MRI Data Analysis of Patients Diagnosed with Dementia

Written by Madhu Raghunath, Yamini Sasidhar, and Shishir Tewari

Acknowledgements

Datasets were provided by OASIS of the Washington University Alzheimer's Disease Research Center. Grant Numbers: P50 AG05681, P01 AG03991, R01 AG021910, P20 MH071616, U24 RR0213.

Table of Contents:

Summary

Our analysis was performed on a dataset containing longitudinal MRI data of patients diagnosed with dementia. Dementia itself is not a specific disease. Its a group of symptoms associated with decline in memory severe enough to reduce a person's ability to perform everyday activities. Alzheimer's is the most common type of dementia. Increase in age is the greatest known risk factor for Alzheimer's and the majority of people with it are 65 and older. Approximately 200,000 Americans under the age of 65 have early-onset Alzheimer's disease.[1]

Our motivation for this data analysis was to identify any indicators for developing dementia in an older subject population. We wanted to characterize the population of older people who develop dementia. As the number of dementia cases increases rapidly, there is a growing need to understand its indicators. In year 2010 number of patients were 4.4M which are estimated to increase to 11.0 M by 2050 in North America region.[2]

Questions and Data Overview

We used the public platform Kaggle to access our dataset, which was uploaded by user Jacob Boysen. This dataset was uploaded as part of the Open Access Series of Imaging Studies (OASIS) project by Washington University Alzheimer's Disease Research Center, making MRI datasets available for public use in clinical and data science project analyses.[3] There were two datasets available for our analysis: cross-sectional and longitudinal MRI data. We chose to use the longitudinal MRI dataset, as longitudinal studies tend to provide more consistent evidence for the effects of aging on intracranial volume.[4]

[1] https://www.alz.org/alzheimers-dementia/what-is-dementia

[2] Alisson Abbott. Nature 2011;475:S2-S4

[3] OASIS: Longitudinal: Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382

[4] Pfefferbaum, A., & Sullivan, E. V. (2015). Cross-sectional versus longitudinal estimates of age-related changes in the adult brain: overlaps and discrepancies. *Neurobiology of Aging*, *36*(9), 2563–2567. http://doi.org/10.1016/j.neurobiolaging.2015.05.005

This dataset includes data from 373 MRI images, across 150 subjects aged 60-96 who visited their physician up to 5 times. These subjects are all right-handed men and women, and separated into three main clinical groups: demented, nondemented, and converted. For the scope of this analysis, we understand the converted group to be patients with mild cognitive impairment (MCI) that become clinically demented.[5] Within this dataset, we find 72 patients that are diagnosed and remain as clinically nondemented throughout the study, and 64 patients that are diagnosed and remain as clinically demented throughout the study. Within this latter population, 51 patients are diagnosed with mild to moderate Alzheimer's Disease (AD). Finally, the remaining 14 patients are designated as clinically nondemented at the start of the study, and found to be clinically demented by the end of the study.

There are several variables under study:

- Subject ID : This is unique identifier to identify patients.
- MRI ID : Unique MRI ID per visit of each patient's.
- Visit (1-5) - Number of times Patients visited to clinic.
- Group : There are mainly three groups Demented, Nondemented and Converted.
- Gender : Male and Female.
- Hand : All patients were only right handed.
- Age (60-96) : Patients were in age range of 60 to 96.
- Years of education (6-23) : All patients had 6 to 23 years of education.
- Socio-economic status (1-5): Socioeconomic status (SES), ranged between 1 and 5, is an economic and sociological combined total measure of a person's work experience and of an individual's or family's economic and social position in relation to others, based on income, education, and occupation.
- Mini-Mental State Examination (MMSE 0-30) : The Mini–Mental State Examination (MMSE) is a 30-point questionnaire that is used extensively in clinical and research settings to measure cognitive impairment.
- Clinical Dementia Rating (CDR 0-2): Severity ratings range along a 5-point scale
  - CDR-0: no cognitive impairment
  - CDR-0.5: questionable or very mild dementia
  - CDR-1: mild
  - CDR-2: moderate
- Quantitative: eTiv and nWBV quantitative variables used in analysis. eTIV ranges from 1106 to 2004 whereas Whole Brain Volume is normalized in scale of 0 to 1
  - Total Intracranial Volume (eTIV)
  - Whole Brain Volume (nWBV)
- Misc: Following two variables were not explaining much so dropped from our analysis.
  - MR Delay (ms)
  - Atlas Scaling Factor (ASF)

Based on these variables, our main questions for this dataset are as follows: (1) Which gender is more at risk of developing dementia at a later age? (2) And at what age group? (3) How do clinical scores

[5] Farias, S. T., Mungas, D., Reed, B. R., Harvey, D., & DeCarli, C. (2009). Progression of Mild Cognitive Impairment to Dementia in Clinic- vs Community-Based Cohorts. *Archives of Neurology*, *66*(9), 1151–1157. http://doi.org/10.1001/archneurol.2009.106

determine the risk factor? (4) Does socioeconomic status and education play a role? (5) How the brain volume factor in?

Data Cleanup and Exploration

Our first steps in cleaning up the dataset involved figuring out which variables were not relevant to our analysis. We decided to remove the columns for: MRI ID, MR Delay, Hand, and ASF. We removed the redundant MRI ID column because we already had identifiers for the patient and each visit (i.e. OAS2_**0001**_MR**2** -> Patient **0001**, Visit **2**). We removed the MR Delay column because the time for image acquisition had no effect on our results. The Hand column showed that all subjects in this sample population were right-handed, so this column was also dropped. The ASF column was dropped as well, as it was included only to prove the validity of the quantitative brain volume variables.

As explained above, for the scope of this analysis we chose to rename any subject instances labeled as “converted” as “demented”. For convenience, we renamed the columns for years of education (“EDUC” -> “EDU”) and subject gender (“M/F” -> “Gender”). We noticed that while socioeconomic status (SES) was given on a 5 point scale, EDU was given as a number of years spanning 6-23. For convenience, we calculated normalized scores for this variable on a similar scale of 1-5, and took the ceiling values to view the data in clusters. The formula we used in our analysis is represented below in Figure 1.

$$Y = 1 + \frac{(X - A)(5 - 1)}{B - A}$$

$$\begin{aligned} A &= minimum\ value \\ B &= maximum\ value \\ X &= original\ value \\ Y &= scaled\ value \end{aligned}$$

Figure 1: Formula for normalizing values on scale of 1 - 5.

In regards to missing data, we were missing 19 entries for SES and 2 entries for MMSE in individual MRI scans. Coincidentally, these missing entries overlapped; when dropping these rows, we found that we only dropped 8 subjects overall. The final dataset left us with 142 out of 150 subjects in our sample population, and 354 out of 373 individual MRI scans.

Finally, we chose to create two separate dataframes in pandas containing the first visits and last visits of each subject by dropping duplicate rows based on the first and last instance of each Subject ID. We made sure to sort the original dataframe by Visit, to make sure we were only grabbing the first and last visits. We made sure to use the last visit dataframe in cases where we wanted to use the patient’s most current status.

Data Analysis

In order to determine if Gender played any role with the risk of developing dementia, we compared each gender with their CDR and MMSE Scores (**Figure 2**). However, we found no significant difference between them. In both cases gender median were identical, with CDR being 0.5 and MMSE being 26.0.

With the given dataset it’s inconclusive if gender has an impact in determining the symptoms of Dementia.

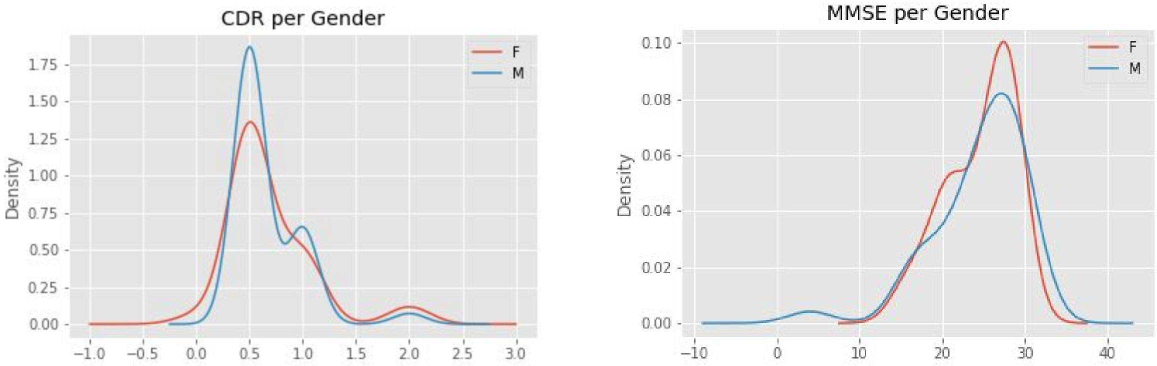


Figure 2: CDR & MMSE per Gender

With Age Vs Group distribution, there’s seem to be a higher concentration of clinically demented patients in the age group of 66-76 (**Figure 3)** than those of the nondemented patients. So perhaps, the onset of symptoms get higher from age 66 and higher.

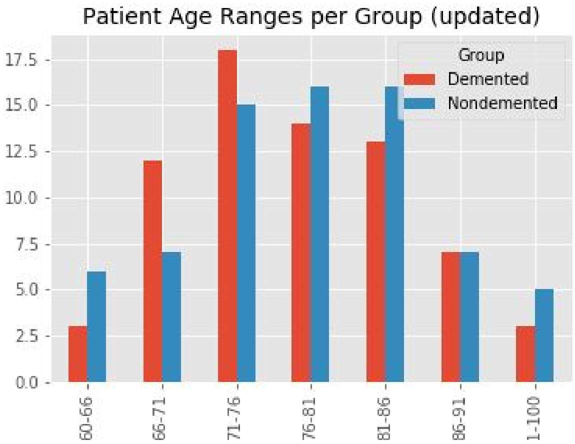


Figure 3: Patients by Group

In the Education Vs Group distribution, clinically demented patients had less median years of education compared to the nondemented. The box plot shows the median between them being 14 for demented and 16 nondemented patients.

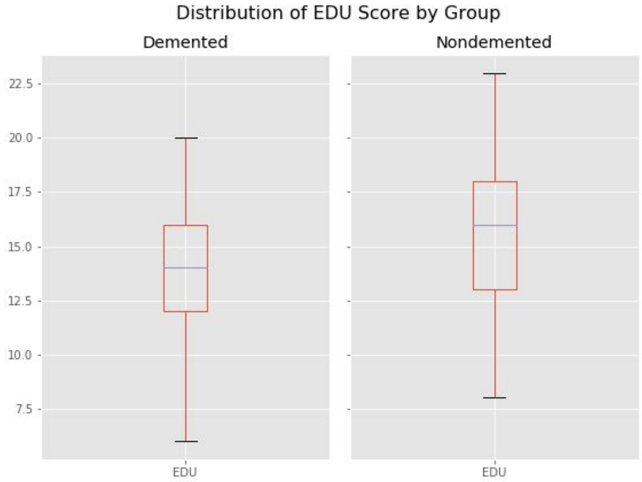


Figure 4: Boxplots depicting the distribution of EDU scores per group.

We can see that there is no obvious relation between clinically demented patients and their socioeconomic status. However very less patients are in higher SES range (**Figure 5**).

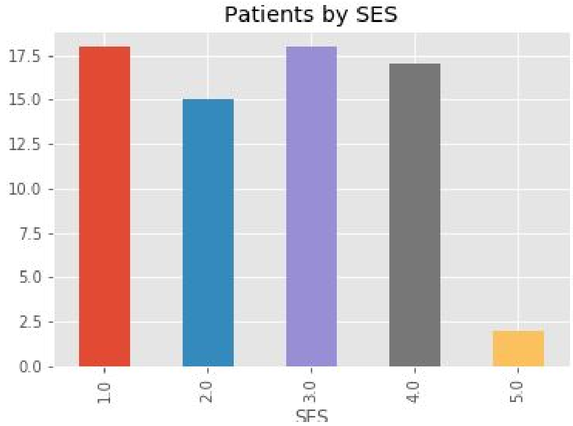


Figure 5: Bar chart depicting the distribution of SES scores for clinically demented patients.

All non demented patients are having 0 CDR ratings whereas demented patients having 0, .5, 1 and 2 ratings. Approximately 95 % demented patients are within the range of 0.5 and 1 (**Figure 6**).

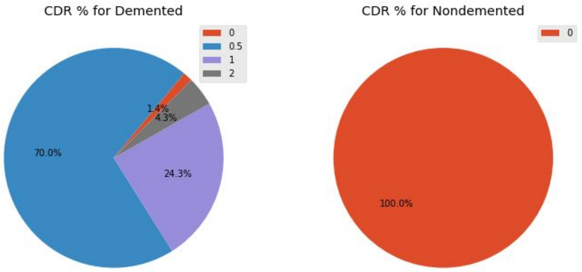


Figure 6: Pie charts depicting the distribution of CDR scores for clinically demented patients.

We chose to depict a density plot for nWBV so we could visualize where the medians and modes of each population subset lie. We can see from **Figure 7** that the clinically demented group has a lower median representing the data subset, showing a lower normalized whole brain volume for these patients.

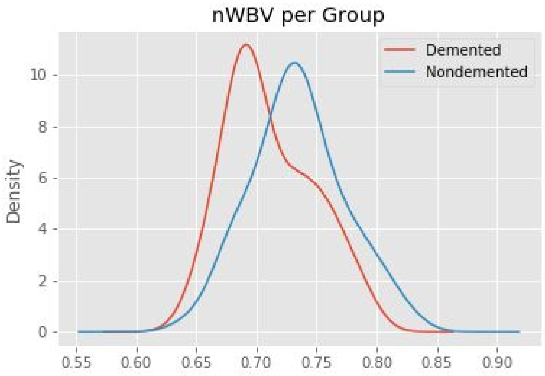


Figure 7: Density plot of nWBV per clinical group with medians;
(a) Demented = 0.6995, (b) Nondemented = 0.7345

Estimated Total Intracranial Volume (eTIV) (**Figure 8**), ranges from 1106 to 2004, has higher number of demented patients in the range of 1300 to 1599.

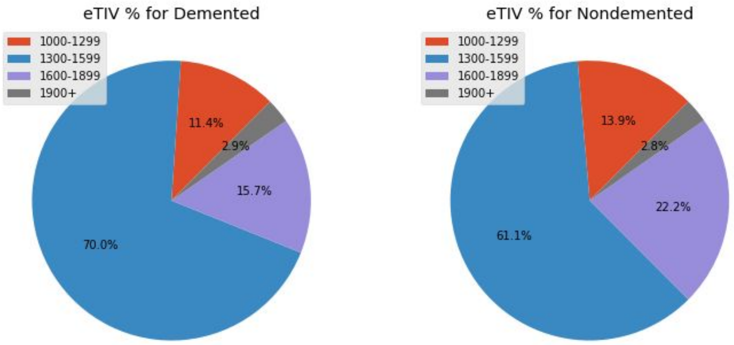


Figure 8: Pie charts depicting the distribution of eTIV quantities for both groups.

We chose to generate a scatter matrix as an efficient method of visualizing the relationships between all variables of interest. This scatter matrix was performed on the group of clinically demented patients, and is seen in **Figure 9**, below. We chose to display density plots on the diagonal, in place of the usual histograms or boxplots; density plots help us to visualize where the median and modes of the data lie.

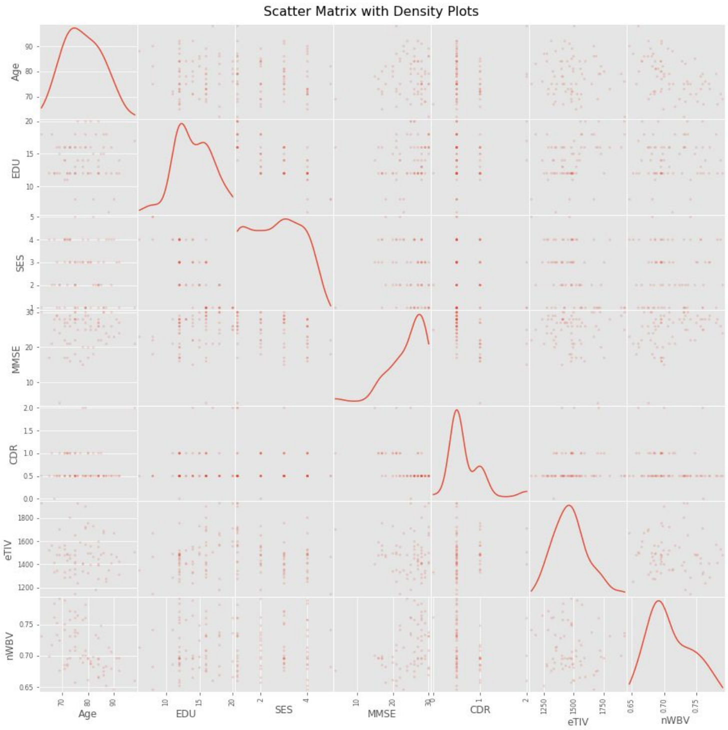
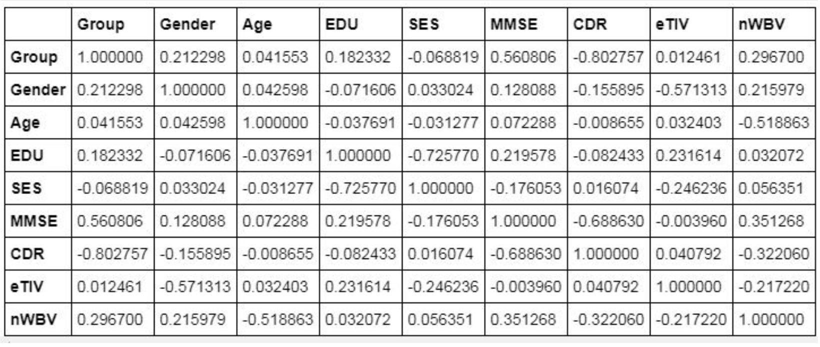


Figure 9: Scatter matrix comparing all variables of interest, density plot along diagonal.

We can see that many of these individual scatter plots have data clustered around number scores, but this does not help our analysis without looking for any linear regression in the data. Instead of plotting individual scatter plots with the linear regression model, we decided to calculate and store the Pearson's R correlation coefficients between all variables of interest in a dataframe (**Table 1**).

Table 1: Tabulated correlation coefficients for linear regression plots.



	Group	Gender	Age	EDU	SES	MMSE	CDR	eTIV	nWBV
Group	1.000000	0.212298	0.041553	0.182332	-0.068819	0.560806	-0.802757	0.012461	0.296700
Gender	0.212298	1.000000	0.042598	-0.071606	0.033024	0.128088	-0.155895	-0.571313	0.215979
Age	0.041553	0.042598	1.000000	-0.037691	-0.031277	0.072288	-0.008655	0.032403	-0.518863
EDU	0.182332	-0.071606	-0.037691	1.000000	-0.725770	0.219578	-0.082433	0.231614	0.032072
SES	-0.068819	0.033024	-0.031277	-0.725770	1.000000	-0.176053	0.016074	-0.246236	0.056351
MMSE	0.560806	0.128088	0.072288	0.219578	-0.176053	1.000000	-0.688630	-0.003960	0.351268
CDR	-0.802757	-0.155895	-0.008655	-0.082433	0.016074	-0.688630	1.000000	0.040792	-0.322060
eTIV	0.012461	-0.571313	0.032403	0.231614	-0.246236	-0.003960	0.040792	1.000000	-0.217220
nWBV	0.296700	0.215979	-0.518863	0.032072	0.056351	0.351268	-0.322060	-0.217220	1.000000

To help visualize these correlation coefficients and instead of scanning each tabulated value one-by-one, we chose to depict each value in a colorplot (**Figure 10**). It is important to note that for this colorplot, we plotted the absolute values of each correlation coefficient so that we could easily spot the plots of interest with the best fits of linear regression.

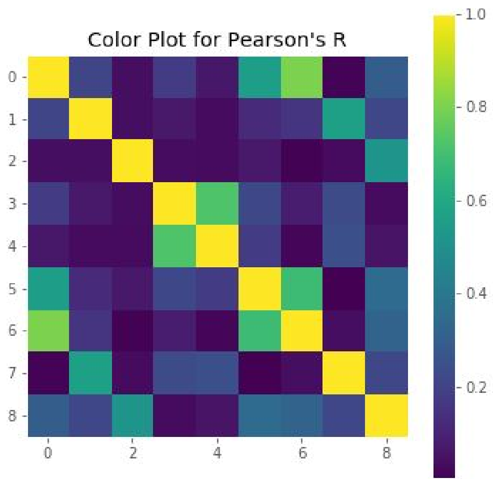


Figure 10: Colorplot depicting linear regression plots of interest based on calculated Pearson's R correlation coefficient.

We investigated the plots of interest as determined in the colorplot by looking at either side of the diagonal, and generated scatter plots with regression lines for each. The three scatter plots are shown below in **Figure 11**. Unfortunately, though the first plot CDR vs. Group had the highest correlation coefficient (r = -0.802), it is unmeaningful. It tells us something we already know: One group has a larger range of CDR scores than the other. In MMSE vs. EDU data points are clustered along with normalized education scores. Although correlation coefficient (r = -0.726) is not higher but a negative linear trend is setup between MMSE and Education. Same kind of negative linear trend(r = -0.689) is also established between MMSE and CDR (MMSE vs. CDR) ratings. Which is meaningful because with higher CDR ratings MMSE should be lower.

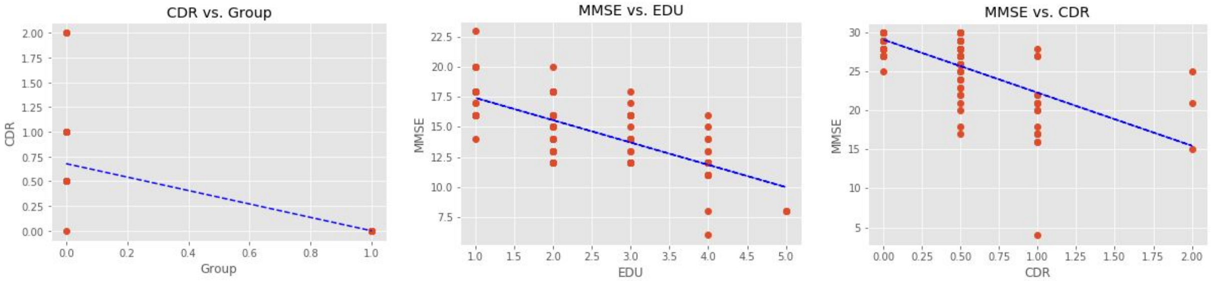


Figure 11: Three generated scatter plots based with highest correlation coefficients: (a) CDR vs. Group, r = -0.802; (b) MMSE vs. EDU, r = -0.726; (c) MMSE vs. CDR, r = -0.689.

It is important to recognize that for the scope of this analysis, we do not aim to reject the null hypothesis in testing the validity of our sample population. We do not expect the two groups of clinically demented and nondemented subjects to be randomly selected; instead, we expect these population subsets to be precise and entirely different. Any variables which show a significant difference in sample means between the two groups proves to be a viable indicator of prevalent dementia within this analysis.

When performing independent student's t-tests, we found the difference in sample means to be significant (p < 0.05) for MMSE, CDR, nWBV, and EDU (**Figure 12**).

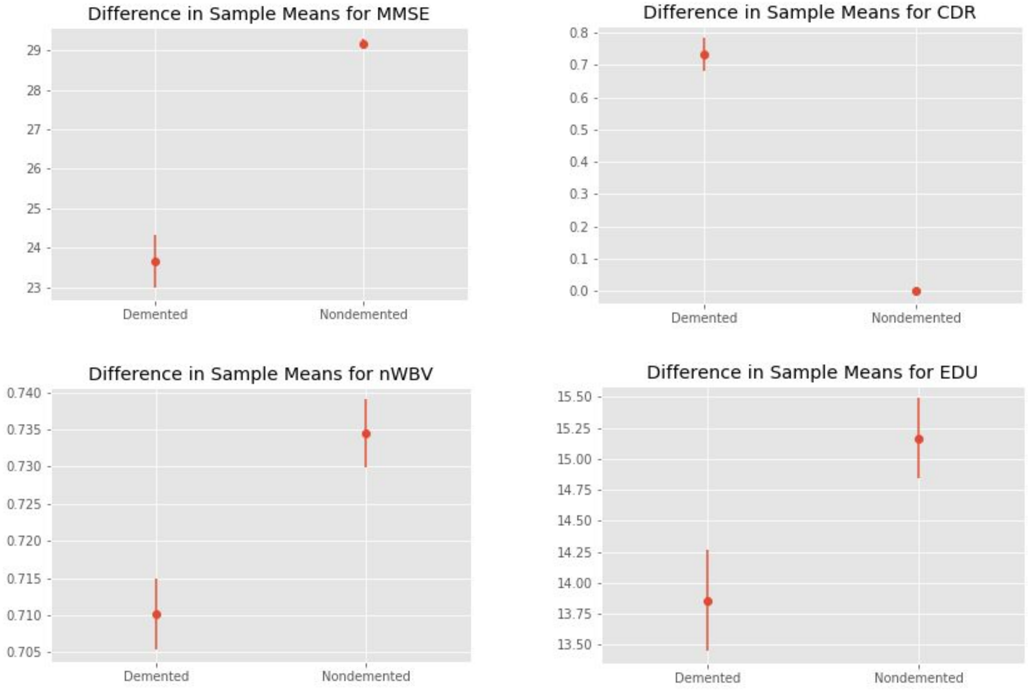


Figure 12: Variables for which the difference in sample means is significant.

On the other hand, we found the difference in sample means to not be significant (p > 0.05) for SES and eTIV (**Figure 13**). We conclude that these variables do not make for viable indicators in older patiets with dementia.

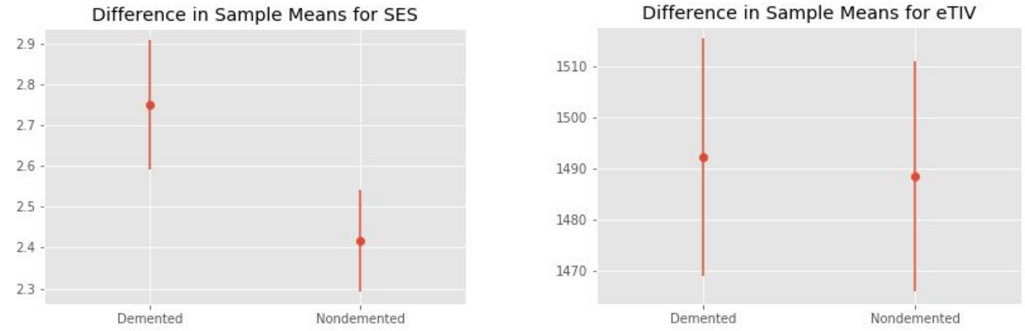


Figure 13: Variables for which the difference in sample means is not significant.

Discussion

We performed this data analysis to see whether we could identify any indicators for the progression of dementia in an older patient population. Our statistical tests couple with our generated plots to accurately depict our population under study. The tools we used include: Pandas, Matplotlib, Numpy, Scipy, and MS Excel PivotTables.

For this sample population, we succeeded in finding: (1) the age group (66 -76) with a higher prevalence of dementia; (2) a negative linear trend between CDR score and MMSE score. A significant difference in normalized whole brain volume between groups. We were not able to find: (1) if one gender had higher prevalence of dementia; (2) a relationship between socioeconomic status and prevalence of dementia, (3) a significant difference in estimated intracranial volume between groups.

Our data was mainly limited in quantity, with only 142 subjects under study and uneven gender counts. Additionally, we felt that some of our variables remained unclear even with research, and we did not have sufficient documentation to access for our questions. A lot of our decisions were made based on reading further literature and discussing until we all agreed on a direction. This mainly affected our choices in which variables applied to our analysis and which did not. We were specifically concerned in whether or not the changes in variables eTIV and nWBV were specific to the effects of aging on dementia status.

In taking steps to expand upon this analysis, we'd be interested in acquiring a larger dataset with additional variables and information, including but not limited to: race, profession, lifestyle, genetic background, and pre-existing conditions. Additionally, we'd be interested in accessing variables for atlas-scaled normalized volumes of different brain regions. Studying the volumes of the temporal lobes, hippocampus, and lateral ventricles compared and normalized to the whole brain may give some insight as to which regions tend to atrophy quicker and the manner in which they do in older patients diagnosed with dementia.[6] A breakdown of these regions of interest is given below in **Table 2**.

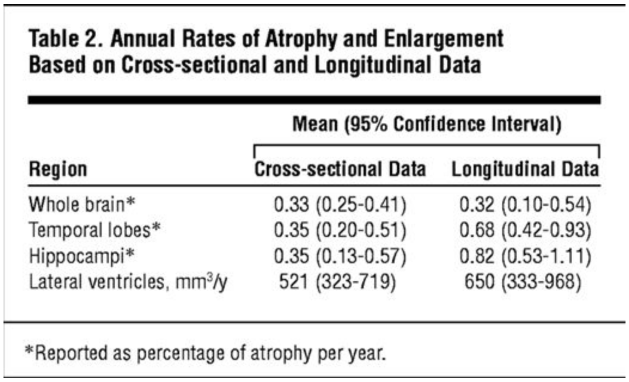


Table 2. Annual Rates of Atrophy and Enlargement Based on Cross-sectional and Longitudinal Data

	Mean (95% Confidence Interval)	
Region	Cross-sectional Data	Longitudinal Data
Whole brain*	0.33 (0.25-0.41)	0.32 (0.10-0.54)
Temporal lobes*	0.35 (0.20-0.51)	0.68 (0.42-0.93)
Hippocampi*	0.35 (0.13-0.57)	0.82 (0.53-1.11)
Lateral ventricles, mm^3/y	521 (323-719)	650 (333-968)

*Reported as percentage of atrophy per year.

[6] Scahill RI, Frost C, Jenkins R, Whitwell JL, Rossor MN, Fox NC. A Longitudinal Study of Brain Volume Changes in Normal Aging Using Serial Registered Magnetic Resonance Imaging. *Arch Neurol.*2003;60(7):989–994. doi:10.1001/archneur.60.7.989