Use of APOE e4 alleles in predicting Alzheimer's disease
PH240C Final Project

Pablo Freyria Duenas

December 17, 2021

Abstract

Alzheimer's disease is recognized as one of the most feared and prevalent diseases in older adults. Despite increasing research efforts, there is an unmet need for treatments that prevent, stop or reverse this disease. Current research has been focused on the use of biomarkers to characterize disease pathways to target with new treatments. Particularly, the APOE e4 allele has been linked to impact the onset and progression of disease. This report studies the additional information of the e4 allele to the variables usually measured in clinic visits, showing that even when accounting for relevant variables, the e4 allele can be used to predict disease status.

1 Introduction

Alzheimer's disease (AD) was first characterized in 1906, when Dr. Alois Alzheimer noticed abnormal buildup and fibers in brain tissue of a woman who died of a mental illness; now called amyloid plaques and neurofibrillary tangles respectively NIH (2021). This disease is estimated to cause between 60% to 80% of dementia cases, a condition that will be experienced by 1 in three older adults Butch (2017); and recent surveys in the U.S. and Europe place it just behind cancer as the most feared disease.

While there have been significant advances in the understanding of the disease, current treatment is still limited to trying to slow down the symptoms Butch (2017), highlighting the need for advancing research to develop treatments that can prevent, stop and reverse the course of the disease. To meet this research obectives, the U.S governement has increased funding in 2021 to $3.2 Bn in AD and dementia research to discover new ways of reduce risk, and uncover biomarkers for early diagnosis and targeting treatments AIM (2021).

One of these biomarkers of interest is the apolipoprotein E (APOE) e4 allele, where *in vitro* experiments have shown its impact in increasing $\beta-$amyloid plaque and filament formation Marra Camilo (2004), the abnormal buildup discovered initially by Dr. Alois Alzheimer.

There is a general consensus that the APOE e4 allele impacts the onset of disease and the rate of brain atrophy; but there are also controversial. As an example, it has been observed how e4 allele carriers have shown more severe memory impairment in non-institutionalized elderly people, and

it has also been observed that it doesn't seem to affect the rate of cognitive decline at later stages of the disease. Moreover, better cognitive functions in very old cognitively intact people have been observed among e4 allele carriers Marra Camilo (2004).

The objective of this study was to study the differences in AD risk attributed to the APOE e4 allele, when accounting for other variables such as demographics, comorbidities, lifestyle, cognitive symptoms, and clinical variables of diagnosis such as predominant symptom, presence of dementia, age begin of symptoms.

It is worth mentioning that other studies usually restrict their study population to subjects with onset less than 3 years ago, absence of other major medical illness, no history of alcohol and drug abuse, no history of depression and no abrupt onset or worsening of disease Marra Camilo (2004). This study aims to assess the impact of the e4 allele when these potentially confounder variables are included

2 Dataset description

The study was made possible by the Uniform Data Set (UDS v.3) curated by the National Alzheimer's Coordinating Center (NACC). The NACC was established in 1999 by the National Institute of Aging (NIA) to facilitate collaborative research by collecting data from NIA funded AD centers across the U.S.

The UDS has been gathered since 2005 using a prospective, standardized, and longitudinal clinical evaluation of the subjects in the NIA program. In each subject's annual visit to the research centers, 16 data collection forms, covering from subjects demographics to neurological examination findings are completed by the clinician. Subjects in the UDS are not a statistical based sample of the U.S. population, but rather a referral or volunteer case series; and each of the research centers enrolled subjects according to its own protocol, most volunteers have normal cognition and tend to be highly educated at enrollment NACC (2021).

The data collected 162,200 observations across 44,300 subjects ranging in age from 18 to 110 years old and 19,000 (42.8%) identified as male; with 1010 variables per observation. Data set had different degrees for identifying AD as the cause of cognitive disorder; ranging from AD being the presumptive etiologic diagnosis of cognitive impairment to AD being probable or possibly a primary cause (contributing or non-contributing) for cognitive impairment. The otcome variable "NACCALZD" chosen in this study is derived by NACC, and defined as "Presumptive etiologic diagnosis of the cognitive disorder — Alzheimer's disease".

Using this outcome variable, 59,000 (36%) of measurements had AD, corresponding to 21,700 subjects (48.9%). Furthermore, 10,500 subjects had no information about the APOE allele, which were removed. Among these subjects, 4,700 (45%) had diagnosis of AD as etiologic cause of cognitive disorder.

2.1 Data processing and selection

The first subsetting of the data was to remove the observations for which no information of the APOE allele was available. UDS had 40 variables related to any medication use, 20 of them consisted on NA's only; so all these variables were dropped and variable the "ANYMEDS" was used as an indicator of being on any medication.

Variables were selected by hand on each category according to the relevance to AD, particularly the ones related to comorbidities Jen-Hung Wang (2018). Initially, the 51 variables seen on Table 1 were chosen as features.

Table 1: Initially selected features

Category	**NACC code**	**Brief description**
Demographics (8)	NACCID, BIRTHYR, SEX, NACCAGE, RACE, EDUC, NACCAPOE, NACCNE4S	Subject ID, Year of birth, Sex, Age at visit race, Years of education, APOE alleles and number of e4 alleles
Comorbidities (16)	CVHATT, HATTYEAR, CVCHF, TBI, CBSTROKE, NACCSTYR, PD, PDYR, DIABETES, B12DEF, ARTHRIT, THYDIS, DEP2YRS, DEPOTHR, ANXIETY, OCD	Heart attack (,Year of), Congestive heart failure, Traumatic Brain Injury, Stroke (, Year of), Parkinson's Disease(, Year of), B12 vitamin deficiency, Arthritis, Thyroid disease, Active depression in last 2 years, or more than 2 yrs ago, anxiety, Obsessive-Compulsive Disorder
Lifestyle (7)	SMOKYRS, QUITSMOK, ALCOCCAS, ALCOHOL, ABUSOTHR, NACCBMI, ANYMEDS	Years smoking cigarettes, Age of quitting smoking, Consumed alcohol in past 3 mo, Alcohol abuse over a 12 mo period, Other abused substances over 12 mo, Body Mass Index, Taking any medication
Symptoms (5)	MEMORY, ORIENT, PERSCARE, SATIS, BORED	Scale of: Memory impairment, Orientation, Personal care. Is subject satisfied with life, Is subject often bored
Cognition scores (6)	NACCMMSE, PENTAGON, LOGIPREV, BOSTON, DIGIF, MOCATOTS	Total MMSE score, intersecting pentagon subscale score, total score from previous administration, Boston naming test, digit span forward correct, MoCA total raw score
Related to diagnosis (9)	DEMENTED, COGFLAGO, NACCCOGF, COGMODE, FDGAD, TAUPETAD,CSFTAU, OTHBIOM	Met criteria for dementia, age of fluctuating cognition begin, predominant symptom first recognized, mode of cognitive symptoms onset, FDG-PET pattern of AD, tau PET evidence of AD, elevated CSF tau or ptau, other biomarker

As a next step, it was inspected how many missing values[1] was there for all of these variables. At first, it was noted that 99% of LOGIPREV was missing; as well as: 95% of TBI, ARTHRIT, ANXIETY, OCD and ALCOCCAS; 74% MOCATOTS; 65% THYDIS; and 64% of FDGAD, TAUPETAD, CSFTAU, OTHBIOM, COGFLAGO. These 13 variables were dropped. Then, 92% of observations were without heart attack, 97% without Parkinson's and 92% without stroke; so the 3 variables recording the year of the event were also removed from data.

Finally, it was noted that on the CVHATT, CVCHF, CBSTROKE, PD, DIABETES, B12DEF, DEP2YRS, DEPOTHR, SMOKYRS, ALCOHOL, ABUSOTHR comorbidity variables, 30.5% of observations were missing. Furthermore, the missingness across these variables came from the same observations and they were dropped. An interesting remark about these missing observations is that these subjects eventually got their measurements; and that the proportion of observations with Alzheimer was lower among those with missing values than in overall, suggesting that there is a correlation of having measured these variables and being diagnosed with AD.

The final data set consisted on 22,000 unique subjects across 53,300 observations with 1 outcome and 35 features; of which, 2 measured the APOE e4 allele and are used to stratify on populations, 13 are binary, 8 are categorical (6 of them ordinal) and 6 are numerical. The variable QUITSMOK, years since quitting smoking, was redefined as the interaction QUITSMOK*SMOKYRS to account that it is only relevant among subjects that used to smoke.

3 Methodology

3.1 Proposed models

The statistical problem can be described as a classification problem: given the previously described 34 features, can we predict whether subject has AD or not? The main challenge with this data is that most of the features are categorical, so classification methods that work on distances such as SVMs, that maximize distance to decision bound, or nearest neighbor classification would be inadequate in this problem. On the other side, classification trees and logistic regression are more adequate for categorical data. Additionally, tree based algorithms can handle missing data by using surrogate splits, that is when a the value for splitting variable is missing, the next best splitting variable is used.

In classification trees, data is sequentially split among feature values or categories so that each leaf under the node has a lower error rate than the node. This method can be modified for further smoothness with the xgboost algorithm, which has shown to have state-of-the-art performance on classification benchmarks with better scalability than other popular algorithms Chen and Guestrin (2016). Xgboost is a gradient tree boosting algorithm with a regularization term in the loss function, penalizing for the number of leafs in each tree and for the weights given to the leafs. The name *gradient* boosting comes from the loss function being optimized iteratively with its second order

[1]Encoded as NA or the UDS form submitted did not collect the data

approximation, involving the first and second order gradient of the loss function at the previous iteration. Besides the regularization term, overfitting is also prevented by shrinking newly added weights by a predifined factor, reducing the influence of each individual tree; and by subsampling columns, a technique also used in random forests.

Finally, logistic regression handles categorical data by imputing dummy variables for each value of the category (removing 1 category to deal with co-linearity issues) and the coefficient associated to that value is the log-odds ratio of belonging to that category vs. belonging to the removed category[2]. This process increases the number of regressor variables, so a LASSO model was used instead. This LASSO model changes the loss function that the logistic regression is trying to minimize, by adding the L_1 norm of the coefficients times a constant λ; so that, minimizing the loss functions also results in setting some effects to zero.

3.2 Model training and tuning

All algorithms are implemented using the statistical software R version 4.0.3. Classification trees were implemented with the rpart package, xgboost with the xgboost package; and the regularized LASSO logistic regression using the glmnet package, for the glmnet model missing values were dropped.

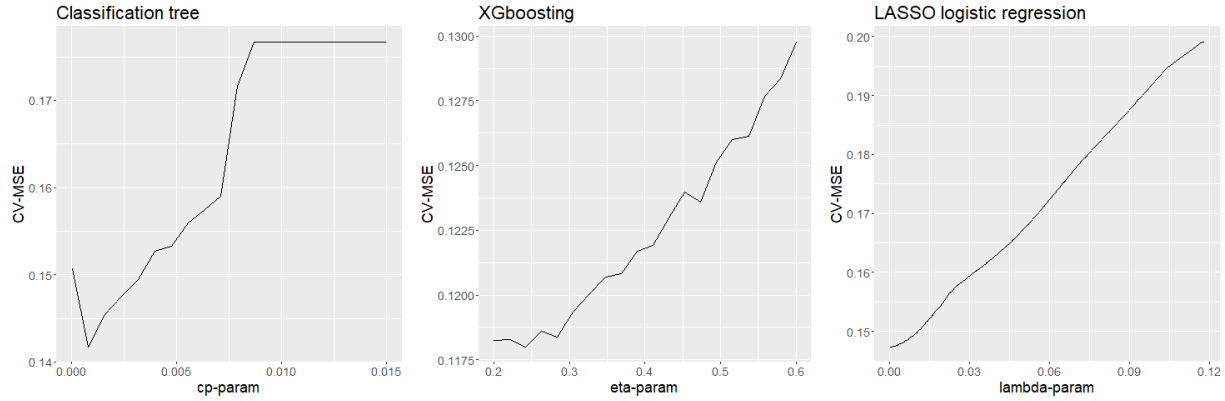
To tune the parameters, a 10-fold cross-validation scheme was implemented. Briefly: data is split into 10 sets, then model is trained on 9 of these sets and evaluated on the remaining set; this process is then repeated 10 times, each with a different testing set, and the model with the lowest average loss is chosen. The loss function is computed with the raw model predictions, that is square difference of the probability of predicting 1 to the actual status is computed. The decision boundary will be studied with the ROC curve.

For the classification trees, the parameter to be tuned is *cp* which governs the minimum decrease in overall lack of fit that any split must have; the default is 0.01, and 20 values between 0.00001 and 0.015 will be tested. For xgboost, the parameter to tune is $\eta \in (0, 1)$, the learning rate, which scales the contribution of each tree; lower learning rates imply training more trees and building a model more robust to overfitting. The default value for this parameter is 0.3, so 20 values between 0.2 and 0.6 will be tested. For LASSO logistic regression, the tuning parameter is λ which is tuned automatically by glmnet, the documentation mentions that training is faster over a sequence of decreasing λ rather than with a single parameter; however the results for each value of λ are given and the average loss can be computed for each value of the parameter. The mean square error loss by the parameter value can be seen in Figure 1.

The plot for the LASSO logistic regression model shows that the regularization term isn't improving the algorithm, with lower values of λ giving lower MSE. In addition to finding the optimal parameters for the proposed algorithms, the tuning process helped in finding the best algorithm: xgboost is the most accurate one; and among interpretable models, classification trees are prefered to LASSO logistic regression. Finally, the ROC curve for the xgboosting algorithm

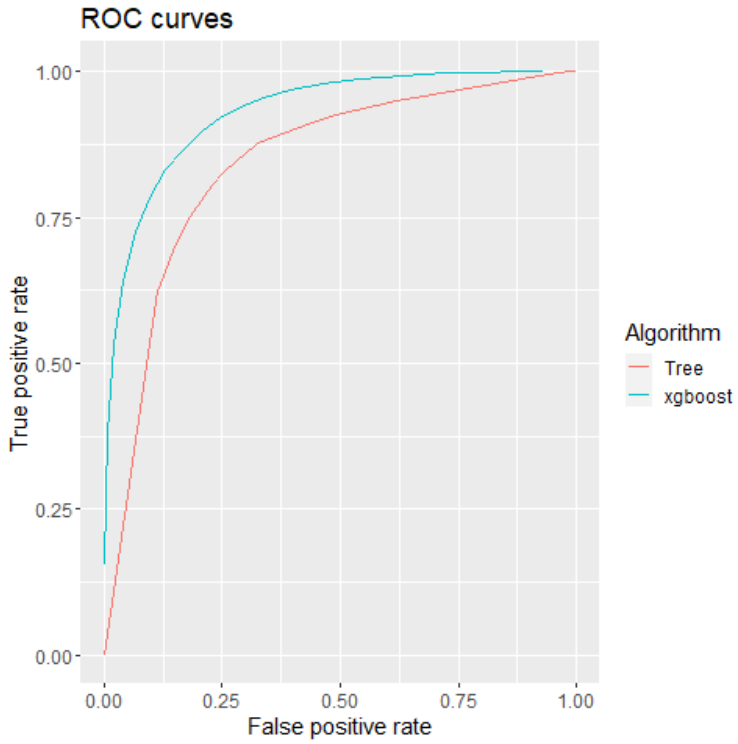
[2]For glm package in R, the first level of the category is removed in the regression

Figure 1: Mean square error by parameter value



and the classification tree can be seen in Figure 2. The xgboost classifier was the strongest classifier with an MSE of 0.09, compared to the 0.14 MSE of the classification tree, both computed with the estimated probabilities. A decision value can be defined using a required sensitivity or using as a threshold the proportion of positive cases in the data. Using the 68.5% of observed positive cases as minimum probability to predict AD, the xgboost classifier achieves a sensitivity of 83% and a specificity of 87%.

Figure 2: ROC of best algorithms



4 Real data analysis

The main objective of this study was to assess the impact of the e4 APOE allele in predictive models in AD and so far performance of some 3 classifiers has been tuned and compared. Now, the objective is to evaluate how models being blinded of APOE information perform among the overall population, and after stratifying by the number of e4 alleles, comparing performance among the "right" and "wrong" population. This was be done with the previoulsy tuned xgboost classifier.

First, after removing the APOE information, the algorithm achieved an MSE of 0.086, slightly lower than with the APOE variables, which might be due to the conflicting results discussed on the introduction. Then, training a model in each subpopulation resulted in an MSE of 0.084, 0.048, 0.006 for the population with 0,1 and 2 alleles respectively. These results suggest that the relationship of the e4 allele to disease status is more complex than a linear term (original models weren't trained with interactions nor any other higher order variables). Model performance for the overall population and the 3 sub-populations can be seen[3]. on Table 2 It is interesting to note how

Table 2: Performance of predictive models on (sub)populations

	All	0 e4 alleles	1 e4 alleles	2 e4 alleles
N	53K	28K	20K	5K
Observed % AD	68.5%	60.6%	75%	85%
Predicted % AD	71.9%	63.2%	78%	86%
MSE	0.086	0.08	0.04	0.006
Miss classification	11.2%	10.6%	5.7%	0.4%
False negatives	7.3%	6.6%	4.2%	0.04%
False positives	3.9%	4%	1.5%	.4 %

the performance improves when stratifying by the e4 allele, particularly with 2 e4 alleles, where model classifies almost perfectly; it's also worth mentioning that this problem is easier as it has the proportion of AD cases is further away from 0.5

Now, model performance can be tested against the "wrong" population, and to analyze the predicted percentage of AD cases vs. the observed one and see if models trained on populations with more e4 alleles tend to overestimate AD cases and vice versa. Results can be seen on Table 3.

Table 3 can be read as follows: reading down on a column, covariates remain the same and difference in predicted cases is explained by a model that was trained on a population with more positive cases, or where interaction of the e4 allele with covariates measuring symptoms, comorbidities, cognitions scores and other relevant information is not captured.

On the other hand, reading on a row captures the changes in variables that might be due to changes in the e4 allele; for example, having more e4 alleles could be reflected on cognition scores. Difference to observed percentage of cases along this direction is due to the models missing the impact of e4 alleles.

[3]To compute predicted AD, the threshold of 0.5 was chosen to avoid introducing bias from the overall population or introducing information that wouldn't be available without the APOE e4 allele

Table 3: Models tested on different populations

	0 e4 alleles	1 e4 alleles	2 e4 alleles
Observed cases	60.6%	75%	85%
	Predicted cases		
Trained on 0 e4 alleles	63.2%	75.2%	81.8%
Trained on 1 e4 alleles	68.6%	78.1%	86.5%
Trained on 2 e4 alleles	74.7%	83.4%	86.9%

The general trend is that below the diagonal; that is, where models are trained on population with more e4 alleles than where tested, models tend to overestimate having AD. On the other hand, above the diagonal, models tend to underestimate number of cases. Predicted cases for the models trained on 0 e4 alleles and on 1 e4 allele seem close the observed cases with 1 more allele, although both models tend to overestimate it on their own population.

These results tend to support the hypothesis that number of e4 alleles is determinant in predicting (diagnosing) AD, with increasing risk along increasing number of alleles. Furthermore, it seems that the pathway through which e4 alleles impact AD is not captured in the 32 variables measuring demographics, comorbidities, lifestyle, symptoms, and cognition scores; supporting its use as a biomarker in clinical practice.

5 Discussion

This document aimed at studying the relevance of the APOE e4 alleles for predicting / diagnosing Alzheimer's disease. It was shown that models with reasonable predictive performance can be built with the main variables collected at clinical visits, even when some of them might be missing. More interestingly, it was shown that having variables describing the pair of alleles in subjects and the number of e4 alleles could result in a little worse MSE: particularly the MSE when having these variables was 0.09 vs. 0.086, suggesting that there is a small segment of the population where the APOE alleles might have the opposite interaction as expected. An example of this case would be the discussion at the beginning where very old people with e4 alleles have been observed to have better cognitive function. Finally, it was shown that the number of e4 alleles increases the risk of Alzheimer's disease, even when accounting for the variables that had shown good predictive power on their own.

The main limitations of this study are the selection bias in the population, as only subjects that attended an Alzheimer's research center and were accepted according to the center protocol were studied. Additionally, it was observed on the study data that variables were measured more frequently on subjects with Alzheimer's, so that the mere fact of having something measured was indicative of having the disease.

It was also interesting to see how models trained on one population could not be applied consistently to other populations. Although, this seems like an obvious result, genetic studies might not conducted equally across populations and races; and models trained on one population might generalize poorly if this is not acknowledged.

Finally, it would be interesting to further study in which subpopulations the impact of the e4 allele is stronger, or even reversed, like the case of very old people having better cognitive function; and to study how this allele can also be used to predict disease progression, measured by the cognitive tests performed in practice.

References

AIM. Alzheimer's and dementia research, 2021. URL https://alzimpact.org/issues/research.

Rachel Butch. We need to talk about alzheimer's disease, 2017. URL https://www.hopkinsmedicine.org/research/advancements-in-research/fundamentals/in-depth/we-need-to-talk-about-alzheimers-disease.

Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Ya-Ju Wu et al Jen-Hung Wang. Medical comorbidity in alzheimer's disease: A nested case-control study. *Journal of Alzheimer's disease*, 63(2):773–781, 2018. doi: 10.3233/JAD-170786. URL https://pubmed.ncbi.nlm.nih.gov/29660933/.

Alezzandra Bizzarro et al. Marra Camilo. Apolipoprotein e e4 allele differently affectsthe patterns of neuropsychologicalpresentation in early and late onset alzheimer's disease patients. *Dementia and Geriatric Cognitive Disorders*, 18:125–131, 2004. doi: 10.1159/000079191.

NACC. About nacc data, 2021. URL https://naccdata.org/requesting-data/nacc-data.

NIH. Alzheimer's disease fact sheet, 2021. URL https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet.