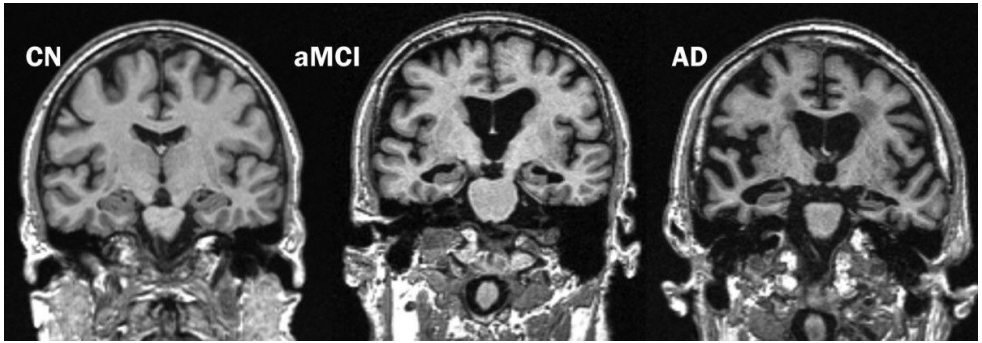
Predicting Alzheimer's Disease

Heather Shen, Zhen Miao, Kristian Eschenburg

Background

- Alzheimer’s Disease (AD) is a neurodegenerative disease characterized by progressive loss of memory
- 1 in 10 people 65 and older have AD
- 6th leading cause of death in United States
- Deaths due to AD nearly doubled from 2000-2014
- In 2017, healthcare costs related to AD were $250B



Research Question

Can we predict whether or not a patient has Alzheimer’s Disease based on neuropathology proteins measured in the brain and traumatic brain injury history?

Identifying challenges in our analysis

- Missing data: Partial data & Entire data
- Too many covariates: the number of variables outnumber our sample size
- The imbalance of some covariates
- Interval data

Methods

- Data consists of:
  - Response variable (whether or not patient had Alzheimer's)
  - Demographic data (education, age, traumatic brain injury history)
  - Neuropathology proteins from 4 regions of the brain
- Use Hot Deck to impute missing data
- Logistic Regression for prediction
- Model Selection, BIC
- 5-fold cross validation for estimating error rates

Assumptions and form of logistic regression

- Binary response variable
  - Demented vs. not demented
  - In reality, dementia exists on a spectrum, but typically studied as factor
- Independent observations
- Linearity of independent variables and log odds

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

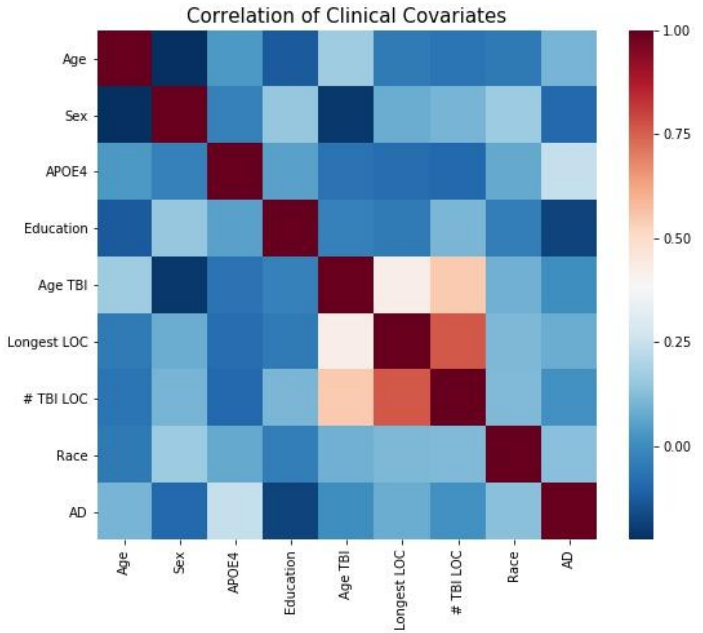
$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Correlation Matrix of Clinical Covariates

- corr(APOE4,AD) = 0.23
  - expected
  - more copies of APOE4 increases risk of AD
- high correlation amongst TBI covariates
  - driven by non-TBI patients
    - i.e. 0 TBI → 0 longest LOC



Proportions of APOE4

	Allele yes	Allele no
Alzheimer's yes	0.1495	0.318
Alzheimer's no	0.065	0.467
Proportion having Alzheimer's given Allele	0.697	0.405

Preliminary Results

- 5 folds cross validation with L2 penalty
- 0.628 accuracy (sd = 0.04) with only clinical data
- 0.46 accuracy (sd = 0.18) with clinical data along with PCA from neuropathology proteins
    - We can see that prediction based only on clinical data performs better
- Apolipoprotein E4 (APOE4) p-value: 0.015
    - Protein related to fat metabolism
    - 3 polymorphisms (variants) (E2,E3,E4)
    - E2 shown protective against AD (but related to Parkinson’s and vascular disease)
    - One copy of E4 raises risk of AD by 2-3 times, two copies by up to 12 times

Moving Forward

- We would like to examine whether or not certain regions of the brain can individually predict whether or not a patient has Alzheimer’s Disease
- We will also be performing prediction based solely on neuropathology data without clinical data

Things We Learned

- We had lots of missing data
  - Examined imputation methods
    - MICE (Multiple Imputation Through Chained Equations)
      - For v in Variables
        - Regress observed(v) on observed(Variables \ v)
        - Predict missing(v) given E(v | Variables \ v)
      - Repeat, permuting order in which variables are examined
        - Generates multiple imputed datasets
    - Since we have low row-rank matrix (n samples < V variables), have identifiability problem
      - MICE failed here
      - How to specify which covariates to use in estimating missing values?

Questions