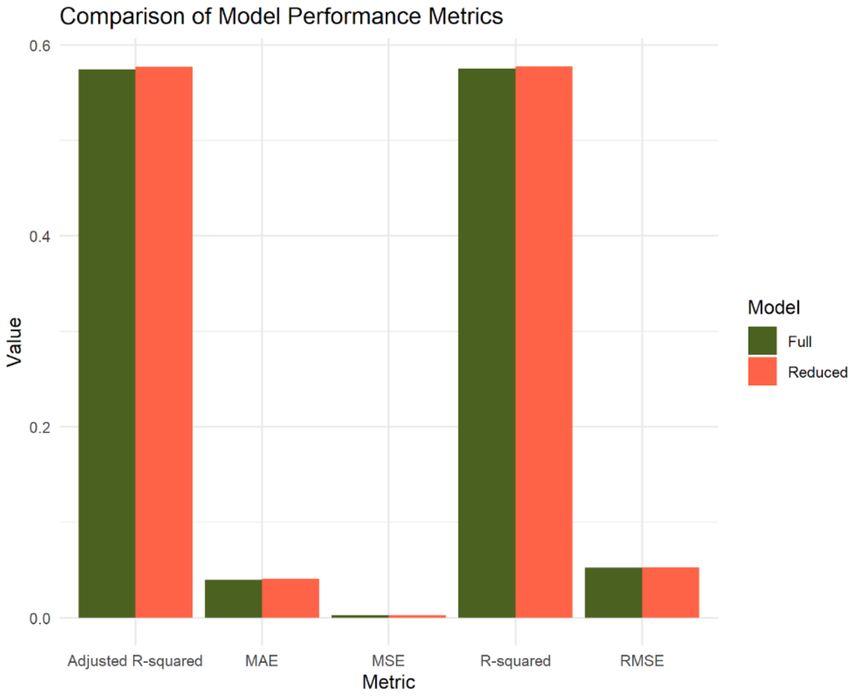

Figure 21: Model Performance Metrics Comparisons

Metric	Full Model	Reduced Model
MSE	0.002721621	0.002788251
RMSE	0.052169156	0.052803895
MAE	0.039942982	0.040496524
R Squared	0.575370810	0.577787743
Adjusted R Squared	0.574461540	0.576893225

Table 5: Linear Model Results Comparison

For detailed explanations and formulas of the metrics used, please refer to *Appendix A: Summary of Statistical Terms.*

The full model, which includes all additional predictors, demonstrates strong performance. This model shows an adjusted R-Squared of 0.57446, indicating the model explains about 57.45% of the variance in amyloid PET levels. It also has a relatively low RMSE of approximately 0.0521 and an MAE of about 0.0399, showing that, on average, the model's predictions are close to the actual values. Despite its complexity, the full model provides a decent fit to the data and good predictive values for amyloid PET levels according to our predictors.

In contrast, the reduced linear regression model with fewer predictors also performs moderately well. An adjusted R-Squared value of around 0.5769 indicates that this model explains about 57.69% of the variance in amyloid PET levels. The model gives slightly higher RMSE and MAE values compared to the full model, suggesting that it is less accurate. However, it still provides a reasonably good fit to the data, albeit less precise, highlighting the importance of some predictors in estimating amyloid PET levels.

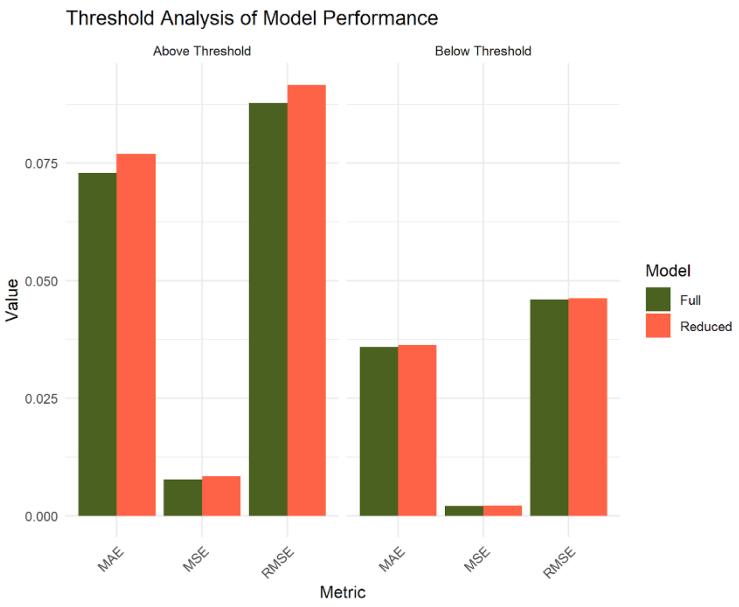

Figure 22: Threshold Analysis for Full and Reduced Models

Below Amyloid positivity threshold:

Metric	Value	Model	Metric	Value	Model
MSE	0.0021050	Reduced	MSE	0.0021016	full
RMSE	0.0458800	Reduced	RMSE	0.0458433	full
MAE	0.0360378	Reduced	MAE	0.0358969	full

Table 6: Performance below Amyloid positivity threshold

Above Amyloid positivity threshold:

Metric	Value	Model	Metric	Value	Model
MSE	0.0085467	Reduced	MSE	0.0074748	full
RMSE	0.0924481	Reduced	RMSE	0.0864570	full
MAE	0.0779106	Reduced	MAE	0.0716533	full

Table 7: Performance above Amyloid positivity threshold

Both models exhibit substantially lower error metrics below the threshold compared to their performance above it. This suggests that the models are more accurate and make closer predictions to the actual values when amyloid levels are below the threshold of 1.17. The reduced model shows slightly stronger performance in all three metrics (MSE, RMSE, MAE) compared to the full model under the same conditions. This is particularly noteworthy as it suggests that, despite the simpler predictor set, the reduced model is quite efficient in conditions below the threshold.

4.5.4 Summary

The trade-offs between the two models are minimal. The full model, with a higher adjusted R-Squared value, suggests a better fit and more accurate predictions, at the cost of increased complexity. The reduced model, while simpler and slightly less accurate, still offers valuable insights and maintains a good balance between simplicity and performance.

Our findings indicate that while neither model is flawless, they provide valuable tools for predicting amyloid PET levels based on plasma biomarkers and demographic factors. The full model suggests that approximately 57.45% of the variance in amyloid PET levels can be explained by its predictors, while the reduced model explains around 57.69%. The most significant biomarkers aiding in these predictions are 'pTau217', 'GFAP', and 'Ab_ratio', as indicated by their coefficients and statistical significance in both models.

These results describe the potential utility of these models in a clinical setting. By incorporating additional predictors, the full model achieves greater predictive accuracy. However, the reduced model demonstrates that even with fewer predictors, significant insights can be drawn. Future improvements might involve refining the selection of predictors to capture additional variance in amyloid PET.

By leveraging statistical methods and metrics, such as adjusted R-Squared, RMSE, and MAE, we ensure that our conclusions are grounded in a solid quantitative analysis, while avoiding overstepping into medical recommendations beyond our expertise.

5 Conclusion / Recommendations

This project set out to address several key questions regarding the prediction of amyloid positivity and amyloid PET levels using plasma biomarkers and demographic factors in the context of Alzheimer's disease. Through a comprehensive comparison of both logistic and linear regression models, we not only aimed to answer these questions but also identified key insights and potential applications for these models.

Our analysis indicates that while both the full and reduced linear and logistic models demonstrate strong predictive power, the inclusion of additional predictors in the full model enhances its accuracy, albeit with increased cost and complexity. Despite this, the reduced models, which exclude more costly variables, still provide valuable insights and maintain robust performance, suggesting that effective predictions can be made with a more accessible set of predictors.

Based on our findings, we recommend the following:

1. <u>Prioritize Model Selection:</u>

   When the primary goal is to achieve the highest possible predictive accuracy and the resources are available, the full models should be employed due to their superior fit and enhanced predictive power. However, in resource-constrained settings, the reduced models offer a viable alternative, maintaining substantial predictive capability without the need for more expensive variables.

2. Ensure Data Quality:

   Ensure consistent and thorough data cleaning and preprocessing to maintain the quality and reliability of the predictive models. Additionally, incorporate additional predictors carefully, making sure to weigh the trade-offs between model complexity and predictive accuracy

3. Regular Evaluation and Refinement:

   Regularly re-evaluate the models with new data to ensure they remain accurate and relevant, adapting to potential changes in underlying data patterns.

4. Practical Application Considerations:

   Select models based on both statistical performance and practical feasibility, ensuring ease of use in real-world settings.

In conclusion, our research underscores the significance of predictive modeling in advancing early detection strategies for Alzheimer's disease using accessible predictors like plasma biomarkers and demographic factors. While recognizing the value of both full and reduced models in predicting amyloid PET positivity, we can emphasize the necessity of continuous refinement and validation efforts. By leveraging statistical methods and expanding our methodological toolkit, we can enhance the reliability and applicability of these predictive models in various settings. Moving forward, it is crucial to prioritize ongoing evaluation and refinement to ensure the relevance and effectiveness of these predictive models in clinical practice or research endeavors.

6 Future Improvements

In the future, with a larger dataset, incorporating a mixed effects model could become a viable option. In lieu of that, additional strategies such as data synthesis or stratification could make the dataset more robust, ensuring adequate representation of all subgroups. Furthermore, some improvements could be made in adopting different methods such as decision trees or random forests. It's important to recognize that random forests are an extension of decision trees, wherein multiple trees are generated and aggregated to create predictions. This ensemble technique combines the qualities of individual decision trees mitigating overfitting while augmenting predictive precision. Furthermore, these methods excel at capturing complex interactions and nonlinear relationships within the data that may not be captured by linear models alone. Additionally, they can handle categorical variables without the need for one-hot encoding which can help simplify the modeling process. Overall, implementing these methods offers an opportunity to explore the data from a different perspective and potentially uncover additional insights that may not be apparent with traditional linear modeling techniques alone.

7 Appendices

7.1 Appendix A: Summary of Statistical Terms

Linear Regression:

Linear regression is a statistical method of modeling the relationship between a dependent variable and one or more independent variables. The model assumes that the relationship between the variables is linear and can be represented with the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

In our case our Y is Amyloid PET and our are coefficients and X are the independent variables; Plasma Biomarkers.

Logistic Regression:

Logistic regression is a statistical method of modeling the probability of a binary outcome based on one ore more predictor variables. It uses the logistic function to model the odds of the dependent variable being positive or negative. The model can be represented as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

- Where p is the probability of the event being positive.

LOOCV:

Leave One Out Cross Validation is a method of cross validation where a single observation from the dataset is used as the validation set, which leaves our remaining observations to be used as the training set. This process is then repeated so that each observation in the dataset is used exactly once as the validation set.

PRESS:

PRESS is a metric used to evaluate the predictive accuracy of a regression model. It assesses how well a model can predict new data, rather than merely fitting the existing data.

PRESS is calculated using leave-one-out cross-validation(LOOCV), where each observation is excluded one at a time, and the model is refitted to predict the excluded observation. At each fold, the model is trained on all data points except one, and the prediction error for the excluded data point is computed. Specifically, the prediction error is the difference between the observed value for the excluded data point. The PRESS statistic is the sum of the squared prediction errors for all observations.

Mathematically, if $\hat{y}_{i(-i)}$ represents the predicted value for the i^{th} observation when it is excluded from the model fitting, and y_i is the observed value, then press is given by:

$$PRESS = \sum_{i=1}^{n} y_i - \hat{y}_{i(-i)}$$

A lower PRESS value indicates a model with better predictive performance.

LOG LOSS:

Log Loss measures the performance of a classification model where the output is a probability value between 0 and 1. It is defined as:

$$Log\ Loss\ = -\frac{1}{n}\sum_{i=1}^{n}[y_i log(p_i) + (1 - y_i)log(1 - p_i)]$$

Where *n* is the number of samples, *y* is the actual label, and *p* is the predicted probability.

ANOVA (Analysis of Variance):

Analysis of Variance (ANOVA) is a statistical method used to compare the means of three or more groups to determine if at least one group mean is significantly different from the others. It partitions the total variance observed in the data into variance components attributable to different sources.

Mixed Effect Model:

Mixed Effect Model is a statistical model that contains both fixed effects and random effects. These models are particularly useful for analyzing data with a hierarchical or clustered structure, where measurements are nested within higher-level units such as subjects, groups, or time points.

Fixed effects are parameters that are assumed to be the same across the entire population. In the context of our study, fixed effects include variables such as demographic characteristics and other covariates that are consistent for all subjects. These effects provide information about the general population-level trends and relationships.

Random effects are parameters that account for the variability within higher level units. When data is collected from multiple subjects, each subject has unique characteristics that influence the outcome. So by including random effects, we can model this variability.

The mixed effect model used is defined as

$$response\_variable = \beta_0 + \beta_1 fixed\_effect_1 + \beta_2 fixed\_effect_2 + \alpha_{id} + \varepsilon$$

Where each are the fixed effect coefficients, is the random intercept for each subject, and is the residual error term.

In specialized medical cases, there are additional random effects included to capture additional levels of variability. Examples of this include using are a random effect for treatments done within each subject.

AIC (Akaike Information Criterion):

Akaike Information Criterion(AIC) is a measure used to compare the goodness of fit of different statistical models. It considers the complexity of the model and the likelihood of the model given the data. A lower AIC value indicates a better model. It is defined as:

$$AIC\ =\ 2k - 2log(L)$$

Where k is the number of parameters in the model and L is the likelihood function.

Calibration Curve:

A Calibration Curve is a graphical representation used in binary classification to assess how well the predicted probabilities of a model align with actual outcomes. The curve plots the fraction of positive cases against the predicted probability, ideally forming a 45-degree line for a perfectly calibrated model.

Precision-Recall Curve:

Precision-Recall curves plot the trade-off between the true positive rate and the positive predictive value using different probability threshold**s.**

ROC/AUC (Receiver Operating Characteristic/ Area Under the Curve):

Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Area Under the Curve (AUC) measures the entire two-dimensional area under the ROC curve, providing a single value to summarize the performance of the classifier. Higher AUC values indicate better performance.

Sensitivity (Recall):

Sensitivity (Recall) measures the proportion of actual positives that are correctly identified by the model. It is defined as:

$$Sensitivity = TruePositives/(TruePositives + FalseNegatives)$$

High sensitivity indicates that the model effectively identifies positive cases.

Specificity:

Specificity measures the proportion of actual negatives that are correctly identified by the model. It is defined as:

$$Specificity = TrueNegatives/(TrueNegatives + FalsePositives)$$

High specificity indicates that the model effectively identifies negative cases.

Precision:

Precision measures the proportion of positive identifications that are truly correct. It is defined as:

$$Precision = TruePositives/(TruePositives + FalsePositives)$$

High Precision indicates that the model has a low false positive rate.

<u>F1 Score</u>:

F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is defined as:

$$F1Score = 2\big(Precision * Recall/(Precision + Recall)\big)$$

The F1 score is useful when needing to balance between precision and recall.

<u>Root Mean Squared Error (RMSE)</u>:

RMSE is a measure of the average deviation between predicted and actual values. Lower values indicated better model performance. Mathematically, it is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y})^2}$$

<u>Mean Absolute Error (MAE)</u>:

MAE is a measure of the average prediction error, where lower values signify better accuracy. It is calculated as:

$$MAE \ = \frac{1}{n}\sum\nolimits_{i=1}^{n}|y_i - \hat{y}|$$

<u>R-Squared (Coefficient of Determination)</u>:

R-squared represents the proportion of variance in the dependent variable that is predictable from the independent variables. Higher values indicate better fit. It is defined as:

$$R^2 \ = \ 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

<u>Adjusted R-Squared</u>:

Adjusted R-squared is an adjusted version of R-Squared that accounts for the number of predictors in the model, providing a more accurate measure of model fit, especially for models with multiple predictors. It is defined as:

$$Adjusted_R^2 = 1 - \frac{(1 - R^2)(n - 1)}{(n - p - 1)}$$

7.2 Appendix B: Merged Dataset Variables

The merged dataset comprises 23 variables across 508 observations from 266 unique subjects. The table below details each variable, its description, and category levels where appropriate.

Variable	Description	Levels	Predictor/ Response
subject_id_number	Unique Identifier for each subject	---	---
visno	Visit number for the Subject	---	---
age_at_appointment	Age at time of Amyloid PET test	---	Predictor
age_at_acquisition	Age at time of plasma biomarker acquisition	---	---
time_diff	Time difference between amyloid PET and matched nearest plasma test	---	Predictor
Amyloid_PET	Amyloid PET results	---	Response
amyloid_binary	Binary indicator of Amyloid Positivity based on PET results with threshold (1.17)	0 : Negative 1: Positive	Response
ptau231	Phosphorylated tau at position 231	---	Predictor
ptau181	Phosphorylated tau at position 181	---	Predictor
pTau217	Phosphorylated tau at position 217	---	Predictor
GFAP	Glial Fibrillary acidic protein	---	Predictor
NFL	Neurofilament light-chain	---	Predictor
Ab40	Amyloid-β 40	---	---
Ab42	Amyloid-β 42	---	---
Ab_ratio	Ratio: Ab42 / Ab40	---	Predictor
apoe_e1	APOE genotype from one biological parent	2, 3, 4	---
apoe_e2	APOE genotype from other biological parent	2, 3, 4	---
apoe_ternary	Indicates count of "4" alleles	0, 1, 2	Predictor
apoe_binary	Indicates any "4" alleles	0, 1	Predictor
Calculated_Consensus_dx	Cognitive status at each visit	Cog_Unimpaired_Stable Cog_Unimpaired_Declining Impaired_Not_MCI Clinical_MCI Dementia	Predictor
sex	Sex of subject	M: Male, F: Female	Predictor
race_primary	Primary race of subject	American Indian or Alaskan Native Asian Black or African American White Other	---
hispanic	Hispanic ethnicity status	Yes, No	---

Table 8: Description of Data of Merged Dataset

7.3 Appendix C: Qualitative Data Merging

Pre-Merging & Removal of Variables

VARIABLE	CATEGORIES	COUNTS			RELATIVE FREQUENCIES (%)		
		Amyloid Negative	Amyloid Positive	TOTAL Count	Amyloid Negative	Amyloid Positive	% of TOTAL
amyloid_binary	Negative (0)	432	---	432	100.0	---	85.04
	Positive (1)	---	76	76	---	100.0	14.96
sex	Female	290	48	338	67.13	63.16	66.55
	Male	142	28	170	32.87	36.84	33.46
race_primary	American Indian or Alaskan Native	4	---	4	0.93	---	0.79
	Asian	3	---	3	0.69	---	0.59
	Black or African American	9	5	14	2.08	6.58	2.76
	Other	3	---	**3**	0.69	---	**0.59**
	White	413	71	484	95.60	93.42	95.28
hispanic	No	429	76	505	99.31	100.0	99.41
	Yes	3	---	**3**	069	---	**0.59**
Calculated_Consensus_dx	Cog_Unimpaired_Stable	361	51	421	83.56	67.11	81.10
	Cog_Unimpaired_Declining	65	16	81	15.05	21.05	15.94
	Impaired_Not_MCI	1	1	2	0.23	1.32	0.39
	Clinical_MCI	5	7	12	1.16	9.21	2.36
	Dementia	---	1	1	---	1.32	0.20
apoe_ternary	0	282	20	302	65.28	26.32	59.45
	1	141	45	186	32.63	59.21	36.61
	2	9	11	20	2.08	14.47	3.94
apoe_binary	0	282	20	302	65.28	26.32	59.45
	1	150	56	206	34.71	73.68	40.55

Table 9: Qualitative Data Pre-Merging & Removal of Variables Not Considered in Models

Post-Merging & Removal of Variables

VARIABLE	CATEGORIES	COUNTS			RELATIVE FREQUENCIES (%)		
		Amyloid Negative	Amyloid Positive	TOTAL Count	Amyloid Negative	Amyloid Positive	% of TOTAL
amyloid binary	Negative (0)	432	---	432	100.0	---	85.04
	Positive (1)	---	76	76	---	100.0	14.96
sex	Female	290	48	338	67.13	63.16	66.55
	Male	142	28	170	32.87	36.84	33.46
Calculated_Consensus_dx	Cog_Unimpaired_Stable	361	51	421	83.56	67.11	81.10
	Unimpaired_Declining_or_Impaired	66	17	83	15.28	22.37	16.33
	MCI_or_Dementia	5	8	13	1.16	10.53	2.56
apoe_ternary	0	282	20	302	65.28	26.32	59.45
	1	141	45	186	32.63	59.21	36.61
	2	9	11	20	2.08	14.47	3.94
apoe_binary	0	282	20	302	65.28	26.32	59.45
	1	150	56	206	34.71	73.68	40.55

Table 10: Qualitative Data Post-Merge & Removal of Variables Not Considered in Models

7.4 Appendix D: Logistic Regression Model Summaries

Best Model (Full)

COEFFICIENTS Of Full Model				
Predictor	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	**-9.422835**	**3.500690**	**-2.692**	**0.00711**
ptau231	**0.083147**	**0.062472**	**1.331**	**0.18321**
ptau181	**-0.734326**	**0.326132**	**-2.252**	**0.02435**
pTau217	**9.456246**	**1.321645**	**7.155**	**8.37e-13**
GFAP	**0.009307**	**0.003709**	**2.509**	**0.01209**
NFL	**-0.108202**	**0.042015**	**-2.575**	**0.01001**
Ab_ratio	**-47.387079**	**19.012820**	**-2.492**	**0.01269**
age_at_appointment	**0.088998**	**0.046712**	**1.905**	**0.05675**
Apoe_ternary1	**1.520416**	**0.533593**	**2.849**	**0.00438**
Apoe_ternary2	**3.261428**	**0.836596**	**3.898**	**9.68e-05**

Table 11: Full Logistic Regression Model Summaries

Best Model (Reduced):

COEFFICIENTS Of Full Reduced Model				
Predictor	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	**2.368254**	**1.385831**	**-1.709**	**0.08747**
ptau181	**-0.454730**	**0.250811**	**-1.813**	**0.06983**
pTau217	**9.740952**	**1.211970**	**8.037**	**9.18e-16**
GFAP	**0.009671**	**0.003289**	**2.941**	**0.00328**
NFL	**-0.068640**	**0.029681**	**-2.313**	**0.02074**
Ab_Ratio	**-57.378951**	**17.619265**	**-3.257**	**0.00113**

Table 12: Reduced Logistic Regression Model Summaries

7.5 Appendix E: Linear Regression Model Summaries

Best Model (Full):

COEFFICIENTS for Full Model				
Predictor	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	**1.088565**	**0.005708**	**190.719**	**< 2e-16**
ptau181	**-0.012874**	**0.005489**	**-2.345**	**0.01941**
ptau217	**0.106610**	**0.006013**	**17.730**	**< 2e-16**
GFAP	**0.013945**	**0.004749**	**2.936**	**0.00347**
NFL	**-0.008711**	**0.004428**	**-1.968**	**0.04968**
Ab_ratio	**-0.020798**	**0.003917**	**-5.310**	**1.66e-07**
Sex	**-0.020066**	**0.008078**	**-2.484**	**0.01332**
Calculated_Consensus_dxMCI	**-0.083275**	**0.026205**	**-3.178**	**0.00158**
Calculated_Consensus_dxUnimpared	**0.008114**	**0.010180**	**0.797**	**0.42582**
Apoe_ternary1	**0.002503**	**0.007985**	**0.314**	**0.75403**
Apoe_ternary2	**0.051814**	**0.019522**	**2.654**	**0.00821**

Table 13: Full Linear Regression Model Summaries

The full model finds significant predictors including pTau217, GFAP, Ab Ratio and Calculate Consensus MCI. While sex shows a smaller significance effect.

Best Model (Reduced):

COEFFICIENTS for Reduced MODEL				
Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.090752	0.004608	236.686	< 2e-16
pTau217	0.093418	0.004322	21.616	< 2e-16
GFAP	0.013836	0.004793	2.887	0.00406
NFL	-0.009427	0.004453	-2.117	0.03475
Ab_Ratio	-0.020211	0.003908	-5.171	3.37e-07
Sex	-0.020175	0.008062	-2.503	0.01265

Table 14: Reduced Linear Regression Model Summaries

The reduced model finds pTau217 and Ab Ratio having the greatest significance, which is coherent with the expensive model. These coefficients provide aid in finding the most important biomarkers for predicting Amyloid PET.

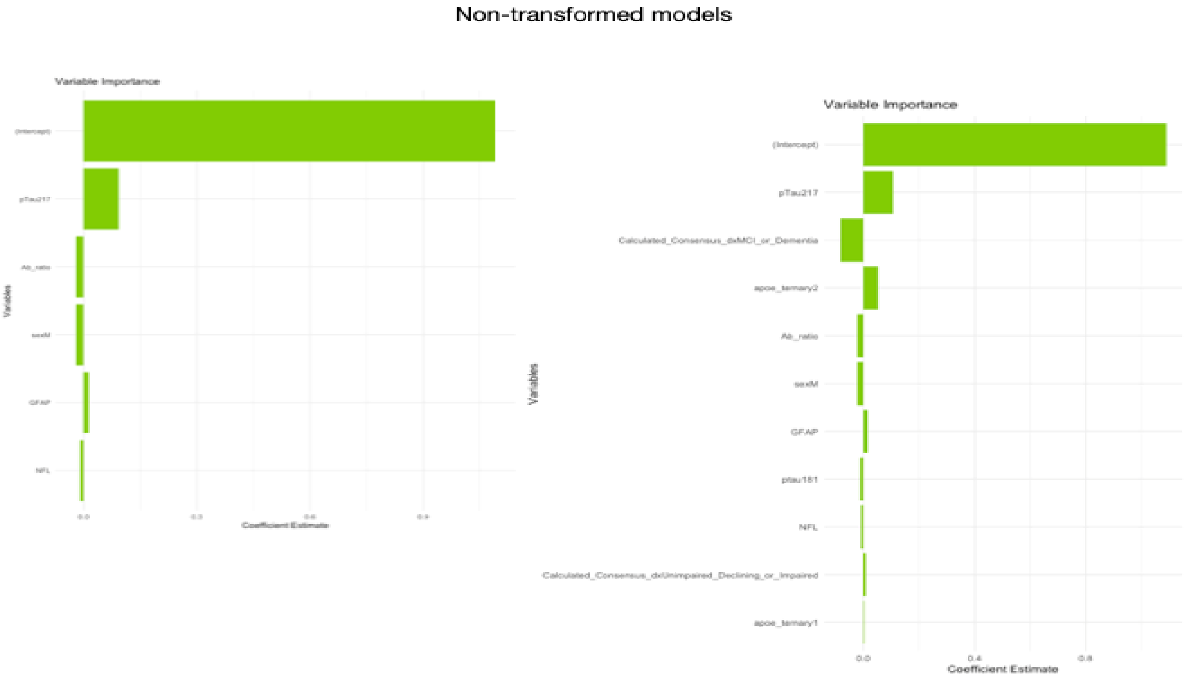

Figure 23: Predictor Influence - Non-Transformed Model

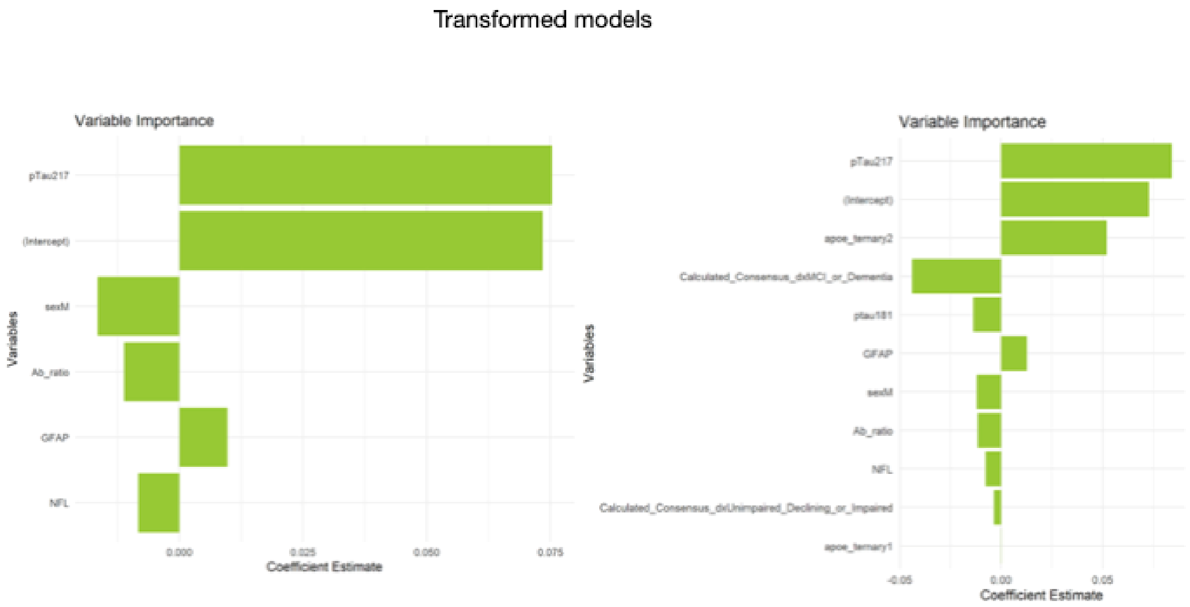

Figure 24: : Predictor Influence - Transformed Model

Metric	Value
Wilcoxon Signed Rank Test V:	54233
Wilcoxon Signed Rank Test p-value	.9835

Table 15: Willcox Full Linear Model

The Wilcoxon signed rank test compares the predictions of the Full Model and the Reduced Model. The high p-value (0.9835) indicates no significant difference between the predictions of the two models, suggesting that both models perform similarly.

Metric	Value
AIC for Full Model	-1436.509
AIC for Reduced Model	-1440.365
BIC for Full Model	-1386.702
BIC for Reduced Model	-1411.237

Table 16: AIC/BIC Linear Models

The table presents the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for two models: the Full Model and the Reduced Model. The Reduced Model has lower AIC and BIC values, indicating a better fit and simpler model compared to the Full Model

7.6 Appendix F: R-Code

Due to the length of the R-Scripts, we have decided not to include them as an appendix. Please reference the R scripts for details on any implementation in R. They have been detailed with comments for clarity throughout the documents.

https://drive.google.com/drive/folders/1NIJhb0L7fTEOCBBJMpBendwNs75T0U8n?usp=drive_link

8 References

Andreasson, U., Zetterberg, H., et al. (2021). Update on ultrasensitive technologies to facilitate research on blood biomarkers for central nervous system disorders. *Molecular Neurodegeneration, 16*(1). https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-021-00451-6

Amft, M., Ortner, M., Eichenlaub, U., Goldhardt, O., Diehl-Schmid, J., Hedderich, D. M., Yakushev, I., & Grimmer, T. (2022). The cerebrospinal fluid biomarker ratio Aβ42/40 identifies amyloid positron emission tomography positivity better than Aβ42 alone in a heterogeneous memory clinic cohort. Alzheimer's research & therapy, 14(1), 60. https://doi.org/10.1186/s13195-022-01003-w

Ashton, N. J., Pascoal, T. A., Karikari, T. K., et al. (2021). Plasma p-tau231: a new biomarker for incipient Alzheimer's disease pathology. *Acta Neuropathol, 141*(5), 709-724. doi:10.1007/s00401-021-02275-6

Barthélemy, N. R., Horie, K., Sato, C., & Bateman, R. J. (2020). Blood plasma phosphorylated-tau isoforms track CNS change in Alzheimer's disease. *Journal of Experimental Medicine, 217*(e20200861). doi:10.1084/jem.20200861

Corder, E. H., et al. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science, 261*(5123), 921-923. https://doi.org/10.1126/science.8346443

Cullen, N. C., et al. (2021). Plasma biomarkers of Alzheimer's disease improve prediction of cognitive decline in cognitively unimpaired elderly populations. *Nature Communications, 12*, 238. https://www.nature.com/articles/s41467-021-23746-0#ref-CR1

Filho, M. (n.d.). Guia Completo da Log Loss: Perda Logarítmica em Machine Learning. https://mariofilho.com/guia-completo-da-log-loss-perda-logaritmica-em-machine-learning/

Hampel, H., et al. (2018). Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology, 14*(11), 639-652. https://doi.org/10.1038/s41582-018-0079-7

Hansson, O. (2021). Biomarkers for neurodegenerative diseases. *Nature Medicine*. doi:10.1038/s41591-021-01382-x

Jack Jr, C. R., et al. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535-562. https://alz-journals.onlinelibrary.wiley.com/doi/full/10.1016/j.jalz.2018.02.018

Jagust, W. J., et al. (2009). The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. *Alzheimer's & Dementia, 6*(3), 221-229. https://doi.org/10.1016/j.jalz.2009.12.003

Jedynak, B. (2024). Client presentation on the Wisconsin Registry for Alzheimer's Prevention (WRAP) study.

Jia, L., et al. (2020). The APOE ε4 exerts differential effects on familial and other subtypes of Alzheimer's disease. *Alzheimer's & Dementia, 16*, 1613-1623. doi:10.1016/j.jalz.2020.06.012

Jia, L., et al. (2023). Plasma biomarkers predict Alzheimer's disease before clinical onset in Chinese cohorts. *Nature Communications, 14*, 6747. https://doi.org/10.1038/s41467-023-42596-6

Klunk, W. E., et al. (2004). Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Annals of Neurology, 55*(3), 306-319. https://doi.org/10.1002/ana.20009

Palmqvist, S., et al. (2018). Plasma Aβ42/40 Ratio Detects Early Stages of Alzheimer's Disease and Correlates with CSF and Neuroimaging Biomarkers in the AB255 Study. *The Journal of Prevention of Alzheimer's Disease*. https://link.springer.com/article/10.14283/jpad.2018.41

Palmqvist, S., et al. (2021). Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. *Nature Medicine, 27*, 1034-1042. https://doi.org/10.1038/s41591-021-01382-x

Passe, M. P., et al. (2019). Assessment of plasma total tau level as a predictive biomarker for dementia and related endophenotypes. *JAMA Neurology, 76*, 598. https://jamanetwork.com/journals/jamaneurology/fullarticle/2727697

"PRESS: Allen's PRESS (Prediction Sum-Of-Squares) statistic, aka... in qpcR: Modelling and Analysis of Real-Time PCR Data." https://rdrr.io/cran/qpcR/man/PRESS.html

"PRESS function - RDocumentation." https://rdocumentation.org/packages/qpcR/versions/1.4-1/topics/PRESS

README file. (2023). Copy of README – plasma_simoa_gothenburg.txt.

Sperling, R. A., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3), 280-292. https://doi.org/10.1016/j.jalz.2011.03.003

Thijssen, E. H., et al. (2021). Diagnostic Accuracy of a Plasma Phosphorylated Tau 217 Immunoassay for Alzheimer Disease Pathology. *JAMA Neurology*. https://jamanetwork.com/journals/jamaneurology/fullarticle/2785187

Wisconsin Alzheimer's Institute. "Detecting Alzheimer's Disease: Brain Proteins and the Wisconsin Registry for Alzheimer's Prevention (WRAP) Community Event." https://wai.wisc.edu/wp-content/uploads/sites/1129/2024/02/DetectingAD_BrainProteins_WRAP_CommunityEventFlyer-1.pdf