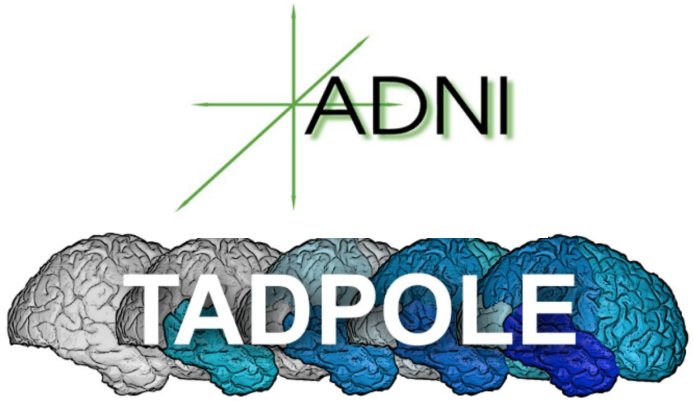
Longitudinal Review of Genomic Biomarkers in Relation to the Progression of Alzheimer's Disease



Computer Science 168 UCLA
Professor Fabien Scalzo

Pratyusha Majumder (UID: 404797253)
Siddharth Satuluru (UID: 404838847)
Saachi Kudtarkar (UID: 604745507)

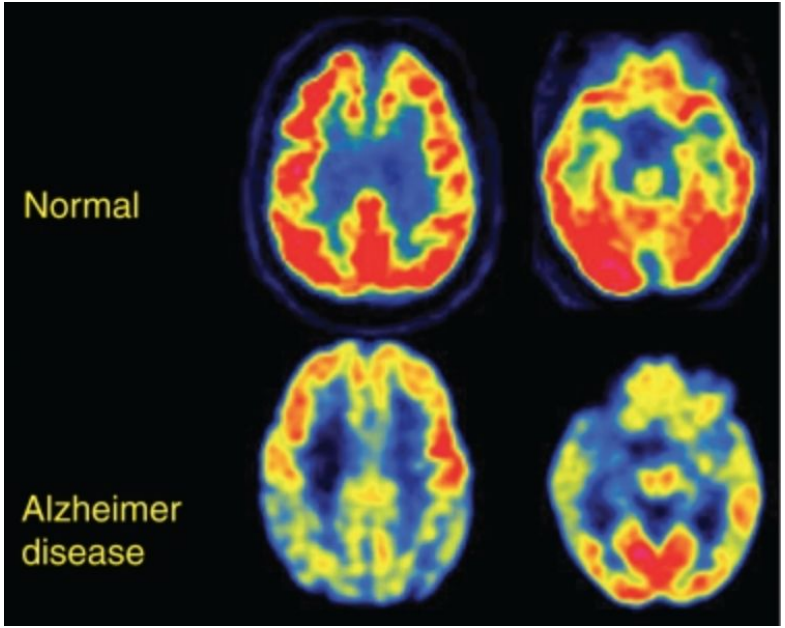
What We're Covering

I. Introduction
II. Methodologies
III. Results
IV. Discussion
V. Conclusion

Introduction

Alzheimer's Disease

- **44 Million** people worldwide are living with AD
  - **5.5 Million** are in the US, where it is the **6th** leading cause of death
- AD is caused by tissue abnormalities called **amyloid plaques** and **neurofibrillary tangles**
  - The amyloid plaques are abnormal deposits of beta-amyloid proteins, and neurofibrillary tangles are abnormal clumps of tau proteins
  - The tau proteins detach from microtubules and clump together, eventually forming tangles inside the neurons and blocking synaptic communication between neurons
- AD is a **neurodegenerative disease** that currently has no cure



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3312396/

Previous Research & ADNI

Current Research:

- **Detecting biomarkers** from Alzheimer's patients, both through the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and through external analysis.
    - A current deep learning approach exists using f-18 PET scans that classifies Alzheimer's disease based on a saliency map
    - Further blood based analysis has been performed to screen for early indication of AD with studies showing evidence for the nerve growth factor precursor protein and amyloid imaging markers as biomarkers for AD

ADNI:

- The Alzheimer's Disease Neuroimaging Initiative (ADNI) has collected large amounts of data related to Alzheimer's disease in elderly people since 2004, on individuals **without Alzheimer's**, individuals with **mild memory impairment**, and individuals with **Alzheimer's or dementia**.

TADPOLE Challenge & Our Project

TADPOLE Challenge:

- The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge aimed to forecast AD indicators for the short to medium term (1-5 years) using the ADNI database
- Participants are asked to find out:
    - If at-risk individuals' progression to AD can be predicted, using the data, processing pipelines and predictive models
    - If these methods can improve cohort selection for clinical trials

Our Project:

We further refined our research goals to answer the following questions:

- Which biomarkers are the most important in determining clinical status?
- How accurately can we classify a patient's clinical status, given specific features?
- How accurately can we forecast the progression of AD in patients?

Methodologies

Overview

*Our goal was to t**rack the progression of the top biomarkers** that contribute to a patient's diagnosis and see if it corresponds to the actual progression that a given patient experiences.*

Steps:

1. Preprocessing Data
2. Feature Selection
3. Building and Selecting a Regression Model
4. Tracking Patient Progression of Select Features
5. Testing Progression Model with Regression Model

Step 1: Preprocessing Data

Dataset Overview:

- Approximately **12,000 entries** and **2,000 features** tracked in the ADNI database.
- Demographic data, cognitive assessment results, extracted features from MRI and PET scans

Dealing with Missing Data:

- Dataset was padded using forward fill
    - Done per patient to ensure that any missing data did not come from a different patient

Making Everything Numerical:

- The following categories were given numerical representations:
    - PTGENDER (gender), PTETHCAT (ethnicity), PTRACCAT (race), PTMARRY (marital status) and DX (diagnosis).
        - **DX_NUMERICAL**: 1 for CN (control) ; 2 for CN to MCI (mild cognitive impairment) or MCI to CN ; 3 for MCI ; 4 for MCI to Dementia or Dementia to MCI ; 5 for Dementia.

Step 2: Feature Selection

The **selectKBest** method from the scikit-learn library was used to select the **20 features** most important in determining the clinical status of a patient.

The f_regression() scoring function was deemed the most suitable as we were testing the target, DX_NUMERICAL, against multiple regressors. This calculated the Pearson correlation coefficient between the features and the target.

The Pearson Correlation Coefficient can be described by the following formula:

$$r = \frac{n\sum xy - \sum x \sum}{\sqrt{(n\sum x^2 - (\sum x)^2)}\sqrt{(n\sum y^2 - (\sum y)^2)}}$$

Step 3: Building and Selecting a Regression Model

The data was split into a training and testing set, with:

- 67% of the data allocated for training
- 33% of the data allocated for testing.

3 different regression models tested : XGBoost Classification, Random Forest Regression, Linear Regression.

XGBoost:

- Gradient Boosting Algorithm, sequentially adds classification trees that account for error of previous model
- Focus on performance, cache awareness, and scalability and uses block structure for parallel learning.
- Minimizes the following loss function:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Step 3: Building and Selecting a Regression Model

Random Forest Regression:

- Supervised learning algorithm which aggregates multiple random decision trees
- As each tree classifies the data, the model classifies based on the majority.
- Contrary to XGBoost that makes use of trees with a smaller number of splits, Random Forest uses trees that are grown to their maximum extent
- Makes a prediction at a new point x using the function shown below:

$$Regression: \hat{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

We chose XGBoost because it had the highest accuracy.

Linear Regression:

- Models the linear relationship between an output variable and a set of input variables
- The 'best fit' for the data is calculated through a process of minimizing the loss.
- Described by the following algorithm:

$$\hat{\theta} = argmin_{\theta} J(\theta) = \|A\theta - y\|^2$$

$$minimizer: \hat{\theta} = (A^T A)^{-1} A^T y$$

Step 4: Tracking Patient Progression of Select Features

Our next goal was to track the progression of AD based on the top 20 features.
Due to variation in data, the data was split into five datasets based on patients' baseline diagnosis:

- Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), Significant Memory Concern (SMC), and Alzheimer's Disease (AD).
- This split was informed by our data visualizations that showed differences in progressions of patients based on baseline diagnosis, which will be shown in the results section

The data was further simplified by using the 'average' patient in each category.

Step 4: Tracking Patient Progression of Select Features

Seasonal Autoregressive Integrated Moving Average with eXogenous regressors model (SARIMAX) was used on each of the 20 features for all five groups.

- Based on the idea that previous data alone can help determine future states
- ‘Auto Regressive’ means it is a linear regression model, and the Y value depends on its own lags.
- ‘Moving average’ means the Y value depends only on the lagged forecast errors
- The two estimates of Y are calculated and combined linearly, alongside an error term, to estimate the final value of Y
- Described using the following formulas:

$$y_t = x_t + u_t$$

$$u_t = \mu + \eta_{t-1} + \zeta_t$$

Step 5: Testing Progression Model with Regression Model

Once predicted values for each of the 20 features for each group were obtained:

- The results of the 20 features were classified using the XGBoost model developed previously, to find the ‘predicted’ diagnosis
- This resulted in ‘predicted’ diagnoses for the ‘average’ patient in each of the five groups
- We compared these ‘predicted’ diagnoses with the actual ‘average’ diagnosis for each group

Results

Step 2 Pearson Coefficient Results

Using PCC, we were able to find the top 20 features:

- **Different neuropsychological assessment scores:**
  - CDRSB, ADAS11, ADAS13, MMSE, RAVLT_Immediate, RAVLT_learning, RAVLT_percent forgetting, Functional Assessment Questionnaire, ECOG_SP_Total
- **Different baseline neuropsychological assessment scores:**
  - FAQ baseline, RAVLT_percent forgetting baseline, RAVLT_learning baseline, RAVLT_Immediate baseline, MMSE baseline, ADAS13 baseline, ADAS11 baseline, CDRSB baseline
- **Genetic Data:**
  - APOE4
- **Data from brain scans:**
  - Hippocampus volume, Entorhinal volume

Step 3 Model Accuracy Results

The XGBoost Classification Model had accuracy of: 88.19%

The Random Forest Regression had accuracy of: 87.66%

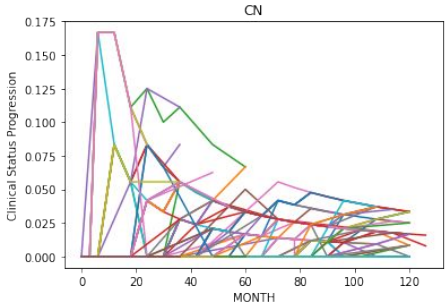
The Linear Regression had accuracy of: 72%

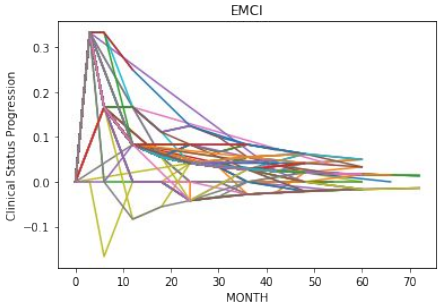
Using this, we selected XGBoost Classification as our classifier of choice.

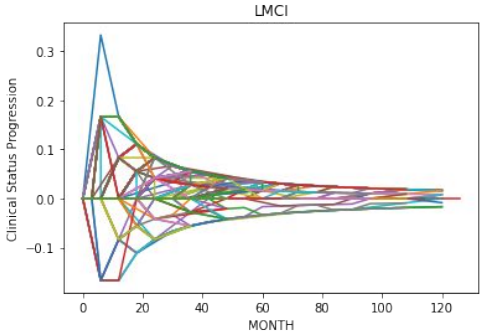
Step 4 Visualizations

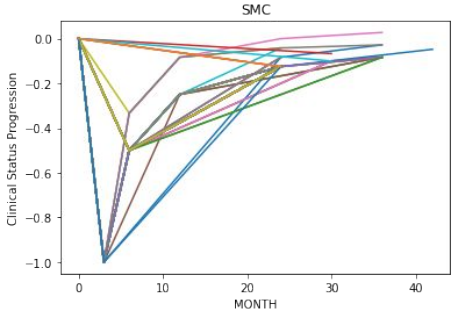
Data visualizations that showed differences in progressions of patients based on baseline diagnosis

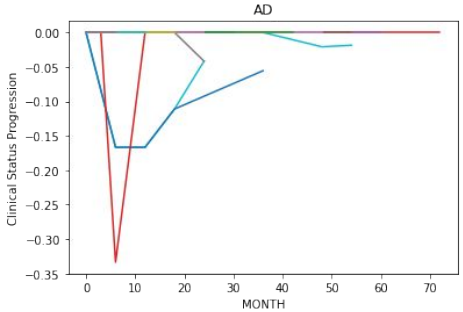
- Informed the decision to split up the data based on baseline diagnosis





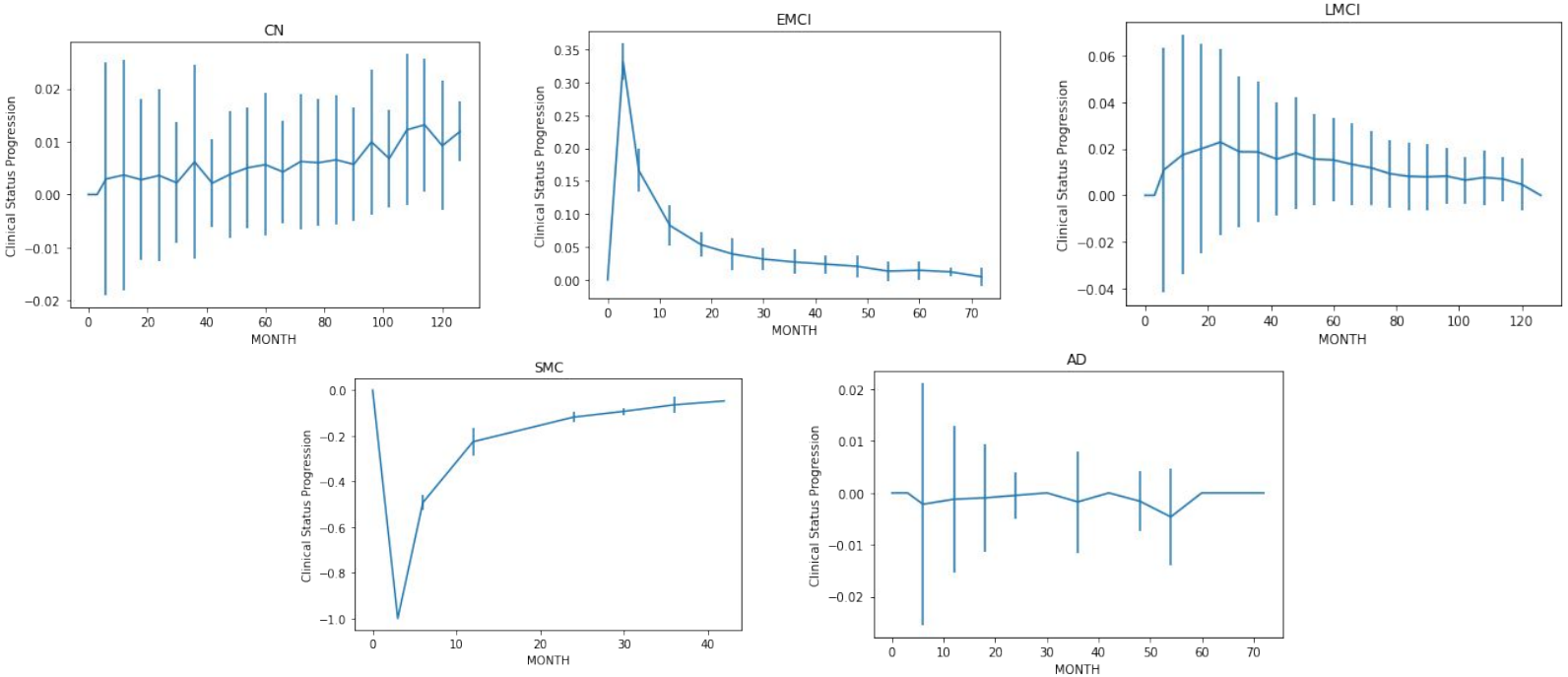






Step 4 Average Patients

The average patient in each of the five groups' progression rates



Step 4 Forecasting Results

The results of each of the progressions on the 20 features for each group are shown in the tables, along with the percent errors.

	Actual	Predicted	Percentage_Error
APOE4	0.212766	0.174990	17.7546
CDRSB	1.648936	1.117618	32.2219
ADAS11	9.553191	9.057013	5.19385
ADAS13	14.978723	14.876513	0.682368
MMSE	27.404255	28.309888	3.30472
RAVLT_immediate	38.829787	33.820302	12.9011
RAVLT_learning	4.276596	4.288399	0.276
RAVLT_perc_forgetting	57.142135	71.567885	25.2454
FAQ	4.787234	2.861832	40.2195
EcogSPTotal	1.653798	1.504630	9.01973
Hippocampus	6774.297872	6343.364704	6.3613
Entorhinal	3486.787234	3093.093844	11.291
CDRSB_bl	0.021277	0.004418	79.2339
ADAS11_bl	5.681064	6.437070	13.3075
ADAS13_bl	8.617234	9.544788	10.7639
MMSE_bl	29.063830	29.098547	0.119451
RAVLT_immediate_bl	45.255319	46.662772	3.11003
RAVLT_learning_bl	6.148936	8.193577	33.2519
RAVLT_perc_forgetting_bl	32.606729	32.588891	0.0547062
FAQ_bl	0.085106	0.060782	28.5808

CN Group

EMCI

	Actual	Predicted	Percentage_Error
APOE4	0.111111	0.403427	263.085
CDRSB	1.000000	1.665895	66.5895
ADAS11	7.444444	6.014643	19.2063
ADAS13	10.444444	12.796458	22.5193
MMSE	28.111111	27.433028	2.41215
RAVLT_immediate	46.555556	36.929746	20.676
RAVLT_learning	3.555556	4.857995	36.6311
RAVLT_perc_forgetting	44.849401	58.564712	30.5808
FAQ	1.333333	3.011111	125.833
EcogSPTotal	1.496552	1.696070	13.3318
Hippocampus	7104.222222	7939.769040	11.7613
Entorhinal	3972.888889	3634.611107	8.51465
CDRSB_bl	1.333333	1.072715	19.5464
ADAS11_bl	6.333333	7.190031	13.5268
ADAS13_bl	9.333333	11.961954	28.1638
MMSE_bl	28.333333	28.280352	0.186992
RAVLT_immediate_bl	43.888889	41.063924	6.43663
RAVLT_learning_bl	6.000000	5.219291	13.0118
RAVLT_perc_forgetting_bl	36.326751	45.570841	25.4471
FAQ_bl	1.000000	1.039348	3.93481

EMCI Group

Step 4 Results (cont.)

	Actual	Predicted	Percentage_Error
APOE4	0.00000	0.412355	100
CDRSB	1.00000	3.244761	224.476
ADAS11	19.00000	16.066208	15.441
ADAS13	30.00000	23.863064	20.4565
MMSE	27.00000	24.503723	9.24547
RAVLT_immediate	13.00000	30.229402	132.534
RAVLT_learning	4.00000	3.366699	15.8325
RAVLT_perc_forgetting	80.00000	77.750376	2.81203
FAQ	2.00000	13.732877	586.644
EcogSPTotal	1.63158	2.643950	62.0484
Hippocampus	7029.00000	5945.023544	15.4215
Entorhinal	3864.00000	3330.948670	13.7953
CDRSB_bl	1.50000	1.286194	14.2537
ADAS11_bl	15.00000	10.481156	30.1256
ADAS13_bl	24.00000	16.765496	30.1438
MMSE_bl	26.00000	27.880323	7.23201
RAVLT_immediate_bl	23.00000	38.407503	66.9891
RAVLT_learning_bl	7.00000	4.175319	40.3526
RAVLT_perc_forgetting_bl	25.00000	56.271835	125.087
FAQ_bl	0.00000	1.747152	100

LMCI Group

SMC

	Actual	Predicted	Percentage_Error
APOE4	1.0	0.379037	62.0963
CDRSB	3.5	0.569568	83.7266
ADAS11	12.0	7.208931	39.9256
ADAS13	18.0	11.652043	35.2664
MMSE	28.0	28.109164	0.389871
RAVLT_immediate	37.0	41.416608	11.9368
RAVLT_learning	5.0	4.896962	2.06075
RAVLT_perc_forgetting	100.0	36.969855	63.0301
FAQ	6.0	2.438051	59.3658
EcogSPTotal	2.0	1.476670	26.1665
Hippocampus	5059.0	7528.533928	48.8147
Entorhinal	2731.0	3531.704717	29.3191
CDRSB_bl	0.5	0.038835	92.2331
ADAS11_bl	8.0	8.227157	2.83947
ADAS13_bl	12.0	12.947299	7.89416
MMSE_bl	30.0	29.474214	1.75262
RAVLT_immediate_bl	41.0	37.795085	7.81687
RAVLT_learning_bl	7.0	4.937673	29.4618
RAVLT_perc_forgetting_bl	100.0	60.267517	39.7325
FAQ_bl	1.0	0.464568	53.5432

SMC Group

AD

	Actual	Predicted	Percentage_Error
APOE4	1.00000	0.935733	6.42668
CDRSB	11.00000	6.073737	44.7842
ADAS11	20.33000	26.662411	31.1481
ADAS13	32.67000	35.086129	7.39556
MMSE	16.00000	18.623604	16.3975
RAVLT_immediate	12.00000	15.304186	27.5349
RAVLT_learning	3.00000	1.953666	34.8778
RAVLT_perc_forgetting	100.00000	98.942425	1.05758
FAQ	28.00000	14.612878	47.8111
EcogSPTotal	3.97436	2.993198	24.6873
Hippocampus	4349.00000	5675.658885	30.5049
Entorhinal	2332.00000	2567.014903	10.0778
CDRSB_bl	3.00000	3.096457	3.21524
ADAS11_bl	16.67000	19.066602	14.3767
ADAS13_bl	29.67000	28.571204	3.70339
MMSE_bl	22.00000	23.293926	5.88148
RAVLT_immediate_bl	14.00000	17.664123	26.1723
RAVLT_learning_bl	1.00000	3.569706	256.971
RAVLT_perc_forgetting_bl	100.00000	71.791965	28.208
FAQ_bl	12.00000	-5.661163	147.176

AD Group

Step 5 Results

The results of the 20 features were classified using the XGBoost model developed previously, to find the ‘predicted’ diagnosis

The table shows the comparison of these ‘predicted’ diagnoses with the actual ‘average’ diagnosis for each group

	Predicted Diagnosis	Real Diagnosis
Clinical Status	3	2.5

Early Mild Cognitive Impairment

	Predicted Diagnosis	Real Diagnosis
Clinical Status	3	2.333333

Late Mild Cognitive Impairment

	Predicted Diagnosis	Real Diagnosis
Clinical Status	3	3.0

Severe Memory Concern

	Predicted Diagnosis	Real Diagnosis
Clinical Status	1	2.0

Alzheimer's Disease Positive

	Predicted Diagnosis	Real Diagnosis
Clinical Status	5	5.0

Discussion

Things to Improve

- **XGBoost classifier resulted in integer-valued predicted diagnoses**. Thus, there was some inherent error when comparing the actual and predicted diagnosis of the patients
    - Even accounting for this error, all predicted values were within one integer value, signifying that the same clinical status would be determined
- **We decided to use only the top 20 features in our model**
    - One can optimize by setting the number of features as a hyperparameter
- We tested multiple models based on classification trees or linear regression, but encountered **long run times and lack of accuracy**
    - XGBoost was the most suitable as it accounted for the long run times of other gradient boosting algorithms and had the highest accuracy of the models we tested
- For time series forecasting, we aimed to predict the average patient overall, as this would allow us to aid in the overall patient treatment. However, there was a **strong variation in data between patients**
    - To account for these differences, the population was categorized based on their baseline diagnosis, which was a decision motivated by both the data visualizations and the knowledge that Alzheimer's Disease progresses more rapidly in later stages

Things to Improve

- Although our predictions and XGBoost model had high accuracy, it was **simplified as we only used data visualizations to determine the progression was different for patients who started with different baseline diagnoses**
    - One could either more strongly prove the distinction between segmentations or adapt the model to allow for all individual patients to be represented, rather than create the average patient in a given group
- The **errors in our final predictions may not be fully attributed to an incorrect SARIMAX forecast**, and could be influenced by the error in our XGBoost model itself
    - One could separate the two so that the accuracy of the forecasting can be more correctly determined
- Simplifying the model to use the ‘average’ patient for each category resulted in **less data for forecasting purposes**. This resulted in higher errors in the forecasting for those groups, compared to the other groups with more time series data points
    - This can be mitigated by using an estimation algorithm to determine the progression of each feature in the years between the assessments. Currently, the data is collected in 3 year intervals, so the 1- and 2- year marks can be estimated and filled in to create more data

Conclusion

Reflections

We were able to successfully build regression models that determine a patient's diagnosis and forecasting models that predict the patient's diagnosis through progressing key features.

Significance:
Present methods of diagnosing AD progression categorize patients into five stages, primarily based on neurophysiological assessments. These assessments often use different thresholds to classify patients

- Our regression models can simplify this process as they use an aggregate of the 20 most relevant features to predict the patient's stage of AD

Furthermore, our progression models were able to track the progression of select features that determine patients' cognitive state. This can help clinicians:

- Determine a patient's subsequent cognitive states based on their current cognitive assessments, neuroimaging results and genetic factors
- Estimate the rate at which a patient may progress given their current state
- Administer individualized treatment at an earlier stage, prolonging cognitive decline in patients with AD

Thank you!