Grant Halver (Statistics, 1st year grad student)
Adam Heike (Industrial Engineering, 1st year grad student)
Yin Jin (Interdisciplinary life science , 2st year grad student)

STAT525 Group 5 Project Report

Introduction

Neurodegenerative disease is one of the most terrifying diagnoses a patient and their family can receive. Of all the conditions in this category, Alzheimer's disease (AD) is perhaps the most well-known and the one most greatly feared. There are no treatments to stop or reverse the progression of this disease and accurately diagnosing the condition is difficult, even among skilled clinicians. (Archer et al., 2017) This project seeks to help physicians in their diagnosis of Alzheimer's patients by testing to see if specific biomarkers and patient demographic information are correlated with the presence of the disease.

The Mini-Mental State Exam (MMSE) is the most common tool used in the diagnosis and assessment of Alzheimer's disease patients. The MMSE is a 30-point test that is used to measure a person's level of cognitive impairment. A score of 23 or lower is indicative of dementia, but this test alone is not enough to confirm a diagnosis. In fact, the MMSE is best used in ruling out a dementia diagnosis. (Mitchell, 2009) Based on how the test is scored, false negatives are a concern, especially among patients with mild cognitive impairment and highly educated individuals. (Erdődi et al., 2020) (Laszlo et al., 2020) Using a data set obtained from the Alzheimer's Disease Neuroimaging Institute, we attempt to predict the MMSE scores of Alzheimer patients using patient information such as age, level of education, and biospecimen concentrations. If a model can be developed that accurately predicts the MMSE scores of AD patients based on their demographic background and biomarkers, then the problem of false negatives can be greatly reduced. This may aid clinicians in interpreting the MMSE for those patients that score well on the exam but are still suspected of having dementia.

Methods

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

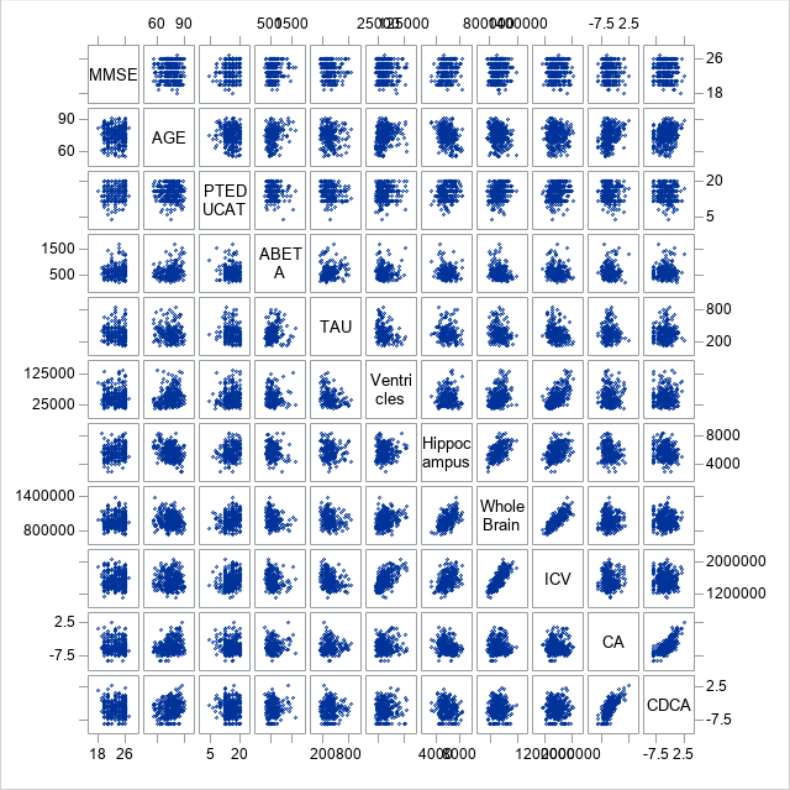
The goal of this project is to identify factors that can be used to predict the Mini-Mental State Exam score of Alzheimer's disease patients. We begin the analysis by evaluating the continuous variables within the ADNI data set with a correlation plot. This displays the relationship between each of these variables and MMSE score so that only those with a potential linear relationship are selected for further analysis. This approach has the additional benefit of helping deal with the problem of missing values within the data set. Patient records that contain missing values in variables that are not linearly related to MMSE can still be used in the analysis, whereas these records might otherwise be excluded for being incomplete. The size of the data set makes it difficult to assess by graphical means if the conditions for inference are met. A Box-Cox transformation is applied to the MMSE scores to determine what transformation, if any, of the response variable will make the residuals of the model approximately normal.

Once the conditions for inference are met, the model selection process can begin. Considering many explanatory variables for our model presents a challenge for model selection. This is addressed by using a forward selection process that tests whether the addition of each additional variable to the model is significant. The criterion used in the model selection is the Schwarz Bayesian, which provides a penalty for added complexity within the model. For each of the variables in the selected model, partial regression plots are used to assess the variances and determine if a transformation of any predictor variable is necessary.

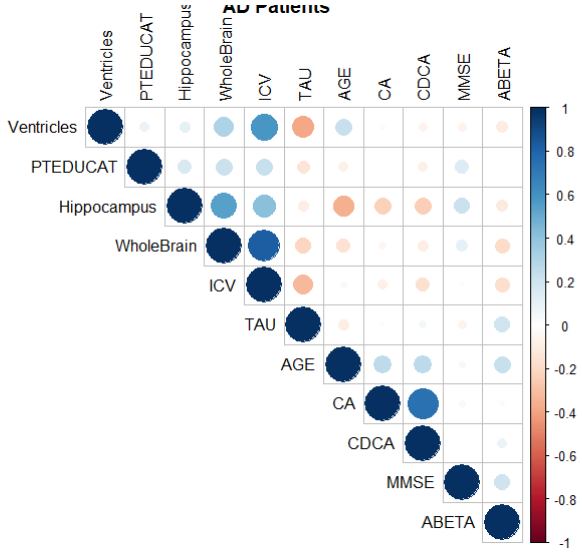
Finally, the potential of influential cases within the data set is considered along with the potential effect that they may have on the model parameter values. Cook's D and DFFITS are used to determine which cases within the data set are influential and then re-evaluate the model with the removal of those influential cases. This will be accomplished by an approximate 90-10 split of the data into a large training set to develop the model and a smaller set with which is used to evaluate the differences between the model predicted values and the actual MMSE scores within the test set. If there is no significant difference between the residual values of the models with and without the points identified by Cook's D & DFFITS then we can be confident that these points are not so influential that they could impact the results.

The last part of the analysis will consider the ability of the model to distinguish between MCI and those with advanced stages of Alzheimer's disease using a linear classifier. If there is some value of the explanatory variables that delineates the difference between the stages of cognitive impairment, it would make this model an extremely valuable tool for clinicians in helping to accurately diagnose dementia patients.

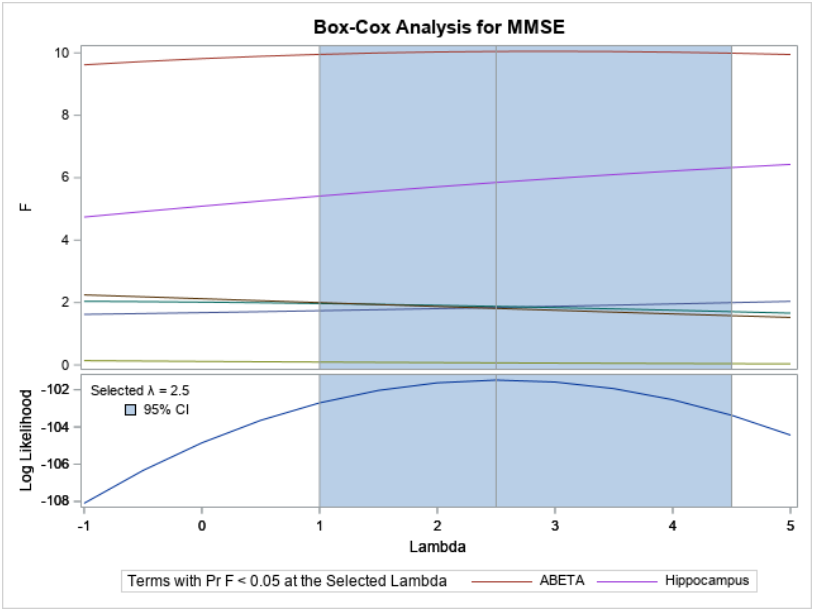
Results



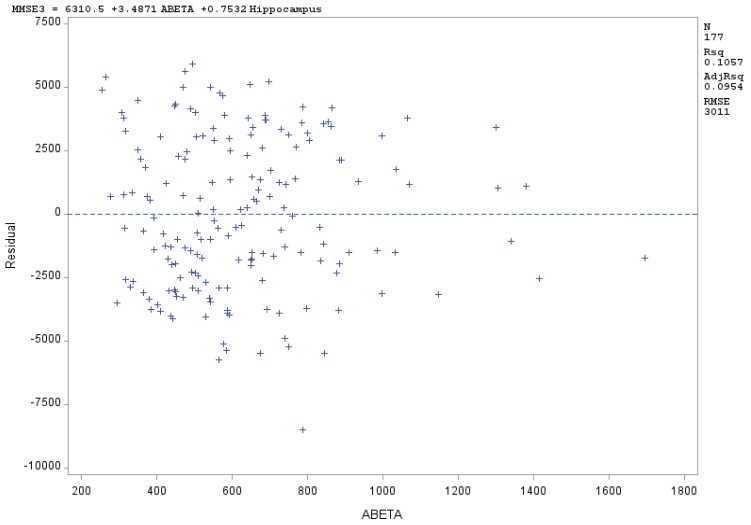
The analysis began by using correlation plots for the Alzheimer's patients. Two different visualizations are provided, one above and one below. From the two visualizations we can confirm that the Age, ICV, CA, and CDCA have almost no correlation with MMSE. Those variables will be removed.



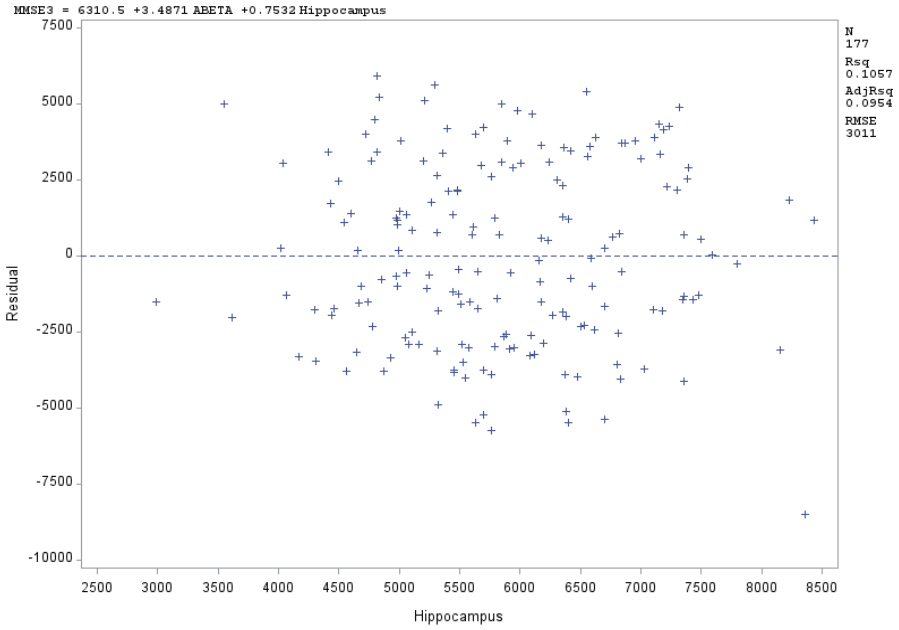
Next we use a Box-Cox plot to check for a transformation. From the plot we can see that the selected λ = 2.5. So we will proceed with the nearest λ of 3, which is the cubic transformation.



The best model was picked using forward selection with the Schwarz Bayesian Criterion (SBC). The use of forward selection compensates for the many incomplete observations in the data set. By building the model up, this avoids having models fit with overly many points removed. The SBC criterion selected the best model based on the minimum of -2log(likelihood)+plog(n) where plog(n) is a penalty for more complex models. The selected model ended up using two variables only: ABETA and Hippocampus.

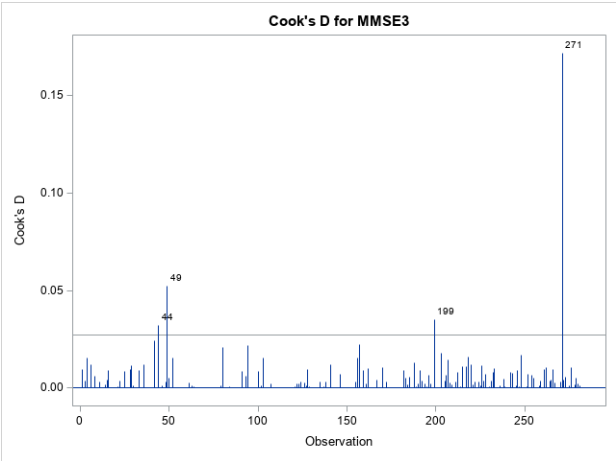


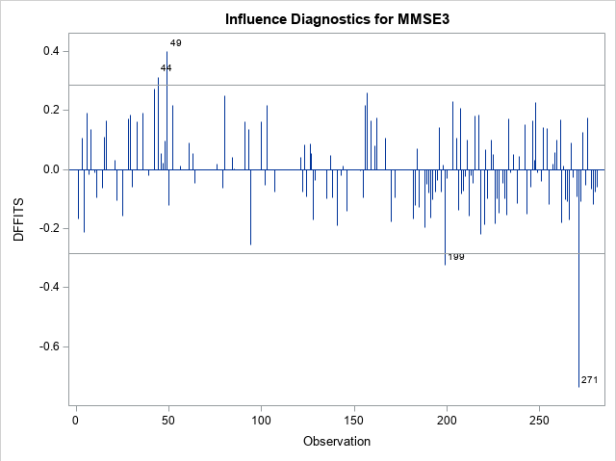
Next partial regression plots will be used to make sure that none of the predictor variables need to be transformed. The plot above possibly shows non-constant variance as ABETA increases. However, there are very few data points with a high ABETA so no conclusion can be made. There is otherwise nothing of concern in the above plot. The plot below has no visible patterns. However, there is a potential influential point on the bottom right of the graph.



The removal of such influential points is now considered to ensure that our model cannot be improved. To test this, the data set was split into two sets: one training set and one validation set. The total data set had 319 observations, including incomplete observations, so an approximate 10% of the total observations of 30 complete observations, plus 6 incomplete observations, were selected for the validation set. The remaining observations were left in the training set.

Cook's D and DFFITS in the graphs below both return 271 as the most influential point so it is clear that the removal of this point should be explored.





The average absolute percent change in the predicted residual for the validation set with the model fit from the validation set was then computed with and without the selected observation. However, the average absolute percent change was -0.85% which is small and less than 1% so the influential point was not removed. Since all other points identified as influential on the graphs above will have a smaller effect on the model than point 271, their removal will not be considered.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: MMSE3

Number of Observations Read	281
Number of Observations Used	148
Number of Observations with Missing Values	133

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	170973526	85486763	9.89	<.0001
Error	145	1252750300	8639657		
Corrected Total	147	1423723826			

Root MSE	2939.32939	R-Square	0.1201
Dependent Mean	12695	Adj R-Sq	0.1080
Coeff Var	23.15358		

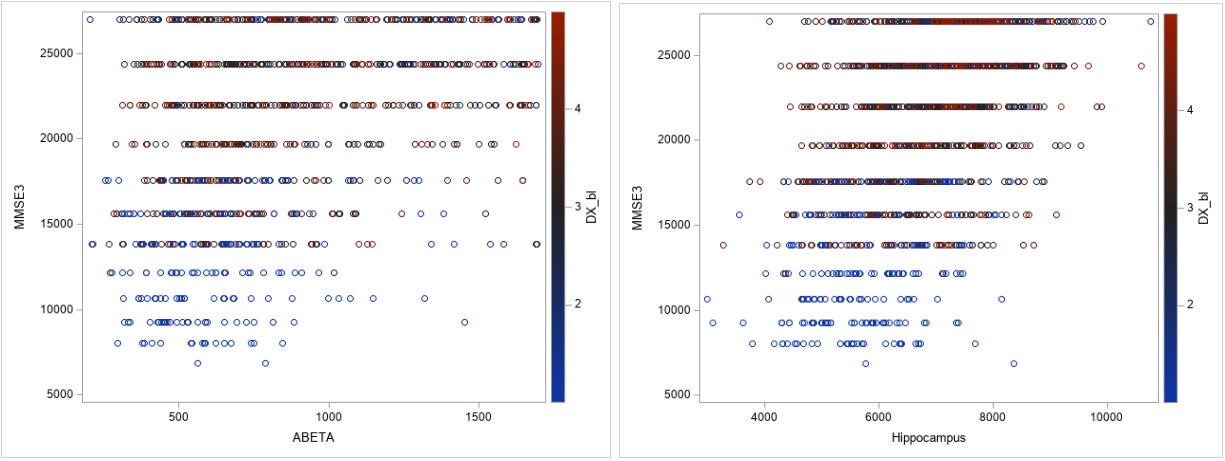
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6087.09191	1632.60758	3.73	0.0003
ABETA	1	3.80813	1.06526	3.57	0.0005
Hippocampus	1	0.73283	0.24053	3.05	0.0027

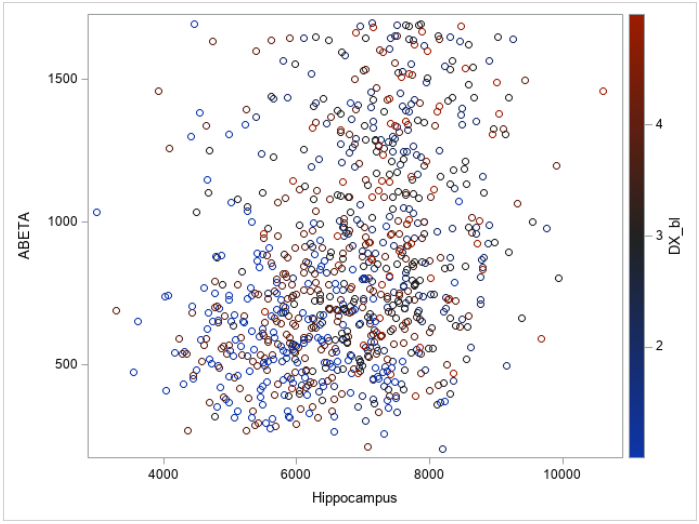
The final model from the analysis is therefore $MMSE^3$ = 6087.09191 + 3.80813*ABETA + 0.73283*Hippocampus. This model has an F-value of 10.28 and a p-value of less than 0.0001

suggesting that model is overall significant. However, the adjusted R^2 is rather low at 0.0954 which suggests that most of the variation in the data is not explained by the model. This result is not surprising given the complexity of Alzheimer's Disease and the fact that our model selection resulted in only two variables being selected. The addition of more variables to the dataset should be considered so that a more complex model may be constructed. On the other hand, the explained variation for only two variables given the number of associations that have been suggested for Alzheimer's Disease is relatively high which suggests that we have identified two highly significant predictor variables.

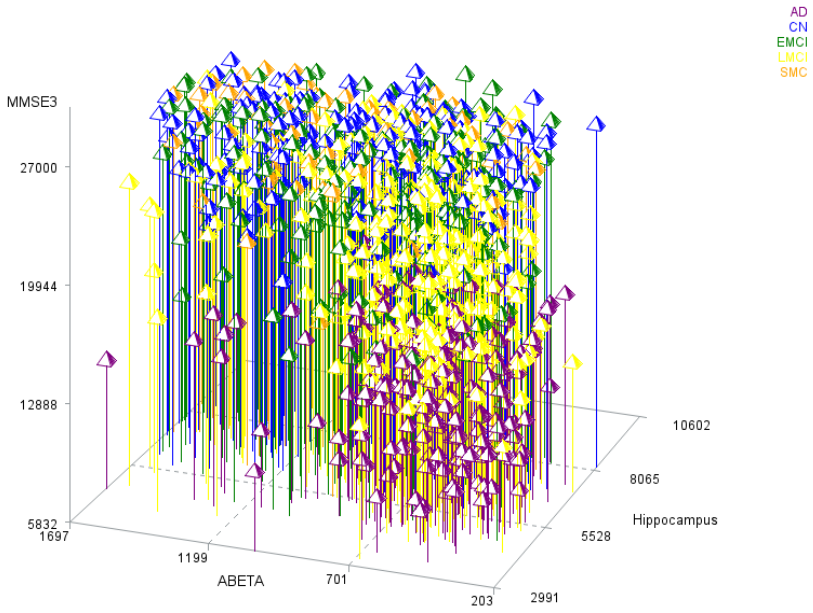
Since the linear model is not good at explaining the total variation, a linear classifier is now considered using the variables in our selected model. When using only one of our predictor variables, the below graphs show that there is no clear separation so a linear classifier using only one of the predictor variables will not work.



Attempting again to create a simple linear classifier, both predictor variables have been used below. However, this also results in no clear separation so no linear classifier is possible.



Changing to a multivariate setting is one final attempt at a linear classifier. From the graph below it can be seen that the Alzheimer's Disease patients (AD or purple on the graph) are not separated cleanly from all of the other groups. Notably the late mild cognitive impairment (LMCI or yellow) and early mild cognitive impairment (EMCI or green) have significant overlap with the AD group. However, the control (CN or blue) and significant memory concern (SMC or orange) groups are nearly separated from the Alzheimer's patients.



There are two groups of patients that could possibly be grouped by a linear classifier: those who are AD, LMCI, or EMCI and those who are CN or SMC. This suggests that as a patient progresses from healthy to receiving an Alzheimer's diagnosis that the changes in ABETA and the Hippocampus become less significant but those changes are a significant factor separating the control and early cases from later progressions of the disease.

Discussion

Our final model indicated that the Cerebrospinal Fluid (CSF) amyloid beta levels and brain hippocampal volume can be used to predict the MMSE score in AD patients. Overall, CSF amyloid beta levels had a positive relationship with MMSE score, indicating that reduced CSF amyloid beta levels can be linked to lower cognitive scores in AD patients. Amyloid beta was derived from amyloid precursor protein (APP), which was abnormally cleaved by β-secretase and γ-secretase enzymes.(Murphy and Levine 2010) The insoluble amyloid beta deposited in aggregated fibrils in senile plaques around neurons causing neurons death. It was hypothesized that the deposition of the amyloid beta in the brain prevents its transit into the CSF and leads to reduced CSF amyloid beta levels. Such decreased CSF amyloid beta levels can be used to reflect the presence of amyloid beta pathology of AD. The phenomenon that concentration of amyloid beta in the CSF is significantly reduced in AD and related with MMSE is consistent with previous studies.(Skoog et al. 2003)(Snider et al. 2009)(Deming et al. 2017) Thus, CSF amyloid beta levels can serve as a reliable predictor for MMSE in AD patients.

Hippocampal volume can also be a promising predictor for MMSE score in AD patients. The feature of Alzheimer's disease is neurons and brain volume loss.(Braak and Braak 1998) Hippocampus, the deep region in the temporal lobe, is related memory, learning, regulation and emotion.(Rubin et al. 2014) Therefore, the progressive hippocampal volume loss is responsible for memory loss and dysfunction of learning and planning, which are the hallmark symptom of Alzheimer's.(Yoshiyama et al. 2007) Many structural MRI studies have also suggested hippocampal atrophy can serve as an accurate predictive marker of progression to AD.(Dubois, Picard, and Sarazin 2009)(Du et al. 2019) In our final model, hippocampal volume had a positive relationship with MMSE score, which indicated hippocampal volume loss can cause lower cognitive score in AD patients.

The adjusted coefficient of determination (adjusted R^2) is rather low in our final model which suggested that CSF amyloid beta levels and brain hippocampal volume accounted for only a small proportion of the MMSE scores in the total sample. Given Alzheimer's disease involves complex brain structural and functional change, only CSF amyloid beta levels and brain hippocampal volume being selected in linear models may not be sufficient to explain MMSE. Besides, the sample size of this study is relatively small. Thus, more complex models, such as longitudinal models or mixed models, may be constructed or larger datasets may be used to achieve a higher power for future model prediction. Nonetheless, the explained variation for only two variables is almost 10 percent, suggesting that we have identified two highly significant predictor variables.

Attempting to explore multivariate classifiers in different conditions, AD patients were not separated cleanly from all the other groups. One possible explanation is that amyloid beta and brain atrophy may also contribute memory variability in healthy elderly and MCI.(Svenningsson et al. 2019)(Hatashita and Wakebe 2017)(Shi et al. 2009) However, the control and SMC groups were roughly separated from the AD patients, whereas LMCI and EMCI had significant overlap with the AD group. This suggests that CSF amyloid beta levels and brain hippocampal volume could be sensitive classifiers to separate the control and significant memory concern from mild cognitive impairment and AD and insensitive to differentiate mild cognitive impairment and AD.

In conclusion, CSF amyloid beta levels and brain hippocampal volume were identified as two highly significant and reasonable predictors for MMSE score in AD patients. Using these two variables as linear classifiers, they can roughly separate the control and SMC groups from MIC and AD. Though encouraging, these findings also indicate that the model would not have a significant impact on the ability of clinicians to detect false positives among high-scoring MMSE patients. Further improvement of the model may be possible with additional research that focuses on variables that are not tracked by the Alzheimer's Disease Neuroimaging Initiative group.

References

Archer, M.C., Hall, P.H. and Morgan, J.C. (2017), [P2–430]: ACCURACY OF CLINICAL DIAGNOSIS OF ALZHEIMER's DISEASE IN ALZHEIMER's DISEASE CENTERS (ADCS). Alzheimer's & Dementia, 13: P800-P801. https://doi-org.ezproxy.lib.purdue.edu/10.1016/j.jalz.2017.06.1086

Braak, H., and E. Braak. 1998. “Evolution of Neuronal Changes in the Course of Alzheimer’s Disease.” *Journal of Neural Transmission, Supplement*.

Deming, Yuetiva et al. 2017. “Genome-Wide Association Study Identifies Four Novel Loci Associated with Alzheimer’s Endophenotypes and Disease Modifiers.” *Acta Neuropathologica* 133(5): 839–56.

Du, Lei et al. 2019. “Identifying Progressive Imaging Genetic Patterns via Multi-Task Sparse Canonical Correlation Analysis: A Longitudinal Study of the ADNI Cohort.” In *Bioinformatics*,.

Dubois, Bruno, Gaetane Picard, and Marie Sarazin. 2009. “Early Detection of Alzheimer’s Disease: New Diagnostic Criteria.” *Dialogues in Clinical Neuroscience*.

Erdődi, L. A., Shahein, A. G., Fareez, F., Rykulski, N., Sabelli, A. G., & Roth, R. M. (2020). Increasing the cutoff on the MMSE and DRS-2 improves clinical classification accuracy in highly educated older adults. Psychology & Neuroscience, 13(1), 93–113. https://doi-org.ezproxy.lib.purdue.edu/10.1037/pne0000185

Hatashita, Shizuo, and Daichi Wakebe. 2017. “Amyloid-β Deposition and Long-Term Progression in Mild Cognitive Impairment Due to Alzheimer’s Disease Defined with Amyloid PET Imaging.” *Journal of Alzheimer’s Disease*.

Laszlo A. Erdodi, Ayman G. Shahein, Katrina J. Kent & Robert M. Roth (2020) The doubtful benefits of giving the benefit of the doubt: Lenient scoring of the spatial orientation items on the mini-Mental Status Exam increases false negative rates, Applied Neuropsychology: Adult, 27:2, 143-149, DOI: 10.1080/23279095.2018.1497990

MahmoudianDehkordi S, Arnold M, Nho K, Ahmad S, Jia W, Xie G, Louie G, Kueider-Paisley A, Moseley MA, Thompson JW, St John Williams L, Tenenbaum JD, Blach C, Baillie R, Han X, Bhattacharyya S, Toledo JB, Schafferer S, Klein S, Koal T, Risacher SL, Kling MA, Motsinger-Reif A, Rotroff DM, Jack J, Hankemeier T, Bennett DA, De Jager PL, Trojanowski JQ, Shaw LM, Weiner MW, Doraiswamy PM, van Duijn CM, Saykin AJ, Kastenmüller G, Kaddurah-Daouk R; Alzheimer's Disease Neuroimaging Initiative and the Alzheimer Disease Metabolomics Consortium. Altered bile acid profile associates with cognitive impairment in Alzheimer's disease-An emerging role for gut microbiome. Alzheimers Dement. 2019 Jan;15(1):76-92. doi: 10.1016/j.jalz.2018.07.217. Epub 2018 Oct 15. Erratum in: Alzheimers Dement. 2019 Apr;15(4):604. PMID: 30337151; PMCID: PMC6487485.

Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. J Psychiatr Res. 2009 Jan;43(4):411-31. doi: 10.1016/j.jpsychires.2008.04.014. Epub 2008 Jun 24. PMID: 18579155.

Murphy, M. Paul, and Harry Levine. 2010. “Alzheimer’s Disease and the Amyloid-β Peptide.” *Journal of Alzheimer’s Disease*.

Rubin, Rachael D., Patrick D. Watson, Melissa C. Duff, and Neal J. Cohen. 2014. “The Role of the Hippocampus in Flexible Cognition and Social Behavior.” *Frontiers in Human Neuroscience*.

Shi, Feng et al. 2009. “Hippocampal Volume and Asymmetry in Mild Cognitive Impairment and Alzheimer’s Disease: Meta-Analyses of MRI Studies.” *Hippocampus*.

Skoog, I. et al. 2003. “Cerebrospinal Fluid Beta-Amyloid 42 Is Reduced before the Onset of Sporadic Dementia: A Population-Based Study in 85-Year-Olds.” *Dementia and Geriatric Cognitive Disorders*.

Snider, Barbara J. et al. 2009. “Cerebrospinal Fluid Biomarkers and Rate of Cognitive Decline in Very Mild Dementia of the Alzheimer Type.” *Archives of Neurology*.

Svenningsson, Anna L. et al. 2019. “β-Amyloid Pathology and Hippocampal Atrophy Are Independently Associated with Memory Function in Cognitively Healthy Elderly.” *Scientific reports*.

Yoshiyama, Yasumasa et al. 2007. “Synapse Loss and Microglial Activation Precede Tangles in a P301S Tauopathy Mouse Model.” *Neuron*.

Appendix A
R Code

```
#Load the package and necessary libraies
install.packages("/Users/jinyin/Downloads/ADNIMERGE_0.0.1.tar.gz", repos=NULL,
type="source")

library(ADNIMERGE)
library(dplyr)

#Load the two datasets from package
testdata<-data.frame(adnimerge)
bile<-data.frame(admcbapp)

#Select variables for analysis
names(bile)
names(baseline)
newbaseline <- baseline%>%
  select(RID,AGE, DX.bl, PTGENDER, PTEDUCAT, PTRACCAT, PTETHCAT, MMSE, ABETA,
TAU, Ventricles, Hippocampus, ,WholeBrain, ICV)
newbile<- bile%>%
  select( RID,CA,CDCA)


#Merge two datasets into one
newbaseline$RID<-as.integer(newbaseline$RID)
biledateaset<- inner_join(newbaseline,newbile,by="RID")


#Rename varialbes and define numeric varialbes
names(biledateaset)[names(biledateaset)=="DX_bl"] <- "DX"
biledateaset$ABETA <-as.integer(biledateaset$ABETA)
biledateaset$TAU <-as.integer(biledateaset$TAU)
biledateaset$Ventricles <-as.integer(biledateaset$Ventricles)
biledateaset$Hippocampus <-as.integer(biledateaset$Hippocampus)
biledateaset$WholeBrain <-as.integer(biledateaset$WholeBrain)
biledateaset$ICV <-as.integer(biledateaset$ICV)



#Split dataset into three subsets based on diagnositic information
CN <- biledateaset[ which(biledateaset$DX=="CN"), ]
AD <- biledateaset[ which(biledateaset$DX=='AD'), ]
MCI <- biledateaset[ which(biledateaset$DX=='SMC'|biledateaset$DX=='EMCI'|
biledateaset$DX=='LMCI'), ]


table(biledateaset$DX)


#Trim dataset for correlation tests and plots
```

```
CN1 <- CN[, c(2,5,8,9,10,11,12,13,14,15,16)]
AD1 <- AD[, c(2,5,8,9,10,11,12,13,14,15,16)]
MCI1 <- MCI[, c(2,5,8,9,10,11,12,13,14,15,16)]


res<-na.omit(MCI1)
res <- cor(res)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", title = "MCI1 Patients")

res<-na.omit(AD1)
res<- AD1
res <- cor(res)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", title = "AD Patients")

res<-na.omit(CN1)
res <- cor(res)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", title = "CN Patients")

#Converting categorical variables to numbers instead of strings
biledateaset$DX.bl <- as.numeric(type.convert(biledateaset$DX.bl))
biledateaset$PTGENDER <- as.numeric(type.convert(biledateaset$PTGENDER))
biledateaset$PTEDUCAT <- as.numeric(type.convert(biledateaset$PTEDUCAT))
biledateaset$PTRACCAT <- as.numeric(type.convert(biledateaset$PTRACCAT))
biledateaset$PTETHCAT <- as.numeric(type.convert(biledateaset$PTETHCAT))

#Save as .csv to import into SAS
write.csv(biledateaset,"C:\\Users\\Grant\\Documents\\STAT525\\biles.csv", row.names = FALSE)
```

Appendix B
SAS Code

```
**Data imported from .csv file;
data _null_;
  infile 'C:\\Users\\Grant\\Documents\\STAT525\\biles.csv' dsd truncover ;
  file 'C:\\Users\\Grant\\Documents\\STAT525\\to_sas.csv' dsd ;
  length word $200 ;
  do i=1 to 16;
    input word @;
    if word='NA' then word=' ';
          if word='>1700' then word='1700';
          if word='>1300' then word='1300';
          if word='<80' then word='80';
          if word='<200' then word='200';
    put word @;
  end;
  put;
run;

**Full data set;
proc import datafile="C:\\Users\\Grant\\Documents\\STAT525\\to_sas.csv"
      out=a1
      dbms=csv
      replace;
run;
quit;

**Confirming which variables to keep;
proc sgscatter data=a1 (where=(DX_bl eq 1));
        matrix MMSE AGE PTEDUCAT ABETA TAU Ventricles Hippocampus  WholeBrain ICV
CA CDCA;
run;
**Age, ICV, CA, CDCA excluded at this step;

**Box-Cox Plot to confirm transformation;
proc transreg data=a1 (where=(DX_bl eq 1));
        model boxcox(MMSE/ lambda=-1 to 5 by 0.5) = identity(PTEDUCAT)identity(ABETA)
identity(TAU)
        identity(Ventricles)identity(Hippocampus) identity(WholeBrain);
run;
**Cube transform suggested;

data a1 (where=(DX_bl eq 1));
        set a1 (where=(DX_bl eq 1));
        MMSE3 = MMSE * MMSE * MMSE;
run;

proc glmselect data=a1 (where=(DX_bl eq 1));    **using FORWARD selection to fit model due
to missing cases;
        class PTGENDER PTRACCAT PTETHCAT;
```

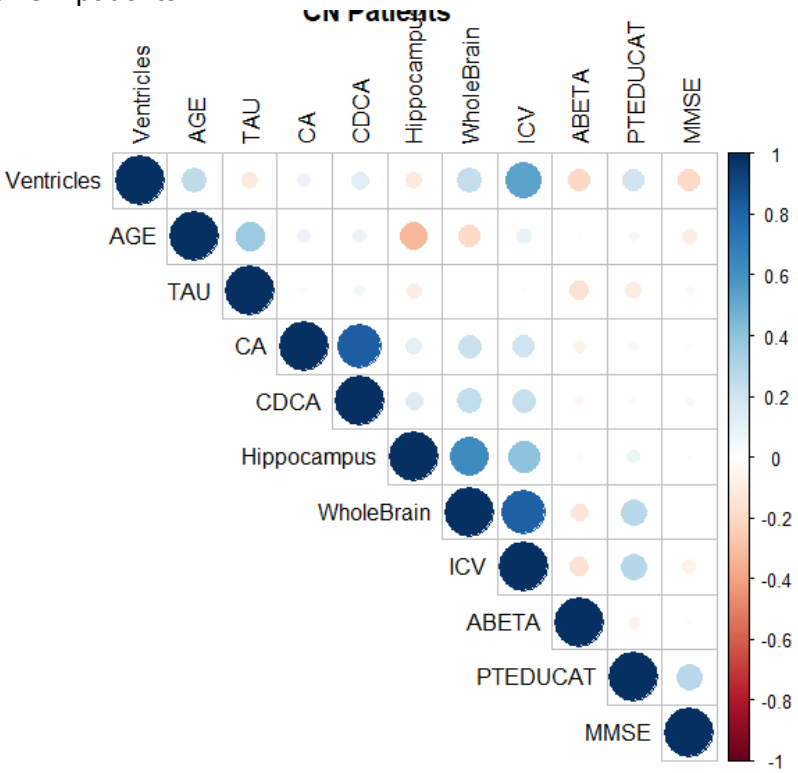
```
        model MMSE3 = PTGENDER PTEDUCAT PTRACCAT PTETHCAT ABETA TAU
Ventricles Hippocampus WholeBrain/selection= FORWARD;
run;
**ABETA and Hippocampus selected;
**F=9.08, df1=2, df2=161, p-value=0.0002;
**Adj R^2 = 0.0902;

**Checking partial regression plots to see if transforming x vars will give better fit;
proc reg data=a1 (where=(DX_bl eq 1));
model MMSE3=ABETA Hippocampus / r partial influence tol;
id ABETA Hippocampus; plot r.*(p. ABETA Hippocampus);
run;
**No transform of x vars suggested;
**MMSE3 = 6310.5 + 3.4871*ABETA + 0.7532*Hippocampus;
**New line due to difference in obs used?;

**Reduced training data set for regression;
proc import datafile="C:\\Users\\Grant\\Documents\\STAT525\\to_sas.csv"
      out=a2
      dbms=csv
      replace;
run;
quit;

data a2;
        set a2 (where=(DX_bl eq 1));
        If RID >= 5029 then Delete;
        MMSE3 = MMSE * MMSE * MMSE;

**Reduced data set for validation;
**30 complete cases;
proc import datafile="C:\\Users\\Grant\\Documents\\STAT525\\to_sas.csv"
   out=a3
   dbms=csv
   replace;
run;
quit;

data a3;
        set a3 (where=(DX_bl eq 1));
        If RID < 5029 then Delete;
        MMSE3 = MMSE * MMSE * MMSE;
run; quit;

**Checking for influential points using training data set;
proc reg data=a2 (where=(DX_bl eq 1)) plots(only) = (CooksD(label) DFFits(label));
  model MMSE3=ABETA Hippocampus;
  output out=RegOut pred=Pred rstudent=RStudent dffits=DFFits cookd=CooksD;
run; quit;
**271 most influential per Cook's D and DFFITS;
**MMSE3 = 6087.09191 + 3.80813*ABETA + 0.73283*Hippocampus;
```

```
**Find RID so only inf points get deleted later;
proc print data = RegOut;
        where CooksD > 0.15;
run;

**If 45 is deleted;
data a4;
        set a2;
        If RID = 4997 then Delete;

proc reg data=a4 (where=(DX_bl eq 1));
  model MMSE3=ABETA Hippocampus;
run; quit;
**MMSE = 5068.02415 + 4.08788*ABETA + 0.88847*Hippocampus;

data a3;
        set a3;
        predres = MMSE3-(6087.09191 + 3.80813*ABETA + 0.73283*Hippocampus);
        deletepredres = MMSE3-(5068.02415 + 4.08788*ABETA + 0.88847*Hippocampus);
        changeresid = abs(deletepredres-predres)/predres*100;

proc means data = a3;
        var changeresid;
run; quit;
**Within 1% so no change is needed;
**5% is threshold;

**Linear classifier?;
proc import datafile="C:\\Users\\Grant\\Documents\\STAT525\\to_sas.csv"
      out=a6
      dbms=csv
      replace;
run;
quit;

data a6;
        set a6;
        MMSE3 = MMSE * MMSE * MMSE;
        length color $8.;
        if DX_bl = 1 then color = "purple";
        if DX_bl = 2 then color = "blue";
        if DX_bl = 3 then color = "green";
        if DX_bl = 4 then color = "yellow";
        if DX_bl = 5 then color = "orange";
run;


proc sgplot data=a6;
        scatter x=ABETA y=MMSE3 / colorresponse=DX_bl;
run;
```

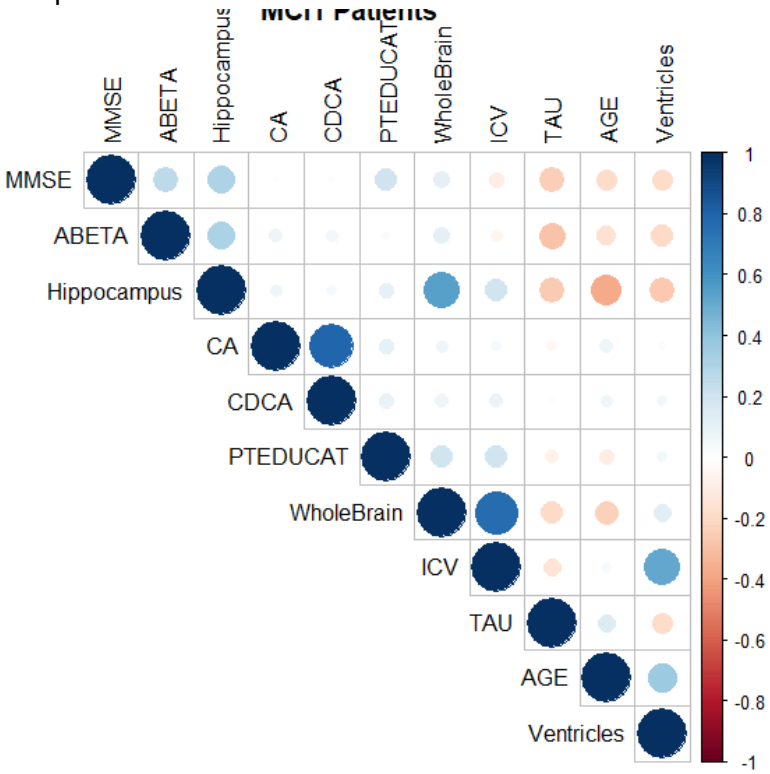
```
proc sgplot data=a6;
	scatter x=Hippocampus y=MMSE3 / colorresponse=DX_bl;
run;
**Neither var is sufficient on their own;

proc sgplot data=a6;
	scatter x=Hippocampus y=ABETA / colorresponse=DX_bl;
run;
**Both predictor variables are insufficient;

proc G3D data=a6;
	 scatter ABETA*Hippocampus=MMSE3 / color=color;
	 note j=r c=purple 'AD' j=r c=blue 'CN' j=r c=green 'EMCI' j=r c=yellow 'LMCI' j=r
c=orange 'SMC';
run; quit;
**DX_bl =1 is mixed with 3 and 4 so no linear classifier from all groups is possible;
**Could distinguish between healthy and Alzheimer's if those are the only two groups of interest;
```

Appendix C
Additional Figures

Correlation plot for CN patients



Correlation plot for MCI1 patients



Regression results for fit with influential point removed

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: MMSE3

Number of Observations Read	280
Number of Observations Used	147
Number of Observations with Missing Values	133

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	210682502	105341251	12.87	<.0001
Error	144	1178751608	8185775		
Corrected Total	146	1389434110			

Root MSE	2861.07935	R-Square	0.1516
Dependent Mean	12735	Adj R-Sq	0.1398
Coeff Var	22.46693		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5068.02415	1624.88779	3.12	0.0022
ABETA	1	4.08788	1.04107	3.93	0.0001
Hippocampus	1	0.88847	0.23979	3.71	0.0003

Output for average absolute percent change in residuals for the validation data set

The SAS System

The MEANS Procedure

Analysis Variable : changeresid

N	Mean	Std Dev	Minimum	Maximum
29	-0.8532942	14.8594112	-62.4655791	32.4059712