Predicting a Neurotoxic protein based on structural magnetic resonance imaging (MRI): an investigation of Alzheimer's Disease

Hunsica J Jayaprakash[1,2], Joseph Yurko[2,3], Helmet T Karim[3], Akiko Mizuno[3], William E Klunk[4], Howard J Aizenstein[4,5]

[1]Department of Computer Science, University of Pittsburgh

[2]Center for the Neural Basis of Cognition, Carnegie Mellon University, University of Pittsburgh

[3]Department of Psychiatry, University of Pittsburgh

[4]Department of Bioengineering, University of Pittsburgh

Introduction

Computational neuroscience is interdisciplinary field that combines computer science, mathematics and neuroscience to study and analyze questions in the field of neuroscience. The research project I conducted with my mentor is an investigation of Alzheimer's disease where I used brain imaging (MRI) and an unsupervised learning technique - principal component analysis (PCA) to reduce the dimensionality of my input data while still maintaining variable interactions. Performing PCA enhances our understanding of variable interactions and their behavior. This information will be useful in finding a suitable predictive model to estimate the neurotoxic protein, beta amyloid (Aβ).

In this paper, I will give a brief background on Alzheimer's disease, beta-amyloid , and modeling preclinical-AD progression. Next, I will discuss the implementation of my research project which includes data collection and reshaping, performing PCA, and clustering and plotting the results. Lastly, I will discuss the limitations of my project, reflections during my research experience, and probable future directions.

Background

Alzheimer's Disease

Alzheimer's Disease (AD) is currently the sixth leading cause of death in the United States which has increased in prevalence due to an aging population. AD is a progressive neurodegenerative disorder of the brain's structure and function. While memory loss is the most predominant symptom, AD also affects other cognitive domains, including language, visuomotor processing, attention, and executive function. This is because AD is a neurodegenerative disorder, which is a slow progression of structural and functional decline of the neural system. This process starts decades prior to AD diagnosis and involves beta-amyloid (Aβ) deposition.

Beta-Amyloid (Aβ)

Aβ is a neurotoxic protein which is found in healthy individuals. However, it is known to be a hallmark of AD pathology at abnormal levels. Aβ is a regional factor that leads to neuronal dysfunction when varying amounts aggregate in different regions. Aβ accumulation, as measured by positron emission tomography (PET) with tracers such as Pittsburgh compound B (PiB) or 18F- florbetapir/ florbetaben, usually occurs primarily before the onset of cognitive deficits. This early stage of significant amyloid pathology without overt cognitive dysfunction that can potentially lead to a later diagnosis of AD is known as pre-clinical AD.

Preclinical – AD modeling

Many individuals with preclinical AD never progress to AD (Sperling, Mormino, & Johnson, 2014), which suggests that levels of Aβ do not have a direct influence in determining cognitive function (Aizenstein et al., 2008) (Figure 1). Figure 1 shows that as AD progresses over time the brain’s structure and function (MRI and PET) increases as well as Aβ. Previous literature has focused on severe cases of AD and healthy controls; however, there are fewer studies on identifying potential AD markers on pre-clinical AD individuals. This early stage of the disease is crucial for prevention and detection of high-risk individuals since overt cognitive decline is still absent.

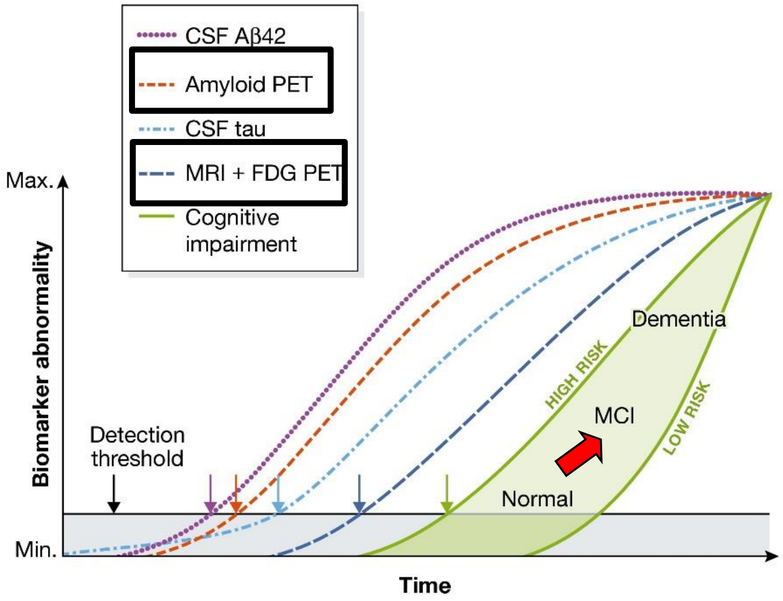


Figure 1.

Purpose of Study

As previously mentioned, the preclinical stage in Alzheimer's disease is important because individuals in this stage have a significant amount of Aβ present in their brains however cognitive deficits are still absent. Thus, this stage is crucial for early detection, identification, and possibly prevention of individuals who are at high-risk of AD. My study aims to enhance our understanding of Aβ accumulation and its relation to brain structure and function. Additionally, developing a predictive model to estimate Aβ will help to create disease model trajectories so that clinicians can advise better treatment plans for individuals.

Implementation

Implementing my research project involved many steps including data collection and reshaping, performing principal component analysis (PCA), visualizing the regions association with each of the principal components, using several clustering methods to analyze patterns between regions, and generating a heatmap to visually see the spatial distribution of active regions per PC. The data that was used for my project was in the format of an excel workbook and contained several variables including: demographic information, multiple measurements of the same subject, across time, and using different measures to quantify brain structure and function (ex. PiB, GMD, FDG). All the analyses that were conducted for this research project were implemented in the statistical software programming language called R. R contains several in built packages and methods allowed me to perform PCA efficiently and generate visualizations to identify patterns in my data.

Data Collection and Reshaping

My research project used data from an ongoing large center study (Normal Aging Study) which includes individuals who are in the pre-clinical Alzheimer's disease stage. I used a sample size of n = 85 unique individuals with a mean age = 76.4±6.1 years. Participants in the dataset underwent brain imaging such as magnetic resonance imaging and positron emission tomography (MRI and PET respectively) to measure brain structure and function.

- MRI to estimate brain structure → (GMD)
- Positron emission tomography (PET) with tracers to estimate brain function→ (FDG)
- Positron emission tomography (PET) with Pittsburgh compound B to estimate Aβ → (PiB)

Please refer to the supplemental figure for more information about the number of regions in the brain that were used and the individual measures that were used in the analysis.

Principal Component Analysis (PCA)

After subsetting the data into 52 unique brain regions, we decided to look at the correlation structure between regions. We created a correlation plot in R using the corr() method as seen in Figure 2. The correlation plot for GMD variables shows three pockets of squares or groupings along the diagonal which indicates that our variables (brain regions) are highly correlated with each other.

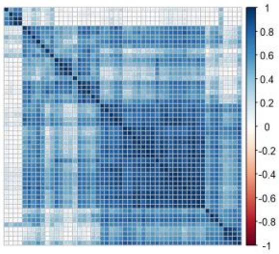
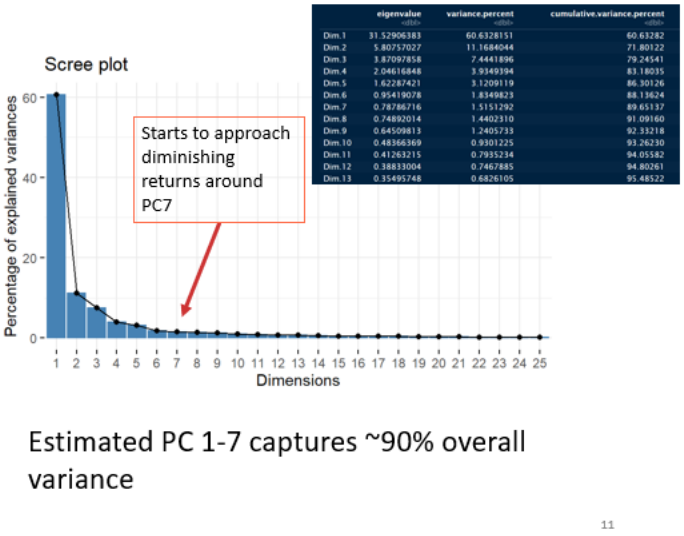


Figure 2.



	eigenvalue	variance.percent	cumulative.variance.percent
	<dbl>	<dbl>	<dbl>
Dim.1	31.52906383	60.6328151	60.63282
Dim.2	5.80757027	11.1684044	71.80122
Dim.3	3.87097858	7.4441896	79.24541
Dim.4	2.04616848	3.9349394	83.18035
Dim.5	1.62287421	3.1209119	86.30126
Dim.6	0.95419078	1.8349823	88.13624
Dim.7	0.78786716	1.5151292	89.65137
Dim.8	0.74892014	1.4402310	91.09160
Dim.9	0.64509813	1.2405733	92.33218
Dim.10	0.48366369	0.9301225	93.26230
Dim.11	0.41263215	0.7935234	94.05582
Dim.12	0.38833004	0.7467885	94.80261
Dim.13	0.35495748	0.6826105	95.48522

Starts to approach diminishing returns around PC7

Estimated PC 1-7 captures ~90% overall variance

Figure 3.

Because of this high correlation we decided to use an unsupervised learning technique called principal component analysis to reduce the dimensionality of our data set while still maintaining the interactions between brain structure (GMD), function (FDG), and Aβ in different brain regions. Because of our high dimensional problem and dataset, PCA is a great tool for simplifying the complexity in our high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.

After performing PCA, we observed that 7 principal components (PCs) were able to capture around 90% of the overall variance (Figure 3). Additionally, we reconstructed the GMD values per region according to PC1 through PC7 and compared it to the actual GMD values per region. Figure 4 demonstrates that as we add PCs to our analysis the error term (reconstructed value – observed value) is pushed down to zero.

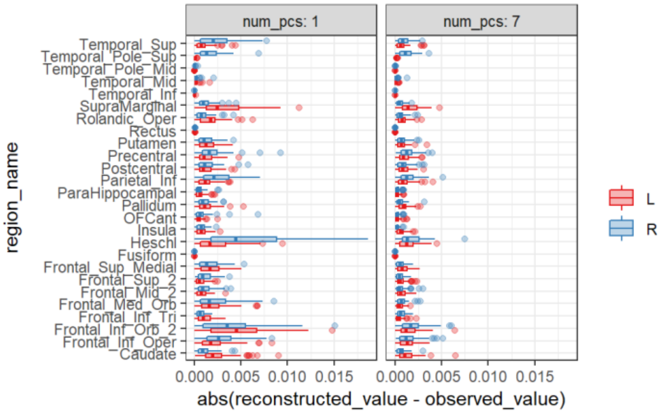


Figure 4.

Analyzing the Active Contribution of Regions

After performing PCA, our next step was to analyze the active regions in the brain that were correlated with each principal component. We did this by establishing a threshold as seen by the red dashed line (Figure 5). In Figure 5, the red line indicates that all brain regions are equally contributing to this PC. Therefore, regions above the red line are active regions and regions below the red line indicate inactive regions for that PC. Since PC1 accounted for 60% of the overall cumulative variance between variables it is reasonable to see that many regions are active in this plot and about 40% of the regions lie below the red threshold line.

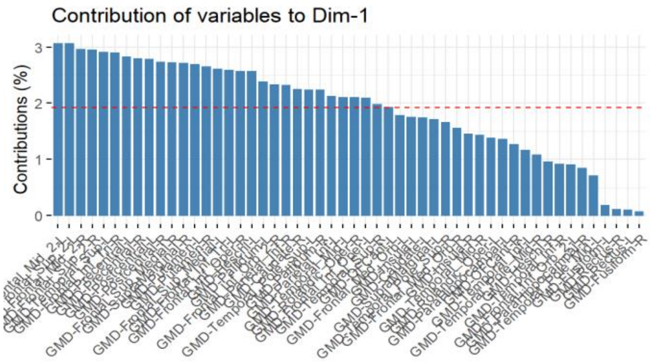


Figure 5.

The horizontal red line represents the threshold of the regions having an equal contribution. Equal contribution = (100 * 1/52). In Figure 5, Dim-1 refers to PC1 and shows that regions above the red line are active regions and regions below the red line are inactive for PC1.

Because the brain is symmetrical, I also performed a hemispheric active contribution analysis to see which regions of the brain are associated with each principal component (Figure 6). Figure 6 shows a breakdown of regions associated with each principal component. Although this does not show us which region belongs to which PC definitively, we can observe the fact that hemispheric differences in the regions are significant enough to be captured by different principal components depending on the hemisphere.

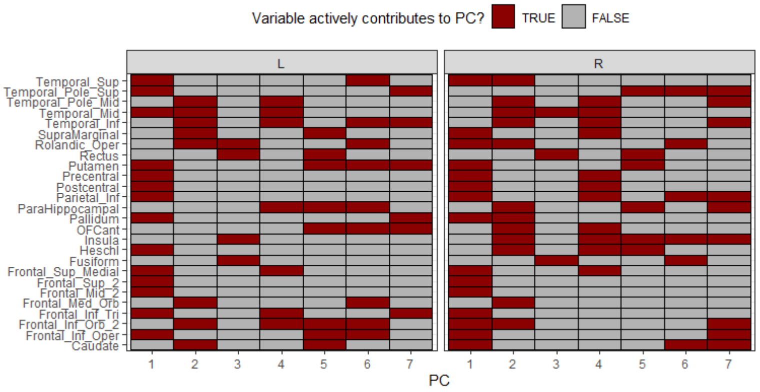
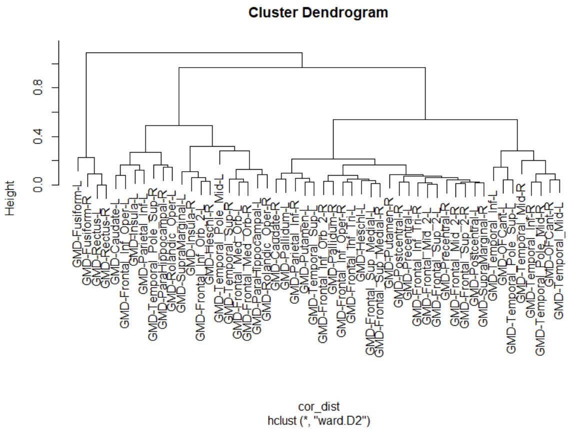


Figure 6.

Hierarchical Clustering

After examining the active regions per PC, we performed several clustering methods to understand the trends in the brain regions. In Figure 7, we visualized these trends using a hierarchical dendrogram based on correlation distance metric of the regions.



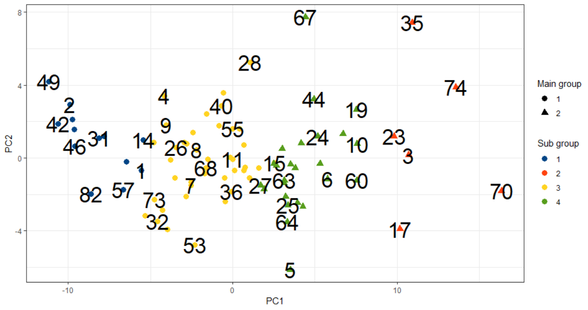


Figure 7. **Figure 8.**

We then cut the dendrogram into Macro and Sub Macro grouping to further cluster the regions based on individual people in PC1 and PC2. In Figure 8, individual 70 and individual 49 are less likely to be associated with PC1 because they both lie on the extreme ends of the plot.

Spatial Distribution of Active Regions

Lastly, we generated a heat map using a brain atlas called ggHo() in R to visualize the spatial distribution of active regions in PC1 through PC4 (Figure 9). Reg regions indicate active regions, blue regions are inactive regions, and grey regions are due to missing region names between our dataset and the brain atlas we used.

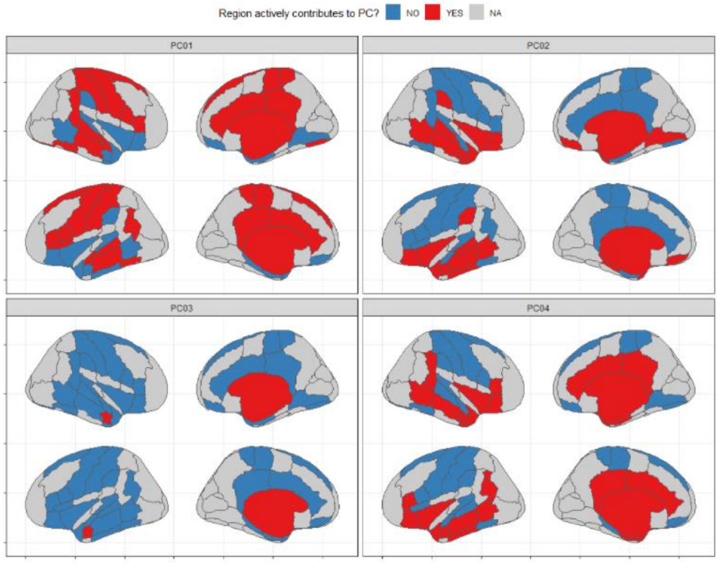


Figure 9.

Results

Our results showed that the first seven PCs from our PCA analysis was able to capture around 90% of the explained cumulative variance among the brain regions. Additionally, the correlation plot showed that many of the brain regions were highly correlated which reaffirmed our use of PCA to reduce the dimensionality of our data set while still maintaining the trends and behavior between variables (brain regions). Lastly, the heat map (Figure 9) showed that regions having similar function are captured by specific PCs. By performing PCA, we can use this information to reconstruct the whole sample space and find a suitable model predict Aβ.

Reflections

Limitations of Study

Limitations of my research project include having a smaller data set size of only 85 unique individuals. This is a limitation because it could lead to an overfitting problem when developing our predictive model. Additionally, this research problem is also a multi-output problem which increases the complexity. Even though we are only predicting beta amyloid, we are predicting it in several PCs, across several individuals and brain regions. Therefore, there is a tradeoff that exists between complexity and interpretability of the overall predictive model.

Reflections of My Research Project and Experience

During my research experience I faced challenges during the cleaning and reshaping data stage. Several of the datapoints had missing values for several input columns. To resolve this, we considered imputing the missing values however for simplicity we only kept the individuals who had no missing values. One thing I did notice was some of the IDs that represented individuals had a frequency greater than 1 in the file. For example, Vault_UID == 900534 had a frequency of 3. This means that the individual that corresponds with Vault UID 900534 had several brain image scans completed at different times. In order to move forward with our analysis, I decided to hold the image scan date time constant across all participants so that there would not be any repetitions of participants in our dataset.

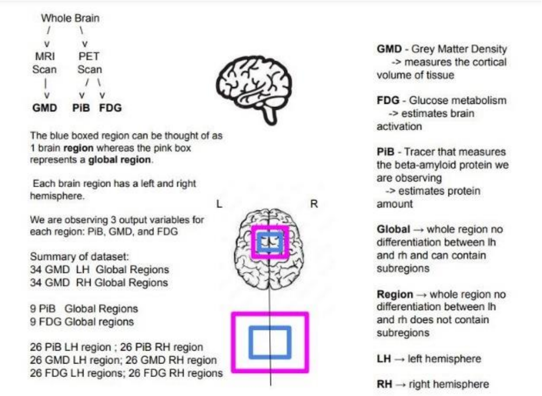
Other conceptual challenges I encountered included figuring out ways to visualize distributions and correlation structures in R. I was unaware of the abundant number of packages that existed in R and with the guidance of my mentor I was able to plot our visualizations after our weekly meetings.

While I have been involved in previous research projects throughout my undergraduate education, I still discovered that the research process is not linear but is rather iterative. At the end of each weekly meeting with my research advisor, I would think about the next step I needed to take to move forward with my project. This often meant breaking up a bigger problem into smaller and manageable chunks that could be implemented in just a few lines of code. Additionally, during my research project my perspective on developing predictive models changed. My advisor showed me that it is not always necessary to use the fanciest neural network to solve a problem but rather use a combination of several simple models that can help explain a much bigger and complex task such as predicting Aβ.

Future Directions

Future directions include using the PC scores from the FDG and GMD analysis to predict Aβ in PC1 using different polynomial models. This approach would use FDG and GMD as inputs to predict the output variable Aβ across different PCs. We then could evaluate models and choose the best one using AIC and BIC metrics. AIC and BIC are metrics that model performance and account for model complexity. Lastly, we could find a suitable predictive model to predict Aβ since the relationship between variables is complex.

Supplemental Information



The blue boxed region can be thought of as 1 brain **region** whereas the pink box represents a **global region**.

Each brain region has a left and right hemisphere.

We are observing 3 output variables for each region: PiB, GMD, and FDG

Summary of dataset:
34 GMD LH Global Regions
34 GMD RH Global Regions

9 PiB Global Regions
9 FDG Global regions

26 PiB LH region ; 26 PiB RH region
26 GMD LH region; 26 GMD RH region
26 FDG LH regions; 26 FDG RH regions

GMD - Grey Matter Density
-> measures the cortical volume of tissue

FDG - Glucose metabolism
-> estimates brain activation

PiB - Tracer that measures the beta-amyloid protein we are observing
-> estimates protein amount

Global → whole region no differentiation between lh and rh and can contain subregions

Region → whole region no differentiation between lh and rh does not contain subregions

LH → left hemisphere

RH → right hemisphere