# Deep learning methods for Alzheimer's disease diagnosis and progression prediction via neuroimaging-genomics integration

Ahmad Abdel-Azim[1,2], Tong Ding[1,2], and Charles Van De Mark[2]

[1] Harvard University, Cambridge, United States
[2] Massachusetts Institute of Technology, Cambridge, United States

**Abstract.** Alzheimer's disease (AD) remains the most pervasive neurodegenerative disorder worldwide, afflicting over 45 million people and imposing an increasing burden on modern health care systems. Current methods of diagnosing AD include analyzing patient medical history, neuropsychological tests, and MRI data. Genetic analysis may provide predictive capability for disease risk. However, due to current gaps in biological understanding of AD and unpredictable disease progression, diagnosis and predicting disease progression remains challenging. Here, we demonstrate that integrating critical insights from genetic and neuroimaging data may facilitate rapid, precise diagnosis and predictions of progression of Alzheimer's disease (AD). We integrate multimodal data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) before and after disease onset to identify specific features that quantify disease progression. We separately employ machine learning (XGBoost) to select important genomic features and employ 3D convolutional neural network (CNN) to build models for imaging (MRI) data. We then combine retained features from both models with clinical information and use a classifier to predict diagnosis and disease progression. We sequentially exclude clinical data, imaging features, and genomic features to quantify the impact of each data modality. To further quantify the predictive capability of these features on disease progression, we attempt prediction at progressive time periods for up to three disease outcomes: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Both F1-score and accuracy score are used to evaluate model performance. When distinguishing between CN and AD patients, our combined model attains a predictive diagnosis accuracy of 99.9% up to 24 months after baseline data input.

**Keywords:** deep learning · Alzheimer's disease (AD) · 3D convolutional neural network (CNN) · genetics · genome-wide association study (GWAS) · genomics · machine learning · multi-Layer perceptron · neurodegeneration · structural MRI · XGBoost

## 1 Introduction

### 1.1 Motivation

Alzheimer's disease (AD) is a heterogeneous neurodegenerative disorder of older age distinguished by $\beta$-amyloid (A$\beta$)-containing extracellular plaques and tau-containing intracellular neurofibrillary tangles (Knopman, et al., 2021). As one of the most frequent neurodegenerative diseases in older individuals above 65, it affects more than 45 million people worldwide, making it an immense economic and psychological burden to society (Masters, et al., 2015). Amnestic cognitive impairment is a common manifestation of AD as well as short-term memory difficulty and the impairment of expressive speech, visuospatial processing, and mental agility. Insufficient understanding of AD pathobiology, and the heterogeneous nature of disease onset and progression make AD especially difficult to diagnose and treat (De Jager, et al., 2018); thus, there is an unmet clinical need for effective AD diagnosis and treatment.

Great progress has been made in detecting and diagnosing AD; PET amyloid (Bohnen, et al., 2012), cerebrospinal fluid biomarkers (Harper, et al., 2014), and tau imaging (Mattsson, et al., 2019) have dramatically improved the sensitivity and specificity of AD diagnosis. However, conventional diagnosis still relies heavily on neurologist experience via clinical and psychometric assessments, such as neuropsychological testing (McKhann, et al., 2011) and the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), inquiry of patient history, and structural MRI (Frisoni, et al., 2010) analysis. Such diagnostic strategies are often adopted as novel modalities and remain limited to research contexts. As revealed by clinicopathological studies, the diagnostic specificity achieved by neurologists ranges between 44.3% and 70.8%, and the sensitivity ranges between 70.9% and 87.3% (Beach, et al., 2012). Further, although characteristic cerebral changes

such as parietal lobe and hippocampal atrophy which are noted in AD can be revealed by structural MRI, it is believed that such imaging-based AD diagnosis is insufficient due to lack of specificity (Frisoni, et al., 2010). Sensitive and specific diagnosis of Alzheimer's disease thus remains a persistent challenge, especially with the paucity of skilled neurologists and the invasive nature of PET and CSF diagnosis.

### 1.2 Prior Work

Previous studies have attempted to draw biological and diagnostic insights related to AD from multi-omics or neuroimaging data. Nativio, et al. (2020) integrated transcriptomic, proteomic, and epigenomic analyses to identify specific chromatin modifications involved with disease pathways in AD; their findings emphasize the power of a multi-omics approach in revealing key features related to AD progression. The ability of genomic information to predict neuroimaging features has also been demonstrated.Zhu et al. (2016) used spatial and linear regression with a GWAS to evaluate the importance of case control sampling when using a GWAS to analyze secondary imaging phenotypes associated with AD. Further, deep learning methods are beginning to show remarkable promise for AD diagnosis. Qiu, et al. (2020) leveraged neuroimaging and clinical data to develop a deep learning framework for the accurate diagnosis of AD. They trained their models on AD and cognitively normal subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (n=417) and validated their approach across other available datasets. Their diagnostic model was successful and more accurate than a team of 11 practicing neurologists, highlighting the benefit of using deep learning approaches for the prognosis and diagnosis of AD. While not specific to AD, Bari et al. (2021) used permutation-based moderation analysis and three way associations in order to combine mutli-omics data with neuroimaging data to study neurological dysfunction in athletes. However, they only used regression models (no deep learning techniques) with a limited sample size (n=23) to investigate specific biological pathways.

## 2 Overview

We extend the work of Qui et al. (2020) in particular to form the basis for a more complex deep learning framework. In their focus on immediate diagnostics, Qui, et al. (2020) did not integrate any omics data into their models, nor did they attempt predicting disease progression. While an individual's state or progression toward the underlying disease phenotype can be reflected in imaging, genomics features contribute to explaining propensity of disease progression and outcome (Bonham, et al., 2019). Thus, the integration of both multi-omics and neuroimaging data in conjunction with genomic data may be critical to AD diagnosis and disease forecasting. Deep learning methods that leverage both genomic and neuroimaging data may facilitate rapid, precise diagnosis of Alzheimer's disease (AD), help predict disease outcome, and may thereby lead to novel therapeutic strategies for effective treatment of individuals with AD. A brief overview of our approach to this problem will be presented here before being elaborated on in more detail in Section 3.

All genomic, MRI, and clinical data was acquired through the Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI1 cohort was used throughout this paper. Genomic data for this cohort included single nucleotide polymorphisms (SNPs), insertion-deletions (indels), and structural variants (SVs). Neuroimaging data included standard 1.5T structural MRI scans. Patients without a baseline scan were excluded. Clinical data inclues age, gender, and Mini-Mental State Exam (MMSE). To utilize this data, important information is extracted separately from patients' genetic files and neuroimaging files. For genomic information, once preprocessing and quality control is performed on raw data, machine learning (XGBoost) is used to select and output an array of critical genomic features. For neuroimaging files, representations are learned by a 3D CNN from preprocessed MRI. Together with clinical data, the selected features from the genomic and neuroimaging stages are fed into a classifier block using either a multi-layer perceptron (MLP) or extreme gradient tree boosting (XGBoost), which informs the selection of 3 diagnostic states: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzhiemer's Disease (AD). We also evaluate model performance in the case of only 2 diagnostic states: CN and AD. As our focus is on identifying specific diagnostically-relevant features that allow for the quantification of disease progression, predicted disease state is compared to known patient disease state at baseline (original diagnosis), 6 months, 12 months, 18 months, and 24 months. The sample sizes for each time period are as follows:

| Baseline | 6 months | 12 months | 18 months | 24 months |
|---|---|---|---|---|
| 756 | 720 | 676 | 304 | 602 |

The variation in sample size is due to patient availability as the study progressed. Predictions for each time period use the baseline MRI data, constant genetic data, and baseline MMSE scores. This basic project workflow is shown in Figure. 1. We then build models that sequentially exclude each of neuroimaging features, genomic features, and clinical data (seen in Figure 1A, 1B, and 1C, respectively) to quantify the impact of each on the performance of the model. Both accuracy and F1-score that considers both recall of a test and precision, and Matthew's correlation are used to evaluate model performance.

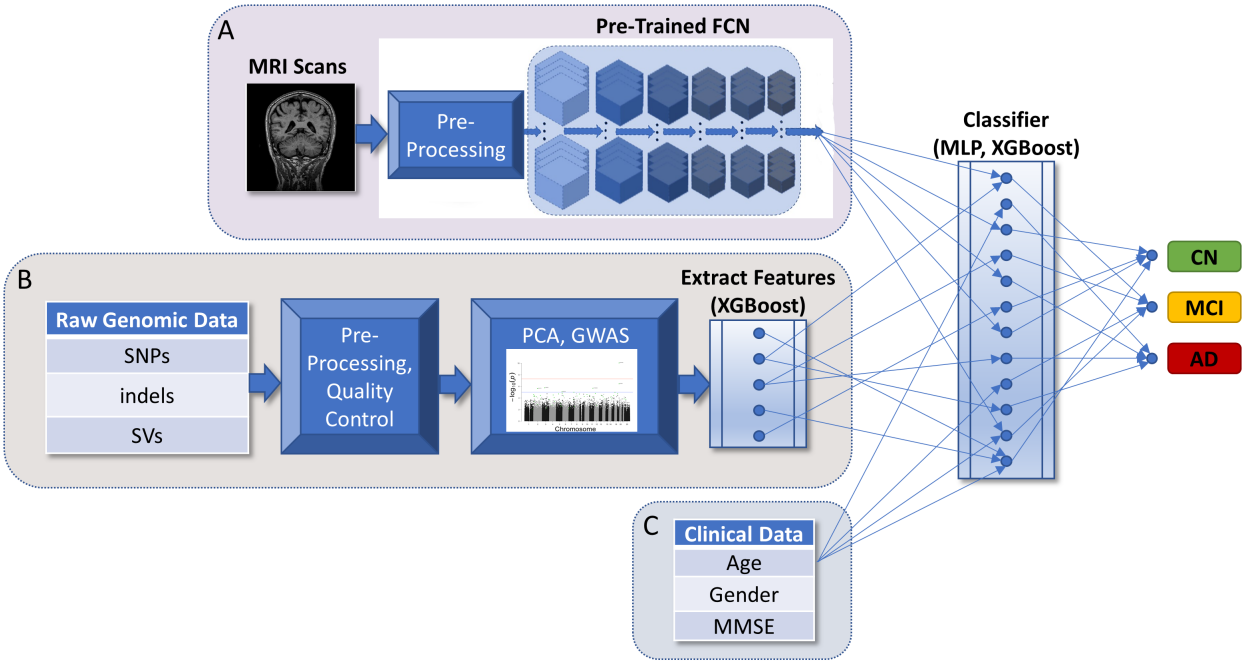



***Fig. 1:*** *System diagram displaying integration of image features (A), genomic features (B), and clinical data (C). Features are input into a classifier (either XGBoost or a multi-layer perceptron) which predicts diagnosis after the specified time period (either at baseline, 6 months, 12 months, 18 months, or 24 months). Diagnosis includes cognitively normal patients (CN), patients with mild cognitive impairment (MCI), and patients with Alzheimer's disease (AD).*

## 3 Methods

### 3.1 Genomics Data Processing

**Preprocessing** Genetic data from the ADNI database, ADNI1 cohort was used in this study. While raw VCF files containing over 3.5 million SNPs are available for all patients, genotype data from the Human610-Quad BeadChip (Illumina, Inc., San Diego, CA) was used here, including 620,901 SNP and CNV markers across the ADNI-1 cohort (Saykin et al., 2020). Patients in this cohort have been diagnosed with NC, MCI, or AD status. Genomic data pre-processing was primarily performed using PLINK (Purcell et al., 2007 and Purcell 2015). Filtering and quality-control measures were implemented to exclude SNPs with minor allele frequency below 0.01, high per-site missing rate, and significant Hardy–Weinberg equilibrium p-value. After all filtering 570,184 SNPs remained and were used in downstream GWA studies.

**GWAS and Feature Selection** We perform feature extraction for genetic data by first running a GWAS using a training subset (80%) of the internal cohort (n = 605 subjects). Note that the results from GWAS

were leveraged for a genomic classification model, so it was important to have a held-out test set. Note that all GWAS regressions were carried out using the R language. A linear model was fit to each SNP variant while adjusting for age, sex, and population stratification. The following model was fit for each variant,

$$Y_i = \beta_0 + \beta_1 G_i + C_i \beta^{COV} + \epsilon_i$$

where $Y_i$ is the phenotype of the $i$th individual, $\beta_0$ is the intercept, $\beta_1$ is the SNP effect, $G_i$ is the genotype for the $i$th individual, $C_i$ is a matrix of additional covariates, including age, gender, and the top 5 genotype principal components, and $\beta^{COV}$ is a vector of the covariate estimates. The same model is fit repeatedly for every variant; p-values and estimates are retained from each model, and multiple-testing correction is performed. Note that PLINK was used to extract the PCs from the genotype matrix, and the top 5 PCs were fit as covariates in the GWAS regressions to help correct for population stratification. PCs were computed after filtering out highly linked variants, resulting in a pruned subset of 479,353 markers that are in approximate linkage equilibrium.

Disease phenotype was recorded at baseline (0 months), 6 months, 12 months, 18 months, and 24 months. The aforementioned GWAS was run 5 times across all phenotypes present across time, and the results were considered for feature selection.

To extract a subset of impactful features from the results of the GWAS, Minimum-Redundancy and Maximum-Relevance (MRMR) was used to select a subset of 50 critical SNPs for phenotype classification on the test set. MRMR (Ding and Peng, 2005) is a method that has garnered traction in the field of genomics as it selects a subset of features with the maximum association with an outcome (i.e. maximum relevance) and minimum correlation between themselves (i.e. minimum redundancy). MRMR features were selected from the top 5,500 variants (ranked by p-value) identified by GWAS. MRMR selected features were found to be largely consistent across all GWAS runs (using the phenotype at different study times); since the 24 month GWAS represents the furthest disease progression captured, the 50 MRMR features extracted from this GWAS run were utilized as the genomic features for subsequent classification and prediction in the test set.

**Genomic Classifier** Several classification algorithms were tested to achieve the highest classification accuracy; extreme gradient tree boosting (XGBoost) models proved to be the most successful in predicting phenotype from genomic features. Rooted in the gradient-boosted decision trees, XGBoost can account for non-linear feature interactions to predict outcomes non-additively (Chen and Guestrin, 2016). An XGBoost model was trained on samples reserved for training (previously included in GWAS cohort) using only the 50 selected MRMR features, followed by hyperparameter tuning. Genomic prediction accuracy was then measured by running this model on the test set; namely, the phenotype of each individual in the test set was predicted by inputting their 50 genotypes into the trained XGBoost model.

**Genome-wide Analysis of Longitudinal Outcomes** While the aforementioned GWAS models provide variants associated with AD staging or onset which is undoubtedly diagnostically relevant, more robust approaches were applied to explicitly capture the genomic patterns associated with AD progression. To capture both cross-sectional genotype effects, namely those associated with disease state at a single time point, and longitudinal genotype effects, namely those associated with disease progression over time, an additional modified GWAS regression was implemented. Note that this model experienced extreme overfitting and was thus not implemented into the multi-omics model integration; however, with additional regularization or penalization strategies, this model may be useful for the identification of variants that generally impact the rate of disease progression.

A linear model was fit to each SNP variant with 5-fold cross-validation (to attempt to decrease overfitting), this time accounting for the interaction of genotype and time. The following model was fit for each variant,

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 t_i + \beta_3 G_i t_i + C_i \beta^{COV} + \epsilon_i$$

where $Y_i$ is the phenotype of the $i$th individual, $\beta_0$ is the intercept, $\beta_1$ is the cross-sectional SNP genotype effect, $G_i$ is the genotype for the $i$th individual, $\beta_2$ is the slope on time, $t_i \in \{0, 6, 12, 18, 24 \text{ months}\}$ is a vector with measurement occasions, $\beta_3$ is the longitudinal SNP effect on time, $C_i$ is a matrix of additional

covariates, including age, gender, and the top 5 genotype principal components, and $\beta^{COV}$ is a vector of the covariate estimates. Similar algorithms have proposed similar methods of longitudinal GWAS analyses, such as the GALLOP algorithm (Sikorska et al., 2018).

The cross-sectional and longitudinal effect sizes are useful here as we can assign aggregate cross-section and longitudinal scores for each subject. Multiplying the genotype matrix $G$ by the longitudinal effect sizes $\beta_3$ and some time $t_i$, and summing across all SNPs, we get an aggregate longitudinal score, where higher scores indicate a higher rate of disease progression. Likewise, multiplying the genotype matrix $G$ by the cross-sectional effect sizes $\beta_1$ and summing across all SNPs, we get an aggregate cross-sectional score, where higher scores correspond to more severe disease (AD diagnosis).

### 3.2 Neuroimaging Processing

**Pre-processing** Prior to training, we performed imaging registration, intensity normalization, and background removal. We downloaded MRI scans from ADNI in the NIFTI file format. The MNI152 template (ICBM 2009c Nonlinear Symmetric template, McGill University, Canada) is used for image registration. Next, the FLIRT tool in the FSL package (Wellcome Center, University of Oxford, UK) is used for scan alignment to the MNI152 template. Intensity normalization is performed to force all voxels to have mean 0 and unit standard deviation, followed by clipping the voxels to the range of [-1, 2.5] to adjust the intensity and other outliers; intensity of all voxels is thereby capped at -1 and 2.5. In addition, to ensure background intensity, background removal is performed, which sets all voxels outside the skull in the background regions to -1.

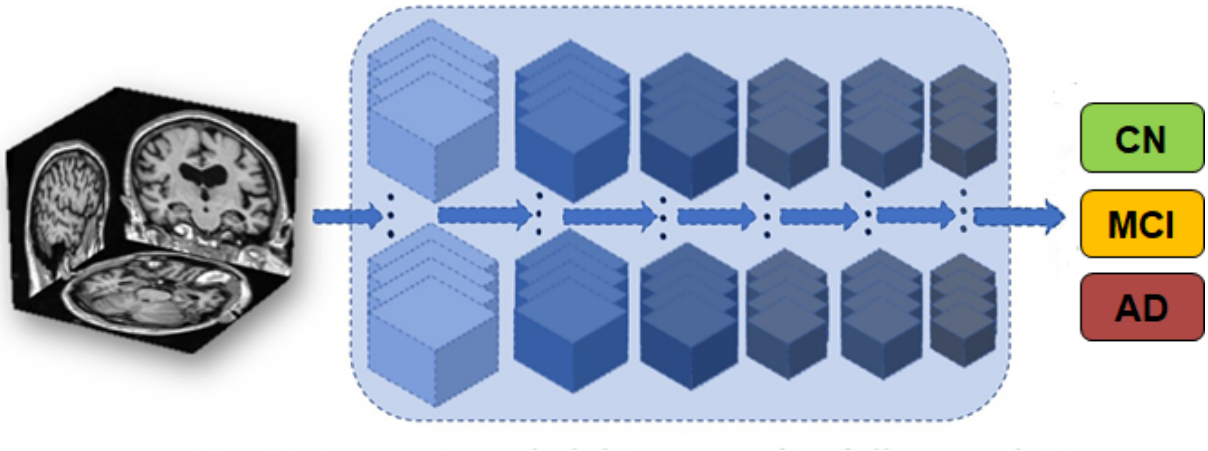



***Fig. 2:*** *Adapted from Qiu et al. (2020) Figure 1. Initial training of the FCN used by Qiu et al. (2020). The FCN was trained on a subsample of the ADNI1 cohort used throughout this project (excluding certain sources of possible causal conflict such as traumatic brain injury, stroke, etc., as well as those patients aged less than 55 years). The MRI scan dataspace was divided into 47 x 47 x 47 voxels; random voxels were sampled to train the FCN.*

**3D Convolutional Neural Network** We are interested in using a 3D CNN to predict subjects' disease status from their first visit to 2 years using their MRI scans from their first visit (baseline). Our outcome of interest has 3 classes, which are AD, MCI, and CN. In addition, we also train a 3D CNN to predict subjects' disease progression for the same time interval but focusing on AD and CN only, where MCI patients are excluded, for comparison.

We utilize a pre-trained 3D CNN model provided by Qiu et al. (2020). The CNN model consists of 4 convolutional layers and 2 dense layers, where each 3D convolutional layer is followed by 3D max pooling, 3D batch-normalization, Leaky Relu activation, and Dropout. Prediction is performed by the last two dense layers, where a softmax function is applied.

Since we are interested in predicting subjects' outcome in 3 classes (AD, MCI and CN) whereas Qiu et al. only focus on AD and CN, we re-design the last dense layer so it has 3 outputs, each corresponding to the probability of each status. Classification is determined by the status with largest probability.

Instead of training from scratch, we perform fine-tuning. For the model predicting only AD and CN, we used Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and initial leaning rate of 1e-4 as the optimizer for fast convergence (given the task is easy for our model). This model is trained for 30 epochs. For the model predicting AD, MCI, and CN, we use stochastic gradient descent (SGD) with a learning rate of 1e-4, momentum=0.9 for more stable convergence. Since including MCI makes the task more difficult to learn, we train for 90 epochs. For both models, we use a mini-batch size of 8. A validation set is used to choose the epoch that has the best performance to avoid over-fitting.

In order for multi-modal data integration, we generate 30 features as representations of each MRI which are the output of the first dense layer of the 3D CNN.

### 3.3 Integration

We aim to integrate the imaging, genomic, and clinical features to achieve more accurate prediction of AD severity and report hidden patterns. The 30 imaging features extracted from 3D CNN, 50 MRMR SNPs selected as described above, and 3 clinical features, which are age, gender and MMSE assessment scores retrieved from subjects' electronic health record (EHR), are integrated. Different combinations of data integration were also tested including clinical with imaging, clinical with genomics, imaging with genomics, and clinical with imaging and genomics. Since XGBoost performed better with genomic data, we implemented XGBoost in the clinical-genomic integration. For other integrated models, a MLP is used as the classifier, where we use 1 hidden layer with 80 neurons, and SGD with learning rate 1e-4 and momentum of 0.9 as optimizer.

## 4 Results

### 4.1 Overfitting observed for genome-wide analysis of longitudinal outcomes

Aggregate longitudinal and cross-sectional scores were calculated across all subjects using the genotype matrix and the results from a GWAS which accounts for the interaction of genotype and time. Higher longitudinal scores indicate a higher rate of disease progression, whereas higher cross-sectional scores correspond to more severe disease (AD diagnosis).

Plotting aggregate cross-sectional and longitudinal scores against each other reveals interesting distinct clusters in the data (Figure 3). First, coloring by baseline status, we see that the cross-sectional aggregate successfully classifies the 3 initial phenotypes (diagnoses) well. Further, those subjects with highly positive longitudinal scores progress from MCI to AD status in the 24-month status; this longitudinal model is not only able to classify phenotypes but even predict faster or slower progressing individuals in the cohort.

However, now considering the test set (colored in red in Figure 3C), which was excluded from the GWAS run, we do not see such distinguishable clusters. While this aggregation method is successful at in-sample cluster separation, it is not able to successfully separate test cases, previously not seen by the model. In fact, using these scores as inputs into the genomic XGBoost model increases in-sample accuracy to 99.2%; however, out-of-sample accuracy drops dramatically to approximately 30%.

The success of this approach in-sample and failure out of sample suggests rampant overfitting, which was a persistent concern throughout this study. Regularization and penalized regression strategies can be applied to reduce this overfitting; however, in this study, these aggregate features were excluded from the multi-omics due to their poor performance on the test set.

### 4.2 Classification of disease status based on genomics

Results from the GWAS run using the phenotype recorded 24 months after inclusion into the study are included in Figure 4. The importance of fitting genotype PCs to account for population stratification is emphasized by Figure 4A; the first two PCs are able to accurately capture the different ethnicities represented

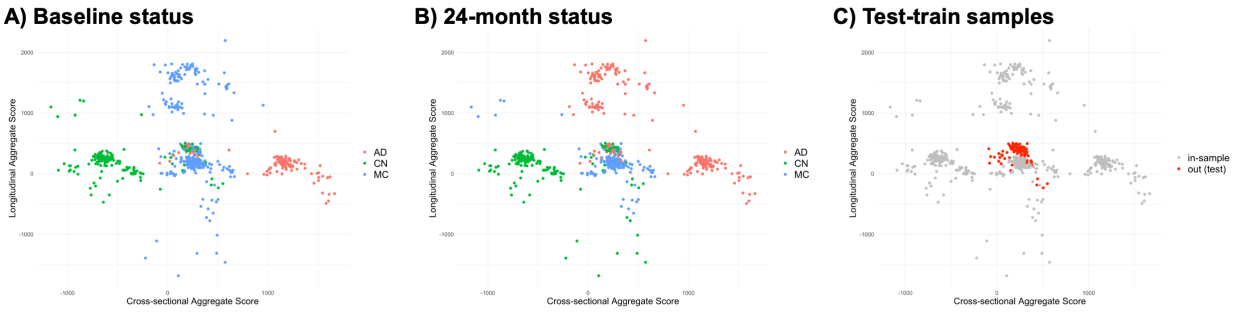



***Fig. 3:*** ***A)*** *Aggregate longitudinal and cross-sectional score calculated for each subject and colored by baseline diagnosis.* ***B)*** *Aggregate longitudinal and cross-sectional score calculated for each subject and colored by 24-month phenotype. Subjects with higher longitudinal scores seem to have progressed, while those with negative longitudinal scores seems to have experienced disease regression.* ***C)*** *Aggregate longitudinal and cross-sectional score calculated for each subject and colored by train/test status. Extreme overfitting is evident here since test cases (red) were not successfully assigned to distinguishable clusters as seen with the training subjects (gray).*

in the ADNI1 cohort, and this variation was accounted for in the GWAS regressions. Panels B and D in Figure 4 include the retained SNP effect sizes and association p-values from the GWAS regression runs. Plotting effect size against p-value reveals that generally SNPs with higher effect sizes also have lower p-values. Further, the 50 MRMR features are colored as green; that these features have high effect sizes and low p-values suggests that the selected MRMR SNPs should optimally predict disease phenotype.

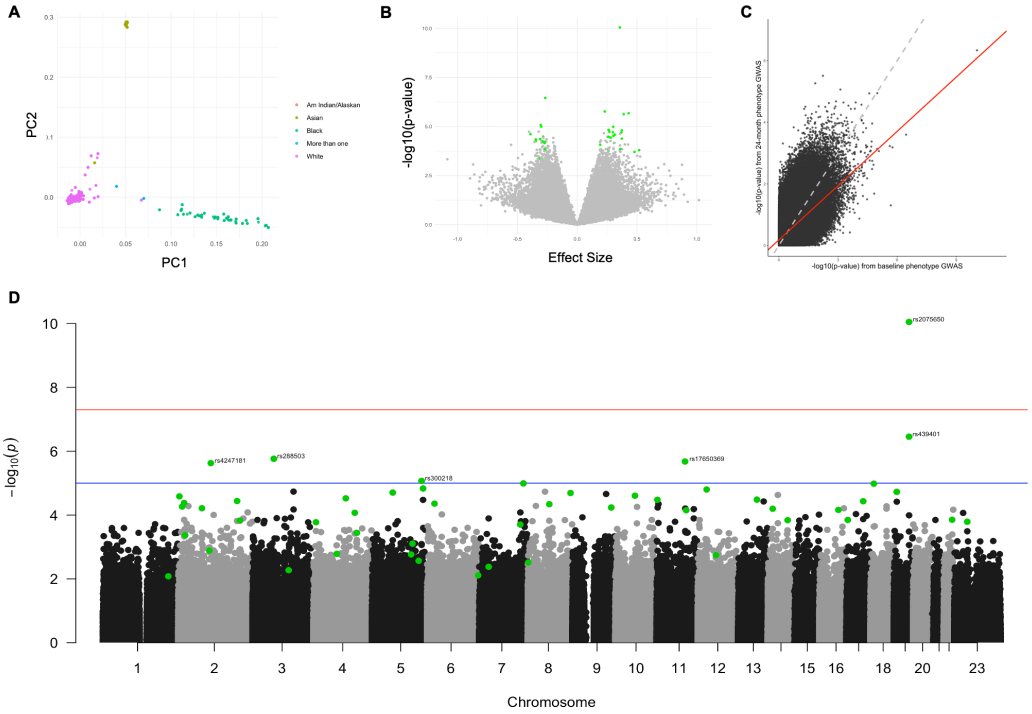



***Fig. 4:*** ***A)*** *PCA plot of patient genotypes projected onto first two PCs, showing clear population stratification based on ethnicity.* ***B)*** *Plot of SNP effect size versus p-values from 24-month GWAS regression. SNPs selected as MRMR features are highlighted in green.* ***C)*** *Plot of SNP p-value from baseline GWAS regression against SNP p-value from 24-month GWAS regression. A slope significantly less than 1 (p-value ¡¡ 0.001) suggests that SNP-phenotype associations become more significant as AD progresses.* ***D)*** *Manhattan plot from 24-month GWAS regression. Red line represents* $5 \times 10^{-8}$ *threshold, and blue line represents* $1 \times 10^{-5}$ *threshold.*

Interestingly, corresponding SNPs had more significant SNP-disease status associations in GWAS runs using later recorded phenotypes; in other words, SNP-phenotype associations became more significant as the disease progressed. Plotting the p-values from the baseline GWAS (using baseline phenotypes) against the p-values from the 24-month GWAS (using phenotypes after 24-months) results in a slope significantly less than 1 (p-value ¡¡ 0.001), again suggesting that SNP-phenotype associations became more significant as the disease progresses (Figure 4).

The 50 selected genomic features were then used for classification of the test set. Accuracy was measured first with genomic features alone and then after explicitly adding clinical features for each patient (age, gender, MMSE score). XGBoost model performed better than MLP for genomic prediction alone, so accuracies from the XGBoost classifier are reported here. While test accuracy based on genetic and clinical features remained fairly constant across all time points (mean accuracy of 71.5%), test accuracy based on genomic features alone increased for phenotypes recorded from later times (Figure 6).

### 4.3 Classification of disease status based on neuroimaging

Accuracy and f1 score of the two 3D CNN models in predicting subjects' status in 2 classes and in 3 classes from their first visit to two years are shown in Figure 5. We find that the CNN does well in predicting disease status in 2 classes when MCI patients are removed, and the accuracy is constant within the 2-year time interval, meaning that the CNN can effectively discriminate AD from CN. Rather, after including MCI patients, we see that it is difficult for the CNN to discriminate MCI from the other two statuses, resulting in low and unstable accuraccies over time. One thing worth noting is that accuracy for 18 months is higher than other time. The reason for this is that there are much fewer samples in the 18 months cohort, and the cohort is also imbalanced. There are only 304 subjects in the 18 months cohort, among which there are only 10 CN patients. Thus, accuracy achieved for 18 months is not persuasive.

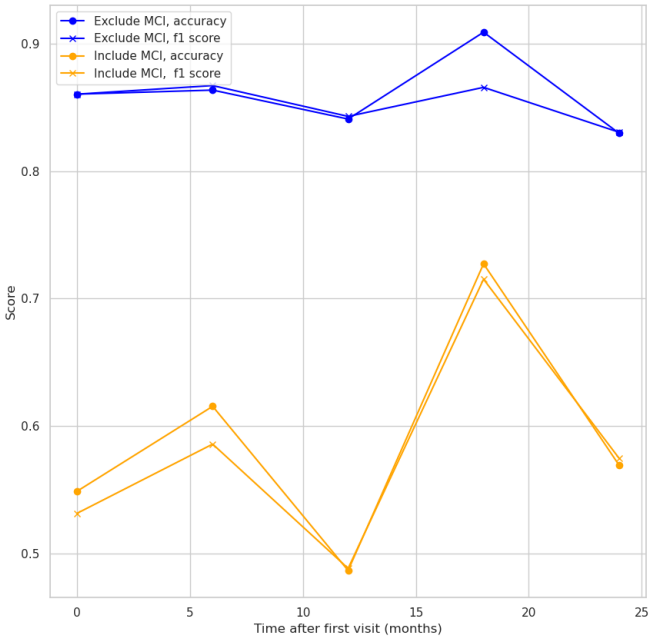



***Fig. 5:*** *Accuracy and f1 score of the two CNN models predicting subjects' disease status in 2 classes and in 3 classes from their first visit to 2 years.*

### 4.4 Classification of disease status based on multi-modal data integration

We try different combinations of data integration using MRI, genomic, and clincal features. The accuracy score of each combination is shown in Figure. 6. The results depart significantly from what we expect to see. From the results, we see that model with clinical features only and model with clinical and genomic features generally perform the best over all models.

The confusion matrix of prediction of disease status made by the multi-modal model is shown in Figure. 7. We see that low accuracy of the multi-modal model is primarily due to its lack of ability in discriminating MCI from AD and CN. Note that the model almost never makes a mistake when discriminating AD from CN, with the exception of a single case in the 24 month time period. This means that the accuracy of our model is nearly 100% .

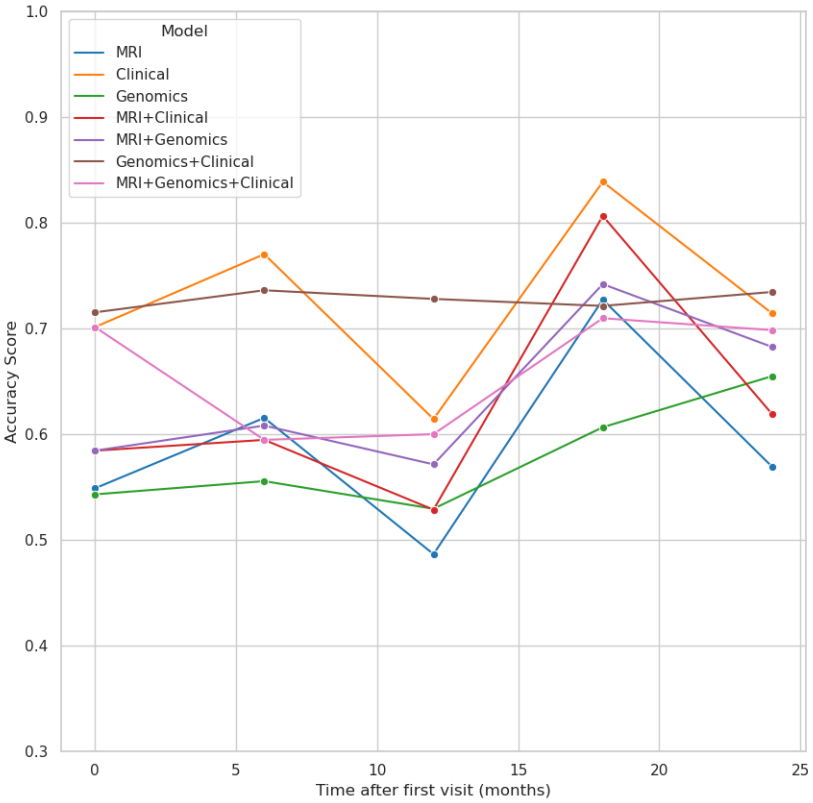
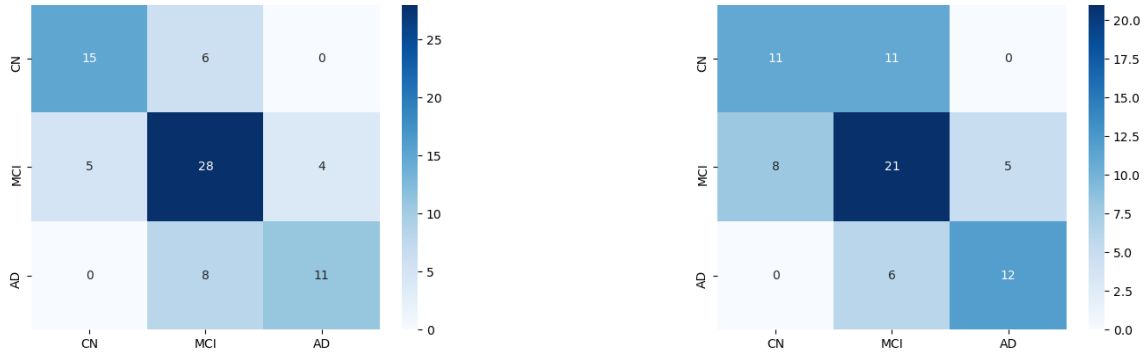



***Fig. 6:*** *Accuracy score of different combinations of different multi-modal features in predicting subjects' disease status in 3 classes from their first visit to 2 years.*

## 5 Conclusion

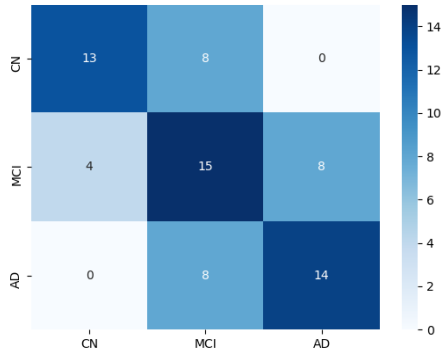
Our results suggest that clinical information and a combination of clinical information and genomic features best discriminate MCI from AD and CN; however, it is important to note that accuracy is model dependent and variable. It is interesting that image data becomes less informative compared to the other two data modalities in classifying 3-class disease status, while image data can definitely discriminate AD from CN effectively. Another interesting finding is that we do not see an "accuracy decay" scaling with time as we expect. Rather, performance of our model does not show any clear relationship with time. A possible reason is that the time interval is simply not long enough; longer longitudinal studies would provide more insight into



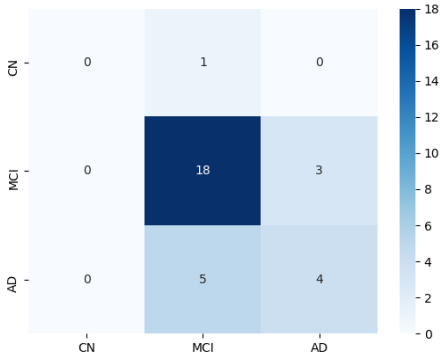

*(a) Confusion matrix for 0 month*



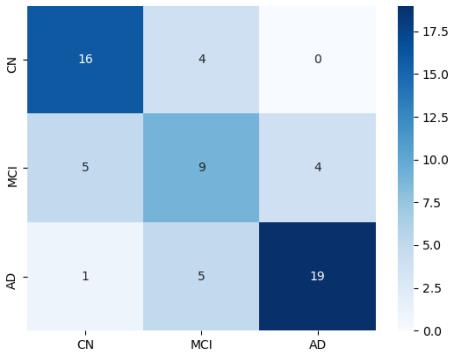

*(b) Confusion matrix for 6 months*




*(c) Confusion matrix for 12 months*




*(d) Confusion matrix for 18 months*




*(e) Confusion matrix for 24 months*

***Fig. 7:*** *Confusion matrix of the multi-modal model based on its prediction on test dataset.*

prediction accuracies over time. While predictions of disease progression based on clinical data were high, combining all data sources produced robust predictions overall. When genomic data, neuroimaging data, and clinical data were used in conjunction with one another, our model could discriminate between CN and AD diagnosis for all patients at baseline, 6 months, 12 months, and 18 months without error. The single error in our two state model occurred when attempting predictive diagnosis at 24 months; this gives the model an overall predictive accuracy of 99.9%. The reason behind the significant increase in diagnostic accuracy when considering only CN vs AD as opposed to CN vs MCI vs AD could be that MCI as a classification is much less rigorously defined than AD. MCI as a label can be applied on a subjective basis in regards to quality of life, and could be representative of conditions other than progression to AD. This would explain our difficulty in handling this category.

Overfitting is a lingering concern, especially when considering the genome-wide analysis of longitudinal outcomes. Across most models and data modalities, in-sample prediction remained extremely high while out-of-sample prediction was drastically lower. Cross-validation was implemented for genomic models, and validation sets and Dropout layers were added to neuroimaging models to reduce overfitting. Regularization, penalized regression methods, and mixed models can be implemented in future work to further address this concern and dramatically improve model accuracy.

Further extrapolations from this project could include RNAseq data throughout disease progression, both for predictions of disease progression and for diagnosis. Comparing the performance of models that used this data to those without could be indicative of the relative importance of this information compared to genetic, imaging, and clinical data. This would have the added benefit of informing novel diagnostics and therapeutics that interact with epigenetic states. Supplementing a deep learning model with the results of emerging diagnostics or therapeutics could also prove informative.

## 6 Data Availability

Reference code in order to build the trained FCN can be found in their repository: https://github.com/vkola-lab/brain2020. Code utilized in this project can be found at https://github.com/HHenryD/MIT6.047.

## 7 Project Reflections

Peer review feedback was a good indicator of some strengths and weaknesses of the proposal that we could address. The influence of peer review feedback predominately manifested itself in the creation of a more detailed project plan, which also formed naturally as the project became more structured. The background to the fundamental problem we were trying to address seemed clear, and our innovation has been recreated to reflect our new project goals. The importance of the objective seemed clear to most of our reviewers as well. Other factors dependent on data availability, such as timeline and division of labor, have been subject to change.

The final project is similar to the initially proposed project; however, instead of using RNAseq data to monitor the transcriptomic profile of patients throughout disease progression, genomics data was used instead. The change has been accounted for throughout this report. The causal factor resulting in this change of project trajectory was access to data. Access to the ROSMAP cohort data, specifically RNAseq data, was required in order to monitor transcriptomic profiles of patients throughout disease progression. Without this element, the focus of the project shifted to using genomics data in conjunction with neuroimaging data and other multi-omics data from ADNI throughout disease progression. Furthermore, as ADNI data did not include RNAseq, and since ROSMAP data was granted for a time before being revoked, access to ADNI data was not requested until rather late in the project timeline, and was only granted on Nov 18th. This created a substantial time constraint on accomplishing what we would have liked to with this project. To account for this scenario, we could have assumed ROSMAP data would not be available, and requested ADNI access earlier under a modified proposal.

Our division of labor was somewhat flexible. Ahmad worked on processing genomic data, Tong worked on processing neuroimaging data, and Charles worked on data management. Many elements were collaborative to ensure data interfaces were compatible and avoided redundancy. Utilizing the knowledge gained through this course on our project has remained very fulfilling throughout our work up through this report.

## 8 References

[1] Bari, S., Vike, N. L., Stetsiv, K., Walter, A., Newman, S., Kawata, K., Bazarian, J. J., Papa, L., Nauman, E. A., Talavage, T. M., Slobounov, S., Breiter, H. C. (2021). Integrating multi-omics with neuroimaging and behavior: A preliminary model of dysfunction in football athletes. *Neuroimage: Reports*, Volume 1, Issue 3, 100032, ISSN 2666-9560.

[2] Beach, T. G., Monsell, S. E., Phillips, L. E., Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *Journal of neuropathology and experimental neurology*, 71(4), 266-273.

[3] Bohnen, N. I., Djang, D. S., Herholz, K., Anzai, Y., & Minoshima, S. (2012). Effectiveness and safety of 18F-FDG PET in the evaluation of dementia: a review of the recent literature. *Journal of Nuclear Medicine*, 53(1), 59-71.

[4] Chen, T. Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. of KDD*, 785–794 (2016)

[5] De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., Klein, H. U., White, C. C., Peters, M. A., Lodgson, B., Nejad, P., Tang, A., Mangravite, L. M., Yu, L., Gaiteri, C., Mostafavi, S., Schneider, J. A., & Bennett, D. A. (2018). A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Scientific data*, 5, 180142. https://doi.org/10.1038/sdata.2018.142

[6] Ding, C., Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(2), 185–205.

[7] Folstein, M. F., Folstein, S. E., McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3), 189-198.

[8] Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2), 67-77.

[9] Harper, L., Barkhof, F., Scheltens, P., Schott, J. M., Fox, N. C. (2014). An algorithmic approach to structural imaging in dementia. *Journal of Neurology, Neurosurgery Psychiatry*, 85(6), 692-698.

[10] Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M., Hyman, B. T., Nixon, R. A., Jones, D. T. (2021). Alzheimer disease. *Nature reviews. Disease primers*, 7(1), 33.

[11] Masters, C. L., Bateman, R., Blennow, K., Rowe, C. C. Sperling 432 RA, Cummings JL (2015) Alzheimer's disease. *Nat Rev Dis 433 Primers*, 1(15056), 434.

[12] Mattsson, N., Insel, P. S., Donohue, M., Jögi, J., Ossenkoppele, R., Olsson, T., ... Hansson, O. (2019). Predicting diagnosis and cognition with 18F-AV-1451 tau PET and structural MRI in Alzheimer's disease. Alzheimer's Dementia, 15(4), 570-580.

[13] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., ... & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's / dementia, 7(3), 263-269.

[14] Nativio, R., Lan, Y., Donahue, G., Sidoli, S., Berson, A., Srinivasan, A. R., Shcherbakova, O., Amlie-Wolf, A., Nie, J., Cui, X., He, C., Wang, L. S., Garcia, B. A., Trojanowski, J. Q., Bonini, N. M., & Berger, S. L. (2020). An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nature genetics*, 52(10), 1024–1035.

[15] Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., ... & Weiner, M. W. (2010). Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology*, 74(3), 201-209.

[16] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based journal of human genetics, 81(3), 559–575.

[17] Purcell, S. CC. PLINK 1.9. 2015. https://cog-genomics.org/plink/. Accessed 18 Nov.

[18] Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., ... & Kolachalama, V. B. (2020). Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain*, 143(6), 1920-1933.

[19] Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., Risacher, S. L., Nho, K., Huentelman, M. J., Craig, D. W., Thompson, P. M., Stein, J. L., Moore, J. H., Farrer, L. A., Green, R. C., Bertram, L., Jack, C. R., Jr, Weiner, M. W., Alzheimer's Disease Neuroimaging Initiative (2010). Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core

aims, progress, and plans. *Alzheimer's dementia : the journal of the Alzheimer's Association*, 6(3), 265–273. https://doi.org/10.1016/j.jalz.2010.03.013

[20] Sikorska, K., Lesaffre, E., Groenen, P., Rivadeneira, F., Eilers, P. (2018). Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization -GALLOP algorithm. *Scientific reports*, 8(1), 6815. https://doi.org/10.1038/s41598-018-24578-7

[21] Zhu W, Yuan Y, Zhang J, Zhou F, Knickmeyer RC, Zhu H. (2017). Genome-wide association analysis of secondary imaging phenotypes from the Alzheimer's disease neuroimaging initiative study. *Neuroimage*. 2017 Feb 1;146:983-1002. doi: 10.1016/j.neuroimage.2016.09.055. Epub 2016 Oct 4. PMID: 27717770; PMCID: PMC5322243.