Predicting Patient Z's Likelihood of Alzheimer's Disease Using Polygenic Score Calculations from GWAS Studies

CBB 472: Biomedical Data Science: Mining and Modeling

Nancy Lu, April 2018

1 Introduction: The Pathology of Alzheimer's Disease

Alzheimer's disease (AD) is a chronic, unrelenting, neurodegenerative disease with a long clinical duration and a high prevalence amongst the elderly population. Epidemiological studies estimate that AD has a prevalence of approximately 10-30% in those over the age of 65 years (Colin 2015). AD specifically targets the cerebral cortex and hippocampus, affecting an individual's memory, thinking, and ability to do some physical tasks. Abnormalities can progress at different rates in people, which makes the disease's progression hard to anticipate. In particular, AD is associated with the buildup of amyloid-β plaques which accumulate in extracellular spaces. Amyloid precursor protein (APP) is cleaved proteolytically to γ-secretases and β-secretases into amyloid-β (Colin 2015). The deposition of amyloid-β ultimately leads to neurofibrillary tangles and neuronal degeneration, which can be seen clearly in molecular imaging. While cleavage of APP occurs through one pathway, another pathway to clear the plaques also exists. Thus, both proteins in the cleavage and the clearing pathway must be considered when designing a model of AD. The complete pathway is illustrated in Figure 1.

AD can occur in two forms - an autosomal dominant inherited form (DIAD), which causes early onset at around age 45 years old, and late-onset which occurs at around the age of 65, as mentioned before. Patient Z is currently 51 years old with no apparent mental decline. Since he continues to be a prolific science writer and professor at Yale University, we can safely assume that does not have early-onset Alzheimer's. Thus, we will only predict his risk of late-onset Alzheimer's in which the pathology less well understood. What is known, however, is that Alzheimer's disease shares many characteristics with Parkinson's disease, which means this model could potentially be extended to other neurodegenerative diseases. Secondly, AD has many comorbidities, such as diabetes mellitus, hypertension, and obesity, which this model may have some predictive power for.

AD has a strong genetic component, that has been well documented in several GWAS studies. According to twin studies performed from 11,489 sample subjects Swedish Twin Registry, Alzheimer's disease has a very high heritability of 79% in the best fitting model and 58% at the lowest (Dudbridge 2013). Thus, more than half of an individual's risk of AD is determined by genetics alone, which makes this particular study interesting and possible. The study also found no significant difference between the genetic factors affecting men and women, so filtering by gender for this study was not necessary.

More specifically, the amyloid-β protein, apolipoprotein E (APOE) and tau are two elements that have been implicated (Lambert 2013). APOE is particularly interesting, as it is involved a mechanism that clears the plaque, along with several other enzymes, such as insulin-degrading enzyme, matrix metalloproteinase, among others. APOE protein is a 299 amino acid protein, whose structure can easily be changed by amino acid substitutions that either affect the total charge or the 3-D conformation of the protein. Numerous studies have shown the profound significance of APOE in acting as a chaperone for the plaques, which directly affects their clearance and deposition. The protein exists in 3 different proteins, with one allele (APOE2) imparting a 12-fold increase in risk and likely early onset, and another allele (APOE3) imparts a 3-fold increase in risk (Colin 2015). APOE4 is the most common isoform, and contributes to approximately 50% of the cases of sporadic AD. From current research, the APOE gene on chromosome 19, has long been know as the strongest genetic risk factor for developing AD, and and will be weighted the most heavily through out the analysis.

Thus, we chose AD as the subject of this study because (1) it has a high prevalence and is a chronic disease, (2) we do not yet know if Patient Z will have it, and (3) there is a strong genetic component.

2 Methods

We focus on Patient Z's data, which was mapped to GRCh37 by Illumina HiSeq data. This patient happened to obtain his complete genome through a loophole and is the first journalist to have his entire genome report, rather than just a partial report containing Illumina's selected SNVs. Patient Z revealed that he is fully Ashkenazi Jewish, meaning that GWAS studies performed of non-Hispanic white men will be the most relevant. The files already contained mostly annotated SNPS, CNVs, and other structural variants.

As a comparison, Brendan Hellweg, a half Ashkenazi male, also kindly allowed us to analyze his 23-and-me read which was obtained from a cheek swab. Brendan's genome was also mapped to GRCh37. 23-and-me is a company that primarily is used to determine ancestry, but may also be used to determine carrier status for a limited number of diseases. 23-and-me sent a full list of Brendan's *single nucleotide polymorphisms* (also known as SNPs), with their rs numbers annotated. Rs numbers are a common way to annotate SNPs instead of directly describing long chromosome and position numbers. These two datasets were analyzed against the GRCh37 reference genome.

To find the most relevant genes in the analysis, *genome-wide association studies* (GWAS, for short) were used. In 2013, Lambert et al. conducted a meta-analysis of 74,046 individuals to determine susceptibility loci for AD. This study is currently the largest of its kind, and is presumably quite statistically robust. The study identified 11 new susceptibility loci, in addition to the 11 currently known, which is a total of 22 genetic loci that seem to have a strong correlation the disease. It is important to note that not all of the genes are detrimental; some likely serve protective functions in the amyloid-β plaque clearing system. This study was performed on individuals of European ancestry, which Patient Z falls into and had 3 stages: (1) first, over 7 million SNPs were genotyped and imputed for the meta-analysis on 17,008 AD cases and 37,154 control studies, (2), approximately 11,000 of those SNPS were genotyped and tested for association, and (3), only SNPs that reached genome-wide significance $p < 5 * 10^{-8}$ were used.

With the information from the GWAS, Marden et al. created a polygenic score calculation to calculate a risk prediction of any individual to get AD. Each of the 22 important SNPS is associated with a meta-analyzed odds ratio (OR) in the GWAs. 4 proxies were used as a more accurate predictor. To calculate the Alzheimer's Disease Genetic Risk Score (AD-GRS), the individual's risk count (how many of the associated SNPs the individual has) is multiplied by the beta coefficient for that polymorphism, which was determined from the Lambert study. One minor change is that beta-coefficient for APOE, which is most highly correlated with Alzheimer's was obtained from AlzGene, since this value has been corroborated by many other studies. Summing these products weights each polymorphism with its potential risk. A positive beta-coefficient means a higher risk of AD, and negative beta-coefficient indicates a protective variant. The weighted allele sum is then exponentiated, and multiplied by by the estimated dementia prevalence in the sample, which is 0.1, as shown in equation 1 below. Finally, the odds are converted into probabilities.

$$odds = 0.1 * e^{\sum_{i=1}^{n} allele\ count * \beta} \tag{1}$$

Strictly speaking, the AD-GRS calculates the probability of dementia predicted by the 22 alleles, since dementia is a well-documented effect closely tied with AD and is more easily documented in a GWAS study. In addition, since APOE was for a long time the main genetic locus associated with AD, an alternative AD-GRS excluding APOE was also performed to see if the other 21 loci had much bearing on the final score. In non-Hispanic white males (which both subjects would fall into), the 21 loci could substantially change the risk prediction, so the original AD-GRS including the APOE gene and the 21 other loci was used. It is important to note that while the AD-GRS does predict memory loss well, it does not predict death due to the disease, indicates that other gene-environment interactions should be considered for this model to be more comprehensive.

The workflow used in this study involved the following: (1) first, we annotate the patient Z's genome with rs numbers, (2) second, we compare the annotated SNPS and search for the ones we believe to be significant in the progression of AD with the Genome Analysis ToolKit (GATK), and (3) finally we calculate the polygenic score using the equations and process specified above. Figure 2 depicts the workflow we used for this analysis.

3 Key Results and Significance

Ultimately, by using the data and the methods section described before, we found that Patient Z has a 13.72% chance of late-onset Alzheimer's which is a slightly elevated percentage compared to the baseline 10% of the population that the study used. The full list of this patient's AD-associated SNPs can be seen in Figure 3. In particular, he has a mutation in APOE, which has a large bearing on AD risk and is thus weighted heavily with a large β-coefficient from the GWAS studies. In total, he has 10 out of the 22 SNPS associated with Alzheimer's, with 5 increasing his risk and 5 decreasing his risk. The most significant was the mutation in the APOE gene, which means that this patient carries an isoform of the protein that increases his risk by threefold. Some SNPs, such as that of NME8, seem to be slightly protective, though their exact mechanism in the AD pathway are not known.

Brendan's data was also analyzed with the same method. Brendan had 21 out of the 22 associated SNPS, including the same mutation in the APOE gene that Patient Z had. While Brendan had almost all of the SNPs, this was not a cause for concern, since he also had more of the protective variants. Due to these protective variants, Brendan's risk score was actually much lower at 9.71%, which is lower than the baseline value of the population. Thus, he has less likely than the general population to have AD. In particular, he had variants of SORL1 and DSG2, which have both been hypothesized as important intermediates in the amyloid-β plaque clearing system. Both Patient Z and Brendan had mutations in PICALM and CELF1. Further analysis could be performed to see if any of the SNPs are more common in Ashkenazi Jewish men and if so, whether certain populations are more prone to certain AD pathologies.

These two calculations have far-reaching consequences in terms of providing a method of actually calculating risk predictions given a person's genome. In addition, specific genes and proteins related to the pathway were identified in both individuals. If AD drugs are developed further, one possibility is personalized medicine - specifically targeting protein mutations that the individual has. For example, it is currently known that the three isoforms of the APOE lipoprotein all have slightly different conformations, and each of these conformations could be targeted with a different drug. Finally, this calculation is significant because it has the potential to be applied to a number of other diseases that have also been studied with GWAS studies, including obesity, cardiovascular disease, and other neurodegenerative diseases.

4 Further Directions

As this study shows, it is not difficult to extract a polygenic score to calculate the risk of an individual having late-onset Alzheimer's disease in the future. This exact procedure could conceivably be extended to any individual genome. Logically, this could also be extended to various other diseases, once enough GWAS studies have been conducted to calculate log-odds scores of all the most important SNPs. For example, cardiovascular disease (CVD) is one example of a disease where calculating the polygenic score is viable. Zhang et al. analyzed GWAS studies for atherosclerotic cardiovascular disease and calculated beta-coefficients for 38 genes (one of which, coincidentally is the APOE gene). These researchers used a lower threshold of significance so a standardized value for significance would need to be determined. While the researchers claim that the top 10 SNPs can explain up to 17% of the CVD risk, it is important to note that CVD has a very strong environmental component which would also need to be incorporated in some way. Environmental risks such as poor diet, lack of exercise, and smoking have a large bearing on CVD. A composite lifestyle index, perhaps the one developed NIH in 2006 (Kaleta), could be incorporated via a canonical correspondence analysis. For this, we would need more information from Patient Z about his exercise, eating, and sleeping habits.

Another avenue of exploration could be examining the gene expression levels for some of the important proteins involved in AD. The GTEx project contains gene expression and quantitative trait loci from 53 different human tissues. By characterizing patient Z's genetic variation in conjunction with this database, gene expression levels of the relevant proteins could be studied, such as apolipoprotein precursor protein. We could also perform powerful statistical tests with the eQTL analysis tools available. It is also possible that other genes and proteins with some significance could be found.

There exists one fairly major problem common to all GWAS studies. GWAS studies are only able to analyze associations, not mechanisms. The mechanisms behind how these genes and their protein products are involved the process must be studied in vitro to deeply understand their importance. On one hand, APOE has been well validated - the protein has been studied extensively. The lipoprotein is known to package cholesterol and most importantly the amyloid-beta peptide and transport them throughout the body to prevent accumulation. The amyloid-beta pepetide is then no longer able to initiate the toxic events that eventually lead to neurofibrillary tangles in the hippocampus and synaptic dysfunction. On the other hand, SORL1 seems to have a strong protective ability against ADD, yet its known functions are mostly transmembrane signaling receptor activity (Lambert 2013). The negative regulation of the amyloid-beta plaques is not well understand, even though a strong correlation has been found. Thus, GWAS studies can only provide correlations and directions towards future potential avenues of research; the mechanisms of each of the proteins should be fully understood using wet-lab experimentation before any sensational conclusions are drawn from the data. The importance of scientific literacy regarding these types of risk predictions cannot be overstated.

All in all, since there is no cure for AD and few recommendations beyond living an active lifestyle and eating a healthy diet, risk predictions such as these may cause more harm than good, causing undue stress to the subject. To conclude, we recommend that Patient Z continue his brilliant writing career and that he take our statistics with a grain of salt.

5 Figures

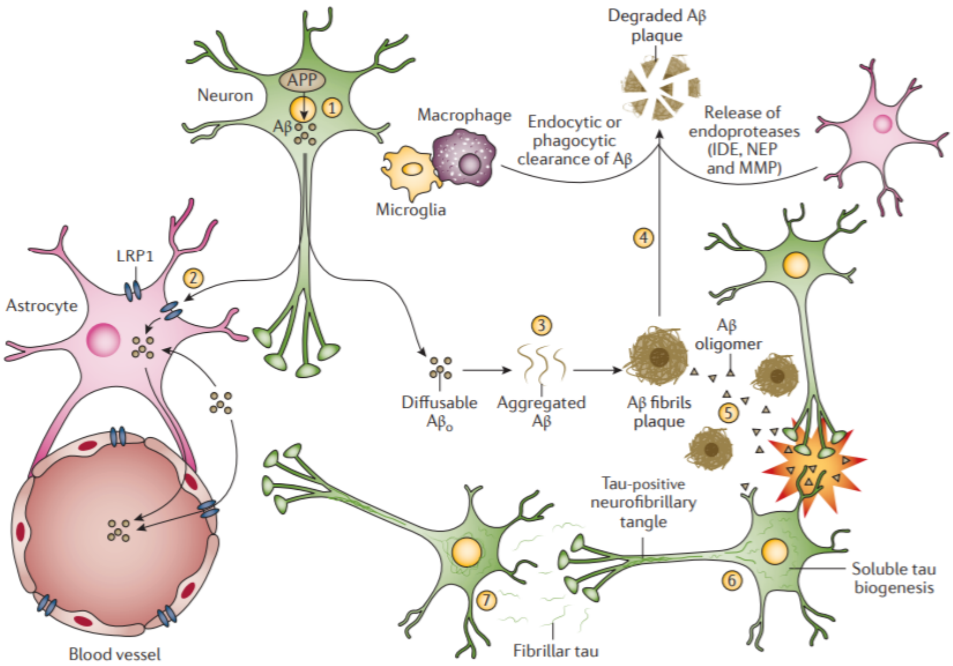


Figure 1: This figure shows the complex molecular pathways occurring that ultimate create neurofibrillary tangles in the hippocampus that lead to the hallmark feature of Alzheimer's disease - memory loss. In the figure, we can see that the amyloid-beta plaques can either be degraded by either endoproteases or endocytic/phagocytic processes. If they are not degraded, then a cascade occurs, which ultimately leads to hippocampal tau protein aggregates. Figure from a Nature Review on AD by Colin et al.

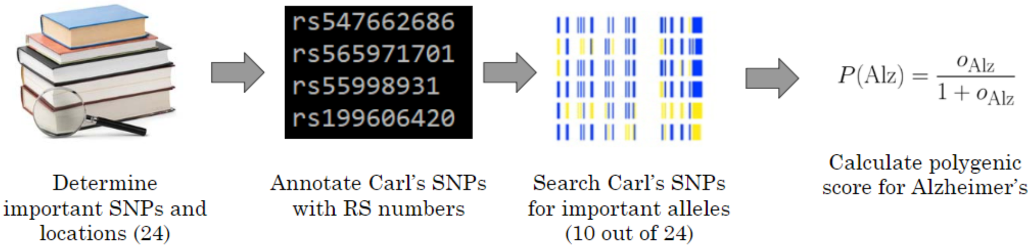
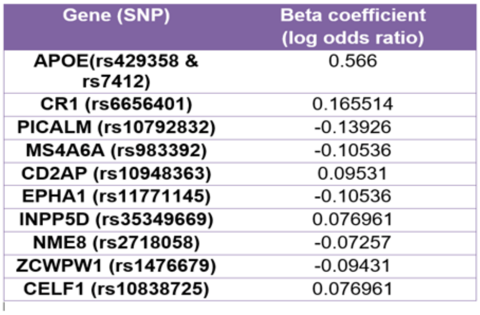
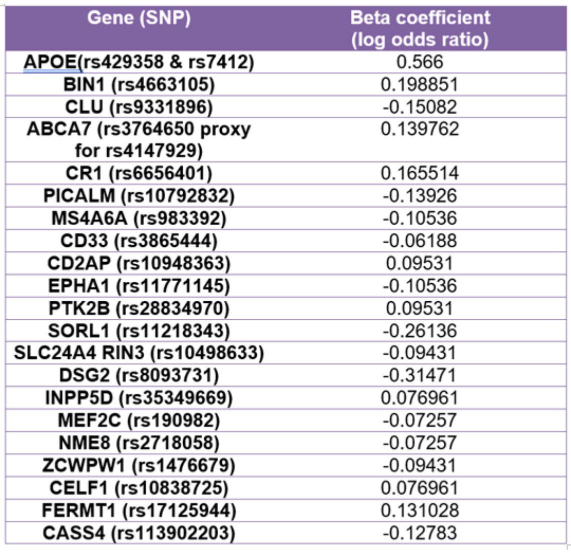

rs547662686
rs565971701
rs55998931
rs199606420
$P(\text{Alz}) = \frac{o_{\text{Alz}}}{1 + o_{\text{Alz}}}$
Determine important SNPs and locations (24)
Annotate Carl's SNPs with RS numbers
Search Carl's SNPs for important alleles (10 out of 24)
Calculate polygenic score for Alzheimer's

Figure 2: The above image shows the workflow used in this study to annotate Patient's Z's genome, compare these SNPs against the known genes that correlated with AD, and then calculate the probability of having the disease.



Gene (SNP)	Beta coefficient (log odds ratio)
APOE(rs429358 & rs7412)	0.566
CR1 (rs6656401)	0.165514
PICALM (rs10792832)	-0.13926
MS4A6A (rs983392)	-0.10536
CD2AP (rs10948363)	0.09531
EPHA1 (rs11771145)	-0.10536
INPP5D (rs35349669)	0.076961
NME8 (rs2718058)	-0.07257
ZCWPW1 (rs1476679)	-0.09431
CELF1 (rs10838725)	0.076961



Gene (SNP)	Beta coefficient (log odds ratio)
APOE(rs429358 & rs7412)	0.566
BIN1 (rs4663105)	0.198851
CLU (rs9331896)	-0.15082
ABCA7 (rs3764650 proxy for rs4147929)	0.139762
CR1 (rs6656401)	0.165514
PICALM (rs10792832)	-0.13926
MS4A6A (rs983392)	-0.10536
CD33 (rs3865444)	-0.06188
CD2AP (rs10948363)	0.09531
EPHA1 (rs11771145)	-0.10536
PTK2B (rs28834970)	0.09531
SORL1 (rs11218343)	-0.26136
SLC24A4 RIN3 (rs10498633)	-0.09431
DSG2 (rs8093731)	-0.31471
INPP5D (rs35349669)	0.076961
MEF2C (rs190982)	-0.07257
NME8 (rs2718058)	-0.07257
ZCWPW1 (rs1476679)	-0.09431
CELF1 (rs10838725)	0.076961
FERMT1 (rs17125944)	0.131028
CASS4 (rs113902203)	-0.12783

Figure 3: The table on the left corresponds to Patient Z's genome, while the table on the right is Brendan's genome. Z has 10 out of the 22 SNPS associated with AD, which lands him at approximately a 13% chance of acquiring AD. Brendan has 21 out of the 22 associated SNPS, which gives his risk score at around 9.7%. Brendan has more protective variants which makes his risk slightly smaller.

6 References

Colin, M. (2015). Alzheimer's disease. Nature Reviews Disease Primers, 1, 1-18. Retrieved from https://www.nature.com/articles/nrdp201556.pdf.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS genetics, 9(3), e1003348.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. Archives of general psychiatry, 63(2), 168-174.

Thomas L. Lenz, Nicole D. Gillespie, Jessica J. Skradski, Laura K. Viereck, Kathleen A. Packard, and Michael S. Monaghan. Development of a Composite Lifestyle Index and Its Relationship to Quality of Life Improvement: The CLI Pilot Study. ISRN Preventive Medicine (2013) https://doi.org/10.5402/2013/481030.

Lambert et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013 Dec;45(12):1452-8. doi: 10.1038/ng.2802.

Maloney, M. T. (2015) One Hundred Years of Alzheimer's Disease: The Amyloid Cascade Hypothesis. Nature Education 8(4):6.

Marden JR, Mayeda ER, Walter S, et al. Using an Alzheimer's Disease polygenic risk score to predict memory decline in black and white Americans over 14 years of follow-up Running head: AD polygenic risk score predicting memory decline. Alzheimer disease and associated disorders. 2016;30(3):195-202. doi:10.1097/WAD.0000000000000137.

Shen, L., Jia, J. (2016). An overview of genome-wide association studies in Alzheimer's disease. Neuroscience bulletin, 32(2), 183-190.

7 Acknowledgements

Thank you to Jacob Levitt, Feimei Liu, David Brandfonbrener, and Dawn Chen for making this group project so interesting and fun. Also special thanks to Prof. Carl Zimmer, who was Patient Z in this study, and Brendan Hellweg for contributing their genomes for this study. And a final thank you to William Meyerson, Xiaotong Li, and Prof. Mark Gerstein for this wonderful class. While sometimes hard, it was very much worth it.