August 2023

Genetic Associations of Alzheimer's Disease and Mild Cognitive Impairment

Scott Hebert
University of Massachusetts Amherst

Genetic Associations of Alzheimer's Disease and Mild Cognitive Impairment

A Thesis Presented

by

SCOTT HEBERT

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

May 2023

Biostatistics and Epidemiology

Genetic Associations of Alzheimer's Disease and Mild Cognitive Impairment

A Thesis Presented

By

SCOTT HEBERT

Approved as to style and content by:

_____
Zhengqing Ouyang, Chair

_____
Chi Hyun Lee, Member

_____
Jing Qian, Member

_____
Lisa Chasan-Taber, Department Head
Biostatistics and Epidemiology

ACKNOWLEDGMENTS

I would like to thank Dr. Zhengqing Ouyang. The dedication, patience, and help he has shown me throughout this process is overwhelming. This work was made possible by his help.

I would also like to thank Dr. Eric Pederson for his help on this thesis. From providing practical advice to helping me understand Bash shell scripts, his assistance is greatly appreciated.

Next, I would like to thank the other members of the Ouyang Lab aside from the two aforementioned members: Siliangyu Cheng, Bejamin Goldberg, Xinyang Liu, Weidong Wang, and Jacob Yun. Their help and collaboration throughout this process has been an immense help.

A thanks is also due, of course, to the members of my thesis committee, Dr. Chi Hyun Lee and Dr. Jing Qian. In the classroom and beyond, they have helped me improve my knowledge and skills in the realm of biostatistics.

Also, I would like to thank the Graduate Program Director for Biostatistics, Dr. Sofija Zagarins. Her teachings during my undergraduate education are part of what motivated me to pursue the field of biostatistics, and her help as Graduate Program Director has guided me during this process.

I would also like to thank Tonya Menard, for whose help with the administrative side of my coursework and beyond at the University I cannot express enough gratitude.

Lastly, I would like to thank my loved ones. Your support and encouragement helped me in ways I cannot describe. Thank you.

ABSTRACT

GENETIC ASSOCIATIONS OF ALZHEIMER'S DISEASE AND MILD COGNITIVE IMPAIRMENT

MAY 2023

SCOTT HEBERT, B.S., SPRINGFIELD COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Zhengqing Ouyang

Over 6 million people are estimated to have been living with Alzheimer's Disease (AD) in 2020, with another 12 million living with Mild Cognitive Impairment (MCI). Research has been conducted to evaluate genetic links to AD, but more research is needed on the subject. The Alzheimer's Disease Neuroimaging Initiative (ADNI) has been conducting a longitudinal study of AD and MCI since 2004 and offering their data to research teams around the world. Diagnostic and demographic data was collected from participants, as well as data regarding single nucleotide polymorphisms (SNPs). SNP data was transformed to a binary format regarding whether the SNP contained the alternative allele for that particular SNP. We performed cross-validation to determine the ideal alpha and lambda values to use in elastic net regularization, which called for LASSO regression, in order to perform feature selection on the SNPs and other predictors, which were systolic and diastolic blood pressure, age, gender, years of education, race, marital status, and handedness. The LASSO regression reduced the number of SNPs from 55,106 to 13 and removed all non-SNP predictors except years of education and marital status.

We used simple logistic regression to assess the relationship between variations in the significant SNPs (as well as years of education and marital status) and diagnosis of AD/MCI, utilizing a separate LASSO regression with conditional selective inference to more accurately calculate the significance of the variables. The adjusted odds ratios for the SNPs are 1.59 (95% CI 1.23, 2.05), 2.37 (95% CI 1.81, 3.12), 0.71 (95% CI 0.54, 0.93), 1.59 (95% CI 1.21, 2.09), 0.55 (95% CI 0.38, 0.79), 2.03 (95% CI 1.27, 3.23), 0.31 (95% CI 0.18, 0.50), 0.43 (95% CI 0.30, 0.60), 0.69 (95% CI 0.53, 0.89), 1.95 (95% CI 1.46, 2.60), 1.89 (95% CI 1.22, 2.90), 1.47 (95% CI 1.13, 1.90), and 0.52 (95% CI 0.37, 0.72) for SNPs rs11086694, rs2075650, rs2094277, rs2261682, rs31887, rs4745514, rs4816158, rs4826619, rs6640551, rs6809370, rs7312407, rs919751, and rs9857853, respectively. The SNPs are located in genes that have clinical significance and may be associated with various diseases that affect cognitive performance. The results propose that the alternative alleles for seven SNPs are associated with an increased risk of Alzheimer's Disease/Mild Cognitive Impairment diagnosis while six SNPs are associated with a decreased risk of diagnosis. This research may have clinical implications and should be further studied.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Impact on Public Health

More than 6 million people are estimated to have been living with Alzheimer's Disease (AD) in 2020 in the United States, and that prevalence is expected to more than double by 2060. Also, an estimated 12 million people had been living with Mild Cognitive Impairment (MCI) in 2020, with a projection of close to 22 million people living with MCI in 2060.[1]

Among Americans aged 65 and older, over 11% had clinical AD in 2020.[1] With such increases in prevalence projected, the burden on disease sufferers, their families and caretakers, and the health care system may also increase.

1.2 Social and Financial Impact

For those individuals and families who care for loved ones with Alzheimer's Disease, providing such care can lead to some spiritual growth, but it is largely categorized by exhaustion and burnout, and some caregivers may neglect disease sufferers at times. Carers also sometimes experience role conflict, mainly with their occupations, which can lead to job loss.[2]

In addition to social stressors, AD can also cause significant financial strain. In the United States in 2020, the estimated cost of dementia health care was $305 billion. The cost of caring for individuals with dementia is expected to increase to over $1.1 trillion dollars in 2050.[3]

1.3 Physiological Mechanisms

Several theories exist in terms of causes of Alzheimer's Disease. The most studied theory is the amyloid-beta (Aβ) hypothesis. According to this, Aβ, which forms from the breakdown of amyloid precursor protein (APP) in the brain, clumps to form plaques. These plaques damage neurons, specifically causing dendritic spine loss[4] (the damage of the spines of neuronal dendrites).[5] However, the timeframe from the formulation of Aβ plaques to the presence of AD symptoms is more than 20 years.[3]

Another theory of the cause of AD is the tau protein hypothesis. Tau protein, usually involved in the stabilization of microtubules in neurons, can become hyperphosphorylated, causing it to lose its shape and form into Paired Helical Filaments,[6] leading to disruption in neuronal microtubules. This disruption causes neurons to starve, leading to neuron death.[4] This starvation occurs because the disrupted microtubules are involved in the transport of various items within neurons.[6]

Inflammation is also a potential cause of Alzheimer's Disease. The aforementioned accumulation of Aβ plaques can cause the activation of microglia in the brain, leading to inflammation.[7]

Alzheimer's Disease may also have genetic causes. The strongest known genetic risk factor involves the Apolipoprotein E gene. The E4 allele, known as APOE4, is thought to influence Aβ plaques, abnormal tau protein tangles, and inflammation of the brain. However, the APOE2 allele decreases risk of AD.[8]

Lack of blood flow in the brain may also play a role in Alzheimer's Disease. People with AD have a reduced level of cerebral blood flow. Also, for individuals who have the APOE4 allele, a reduction of blood flow has been seen. This lack of blood flow

can also hamper the removal of Aβ plaques, specifically on endothelial cells,[9] potentially furthering issues.

Alzheimer's Disease is the most common cause of dementia.[8] The disease is characterized by memory loss, difficulty with problem-solving, delusions, and performing basic tasks. People with AD may also experience behavioral issues, and the late stages of the disease may cause an inability to communicate and seizures.[10]

Mild Cognitive Impairment may include symptoms of cognitive decline, but the symptoms do not significantly impact the sufferer's basic functioning.[11] MCI may lead to Alzheimer's Disease, but some individuals with MCI will never develop AD.[10,11]

1.4 Alzheimer's Disease Neuroimaging Initiative-Driven Prior Research

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal study with the goal of improving early detection of AD through the study of imaging, genetic, and other data collected from participants throughout the United States and Canada. ADNI receives public and private funding and has been ongoing since 2004. The first phase of the study, ADNI-1, was conducted from 2004 to 2009 with the goal of developing biomarkers that could be used as outcomes for clinical trials and included 800 participants. [12]

ADNI-GO was a two-year extension of the project started in 2009, and ADNI-2 was a 5-year extension of the project started in 2011. ADNI-GO added 200 new participants to the study, and ADNI-2 added another 550 participants, with both also including participants from previous study versions. ADNI-GO sought to examine biomarkers in early disease stages, and ADNI-2 sought to use biomarkers as predictors of disease. ADNI-3 was a 5-year extension started in 2016, though its results have not all

been released. It added another 371 participants and began studying tau scans for use in clinical trials.[12]

The study overall includes 1,921 participants: 483 healthy elderly control participants; 1,001 participants with MCI; and 437 participants with AD. Participants were followed over time to study various diagnostic measurements, genetic testing, and various scans, including cognitive tests, magnetic resonance imaging (MRI) scans, and positron emission tomography (PET) scans.[13]

Over 3,700 articles have been published that use ADNI data.[14] This research has indicated various associations with biomarkers and AD and MCI diagnosis. One study found potential associations between multiple mitochondrial single nucleotide polymorphisms (SNPs) and AD.[15] Another study found that, using ADNI and another data source, the medial temporal lobe was typically the first location of issues regarding tau protein due in part to higher expression of certain genes (including APOE) in that brain region.[16]

1.5 Goals of Research

Current testing for AD (through PET scans and cerebrospinal fluid tests to test for Aβ, for example) is expensive, have the potential to produce adverse effects, and are not widely available.[3] Better, easier testing for AD and MCI would therefore be beneficial for providing more widespread screening. Also, since many clinical trials involving participants who already have clinical AD have led to poor results, it would be beneficial to conduct testing and provide treatment as early as possible.[3] This provides motivation for genetic testing, as it can be done through easier testing (ADNI mainly used peripheral blood tests)[17] and can be done at any time.

APOE, as mentioned, is one potential risk factor involved in AD diagnosis. However, there is potential for more genes to be involved in AD and MCI risk. This research seeks to examine relationships between various single nucleotide polymorphisms and diagnosis of Alzheimer's Disease or Mild Cognitive Impairment, with further review being conducted on genes that contain any associated SNPs.

CHAPTER 2

METHODS

2.1 Study Design

In ADNI-1 and ADNI-GO/2, in addition to various scans and biospecimen collection, genetic data was collected through blood samples (ADNI-GO and ADNI-2 had a combined genetic data collection period, with no participant overlap). These samples were sequenced using Illumina sequencing, though ADNI-1 used the Illumina Human610-Quad BeadChip while ADNI-GO/2 used the Illumina HumanOmniExpress BeadChip for genotyping.[18] Each ADNI version sequenced over 600,000 genetic markers,[19] including single nucleotide polymorphism (SNP) and copy number variation (CNV) data. Data regarding chromosome and position, B allele frequency, and other data about SNPs and CNVs were collected. In ADNI-1, participants attended a screening visit, a baseline visit, and follow-up visits at 6, 12, 18, 24, and 36 months post-baseline visit.[20] In ADNI-2, new patients attended a screening visit, a baseline visit, and follow-up visits. For controls and patients with MCI, follow-up visits were conducted at 3 and 6 months post-baseline visit and every 6 months thereafter. For patients with AD, follow-up visits were conducted at 3, 6, 12, 18, and 24 months post-baseline visit and every 6 months thereafter.[21]

Diagnostic data were collected through various cognitive tests, including memory tests, the Boston Naming Test, and others, and diagnostic summaries were created for patients throughout the study and categorized as being normal controls, having MCI, or having AD.[20,21] In ADNI-2, MCI was subcategorized into early and late MCI.[21]

Genetic data was collected for 1,550 participants, 757 from ADNI-1 and 793 from ADNI-GO/2. Diagnostic data was collected from 2,920 participants, including ADNI-3. Genetic data from ADNI-3 has not yet been released.

2.2 Participant Inclusion and Exclusion Criteria

Participants were required to be between the ages of 55 and 90 at baseline have a Geriatric Depression Scale score of less than 6, have a study partner who accompanies them to visits, have proper visual and auditory acuity, good general health, completed at least six grades of education or a sufficient work history, agree to collect blood and other samples for testing, and several other requirements. Further requirements were placed upon participants based on their categorization of diagnosis. For example, normal controls must have been free of memory complaints.[20,21]

Certain exclusion criteria were put in place throughout the study. In ADNI-1, the main criteria for exclusion centered around specific medications, including certain antidepressants and analgesics.[20] In ADNI-2, the exclusion criteria included diagnosis of certain conditions, including major depression, as well as a history of alcohol or substance abuse.[21] There were various other criteria for exclusion throughout the study.

2.3 Data Collection and Manipulation

Permission to access and download ADNI study data was obtained through ADNI directly via an electronic application. Data containing SNP and CNV information were downloaded from the ADNI website from https://ida.loni.usc.edu/pages/access/geneticData.jsp, as well as data containing information regarding patient diagnostic and demographic data, including blood pressure, age at the baseline visit, gender, years of education, race, marital status, and handedness.

Genetic data was separated by ADNI version, with one genome-wide association study (GWAS) being held for ADNI-1 and another being held for ADNI-GO/2. Genetic data were filtered to only include SNP data and were used in A/B notation, with the A allele being the reference allele. The genetic data were tidied, as they were originally organized as separate files for each participant, with each SNP being a separate row. The tidying of the files was completed using Bash shell scripts in a Unix environment. The tidying included combining the files for each participant into two files (one for ADNI-1 and one for ADNI-GO/2), selecting and copying each relevant variable from the new files, combining the relevant variables into cohesive files (this was done because typical filtering was computationally inefficient in comparison to this method), and pivoting the data so that each participant only was included in one row, and each SNP was a distinct variable. The data from ADNI-1 and ADNI-GO/2 were initially kept separate due to their large file size, but they were later combined after being filtered to only include relevant data, keeping only SNPs that were present in both versions of ADNI. Also, data regarding the chromosome and position (location on the relevant chromosome) of each SNP was extracted from the data and kept separately for reference after analysis was completed in order to determine the genes any relevant SNPs were located on or near. Also, SNPs were initially transformed into a numerical representation depending on the number of B alleles (i.e., “AA,” “AB,” and “BB” were transformed into 0, 1, and 2, respectively) in order to reduce file size and increase computational efficiency when handling the data.

The genetic data was then combined with the other demographic and diagnostic data. The various data files used one of two (or, in some cases, both) versions of the

patient ID number, which were matched, confirming the matching method with ADNI personnel via email communication. The most recent available diagnosis for each participant was used. SNPs were transformed into a binary format, based on the presence or absence of at least one B allele. SNPs with less than 5% variability (i.e., greater than 95% of participants having the same version of the binary SNP) were removed from the data, and only complete cases were kept in the data set, meaning any variables, genetic or otherwise, with any missing values were removed from the data.

Some demographic variables were transformed to a binary format due to a lack of variability of the data, with some factor levels having very few options. Specifically, race was made binary (white versus non-white), as was marital status (married versus unmarried). Also, diagnosis was transformed into a binary variable as well, with the presence of AD or MCI being considered a "1" and the absence of either diagnosis being considered a "0" in order to simplify the interpretability of results and allow for logistic regression to be performed.

After the genetic data were combined with the diagnostic and demographic data, only 1,465 patients remained, as an inner join was used so as to have complete data for the demographic and diagnostic variables. However, the loss of patients compared to the full genetic data set was only 85 patients, or 5% of those participants. Also, after the binary SNPs were filtered based on B allele frequency, 55,106 SNPs remained.

In order to perform conditional selective inference on the data, duplicate columns needed to be removed from the data. This removed 66 duplicated SNPs, leaving 55,040 to be used for that comparative method.

2.4 Data Analysis

The data were first viewed based on a breakdown of participant demographic and diagnostic data. Then, elastic net cross-validation was conducted on the data in order to choose the optimal α and λ values for elastic net regression and, thus, variable selection. Based on the results of the cross-validation, LASSO regression was performed on the data, selecting a subset of the SNPs and other predictors, specifically 13 SNPs, years of education, and marital status. Simple regression was then performed on that variable subset with diagnosis as the outcome variable. In order to calculate more accurate p-values for the predictors, a separate LASSO regression was conducted utilizing conditional selective inference. This method involved re-running the data through elastic net cross-validation (since there were 66 fewer SNPs) and utilizing those results to perform a separate LASSO regression that performed conditional selective inference (which calculated more accurate and less biased p-values). Then, a test for multicollinearity was performed on the model. Adjusted odds ratios and their corresponding 95% confidence intervals were calculated. Lastly, the genes in or near which the SNPs were located were determined using the dbSNP database and the bedtools utility software.

Aside from the initial data tidying using Bash scripts and the use of the bedtools software to determine some SNP gene locations (and some mutations and other minor data adjustments using Microsoft Excel), all data manipulation and analysis was performed using R version 4.2.0 with RStudio version 2022.07.2.

CHAPTER 3

RESULTS

3.1 Participant Demographics

The basic participant demographic data are shown in Table 1 (Appendix). Of the 1,465 participants included in the study, 56.7% were male and 43.3% were female. 93.0% of participants were white, while the rest of the participants, 7.0%, were non-white. 76.0% were married while the remaining 24.0% were unmarried. 91.1% were right-handed while 8.9% were left-handed. The majority of patients had a diagnosis of either AD or MCI, specifically 72.4%. 27.6% of participants were normal controls.

The mean age of patients was 73.7 years, with a range from 54.4 years to 91.4 years. On average, participants received 15.9 years of education, with a range from 4 years to 20 years. Average systolic blood pressure was 134.9, with a range of 83 to 201. For diastolic blood pressure, the mean was 74.4, with a range of 43 to 108.

Demographic data broken down by diagnosis are shown in Table 2 (Appendix). Diagnosis of AD or MCI was higher among males (75.7%) compared to females (68.0%). Diagnosis was also higher among white participants (73.1%) compared to non-white participants (62.1%). Married participants had a higher percentage who were diagnosed with AD or MCI (75.1%) compared to non-married participants (63.5%). Also, diagnosis was more common among right-handed participants (72.8%) compared to left-handed participants (67.9%).

The means of age, years of education, systolic blood pressure, and diastolic blood pressure were all very similar among diagnosed participants and controls, with their mean values all being within roughly one unit when comparing between the two groups.

3.2 Elastic Net Cross-Validation and LASSO Regression

Elastic net cross-validation was conducted on the data, excluding participant ID as a variable. Multiple values of α were tested, as well as multiple values of λ. This was done to minimize the cross-validation error of the data when performing elastic net regression for variable selection. The ideal value of α was 1, which pointed to performing LASSO regression on the data. The ideal value of λ was 0.0412638. Results of the elastic net cross-validation are illustrated in Figure 1.

Then, LASSO regression was run on the full data set, again excluding participant ID as a variable. The ideal values of α and λ obtained from elastic net cross-validation were used. The LASSO regression left 13 SNPs and 2 other predictors and reduced the coefficients of all other predictors to 0. The 13 remaining SNPs were rs11086694, rs2075650, rs2094277, rs2261682, rs31887, rs4745514, rs4816158, rs4826619, rs6640551, rs6809370, rs7312407, rs919751, and rs9857853. The other remaining predictors were years of education and marital status.

3.3 Logistic Regression and Associations

Simple logistic regression was then performed on the 15 predictors remaining after LASSO regression with the diagnosis of AD or MCI as the outcome variable. Adjusting for all other predictors included in the logistic regression model, the adjusted odds ratios were 1.59 (95% CI 1.23, 2.05), 2.37 (95% CI 1.81, 3.12), 0.71 (95% CI 0.54, 0.93), 1.59 (95% CI 1.21, 2.09), 0.55 (95% CI 0.38, 0.79), 2.03 (95% CI 1.27, 3.23), 0.31 (95% CI 0.18, 0.50), 0.43 (95% CI 0.30, 0.60), 0.69 (95% CI 0.53, 0.89), 1.95 (95% CI 1.46, 2.60), 1.89 (95% CI 1.22, 2.90), 1.47 (95% CI 1.13, 1.90), and 0.52 (95% CI 0.37, 0.72) for rs11086694, rs2075650, rs2094277, rs2261682, rs31887, rs4745514,

rs4816158, rs4826619, rs6640551, rs6809370, rs7312407, rs919751, and rs9857853, respectively. The adjusted odds ratio of years of education was 0.87 (95% CI 0.83, 0.91), and the adjusted odds ratio of marital status (with being married as the reference level) was 0.67 (95% CI 0.50, 0.89). The adjusted odds ratios and their corresponding 95% confidence intervals are shown in Table 3 (Appendix).

Using the predictors found after LASSO regression to inform simple logistic regression is a naïve approach that can lead to biased (artificially low) p-values regarding the observed associations. This is why a separate regression was performed to calculate more accurate p-values utilizing conditional selective inference, and conditional selective inference conditions the inference on the LASSO variable selection, which allows for more valid p-values.[22] This method's cross-validation led to the LASSO regression including 12 extra predictors (all SNPs), which may be due to the data set including 66 fewer SNPs initially, which could affect the results. Thus, those extra SNPs can be ignored for these purposes, especially considering their high p-values, some of which were close to 1. The p-values obtained through simple logistic regression and the p-values obtained through conditional selective inference can be found in Table 4 (Appendix).

A check for multicollinearity was performed on the simple logistic regression model, using the calculation of the variance inflation factor (VIF) of each model predictor. After this test, the VIF of each model predictor was close to 1, indicating that there is no multicollinearity in the model. The results of the check for multicollinearity can be found in Table 5 (Appendix).

3.4 Interpretation of Genetic Predictors

The model found significant associations between the 13 SNPs, years of education, and marital status. However, the presence of B allele in the SNPs rs11086694, rs2075650, rs2261682, rs4745514, rs6809370, rs7312407, and rs919751 were associated with an increased risk of AD or MCI while the presence of the B allele in the SNPs rs2094277, rs31887, rs4816158, 4826619, rs6640551, and rs9857853 were associated with a decreased risk of AD or MCI. An increase in years of education was associated with a decreased risk of diagnosis. Lastly, being married was associated with an increased diagnosis risk (or, conversely, being unmarried was associated with a decreased risk).

In order to determine the genes in which the SNPs were located, dbSNP was used,[23] except for rs11086694 and rs2094277, which required the use of the bedtools software. For bedtools, the GRCh37 locations of the SNPs were used, as that is what ADNI data provided. Also, data regarding gene function, gene expression, and potential gene disease associations were obtained through the National Center for Biotechnology Information's Gene database.[24]

The SNP rs2075650 is located in chromosome 19 in the TOMM40 gene.[25] The TOMM40 gene has the function of importing protein precursors into mitochondria.[26] With the relevance of amyloid precursor protein in Alzheimer's Disease, as mentioned before, this SNP may play a role in the disease. This is further exemplified by previous research finding an association between rs2075650 and AD.[27] This association was found to be significant with conditional selective inference, so it should be taken more seriously.

Rs4816158 is located in chromosome 20 in the PAK5 gene.[28] This gene induces microtubule stabilization, promotes neurite growth, and regulates cytoskeleton dynamics. It is also mostly expressed in the brain.[29] Considering the impact the tau protein and microtubule stabilization in the development of Alzheimer's Disease (coupled with this association being found to be significant through conditional selective inference), this SNP and its gene are particularly in need of further study.

Rs6640551, located in the X chromosome, is in the gene known as SHROOM2.[30] This gene functions in the formation of new blood vessels and the formation of contractile networks in endothelial cells. It is associated with ocular albinism type 1 syndrome.[31] Given the association between AD and the loss of blood flow, a gene that controls the formation of new blood vessels is of interest.

Rs6809370 is located on chromosome 3 and is in the PARL gene.[32] It is involved in mitochondrial remodeling and apoptosis, and it is has a potential association with Parkinson's disease.[33] Rs31887 is in chromosome 5 in the CTNND2 gene,[34] which is involved in brain and eye development and is expressed mostly in the brain.[35] The CTNND2 gene is of interest due to its function in brain development, and due to the rs6809370 SNP having a significant association under conditional selective inference.

The gene abbreviated as PRKACG is the location of the SNP rs4745514, which is located on chromosome 9.[36] This gene encodes the gamma form of one of its subunits.[37] The SHROOM4 gene houses the SNP rs4826619 on the X chromosome[38] (the same chromosome as the SHROOM2 gene) and may be involved in cytoskeletal architecture,[39] similarly to the PAK5 gene. Rs4826619 was found to be significant through the conditional selective inference method used.

Similarly, rs919751, located in chromosome 5 (the same chromosome as rs31887) and part of the PDGFRB gene,[40] is involved in actin cytoskeleton, but it is also involved in the development of the cardiovascular system. It is potentially associated with 5q- syndrome,[41] which is a condition that affects bone marrow cells and lead to a form of anemia.[42] The associations regarding the cardiovascular system this SNP may have make it of interest for further study.

Rs2261682 is located in chromosome 2 in the AMMECR1L gene.[43] This gene is expressed fairly evenly throughout most human tissues, but its highest expression is in testis tissue.[44] This gene is similar to the AMMECR1 gene, which has an unknown function. However, the AMMECR1 gene is associated with Amme Complex,[45] which is shorthand for a condition that includes Alport syndrome, intellectual disability (the second "M" in the abbreviation used to stand for "mental retardation," though the abbreviation has not changed with the use of "intellectual disability"), midface hypoplasia, and elliptocytosis syndrome.[46] Elliptocytosis has to do with red blood cells being shaped as ellipses instead of having a round shape. The typical round shape of red blood cells is held in part by cytoskeleton proteins.[47] The cardiovascular implications of this gene and the similarities among it and the PAK5 and SHROOM4 genes regarding cytoskeletal structure make it worthy of further investigation.

The aforementioned genes have all been protein coding genes. However, the model included 4 non-coding RNA genes. Rs11086694 is located in chromosome 20 (the same as rs4816158) in or near the LINC02910 gene, and rs2094277 is in chromosome 13 in the LINC00347 gene. Though not much is known about these genes, LINC02910 is

expressed mostly in bone marrow,[48] which draws a vague link between it and the PDGFRB gene. Also, LINC00347 is expressed almost entirely in testis tissue.[49]

Rs7312407 is located in chromosome 12 in another non-coding RNA gene, LINC01479.[50] This gene is expressed mostly in the heart,[51] and, given the aforementioned impact the cardiovascular system plays on the development of AD, could make it a significant gene to further study. Rs9857853 is located in chromosome 3 (the same as rs6809370) and is located in the non-coding RNA gene LOC 105374313,[52] which is expressed mostly in testis tissue.[53]

Despite the unknown functions of non-coding RNA genes, they may play a role in gene expression regulation and have potential impacts on AD.[54] Thus, they should not be discounted in their relationships to AD due to the lack of knowledge regarding their functions.

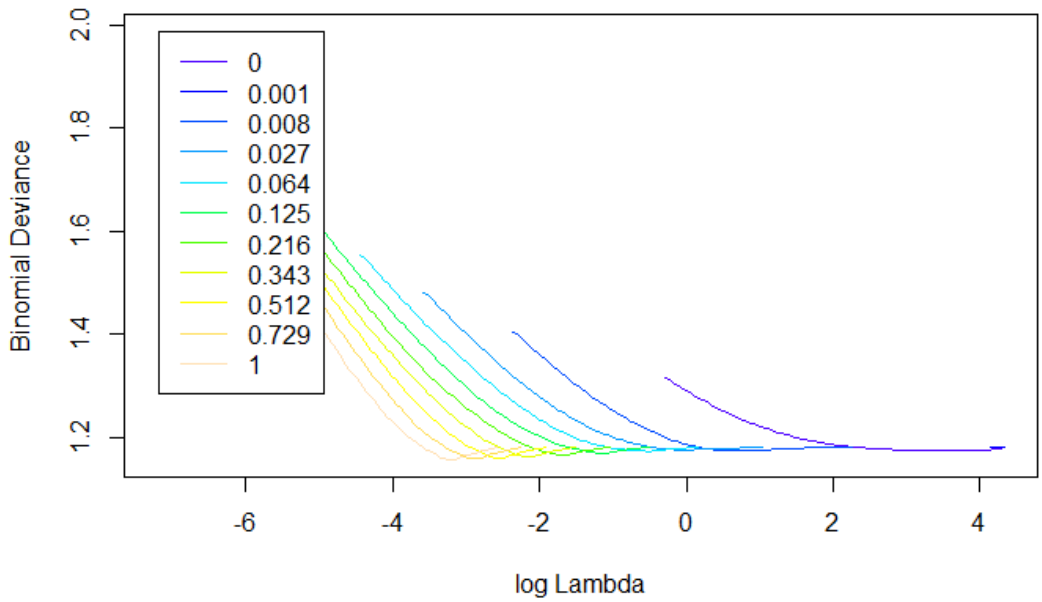
3.5 Interpretation of Non-Genetic Predictors

Years of education was significantly associated with a decreased risk on diagnosis of AD or MCI, based on the simple logistic regression model. While this aligns with current research, which states that higher levels of education are causally associated with reduced risk or delayed onset of AD,[55] although some research skeptically states that the relationship between education and AD may be affected by intelligence,[56] it is difficult to interpret the adjusted odds ratio of the variable, due to its continuous nature. However, since this association was found to be significant after performing conditional selective inference, this association should be further studied.

Marital status was associated with AD, with unmarried people being at a lower risk of diagnosis compared to married people. However, this is in stark contrast to

existing literature. One study found that unmarried people were at a greater risk of developing dementia compared to married individuals.[57] The reason for the simple logistic regression model giving the opposite finding is unknown. It is possible that marital status is a confounder for another variable not shown in the data, or that the sample was not representative in terms of marital status and diagnosis. It also may have been due to the transformation of the variable into a binary format. Regardless, this association, having been found to not be significant using the conditional selective inference method, this association should be viewed warily.

Figure 1: Results of elastic net cross-validation



CHAPTER 4

DISCUSSION

While ADNI was a thoroughly conducted study, it and the research shown here have limitations. ADNI may suffer from selection bias, due to the fact that some participants who otherwise would have developed AD or MCI could have died or left the study before being diagnosed or showing symptoms. Also, since late-onset Alzheimer's Disease is defined as having its onset be at age 65 or older,[58] patients who entered ADNI at age 55 and only stayed in the study for under 10 years would not have even been old enough to develop late-onset Alzheimer's Disease, further adding to the selection bias for participants who may have developed AD later in life, even if they survived long enough to develop symptoms and be diagnosed, as they would have no longer been in the study.

Alzheimer's Disease disproportionately affects minority groups compared to white individuals in terms of frequency.[3] Given this, the ADNI study having overwhelmingly white participants (the data here included 93% white participants) points to potential issues with generalizability, as the sample does not seem to be representative in that regard.

One minor issue in the data is that some of the ages and years of education were slightly outside of the range specified in the inclusion criteria, although those values were not far from that range. The age range was not violated by more than 1.4 years in either direction, and the education range was only violated by 2 years on the lower end. This does not pose a significant threat to the validity of the data, but it may be problematic to some degree.

The education variable (measured in years) was a bit left-skewed, but due to its uneven nature (most people seemed to complete either 0, 2, 4, or 6, or 8+ years of higher education), most transformations of the data were not effective at reshaping it into a normal distribution. This posed a potential issue with the data, as, while logistic regression does not require normality of predictors, it does require continuous predictors to be linear in relation to their logits.[59] However, after performing a Box-Tidwell test, which is an appropriate method of checking that assumption,[60] on the education variable, the result showed that the education variable did not violate that assumption (p-value = 0.83, showing a small chance of violating the assumption).

The binary transformations involved throughout this process, while allowing for simpler interpretation, inherently also lead to some loss of information. For example, the distinction between AD and MCI is lost in this analysis, though it is important. Further research into this topic could include multinomial regression. Further research could also view the distinction between SNPs having one copy versus two copies of the B allele of each SNP.

As mentioned previously, the conditional selective inference method led to a dozen extra predictors being included after the comparative LASSO regression. While this is likely due to some of the SNPs being removed from the data to allow for the method to be conducted, this does have an impact on the resulting p-values for the SNPs included in the initial LASSO and simple logistic regressions. However, these p-values are still likely more valid in comparison to the p-values conducted directly through the logistic regression method (and any that were conducted through the initial LASSO regression, as those p-values would be inherently biased).

Given the detrimental impact of Alzheimer's Disease on the physical, emotional, and financial aspects of the lives of sufferers and their caregivers as well as the impact on the health care system in the United States and beyond, and given the need for early, accurate, and inexpensive testing, genetic associations of Alzheimer's Disease and Mild Cognitive Impairment should be further studied. This research points mostly to a need to further research of the SNP rs2075650 and its associated TOMM40 gene and the SNP rs4816158 and its associated PAK5 gene, due to the functions of those genes that have the potential to significantly impact AD risk. The PAK5 gene is of particular interest, as it has not been previously researched in its relationship with AD/MCI risk, with this research being the first to detect such an association. Further study of these and other potential associations could lead to improved early detection and, thus, early treatment of Alzheimer's Disease, and could reduce the burden of the disease.

APPENDIX

TABLES

Table 1: Participant demographic data

Categorical variables	N	%
Gender		
Female	635	43.3%
Male	830	56.7%
Race		
White	1362	93.0%
Nonwhite	103	7.0%
Marital Status		
Married	1114	76.0%
Unmarried	351	24.0%
Handedness		
Right-handed	1334	91.1%
Left-handed	131	8.9%
Diagnosis		
AD or MCI	1060	72.4%
None/control	405	27.6%
Continuous variables	**Mean**	**Range**
Age	73.7	(54.4, 91.4)
Education (years)	15.9	(4, 20)
Blood pressure (systolic)	134.9	(83, 201)
Blood pressure (diastolic)	74.4	(43, 108)

Table 2: Participant demographic data, by diagnosis

Categorical variables	AD or MCI diagnosis		Normal control	
	n	%	n	%
Gender				
Female	432	68.0%	203	32.0%
Male	628	75.7%	202	24.3%
Race				
White	996	73.1%	366	26.9%
Nonwhite	64	62.1%	39	37.9%
Marital Status				
Married	837	75.1%	277	24.9%
Unmarried	223	63.5%	128	36.5%
Handedness				
Right-handed	971	72.8%	363	27.2%
Left-handed	89	67.9%	42	32.1%
Continuous variables	**Mean**	**Range**	**Mean**	**Range**
Age	73.9	(54.4, 91.4)	73.1	(55.0, 89.3)
Education (years)	15.7	(4, 20)	16.5	(6, 20)
Blood pressure (systolic)	135.2	(86, 201)	134.1	(83, 192)
Blood pressure (diastolic)	74.5	(43, 108)	74.3	(49, 100)

Table 3: Adjusted odds ratios and corresponding 95% confidence intervals

Predictor	Adjusted Odds Ratio	95% Confidence Interval
rs11086694	1.59	(1.23, 2.05)
rs2075650	2.37	(1.81, 3.12)
rs2094277	0.71	(0.54, 0.93)
rs2261682	1.59	(1.21, 2.09)
rs31887	0.55	(0.38, 0.79)
rs4745514	2.03	(1.27, 3.23)
rs4816158	0.31	(0.18, 0.50)
rs4826619	0.43	(0.30, 0.60)
rs6640551	0.69	(0.53, 0.89)
rs6809370	1.95	(1.46, 2.60)
rs7312407	1.89	(1.22, 2.90)
rs919751	1.47	(1.13, 1.90)
rs9857853	0.52	(0.37, 0.72)
Education	0.87	(0.83, 0.91)
Marital Status	0.67	(0.50, 0.89)

Note: The reference level for marital status is "Married."

Table 4: P-values, before and after conditional selective inference

Predictor	β (Coefficient) (Logistic Regression)	P-value (Logistic Regression)	P-value (Conditional Selective Inference
rs11086694	0.46119	<0.001	0.442
rs2075650	0.86299	<0.001	<0.001
rs2094277	-0.34303	0.013	0.709
rs2261682	0.46152	<0.001	0.473
rs31887	-0.60574	0.001	0.192
rs4745514	0.70905	0.002	0.573
rs4816158	-1.18076	<0.001	0.026
rs4826619	-0.84584	<0.001	0.003
rs6640551	-0.37642	0.004	0.504
rs6809370	0.66735	<0.001	0.023
rs7312407	0.63408	0.004	0.541
rs919751	0.38252	0.004	0.517
rs9857853	-0.65483	<0.001	0.259
Education	-0.13689	<0.001	0.003
Marital Status	-0.40626	0.005	0.372
rs10942262	(Not included in original LASSO regression or logistic regression)		0.513
rs12422895			0.557
rs1866361			0.844
rs1873442			0.871
rs323467			0.918
rs4936046			0.887
rs4977761			0.670
rs714180			0.909
rs720202			0.754
rs7214481			0.687
rs7428265			0.970
rs7933268			0.621

Note: The reference level for marital status is "Married."

Table 5: Variance inflation factors of model predictors

Predictor	Variance Inflation Factor
rs11086694	1.010628
rs2075650	1.017578
rs2094277	1.018341
rs2261682	1.017602
rs31887	1.014424
rs4745514	1.014226
rs4816158	1.016831
rs4826619	1.033316
rs6640551	1.032130
rs6809370	1.019630
rs7312407	1.015184
rs919751	1.022981
rs9857853	1.017855
Education	1.036740
Marital Status	1.023450

Note: The reference level for marital status is "Married."

BIBLIOGRAPHY

1 Rajan, K. B., Weuve, J., Barnes, L. L., McAninch, E. A., Wilson, R. S., & Evans, D. A. (2021). Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimer's & dementia : the journal of the Alzheimer's Association*, *17*(12), 1966–1975. https://doi.org/10.1002/alz.12362

2 Ashrafizadeh, H., Gheibizadeh, M., Rassouli, M., Hajibabaee, F., & Rostami, S. (2021). Explain the experience of family caregivers regarding care of Alzheimer's patients: a qualitative study. *Frontiers in psychology*, *12*, 699959.

3 Wong, W. (2020). Economic burden of Alzheimer disease and managed care considerations. *The American journal of managed care*, *26*(8 Suppl), S177-S183.

4 Kocahan, S., & Doğan, Z. (2017). Mechanisms of Alzheimer's Disease Pathogenesis and Prevention: The Brain, Neural Pathology, N-methyl-D-aspartate Receptors, Tau Protein and Other Risk Factors. *Clinical psychopharmacology and neuroscience : the official scientific journal of the Korean College of Neuropsychopharmacology*, *15*(1), 1–8. https://doi.org/10.9758/cpn.2017.15.1.1

5 Dorostkar, M. M., Zou, C., Blazquez-Llorca, L., & Herms, J. (2015). Analyzing dendritic spine pathology in Alzheimer's disease: problems and opportunities. *Acta neuropathologica*, *130*(1), 1–19. https://doi.org/10.1007/s00401-015-1449-5

6 Muralidar, S., Ambi, S. V., Sekaran, S., Thirumalai, D., & Palaniappan, B. (2020). Role of tau protein in Alzheimer's disease: The prime pathological player. *International journal of biological macromolecules*, *163*, 1599-1617.

7 Ozben, T., & Ozben, S. (2019). Neuro-inflammation and anti-inflammatory treatment options for Alzheimer's disease. *Clinical biochemistry*, *72*, 87-89.

8 Parhizkar, S., & Holtzman, D. M. (2022, February). APOE mediated neuroinflammation and neurodegeneration in Alzheimer’s disease. In *Seminars in immunology* (p. 101594). Academic Press.

9 Korte, N., Nortley, R., & Attwell, D. (2020). Cerebral blood flow decrease as an early pathological mechanism in Alzheimer's disease. *Acta Neuropathologica*, *140*, 793-810.

10 National Institute on Aging. (2022, October 18). *What are the signs of alzheimer’s disease?* National Institute on Aging. https://www.nia.nih.gov/health/what-are-signs-alzheimers-disease

11 Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., ... & Winblad, B. (2006). Mild cognitive impairment. *The lancet*, *367*(9518), 1262-1270.

12 Alzheimer's Disease Neuroimaging Initiative. (2017). *ADNI | About*. Adni.loni.usc.edu. https://adni.loni.usc.edu/about/

13 Alzheimer's Disease Neuroimaging Initiative. (2017). *ADNI | Study Design.* Adni.loni.usc.edu. https://adni.loni.usc.edu/study-design/#study-objective-container

14 Alzheimer's Disease Neuroimaging Initiative. (2017). *ADNI | Publications.* Adni.loni.usc.edu. https://adni.loni.usc.edu/news-publications/publications/

15 Xu, X., Wang, H., Bennett, D. A., Zhang, Q. Y., Wang, G., & Zhang, H. Y. (2022). Systems Genetic Identification of Mitochondrion-Associated Alzheimer's Disease Genes and Implications for Disease Risk Prediction. *Biomedicines*, *10*(8), 1782.

16 Wen, Q., Yu, M., Shahid, S. S., Zhao, Y., Risacher, S. L., Saykin, A. J., & Wu, Y. C. (2022). Genetic and Structural Network Contributions to The Regional Vulnerability of Tauopathy. *Alzheimer's & Dementia*, *18*, e067455.

17 Alzheimer's Disease Neuroimaging Initiative. (2017). *ADNI | Genetic Data Methods.* Adni.loni.usc.edu. https://adni.loni.usc.edu/methods/genetic-data-methods/

18 Alzheimer's Disease Neuroimaging Initiative. (2017). *ADNI | Genetic Data.* Adni.loni.usc.edu. https://adni.loni.usc.edu/data-samples/data-types/genetic-data/

19 Alzheimer's Disease Neuroimaging Initiative. (2017). *ADNI | METHODS AND TOOLS.* Adni.loni.usc.edu. https://adni.loni.usc.edu/methods/

20 Alzheimer's Disease Neuroimaging Initiative. (2006). *ADNI 1 Procedures Manual*. ADNI. https://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf

21 Alzheimer's Disease Neuroimaging Initiative. (2006). *ADNI 2 Procedures Manual*. ADNI. https://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf

22 Duy, V. N. L., & Takeuchi, I. (2021). More powerful conditional selective inference for generalized lasso by parametric programming. *arXiv preprint arXiv:2105.04920*. https://arxiv.org/abs/2105.04920

23 Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.*, 9, 677–679.

24 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/

25 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs2075650; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs2075650

26 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens translocase of outer mitochondrial membrane 40 (TOMM40); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/10452

27 Potkin, S. G., Guffanti, G., Lakatos, A., Turner, J. A., Kruggel, F., Fallon, J. H., Saykin, A. J., Orro, A., Lupoli, S., Salvi, E., Weiner, M., & Macciardi, F. (2009). Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease. *PLoS ONE*, *4*(8), e6501. https://doi.org/10.1371/journal.pone.0006501

28 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs4816158; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs4816158

29 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens p21 (RAC1) activated kinase 5 (PAK5); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/57144

30 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs6640551; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs6640551

31 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens shroom family member 2 (SHROOM2); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/357

32 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs6809370; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs6809370

33 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens presenilin associated rhomboid like (PARL); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/55486

34 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs31887; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs31887

35 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens catenin delta 2 (CTNND2); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/1501

36 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs4745514; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs4745514

37 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens protein kinase cAMP-activated catalytic subunit gamma (PRKACG); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/5568

38 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs4826619; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs4826619

39 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens shroom family member 4 (SHROOM4); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/57477

40 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs919751; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs919751

41 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens platelet derived growth factor receptor beta (PDGFRB); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/5159

42 National Cancer Institute. (2011, February 2). *5q minus syndrome*. Www.cancer.gov. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/5q-minus-syndrome

43 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs2261682; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs2261682

44 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens AMMECR1 like (AMMECR1L); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/83607

45 GeneCards. (2023). *AMMECR1 Gene - GeneCards | AMMR1 Protein | AMMR1 Antibody*. Www.genecards.org. https://www.genecards.org/cgi-bin/carddisp.pl?gene=AMMECR1

46 National Center for Biotechnology Information. (n.d.). *Alport syndrome-intellectual disability-midface hypoplasia-elliptocytosis syndrome - NIH Genetic Testing Registry (GTR) - NCBI*. Www.ncbi.nlm.nih.gov. Retrieved March 16, 2023, from https://www.ncbi.nlm.nih.gov/gtr/conditions/C1846242/

47 Jha SK, Vaqar S. Hereditary Elliptocytosis. [Updated 2022 Nov 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. https://www.ncbi.nlm.nih.gov/books/NBK562333/

48 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens long intergenic non-protein coding RNA 2910 (ABR); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/284756

49 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens long intergenic non-protein coding RNA 347 (LINC00347); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/?term=linc00347

50 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs7312407; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs7312407

51 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens long intergenic non-protein coding RNA 1479 (LINC01479); [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/?term=linc01479

52 dbSNP [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1998 – Homo sapiens rs9857853; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/snp/rs9857853

53 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – Homo sapiens LOC105374313; [cited 2023 Mar 16]. https://www.ncbi.nlm.nih.gov/gene/?term=LOC105374313

54 Wang, E., Lemos Duarte, M., Rothman, L. E., Cai, D., & Zhang, B. (2022). Non-coding RNAs in Alzheimer's disease: perspectives from omics studies. *Human Molecular Genetics*, *31*(R1), R54-R61.

55 Zhang, X. X., Tian, Y., Wang, Z. T., Ma, Y. H., Tan, L., & Yu, J. T. (2021). The epidemiology of Alzheimer's disease modifiable risk factors and prevention. *The journal of prevention of Alzheimer's disease*, *8*, 313-321.

56 Anderson, E. L., Howe, L. D., Wade, K. H., Ben-Shlomo, Y., Hill, W. D., Deary, I. J., ... & Hemani, G. (2020). Education, intelligence and Alzheimer's disease: evidence from a multivariable two-sample Mendelian randomization study. *International journal of epidemiology*, *49*(4), 1163-1172.

57 Liu, H., Zhang, Z., Choi, S. W., & Langa, K. M. (2020). Marital status and dementia: Evidence from the Health and Retirement Study. *The Journals of Gerontology: Series B*, *75*(8), 1783-1795.

58 Rabinovici, G. D. (2019). Late-onset Alzheimer Disease. *CONTINUUM: Lifelong Learning in Neurology*, *25*(1), 14–33. https://doi.org/10.1212/con.0000000000000700

59 Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, *18*(10), 1099–1104.

60 Shrestha, N. (2019). Application of binary logistic regression model to assess the likelihood of overweight. *Am J Theor Appl Stat*, *8*(1), 18-25.