Decoding Dementia

Andre Buser, Victor Adafinoaiei
Dec 11, 2023

1. Introduction and Objectives

Traumatic brain injuries (TBI) are a serious health issue that affects hundreds-thousands of people each year according to the Centers for Disease Control and Prevention (CDC). These injuries can range from a mild bump on the head to severe trauma, and they can lead to long-term problems, including the risk of developing dementia – a condition that impairs memory and thinking skills. As our population ages, understanding the connection between TBI and dementia is becoming increasingly important.

The brain is a delicate and complex organ, and how it reacts to injuries is not fully clear. After a TBI, there can be immediate damage as well as a series of changes in the brain that may contribute to diseases like dementia. Studies have shown that people who have had a TBI may be more likely to develop dementia, but we don't yet know exactly why this is the case.

Our project was designed to enhance our understanding of the outcomes following traumatic brain injuries, focusing on their determinants and consequences. We used causal inference to investigate the complex relationships between TBI and the subsequent risk of dementia, considering a range of demographic and clinical factors.

Our project is split into two main parts:

Step 1: Looking at People's Backgrounds and TBI Details
First, we're studying how TBI is linked to dementia and how factors like age, gender, education, genetics, and brain health scores play a role. We'll know we're on the right track if we:

- Find clear evidence that TBI and dementia are connected.
- Measure how strong this connection is and show that it's meaningful.

Step 2: Studying Proteins and Brain Changes
Next, we're examining the proteins[1] in the brain and other changes to find patterns that could explain the link between TBI and dementia. We'll consider this part successful if we:

- Group the data in a way that makes sense and is backed up by the numbers.
- Show clear differences in protein levels between these groups.
- Present our findings in easy-to-understand format

2. Related Work

In our research, we're building on the insights of several studies that use artificial intelligence (AI) and machine learning (ML) to delve into the complexities of dementia. We're particularly interested in how these technologies can help us make sense of the extensive dataset provided by the Allen Institute. Here's an overview of the influential work we've referenced.

Moura and Oliveira (2021) presented an article titled "What Do Machines Tell Us About Dementia? Machine Learning Applied to Aging, Dementia, and Traumatic Brain Injury Study." They used machine learning to analyze patterns in the way genes are expressed in the brain, which could indicate the

[1] For detailed explanations of specific technical terms please refer to the Appendix C: Glossary of Technical Terms

presence of dementia. Their research is crucial for us as we examine similar genetic data to identify potential markers of dementia.

In the 2023 study by Ranson et al., “Harnessing the potential of machine learning and artificial intelligence for dementia research,” the authors discuss the potential of AI to unlock new insights into disease mechanisms and aid in the discovery of new medications. They emphasize the value of analyzing diverse types of data to better understand and categorize the disease, which is particularly relevant to our project that utilizes a variety of data such as brain tissue images, gene activity measurements, and protein levels.

Lastly, Pölsterl and colleagues (2022) in their paper “Identification of causal effects of neuroanatomy on cognitive decline requires modeling unobserved confounders,” address the challenge of determining how changes in brain structure can influence cognitive decline in Alzheimer’s disease. They introduce a new method to account for factors that are not directly observed but can influence the results of brain imaging studies. This method is especially pertinent to our investigation into the relationship between brain injuries and dementia.

3. Data Description

We focused on three key data files from the Aging, Dementia, and Traumatic Brain Injury Study (Allen Institute for Brain Science, 2017):

Table 1: Dataset Files Utilized – full description in Appendix A: Extended Data Description

Description and Importance	**Rec.**[2]	**Attr.**[3]
DonorInformation.csv contains detailed information about individual donors, including various age-related characteristics.	107	19
ProteinAndPathologyQuantifications.csv offers quantified measurements related to proteins and pathologies[4] associated with brain aging or neurodegenerative disorders[4].	377	33
group_weights.csv contains subject and sampling weights as calculated and described in the Allen Institute's 'Technical White Paper: Weighted Analyses' (2016).	107	2

3.1 Data Cleaning and Exploration

DonorInformation.csv

To guarantee the reliability and usability of the "DonorInformation.csv" file, we initiated preliminary validation checks to confirm the trustworthiness and suitability of the data for our research requirements. The main actions carried out during this process are summarized in Table 2 (refer to Appendix B: Data Cleaning and Exploration: DonorInformation.csv for all details). Detailed distributions can be found in Appendix F: EDA - Feature Distributions.

[2] Records
[3] Attributes
[4] For detailed explanations of specific technical terms please refer to the Appendix C: Glossary of Technical Terms

Table 2: Data Cleaning and Transformation - DonorInformation.csv | *new feature

Attribute/Feature	**Insights and transformations**
act_demented_clean*	● Mapped the 'No Dementia' and 'Dementia' values to 0 and 1, respectively, in the 'act_demented_clean' attribute. ● Nearly evenly distributed between ”No dementia” (57) and “Dementia” (50).
age_at_first_tbi_bin*	● A significant number of TBI incidents occur during the early stages of life, as the majority of donors encounter their initial TBI before reaching the age of 30. Another notable concentration can be observed from the age of 60 onwards. ● To reflect this observation, we established three categories: early, mid, late: "early_years (1-30)", "mid_years (31-60)", "late years (61-90)".
age_clean*	● Created a new attribute to retain the quantitative data of the age using the median.
apo_e4_allele_clean*	● Mapped with 0 for 'N', 1 for 'Y', and -1 for 'unknown' in the 'apo_e4_allele_clean' attribute. ● Contains 7 NANs
education_years_stages_bin*	● Created grouped education years following the typical educational stages in the US: ● "Less than High School (0-11 years)" ● "High School Graduate (12 years)" ● "Some College (no degree) (13-15 years)" ● "Associate's or Bachelor's Degree (16-18 years)" ● "Graduate or Professional Degree (19+ years)"
longest_loc_duration_bin*	● Merged different groups to achieve a more balanced distribution: ○ Combining "10 sec - 1 min" with "1-2 min" to create "10 sec - 2 min" ○ Combining "6-9 min" with "10 min - 1 hr" to create "6 min - 1 hr" ○ Further combining "10 sec - 2 min" with "3-5 min" to create "10 sec - 5 min"
name	● Updated TBI information for four donors who were initially thought not to have had a TBI (H14.09.072, H14.09.018, H14.09.038, H14.09.034); changed to TBI “Y”. ● Source
sex_clean*	● Converted ”F” and “M” values to 0 and 1. ● Sex is slightly unbalanced, with 63 donors identified as male (M) and 44 donors identified as female (F).

ProteinAndPathologyQuantifications.csv

We began by examining the basic structure of the dataframe, which included a range of information such as donor ID, donor name, structure ID, structure acronym, and various immunohistochemistry (IHC) markers.

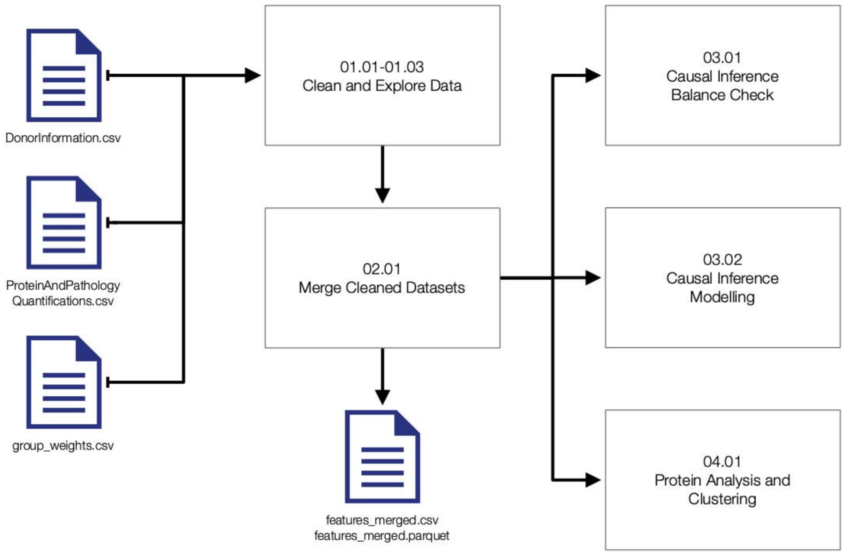
We streamlined the dataset by transforming it from a long format, where each row represented a single scan, to a wide format. Initially, the dataset had 377 rows and 33 columns, with multiple rows per donor. After pivoting, each donor's scans across different brain regions like Frontal White Matter (FWM), Hippocampus (HIP), Posterior Cingulate Cortex (PCx), and Temporal Cortex (TCx) were combined into a single row. This reconfiguration provided a more comprehensive and accessible view of the data for analysis.

To maintain the integrity of the dataset and ensure consistency in subsequent analyses, we replaced these missing values with -1.

4. Methodology

Our workflow for developing the final models and insights, which were used for this project, is depicted in the figure below:

Figure 1: Decode Dementia Workflow


DonorInformation.csv
ProteinAndPathology Quantifications.csv
group_weights.csv
01.01-01.03 Clean and Explore Data
02.01 Merge Cleaned Datasets
features_merged.csv
features_merged.parquet
03.01 Causal Inference Balance Check
03.02 Causal Inference Modelling
04.01 Protein Analysis and Clustering

4.1 Covariance Balance

To ensure the comparability of control and treatment groups within the "DonorInformation.csv" dataset, we conducted covariance balance[5] checks with the following steps:

Filtering for Matched Pairs: We began by filtering the data to isolate matched pairs, which is a critical step for maintaining the relevance and validity of subsequent statistical tests. This filtering process helped to eliminate records with missing or unspecified values that could potentially distort our findings.

Paired T-Tests for Continuous Variables: For continuous variables such as 'age_clean' and 'education_years', we employed paired t-tests. These tests are particularly suited for analyzing matched data and were used to detect any statistically significant differences in means between the control and treatment groups.

Chi-Squared Tests for Categorical Variables: We applied chi-squared tests to categorical variables, including 'sex_clean', 'apo_e4_allele_clean', 'cerad', and 'act_demented_clean'. These tests are designed to evaluate the distribution and association of categorical data across groups.

4.2 Causal Inference

To elucidate the causal relationships between various factors and the outcome (actual dementia) of traumatic brain injury (TBI), we employed a structured approach to causal inference modeling[5]:

Unadjusted Models: Our initial models were unadjusted, serving as a baseline to identify direct associations between TBI and the likelihood of dementia diagnosis. These models provided an initial understanding of the relationship without accounting for additional variables. To visually articulate the hypothesized causal pathways, we constructed a causal graph (Figure 2).

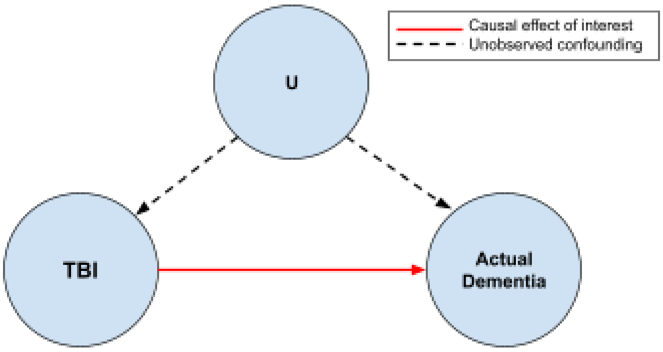
[5] For detailed explanations of specific technical terms please refer to the Appendix C: Glossary of Technical Terms

Adjusted Models: We then advanced to adjusted models, which included confounders such as age, sex, genetic predispositions, and education. These models allowed us to control for these factors and better isolate the effect of TBI on dementia outcomes.

Comprehensive Model Analysis: A comprehensive model was also developed, integrating all variables and group weights. This model was crucial for capturing the multifaceted nature of the factors contributing to TBI outcomes and was evaluated for its goodness of fit. For this final adjusted model we considered all pre-mortality variables (group weights considered and using binned attributes).

Emphasis on Group Weights: The inclusion of group weights in our models was a critical methodological decision that significantly enhanced the validity of our results. It was through the application of these weights that we uncovered statistically significant relationships, which remained undetected in the unweighted models.

Figure 2: Causal graph for unadjusted models



To interpret the coefficients from our logistic regression model in a more accessible form, we utilize a specific approach to translate them into percentage changes. In logistic regression, each coefficient indicates how the log-odds of the outcome variable shift with a one-unit increase in a predictor. To convert these log-odds changes into percentage changes, which are easier to understand, we apply the following formula:

$$Percentage\ Change\ in\ Odds\ =\ (exp(Coefficient)\ -\ 1) \times 100\%$$

4.3 Cluster Analysis

To discern additional patterns within the biomarker[6] data, we continued with a cluster analysis, detailed in the following steps:

Cluster Identification and Visualization in Biomarker Analysis: Using the silhouette score, we determined that three clusters were optimal for grouping the data points based on their biomarker profiles. We then employed Principal Component Analysis (PCA)[7] to visually confirm the separation between these clusters. By reducing the data dimensionality to two principal components, we were able to clearly visualize and validate the distinct segments (Figure 3).

Statistical Analysis (ANOVA and Tukey's HSD Test) on PCA Components: To confirm the distinctiveness of the identified groups, we performed ANOVA and Tukey's HSD tests on the PCA components. These analyses confirmed that the groups were significantly different in terms of their protein profiles and the presence or absence of brain injury or dementia, thereby validating our data grouping method.

[6] For detailed explanations of specific technical terms please refer to the Appendix C: Glossary of Technical Terms

Correlation Heatmaps for Clustered Biomarker Profiles: For each of the three clusters, we created correlation heatmaps to visualize the relationships among key biomarkers. These heatmaps were informed by PCA, which highlighted the top contributing features for each cluster, providing insights into the interplay of biomarkers within each group (Figure 4).

Figure 3: Scatter Plot - PCA Visualization after K-Means Clustering

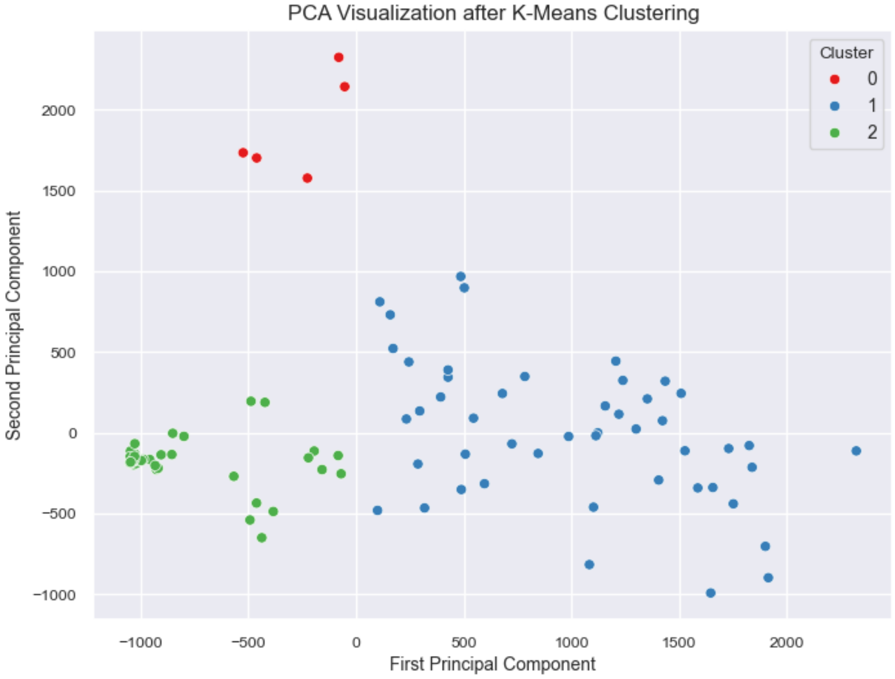
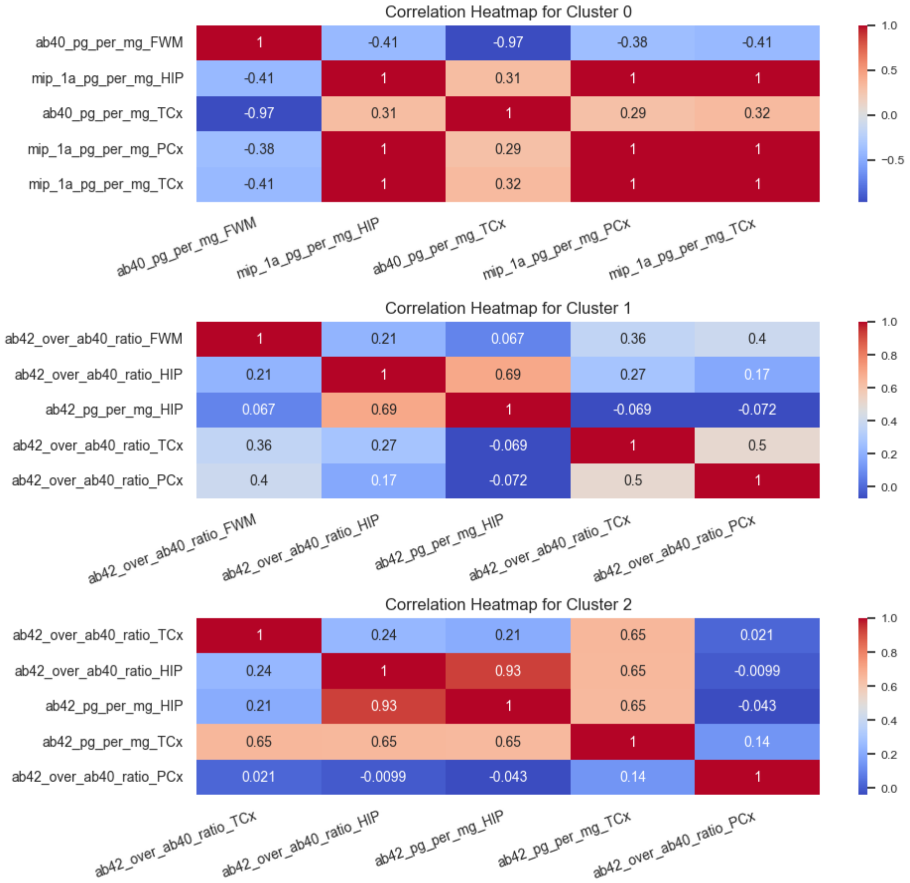
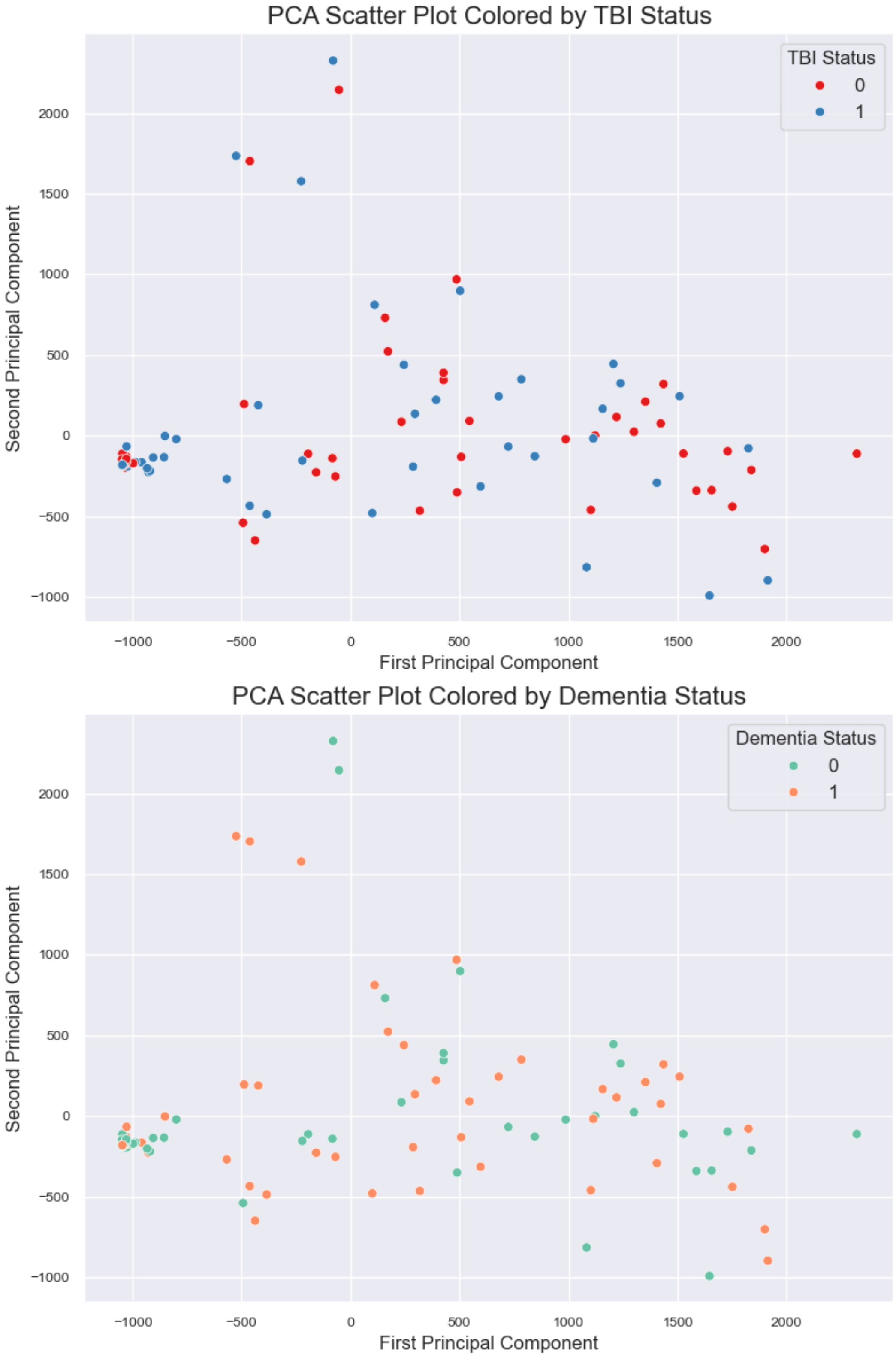


Figure 4: Heatmap - Correlation for each cluster



PCA Scatter Plots Colored by Health Status: We generated PCA scatter plots to illustrate the distribution of data points based on their biomarker profiles and health conditions. One plot displayed the first and second principal components with points colored according to TBI status, while another plot used the same components but with points colored by dementia status. These visualizations helped to elucidate the relationship between biomarker profiles and health outcomes (Figure 5).

Figure 5: Scatter Plot - Colored by TBI Status and Dementia Status



PCA Scatter Plot Colored by TBI Status

PCA Scatter Plot Colored by Dementia Status

5. Results

5.1 Covariance Balance

The balance checks confirmed that there were no significant differences between the groups for all variables tested, indicating that the groups were well-matched in terms of age, sex, APOE ε4 allele status, education years, CERAD scores[7], and dementia status (for further information, please see Appendix E: Covariance Balance Checks).

5.2 Causal Inference

Unadjusted Model (Considering Group Weights): We utilized a Generalized Linear Model (GLM) with a binomial family and Logit link function, suitable for binary outcomes like dementia presence or absence. For the 107 observations, the model showed a Pseudo R-squared of 0.3663. The significant coefficients, including -1.1238 for the constant and 0.5521 for ever_tbi_w_loc_clean, demonstrated that a history of TBI with loss of consciousness significantly increased the odds of being diagnosed with dementia. Both coefficients had p-values less than 0.000, confirming their statistical significance. The results suggested that for every one-unit increase in ever_tbi_w_loc_clean, the odds of being act_demented_clean increased by about 73.70%, holding other variables constant.

Figure 6: Forest Plot - Logistic Regression Coefficients with 95% Confidence Intervals (CI)

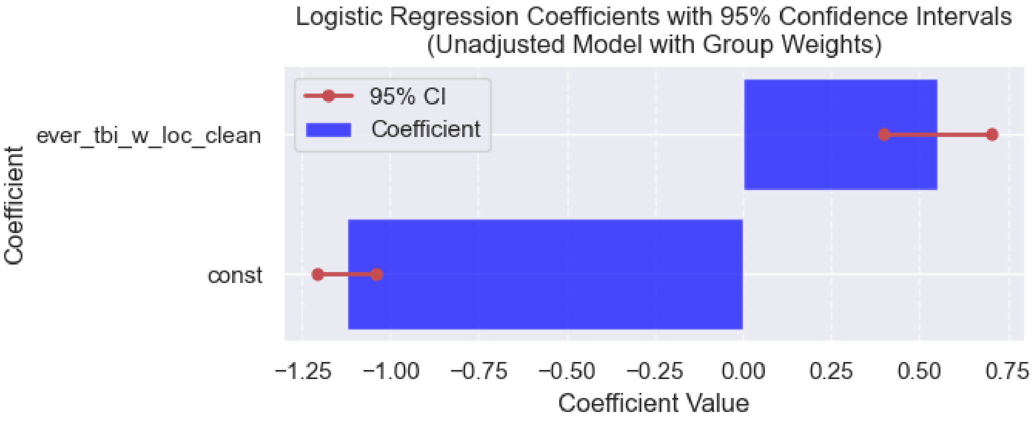


Table 3: Logistic Regression Analysis of the Association Between History of TBI with LOC considering group weights.

Variable	Coeff.[7]	SE7	z-Value[7]	P-Value[7]	95% CI[7]	%
Constant	-1.1238	0.043	-26.299	<0.001	(-1.208, -1.040)	Baseline
ever_tbi_w_loc_clean (Y)	0.5521	0.078	7.065	<0.001	(0.399, 0.705)	73.70%

Coeff.: Coefficient | SE: Standard Error | CI: Confidence Interval

Adjusted model with all pre-mortality variables: In our refined analysis, we considered all available variables **prior to mortality**, excluding 'num_tbi_w_loc' due to its skewed distribution (45 donors reported one TBI incident, while only 9 reported multiple). We also omitted records missing 'apo_e4_allele_clean' data, resulting in a dataset of 100 donors.

Utilizing a Generalized Linear Model (GLM) with a binomial family and a Logit link function, we examined factors influencing dementia (details in Table 4). Each coefficient was statistically significant with p-values below 0.05, affirming the robustness of our model.

[7] For detailed explanations of specific technical terms please refer to the Appendix C: Glossary of Technical Terms

Figure 7: Forest Plot - Logistic Regression Coefficients with 95% Confidence Intervals

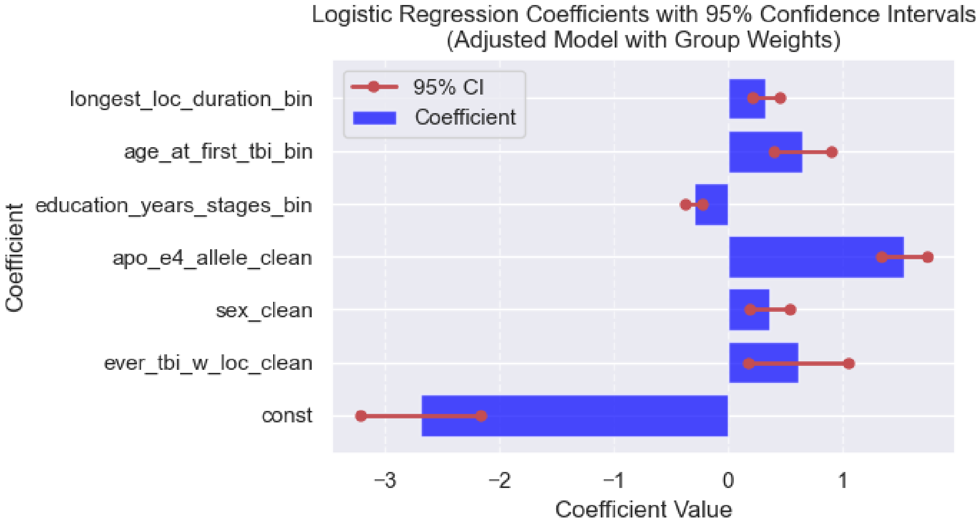


Table 4: Logistic Regression Analysis of the Association Between History of TBI with LOC considering group weights

Variable	Coeff.	SE	z-Value	P-Value	95% CI	%
Constant	-2.6894	0.267	-10.077	<0.001	(-3.212, -2.166)	Baseline
ever_tbi_w_loc_clean (Y)	0.6136	0.221	2.776	0.006	(0.180, 1.047)	+84.71%
Sex_clean (M)	0.3655	0.088	4.140	<0.001	(0.192, 0.539)	+44.13%
apo_e4_allele_clean	1.5415	0.103	15.039	<0.001	(1.341, 1.742)	+367.18%
education_years_stages _bin	-0.3041	0.038	-8.029	<0.001	(-0.378, -0.230)	-26.22%
age_at_first_tbi_bin	0.6466	0.128	5.055	<0.001	(0.396, 0.897)	+90.90%
longest_loc_duration_bin	0.3323	0.062	5.353	<0.001	(0.211, 0.454)	+39.42%

However, when analyzing the GLM model, we noted that it reported a Pseudo R-squared of 0.9828. This value is unusually high for logistic regression models, especially given our modest sample size of 100 observations. A Pseudo R-squared of this magnitude often suggests overfitting. Overfitting occurs when a model captures noise as if it were a significant pattern, leading to poor performance on new, unseen data.

To address this concern and to validate our insights, we took the following steps to run a second model:

Data Preprocessing: We expanded the dataset by duplicating records according to their respective rounded 'group_weight' to make it more reflective of the real-world distribution and ensuring a more robust analysis resulting in 3969 records, compared to the 107 from the original dataset.

Modeling with Logit: We applied a logistic regression model using sm.Logit from the statsmodels library. This method allowed us to calculate the coefficients and p-values for each predictor, providing us with a statistical foundation to gauge their impact on our outcome variable, 'act_demented_clean'.

Comparative Analysis with GLM: We compared the results from sm.Logit with those obtained from a Generalized Linear Model (GLM) using the same library (Table 5).

Table 5: Comparing Logit Vs GLM Coefficients

Variable	Logit Coeff.	Logit %	GLM Coeff.	GLM %
Constant	-2.6530	Baseline log odds	-2.6894	Baseline
History of TBI with LOC	0.6046	+83.06%	0.6136	+84.71%
Gender (Male)	0.3683	+44.52%	0.3655	+44.13%
Presence of APOE ε4 Allele	1.5550	+373.50%	1.5415	+367.18%
Education Years (Binned)	-0.3103	-26.68%	-0.3041	-26.22%
Age at First TBI (Binned)	0.6430	+90.21%	0.6466	+90.90%
Longest Duration of LOC (Binned)	0.3233	+38.17%	0.3323	+39.42%

The insights derived from the Generalized Linear Model (GLM) have been substantiated and reinforced by the findings from the less overfitted Logit model (see Table 5). This confirmation adds robustness to our understanding and interpretation of the data, enhancing the reliability of the conclusions drawn from the GLM.

5.3 Cluster Analysis

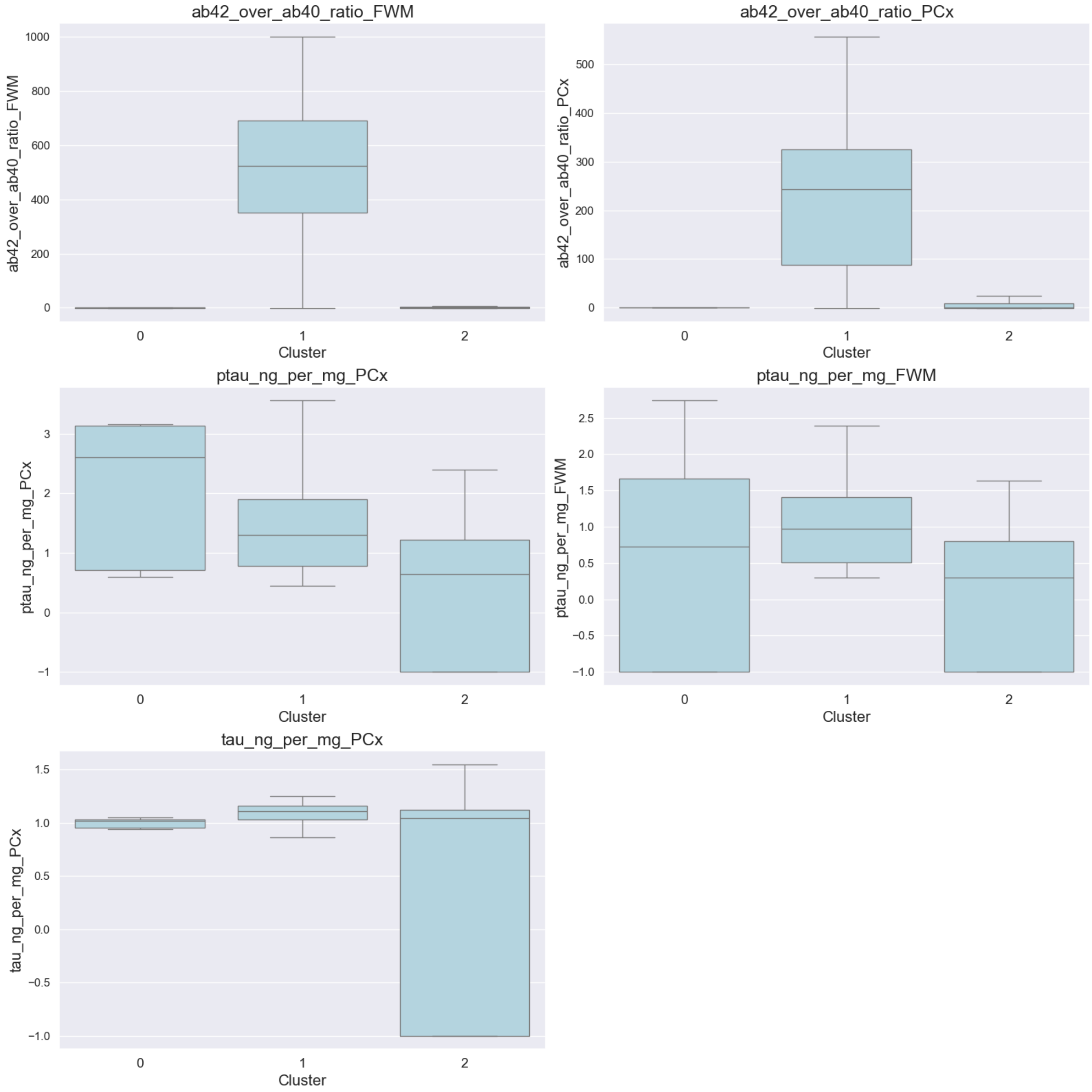
ANOVA Analysis of Biomarker Clusters: We identified the top 5 most significant findings and illustrated them through box plots, effectively demonstrating how biomarkers exhibit varying patterns across the identified clusters (Figure 8). Our analysis revealed meaningful disparities in factors such as proteins, indicating distinctions that are unlikely to occur by chance. These differences likely signify distinct biological or pathological characteristics, substantiated by rigorous statistical tests like ANOVA and their corresponding p-values.

All five biomarkers exhibit significant differences across the clusters, as indicated by high F-values and extremely low p-values:

Table 6: Biomarkers with F- and P-values

Biomarker	F-value	P-Value	Description
ab42_over_ab40_ratio_FWM	81.99	p < 1x10^-22	Aβ42/Aβ40 ratio in Frontal White Matter, indicating Alzheimer's disease risk.
ab42_over_ab40_ratio_PCx	24.02	p < 1x10^-9	Aβ42/Aβ40 ratio in Posterior Cortex, a marker for amyloid plaque formation.
ptau_ng_per_mg_PCx	20.12	p < 1x10^-8	Elevated phosphorylated tau in Posterior Cortex, associated with Alzheimer's.
ptau_ng_per_mg_FWM	15.45	p < 1x10^-6	Phosphorylated tau levels in Frontal White Matter, indicating neurodegeneration.
tau_ng_per_mg_PCx	15.17	p < 1x10^-6	Total tau protein in Posterior Cortex, a marker for neuronal damage.

Figure 8: Box Plots for the top five proteins exhibit varying patterns across the identified clusters



6. Conclusion & Discussion

In our study, we established a substantial link between traumatic brain injuries (TBI) and an elevated risk of dementia. Key findings indicate that TBIs, especially those involving loss of consciousness, alongside the presence of the APOE ε4 allele and male gender, significantly increase dementia risk, while higher education levels seem to offer some protective effects.

Utilizing ANOVA, we identified five critical brain biomarkers related to protein patterns, with particular emphasis on the Aβ42/Aβ40 ratio and tau protein levels. These findings underscore the complex impact of TBI on cognitive health, supported by strong statistical evidence.

While our study provides extensive insights, it is not without limitations. The potential presence of unseen confounding variables and the reliance on retrospective data suggest a need for further research. Future studies should aim to include longitudinal analyses to better understand the connections between TBI and dementia and to explore the underlying mechanisms.

The implications of our findings could be far-reaching for public health policies, clinical practices, and individual decision-making. They emphasize the need for improved TBI prevention strategies, heightened awareness of its long-term risks, and the adoption of personalized medical approaches that take genetic and demographic factors into account. Additionally, our study calls for the establishment of ethical guidelines in research and application to prevent discrimination.

In conclusion, this study not only contributes to the growing body of knowledge on TBI and dementia but also opens new avenues for research and policy-making that can profoundly impact public health and individual well-being.

7. Ethical Considerations

Our analysis recognized that ethical considerations were intrinsic to this project, particularly given its intersection with sensitive health data and the dynamics of the health insurance industry. The collection and use of data, especially biomarkers associated with Traumatic Brain Injury (TBI), carried inherent risks. These included potential privacy breaches, data misuse, and the ethical dilemma of balancing patient confidentiality with data sharing for research and treatment optimization. Additionally, there was a risk of discrimination or stigmatization of individuals with a history of TBI in the context of health insurance coverage and costs.

To mitigate these concerns, we applied the following guiding principles as the project progressed:

- **Consent and Anonymity**: We checked that the data from these Aging, Dementia and Traumatic Brain Injury Study were collected with consent. This was confirmed in the Allen Institute's 'Technical White Paper: ACT Cohort' (2016).
- **Ethical Use of Donor Data**: We were mindful of the implications of our findings on the donors' privacy and continuing to maintain the de-identified property of the data**.**
- **Transparency and Accountability**: We openly disclosed our methodologies for data cleaning, transformation, and modeling to ensure complete transparency regarding the utilization of data.
- **Public Benefit**: We published this project to share our insights and to contribute to the public good, particularly in improving understanding and treatment of TBI.

8. References

- Allen Institute for Brain Science. (2017). Technical White Paper: Overview of the Aging, Dementia and Traumatic Brain Injury (TBI) Project. http://aging.brain-map.org/
- Allen Institute. (2016). Technical White Paper: Weighted Analyses. Retrieved from https://help.brain-map.org/download/attachments/9895983/Weighted_Analyses.pdf?version=1&modificationDate=1456179403835&api=v2
- Allen Institute. (2016). Technical White Paper: ACT Cohort. Retrieved from https://help.brain-map.org/download/attachments/9895983/ACT_Cohort.pdf?version=2&modificationDate=1492728684163&api=v2
- Moura, D. A. P., & Oliveira, J. R. M. de. (2021). What Do Machines Tell Us About Dementia? Machine Learning Applied to Aging, Dementia, and Traumatic Brain Injury Study. [Preprint]. https://doi.org/10.21203/rs.3.rs-840907/v1
- Pölsterl, S., Wachinger, C., the Alzheimer's Disease Neuroimaging Initiative, & the Japanese Alzheimer's Disease Neuroimaging Initiative. (2022). Identification of causal effects of neuroanatomy on cognitive decline requires modeling unobserved confounders. Alzheimer's & Dementia. https://doi.org/10.1002/alz.12825
- Ranson, J. M., Bucholc, M., Lyall, D., Newby, D., Winchester, L., Oxtoby, N. P., Veldsman, M., Rittman, T., Marzi, S., Skene, N., Al Khleifat, A., Foote, I. F., Orgeta, V., Kormilitzin, A., Lourida, I., & Llewellyn, D. J. (2023). Harnessing the potential of machine learning and artificial intelligence for dementia research. Brain Informatics, 10, Article 6. https://doi.org/10.1186/s40708-022-00183-3

9. Statement of Work

Table 7: Statement of Work

Activities	Lead	Support
Development Environments and GitHub Repository	Andre	Victor
Related Work	Victor	Andre
Data Description	Jointly	
Data Cleaning	Andre	Victor
Causal Inference	Andre	Victor
Clustering	Victor	Andre
Conclusion & Discussion	Jointly	
Ethical Consideration	Jointly	
Project Report	Jointly	
Exhibition Poster	Victor	Andre

Appendices

Appendix A: Extended Data Description

Our analysis incorporated a subset of the dataset from the Aging, Dementia, and Traumatic Brain Injury Study, a collaborative effort spearheaded by the University of Washington, Kaiser Permanente Washington Health Research Institute, and the Allen Institute for Brain Science."

The dataset originates from a unique aged cohort from the Adult Changes in Thought (ACT) study, a longitudinal investigation into brain aging and dementia within the Seattle metropolitan area. The ACT study is managed by the Kaiser Permanente Washington Health Research Institute and has established protocols for sharing its data with external researchers.

We focused on three key data files from the Aging, Dementia, and Traumatic Brain Injury Study (Allen Institute for Brain Science, 2017):

Table: Extended Data Description

Description and Importance	**Rec.**[8]	**Attr.**[9]
DonorInformation.csv contains detailed information about individual donors, including various age-related characteristics. This file allowed us to understand the baseline characteristics of our study population, providing a foundation for any associations or patterns we might observe in relation to TBI and dementia.	107	19
ProteinAndPathologyQuantifications.csv offers quantified measurements related to proteins and pathologies associated with brain aging or neurodegenerative disorders. The data from this file were instrumental in correlating molecular and pathological changes with clinical outcomes, helping us to uncover potential biomarkers or pathological processes associated with dementia post-TBI.	377	33
group_weights.csv contains subject and sampling weights as calculated and described in the Allen Institute's 'Technical White Paper: Weighted Analyses' (2016). These weights are crucial for adjusting our analyses to be representative of the larger ACT cohort, thereby enhancing the validity of our findings. The Allen Institute (2016) provides detailed methodologies for the calculation of subject and sampling weights.	107	2

[8] Records

[9] Attributes

Appendix B: Data Cleaning and Exploration: DonorInformation.csv

Table: Data Cleaning and Exploration - DonorInformation.csv

Attribute/Feature	**Insights and transformations**
act_demented	● "Actually demetend" is nearly evenly distributed between "No dementia" (57) and "Dementia" (50).
act_demented_clean*	● Mapped the 'No Dementia' and 'Dementia' values to 0 and 1, respectively, in the 'act_demented_clean' attribute.
age	● Most donors fall within the age range of 90-94. ● Age ranges like '90-94' and '95-99' suggest that exact ages within these intervals were not recorded. ● The '100+' category indicates donors who are 100 years old or older, but the exact age is not specified.
age_at_first_tbi	● A large number of donors (54) have an age value of 0 for their first TBI. ● Based on the study documentation, we discovered that excluding 0 was necessary to ensure an accurate distribution, as 0 indicated the absence of TBI. This observation also aligned with the ever_tbi_w_loc column. ● The data spans a wide range of ages, from early childhood to advanced age.
age_at_first_tbi_bin*	● A significant number of TBI incidents occur during early stages of life, as the majority of donors encounter their initial TBI before reaching the age of 30. Another notable concentration can be observed from the age of 60 onwards. ● To reflect this observation, we established three categories: early, mid, late: "early_years (1-30)", "mid_years (31-60)", "late_years (61-90)".
age_bin*	● Binned the age information in order to obtain a series that is more evenly distributed, while also ensuring the preservation of the current groups: "90-94", "95-99", "87-89", "81-86", "77-79", "100+"
age_clean*	● Created a new attribute to retain the quantitative data of the age using the median.
apo_e4_allele	● Contains 7 NANs ● Replaced NAN with "Unknown"
apo_e4_allele_clean*	● Mapped with 0 for 'N', 1 for 'Y', and -1 for 'unknown' in the 'apo_e4_allele_clean' attribute.
control_set	● Control set no 53 (H15.09.106) is the only one with an incomparable individual who had a TBI.
education_years	● The most common education duration is 12 years, which typically corresponds to the completion of high school in many education systems. ● The mean years of education is slightly over 14 years, indicating that, on average, the donors have some education beyond high school. ● The wide range, from 6 to 21 years, shows a diverse group of donors in terms of educational background.
education_years_quartiles_bin*	● Created quartile-based bins (creating 4 equally sized bins) for a better distribution of the feature.

Attribute/Feature	Insights and transformations
education_years_stages_bin*	● Created grouped education years following the typical educational stages in the US: ● "Less than High School (0-11 years)" ● "High School Graduate (12 years)" ● "Some College (no degree) (13-15 years)" ● "Associate's or Bachelor's Degree (16-18 years)" ● "Graduate or Professional Degree (19+ years)"
ever_tbi_w_loc	● “Ever TBI” with loss of consciousness (LOC) is nearly evenly distributed between yes (57) and no (50).
ever_tbi_w_loc_clean	● Mapped the values 'N' and 'Y' to 0 and 1.
longest_loc_duration	● A significant number of donors (61) have "Unknown or N/A" as their longest LOC duration, which could be due to either not having experienced a LOC or missing data. ● Shorter LOC durations, especially less than 10 seconds, are relatively common with 19 entries. ● There are also donors who have experienced longer durations of LOC, ranging from 10 minutes to more than an hour.
longest_loc_duration_bin*	● Merged different groups to achieve a more balanced distribution: ○ Combining "10 sec - 1 min" with "1-2 min" to create "10 sec - 2 min" ○ Combining "6-9 min" with "10 min - 1 hr" to create "6 min - 1 hr" ○ Further combining "10 sec - 2 min" with "3-5 min" to create "10 sec - 5 min"
longest_loc_duration_clean*	● 7 records where a TBI event was recorded, however, the longest_loc_duration is "Unknown or N/A". For these records, we set the longest duration to -1 ● 50 records with no TBI, for those we set the longest_loc_duration to 0 ● Created "longest_loc_duration_clean" to retain the quantitative data of the longest_loc_duration. ● Convert all time ranges to seconds and use the median.
name	● Updated TBI information for four donors who were initially thought not to have had a TBI (H14.09.072, H14.09.018, H14.09.038, H14.09.034); changed to TBI “Y”. ● Source
sex	● Sex is slightly unbalanced, with 63 donors identified as male (M) and 44 donors identified as female (F).
sex_clean*	● Converted ”F” and “M” values to 0 and 1.

Appendix C: Glossary of Technical Terms

Table: Technical Terms and Definitions

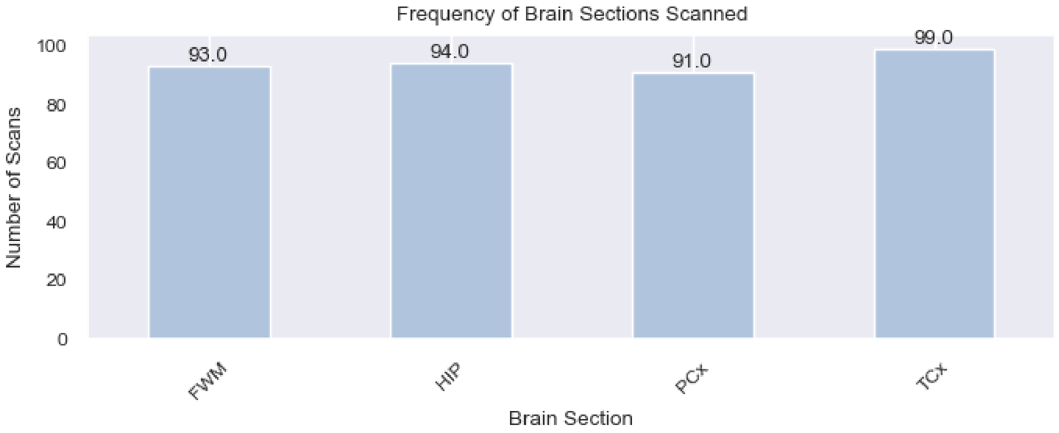
Term	Definition
Aβ42/Aβ40	The Aβ42/Aβ40 ratio refers to the proportion of two forms of amyloid beta peptides, Aβ42 (42-amino acid peptide) and Aβ40 (40-amino acid peptide), where Aβ42 is more prone to aggregation and linked to Alzheimer's disease, and a higher ratio is associated with increased Alzheimer's risk.
APOE ε4 allele	The APOE ε4 allele is a specific version of a gene called APOE (apolipoprotein E) that is associated with an increased risk of developing Alzheimer's disease, particularly the late-onset form, which is the most common type of Alzheimer's.
Biomarkers	Biomarkers are measurable indicators of a biological state or condition. In medical research, they are used to detect or monitor diseases. For dementia, biomarkers in blood or brain scans can show the presence or progression of the disease.
Causal Inference Modeling	This statistical approach helps determine if one thing causes another. In dementia research, it's used to understand whether certain factors (like genetics or lifestyle) are likely causes of the disease.
CERAD score	The CERAD (Consortium to Establish a Registry for Alzheimer's Disease) score is a tool used in medical research and practice to evaluate the severity of Alzheimer's disease. It is designed to assess the extent and distribution of neurofibrillary tangles and neuritic plaques, which are hallmark features of Alzheimer's disease in the brain. Neurofibrillary Tangles and Neuritic Plaques: In Alzheimer's disease, certain proteins in the brain form abnormal structures called neurofibrillary tangles and neuritic plaques. These are believed to contribute to the death of brain cells and the symptoms of Alzheimer's.
Coefficient (Coeff.)	In statistics and mathematics, a coefficient is a number that multiplies a variable in an equation. For example, in a linear regression model, coefficients represent the degree of change in the dependent variable for one unit of change in an independent variable.
Confidence Interval (CI)	A Confidence Interval is a range of values, derived from sample statistics, that is likely to contain the value of an unknown population parameter. It gives an estimated range of values which is likely to include an unknown parameter, calculated from a given set of sample data.
Covariance Balance	In statistics, covariance balance is a method to ensure variables in different groups (like treatment and control groups in research) are comparable. This helps researchers draw more accurate conclusions about the effects of a treatment or condition.
F-value	The "F-value" in an ANOVA table quantifies the statistical significance of group differences for a given variable, with higher values indicating greater differences.
Frontal White Matter (FWM)	This refers to the white matter in the frontal lobe of the brain, which contains nerve fibers. These fibers are crucial for communication between different

Term	Definition
	brain regions. Damage or changes in the FWM can affect cognitive functions and behavior.
Hippocampus (HIP)	The hippocampus is a small, curved region in the brain involved in memory formation and navigation. It is one of the first regions affected in Alzheimer's disease, leading to memory loss.
Neurodegenerative Disorders	These are diseases where nerve cells (neurons) in the brain gradually deteriorate or die. This leads to problems with movement or mental functioning. Alzheimer's disease is a well-known example, characterized by memory loss and cognitive decline.
p-value	The p-value is a measure used in statistical hypothesis testing. It tells you the probability of obtaining test results at least as extreme as the ones observed, under the assumption that the null hypothesis is true. A low p-value (typically <0.05) suggests that the observed data are unlikely under the null hypothesis, leading to its rejection.
Pathologies	Pathology is the scientific study of diseases. It involves examining the causes, development, and effects of diseases. In the context of dementia, it refers to the study of brain changes and abnormalities caused by the disease.
Posterior Cingulate Cortex (PCx)	This part of the brain is involved in memory and emotional processing. Changes in the PCx are often associated with various neurological conditions, including dementia.
Principal Component Analysis (PCA)	PCA is a statistical method that simplifies complex data sets by reducing their dimensions. It helps in identifying patterns and making data easier to explore and visualize, especially in large datasets like those used in AI and ML for medical research.
Proteins	Proteins are large, complex molecules vital to the body's functioning. They perform various roles, like supporting immune responses and enabling movement. In dementia research, specific proteins can indicate brain changes related to the disease.
Standard Error (SE)	The Standard Error measures the accuracy with which a sample represents a population. In simpler terms, it tells us how much the sample mean (average) is likely to vary from the true population mean. A smaller SE indicates more precise estimates.
Temporal Cortex (TCx)	The temporal cortex is part of the cerebral cortex, involved in processing sensory input and is important for understanding language, recognizing faces, and memory formation.
z-value	A z-value, or z-score, is a statistical measurement that describes a value's relationship to the mean of a group of values, measured in terms of standard deviations from the mean. It's used in standard normal distribution to determine the probability of a value occurring within a normal range and to compare scores from different samples.

Appendix D: Brain Scan Distributions

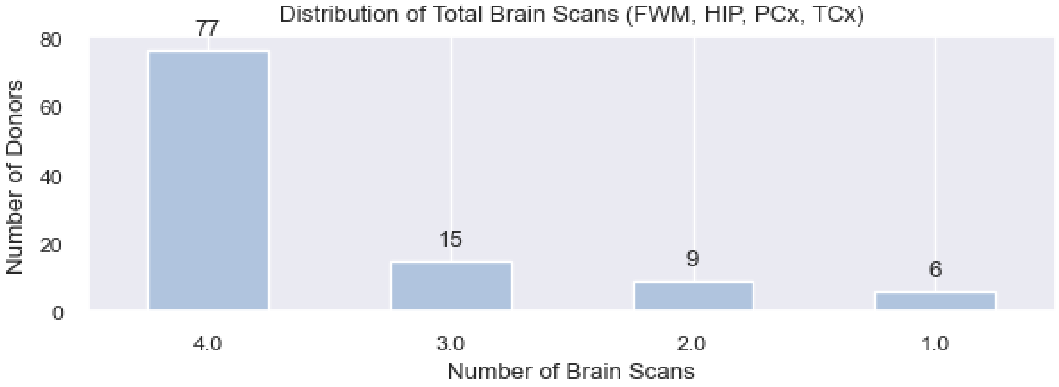
We created tables to compare which brain regions were scanned for each donor, using the available regions Frontal White Matter (FWM), Hippocampus (HIP), Posterior Cingulate Cortex (PCx), and Temporal Cortex (TCx). The tables helped us visualize the variability in scans across these regions for each donor, revealing inconsistencies in scanning all four regions, which suggests a potential gap in the dataset's coverage (Figure 1).

Figure: Bar Plot - Frequency of Brain Sections Scanned



This discrepancy becomes particularly significant when we consider the importance of analyzing and comparing scans across different donors. Nevertheless, it's worth mentioning that for 77 out of 107 donors, all four brain sections were scanned. The Temporal Cortex (TCx) emerged as the most frequently scanned region, with a total of 99 scans, suggesting a heightened emphasis or prevalence in data acquisition within this specific area. The Hippocampus (HIP) closely followed with 94 scans, indicating its substantial representation in the dataset (Figure 2).

Figure: Bar Plot - Distribution of Total Brain Scans

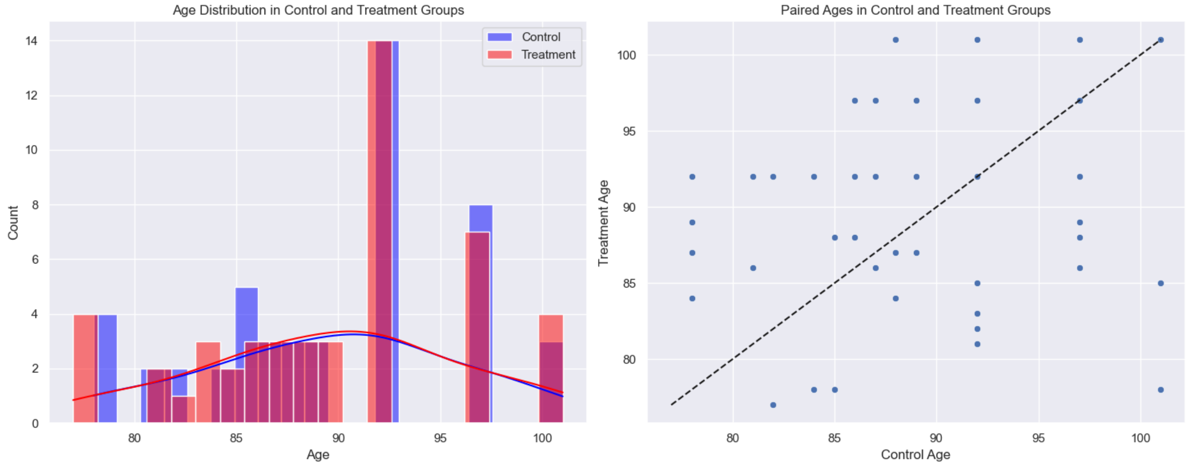


Appendix E: Covariance Balance Checks

Age

The paired t-test for the continuous variable age_clean yields a p-value of approximately 0.945. This high p-value suggests that there is no statistically significant difference in the age_clean variable between the treatment and control groups within the paired data, indicating that the covariate is balanced with respect to age.

Figure: Paired Ages in Control and Treatment Groups



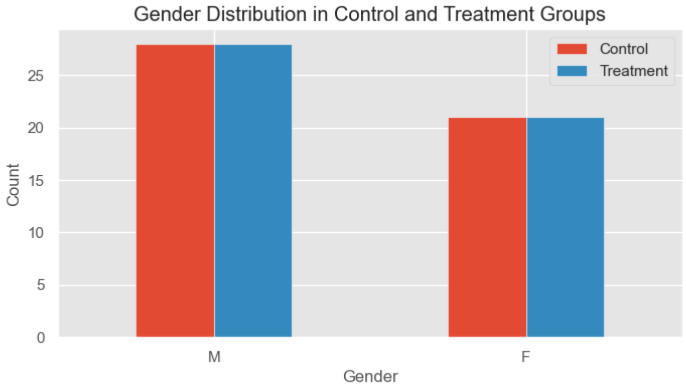
Histograms (Left Plot): This plot shows the age distributions of the control (blue) and treatment (red) groups. The overlapping areas indicate where the age ranges of the two groups intersect. This visualization helps to understand the distribution and range of ages in both groups.

Scatter Plot (Right Plot): Each point represents a pair of individuals from the control and treatment groups. The closer the points are to the diagonal dashed line, the more similar the ages are within each pair. This plot illustrates the direct age comparisons within each pair.

Gender (sex)

These results indicate that there is no significant difference in the distribution of the 'sex_clean' variable between the control and treatment groups. The chi-squared statistic of \(0.0\) and the p-value of \(1.0\) suggest that the observed frequencies in your data perfectly match the expected frequencies under the null hypothesis of no association between group (control or treatment) and sex.

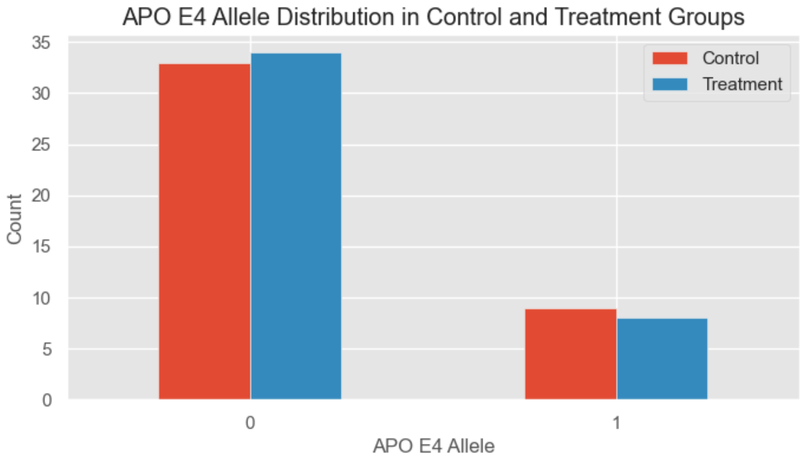
Figure: Gender Distribution in Control and Treatment Groups



APO E4 Allele

The Chi-squared test results for the 'apo_e4_allele_clean' variable between control and treatment groups yielded a chi-squared statistic of (0.0) and a p-value of (1.0), indicating no significant difference between the observed and expected frequencies. The degrees of freedom for the test were (1), reflecting the number of categories in the variable minus one. The test's expected frequencies, an identical distribution of 'apo_e4_allele_clean' across both groups, matched perfectly with the observed data. These results imply that there is no statistically significant association between the group type (control or treatment) and the distribution of the 'apo_e4_allele_clean' allele. Consequently, the factor differentiating the groups does not appear to influence the distribution of this allele within the context of this study.

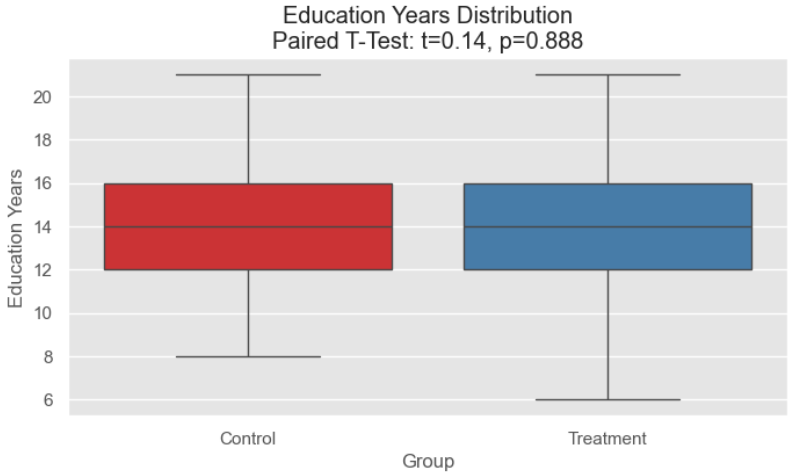
Figure: APO E4 Allele Distribution in Control and Treatment Groups



Education Years

The paired t-test conducted to compare the 'education_years' variable between control and treatment groups yielded a test statistic of (0.14154) and a p-value of (0.88804). These results indicate a minimal difference between the mean years of education in both groups. The high p-value, significantly exceeding the standard alpha level of 0.05, suggests that the observed difference is not statistically significant. With 48 degrees of freedom, the test provides a reliable analysis of the data. In conclusion, the factor differentiating the groups does not appear to have a significant influence on the years of education within the context of this study.

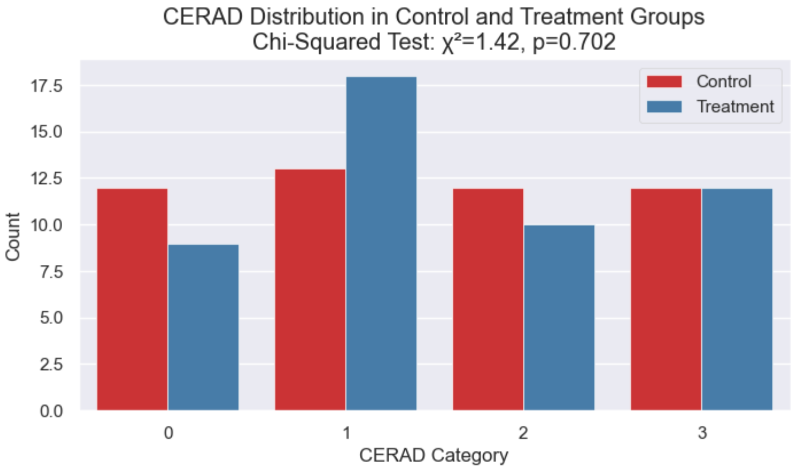
Figure: Education Years Distribution Paired T-Test



CERAD

The Chi-squared test for comparing the 'cerad' variable between control and treatment groups resulted in a chi-squared statistic of (1.41684) and a p-value of (0.70159). These results indicate a low degree of difference between the observed and expected frequencies under the null hypothesis. The test's degrees of freedom, (3), are determined by the number of categories in the 'cerad' variable minus one. The expected frequencies array shows a uniform distribution across both groups, suggesting no significant association between group type and 'cerad'. In summary, the factor differentiating the control and treatment groups does not significantly affect the distribution of the 'cerad' variable within this study.

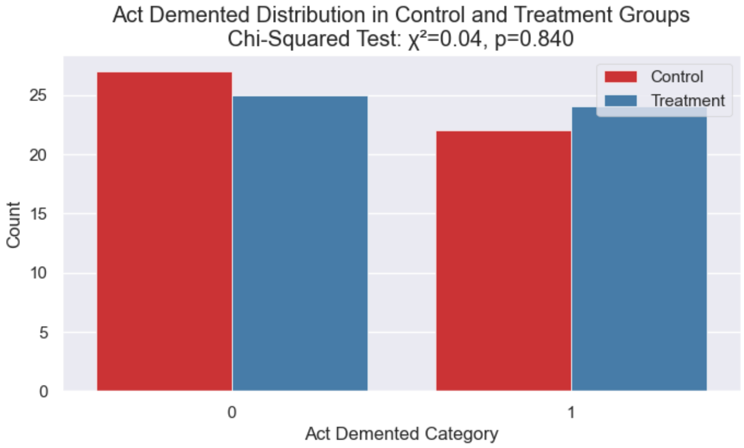
Figure: CERAD Distribution in Control and Treatment Groups Chi-Squared Test



Act Demented

The Chi-squared test conducted on the 'act_demented_clean' variable resulted in a chi-squared statistic of approximately (0.041). The associated p-value is (0.840), indicating that there is no significant difference in the distribution of 'act_demented_clean' between the control and treatment groups. The degrees of freedom for the test were (1), reflecting the categorical nature of the variable. The expected frequencies, nearly equal in both groups, match the observed frequencies closely. In summary, these results suggest no statistically significant association between the group type (control or treatment) and the 'act_demented_clean' variable, affirming a balanced distribution between the groups for this variable.

Figure: Act Demented Distribution in Control and Treatment Groups\nChi-Squared Test

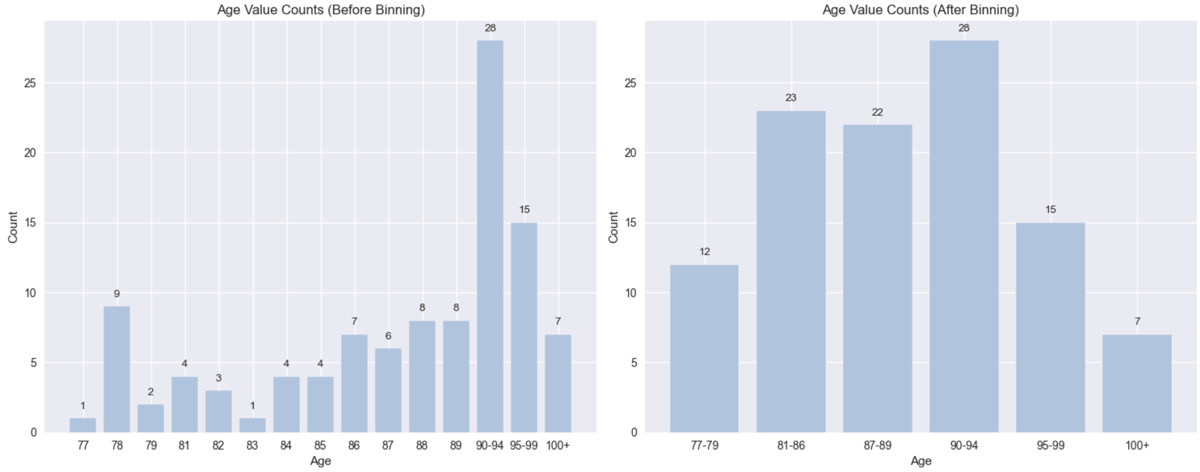


Appendix F: EDA - Feature Distributions

Age Binning

We strive to categorize the age information in order to obtain a dataset that is evenly distributed, while also ensuring the preservation of the current groups: 90-94, 95-99, 100+.

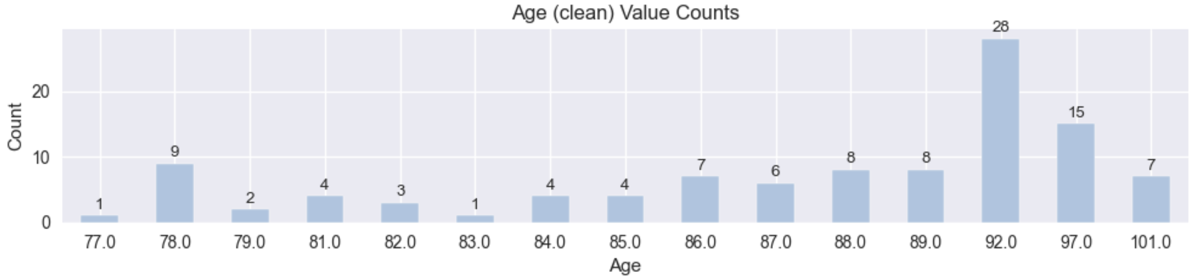
Figure: Age Binning



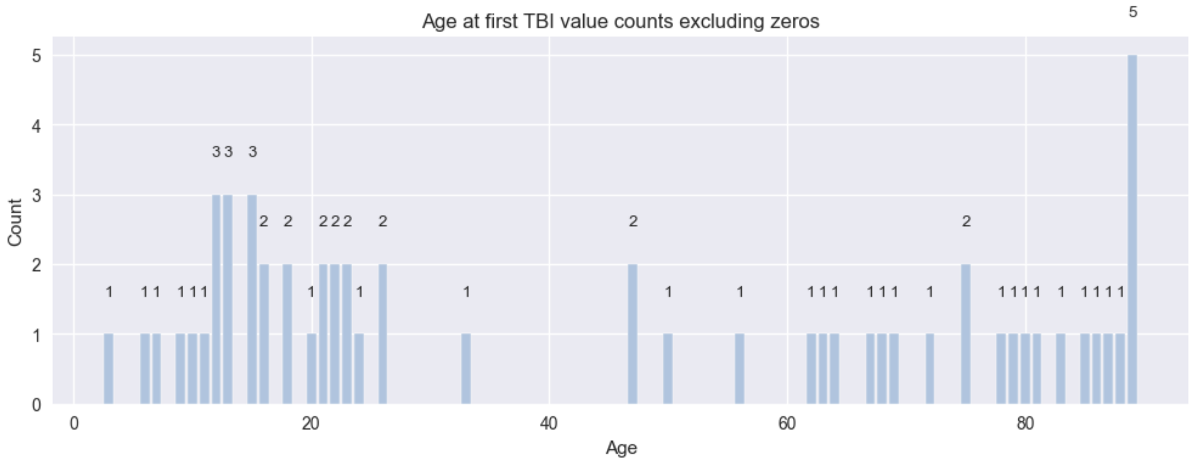
Age Clean

We are also generating a new attribute called "age_clean" to retain the quantitative data of the age. We will use the median.

Figure: Age Clean



Age at first TBI at First TBI

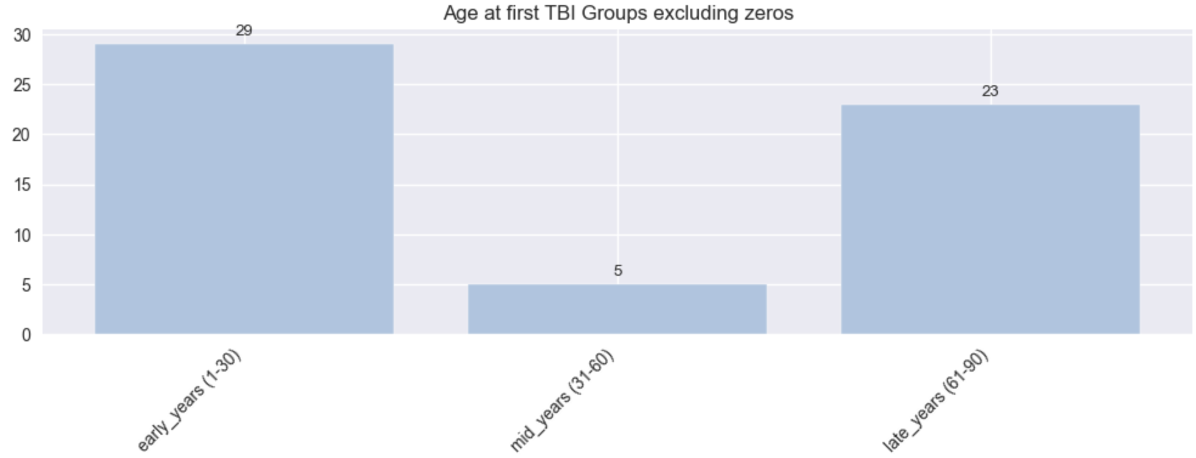


A significant number of TBI incidents occur during early stages of life, as the majority of donors encounter their initial TBI before reaching the age of 30. Another notable concentration can be observed from the age of 60 onwards.

To reflect this observation, we will establish three categories: early, mid, late:

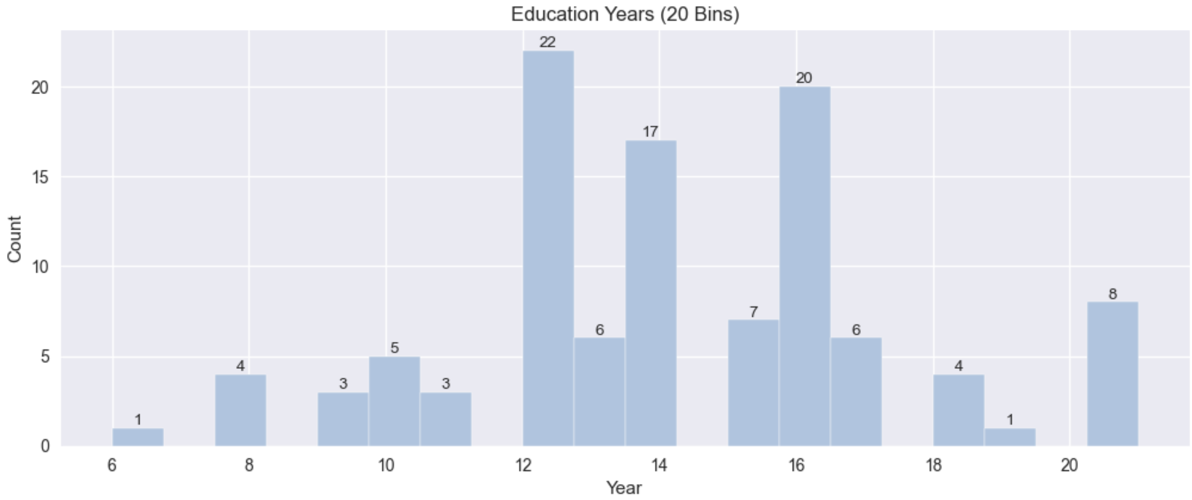
- 1 <= age <= 30: "early_years (1-30)"
- 31 <= age <= 60: "mid_years (31-60)"
- 61 <= age <= 90:l "late_years (61-90)"
- Otherwise "NA"

Figure: Age at first TBI Groups excluding zeros



Education Years

Figure: Education Years (20 Bins)

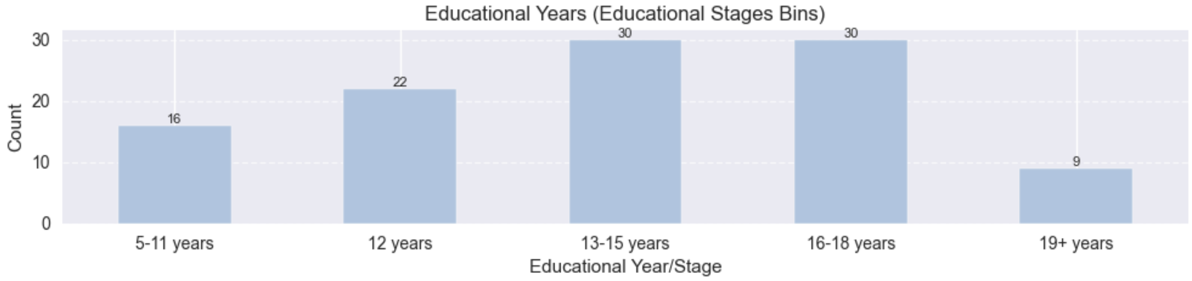


- The most common education duration is 12 years, which typically corresponds to the completion of high school in many education systems.
- The mean years of education is slightly over 14 years, indicating that, on average, the donors have some education beyond high school.
- The wide range, from 6 to 21 years, shows a diverse group of donors in terms of educational background.

We are creating an education years grouping which follows the typical educational stages in the US:

- 'Less than High School (0-11 years)'
- 'High School Graduate (12 years)'
- 'Some College (no degree) (13-15 years)'
- 'Associate's or Bachelor's Degree (16-18 years)'
- 'Graduate or Professional Degree (19+ years)'

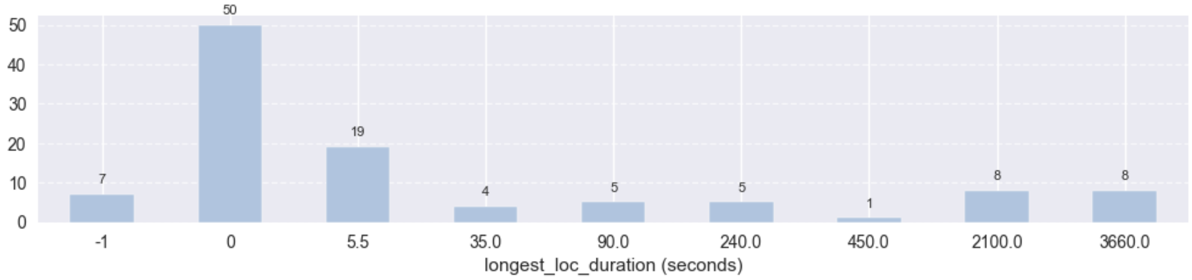
Figure: Educational Years (Educational Stages Bins)



Longest Loc Duration

We are also generating a new attribute called "longest_loc_duration_clean" to retain the quantitative data of the longest_loc_duration. We convert all time ranges to seconds and use the median. Missing values are replaced with “-1”.

Figure: longest_loc_duration (seconds)


longest_loc_duration (seconds)