Checkmating Alzheimer's Disease:

Using Chess, Biology, and Data Science to Diagnose the Most Common Cause of Dementia

Michael Murphy[1], Aryan Mittal[2], John Ivanov[3]

Palo Alto High School[1], Portola High School[2], Tesoro High School[3]

Abstract

The purpose of this research project is to identify and evaluate different biological and cognitive signals of Alzheimer's Disease (AD). After reviewing existing literature about the different factors associated with AD, 23 variables of interest were selected from a dataset provided by the National Alzheimer's Coordinating Center (NACC) and used to construct a logistic regression model in R. Subsequently, the model was generalized using Lasso (Fig. 1a), and the resulting model (Fig. 1b) utilized only the 6 most significant variables to predict AD diagnoses with an 83% accuracy rate. These 6 variables reflect that the time taken to finish a trail-making test, difficulty traveling and playing strategy games, and cerebrospinal fluid volume are all positively associated with AD, while left hippocampus volume is negatively associated with AD. These findings are crucial because they can be used to ensure that patients are quickly and correctly diagnosed and receive the necessary care to remain healthy.

Background

Alzheimer's Disease (AD) is the most common neurodegenerative disease in the world, with over 47.5 million people currently diagnosed and an average of 7.7 million new cases declared each year [2]. Although there is currently no cure for AD, various treatments exist to temporarily mitigate its cognitive effects. Despite the availability of these treatments, AD remains a serious threat to many elderly people particularly due to its difficulty to diagnose [1]. The primary symptom of AD, memory impairment, is something that is already common for elderly people [1], so finding other signals of the disease is important to be able to correctly diagnose patients and provide proper care. Furthermore, since AD (like other neurodegenerative diseases) is a progressive disease with certain thresholds of "no return", it is imperative that it is detected in its early stages to delay significant damage. This project investigated the extent to which various cognitive and biological factors can be used to accurately detect AD in patients.

Methods

1. Literature Review
   a. Existing literature on AD reveals that there are a variety of potential biological and cognitive signals for the disease, but in this project we decided to focus specifically on:
      i. A person's ability to play strategy games without assistance such as chess [2]
      ii. A person's severity of hallucinations, motor disturbances, and other psychotic symptoms [3]
      iii. A person's ability to perform basic tasks such as taxes [1]
      iv. A person's brain volume and cerebrospinal fluid volume [4]
2. Exploratory Data Analysis
   a. Variables in an NACC dataset containing information about these factors were subsequently examined in R
   b. Using visualizations (Fig. 2a & 2b) as well as Chi-Squared and T-Tests, we were able to finalize our selection on 23 variables of interest
3. Logistic Regression with Lasso
   a. A logistic regression model, with Alzheimer's diagnosis as the response variable, was created using the training data, and normalized with the "Lasso" technique
   b. (Fig. 1a) Equation for Lasso logistic regression using the negative log likelihood (L) of Beta, also known in Machine Learning as the Energy Function

$$-\log(L(\beta)) + \lambda \sum_{j=1}^{P} |\beta_j|$$

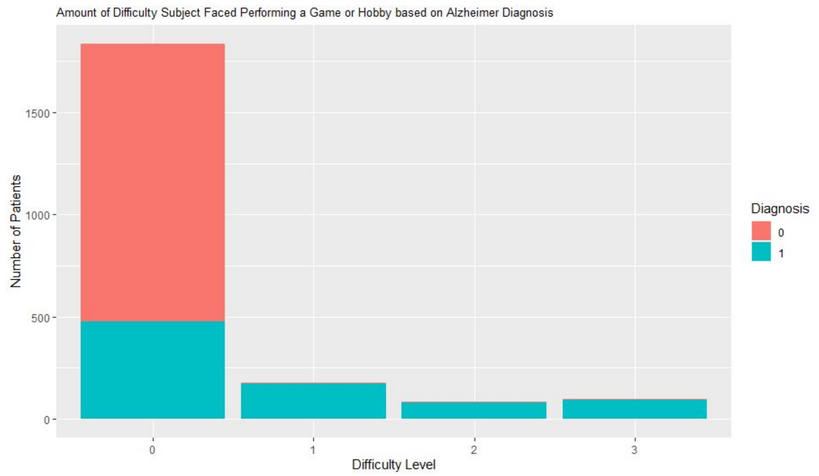
   c. Lasso regression is used to obtain the subset of predictors that minimizes prediction error for a response variable by imposing a constraint (lambda) on the model parameters that causes regression coefficients for insignificant variables to shrink toward zero (Fig. 3a, Fig. 3b)
   d. Post Lasso, the logistic regression model was generalized to use only the 6 most significant variables to generate a prediction for diagnosis.
   e. (Fig. 1b) Equation for the logistic regression model

$$\text{logit}(y) = -0.921 + 0.477(games) + 0.003(traila) + 0.005(trailb) + 0.001(csfvol) - 0.238(lhippo) + 0.888(taxes) + 0.106(travel)$$

$$P = exp(y) \,/\, (1 + exp(y))$$

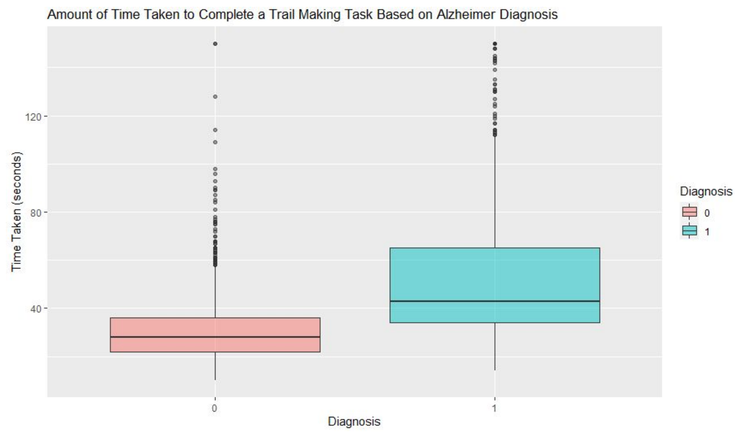
   f. The NACC dataset was divided into 2 subsets for cross-validation, with 80% of it devoted to training data and 20% of it to testing data
4. Testing
   a. Tested on the designated test data, the model was able to correctly predict whether a patient had AD **83%** of the time

Data and Results



(Fig. 2a)

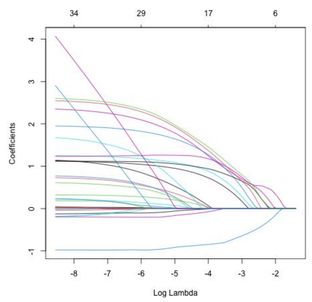
(Fig. 2a) This stacked bar plot illustrates that as a patient's difficulty performing a strategy game or hobby increased, they are more likely to have AD (Diagnosis of 1). This positive association is reflected both existing research [3] and by the coefficient for "games" in the logistic regression model (Fig. 1b).



(Fig. 2b)

(Fig. 2b) These box plots illustrate that those diagnosed with AD have a higher median for the amount of time taken to complete a trail-making test, which suggests that those who take longer on the trail-making test are more likely to have AD. This positive association was also present in the logistic regression model, as "traila" and "trailb" both have positive coefficients (Fig. 1b).

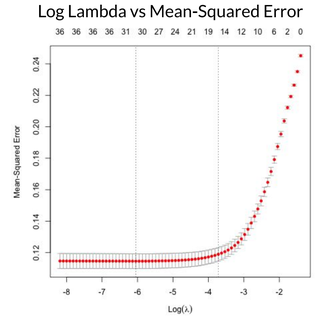
Log Lambda vs Coefficients



(Fig. 3a)

(Fig. 3a) This plot illustrates the how the different coefficients in the logistic regression model shrink towards zero as the logarithmic value of lambda (the penalty term) increases.

Log Lambda vs Mean-Squared Error



(Fig. 3b)

(Fig. 3b) This plot illustrates how the mean squared error (the total difference between the model's predicted values and the true values) increases as the logarithmic value of lambda increases. This makes intuitive sense, since as more coefficients shrink to zero (Fig. 3a), the model will have to rely on less variables and less information to make predictions. The left line indicates the optimal lambda value (~ -6) and the right line indicates this value plus one standard error value.

Discussion

1. Findings
   a. The goal for this experiment was to determine factors that effectively signal Alzheimer's Disease in patients. Fig. 2 demonstrates that the logistic regression model generates the response variable based on just these 6 factors:
      i. "traila" - the number of seconds a patient took to complete part A of trail-making test
      ii. "trailb" - the number of seconds a patient took to complete part B of trail-making test
      iii. "csfvol" - the volume of a patient's cerebrospinal fluid
      iv. "lhippo" - the volume of a patient's segmented left hippocampus
      v. "games" - a patient's difficulty at play strategy games such as chess without assistance
      vi. "travel" - a patient's difficulty at travel without assistance
   b. The positive coefficients in the model for all variables besides "lhippo" signify a positive association between these variables and AD, and that as these variables increase in value, the patient is more likely to have AD. For example, a patient who takes longer on the trail making test is more likely to have AD.
   c. The negative coefficient in the model for "lhippo" suggests a negative association between a patient's left hippocampus volume and AD, and that as a patient's left hippocampus volume decreases, they are more likely to have AD. This makes intuitive sense, since a loss of brain volume is a known effect of AD [4].
2. Limitations
   a. One limitation for our research is that the NACC dataset that was used is comprised of participants from only 26 US States. Thus, our findings may not be applicable to the United States as a whole, nor to the entire world.
3. Continuation
   a. With access to more data, our group would like to investigate the association between video game proficiency and AD. Our findings and existing research suggest that a person's ability to play strategy board games such as chess effectively signals AD, but whether or not a person's ability to play video games signals AD remains unexplored.

References

[1] Alzheimer's Association, "10 Warning Signs of Alzheimer's," p. 12, 2019.

[2] M. Lillo-Crespo, M. Forner-Ruiz, J. Riquelme-Galindo, D. Ruiz-Fernández, and S. García-Sanjuan, "Chess Practice as a Protective Factor in Dementia," *Int J Environ Res Public Health*, vol. 16, no. 12, p. 2116, Jun. 2019, doi: 10.3390/ijerph16122116.

[3] J. T. Fuller, T. K. Choudhury, D. A. Lowe, S. Balsis, and Alzheimer's Disease Neuroimaging Initiative, "Hallucinations and Delusions Signal Alzheimer's Associated Cognitive Dysfunction More Strongly Compared to Other Neuropsychiatric Symptoms," *J Gerontol B Psychol Sci Soc Sci*, vol. 75, no. 9, pp. 1894–1904, Oct. 2020, doi: 10.1093/geronb/gbz032.

[4] E. C. B. Johnson *et al.*, "Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation," *Nat Med*, vol. 26, no. 5, pp. 769–780, May 2020, doi: 10.1038/s41591-020-0815-6.

Acknowledgements

We would like to thank:
- Dr. Babak Shahbaba
- Dr. Sam Behseta
- Dimitri Kaviani
- Zahra Moslemi
- Brian Schetzle

For mentoring us through this project and teaching us Data Science and R Programming this summer at UCI. We would also like to thank the National Alzheimer's Coordinating Center (NACC) for allowing us to use their data to conduct this research.