Analysis Of Statistical Data on Dementia and Seeking Potential Correlations and Research Directions

Chenlai Zheng*

Department of Biology, Portland State University, Oregon, US

* Corresponding Author Email: chenlai@pdx.edu

Abstract. Modern human society is plagued by many diseases that were previously unappreciated, dementia being one of these diseases. Due to advances in modern medicine, the average human lifespan has increased greatly. However, many diseases have also emerged as a result of the longevity of humans. This article tries to find the potential factors that lead to dementia by analyzing data from a dementia study of 150 subjects at Washington University. The paper focuses on screening and analyzing the data by using R language and programs, and the proportions of subjects were calculated and compared according to gender. In addition, age, eTIV, nWBV, and ASF were subjected to binary analysis and scatterplots in permutations and combinations. The study did not find potential associations for variables other than eTIV and ASF. This study primarily used data statistics and analysis to provide possible potential research directions for clinical researchers in this field, with the hope that future research will lead to breakthroughs in Alzheimer's disease (AD).

Keywords: Dementia; Alzheimer's disease; statistics; data analysis; R language.

1. Introduction

With the development of modern medicine, people's life expectancy has been significantly improved compared with the past, but more diseases are manifested and paid attention to as people live longer. Dementia is an increasingly prevalent public health problem plaguing human society [1]. As of 2017, approximately 50 million people worldwide were diagnosed with dementia, and this number continues to grow [1]. When people's memory and cognition decline, impaired thinking and decision-making skills, it often means that dementia has occurred. Many diseases cause dementia, such as Down syndrome caused by a genetic defect of chromosome 21, or brain damage caused by trauma to the brain, but the most common is Alzheimer's disease. Dementia is mostly manifested by brain atrophy, cognitive impairment, and Alzheimer's disease (AD) [1]. Alzheimer's disease (AD) is a chronic disease that will continue to deteriorate over time, and its main symptoms include memory loss, language impairment, and cognitive impairment [2-3]. Eventually, patients will lose their ability to take care of themselves and slowly move toward death. The high-risk population for AD is mainly the elderly, and AD is currently one of the leading causes of death among people over the age of 65[4]. According to the US CDC statistics, Alzheimer's disease patients account for 60% to 80% of dementia; vascular dementia accounts for about 10% of dementia. Lewy body dementia (LBD), frontotemporal dementia, and other diseases account for about 10% of dementia [5]. The traditional clinical diagnosis method is that doctors label patients with dementia and cognitive impairment as "probable AD" until the patients undergo autopsy and doctors find amyloid plaques and tau-based neurofibrillary tangles in the patient's brain, then doctors were able to determine that a patient had AD [6]. Since it is difficult to accurately diagnose AD in patients, the mainstream diagnostic methods are Mini-Mental State Examination (MMSE), The Herth Hope Index (HHI), and Adult Hope Scale (AHS) through psychological methods [7]. According to the current research, most scholars agree with the view that AD is caused by the deposition of amyloid beta (Aβ) in the brain of patients, causing cognitive degeneration, and there is a family genetic risk of AD [8]. Furthermore, it is believed that chromosomal clinically, abnormalities lead to an increased risk of AD [9]. However, effective means of diagnosing and treating AD are still lacking [10]. This means that how to diagnose and treat AD will be a good research direction. The purpose of this paper is to systematically analyze the data from patients with dementia to identify potential factors associated with AD. R language and

software will be used for data analysis and calculations. The paper hopes that the analysis of these data will lead to finding some statistical associations related to AD and to predict or evaluate AD in the future.

2. Methodology

The data comes from a larger database of individuals who had participated in MRI studies at Washington University [11]. A total of 150 subjects, aged 60 to 96, participated in the one-year survey. Since AD is the main cause of dementia in the elderly, this data can be used for AD research [4]. This data records the subject's MR delay, gender, age, social economic status (SES), education level (EDUC), mini-mental state examination (MMSE), clinical dementia ratio (CDR), estimated total intracranial volume (e-TIV), normalized whole brain volume (WBV), and atlas scaling factor (ASF). Table 1 shows the title, full name, and range of various variables in the data in detail. There are 150 subjects were divided into three groups: Demanted, Nondemanted, Converted. The higher the value of SES, the higher the income; the higher the EDUC value, the more years of education; the value of MMSE high than 26 is normal; the CDR value is Degree of dementia (0 is non-dementia, greater than 0 is dementia).

Table 1. Data dictionary

Variable name	Variable type	Variable description	Range
Group	Categorical	Group of dement	Demanted, Nondemanted, Converted
Visit	Numeric	Visit times	1~5
MR Delay	Numeric	delayed MR Scan	0~2639
Sex	Categorical	Gender	M=Male, F=Female
Age	Numeric	Age (years)	60~98
SES	Categorical	Social Economic Status	1~5
EDUC	Categorical	Education level	6~23
MMSE	Numeric	Mini-Mental State Examination	4~30 (higher than 26 being considered normal)
CDR	Numeric	Clinical Dementia Ratio	0~2
e-TIV	Numeric	Total Intracranial Volume	11106~2004
n-WBV	Numeric	Whole Brain Volume	0.644~0.837
ASF	Numeric	Scaling Factor	0.876~1.587

To search for potential relationships leading to AD, R language and software were used for data analysis in this paper. The first and most important step is to screen the data. Since the data come from the one-year of visits and records from 150 subjects, with some subjects leaving multiple records, it became critical to filter the data. Taking the last visit for this study was the most effective. This is because in general, if someone is in the borderline range of AD, then later records over time are more likely to be diagnosed with dementia. However, the earlier records may not be diagnosed with dementia. In addition, patients diagnosed as Converted in the data were also excluded when data cleaning was performed. This is because they were described as patients who were diagnosed with dementia in the early stages but were diagnosed with non-dementia in the later stages. The reason for the occurrence of Converted patients may be the result of misdiagnosis, as dementia is widely believed to be an irreversible and incurable disease [12]. However, misdiagnosis may be due to the fact that there is no effective and direct means to diagnose and treat dementia except a mental exam or brain CT scan.

After data screening, the data were categorized, studied and discussed from several variables based on their different attributes. First, filter the data and remove the unusable parts. Since some subjects in this data took multiple exams, they were recorded repeatedly. If these data are used directly, the data will lose independence. In order to avoid this situation, data screening will select the results of the subject's last exam as the available database. And the data were categorized and analyzed according to gender, and the data from males and females were divided into two groups for research and comparison. Afterwards, dementia and non-dementia were divided into two groups for comparison and research. Basic statistical methods were used to calculate their numbers and respective percentages. In addition, to look for differences in dementia between men and women (men are more likely to have dementia than women, or women are more likely to have dementia than men), two-tailed t-tests will be used to test the validity of the hypotheses.

In addition, four variables (age, e-TIV, WBV, ASF) are permutated to generate correlated scatterplots: age vs e-TIV, age vs WBV, age vs ASF, e-TIV vs WBV, e-TIV vs ASF, and WBV vs ASF. In this way, the possible relationship between these variables will be more easily found, and I hope that some linear correlations between variables can be found.

3. Result

R Studio and R language were used for data analysis. After data cleaning 128 subjects remained (54 males and 74 females) of which 56 were classified as demented and 72 as non-demented (Figure 1).

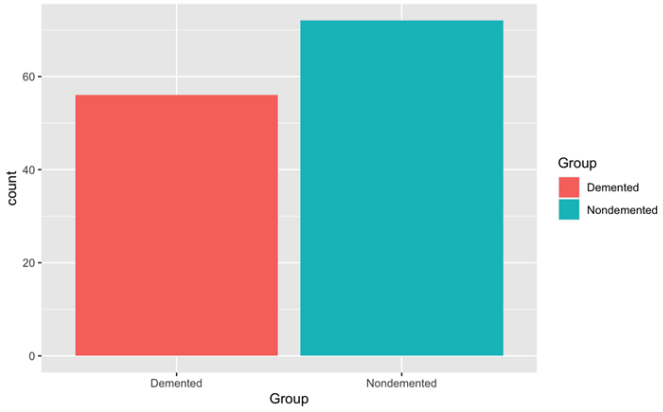

Fig. 1 Available data sets (dementia vs non-dementia)

Next, the data were grouped by two factors, gender and presence of dementia, and their percentages were calculated. There were 54 males and 74 females out of 128 subjects (Table 2). Out of these 128, 56 were diagnosed with dementia (43.75%) and the remaining 72 were non-dementia (56.25%). Males with dementia accounted for 57.14% of all dementia patients and 42.86% of females.

While 59.26% of males were diagnosed with dementia, the remaining 40.74% were non-dementia. Among females, 32.43% were diagnosed with dementia and 67.57% were non-dementia.

Table 2. Percentage of dementia in males and females

	Demented	Nondemented	Total	% of Demented	% of Nondemented
Male (M)	32	22	54	59.26%	40.74%
Female (F)	24	50	74	32.43%	67.57%
Total	56	72	128	43.75%	56.25%
% of Male	57.14%	30.56%			
% of Female	42.86%	69.44%			

After preliminary data screening, these data were also consistent with the independence of data. Data with both randomness and independence implied that the data obeyed a normal distribution and could be subjected to a two-tailed T test. The data were paired and combined by permutation and combination and plotted in the graph in Figure 2.

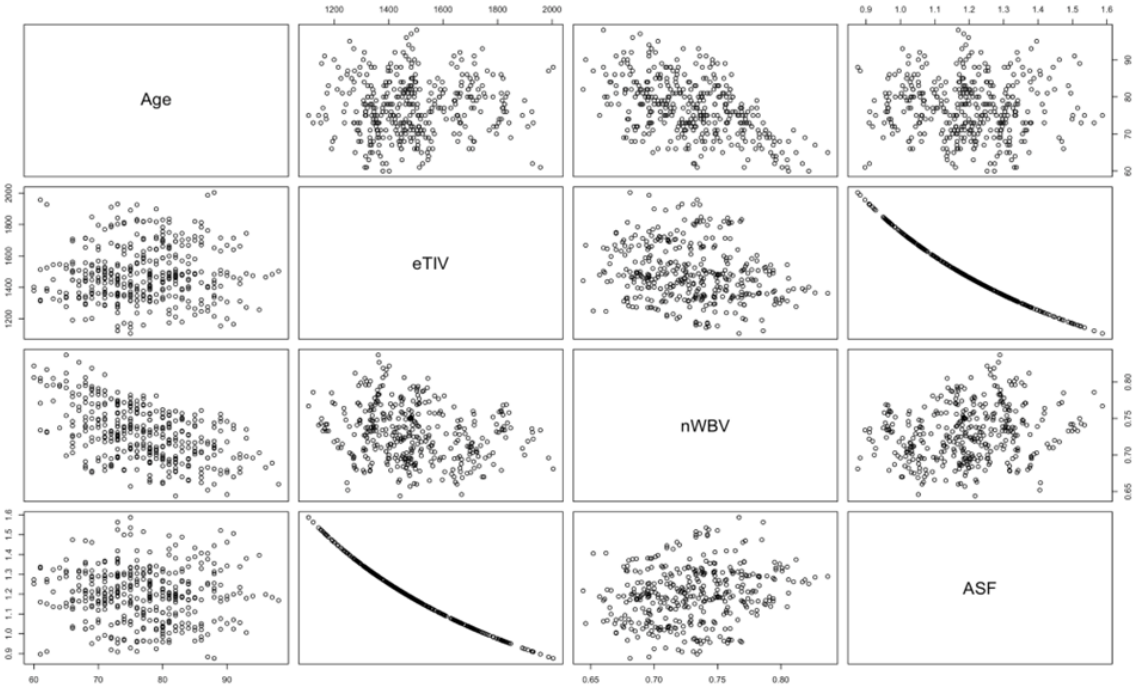

Fig. 2 Continuous Variables Correlation Charts

Figure 2 shows Continuous Variables Correlation Charts: age vs e-TIV, age vs WBV, age vs ASF, e-TIV vs WBV, e-TIV vs ASF, and WBV vs ASF.

These charts allow for quick screening of linearly correlated variables to quickly determine the association between these variables. The scatter plot of age vs eTIV obeys a normal distribution and shows some horizontal symmetry, so it is not possible to quickly determine whether there is a correlation between them. The scatterplot of age vs nWBV is also subject to a normal distribution, and the distribution of both is on both sides of the diagonal line, which shows that nWBV decreases linearly with increasing age. The scatterplot of age vs ASF showed a similar distribution to age vs eTIV, making it difficult to quickly identify the correlations that existed between them. However, the distribution of the scatter plots of eTIV vs nWBV and nWBV vs ASF is difficult to find, so further data analysis is needed to determine whether there is a correlation between them. However, the most

striking of these scatterplots is eTIV vs ASF, which shows a surprisingly linear correlation, with the points arranged as a curve. The ASF is a one-parameter scaling factor that allows for comparison of the estimated total intracranial volume (eTIV) based on differences in human anatomy. A high correlation between eTIV and ASF can be observed from the above figure, therefore, only ASF is used as one of the variables of interest in this paper.

4. Discussion

The results based on the data analysis showed a higher rate of dementia in women, which is inconsistent with the results of other studies. Other research results have shown that women are more likely to develop AD than men [13]. Because females are older than males on average, which means that women are more likely to develop AD during their lifetime. This remarkable result may be due to several underlying factors. First of all, the sample size for this paper is too small, with only 150 subjects participating in the data. Only 128 subjects actually met the research requirements of this paper. Smaller samples will lead to bias. It is quite possible that the females in this data happen to be in the two tails of the normal distribution in the population. Likewise, males may be at this extreme. Due to the existence of this deviation, the results of data analysis are inconsistent with other research results. Secondly, the smaller sample size may lead to another situation that may lead to deviations in the research results, that is, the existence of serious outliners. This kind of situation often occurs when the average value of a set of data is counted. When this happens, the calculation results will deviate due to the appearance of extreme subjects. It is clear that this study did not remove the maximum and minimum values when filtering the data. Third, the logistic regression model operation was not performed for the data analysis. The reason for the error in performing this operation is that the algorithm did not converge due to too many predictors. Since the logistic regression operation was not performed, it was difficult to obtain the prediction results of the regression equation. This makes it impossible to obtain the exact correlation between the variables. Finally, this analysis also did not perform a lasso regression operation. If a lasso regression had been done in this paper, more accurate prediction results might have been obtained or potential correlations with other variables might have been found.

Furthermore, there was no specific linear correlation between age and eTIV, which the author considered to be expected since AD causes brain atrophy in patients [14]. However, the negative correlation between age and nWBV is very interesting. One possible explanation is that the mean nWBV of the older population will be lower than the mean nWBV of the younger population because older people are more likely to suffer from AD. This explanation would be contradicted by the scatterplot of age vs eTIV. If this hypothesis is true, theoretically eTIV should also decrease with age and show a negative correlation. But the fact is that they do not show a negative correlation.

In addition, no significant positive or negative associations emerged between age and ASF. (Note: ASF is a unit to measure the expansion or contraction of the brain.) The most prominent feature of AD is the patient's cognitive decline, memory loss, and decreased mobility. But the main cause of these is the patient's brain. The older the brain atrophies or ages, the greater the risk of developing AD [9]. According to the current academic research results, the conclusion should be that age is negatively correlated with ASF, because older people are more susceptible to AD [15].

It is worth noting that the providers (subjects) of these data are all right-handed. This may cause errors in the study, because the left and right hands of humans are controlled by different hemispheres, and perhaps differences in the use of the brain (nerve activity) will lead to different manifestations of AD [16].

5. Conclusion

This study focused on the data analysis of the correlation between the odds of Alzheimer's disease and gender. The data were analyzed and processed by using R language to produce scatter plots of

correlations for each variable data. The results of this study showed that there was no strong correlation between gender and AD, but there was a strong negative correlation between eTIV and ASF. Although the conclusion of gender correlation is not in line with the expectations of this study, but this result is reasonable. The main reason for the results not meeting expectations is the small sample size, which greatly increases the possibility of error. In addition to the small sample size, limitations of this experiment include the fact that all subjects were right-handed and the experiment was unable to perform logistic regression and create a predictive model. Lasso regression seems to be a good way to study the relationship between these variables. If there is enough data, theoretically it should be possible to find the potential linear relationship between variables through lasso regression and build a predictive model. Unlike other regression equation predictions, lasso regression can solve the relationship between multiple complex variables well, and it can give more accurate prediction results. Theoretically, the more times lasso regression is used to simulate, the closer to the real result. Unfortunately, this study did not find evidence of a strong correlation between AD and age, eTIV, nWBV, and ASF. This study is just trying to find the correlation between them through data analysis, but how AD is formed is still unknown. Data analysis can only provide a research direction for subsequent clinical researchers, and how the brain lesions of patients eventually develop into AD is still a valuable research topic. Hopefully, in the near future, human brain science research will make great progress. People will be able to understand how the brain works and how brain diseases develop, which will be a breakthrough achievement for the prevention and treatment of these diseases.

References

[1] Olga Włoczkowska, Joanna Perła-Kaján, A. David Smith, et al. Anti‐N‐homocysteine‐protein autoantibodies are associated with impaired cognition. Alzheimer's & Dementia : Translational Research & Clinical Interventions, 2021, 7(1)

[2] Iliffe Steve; Burns Alistair. Alzheimer’s disease. BMJ. 2009-02-05, 338: b158 [2019-01-01]. ISSN 1468-5833. PMID 19196745

[3] World Health Organization, Dementia. Sep 20, 2022. Retrieved on Dec 24, 2022. Retrieved form: https://www.who.int/news-room/fact-sheets/detail/dementia

[4] Jeffrey Cummings, Garam Lee, Kate Zhong, et al. Alzheimer's disease drug development pipeline: 2021. Alzheimer's & Dementia : Translational Research & Clinical Interventions, 2021, 7(1)

[5] Centers for Disease Control and Prevention, Alzheimer's Disease and Healthy Aging. Retrieved on April 5, 2019. Retrieved form: https://www.cdc.gov/aging/dementia/index.html

[6] Ronald C. Petersen. How early can we diagnose Alzheimer disease (and is it sufficient)?: The 2017 Wartenberg lecture. Neurology, 2018, 91(9), 395–402.

[7] Benjamin T. Mast, Sheila L. Molony, Nicholas Nicholson, et al. Person‐centered assessment of people living with dementia: Review of existing measures. Alzheimer's & Dementia : Translational Research & Clinical Interventions, 2021, 7(1)

[8] Han Ly, Nirmal Verma, Savita Sharma , et al. The association of circulating amylin with β‐amyloid in familial alzheimer's disease. Alzheimer's & Dementia : Translational Research & Clinical Interventions, 2021, 7(1)

[9] Wayne Silverman, Sharon Krinsky-McHale. Alzheimer's risk and quality of life: History of down syndrome as a case in point. Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring, 2021, 13(1)

[10] Mohamed Raâfet Ben Khedher, Mohamed Haddad, Danielle Laurin, Charles Ramassamy. Apolipoprotein E4–driven effects on inflammatory and neurotrophic factors in peripheral extracellular vesicles from cognitively impaired, no dementia participants who converted to alzheimer's disease. Alzheimer's & Dementia : Translational Research & Clinical Interventions, 2021, 7(1)

[11] Battineni, Gopi; Amenta, Francesco; Chintalapudi, Nalini . "Data for: MACHINE LEARNING IN MEDICINE: CLASSIFICATION AND PREDICTION OF DEMENTIA BY SUPPORT VECTOR MACHINES (SVM)", 2019, Mendeley Data, V1, doi: 10.17632/tsy6rbc5d4.1

[12] Cleveland Clinic, Dementia. March 12, 2022. Retrieved on Dec 24, 2022. Retrieved form: https://my.clevelandclinic.org/health/diseases/9170-dementia

[13] Michelle M. Mielke. Sex and Gender Differences in Alzheimer's Disease Dementia. The Psychiatric times, 2018, 35(11), 14–17.

[14] Atri Chatterjee, Veronica Hirsch-Reinshagen, Syed Ali Moussavi, et al. Clinico‐pathological comparison of patients with autopsy‐confirmed alzheimer's disease, dementia with lewy bodies, and mixed pathology. Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring, 2021, 13(1)

[15] Randy L. Buckner, Denise Head, Jamie Parker, Anthony F. Fotenos, Daniel Marcus, John C. Morris, Abraham Z. Snyder, A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume, NeuroImage, Volume 23, Issue 2, 2004, Pages 724-738,

[16] Florian P. Kolb, Dieter F. Kutz, Jana Werner, et al. Stimulus‐dependent deliberation process in left‐ and right‐handers obtained via current source density analysis. Physiological Reports, 2022, 10(23)