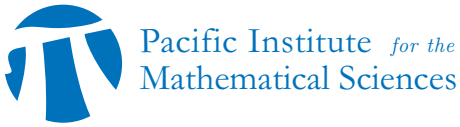

PACIFIC INSTITUTE FOR THE MATHEMATICAL SCIENCES VIRTUAL EXPERIMENTAL MATHEMATICS LAB (PIMS VXML) FINAL REPORT: GENOME-WIDE ASSOCIATION STUDY ON GENE PATHWAY IDENTIFICATION AND COGNITIVE FUNCTION PREDICTION

LI XING, KYLE GARDINER
MATHEW ZBITNIFF, ROHAM ASGARI, HANYE ZHONG

1. Introduction

Genome-wide association studies (GWASs) are a research approach used to identify genetic variations associated with complex diseases. This process works by scanning many individual single nucleotide polymorphisms (SNPs), that make up a genome, to identify which SNPs are statistically significant to the trait/disease of interest[1]. However, traditional single SNP-wise testing methods for identifying associations tend to lack power when sample sizes are small. Large sample sizes are needed when testing millions of SNPs to ensure reliable statistical power[1][2][3].

Knowing this information, we can utilize machine learning (ML) methods to help bridge the gap caused by small sample sizes. ML has the ability to identify potentially relevant SNPs that may have been missed[4].

This report will outline the ML methods used on a Department of Defense (DoD) Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which is a collaborative study with a small sample size, to identify relevant SNPs to Clinical Dementia Rating (CDR) scores of participants. We will also explore the biological importance of gene pathways associated with identified SNPs.

2. Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner,

Date: 2022-23.

MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see `www.adni-info.org`.

The datasets compiled by ADNI contain the genotypic, phenotypic, and demographic information of participants in addition to cognitive measurements like CDR, Mini-Mental State Examination (MMSE), and Alzheimer's Disease Assessment Scale (ADAS).

2.1. **Data Preprocessing.** First, we had to perform quality control (QC) on the data set. We used a popular toolset called PLINK[5]. This software cleans the data by checking for low minor allele frequencies, deviations from the Hardy-Weinberg equilibrium, population stratification, and any other confounding variables such as unequal sex proportions. Next, we used imputation to deal with missing values. We used a simple imputation method. If the variable was categorical, the missing value in the column would be replaced with the mode, whereas if the variable was continuous, the missing value would be replaced with the mean.

2.2. **Outcome.** On top of all that, there was a severe positive skew for our outcome variable, CDR. When a severe skew exists, linear regression can lead to biased estimates and inaccurate results. However, we dichotomized the CDR scores to use in logistic regression. The scores were adjusted using a predetermined threshold of 0.5[6], where scores below 0.5 have no cognitive impairment, while scores equal to or greater than 0.5 have some degree of cognitive impairment (mild to severe).

Table 1 outlines the statistics of the newly defined CDR categories.

	Normal Cognition (N=122)	Some Cogntive Impairment (N=75)	Total (N=197)
Age (years)			
Mean (SD)	69.0 (4.15)	69.5 (5.06)	69.2 (4.51)
Median [Min, Max]	67.9 [60.9, 82.2]	68.3 [61.6, 85.2]	67.9 [60.9, 85.2]
Ethnicity			
Not Hispanic/Latino	114 (93.4%)	63 (84.0%)	177 (89.8%)
Other	8 (6.6%)	12 (16.0%)	20 (10.2%)
APOE.e4			
Zero Alelles	88 (72.1%)	56 (74.7%)	144 (73.1%)
At Least One Alelle	34 (27.9%)	19 (25.3%)	53 (26.9%)
MMSE			
Mean (SD)	28.4 (1.63)	27.8 (1.83)	28.2 (1.73)
Median [Min, Max]	29.0 [21.0, 30.0]	28.0 [22.0, 30.0]	28.0 [21.0, 30.0]
ADAS			
Mean (SD)	10.3 (4.45)	13.7 (5.03)	11.6 (4.95)
Median [Min, Max]	10.0 [3.00, 23.0]	13.0 [4.00, 28.0]	11.0 [3.00, 28.0]

TABLE 1. Summary statistics of the dichotomized CDR categories

3. METHOD

3.1. **Traditional Single SNP Association Testing.** Once we had obtained the clean dataset, we wanted to utilize traditional GWAS regression methods to find associations. Normally linear regression would be used, but since we dichotomized the CDR scores, logistic regression was used.

The following logistic regression model:

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{SNP}_i$$

where π is the probability of having some degree of cognitive impairment, adjusted by age, MMSE, ADAS, Apolipoprotein E4 (ApoE4), ethnicity, and top 5 principal components (pcs), was used to screen the significance of the SNPs. Since none of the SNPs were concluded to be statistically significant, we moved on to using ML to select the important biomarkers from the top 1000 SNPs from the logistic regression.

3.2. **Machine Learning: Elastic Net.** We took the top 1000 SNPs from logistic regression and input them into the elastic net penalized regression which operates by utilizing the elastic net penalty to minimize the following objective function [7]:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{1}$$

The elastic net penalized regression is particularly helpful due to its ability to do shrinkage and variable selection simultaneously. However, this variable selection is quite dependent on the values of λ_1 and λ_2 which can be k-fold cross-validated (i.e 5 fold) using the *glmnet* [8] package in R.

3.3. **Gene Enrichment Analysis.** Mapping the SNPs to the genes of an associated pathway is essential to understand the underlying mechanisms of the disease/trait of interest. Here, we used SNPnexus (`https://www.snp-nexus.org/v4/`) to first map the significant SNPs to their genes. We then used the web-based tool for gene enrichment analysis, EasyGSEA (`https://tau.cmmt.ubc.ca/eVITTA/easyGSEA_demo/`). The EasyGSEA database contains KEGG, WikiPathways, DrugBank v5-1-8, and DisGeNET, which are disease-related gene sets to map the selected genes from SNPnexus to their biological pathways.

4. Results

4.1. **Single SNP Association Testing.** After applying the traditional logistic regression GWAS method on our dataset, none of the SNPs tested appear to be significant since they don't extend beyond the genome-wide significance threshold in figure 1. We will now explore the results when using ML techniques.

4.2. **Machine Learning: Elastic Net.** The logistic regression model allows us to screen the SNPs from most to least significant. The top 1000 SNPs from logistic regression were used in the elastic net penalized regression. By taking the top 1000, we can avoid the computational limitations of applying elastic net to millions of SNPs.

Using the elastic net criterion, 136 out of the top 1000 SNPs were selected to be significant to CDR/cognitive decline. The elastic net selected SNPs are shown in figure 1 as red dots.

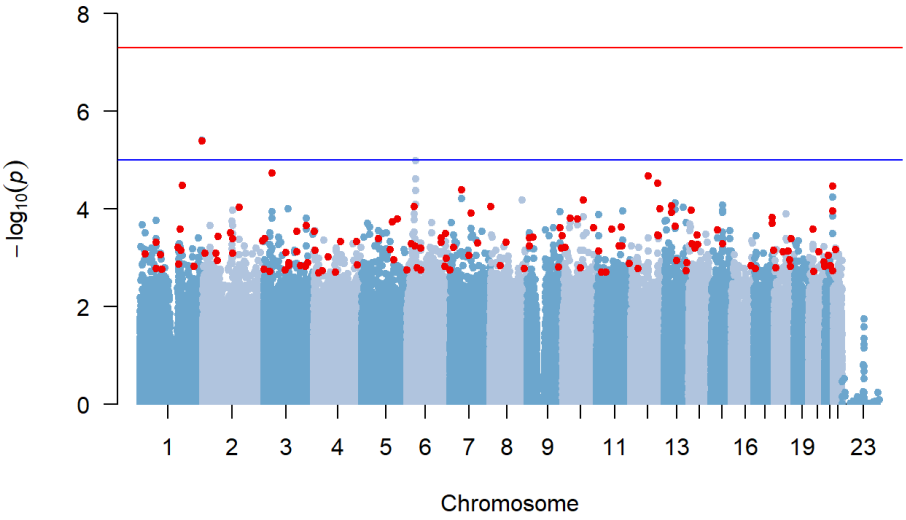

FIGURE 1. Manhattan plot from the single SNP association testing. The blue horizontal line represents a suggestive line (p-value=$1 * 10^{-5}$). The horizontal red line represents the genome-wide significance threshold (p-value=$5 * 10^{-8}$). Red dots represent SNPs selected by elastic net.

4.3. **Gene Enrichment Analysis.** Using the 136 selected SNPs from elastic net in gene enrichment analysis, some prominent gene pathways were identified. The most notable pathway was the positive regulation of interleukin-2 (IL-2). This pathway refers to the activation of IL-2. A couple of studies have identified the partial role of IL-2 in cognitive decline. For example, Liang et al. identified strong correlations between the levels of IL-2 in amnestic mild cognitive impaired participants and their cognitive scores. They concluded lower levels of IL-2 are associated with declined cognitive scores. They even found that IL-2 levels may be better at identifying cognitive impairment compared to the common Aβ and tau biomarkers.[9]

5. CONCLUSION

This report demonstrates the ability to use ML methods to identify significant SNP to disease/trait associations that may have been missed in traditional single SNP association testing due to small sample sizes. We have shown this using the DoD ADNI dataset with a small sample

size. It shows that ML can become an effective tool that fits into the repertoire of future GWASs. In addition, we have illustrated the use of gene enrichment analysis to identify prominent gene pathways related to cognitive decline. Hopefully, this research will spark the interest to use other ML methods in GWASs when traditional methods are inconclusive.

6. Acknowledgements

We would like to acknowledge Li Xing for providing guidance and resources throughout the duration of this project. We would also like to extend appreciation and thanks to the members of PIMS VXML for organizing this opportunity and ensuring everything went smoothly and stayed on track.

References

[1] Emil Uffelmann, Qin Q. Huang, Nchangwi S. Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 2021.

[2] Eun P. Hong and Ji W. Park. Sample size and statistical power calculation in genetic association studies. *Genomics amp; Informatics*, 10(2):117, 2012.

[3] Yogasudha Veturi and Marylyn D. Ritchie. How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? *Biocomputing 2018*, 2017.

[4] Kevin P. Murphy. *Machine learning: A probabilistic perspective*. MIT Press, 2021.

[5] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[6] Sid O'Bryant. Staging dementia using clinical dementia rating scale sum of boxes scores. *Archives of Neurology*, 65(8):1091, 2008.

[7] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[8] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[9] Chih-Sung Liang, Chia-Lin Tsai, Guan-Yu Lin, Jiunn-Tay Lee, Yu-Kai Lin, Che-Sheng Chu, Yueh-Feng Sung, Chia-Kuang Tsai, Ta-Chuan Yeh, Hsuan-Te Chu, and et al. Better identification of cognitive decline with interleukin-2 than with amyloid and tau protein biomarkers in amnestic mild cognitive impairment. *Frontiers in Aging Neuroscience*, 13, 2021.