Exploring the Impact of Demographic Factors, Medical History, and Medicine Usage on Alzheimer's Disease Treatment Outcome

Yunxin Gan

August 28, 2023

Abstract

Background As the global population continues to age, the prevalence of people with Alzheimer's disease (AD) is expected to rise significantly. AD is a progressive neurodegenerative disorder primarily affecting older people and placing a burden on families and societies worldwide. Therefore, there is an urgent need for effective strategies addressing these burdens and ensuring the well-being of ageing populations around the world. This project will explore the impact of demographic factors, medical history, and medicine usage on AD treatment outcome, providing insights for understanding treatment responses, identifying risk factors, improving treatment strategies, and informing healthcare policy-making and decision-making. These insights have the potential to enhance patient care, optimise resource allocation, and ultimately improve AD's overall management.

Methods Data were collected from the Alzheimer's Disease Neuroimaging Initiative, including demographic information, clinical and visit information, medical history, and medicine usage. We used cognitive decline as the indicator to demonstrate the treatment outcome. 0 indicated a negative treatment outcome where the cognitive decline was progressive, and 1 indicated a positive treatment outcome where the cognitive function of the patient remained stable or improved. Initially, logistic regression analysis was applied to identify factors that had a relatively high association with the treatment outcome. Machine learning was then conducted to predict the treatment outcome of AD, using the factors identified from statistical analysis.

Results From the logistic regression analysis, we learnt that age (AGE), independence level (INDPEND), residence type (RESIDENC) and living situation (NACCLIVS), medical history of cognition decline symptoms (NACCOGF), indicator of currently manifesting changes in behaviour (BEAPATHY), donepezil (DONEPEZIL), and memantine (MEMANTINE) were relatively stronger predictors of AD positive treatment outcome. Moreover, the random forest model, in which the subsets of features involved 10 demographic information and medical histories and three donepezil, memantine, and calcium carbonate, achieved the best performance regarding identifying positive treatment outcomes. The precision, recall, and F1-score were 0.62, 0.85, and 0.72, respectively.

Conclusions Older patients tend to have lower chances of positive AD treatment outcome. Additionally, higher independence levels correlate with a greater likelihood of positive outcomes, possibly because increased personal freedom and self-determination benefit patients' well-being. Regarding medical conditions, patients manifesting behaviour changes like social withdrawal are more prone to declined cognitive treatment outcomes. This behaviour might indicate a more advanced AD stage, reducing the likelihood of positive outcomes. Last, donepezil and memantine have a relatively higher association with AD treatment outcome.

Contents

1 Introduction

1.1 Background and Significance of Alzheimer's Disease

Currently, there are more than 55 million people with dementia worldwide, with 60% of them living in low- or middle-income countries. The most common form of dementia is Alzheimer's Disease (AD). According to an article about dementia the World Health Organization [1]. published, dementia is the seventh leading cause of death and one of the main reasons for disability and dependence among elderly people around the world.

There are some burdens associated with AD. First, AD greatly affects quality of life. AD is a progressive disease, meaning that it gets worse with time. Patients' cognitive function may be impaired progressively, leading to a decline in their quality of life. Memory loss, confusion, and difficulty communicating are all forms of cognitive function impairments that affect the individual's ability to engage in daily activities. Second, AD puts a burden on caregivers in the family. Patients with AD require extensive care, and the people who provide this care, such as family members or friends, may go through physical, mental, and financial strain. Caregivers may also face challenges in assisting individuals with AD with daily activities and in managing their behaviours. Last, there is a significant economic burden that AD places on patients, families, and healthcare systems. According to the article '2023 Alzheimer's Disease Facts and Figures' published by Alzheimer's Association [2], in the United States, the total payment for all individuals with AD or other forms of dementia was estimated to be $345 million. The costs associated with medical care, long-term care, and caregiver support continue to rise as the prevalence of individuals with AD increases.

According to the 'World Alzheimer Report 2015' by Alzheimer's Disease International, it was estimated that there would be one AD case every three seconds in the future. By 2030, there would be 74.7 million people with dementia, double the number in 2015 [3]. Based on the estimates for the absolute number of dementia cases, China and the Western Pacific region, Western Europe, and the North Pacific region have a higher number of dementia cases than other regions [4]. Therefore, given the prevalence of AD and its projected growth in the coming years, understanding how different factors such as demographic information, medical history, and medicine usage affect AD treatment outcome can provide valuable insights for improving patient care and developing targeted interventions. Additionally, this research has the potential to help professionals optimise treatment strategies, improve patients' treatment outcomes, and enhance the management of AD overall, ultimately improving the quality of care provided to patients with AD.

1.2 Problem Statement and Objectives

This project will explore how demographic information (i.e. sex, age, education level, independence level, and residence type), medical history, and medicine usage affect the treatment outcome of AD. This research aims to answer several questions of interest related to AD treatment outcome:

- How do demographic factors such as age and gender affect the treatment outcome of AD ? Are there any differences in treatment outcomes with different demographic groups?
- What is the association between certain medical conditions or comorbidities and AD treatment outcome?
- Are there any specific medicines that are more commonly associated with positive treatment outcomes?

1.3 Structure

This dissertation is structured as follows: Chapter 2 contains the literature review, including an overview of AD and its existing treatments and an exploration of previous research on the impact of different factors on AD. For example, existing research on the effect of various comorbidities will be presented. The limitations of these studies will also be addressed. Next, in Chapter 3, the methodology of the research will be discussed, and a flow chart showing the sequence of steps will be presented. Chapter 4 will focus on data collection and pre-processing, including where the data used in this research were obtained from and what work was done to filter the data. Chapter 4 will also present some visualisations of our explanatory data analysis. For instance, we will analyse the overall treatment outcome and how it varies across different factors. In Chapter 5, statistical analysis will be conducted, including the interpretation of the logistic regression model and the identification of factors that have a relatively higher association with the AD treatment outcome. Chapter 6 will consist of the selection and interpretation of the machine learning model and the resulting model and feature performance. Chapter 6 will mainly focus on addressing the relationship between different factors and AD treatment outcome based on the findings. Chapter 7 will contain a summary of the findings and will present the limitations. Last, Chapter 8 will present the research conclusions.

2 Literature Review

2.1 AD Overview and Treatment

AD is a type of brain disease that damages the neurons in our brains. The first neurons that are damaged are those responsible for memory, thinking, and language. Therefore, the first symptoms of people with AD usually tend to be memory, thinking, and language problems. AD is characterised by specific changes in the brain, including the formation of clumps of beta-amyloid protein fragments (known as beta-amyloid plaques) outside neurons and the accumulation of an abnormal form of the tau protein (referred to as tau tangles) inside neurons [2].

Most people with AD develop it at age 65 or older. This is called late-onset AD. Regarding the risk factors for late-onset AD, age, genetics, and family history are believed to have significant effects on developing AD. Age is the most significant risk factor among these. As age increases, the percentage of people with AD increases. It is estimated that 6.7 million Americans have AD in 2023, and 73% of them are 75 years old or older. Regarding genetics, experts have found that there are genes that increase the risk of developing AD. APOE-e4 is the gene that affects the risk of late-onset AD the most. People having the e4 form of the APOE gene have higher risks of developing AD than people with other forms of the gene. Although family history is not essential for an individual to develop AD, individuals with parents or siblings with AD are more likely to develop AD [2].

Currently, two types of treatment for AD exist: medicine treatments and non-medicine treatments. The U.S. Food and Medicine Administration (FDA) has approved seven medicines, five of which are aimed at improving symptoms. For example, memantine protects the brain from an excessive number of neurotransmitters, and donepezil alleviates symptoms by increasing the number of neurotransmitters in the brain. Meanwhile, non-medicine treatment aims at maintaining or improving cognitive function, overall quantity of life, engagement, and ability to perform daily activities. Non-medicine treatment includes physical activities, memory orientation exercises, and so on [2].

Age, genetics, and family history are three risk factors that cannot be modified, but they have a significant impact on developing AD. There are also some other risk factors such as

various medical conditions that will be discussed later. More research on medicine treatment, including medicine efficacy in the treatment of AD will be reviewed.

2.2 Existing Research

Cognitive impairment is a crucial indicator in the diagnosis of AD; more specifically, cognitive impairment is a significant feature of AD that usually gets worse with time. Tracing the rate of cognitive decline can provide valuable insights into the disease's progression [5]. Furthermore, cognitive impairment is an essential treatment outcome measurement in clinical trials or studies evaluating potential treatment of AD. Slowing or stabilising cognitive decline is the ultimate goal in therapeutic interventions of AD treatment. Individuals with memory impairments who did not have dementia are said to have mild cognitive impairment (MCI). According to the AD continuum, patients with MCI have very mild symptoms that do not interfere with everyday activities [2]. These patients have higher risks of developing dementia caused by AD than other people of similar ages who do not have MCI [6]. To reduce the risks of developing dementia, people with cognitive impairment should keep active both mentally and socially. Various approaches can help people improve or maintain their cognitive function and ultimately improve the treatment outcome of AD. For example, medicine treatments are used for different risks factors such as heart disease, diabetes, and high blood pressure [7].

Researchers have known for years that there is a connection between blood pressure and AD. More specifically, elderly people with hypertension have an increased risk of developing dementia [8]. Therefore, for patients who have high blood pressure, antihypertensive treatment may be an appropriate measure to reduce their risk of developing AD. Apart from hypertension, depression is a common comorbidity in patients with AD. The reason for that is when patients find out that they have memory loss or when their ability to perform daily activities weakens, they tend to become depressed [9]. Similar to people with hypertension, people with depression face the risk of developing dementia as their risk of developing MCI increases. Last, people with diabetes have a higher risk of dementia for several reasons: first, diabetes may cause higher risk of heart disease and stroke. If blood vessels are damaged, there may be cognitive decline. Moreover, high blood sugar may cause inflammation, which in turn may damage brain cells and lead to dementia [10].

We discuss key findings from existing research regarding the impact of risk factors on the development of AD. These findings show that MCI increases the risk of developing dementia. Additionally, people with comorbidities such as hypertension, diabetes, and depression face a higher risk of developing AD than people without these comorbidities. These findings provide significant insights into the multifactorial nature of the disease. They also emphasise the need for comprehensive healthcare strategies for the treatment of not only AD but also those comorbidities.

2.3 Research Gaps and Limitations

Although the existing research on AD provides insights into its multifactorial nature, some research gaps and limitations remain. First, as we discussed previously, existing studies mainly focus on risk factors, diagnostic techniques, and disease progression. Few studies specifically investigate the impact of various risk factors on the treatment of AD. The current project aims at addressing this research gap by exploring the association among multiple factors on AD treatment outcome. This can provide valuable insights into the personalised treatment approach. Further, research on cognitive impairment is not adequate, though we know that cognitive improvement is a crucial indicator for the diagnosis of AD. Research that specifically explores the effect of MCI on the treatment of AD is required. Similarly, there is a need for

more comprehensive investigations that focus on the impact of various comorbidities on the AD treatment outcome. By understanding this impact, we can develop potential treatment strategies that target both AD and common comorbidities such as hypertension, hypercholesterolemia, and diabetes. Most existing projects also face data limitations. For example, Barry Reisberg et al. [11] investigated the efficacy of memantine in moderate-to-severe AD. The sample size was 252 patients (67% women) with a mean age of 76. A total of 181 patients completed the study. A small sample size and limited data availability may have several implications for research results. Specifically, in Barry Reisberg et al. [11], 67% of the sample were women. This suggests that data were not diverse in the study, limiting the applicability of the results. A small sample size leads to higher uncertainty. The study's results may not generalise well a larger data set.

The significance of this project lies in enhancing our understanding of the impact of multiple factors, including demographic information, comorbidities, and medicine usage, on the AD treatment outcome. Thus, this project contributes to the development of more effective and personalised treatment approaches for patients with AD.

3 Research Design

3.1 Methodology

According to the article 'Following data science methodology' by IBM [12], typically, a data science project consists of the following steps: problem identification, data collection, data preparation, model construction and evaluation, and model deployment. First, as discussed in the previous section, this project aims to understand the impact of different factors, including demographic information, medical history, and medicine usage, on treatment outcomes of AD. Next, the data contain demographic information such as sex; age and gender; and medical history, namely comorbidities and medical history of patients. Additionally, it is necessary to perform data pre-processing to obtain clean data and avoid missing values. For instance, some information about patients may not be recorded. Logistic regression analysis will be conducted to identify factors that have a high association with treatment outcomes. These identified factors will be used as predictors in the machine learning model to predict the treatment outcome. In terms of model evaluation, the performance of the model will be assessed using metrics such as precision, recall, and F1-score. Cross-validation will be employed to estimate the generalisability of the model and assess its potential for overfitting.

3.2 Flow Chart

A flow chart is a visual representation of the sequence of steps and decisions required to perform a process. Each step in the sequence is represented using a diagram and linked using lines and arrows. Figure 1 shows the flow chart for this project. First, data containing all required information need to be collected and pre-processed. Following that, explanatory data analysis is conducted to learn about the impact of different factors on AD treatment outcome. Specifically, we investigate how the treatment outcome varies across different factors. Furthermore, statistical analysis is conducted to select features based on their association with the AD treatment outcome. These selected features are used to train a machine learning model. Further feature selection is implemented to obtain the most influential factors of demographic information, medical history, and medicine usage.

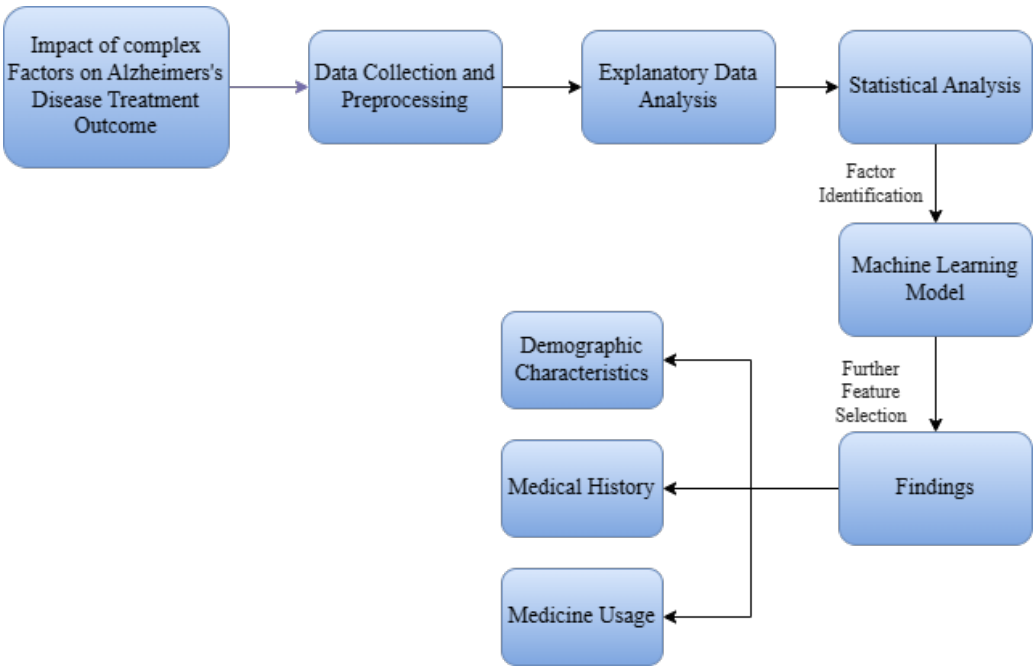

Figure 1: *Flow chart. The steps in this project can be summarised as follows: collecting and cleaning data, performing initial data analysis, identifying factors by statistical analysis, making predictions using machine learning algorithms with identified factors, and presenting findings by showing the most dominant factors of different types of information.*

4 Data Collection and Preprocessing

4.1 Data Sources

Data were collected from Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a research initiative that allows researchers to access study data as they work to define the progression of AD [13]. ADNI is a longitudinal multicentre study aiming at developing biomarkers for the early detection and monitoring of AD. It focuses on clinical, imaging, genetic, and biochemical markers to better understand and track the progression of AD. One of the significant goals of ADNI is supporting advances in AD intervention, prevention, and treatment by applying new diagnostic methods at the earliest possible stages (when intervention may be most effective) [13]. On the official ADNI website, clinical data, genetic data, MRI and PET image data, and biospecimen data are available. Because demographic, medical history, and medicine usage data are required for this project, we collected clinical data from the website.

4.2 Data Cleaning

Clinical data involves demographics, physical examinations, and cognitive assessment data. Regarding demographic information, apart from personal details such as sex, age, race, and language details, information on education level, marital status, living conditions, independence level, and residence type of each individual was collected. The independence level and residence type can reflect an individual's independence living ability. For example, individuals living in private residences are likely to have higher levels of independence living ability than individuals living in nursing houses or assisted living houses. The living situation of individuals in the data set can provide insights into their living arrangements. Additionally, medical history involves the medical histories of individuals and their families. Individual medical histories include common comorbidities such as hypertension, hypercholesterolemia, diabetes, depression,

and urinary incontinence. It also includes onset symptoms, including the mode of cognitive symptoms and changes in behaviour. Medical histories of families involve indicators of a father or mother with cognitive impairment and evidence of FTLD mutation and AD mutation in families. Regarding medicine usage, each medicine was represented as a feature with a value of 0 or 1, indicating whether the patient has taken that particular medicine or not.

After selecting all the required features from the complete set of raw data features, it was necessary to clean the data set to make sure there were no missing or invalid data. Individuals with any unknown medical conditions or demographic information needed to be filtered out from the data set. The cleaning process was simply carried out by removing those patients with either unknown or unavailable data. The treatment outcome was represented using the change in cognitive function, which had two outcomes: declined cognitive treatment outcome and not declined cognitive treatment outcome.

Feature	Description	Data Type	Values
SEX	Sex	Numeric value	1 = Male 2 = Female
AGE	Age	Numeric value	28 - 111
HISPANIC	Hispanic/Latino ethnicity	Numeric value	0 = No 1 = Yes
RACE	Race	Numeric value	1 = White 2 = Black or African American 3 = American Indian or Alaska Native 4 = Native Hawaiian or Other Pacific Islander 5 = Asian 50 = Other (specify)
EDUC	Education level	Numeric value	0 - 36
MARISTAT	Marital state	Numeric value	1 = Married 2 = Widowed 3 = Divorced 4 = Separated 5 = Never married (or marriage was annulled) 6 = Living as married/domestic partner
PRIMLANG	Primary language	Numeric value	1 = English 2 = Spanish 3 = Mandarin 4 = Cantonese 5 = Russian 6 = Japanese 8 = Other primary language (specify)
INDEPEND	Independence level	Numeric value	1 = Able to live independently 2 = Requires some assistance with complex activities 3 = Requires some assistance with basic activities 4 = Completely dependent
RESIDENC	Residence type	Numeric value	1 = Single- or multi-family private residence (apartment, condo, house) 2 = Retirement community or independent group living 3 = Assisted living, adult family home, or boarding home 4 = Skilled nursing facility, nursing home, hospital, or hospice
NACCLIVS	Living situation	Numeric value	1 = Lives alone 2 = Lives with spouse or partner 3 = Lives with relative or friend 4 = Lives with group 5 = Other

Figure 2: *Feature table of demographic information. The demographic information in the cleaned data set includes sex, age (AGE), race (RACE), education level (EDUC), marital status (MARISTAT) and primary language (PRIMLANG), living situation (NACCLIVS), independence level (INDEPEND), and residence type (RESIDENC).*

The table shown in Figure 2 displays the description, data type, and values of each feature of the demographic information from the cleaned data set. Categorical values were encoded as numeric values. As mentioned previously, the individual's independence level can reflect their ability to perform daily activities without assistance and their functional capabilities and care

requirements. Residence type indicates the type of housing. It can be helpful for assessing living conditions and social environments. Both these factors provide an understanding of the well-being and quality of life of individuals. Additionally, individuals' living situation can indicate the social support and network available to them. For instance, individuals who live alone may have less direct social support than those living with their spouses or partners.

Similar to the data type of demographic variables, the variables of medical histories and medicine were also numeric values. For example, the variable HYPERTEN had numeric values of 0, 1, and 2 indicating the subject did not have hypertension, active hypertension, or inactive hypertension; and the variable DONEPZIL had numeric values of 0 and 1, suggesting that the patient had taken the medicine donepezil.

The cleaned data set consisted of 49,639 patients and 228 features. The data summary included information on the number and prevalence of individuals with declined and not declined cognitive treatment outcome. This information was categorised by various demographic factors (Figure 3).

First, the total number of people whose cognitive symptoms had been declining over time in a gradual and continuous manner was 21,514. This accounted for 95% of the total number of patients. Meanwhile, only 5% of patients found that their cognitive syndromes remained relatively stable, or that there had been an improvement in their cognitive function. This is because AD is a progressive disease, meaning the cognitive function of patients worsens over time.

	Declined		Not Declined	
	No.	%	No.	%
SEX				
Male	24006	95.0	1256	5.0
Female	23951	95.1	1233	4.9
AGE				
Under 65	7770	94.5	453	5.5
65-85	33310	94.8	1835	5.2
Older 85	6877	97.0	201	3.0
TOTAL	47957	95.0	2489	5.0

Figure 3: *Demographics of patients by treatment outcome. This table displays the number of patients with two different treatment outcomes categorised by sex (SEX) and age (AGE).*

Next, the number of men with cognitive treatment outcome was higher than the number of women with not declined cognitive treatment outcome, relative to their respective gender groups. Additionally, the number of individuals aged 65–85 with not declined cognitive treatment outcome was the highest among their respective age groups.

5 Explanatory Data Analysis

5.1 Overall Treatment Outcome Distribution

Figure 4, show the distribution of the overall treatment outcome. It is a pie chart displaying the fraction of both declined and not declined cognitive treatment outcomes. The pie chart clearly shows that this data set contained imbalanced data. Only 5% of all treatment outcomes had not resulted in declined in cognitive function.

Oversampling technique was employed to balance the data set and deal with the class imbalance problem. Synthetic minority oversampling technique (SMOTE) is one of the most

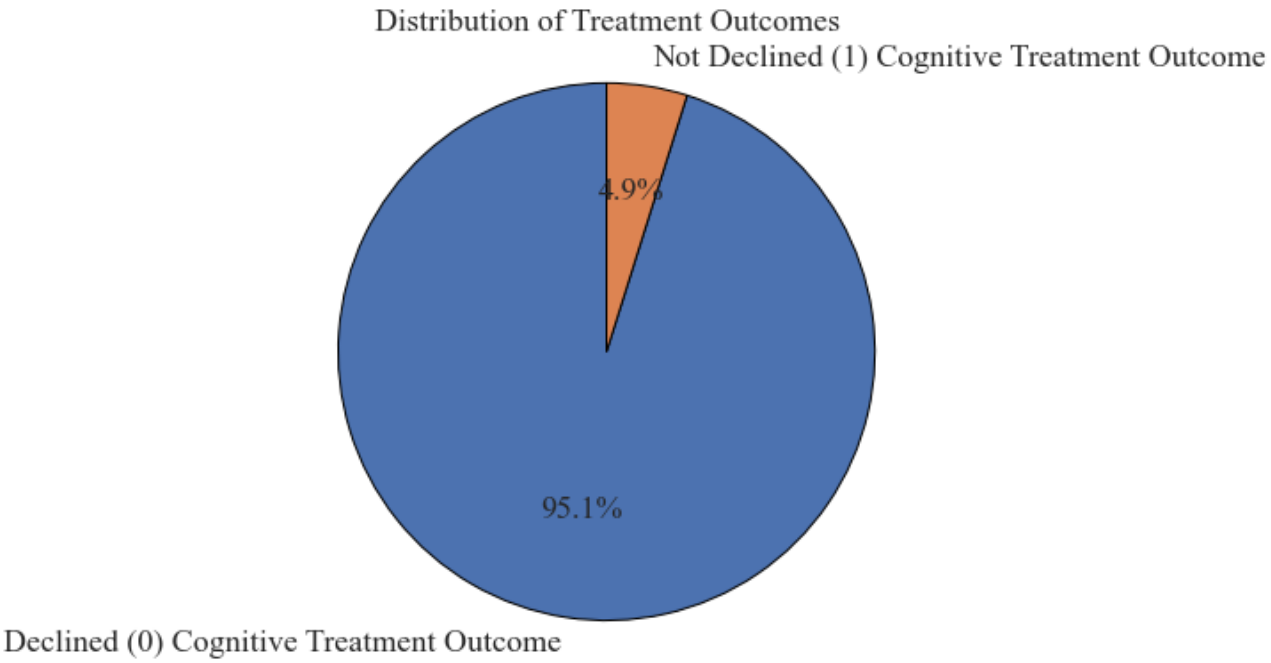

Figure 4: *Distribution of treatment outcome. The orange colour shows the proportion of patients with not declined cognitive treatment outcomes, and the blue colour shows the proportion of patients with declined cognitive treatment outcomes.*

common algorithms used to address the class imbalance problem by increasing the sample size of the minority class [13], Adaptive synthetic sampling (ADASYN) is another algorithm that can be used to balance the data set by generating synthetic samples for minority classes while taking into account each minority class sample's density distribution [14].

5.2 Association with Treatment Outcome

Given the progressive nature of AD, a treatment outcome of 0, indicating progressively declined cognitive function, was more prevalent in the data used in this project. In the data set, the minority class treatment outcome 1, indicating a not declined cognitive treatment outcome, was of particular interest. By analysing and understanding the factors associated with a positive treatment outcome, we can gain insights into potential factors that contribute to improved or static cognitive function. Furthermore, determining whether there are any medicines that are more associated with positive treatment outcomes is useful. This information can help tailor treatments and interventions to patients who are likely to show positive treatment outcomes.

The bar charts in Figure 5 display the fraction of people with not declined cognitive treatment outcomes, grouped by different demographics (sex, age, race, year of education, marital status, living situation, independence level, residence type, Hispanic or not, and primary language).

First, there were few differences in the fraction of male and female patients with positive treatment outcomes. Second, among all races, American Indians were the race with the highest fraction of patients with positive treatment outcomes. All the other races had a similar fraction of patients with not declined cognitive treatment outcomes. From grouping by year of education, we could see that education year group 5 had the highest fraction of positive treatment outcome. One potential reason for this is that patients with only 5 years of education are younger ages. Among patients with more than 10 years of education, the year of education 20 group had the highest fraction of not declined cognitive treatment outcome. The marital status bar chart shows that the 'Never Married' group had the highest fraction of patients with not declined

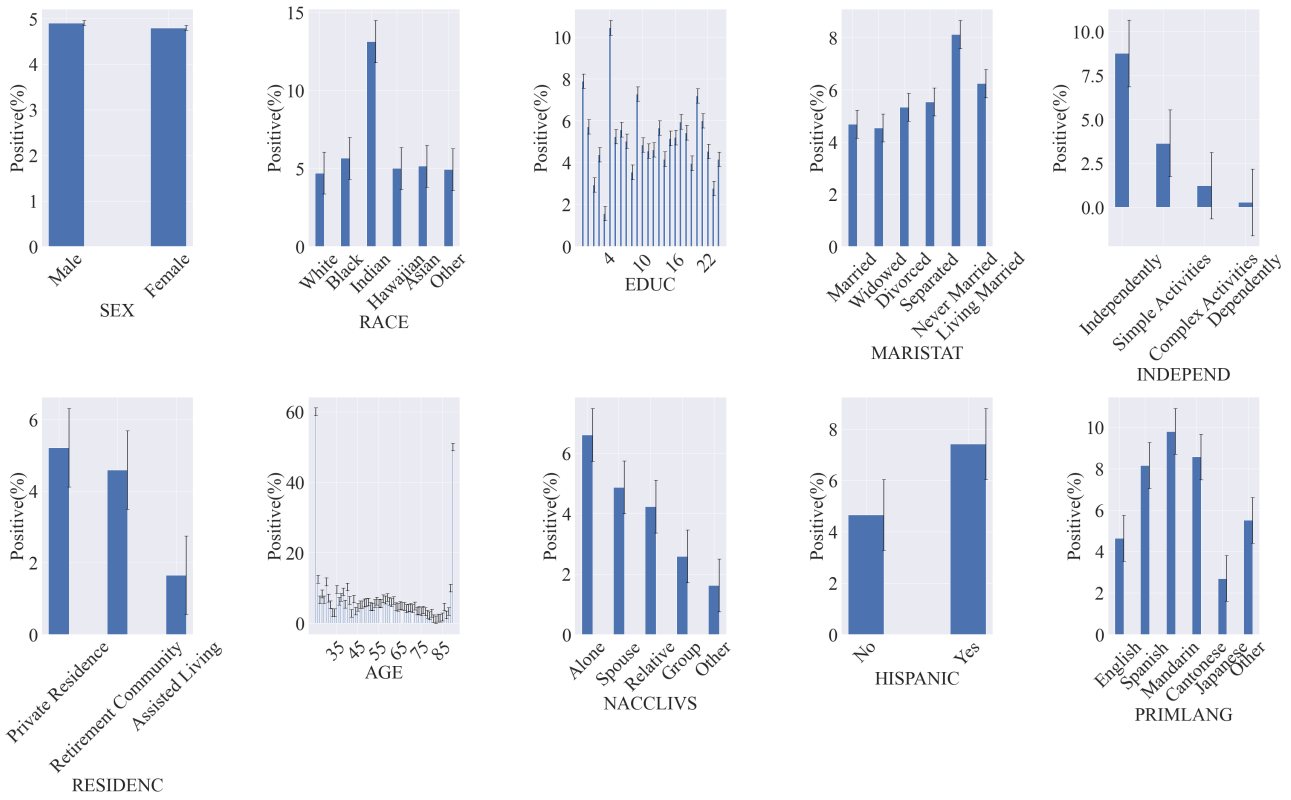

Figure 5: *Fraction of patients with not declined cognitive treatment outcome, including confidence intervals, grouped by 10 demographic features. In the first row, from left to right is the fraction of patients with positive treatment outcomes grouped by sex, race, year of education, marital status, and independence level. In the second row, from left to right is the fraction of patients with not declined cognitive treatment outcome grouped by residence type, age, living situation, Hispanic or not, and primary language.*

cognitive function treatment outcome. There may be various potential reasons for this. For instance, the age of patients who belonged to the 'Never Married' group may have been less than that of patients who belonged to other groups. This group thus had a lower risk of cognitive decline compared to other groups. As can be seen from the bar charts for independence state and residence type, there was a clear pattern of association between treatment outcome. There was a decreasing trend of patients with not declined cognitive treatment outcome. Specifically, as the independence level of individuals became lower, they became less likely to have a positive treatment outcome.

Furthermore, the bar charts show the fraction of individuals with positive treatment outcomes at different ages. As the age of individuals increased, they became less likely to have either static or improved cognitive treatment outcomes. There was a decreasing pattern within the living situation bar chart. The Living Alone group had the highest fraction of individuals with positive treatment outcomes. Regarding the association between Hispanic or not and treatment outcome, individuals who were Hispanic tended to have a higher fraction of positive treatment outcomes than individuals who were not Hispanic.

Last, because the primary language of individuals was not English, they tended to be more likely to have a not declined cognitive treatment outcome. According to Viorica Marian et al. [15], seniors who are bilingual can experience less cognitive decline compared to monolingual seniors.

Figure 6 contains 12 bar charts providing insights into the association between medical histories and not declined cognitive treatment outcomes. First, bar charts visualising the frac-

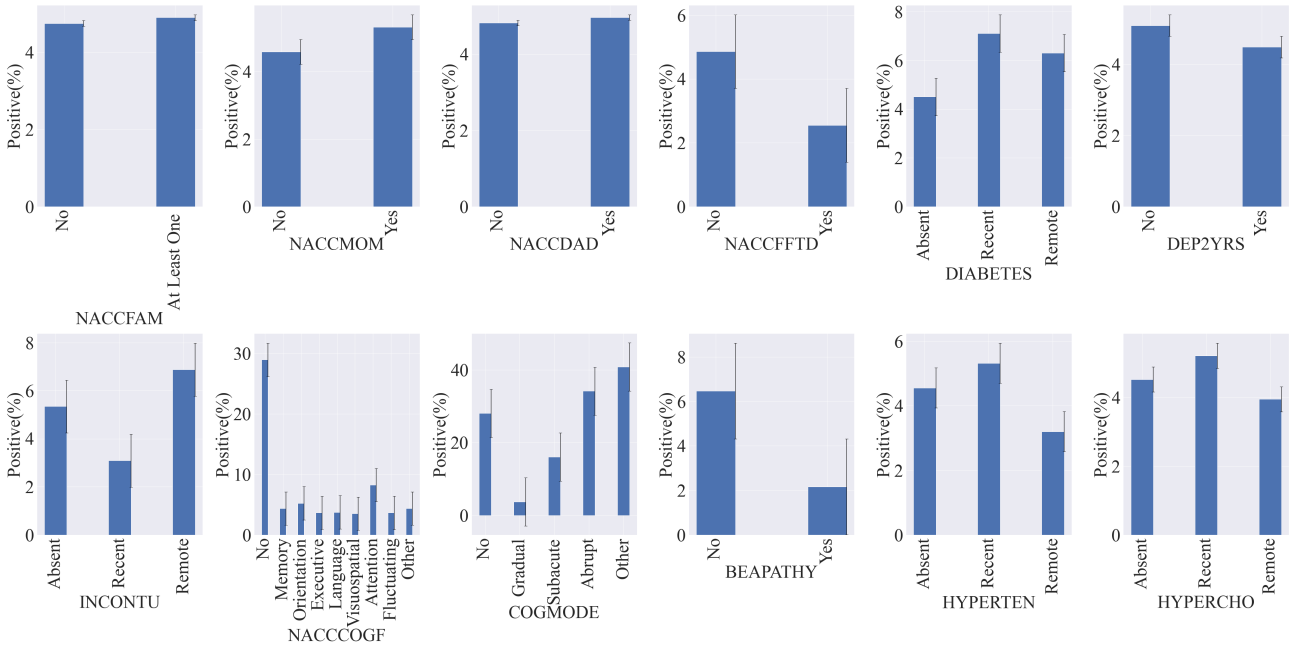

Figure 6: *The top row shows the fraction of patients with stable cognitive treatment outcomes along with confidence intervals, categorised by 12 medical history features. The left-to-right sequence in the first row includes indicators like NACCFAM, NACCMOM, NACCDAD, NACCFADM, NACCFFTD, and DIABETES, which respectively represent whether there's a family history of cognitive impairment, maternal cognitive impairment, or paternal cognitive impairment; a dominantly inherited AD mutation; a dominantly inherited FTLD mutation; and diabetes. In the bottom row, the fraction of patients with stable cognitive outcomes is displayed based on the variables DEP2YRS, NACCCOGF, COGMODE, BEAPATHY, HYPERTEN, and HYPERCHO. These correspond to the presence of depression in the last two years, the first recognised cognitive decline symptom, mode of onset of cognitive symptoms, manifestation of meaningful behaviour change, presence of hypertension, and presence of hypercholesterolemia.*

tion of patients with positive treatment outcomes were grouped according to evidence that the patients' first-degree family members had cognitive impairment (NACCFAM) and that the patients' mother (NACCMOM) and father (NACCDAD) had cognitive impairment. There was no obvious association between the fraction of positive treatment outcome and whether a patient's first-degree family member and father had cognitive impairment. However, there was a slightly greater difference between the fraction of positive treatment outcome of patients with mothers who did and did not have cognitive impairment. Next, the bar charts of the fraction of not declined cognitive treatment outcomes grouped by evidence of a dominantly inherited AD mutation (NACCFADM) and FTLD (NACCFFTD) mutation in the patient's family showed that such a patient is less likely to obtain a not declined cognitive treatment outcome. Additionally, in terms of comorbidities, it was unclear from the bar charts of the fraction of patients with positive treatment outcome grouped by diabetes (DIABETES), depression in the last two years (DEP2YRS), hypertension (HYPERTEN), and hypercholesterolemia (HYPERCHO) whether patients with diabetes, hypertension, and hypercholesterolemia were less likely to have not declined cognitive treatment outcome. Patients with depression in the last two years were less likely have a positive treatment outcome than patients who did not have depression in the last two years.

Finally, in the bar charts of the fraction of not declined treatment outcomes grouped by mode of onset of cognitive symptoms (COGMODE), the predominant symptom of a decline in the patient's cognition (NACCCOGF) was that the patient manifested a meaningful change in

behaviour (BEAPATHY). Patients who did not have any cognitive symptoms and impairments or did not manifest any changes in behaviours tended to have a much higher chance of obtaining positive treatment outcome.

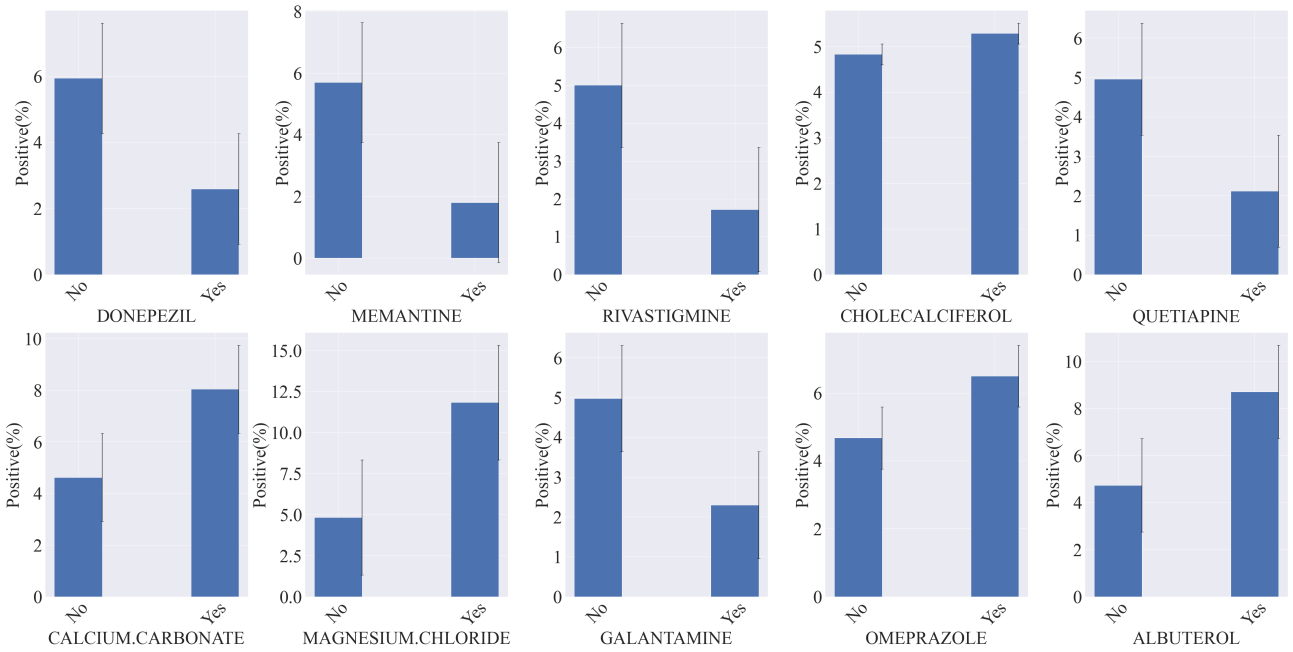

Figure 7: *Fraction of patients with not declined cognitive treatment outcome including confidence intervals, grouped by 10 selected medicines. From top left to right bottom is the fraction of patients with positive treatment outcome grouped by donepezil (DONEPEZIL), memantine (MEMANTINE), rivastigmine (RIVASTIGMINE), cholecalciferol (CHOLECALCIFEROL), quetiapine (QUETIAPINE), calcium carbonate (CALCIUM CARBONATE), magnesium.chloride (MAGNESIUM.CHLORIDE), galantamine (GALANTAMINE), omeprazole (OMEPRAZOLE), albuterol (ALBUTEROL).*

The bar charts in Figure 7 show the fraction of patients with positive treatment outcome grouped by different medicines. The fraction of patients taking common medicines, including donepezil, memantine, rivastigmine, or galantamine, used to treat AD who had a positive treatment outcome was not higher than that of patients not taking these medicines. One potential reason for this is that the sample size of patients who took these medicines was relatively small. Only 30%, 20%, 5%, and 5% of patients took donepezil, memantine, rivastigmine, and galantamine, respectively. Additionally, patients taking these medicines were in a moderate to severe stage of AD. Both donepezil and memantine can be used to treat patients with moderate to severe AD. Regarding the association between other medicines and AD treatment outcome, the fraction of patients taking cholecalciferol, calcium carbonate, magnesium chloride, omeprazole, or albuterol who had a positive treatment outcome was higher than the fraction of patients not taking any of these medicines.

Furthermore, the confidence interval for patients who took donepezil, memantine, rivastigmine, or galantamine did not overlap with that of patients who did not take these medicines. This indicated that there was a significant difference between the group that took these medicines and the group that did not. Additionally, the confidence interval for the group that took donepezil, memantine, rivastigmine, or galantamine was lower than that for those who did not. This suggests that the estimated fraction of positive treatment outcomes for the group that took these medicines was lower than that for the group that did not. A small sample size can result in a lower estimated fraction of positive treatment outcomes for patients who took these medicines. This is because the small size can make the estimates more sensitive to the

treatment outcome of a few patients. For instance, the group of patients who took donepezil was small. If only a couple of patients had positive treatment outcomes, the estimate of the fraction of positive treatment outcomes would be affected.

5.3 Correlation Between Feature and Treatment Outcome

Figure 8 contains three heatmaps that show the correlation between each pair of features of demographics, medical history, and medicine, and the correlation between each feature with the treatment outcome.

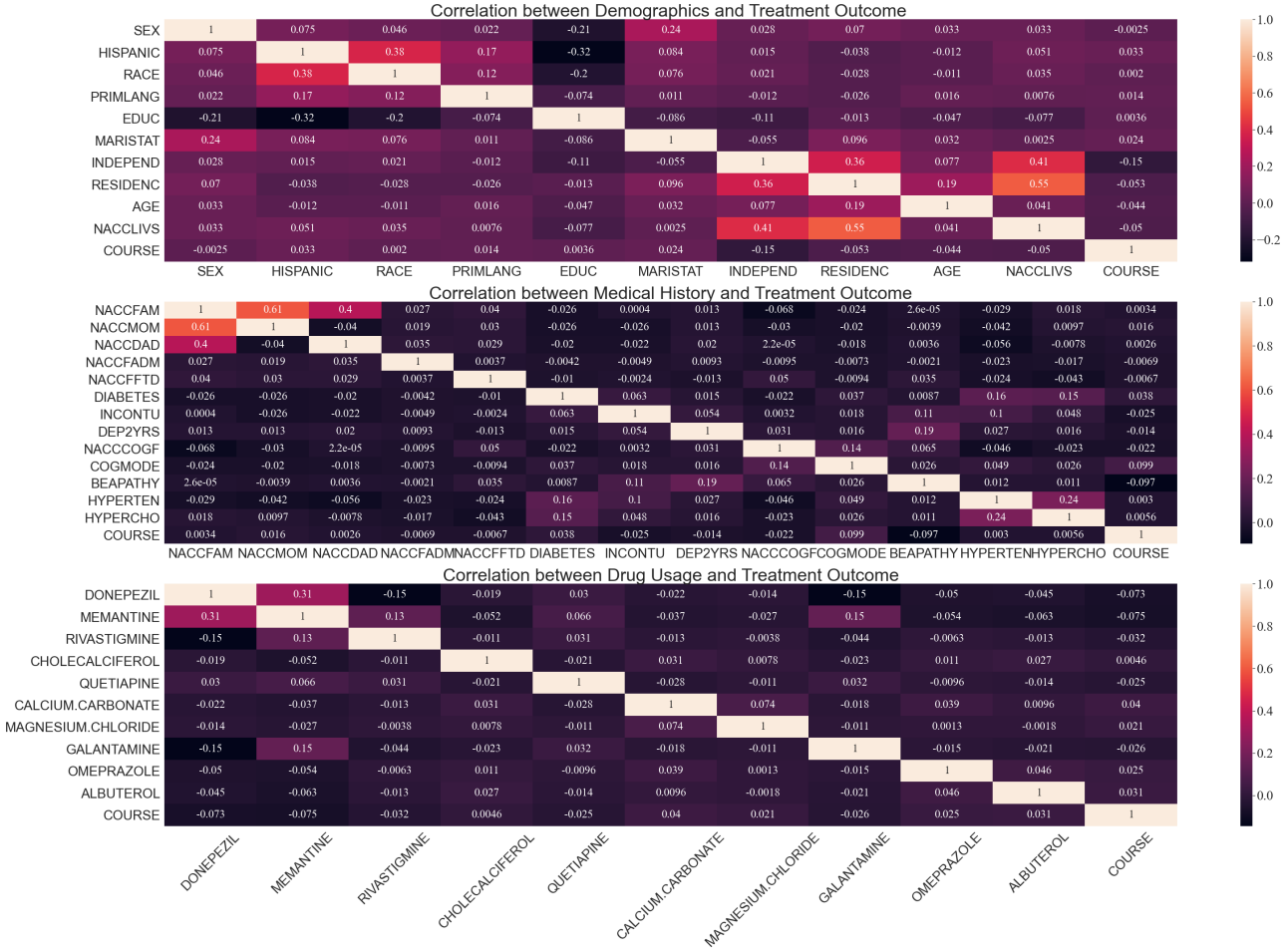

Figure 8: *Correlation between features and treatment outcome. The heatmap shown at the top shows the correlation between each pair of variable of demographic information and between each feature with the treatment outcome, which is the COURSE variable shown in the last column. The heatmaps in the middle and bottom show the correlation between each pair of medical history and medicine variables and between each feature with the overall treatment outcome, respectively.*

From the topmost heatmap, we can see that there were strong correlations between some pair of variables of demographics. For instance, the correlation between any pair of variables among INDEPEND, RESIDENC, and NACCLIVS was relatively stronger than between other pairs. In particular, the correlation between INDEPEND and NACCLLIVS was the highest, indicating that the individual's independence level was correlated with their living situation. Regarding the demographic feature correlations with the COURSE treatment outcome, the correlation among INDEPEND, NACCLIVS, and RESIDENC was negative. This means that as the level of independence or as the living condition becomes more dependent or less favourable, a higher

likelihood of declined cognitive treatment outcome is obtained. Additionally, there is a negative correlation between age and treatment outcome, illustrating that individuals of older ages tend to be less likely to have either a static or improved cognitive treatment outcome.

The heatmap in the middle displays correlations between each pair of variables of medical history, as well as the correlation between each feature and the treatment outcome. To begin with, the correlation between the variable NACCFAM and variable NACCMOM is the highest among all correlation values meaning that there is a strong positive correlation between them. Specifically, if an individual has any first-degree family member with cognitive impairment, it is likely that this family member is his/her mother. Among the correlation between each feature with treatment outcome, we can see that the correlation between variable NACCCOGF and the treatment outcome is the highest, indicating that individuals with any cognition impairment symptoms are less likely to have a positive treatment outcome.

Last, the heatmap at the bottom shows the correlation between each of the 20 selected medicines and between each medicine and treatment outcome. These 20 medicines were selected based on their correlation with the AD treatment outcome. Only the top 20 medicines with the highest correlation values between the AD treatment outcome were selected for inclusion in the figure. Regarding the correlation between each medicine, the correlation value between MEMANTINE and DONEPENZIL was the highest, indicating that individuals who take memantine are more likely to take donepezil. This is because both two medicines are used to treat AD. Further, there was a negative correlation between RIVASTIGMINE and GALANTAMINE (the other two medicines used to treat patients with mild-to-moderate dementia caused by AD) and memantine, with correlation values of 0.13 and 0.15. Regarding the correlation between each medicine and treatment outcome, there seemed to be a negative correlation between memantine, donepezil, rivastigmine, and galantamine and AD treatment outcome. The sample size of patients who took these four medicines was relatively small, which may have affected the direction and strength of the correlation analysis. Another potential reason for the effect is that memantine and donepezil can both be used to treat moderate-to-severe AD, meaning that patients taking this medicine may have a higher likelihood of declined cognitive treatment outcome regardless of treatment.

6 Statistical Analysis

6.1 Logistic Regression

The aim of performing a statistical analysis is to identify variables that have a high association with the AD treatment outcome. These features will be used to train the machine learning algorithms. Statistical analysis was conducted using a logistic regression algorithm. A logistic regression model measures the relationship between the independent variable (covariate) and dependent variable (response) by estimating probabilities using the underlying logit function [16]. A logit function can be defined as follows

$$logit(p) = log(\frac{p}{1-p})$$

where the p is the given probability and $\frac{p}{1-p}$ is the corresponding odds.

The logistic function is a linear generalised model also known as GLM. There are three components of a GLM: random component, systematic component, and link function [17]. The random component refers to the probability distribution of the response variable. For instance, it refers to the binomial distribution of response variable Y in logistic regression. The systematic component refers to the explanatory variable as a combination of linear predictors, such as $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ where β_0, β_1 and β_2 are parameters and X_1,X_2 are covariates. The

link function η refers to the link between the random component and systematic component, which shows how the expected value of the response variable is related to linear predictors of the explanatory variable, such as $\eta = logit(\pi)$ for logistic regression.

In this project, the random response component was the probability distribution of the response variable treatment outcome of AD, which was binomial distribution:

$$Y \sim Bin(N, \pi)$$

where N is the total number of patients, and π is the probability where there is a positive treatment outcome.

The response variable has two values, 0 and 1, with 0 indicating a declined cognitive treatment outcome and 1 indicating a not declined cognitive treatment outcome. In terms of systematic components, the explanatory variables were demographic information, medical histories, and medicines. For example, given the explanatory variable X_{AGE}, which we defined as X_{AGE}, the combination of a linear predictor wasthe combination of a linear predictor will then be

$$\beta_0 + \beta_1 X_{AGE}$$

where β_0 and β_1 are two parameters.
With logit as a link function, the link function can be written as follows:

$$\eta = logit(\pi) = \frac{\pi}{1 - \pi}$$

where π is the probability as there is a not declined cognitive treatment outcome.

6.2 Factors Identification

To identify factors that had a relatively higher association with the treatment outcome, each logistic regression model as covariate is each demographic, medical history, and medicine was constructed. After each bivariate logistic regression model with each of the independent variables was trained using training data, testing data were used to make predictions. The performance of each model was evaluated using different matrices. In this project, the area under the receiver operating characteristics curve (AUC-ROC) curve was applied. The AUC value of each logistic regression model as covariate is each factor was used to assess the predictive power of each variable. To be specific, the higher the AUC value, the stronger predictive power the factor had. To deal with the imbalanced training data, samples with positive treatment outcomes were generated randomly so that the sample size of the positive treatment outcome class was the same as that of the negative treatment outcome class. Loops controlled by list of demographics, medical histories, and medicines, respectively, could evaluate the predictive power of different variables on the AD treatment outcome. They generated a ROC-AUC curve, allowing us to compare how well each variable discriminated between positive and negative treatment outcomes. By visualising ROC curves and observing AUC values, we could identify which variables had a stronger influence on predicting AD treatment outcome.

As Figure 9 shows, there were AUC-ROC curves of 12 demographic information including sex, Hispanic or not, race, primary language, year of education, marital status, independent state, residence type, age and living situation. The logistic regression model that predicted the treatment outcome COURSE based on the covariate INDEPEND had the best performance compared to other logistic models, as we can tell from its AUC value of 0.69. Next, the logistic regression models that predicted the treatment outcome based on the covariates AGE and NACCLIVS had the same AUC value, 0.56. This was the second highest AUC value. The higher AUC value indicated that a model that took INDEPEND as an indicator had

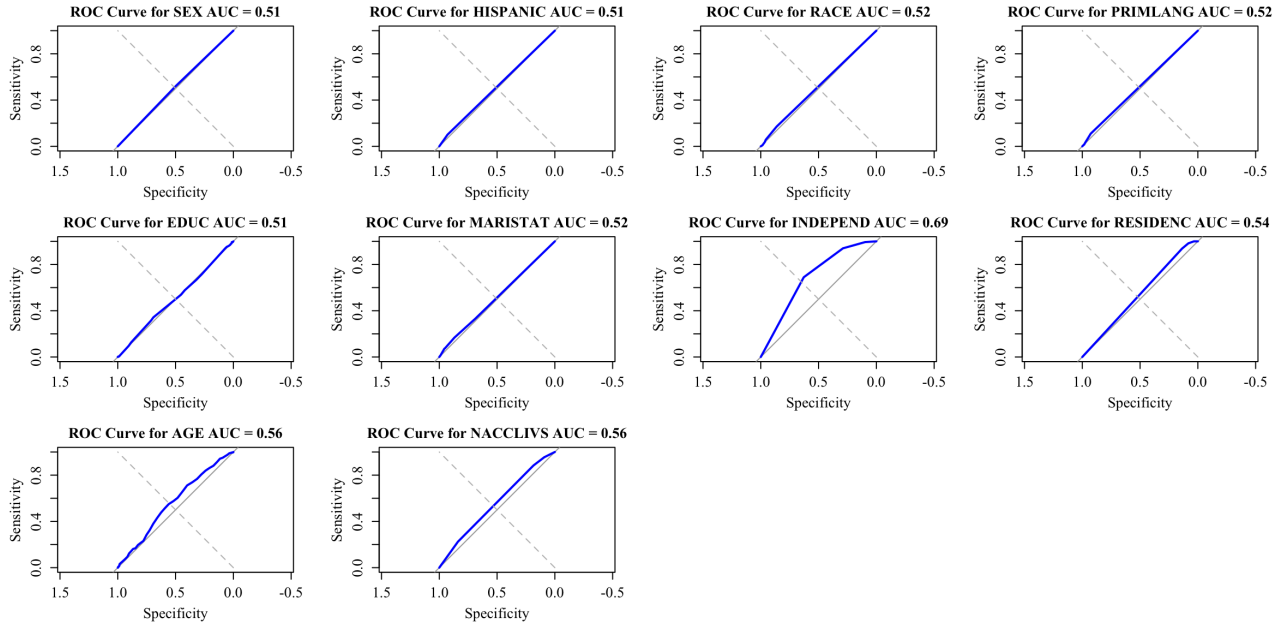

Figure 9: *AUC-ROC curves of the models with demographic covariates is presented. In the top row, from left to right are the ROC-AUC curves of models predicting the treatment outcome COURSE with predictors SEX, HISPANIC, RACE, PRILANG, and EDUC, respectively. In the second row, from left to right, the ROC-AUC curves correspond to models with covariates MARISTAT, INDEPEND, and RESIDENC. The third row displays ROC-AUC curves for models with covariates AGE and NACCLIVS.*

more differentiation power and was more effective at predicting treatment outcomes than the covariates NACCLIVS and AGE. Because the AUC value represented the probability that a randomly chosen positive treatment outcome would be ranked higher than a randomly chosen negative treatment outcome, an AUC value of close to 0.5 meant the model performed like a random classifier. For instance, models that made a prediction based on SEX, HISPANIC, RACE, PRIMLANG, EDUC, and MARISTAT had AUC values close to 0.5, indicating that these models performed poorly as a random classifier.

Similarly, Figure 10 involves the ROC-AUC curve of 12 medical histories, where we can tell from the AUC value of each model that the model as the covariate is BEAPATHY indicating whether the patient currently manifests changes in behaviour has the highest AUC value Of 0.61. Additionally, the model that predicts the treatment outcome COURSE based on the predictor NACCCOGF that indicates the patient's predominant symptom that was first recognised as a decline in cognitive function has the second highest AUC value of 0.57. The rest of the ROC curves show that the model that takes other variables such as NACCDAD, NACCFAM, NACCFFTD, HYPERTEN and COGMODE has an AUC value close to 0.5, which indicates father with cognitive impairment, a family member with inherited AD and FTLD mutation, mode of onset of cognitive decline and hypertension are weaker predictors of treatment outcome compared to other covariates.

Similarly, Figure 10 shows the ROC-AUC curves of 12 medical histories. We can tell from the AUC value of each model that the model as the covariate is BEAPATHY, indicating whether the patient currently manifested changes in behaviour, had the highest AUC value of 0.61. Additionally, the model that predicted the treatment outcome COURSE based on the predictor NACCCOGF – indicating the patient's predominant symptom, a decline in cognitive function – had the second highest AUC value of 0.57. The rest of the ROC curves showed that the model that took other variables such as NACCDAD, NACCFAM, NACCFFTD, HYPERTEN, and COGMODE had an AUC value close to 0.5. This indicated that having a father with cognitive

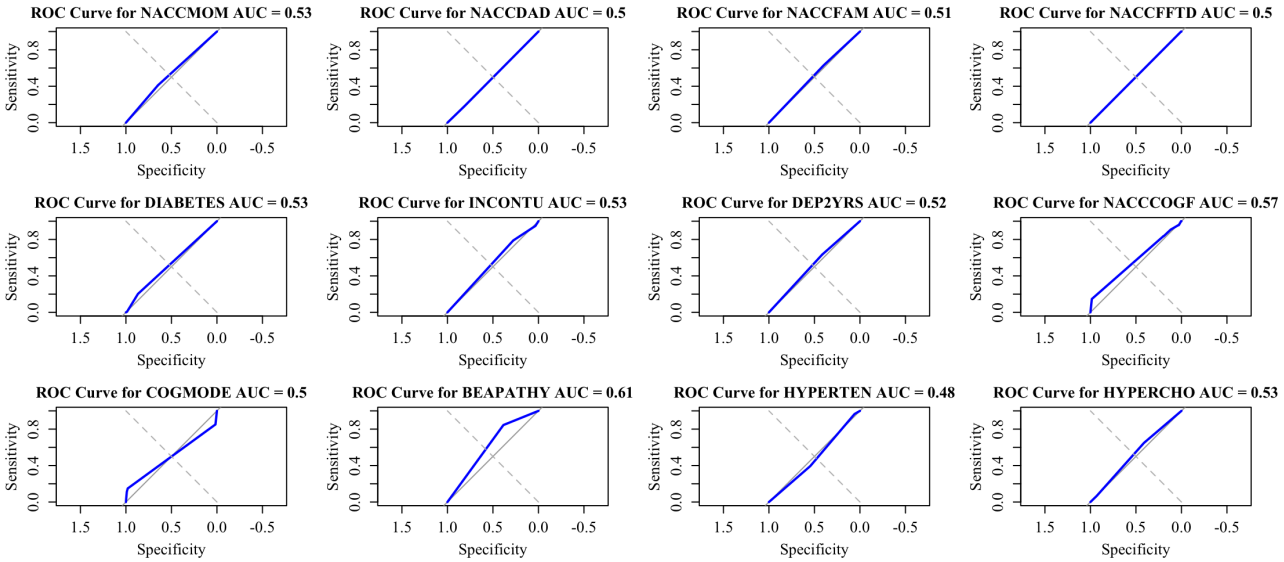

Figure 10: *ROC-AUC curve of each logistic regression model using medical history as covariates. The first row shows NACCMOM, NACCDAD, NACCFADM, and NACCFFTD as predictors. The second row shows DIABETES, INCONTU, DEP2YRS, and NACCCOGF as predictors. The third row shows COGMODE, BEAPATHY, HYPERTEN, HYPERCHO as covariates.*

impairment, a family member with inherited AD and FTLD mutation, and mode of onset of cognitive decline and hypertension were weaker predictors of treatment outcome compared to other covariates.

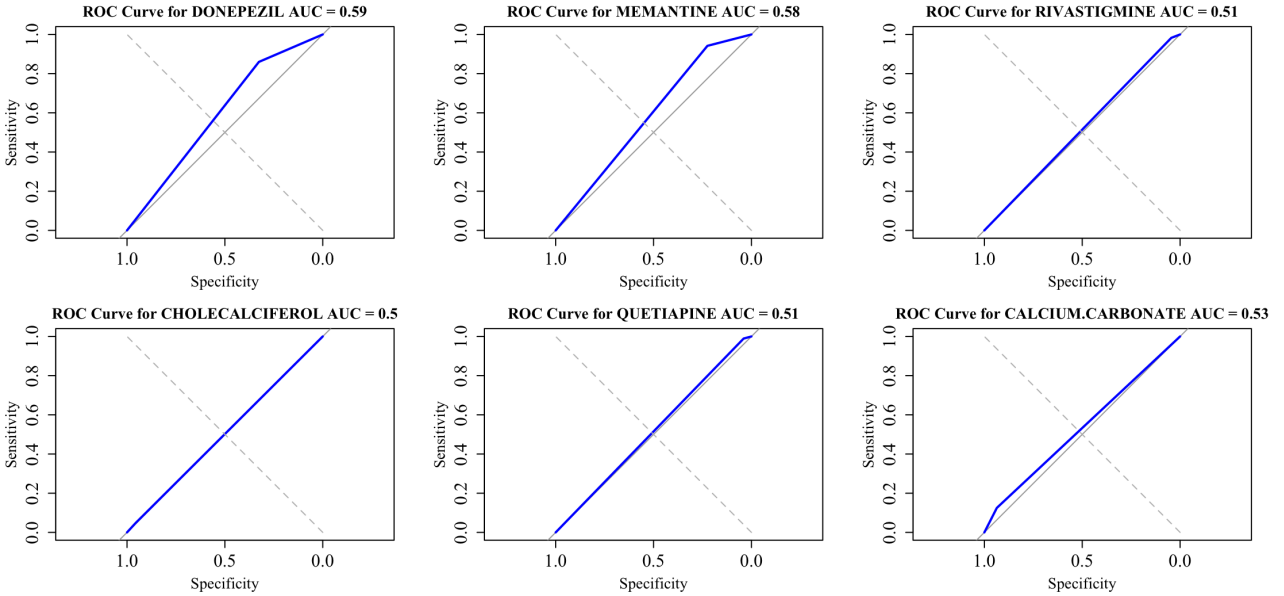

Figure 11: *ROC curves of models as covariate are medicine. At the row, from left to right is the AUC value of model that predicts the treatment outcome as the covariate is medicine donepezil, memantine and rivastigmine subsequently. At the second row, from left to right is the model AUC value as covariate is cholecalciferol, quetiapine and calcium carbonate accordingly.*

Figure 11 shows ROC-AUC curves of each model that predicted AD treatment outcome. The predictor was medicine, and the medicines were donepezil, memantine, rivastigmine, cholecalciferol, quetiapine, and calcium carbonate. These medicines were selected based on their correlation values with the treatment outcome. They were the top six medicines that had the

highest correlation with the treatment outcome. The two models where the predictors were memantine and donepezil, respectively, had the highest AUC-ROC. This means that memantine and donepezil were the two strongest predictors among all six medicines. Rivastigmine is used to treat mild to severe dementia associated with AD. It may be helpful in improving thinking ability in some patients but will not cure or stop AD from getting worse [18]. Therefore, RIVASTIGMINE was a weak predictor that did not have much power in discriminating positive and negative treatment outcome. This was apparent from the AUC value of 0.51 for the model with RIVASTIGMINE as the covariate.

7 Machine Learning Algorithm

7.1 Algorithm Overviews

In this project, different models using different algorithms with different subsets of features were built, including logistic regression, decision tree, and random forest. All models that were built based on different algorithms used the scikit-learn library in Python, an open source of machine learning supporting supervised and unsupervised machine learning. Scikit-learn also supports tools for model fitting, selection and evaluation, and other utilities [19]. The procedure of developing a machine learning model involves training it using the training data set and evaluating it using the testing data set. Model evaluation matrices involve precision, recall, and F1-score. The following are the mechanisms of each of the models:

- Logistic Regression: It works by calculating the probability of each instance belonging to each class. If the probability of a sample belonging to class 1 is greater than or equal to a threshold (i.e. 0.5), this sample will be classified as class 1; otherwise, it will be classified as 0 [14].
- Decision tree: It can be used to solve both regression and classification problems. The classification decision tree was applied in this research. A decision tree uses decisions to make predictions and assess their consequences [20].
- Random forest: It involves growing multiple individual decision trees for more accurate predictions. When applying random forest algorithm for a classification problem, each individual tree casts a 'vote', and then the random forest chooses a classification based on the majority of the 'votes' cast [21].

7.2 Model Performance

Because of the progressive nature of AD, the data set used in this project was an imbalanced data set, where the declined cognitive treatment outcome class (negative treatment outcome) was much more prevalent than the not declined cognitive treatment outcome (positive treatment outcome) class. Traditional model accuracy may not be the most suitable metric for assessing each model's performance. The reason for this is that a model that always predicts the majority class (declined cognitive treatment outcome) can still achieve high model accuracy, regardless of its lack of value in real-world cases. Therefore, to assess the performance of each model, evaluation metrics involving precision, recall, F1-score, and area under the precision-recall curve (AUC-PR) of positive treatment outcome class were focused on.

- Precision: Precision is the ratio of true positive predictions to all the positive predictions the model makes. It can indicate the accuracy of those positive predictions.

- Recall: Recall is the ratio of true predicted positive treatment outcomes to all actual positive treatment outcomes. It represents the ability of the model to correctly capture all positive treatment outcomes.

- F1-score: F1 is the harmonic mean of precision and recall. It can provide insights into the balance between precision and recall. A higher F1-score indicates the model has achieved a better balance between precision and recall.

- AUC-PR: the area under the precision-recall curve is defined as average precision. Average precision indicates model performance in predicting positive treatment. A higher average precision indicates the model is better at correctly distinguishing positive treatment outcome samples from negative treatment outcome samples, even though the positive treatment outcome is the minority class.

To deal with the class imbalance problem existing in the data set, the training data were oversampled before fitting to the machine learning models. ADASYN was used to oversample the training data. Unlike the SMOTE algorithm, which simply increases the minority size to make both the declined and not declined cognitive treatment outcome classes the same size by randomly generating not declined cognitive treatment outcome class samples with replacement, ADASYN balances the data set by generating synthetic samples [22] based on the difficulty of classification for each instance. ADASYN can improve model performance in regard to predicting positive treatment outcomes by helping the model learn better from the positive treatment outcome class. In the current study, unlike in the statistical analysis, each logistic regression model was evaluated using the original test data set, which was not oversampled to make sure the test data were real-world data. Machine models were evaluated using the test data set obtained from the stratified shuffle split. The difference between stratified shuffle split and random train and test split is that the former can ensure the distribution of class remains unchanged while the split is being performed [23]. Hence, the test data are more representative as the original data set.

Model Performance / Oversampling Techniques	Precision	Recall	F1-score
SMOTE & Normal Train-Test Split	0.18	0.2	0.19
ADASYN & Normal Train-Test Split	0.14	0.33	0.19
ADASYN & Normal Train-Test Split & StratifiedShuffleSplit	**0.35**	**0.58**	**0.44**

Figure 12: *Comparing decision tree models' performance in predicting positive treatment outcomes using different oversampling techniques. The table presents precision, recall, and F1-scores for non-declined treatment outcome predictions under various oversampling methods. SMOTE & Normal Train-Test Split involve upsampling the positive outcome class to balance sizes using the SMOTE algorithm. It is followed by model evaluation of test data from the normal Train-Test split. ADASYN & Normal Train-Test Split oversamples training data using ADASYN. The model is evaluated using non-oversampled test data. ADASYN & Normal Train-Test Split & StratifiedShuffleSplit uses StratifiedShuffleSplit for model evaluation, ensuring representative test data from real-world scenarios..*

The table shown in Figure 12 displays the precision, recall, and F1-score of the positive treatment outcome of a decision tree model. The subset of features involves all demographics,

medical histories, and top six medicines with the highest correlation values with the treatment outcome. Different oversampling techniques were applied, including SMOTE and normal train-test split, ADASYN and normal train-test split, and ADASYN and normal train-test split & stratified shuffle split. The decision tree model was trained using oversampled training data from the ADASYN algorithm and then evaluated. The decision tree model using test data obtained from the StratifiedShuffleSplit function had the best performance at predicting not declined cognitive treatment, as the precision, recall, and F1-score values show.

Before assessing the performance of different models of various algorithms and subsets of features, hyperparameter tuning was applied in each model to find the best value of the hyperparameter. This helped achieve the highest F1-score of the positive treatment outcome class. The reason for considering F1-score as a performing hyperparameter is that F1-score can ensure a good trade-off between precision and recall, thus achieving high precision and high recall at the same time. Grid search algorithm is an approach that uses a different combination of all specified hyperparameters and their values. The performance of each combination is calculated to find the best value for each hyperparameter [24].

Figure 13 shows precision, recall, and F1-score values of positive treatment outcomes of random forest, decision tree, and dummy classifier with different subsets of features. The dummy classifier serves as the baseline for comparison against more complex models, of which precision, recall, and F1-score are the lowest among all models with different subsets of features. First, the feature sets involved all demographic information, medical histories, and top six medicines with the highest correlation with the treatment outcome. The decision tree model achieved higher precision, recall, and F1-score values of 0.35, 0.58, and 0.44, respectively, compared to other models.

After that, it can be seen that random forest model achieved much higher precision, recall and F1-score as the feature sets involve all demographic information, 8 medical histories involving BEAPATHY, NACCCOGF, NACCMOM, NACCFAM, DIABETES, INCONTU, DEP2YRS, HYPERCHO, 6 medicines including DONEPEZIL, MEMANTINE, CALCIUM CARBONATE, RIVASTIGMINE, QUETIAPINE. The features were selected based on the AUC-ROC values as we have seen in the Figure 10, 9 and 11, only the features with AUC values of higher than 0.5 were included in the input features. Furthermore, it is easy to see that the random forest model where the subset of input features include demographic independence level (INDEPEND),residence type (RESIDENC), age (AGE), primary language (PRIMLANG), marital status (MARISTAT), race (RACE), year of education (EDUC), sex (SEX), Hispanic or not (HISPANIC), living situations (NACCLIVS), 10 selected medical history involving indicator of mum and dad with cognitive impairment (NACCMOM and NACCDAD), indicator of first-degree family member with cognitive impairment (NACCFAM), mode of onset of cognitive impairment (COGMODE), indicator of diabetes, urinary incontinence, hypercholesterolemia, depression in the last two years (DIABETES, INCONTU, HYPCHO and DEP2YRS), the predominant symptom that was first recognized as a decline in the subject's cognition and if the patient currently manifests meaningful change in behaviour (NACCCOGF and BEAPATHY), 3 selected medicines involve donepezil (DONEPEZIL), memantine (MEMANTINE) and calcium carbonat (CALCIUM CARBONAT) is the best at correctly identifying and capturing patients with positive treatment outcome and achieving a trade-off between precision and recall, which we could tell from its precision, recall and F1-score of 0.62, 0.85 and 0.72 respectively.

Next, random forest model achieved much higher precision, recall and F1-score. The feature sets involved all demographic information, eight medical histories (BEAPATHY, NACCCOGF, NACCMOM, NACCFAM, DIABETES, INCONTU, DEP2YRS, HYPERCHO), and six medicines (DONEPEZIL, MEMANTINE, CALCIUM CARBONATE, RIVASTIGMINE, QUETIAPINE). The features were selected based on AUC-ROC values, as Figures 9, 10, and 11, show. Only the features with AUC values of higher than 0.5 were included in the input

Model	Subeset of features	Precison	Recall	F1-score
Dummy Classifier	Demographics: INDEPEND, RESIDENC, AGE, PRIMLANG, MARISTAT, RACE, EDUC, SEX, HISPANIC, NACCLIVS	0.05	0.53	0.09
Logistic Regression	Medical histories: NACCMOM, NACCDAD, NACCFAM, NACCFFTD, DIABETES, INCONTU, DEP2YRS, NACCCOGF, COGMODE, BEAPATHY, HYPERTEN, HYPERCHO	0.09	0.51	0.16
Decision Tree	Medicines: DONEPEZIL, MEMANTINE, RIVASTIGMINE, CHOLECALCIFEROL, QUETIAPINE, CALCIUM CARBONATE	0.35	0.58	0.44
Random Forest		0.2	0.32	0.25
Dummy Classifier	Demographics: INDEPEND, RESIDENC, AGE, PRIMLANG, MARISTAT, RACE, EDUC, SEX, HISPANIC, NACCLIVS	0.05	0.49	0.09
Logistic Regression	Medical histories: BEAPATHY, NACCCOGF, NACCMOM, NACCFAM, DIABETES, NCONTU, DEP2YRS, HYPERCHO	0.09	0.53	0.15
Decision Tree	Medicines: DONEPEZIL, MEMANTINE, CALCIUM CARBONATE, RIVASTIGMINE, QUETIAPINE	0.34	0.64	0.45
Random Forest		0.59	0.85	0.7
Dummy Classifier	Demographics: INDEPEND, RESIDENC, AGE, PRIMLANG, MARISTAT, RACE, EDUC, SEX, HISPANIC, NACCLIVS	0.05	0.51	0.09
Logistic Regression	Medical histories: NACCMOM, NACCFAM, COGMODE, NACCDAD, DIABETES, INCONTU, DEP2YRS, NACCCOGF, BEAPATHY, HYPERCHO	0.09	0.54	0.16
Decision Tree	Medicines: DONEPEZIL, MEMANTINE, CALCIUM CARBONATE	0.33	0.58	0.42
Random Forest		**0.62**	**0.85**	**0.72**
Dummy Classifier	Demographics: INDEPEND, RESIDENC, AGE, PRIMLANG, MARISTAT, RACE, NACCLIVS	0.05	0.51	0.09
Logistic Regression	Medical histories: NACCMOM, DIABETES, INCONTU, DEP2YRS, NACCCOGF, BEAPATHY, HYPERCHO	0,.09	0.61	0.16
Decision Tree	Medicines: DONEPEZIL, MEMANTINE, CALCIUM.CARBONATE	0.23	0.64	0.34
Random Forest		0.31	0.81	0.45

Figure 13: *Comparison of precision, recall, and F1-score of positive treatment outcome class grouped by each model and involving logistic regression model, decision tree, random forest, and dummy classifier with different subsets of features. Demographic features involve independence level (INDEPEND), residence type (RESIDENC), age (AGE), primary language (PRIMLANG), race (RACE), year of education (EDUC), sex (SEX), Hispanic or not (HISPANIC), and living conditions (NACCLIVS). Medical history involves indicator of mother and father with cognitive impairment (NACCMOM and NACCDAD), indicator of FTLD mutation in the patient's family (NACCFTLD), indicator of first-degree family member with cognitive impairment (NACCFAM), mode of onset of cognitive impairment (COGMODE), indicator of diabetes (DIABETES), urinary incontinence (INCONTU), hypercholesterolemia (HYPERCHO), hypertension (HYPERTEN) and depression in the last two years (DEP2YRS), and the predominant symptom of decline in the subject's cognition (NACCCOGF), as well as whether the patient currently manifests meaningful change in behaviour (BEAPATHY). The medicines involve donepezil (DONEPEZIL), memantine (MEMANTINE), rivastigmine (RIVASTIGMINE), cholecalciferol (CHOLECALCIFEROL), quetiapine (QUETIAPINE), and calcium carbonate (CALCIUM CARBONATE).*

features. Furthermore, it is easy to see that the random forest model where the subset of input features included demographic independence level (INDEPEND), residence type (RESIDENC), age (AGE), primary language (PRIMLANG), marital status (MARISTAT), race (RACE), year of education (EDUC), sex (SEX), Hispanic or not (HISPANIC), living situation (NACCLIVS), 10 selected medical histories involving indicators of father and mother with cognitive impairment (NACCMOM and NACCDAD), indicator of first-degree family member with cognitive impairment (NACCFAM), mode of onset of cognitive impairment (COGMODE), indicator of

diabetes, urinary incontinence, hypercholesterolemia, depression in the last two years (DIABETES, INCONTU, HYPCHO and DEP2YRS), and the predominant symptom of a decline in the subject's cognition, as well as of the patient currently manifests meaningful change in behaviour (NACCCOGF and BEAPATHY). Donepezil (DONEPEZIL), memantine (MEMANTINE), and calcium carbonate (CALCIUM CARBONAT) were the best at correctly identifying and capturing patients with positive treatment outcome and achieving a trade-off between precision and recall, as the precision, recall, and F1-score of 0.62, 0.85 and 0.72, respectively, showed.

Last, the covariates with AUC values of 0.51 or lower were not included in input feature sets. These models tend to perform worse at achieving a balance between precision and recall. For instance, the random forest model as the input feature set involved demographic information, independence level, residence type, age, primary language, marital status, race and living situation, medical history of mother and father with cognitive impairment, indicator of diabetes, urinary incontinence, hypercholesterolemia, hypertension and depression in the last two years, and the predominant symptom of a decline in the subject's cognition, as well as whether the patient currently manifested meaningful change in behaviour. Donepezil, memantine, and calcium carbonate had a much lower precision of 0.31, a slightly lower recall of 0.81, and a much lower F1-score of 0.45, indicating that it had lower accuracy at correctly identifying patients with positive treatment outcome and that it failed at achieving a good balance between precision and recall for positive predictions.

		Logistic Regression			Decision Tree			Random Forest		
		Precison	Recall	F1-score	Precison	Recall	F1-score	Precison	Recall	F1-score
Demographic										
Age	Under 65	0.14	0.56	0.23	0.35	0.62	0.45	0.70	0.84	0.77
	65-85	0.09	0.51	0.10	0.34	0.59	0.43	0.59	0.84	0.69
	Older 85	0.06	0.45	0.10	0.34	0.62	0.44	0.76	0.84	0.80
Sex	Male	0.10	0.56	0.16	0.34	0.64	0.44	0.59	0.86	0.70
	Female	0.10	0.58	0.17	0.40	0.66	0.49	0.74	0.84	0.78
Medical History										
BEAPATHY	Yes	0.10	0.50	0.16	0.35	0.63	0.45	0.61	0.84	0.71
	No	0.06	0.61	0.16	0.31	0.52	0.39	0.72	0.81	0.76
Medicine										
DONEPEZIL	Yes	0.10	0.52	0.17	0.38	0.63	0.48	0.65	0.86	0.74
	No	0.05	0.44	0.09	0.35	0.54	0.42	0.66	0.80	0.72

Figure 14: *Precision, recall, and F1-score of positive treatment outcome class of logistic regression, decision tree, and random forest model grouped by age (AGE), sex (SEX), medical history of patient currently manifesting as changes in behaviour such as social apathy (BEAPATHY), and donepezil (DONEPEZIL).*

Figure 14 shows the performance of each model grouped by age and sex, medical history of patients manifesting changes in behaviour such as social withdrawal, and donepezil. In terms of age, logistic regression showed better performance at predicting positive treatment outcomes for patients under 65 than patients aged 65–85 and older than 85. The decision tree model showed similar precision, recall, and F1-score for the positive treatment outcome class across patients with different age groups. The random forest model had the lower precision and F1-score of the positive treatment outcome class for the 65–85 age group compared to the age groups that were under 65 and older than 85. Regarding sex, there was not much difference between men and women in the precision, recall, and F1-score of all three models. Regarding the medical history of patients currently manifesting changes in behaviour, we could see that

the random forest model had a slightly higher F1-score grouped by women than men, indicating it a better balance between precision and recall. Next, both the decision tree and random forest models had similar performance among patients who took donepezil and those who did not. In conclusion, random forest had the highest predictive power in regard to predicting positive treatment outcomes. Its performance was similar across different features rather than heavily dependent on specific features.

As Figure 15 shows, from the average precision value, which could be retrieved from the precision-recall curve of the random forest model and decision tree model as a subset of features involving all demographics, 10 medical histories, and three medicines, we learnt that the random forest model had a much higher average precision value (0.7) than the logistic regression model (0.09) or decision tree model (0.46). This indicated that the random forest model had much higher precision at various recall levels and hence was much better at correctly identifying patients with not declined cognitive treatment outcome.

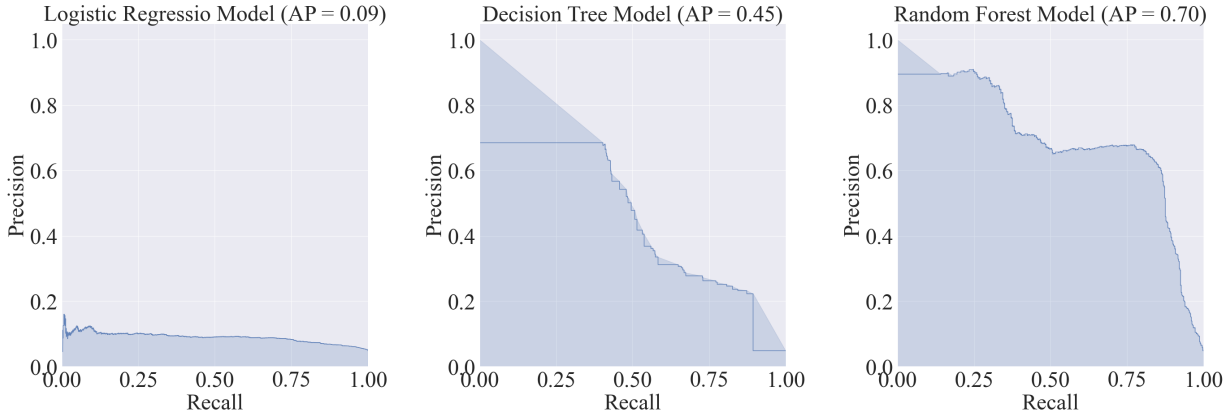

Figure 15: *The precision-recall curve of logistic regression, decision tree, and random forest as the subset of features involves all demographics, 10 medical histories, and three medicines. From left to right, it shows the precision-recall curve of the logistic regression model, decision tree model, and random forest model, respectively.*

7.3 Feature Importance

The bar chart shown in the Figure 16 displays the feature importance scores of the best random forest model, where the input feature involves demographic information sex, age, individual independence level, residence type, living situations, age, primary language, marital state, year of education, and Hispanic or not, medical history indicator of mum and dad with cognitive impairment, indicator of the first-degree family member with cognitive impairment, mode of onset of cognitive impairment, indicator of diabetes, urinary incontinence, hypercholesterolemia, depression in the last two years, the predominant symptom that was first recognized as a decline in the subject's cognition and if the patient currently manifests meaningful change in behaviour, medicine donepezil, memantine and calcium. carbonate. These feature importance scores can be used to determine the relative feature importance in building the best random forest model [25]. The ranking of the features can provide insights into the ability of the best random forest model to correctly classify patients with positive treatment outcomes. Therefore, features with higher feature scores indicate that changes in their features can have more impact on the likelihood of a not declined cognitive treatment outcome. For instance, AGE is the feature with the highest importance score which illustrates age has the most significant impact on the model's ability to correctly classify the patients with not declined cognitive treatment outcomes. Additionally, INDEPEND, and EDUC, are the other two features that have rela-

tively higher influences on the model's ability to classify individuals with positive treatment outcomes. In terms of medical history, BEAPATHY, NACCCOGF and COGMODE are three features that have the higher importance scores which suggest that the if the patient manifests changes in behaviour, the predominant symptom that was first recognized as a decline in the patient's cognition and mode of onset of cognitive impairment are three medical indicators that more relevant to the not declined cognitive treatment outcome. Concerning the medicines, it is not surprising to see that memantine and donepezil have much higher importance scores than calcium. carbonate is one of the features with the lowest importance scores.

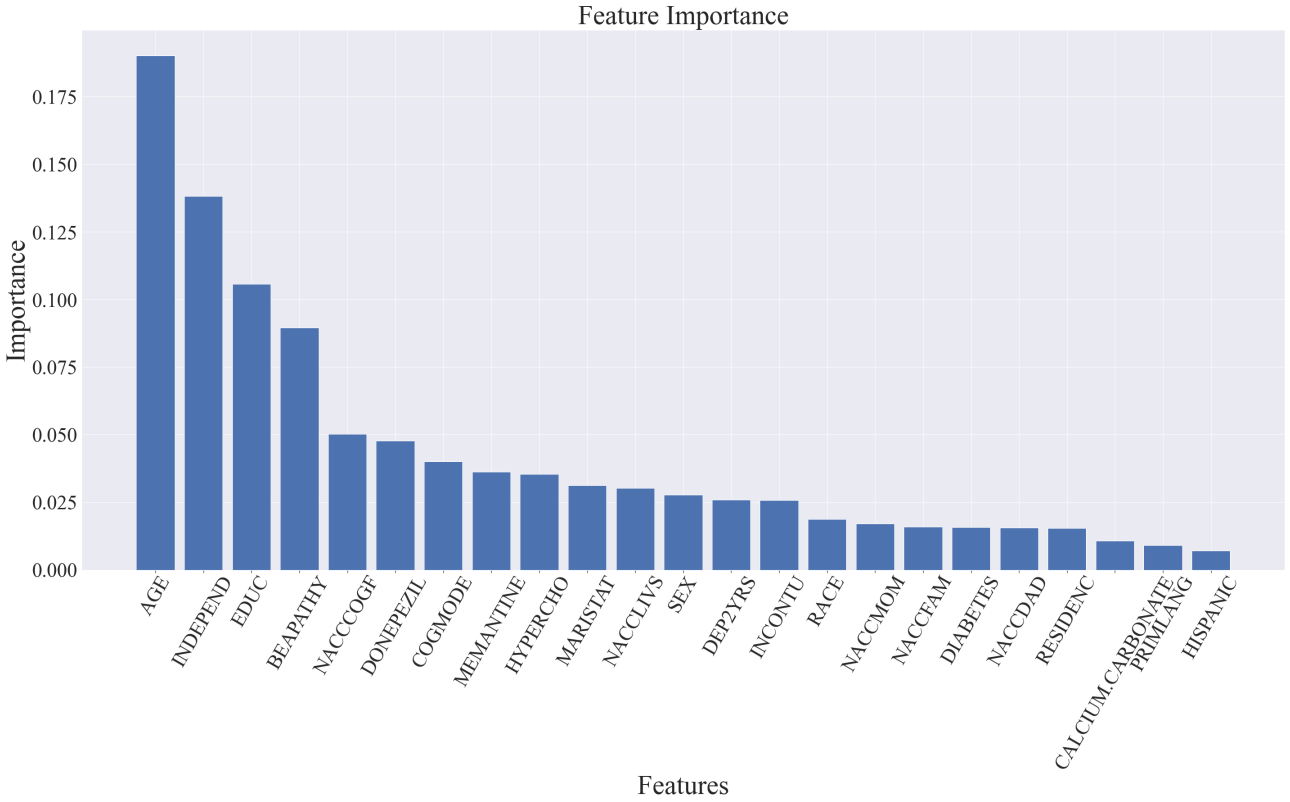

Figure 16: *Feature importance of the random forest model. The subsets of features involve demographic information (INDEPEND, RESIDENC, AGE, PRIMLANG, MARISTAT, RACE, EDUC, SEX, HISPANIC, NACCLIVS); medical history (NACCMOM, NACCFAM, COGMODE, NACCDAD, DIABETES, INCONTU, DEP2YRS, NACCCOGF, BEAPATHY, HYPERCHO); and medicine (DONEPEZIL, MEMANTINE, CALCIUM CARBONATE).*

The bar chart in Figure 16 displays the feature importance scores of the random forest model. The input features involved sex, age, individual independence level , residence type, living situation, age, primary language, marital status, year of education, Hispanic or not, medical history of parents with cognitive impairment, indicator of first-degree family member with cognitive impairment, mode of onset of cognitive impairment, indicator of diabetes, urinary incontinence, hypercholesterolemia, depression in the last two years, and the predominant symptom of a decline in the subject's cognition, as well as whether the patient currently manifests meaningful change in behaviour. The featured medicines were donepezil, memantine and calcium carbonate. These feature importance scores can be used to determine relative feature importance in building the best random forest model [26]. The ranking of the features can provide insights into the ability of the best random forest model to correctly classify patients with positive treatment outcomes. Therefore, features with higher feature scores indicated that changes in the features could have a greater impact on the likelihood of a not declined cognitive treatment outcome. For instance, AGE was the feature with the highest importance

score. It illustrated that age has the most significant impact on the model's ability to correctly classify patients with not declined cognitive treatment outcomes. Additionally, INDEPEND and EDUC were the other two features that had relatively higher influences on the model's ability to classify individuals with positive treatment outcomes. In terms of medical history, BEAPATHY, NACCCOGF, and COGMODE were three features with the higher importance scores. This suggests that the patient's manifestation of changes in behaviour, the predominant symptom of a decline in the patient's cognition, and the mode of onset of cognitive impairment were the three medical indicators more relevant to the not declined cognitive treatment outcome. Concerning medicines, it is not surprising that memantine and donepezil had much higher importance scores than calcium carbonate.

8 Results

8.1 Demographic Characteristics

Initially, as reflected in the bar charts in Figure 5, we can see that there is a strong negative association between factor age, individual independence level , living situation, and residence type and positive AD treatment outcome. For instance, as the patient gets an order or develops a decreased independence level , they are less likely to obtain a not declined cognitive treatment outcome. Our findings from the first heatmap of Figure 8 show that independence level , residence type, age, living situation, and marital status are the top five variables that have the highest correlation value with the treatment outcome. Additionally, they have a negative correlation with the treatment outcome. Furthermore, according to the AUC-ROC curve of each logistic regression model as a covariate in each demographic in Figure 9, independence level , age, residence type, and living situation have AUC values of 0.69, 0.54, 0.56, and 0.56, respectively. Thus, age, independence level , residence type, and living situation are relatively stronger predictors among all demographics. This also aligns with our findings from the bar charts showing the fraction of patients with both declined and not declined cognitive treatment outcomes grouped by independence level , age, residence type, and living situation. The heatmap shows the correlation values among the demographic factors of independence level , age, residence type and living situation, and treatment outcome. Additionally, as we can see from the bar charts displaying feature importance rank in Figure 16, age is the feature that has the highest important score. This illustrates that age and individual independence level are two features that have the most significant impact on predicting positive treatment outcomes.

In conclusion, by comparing results from different analyses, we identify demographic factors that consistently show a relatively higher association with AD treatment outcome. These factors involve age, independence level , residence type, and living situation. Elderly patients who are not able to able live independently tend to be less likely to have a positive treatment outcome than patients who are able to live independently either in a private residence or a retirement community.

8.2 Medical History

When comparing the results from different analyses, we focused on those medical histories that were consistently highlighted because they may be more robust and significant.

For instance, it is easy to see from the bar charts of the fraction of patients with declined and not declined treatment outcomes grouped by different medical histories in Figure 6 that individuals without any cognitive decline have a much higher likelihood of a positive treatment outcome than individuals with cognitive decline. Additionally, an individual currently manifesting a change in their behaviour seems to be less likely to have a not declined positive

treatment outcome. Regarding common comorbidities, patients with depression in the last two years tend to be less likely to obtain a positive treatment outcome. Moreover, patients who do not have a family member with FLTD mutation tend to have a higher likelihood of a not declined treatment outcome than patients who do not.

According to the second heatmap in Figure 8, which shows the mode of cognitive impairment, three variables that have the highest correlation with the treatment outcome are indicators of patient currently manifesting a change in their behaviour and indicators of diabetes. Meanwhile, one of the variables that have the lowest correlation with the treatment outcome is the indicator of patient's family member with FLTD. According to the AUC-ROC curve of each model as a covariate in Figure 10, if there are symptoms of a decline in the patient's cognition and if the patient currently manifests changes in behaviour, then these are two stronger predictors of treatment outcome. A medical history of hypertension, an indicator of a first-degree family member with cognitive impairment, an indicator of a father with cognitive impairment, and an indicator of a family member with FLTD are four very weak predictors of treatment outcome. Last, the bar chart in Figure 16 illustrates the feature importance of different medical histories. NACCCOGF and BEAPATHY have the highest scores among all medical history features. They denote indicators of any symptoms of a decline in the patient's cognition and they denote whether the patient currently manifests changes in behaviour. These have a more significant influence on the prediction of positive treatment outcome.

In summary, by comparing findings obtained from all analyses that were conducted, indicator of cognitive impairment (NACCCOGF) and indicator of currently going through changes in behaviour (BEAPATHY) were shown to be two medical histories that tend to have a relatively higher association between the AD treatment outcome compared to all other medical histories. Additionally, in terms of the impact of different comorbidities on the AD treatment outcome, as the second correlation heatmap in Figure 8 shows, indicators of diabetes (DIABETES), urinary incontinence (INCONTU), and depression in the last two years (DEP2YRS) have relatively higher correlation values with treatment outcome. This also aligns with our findings from the ROC-AUC plots shown in Figure 10. Indicators of diabetes and urinary incontinence are two of the strongest predictors concerning all comorbidities.

8.3 Medicine Usage

After combining results from different analyses, we could identify medicines that are more closely correlated with the AD treatment outcome. The bar charts that display the fraction of patients with positive treatment outcomes grouped by different medicines in Figure 7 show that out of all of the medicines that can be used to treat AD, memantine, donepezil, rivastigmine, and galantamine have a negative association with the treatment outcome. One possible reason for this is the relatively small sample size of patients who took these medicines. This might have resulted in a lower estimated fraction of positive treatment. Further, the medicines may not work for everyone and may lose effectiveness over time. These medicines tend to be more effective for early to moderate AD [26]. This indicates that if these medicines are used to treat patients with severe AD, whose cognitive impairments are more severe, the patients are more likely to have a declined cognitive function regardless of treatment. From the AUC-ROC curve of the logistic model, as each covariate is each medicine, it is easy to see that memantine and donepezil are the two strongest predictors of treatment outcomes among all medicines. This aligns with the fact that donepezil and memantine are two medicines with the highest correlation values with AD treatment outcomes, as shown in the heatmap at the bottom of Figure 8.

To conclude, our findings show that memantine and donepezil are two medicines that have a relatively higher association with AD treatment. The negative correlation between memantine

and treatment outcome does not imply a cause-and-effect relationship, which suggests there might be other factors that contribute to the observed correlation.

9 Discussions

9.1 Findings Summary

According to the explanatory data analysis conducted to illustrate the association among demographic information, medical history, and medicine and positive AD treatment outcome, age, independence level, residence type, and living situation are demographics that have a relatively higher association with treatment outcome. The negative correlation value between each of these demographics and treatment outcome indicates that patients with older age and a reduced ability to live independently tend to have a positive AD treatment outcome. At the same time, sex, year of education, and race have the lowest correlation with AD treatment outcome. For instance, the correlation value between race and AD treatment outcome is 0.002, as shown in the first heatmap of Figure 8, indicating that race has little impact on AD treatment outcome. The findings from the explanatory data analysis also align with the findings from the statistical analysis. From Figure 9, we find that independence level, age, residence type, and living situation, respectively, have the highest AUC values. Meanwhile, sex, Hispanic or not, and year of education, respectively, have the lowest AUC values. This indicates that age, independence level, residence type, and living situation are relatively stronger predictors whereas sex, Hispanic or not, and year of education are weaker predictors of AD treatment outcome.

Similarly, by comparing the findings from the explanatory data analysis and statistical analysis, we learn that regarding medical history, indicator of any cognitive decline symptoms and indicator of the individual currently manifesting changes in behaviour are two medical conditions that have a relatively higher association with the AD treatment outcome. At the same time, indicator of hypertension, indicator of father and first-degree family member with cognitive impairment, and indicator of family member with FTLD mutation are three factors with the lowest association with AD treatment outcome.

Finally, with respect to the association between medicines and AD treatment outcome, the bar charts in Figure 7 reveals the fraction of positive treatment outcomes grouped by each medicine. The negative correlation between donepezil, memantine, rivastigmine, galantamine, and AD treatment outcome does not imply causation, suggesting that there might be some other factors that contribute to the association between donepezil, memantine, rivastigmine, galantamine, and AD treatment outcome. Despite this, according to our findings from the AUC-ROC curve of each logistic regression model as the covariate is each medicine, we learnt that the AUC-ROC curves of the logistic regression model as the covariates are donepezil and memantine, respectively. Donepezil and memantine are relatively stronger predictors compared to other medicines.

9.2 Limitations and Future Work

One of the significant limitations was the relatively small sample size of patients. There was a limited sample size of patients who took common medicines to treat AD, such as donepezil, memantine, rivastigmine, and galantamine. This might have affected the relationship between these medicines and AD treatment outcome. A large sample size of patients taking donepezil, memantine, rivastigmine, and galantamine could provide more robust insights into the association between them and AD treatment outcome. Next, because of the nature of AD, the data set that was used in this project was imbalanced. The majority of patients had negative

treatment outcomes, which could potentially introduce bias into the predictive model. To deal with the class imbalance problem, different upsampling techniques were applied, as shown in the table in Figure 12. We could see that the oversampling skill that trains the model using upsampled training data with the ADASYN algorithm and that evaluates the model using test data obtained from StratifiedShuffleSplit could improve the decision tree model performance in identifying patients with positive AD treatment outcome. Apart from different oversampling skills, more advanced machine learning techniques such as deep learning and ensemble methods can be applied to improve the model's predictive accuracy. This will allow researchers to capture more complex relationships between different factors and AD treatment outcomes that other models might fail to observe.

Additionally, this project mainly focused on exploring the effect of demographics, medical history, and medicine usage on AD treatment outcome. Nevertheless, other factors such as genetic disposition, lifestyle, and social support system could contribute to the AD treatment outcome. For instance, the factor that indicates whether the patient has the APOE gene and that indicates the form of the APOE gene can be incorporated into future analyses. This might provide a more extensive understanding regarding the impact of different factors on the AD treatment outcome. Furthermore, data on the severity of cognitive decline and the stage of AD can be collected to gain more deep insights into the impact of the medical condition on the treatment outcome. Other patient-centred outcomes such as quality of life and caregiver burden could be included to assess the AD treatment outcome, given the fact that AD lowers the quality of life of individuals and puts a heavy burden on caregivers.

Last, external validation can be applied to ensure the models can be generalised to unseen data from a different context. The data set used in external validation usually comes from a different data source or period. The difference between the method used in this project and external validation is that when evaluating machine learning models, we used testing data that were not oversampled, whereas external validation will evaluate models using an entirely different data set collected from other sources.

10 Conclusions

Through our investigation into the impact of different demographics, medical histories, and medicines on the AD treatment outcome, we gained insights that could be useful in improving treatment outcomes and treatment strategies. There are several key findings. We observed that there are some demographics that have a significant impact on predicting AD treatment outcome, including age, independence level, residence type and living situation. Besides, some specific medical conditions that have a relatively higher association with the AD treatment outcome, such as an indicator of behaviour change and symptoms of cognitive decline, were identified. There are also some medicines that contribute more to the AD treatment outcome such as donepezil, memantine, and rivastigmine.

However, there are certain limitations such as the small sample size for certain medicines groups. Furthermore, other factors such as the stage of the disease may influence the AD treatment outcome and need further investigation in the future. In summary, this project highlights the significance of applying the multidimensional approach to understanding AD treatment outcome. The findings have the potential to enhance patient care, optimise resource allocation, and ultimately improve the overall management of AD.

11 Appendices

```
1 par(mfrow=c(3,4))
```

```r

set.seed(123) # Set seed for reproducibility
train_indices <- createDataPartition(filtered_data$COURSE, p = 0.8, list =
    FALSE)
train_data <- filtered_data[train_indices, ]
test_data <- filtered_data[-train_indices, ]

train_data_new <- upSample(x=train_data[, 1:ncol(train_data)-1],y=as.factor(
    train_data$COURSE))
train_data_new <- train_data_new %>% rename(COURSE='Class')

# Create a vector of medicine names
demo_names <- c('SEX','HISPANIC','RACE','PRIMLANG','EDUC','MARISTAT','
    INDEPEND','RESIDENC','AGE','NACCLIVS')

# Create a list to store the ROC curves and AUC values
roc_auc <- numeric(length(medical_names))

# Loop through each medicine and build the logistic regression model
for (i in 1:length(demo_names)) {
  # Extract the current medicine name
  demo <- demo_names[i]

  # Build the model with the current medicine as a covariate
  formula <- paste("COURSE ~", demo)

  model <- glm(formula, family = binomial, data = train_data_new)

  # Make predictions using the model
  predictions <- predict(model, newdata = test_data, type = "response")
  predicted_classes <- ifelse(predictions > 0.5, 1, 0)

  # Calculate the ROC curve and ROC-AUC
  roc_curve <- roc(test_data$COURSE, predictions)
  roc_auc[i] <- auc(roc_curve)


   # Set font family
   par(family = "Times New Roman")
   # Plot the ROC curve for each logisic regressio model as covariate is
    each variable of demograpgic information
    plot(roc_curve, main = paste("ROC Curve for", demo,"AUC =", round(roc_
    auc[i], 2)), col = "blue", lwd = 2,  cex.main = 1.3,title.adj = 0.5,cex.
    lab = 1.3, cex.axis = 1.3)
    lines(x = c(0, 1), y = c(0, 1), col = "gray", lty = 2)
    par(cex.axis = 1.5)
}

# Create a vector of medical histories
medical_names <- c('NACCMOM','NACCDAD','NACCFAM',
'NACCFFTD' ,'DIABETES', 'INCONTU','DEP2YRS',
'NACCCOGF','COGMODE','BEAPATHY','HYPERTEN','HYPERCHO')

# Create a list to store the ROC curves and AUC values
roc_auc <- numeric(length(medical_names))

# Loop through each medicine and build the logistic regression model
for (i in 1:length(medical_names)) {
  # Extract the current medicine name
  medical <- medical_names[i]
```

```
56
57    # Build the model with the current medicine as a covariate
58    formula <- paste("COURSE ~", medical)
59    model <- glm(formula, family = binomial, data = train_data_new)
60
61    # Make predictions using the model
62    predictions <- predict(model, newdata = test_data, type = "response")
63    predicted_classes <- ifelse(predictions > 0.5, 1, 0)
64
65    # Calculate the ROC curve and ROC-AUC
66    roc_curve <- roc(test_data$COURSE, predictions)
67    roc_auc[i] <- auc(roc_curve)
68
69     # Set font family
70     par(family = "Times New Roman")
71     # Plot the ROC curve
72      plot(roc_curve, main = paste("ROC Curve for", medical ,"AUC =", round(
      roc_auc[i], 2)), col = "blue", lwd = 2,  cex.main = 1.3,title.adj = 0.5,
      cex.lab = 1.3, cex.axis = 1.3)
73      lines(x = c(0, 1), y = c(0, 1), col = "gray", lty = 2)
74      par(cex.axis = 1.5)
75  }
76
77  # Create a vector of medicine names
78  medicine_names_new <- c('DONEPEZIL',
79   'MEMANTINE',
80   'RIVASTIGMINE',
81   'CHOLECALCIFEROL',
82   'QUETIAPINE',
83   'CALCIUM.CARBONATE'
84  )
85
86  # Create a list to store the ROC curves and AUC values
87  roc_auc <- numeric(length(medicine_names_new))
88
89  # Loop through each medicine and build the logistic regression model
90  for (i in 1:length(medicine_names_new)) {
91    # Extract the current medicine name
92    medicine <- medicine_names_new[i]
93
94    # Build the model with the current medicine as a covariate
95    formula <- paste("COURSE ~", medicine)
96    model <- glm(formula, family = binomial, data = train_data_new)
97
98    # Make predictions using the model
99    predictions <- predict(model, newdata = test_data, type = "response")
100   predicted_classes <- ifelse(predictions > 0.5, 1, 0)
101
102   # Calculate the prediction accuracy
103   accuracy[i] <- mean(predicted_classes == test_data$COURSE)
104
105     # Calculate the ROC curve and ROC-AUC
106   roc_curve <- roc(test_data$COURSE, predictions)
107   roc_auc[i] <- auc(roc_curve)
108
109
110    par(family = "Times New Roman")
111     # Plot the ROC curve
112     plot(roc_curve, main = paste("ROC Curve for", medicine,"AUC =", round(
      roc_auc[i], 2)), col = "blue", lwd = 2,  cex.main = 1.3,title.adj = 0.5,
```

```
    cex.lab = 1.3, cex.axis = 1.3)
113     lines(x = c(0, 1), y = c(0, 1), col = "gray", lty = 2)
114     par(cex.axis = 1.5)
115
116 }
```

Listing 1: R code of producing AUC-ROC of each logistic regression model as covariate is each variable of demographic information and medical history and medicine.

```
1
2 # Define features and target variables
3 X = filtered_data.drop('COURSE', axis=1)
4 y = filtered_data['COURSE']
5
6 # Subset of features
7 X_new = X[['INDEPEND','RESIDENC','AGE',
8 'PRIMLANG','MARISTAT','RACE',
9 'EDUC','SEX','HISPANIC',
10 'NACCLIVS','NACCMOM','NACCFAM',
11 'DIABETES','INCONTU','DEP2YRS',
12 'NACCCOGF','BEAPATHY','HYPERCHO','DONEPEZIL', 'MEMANTINE', 'RIVASTIGMINE','
    QUETIAPINE','CALCIUM.CARBONATE',]]
13
14 # Split the data into training and testing sets
15 X_train, X_test, y_train, y_test = train_test_split(X_new, y, test_size=0.2,
     random_state=42)
16
17 # Initialize ADASYN
18 adasyn = ADASYN(sampling_strategy='auto', random_state=42)
19
20 # Apply ADASYN to balance the dataset
21 X_train_resampled, y_train_resampled = adasyn.fit_resample(X_train, y_train)
22
23 # Use StratifiedShuffleSplit to obtain the test dataset that is
    representative of the real-world data
24 splitter=StratifiedShuffleSplit(n_splits=1,random_state=12, test_size=0.20)
25 for train,test in splitter.split(X_new,y):
26     X_train = X_new.iloc[train]
27     y_train = y.iloc[train]
28     X_test = X_new.iloc[test]
29     y_test = y.iloc[test]
30
31 # Create a Random Forest Classifier
32 clf = RandomForestClassifier(max_depth=None, random_state=5,
33 n_estimators=500)
34
35 # Fit the classifier on the resampled training data
36 clf.fit(X_train_resampled, y_train_resampled)
37
38 # Make predictions on the test data
39 y_pred = clf.predict(X_test)
40
41 # Evaluate the model
42 # Produce classification matrix
43 print(classification_report(y_test,y_pred))
```

Listing 2: Python code of developing a random forest model and evaluating it.

```
1
2 # Define features and target variables
3 X = filtered_data.drop('COURSE', axis=1)
```

```
y = filtered_data['COURSE']

# Subset of features
X_new = X[['INDEPEND','RESIDENC','AGE','PRIMLANG','MARISTAT','RACE','EDUC','SEX','HISPANIC','NACCLIVS',
        'NACCMOM','NACCFAM','COGMODE','NACCDAD'
,'DIABETES', 'INCONTU','DEP2YRS','NACCCOGF','BEAPATHY','HYPERCHO','DONEPEZIL', 'MEMANTINE','CALCIUM.CARBONATE',]]

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_new, y, test_size=0.2, random_state=42)

# Initialize ADASYN
adasyn = ADASYN(sampling_strategy='auto', random_state=42)

# Apply ADASYN to balance the dataset
X_train_resampled, y_train_resampled = adasyn.fit_resample(X_train, y_train)

# Use StratifiedShuffleSplit to obtain the test dataset that is representative of the real-world data
splitter=StratifiedShuffleSplit(n_splits=1,random_state=12, test_size=0.20)
for train,test in splitter.split(X_new,y):
    X_train = X_new.iloc[train]
    y_train = y.iloc[train]
    X_test = X_new.iloc[test]
    y_test = y.iloc[test]

# Train a random forest classifier
rf_classifier = RandomForestClassifier(max_depth=None, random_state=5, n_estimators=500)
rf_classifier.fit(X_train_resampled, y_train_resampled)

# Predict probabilities of positive class on the test set
y_scores_1 = rf_classifier.predict_proba(X_test)[:, 1]

# Compute precision-recall curve
precision_1, recall_1, thresholds_1 = precision_recall_curve(y_test, y_scores_1)

# Compute average arecision score
avg_precision_1 = average_precision_score(y_test, y_scores_1)

# Train a decision tree Classifier
dt_classifier = DecisionTreeClassifier(max_depth= None, max_features= 'log2', min_samples_leaf= 1, min_samples_split=10)
dt_classifier.fit(X_train_resampled, y_train_resampled)

# Predict probabilities of positive class on the test set
y_scores_2 = dt_classifier.predict_proba(X_test)[:, 1]

# Compute precision-recall curve
precision_2, recall_2, thresholds_2 = precision_recall_curve(y_test, y_scores_2)

# Compute average precision score
avg_precision_2 = average_precision_score(y_test, y_scores_2)

# Train a logistic regression classifier
clf_lr = LogisticRegression(random_state=0,C=0.1,penalty='l2').fit(
```

```
    X_train_resampled, y_train_resampled)

# Predict probabilities of positive class on the test set
y_scores_3 = clf_lr.predict_proba(X_test)[:, 1]

# Compute precision-recall curve
precision_3, recall_3, thresholds_3 = precision_recall_curve(y_test,
    y_scores_3)

# Compute average precision score
avg_precision_3 = average_precision_score(y_test, y_scores_3)


# Create two subplots side by side
fig, axes = plt.subplots(1, 3, figsize=(40, 15))

# Set font family
plt.rcParams['font.family'] = 'Times New Roman'

# Plot the precision-recall curve of logistic regression model
axes[0].step(recall_3, precision_3, color='b', alpha=0.8, where='post')
axes[0].fill_between(recall_3, precision_3, alpha=0.2, color='b')
axes[0].set_xlabel('Recall',fontsize=50)
axes[0].set_ylabel('Precision',fontsize=50)
axs[0].set_xticklabels(axs[0].get_xticklabels(), fontsize=50)
axes[0].set_title('Logistic Regressio Model (AP = {:.2f})'.format(
    avg_precision_3),fontsize=50)
axes[0].set_ylim([0.0, 1.05])
axes[0].set_xlim([0.0, 1.0])
axes[0].tick_params(axis='both', labelsize=50)
axes[0].grid(True)

# Plot the precision-recall curve of decision tree model
axes[1].step(recall_2, precision_2, color='b', alpha=0.8, where='post')
axes[1].fill_between(recall_2, precision_2, alpha=0.2, color='b')
axes[1].set_xlabel('Recall',fontsize=50)
axes[1].set_ylabel('Precision',fontsize=50)
axes[1].set_title('Decision Tree Model (AP = {:.2f})'.format(avg_precision_2
    ),fontsize=50)
axs[1].set_xticklabels(axs[1].get_xticklabels(), fontsize=50)
axes[1].set_ylim([0.0, 1.05])
axes[1].set_xlim([0.0, 1.0])
axes[1].tick_params(axis='both', labelsize=50)
axes[1].grid(True)

# Plot the precision-recall curve of random forest model
axes[2].step(recall_1, precision_1, color='b', alpha=0.8, where='post')
axes[2].fill_between(recall_1, precision_1, alpha=0.2, color='b')
axes[2].set_xlabel('Recall',fontsize=50)
axes[2].set_ylabel('Precision',fontsize=50)
axes[2].set_title('Random Forest Model (AP = {:.2f})'.format(avg_precision_1
    ),fontsize=50)
axs[2].set_xticklabels(axs[2].get_xticklabels(), fontsize=50)
axes[2].set_ylim([0.0, 1.05])
axes[2].set_xlim([0.0, 1.0])
axes[2].tick_params(axis='both', labelsize=50)
axes[2].grid(True)
```

Listing 3: Python code of producing the precision-recall plot of logistic regression and decision tree and random forest model.

References

[1] World Health Organization. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia, 2023.

[2] Alzheimer's Association. 2023 alzheimer's disease facts and figures. https://www.alz.org/media/Documents/alzheimers-facts-and-figures.pdf, 2023.

[3] Alzheimer's Disease International. World alzheimer report 2015: The global impact of dementia. https://www.alzint.org/u/WorldAlzheimerReport2015.pdf, 2015.

[4] L Rizzi, I Rosset, and M Roriz-Cruz. Global epidemiology of dementia: Alzheimer's and vascular types. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4095986/, 2014.

[5] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, and et al. The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. https://www.sciencedirect.com/science/article/abs/pii/S1552526011001014.

[6] R. C. Petersen, J.C. Stevens, M. Ganguli, E. G. Tangalos, J.L. Cummings, and S. T. DeKosky. Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) [retired]. https://n.neurology.org/content/56/9/1133.short.

[7] Alzheimer's Society. How is mild cognitive impairment treated? https://www.alzheimers.org.uk/about-dementia/types-dementia/mild-cognitive-impairment-treatments.

[8] Blood pressure and alzheimer's risk: What's the connection? https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/blood-pressure-and-alzheimers-risk-whats-the-connection, Mar 2022.

[9] Beth Oller. Depression and alzheimer's disease. https://familydoctor.org/depression-and-alzheimersdisease/: :text=DepressionApr 2022.

[10] Alzheimer's Association. Diabetes and cognitive decline. https://alz.org/media/Documents/alzheimers-dementia-diabetes-cognitive-decline-ts.pdf, journal=Diabetes and cognitive decline, 2023.

[11] Barry Reisberg, Rachelle Doody, Albrecht Stöffler, Frederick Schmitt, and Hans Jörg Möbius. Memantine in moderate-to-severe alzheimer's disease. https://www.nejm.org/doi/full/10.1056/NEJMoa013128, 2003.

[12] Mohammad Fawaz Siddiqi. Following the data science methodology. https://developer.ibm.com/blogs/following-the-data-science-methodology/, 2021.

[13] Handling imbalanced data with smote and near miss algorithm in python. https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/, Jan 2023.

[14] What is the default threshold in sklearn logistic regression? https://www.geeksforgeeks.org/what-is-the-default-threshold-in-sklearn-logistic-regression/, Jun 2023.

[15] Viorica Marian and Anthony Shook. The cognitive benefits of being bilingual. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583091/, 2012.

[16] Explaining logistic regression as generalized linear model (in use as a classifier). https://www.datasciencecentral.com/explaining-logistic-regression-as-generalized-linear-model-in-use/: :text=Logisticjournal=Data Science Central, publisher=TechTarget, author=Ajitjaokar, year=2019,.

[17] Beyond logistic regression: Generalized linear models (glm). https://online.stat.psu.edu/stat504/lesson/beyond-logistic-regression-generalized-linear-models-glm.

[18] Rivastigmine (oral route). https://www.mayoclinic.org/drugs-supplements/rivastigmine-oral-route/side-effects/drg-20065860?p=1, Apr 2023.

[19] Get started. https://scikit-learn.org/stable/getting$_s tarted.html$.

[20] Anshul Saini. Decision tree algorithm - a complete guide. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/: :text=AJul 2023.

[21] Rachel Meltzer. What is random forest? [beginner's guide + examples]. https://careerfoundry.com/en/blog/data analytics/what-is-random-forest/: :text=2.-,HowJul 2021.

[22] Adaptive synthetic sampling (adasyn). https://www.activeloop.ai/resources/glossary/adaptive-synthetic-sampling-adasyn/.

[23] sklearn.model$_s election.stratifiedshuffle split.https : //scikit - learn.org/stable/modules/generate$

[24] Rahul Shah. Tune hyperparameters with gridsearchcv. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/, Jul 2022.

[25] Noa Azaria. Feature importance: 7 methods and a quick tutorial. https://www.aporia.com/learn/feature-importance/feature-importance-7-methods-and-a-quick-tutorial/, 2023.

[26] Alzheimer's: Drugs help manage symptoms. https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048103: :text=Donepezil2022.