Project Report

Uncovering the Predictors of Alzheimer’s Using Statistical Methods

Abstract

This project investigates the cognitive and behavioral predictors of Alzheimer's disease through the application of rigorous statistical analysis techniques. Alzheimer's disease is a progressive and irreversible neurodegenerative condition that significantly impairs memory, thinking, and behavior, affecting millions of individuals worldwide. Timely diagnosis is critical, as early intervention can slow disease progression and improve quality of life. Our study utilizes a cross-sectional dataset comprising 2,149 patient records obtained from Kaggle, incorporating 35 variables encompassing demographic details, lifestyle factors, cognitive assessments, and neurobehavioral symptoms.

The primary focus was to assess the association between MMSE (Mini-Mental State Examination) scores and behavioral symptoms, specifically memory complaints, confusion, forgetfulness, and difficulty completing tasks, and their combined influence on Alzheimer's diagnosis. A range of statistical methods was employed, including logistic regression, Spearman correlation, Chi-square tests, Mann-Whitney U tests, and proportion tests. Findings revealed that both MMSE scores and memory complaints were statistically significant predictors of a positive Alzheimer's diagnosis. In contrast, symptoms such as confusion, forgetfulness, and difficulty completing tasks did not show strong predictive value.

The study highlights the utility of MMSE as a reliable screening tool and emphasizes the importance of subjective memory complaints in identifying individuals at risk. Limitations include reliance on self-reported data, which may introduce bias, and limited generalizability due to the dataset's demographic homogeneity. Future research should employ longitudinal designs and more diverse populations to validate and expand upon these findings.

Uncovering the Predictors of Alzheimer's Using Statistical Methods Alzheimer's disease is a brain disorder that slowly damages memory, thinking, and behavior. It is the most common cause of dementia and affects millions of people around the world (Choe et al., 2020). Early detection of Alzheimer's is very important because it allows patients to get proper care, support, and treatment sooner. This study focuses on understanding whether certain symptoms both cognitive and behavioral can help predict if a person is likely to be diagnosed with Alzheimer's. The main cognitive symptom studied is the Mini-Mental State Examination (MMSE) score, while the behavioral symptoms include memory complaints, confusion, forgetfulness, and difficulty completing tasks (Mayo Clinic, 2024).

The **research question** is: Which cognitive (MMSE) and behavioral symptoms are the strongest predictors of an Alzheimer's disease diagnosis?

Dataset Description

The dataset used for this study is titled the Alzheimer's Disease Dataset, sourced from Kaggle, and contains 2,149 patient records with 35 variables related to demographic, lifestyle, cognitive, behavioral, and medical characteristics. This cross-sectional dataset is well-suited for analyzing factors associated with Alzheimer's disease diagnosis. The **variables** include: Patient_ID, Age, Gender, Education_Level, Marital_status, Smoking, Alcohol, Physical_Activity, BMI, High_Cholesterol, Hypertension, Diabetes, Stroke, Cardiovascular_Disease, Respiratory_Problems, Depression, Memory_Complaints, Forgetfulness, Confusion, Difficulty_Completing_Tasks, Trouble_Concentrating, Irritability, Anxiety, Sleep_Problems, MMSE, Diagnosis, Number_of_Children, Living_Condition, Country, Occupation, Financial_Status, Family_History_AD, Language, Hearing_Impairment, and Vision_Impairment (El Kharoua, 2024).

The **target variable** in this study is Diagnosis (0 = No Alzheimer's, 1 = Yes), and the main predictors of interest include MMSE scores and behavioral symptoms such as Memory_Complaints, Confusion, Forgetfulness, and Difficulty_Completing_Tasks. The dataset contains both numerical and categorical variables, with MMSE and BMI being continuous, while most others are binary indicators. Before analysis, preprocessing steps such as missing value handling, type conversion, and normality testing were conducted to ensure data quality and model readiness. ***Link to the dataset***

https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset

Pre-Processing Steps

Checked for missing values and found none. We also converted categorical variables into binary indicators for further analysis. The code snippet for checking missing values and converting can be seen in the Appendix, Figure A1, and Figure A2, respectively.

Summary statistics

Figure A3 in the Appendix presents summary statistics. MMSE Mean: 14.76 (range: \~0 to 30), Behavioral symptoms had low average values, indicating low prevalence. Diagnosis: The Majority of individuals were not diagnosed with AD.

Visualizations

Figure A4 in the Appendix presents bar charts for diagnosis, memory complaints, confusion, forgetfulness, and difficulty completing tasks, along with a histogram for MMSE scores. The diagnosis chart reveals a class imbalance, with most individuals not diagnosed with Alzheimer’s. Similarly, the bar charts for behavioral symptoms show that the majority of participants did not report memory complaints, confusion, forgetfulness, or task difficulty, highlighting that these symptoms are relatively infrequent in the dataset.

The histogram of MMSE scores shows a broad distribution ranging from near 0 to 30, indicating considerable variability in cognitive function. The shape of the distribution does not suggest a strong skew, which supports the need for non-parametric analyses in later sections.

Statistical Methods

Logistic Regression

A binary logistic regression model was built to assess the relationship between the likelihood of an Alzheimer’s diagnosis (dependent variable) and several predictors, MMSE score and binary behavioral symptoms (memory complaints, confusion, forgetfulness, and difficulty completing tasks).

Justification. Logistic regression is appropriate when the dependent variable is binary (i.e., diagnosed or not diagnosed). It quantifies the effect of each predictor on the log-odds of the outcome and provides interpretable odds ratios.

Relevance. This method allows us to estimate how a unit change in a predictor (e.g., MMSE score) affects the odds of an Alzheimer’s diagnosis.

Chi-Square Tests

Chi-square tests were conducted to examine the independence between categorical behavioral symptoms and Alzheimer's diagnosis.

Justification. This test is ideal for evaluating associations between two categorical variables.

Relevance. Helps determine if behavioral symptoms like memory complaints are statistically associated with diagnosis outcomes.

Mann-Whitney U Test

This non-parametric test was used to compare MMSE scores between diagnosed and nondiagnosed groups.

Justification. Since MMSE is a continuous variable but not normally distributed (confirmed via Shapiro-Wilk test), a non-parametric test was appropriate.

Relevance. Identifies whether cognitive performance, as measured by MMSE, significantly differs between the two diagnostic groups.

Spearman Rank Correlation

Spearman's correlation was used to examine the monotonic relationship between MMSE scores and the diagnosis variable.

Justification. Suitable for non-normal data and ordinal/continuous variables.

Relevance. Measures the strength and direction of the relationship between cognitive scores and diagnosis likelihood.

Proportion Test (prop.test)

This test evaluated whether the proportion of individuals reporting memory complaints significantly differed from 50%.

Justification. Useful for estimating confidence intervals and testing hypotheses for proportions in binary variables.

Relevance. Assesses how prevalent memory complaints are within the dataset, aiding in contextualizing their role as a predictor.

Results

Since the **Shapiro-Wilk test** (W = 0.95253, $p < 2.2e-16$) and **QQ-plot** indicated that MMSE scores were not normally distributed, we decided to proceed with non-parametric methods (See Figures A5 and A6 in the Appendix).

We began with a **Spearman rank correlation** to assess the relationship between MMSE scores and diagnosis. There was a moderate negative correlation ($\rho = -0.24$), which suggests that lower MMSE scores are linked with a higher likelihood of an Alzheimer's diagnosis. We also observed a positive correlation between memory complaints and diagnosis ($\rho = 0.31$), indicating that individuals reporting memory issues were more likely to be diagnosed. The corresponding Spearman correlation graph is provided in the appendix, Figure A7, and correlation heatmap, Figure A8.

Next, we performed a **logistic regression** with diagnosis as the dependent variable as it was categorical(binary) and included MMSE, memory complaints, confusion, forgetfulness, and task difficulty as predictors. Among these, MMSE ($p < 2e-16$) and memory complaints ($p < 2e16$) were significant. The **odds ratio for MMSE** was 0.93, meaning that each additional point in MMSE score reduced the odds of diagnosis by about 6.6%. For memory complaints, the odds ratio was 5.26, suggesting that individuals reporting memory issues were over five times more likely to be diagnosed. The model fit significantly improved, with the residual deviance dropping from 2792.3 to 2457.0 ($p < 0.001$). The logistic regression results are shown in Figure A10, and the odds ratio in Figure A11 in the Appendix.

We also calculated the **predicted probabilities** from the model and found that as MMSE scores increase, the predicted probability of diagnosis generally decreases. This aligns with the negative association we observed in the model, where the odds ratio for MMSE is less than 1 (0.9342707). Specifically, the negative coefficient (-0.067989) for MMSE suggests that better cognitive performance (higher MMSE) is linked to a lower likelihood of receiving a positive diagnosis. The predicted probability is shown in Figure A12 in the Appendix.

To further investigate, we used **Chi-square tests** to examine the association between each symptom and diagnosis. Memory complaints showed a strong and significant association ($X^2 = 200.62$, $p < 2.2e-16$), while confusion, forgetfulness, and task difficulty were not significantly associated (p-values > 0.5). The Chi-square test results are shown in Figure A13 in the Appendix.

To compare MMSE scores between diagnosed and non-diagnosed individuals, we ran a **Mann-Whitney U test**. The test result ($W = 678416$, $p < 2.2e-16$) confirmed that there was a significant difference in MMSE scores between the two groups. The result of the Mann-Whitney U test is shown in Figure A9 in the Appendix.

Lastly, we conducted **a one-sample proportion test** to check if the proportion of patients reporting memory complaints differed from 0.5. The observed proportion was 0.208 (95% CI: 0.1911-0.2259), and the result was highly significant ($p < 2.2e-16$), indicating that fewer individuals reported memory complaints than what we would expect under the null hypothesis.

The result of the proportion test can be viewed in the Appendix, Figure A14.

Limitations

Logistic regression models can be sensitive to multicollinearity among predictors, which may affect the stability of the estimated coefficients. Additionally, potential bias in self-reported symptoms, such as memory complaints and forgetfulness, should be considered, as these are subjective and influenced by individual perception, which may not always reflect the true clinical condition. The model may also be subject to confounding variables that were not measured or included, such as medication use or genetic predisposition, which could influence the diagnosis of Alzheimer's. Furthermore, the results of this analysis are based on a specific dataset, and therefore, the findings may not generalize to other populations without further external validation.

Conclusion

This study highlights the importance of MMSE scores and memory complaints in predicting Alzheimer's disease. While other behavioral symptoms were not significant predictors, MMSE and memory complaints offer meaningful insights into early diagnosis. These results contribute to clinical efforts in refining screening processes and directing attention to effective early indicators.

References

Choe, Y. M., Lee, B. C., Choi, I. G., Suh, G. H., Lee, D. Y., & Kim, J. W. (2020). MMSE subscale scores as useful predictors of AD conversion in mild cognitive impairment. *Neuropsychiatric Disease and Treatment*, *16*, 1767-1775. https://doi.org/10.2147/NDT.S263702

El Kharoua, R. (2024). *Alzheimer's disease dataset: Comprehensive health information for Alzheimer's disease* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8668279

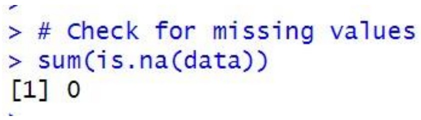
Mayo Clinic. (2024, November 8). Alzheimer's disease - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptomscauses/syc-20350447

Appendix

Code Snippets, Statistical Results, and Visualizations

Figure A1

Missing Values

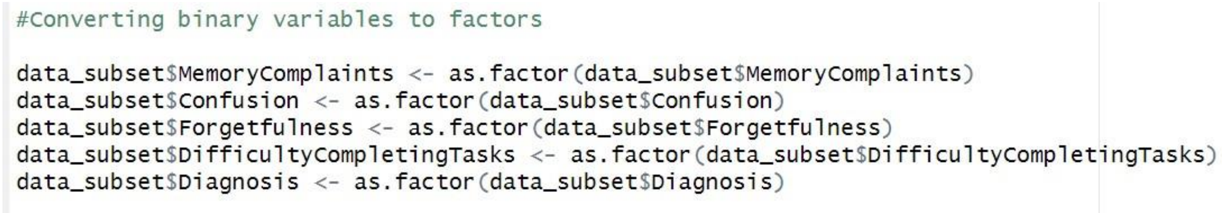


```
> # Check for missing values
> sum(is.na(data))
[1] 0
```

Note. Used the is.na code to check null values, and there were none in the dataset.

Figure A2

R code for converting binary variables to factors

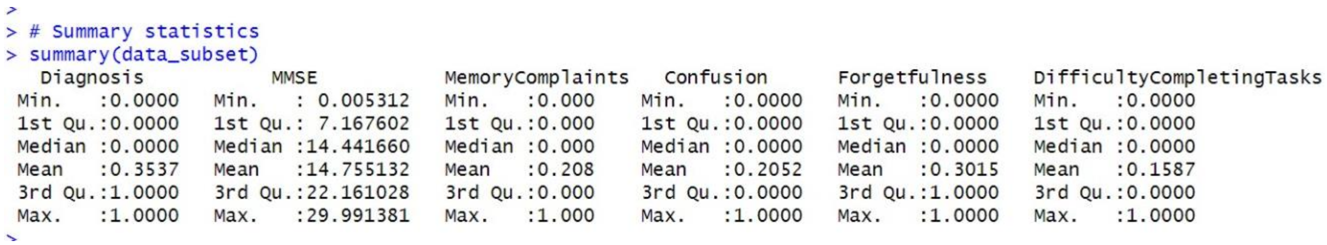


```
#Converting binary variables to factors

data_subset$MemoryComplaints <- as.factor(data_subset$MemoryComplaints)
data_subset$Confusion <- as.factor(data_subset$Confusion)
data_subset$Forgetfulness <- as.factor(data_subset$Forgetfulness)
data_subset$DifficultyCompletingTasks <- as.factor(data_subset$DifficultyCompletingTasks)
data_subset$Diagnosis <- as.factor(data_subset$Diagnosis)
```

Note. We converted binary variables to factors before logistic regression to treat them as categorical variables, not numeric ones, ensuring correct modeling and interpretation.

Figure A3

Summary Statistics

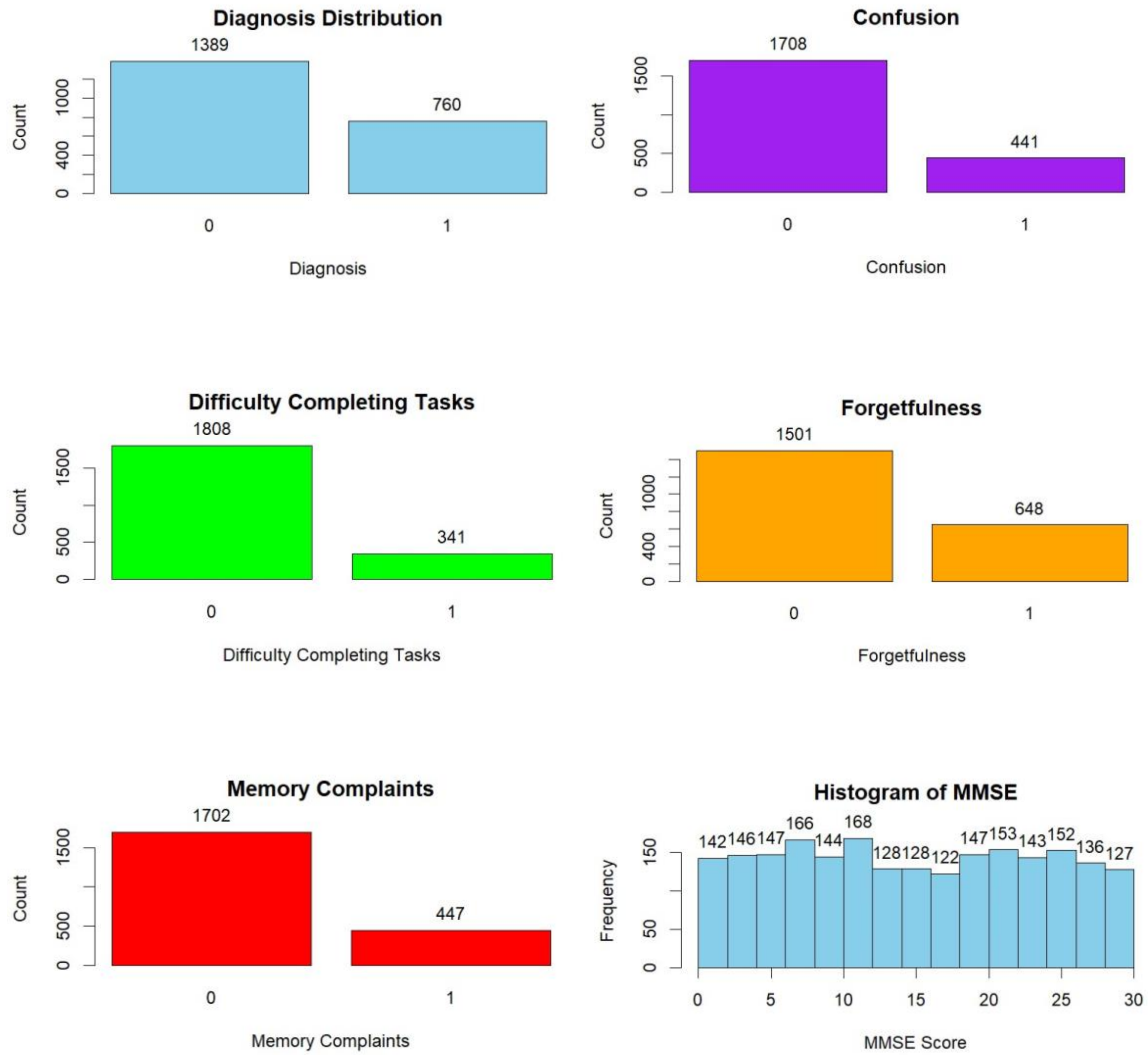


```
>
> # Summary statistics
> summary(data_subset)
   Diagnosis            MMSE           MemoryComplaints   Confusion        Forgetfulness     DifficultyCompletingTasks
 Min.   :0.0000   Min.   : 0.005312   Min.   :0.000    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.: 7.167602   1st Qu.:0.000    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :14.441660   Median :0.000    Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   :0.3537   Mean   :14.755132   Mean   :0.208    Mean   :0.2052   Mean   :0.3015   Mean   :0.1587  
 3rd Qu.:1.0000   3rd Qu.:22.161028   3rd Qu.:0.000    3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :29.991381   Max.   :1.000    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
>
```

Note. We performed summary statistics to understand the overall distribution, central tendencies, and patterns in the dataset before further analysis.

Figure A4

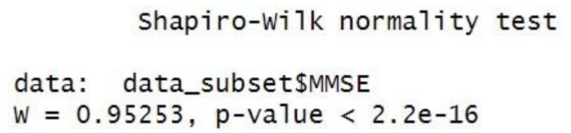
Visualizations



Note. Most individuals were not diagnosed with Alzheimer's and did not report symptoms like memory complaints, confusion, forgetfulness, or task difficulty, indicating class imbalance and low symptom prevalence. The MMSE score histogram shows a wide, fairly uniform distribution, reflecting diverse cognitive performance levels.

Figure A5

Shapiro-Wilk test results

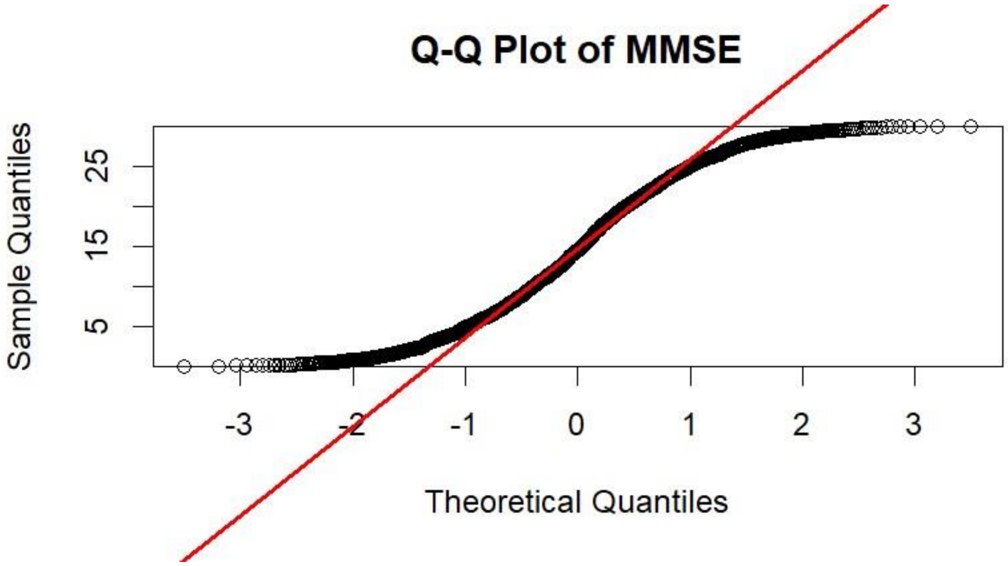


```
        Shapiro-Wilk normality test

data:  data_subset$MMSE
W = 0.95253, p-value < 2.2e-16
```

Note. The Shapiro-Wilk normality test indicated that the data is not normally distributed, suggesting that parametric tests may not be appropriate.

Figure A6

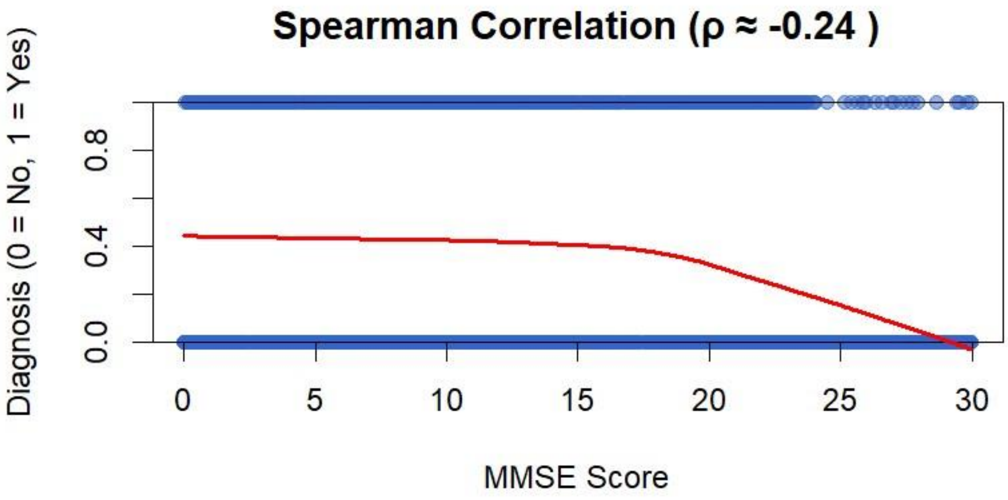
Q-Q Plot



Note. The QQ plot showed noticeable deviation from the red reference line, especially at the tails, forming a typical 'S' curve. This indicates that the MMSE scores do not follow a normal distribution.

Figure A7

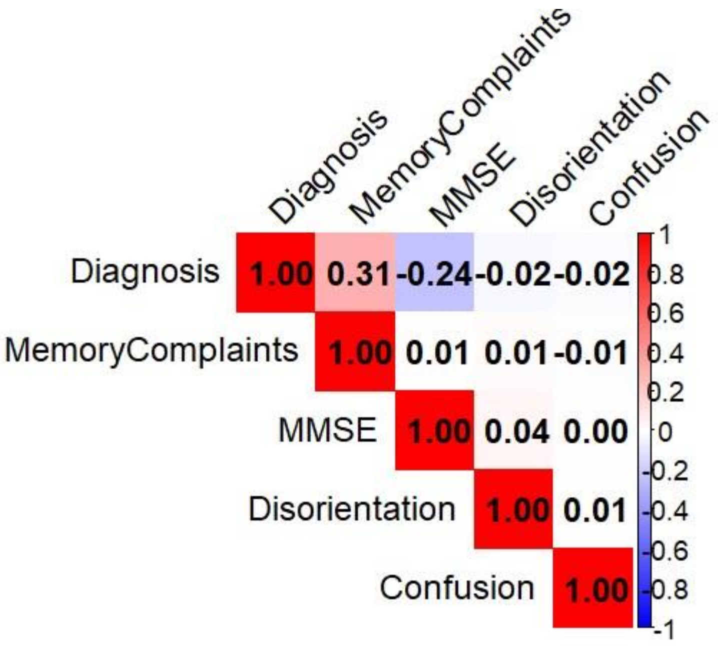
Spearman Correlation result for MMSE



Note. The graph visually represents a weak to moderate negative monotonic relationship. As the MMSE scores increase, the likelihood of an Alzheimer's diagnosis decreases. This is shown by the downward trend in the graph.

Figure A8

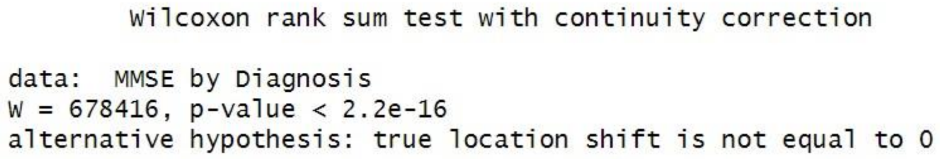
Correlation Heatmap



Note. The correlation heatmap shows a moderate positive correlation between Diagnosis and Memory Complaints (0.31), and a moderate negative correlation with MMSE scores (-0.24)

Figure A9

Mann-Whitney U Test results

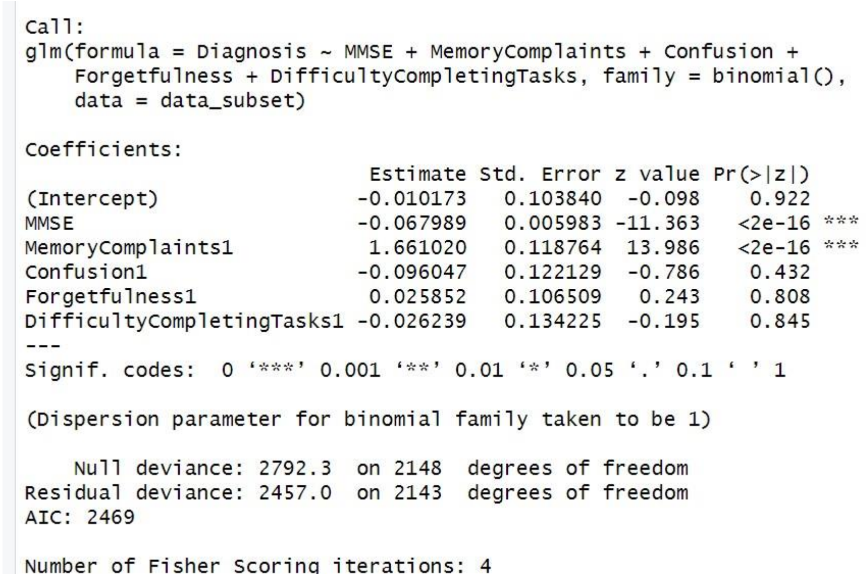


```
        Wilcoxon rank sum test with continuity correction

data:  MMSE by Diagnosis
W = 678416, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Note. The Mann-Whitney U test (W = 678416, p < 2.2e-16) shows a significant difference in MMSE scores between Alzheimer's and non-Alzheimer's groups, so we reject the null hypothesis.

Figure A10

Logistic Regression results

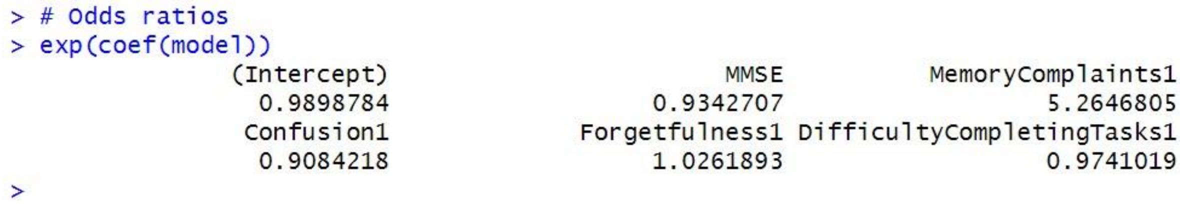


```
Call:
glm(formula = Diagnosis ~ MMSE + MemoryComplaints + Confusion +
    Forgetfulness + DifficultyCompletingTasks, family = binomial(),
    data = data_subset)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.010173   0.103840  -0.098    0.922
MMSE                      -0.067989   0.005983 -11.363   <2e-16 ***
MemoryComplaints1          1.661020   0.118764  13.986   <2e-16 ***
Confusion1                -0.096047   0.122129  -0.786    0.432
Forgetfulness1             0.025852   0.106509   0.243    0.808
DifficultyCompletingTasks1 -0.026239   0.134225  -0.195    0.845
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2792.3  on 2148  degrees of freedom
Residual deviance: 2457.0  on 2143  degrees of freedom
AIC: 2469

Number of Fisher Scoring iterations: 4
```

Note. Logistic regression predicts the binary outcome (Diagnosis) using MMSE scores and symptom variables, estimating each predictor's effect on the odds of an Alzheimer’s diagnosis.

Figure A11 *Odds*

Ratio

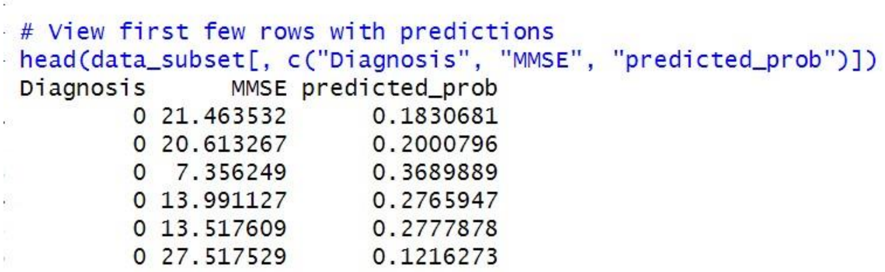


```
> # Odds ratios
> exp(coef(model))
               (Intercept)                       MMSE          MemoryComplaints1
                 0.9898784                  0.9342707                  5.2646805
                Confusion1             Forgetfulness1 DifficultyCompletingTasks1
                 0.9084218                  1.0261893                  0.9741019
>
```

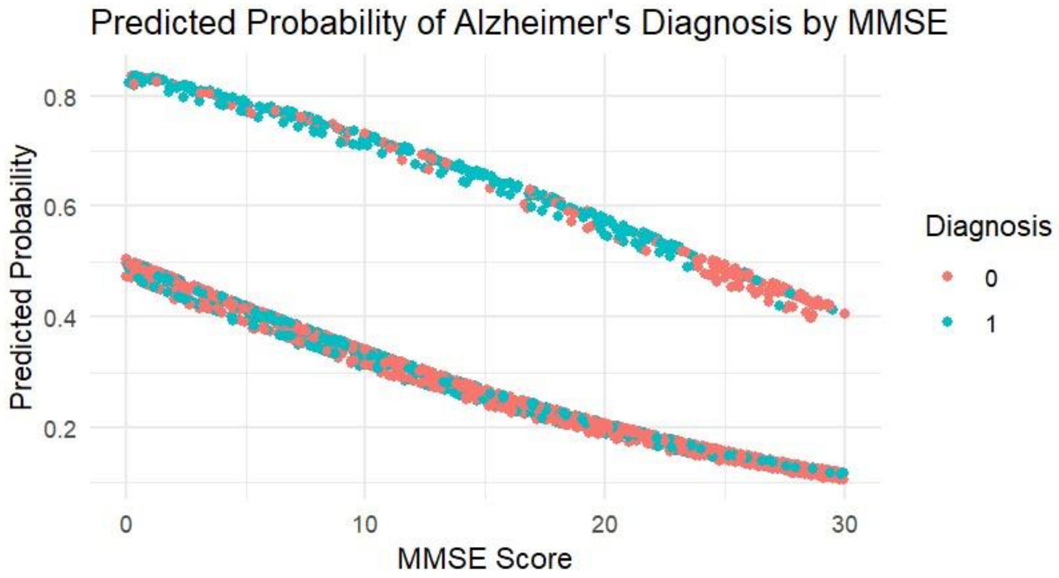
Note. The image shows that higher MMSE scores lower diagnosis odds, while memory complaints increase them. Other symptoms have minimal impact.

Figure A12

Predicted Probability results



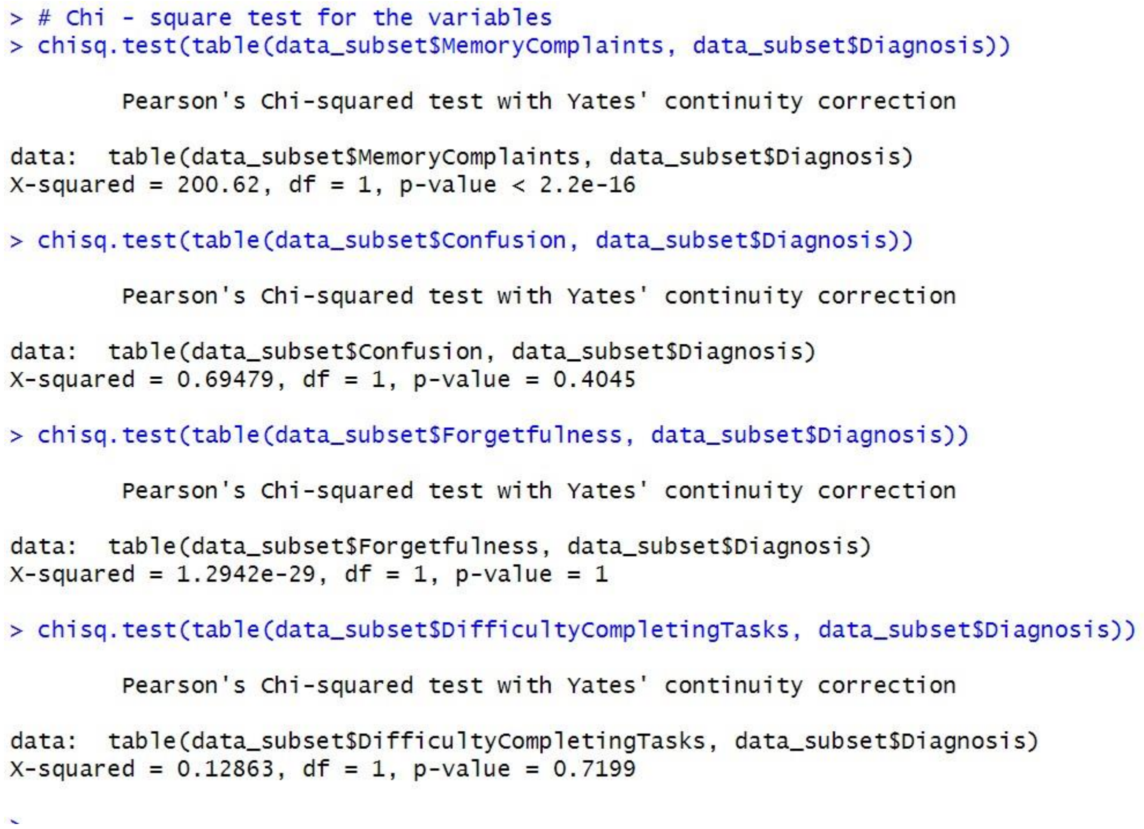
```
# View first few rows with predictions
head(data_subset[, c("Diagnosis", "MMSE", "predicted_prob")])
Diagnosis      MMSE predicted_prob
        0 21.463532      0.1830681
        0 20.613267      0.2000796
        0  7.356249      0.3689889
        0 13.991127      0.2765947
        0 13.517609      0.2777878
        0 27.517529      0.1216273
```



Note. The predicted probability for each patient shows that as MMSE scores increase, the likelihood of an Alzheimer’s diagnosis decreases, supporting the negative association found in the model (odds ratio < 1).

Figure A13

Chi-Square test results

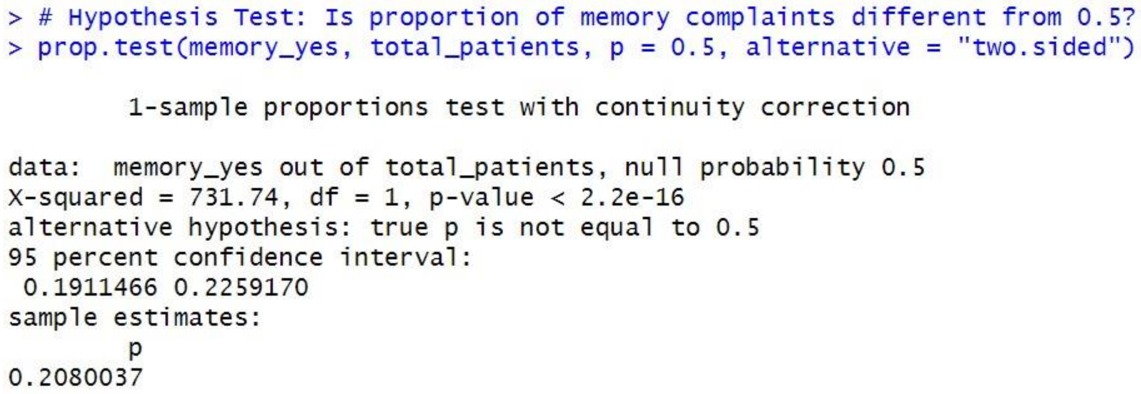


```
> # Chi - square test for the variables
> chisq.test(table(data_subset$MemoryComplaints, data_subset$Diagnosis))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(data_subset$MemoryComplaints, data_subset$Diagnosis)
X-squared = 200.62, df = 1, p-value < 2.2e-16

> chisq.test(table(data_subset$Confusion, data_subset$Diagnosis))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(data_subset$Confusion, data_subset$Diagnosis)
X-squared = 0.69479, df = 1, p-value = 0.4045

> chisq.test(table(data_subset$Forgetfulness, data_subset$Diagnosis))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(data_subset$Forgetfulness, data_subset$Diagnosis)
X-squared = 1.2942e-29, df = 1, p-value = 1

> chisq.test(table(data_subset$DifficultyCompletingTasks, data_subset$Diagnosis))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(data_subset$DifficultyCompletingTasks, data_subset$Diagnosis)
X-squared = 0.12863, df = 1, p-value = 0.7199
```

Note. The Chi-Square Test shows a significant association between memory complaints and diagnosis, while confusion, forgetfulness, and difficulty completing tasks have no significant association.

Figure A14

Proportion test results for memory complaints



```
> # Hypothesis Test: Is proportion of memory complaints different from 0.5?
> prop.test(memory_yes, total_patients, p = 0.5, alternative = "two.sided")

        1-sample proportions test with continuity correction

data:  memory_yes out of total_patients, null probability 0.5
X-squared = 731.74, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1911466 0.2259170
sample estimates:
        p 
0.2080037 
```

Note. The proportion test reveals that 20.8% of patients have memory complaints, with a 95% confidence interval of 0.1911 to 0.2259. The extremely small p-value (< 2.2e-16) indicates a significant difference from the null hypothesis, confirming that the proportion of patients with memory complaints is not