**Phase One – Business Understanding**

Alzheimer’s Disease is the leading cause of dementia worldwide, and is characterized by the pathological deposition of tau and amyloid-beta protein aggregates in the cerebral cortex. I work in an Alzheimer’s Disease research lab studying the protein tau, so I’m particularly interested in the role of tau. In AD, tau misfolds into “neurofibrillary tangles” (NFTs), following a typical pattern of propagation known as the Braak stages. While such staging has traditionally only been possible at autopsy, recent years have seen novel PET tracers specific for tau NFTs that can track tau spreading. Cross-sectional studies have also associated regional tau-PET uptake with cognitive status, but there are not yet many (if any) longitudinal studies using tau-PET. It would be useful to directly examine how changes in regional tau burden are associated with and predictive of cognitive changes.

**Phase Two – Data Understanding**

For my project, I have chosen to compare longitudinal tau-PET neuroimaging with longitudinal cognitive test scores for n=259 subjects, each with at least two PET scans spaced approximately one year apart. The tau-PET data was pre-processed by aligning the PET standardized uptake value (SUV) image to structural MRI (MPRAGE) acquired within the same session; regions of interest (ROIs) are quantified based on the co-registered images using FreeSurfer. This data was acquired from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), a data consortium between universities and medical centers around the country. Cognitive status was measured via a battery of neuropsychological evaluations, but for simplicity, I have selected the Alzheimer's Disease Assessment—Cognitive 13-item scale (ADAS13). This scale ranges from 0 to 70, with 0 indicating no cognitive impairment and 70 indicating very severe cognitive impairment. Additionally, each subject received a cognitive diagnosis at each visit. The data is free to access once a user signs up as a “Principal Investigator” along with their proposed analysis.

**Phase 3 – Data Preparation**

The tau-PET and cognitive score data pertain to individual time points, which I will convert to slopes to compare change in tau burden with change in cognitive ability. This will yield a delta-SUVR value and a delta-ADAS13 score. For subjects with more than two time points, I will calculate the geometric mean of the tau-PET values and the ADAS13 scores. For the cognitive diagnosis, I will create three columns: one for the first time-point diagnosis, one for the second time-point diagnosis, and one summarizing the status (e.g. stable MCI, MCI decline to dementia, etc.). I anticipate missing data in certain ROIs that are not detected in all subjects – for example, the ‘fifth ventricle” ROI is only reported in a small fraction of subjects. Depending on the overall missingness of a variable, I will either omit it or impute based on tau-PET values in the other ROIs.

**Phase 4 – Modeling**

I will explore two different predictive models. The first will use annual rate of regional tau-PET accumulation to predict change in ADAS-13 score. I will create an ensemble for this prediction model with elastic net regression, kNN, and a random forest. Since there are many ROIs (n=33), I will also explore if dimension reduction with PCA improves model performance. The second model will predict whether an individual’s cognitive status will remain stable or decline between clinical visits based on the rate of regional tau-PET accumulation. I haven’t decided yet whether to consider CN and MCI subjects separately or all as one; if it’s the latter, this is a binary classification problem, for which I will use logistic regression combined with random forest. I’m sticking with “white-box” models as I’m interested in the individual regional contributions of tau deposition to cognitive decline, though this aim may evolve as I continue developing the models.

**Phase 5 – Model evaluation**

I will evaluate the continuous model predicting ADAS13 scores using MSE and correlation coefficient (Spearman, unless data is approximately normal). I will evaluate the binary cognitive status classification model according to ROC and precision vs. recall.

**Phase 6 – Deployment**

I will create a Shiny app to interactively display the results of my analysis. The user will switch between models (continuous ADAS13 vs. stable/decline prediction). For the continuous ADAS13 model, I will show a correlation scatterplot and a bar graph of feature importance. I will also use packages from the neuroconductor library to