

Module_1: (Alzheimer's)

Team Members:

Addison Buck, Marielle Miranda

Project Title:

Comparing the Impact of A β 42 and pTau Levels on CERAD Scores Using Computational Methods

Project Goal:

These are the questions that we considered

- While it is known that ABeta42 and pTau are the main contributors to Alzheimers, does the ratio/measurements of ABeta40 and tTau show the same results as those two?
- Is there a correlation between age of onset symptoms for Alzheimer's and age of death?
- Is there a correlation between patient race and ABeta42 or pTau levels?

The Question we chose:

- Does ABeta42 or does pTau have a bigger contributing effect on Alzheimer's CERAD scores, or is it even?

For the bar graph, we also had a sub question: Is there a difference in pTau levels by sex?

Disease Background:

Fill in information about 11 bullets:

- Prevalence & incidence: *It is estimated that 7.2 million Americans aged 65 and older have Alzheimer's, with 200,000 cases in the 30-64 aged category. (ChatGPT Prompt: How many current cases of Alzheimer's are there in the United States?). It is also estimated that 900,000 people aged 65 and older develop Alzheimer's per year. (ChatGPT Prompt: How many new diagnoses of Alzheimer's are there per year?)*
- Economic burden: *The total cost of Alzheimer's and related diseases to the United States is 781 billion dollars. Those with Alzheimer's pay about 10,200 dollars in treatments, doctors, and living expenses out-of-pocket annually, 7700 dollars more than the average. Additionally, insurances such as Medicaid pay patients around \$43,445 annually. (ChatGPT Prompts: How much does it cost Americans per year to deal with Alzheimer's, and On average, how much does a patient have to spend living with Alzheimer's, factoring in treatments, doctors, assisted living, etc.)*
- Risk factors (genetic, lifestyle): *The strongest genetic risk factor is APOE ϵ 4 allele. It is shown that an increase of APOE ϵ 4 allele can lead to a higher risk of late on-set*

Alzheimer's. Additionally, having family members (first-degree relatives) who have been diagnosed with Alzheimer's can increase this risk. For lifestyle/environments, the biggest risk can be age. After every 5 years after the age of 65 the risk of obtaining Alzheimer's doubles. In all, a decrease and poor lifestyle choices, such as physical inactivity, lack of sleep, poor cardiovascular health, and low cognitive engagement lead to an increase of Alzheimer's. ChatGPT prompt: "what are genetic and lifestyle risk factors of Alzheimer's"

- Societal determinants: Social determinants of Alzheimer's can be mostly linked to one's socioeconomic status. A patient's socioeconomic status can affect their availability to healthcare access (an inability to healthcare can lead to a later diagnosis and treatment), exposure to environmental factors (higher exposure to pollution and toxins can lead to cognitive decline), education (fewer years of education has been linked to a higher increase of Alzheimer's). Some other factors that can affect Alzheimer's can be social engagement since a lack of social engagement can lead to an increase of dementia due an increase of isolation. ChatGPT prompt: "what are societal determinants for Alzheimer's"
- Symptoms: Signs of Alzheimer's dementia are memory loss disrupting daily life, challenges in planning or problem-solving, difficulty completing familiar tasks, confusion, trouble understanding images, new problems with speaking or writing, misplacing things, poor judgement, withdrawal from work, and changes in mood (<https://www.alz.org/getmedia/ef8f48f9-ad36-48ea-87f9-b74034635c1e/alzheimers-facts-and-figures.pdf>). Patients may also experience depression, anger, changes in sleeping, and delusions. (<https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447>)
- Diagnosis: For Alzheimer's many common symptoms are memory loss, cognitive decline, and behavioral changes. Link: (<https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048075#:~:text=Diagnosing%20Alzheimer's%20dementia%0A%0AThe%20doctor%20or%20other%20healthcare,have%20a%20physical%20exam%20and%20several%20tests>)
- Standard of care treatments (& reimbursement): To treat Alzheimer's most patients will use cholinesterase inhibitors, memantine, anti- amyloid antibodies link: [https://www.brain.northwestern.edu/dementia/ad/treatment.html#:~:text=While%20there%20are%20no%20cures,Other%20Medications,When it comes to reimbursement, in 2021, 42.5 million dollars were spent on research and development on drugs alone. Additionally, the amount patients pay per prescription may range from 11 dollars to 724 dollars\(annually, it can range from 26,500 to 32,000 dollars\)](https://www.brain.northwestern.edu/dementia/ad/treatment.html#:~:text=While%20there%20are%20no%20cures,Other%20Medications,When it comes to reimbursement, in 2021, 42.5 million dollars were spent on research and development on drugs alone. Additionally, the amount patients pay per prescription may range from 11 dollars to 724 dollars(annually, it can range from 26,500 to 32,000 dollars))
https://www.goodrx.com/conditions/alzheimers-disease/drugs?srsId=AfmBOop5QxJ78zYTsO6fN7zl7jPm_fe8CgV2B3ZnPKVS4wUpSa1QdiwH
- Disease progression & prognosis: The progression of Alzheimer's is typically slow and gradual with the disease advancing over the course of seven years. Additionally, Alzheimer's is divided into seven stages: preclinical, mild cognitive impairment, mild

Alzheimer's, moderate Alzheimer's, and severe Alzheimer's. **PRECLINICAL:** This stage is before symptoms arrive. There is no noticeable memory loss/cognitive problems, but protein buildup of ABeta and Tau start to begin. **MILD COGNITIVE IMPAIRMENT:** There are subtle lapses in memory (i.e. the patient may start to forget names), but the patient is still able to live independently and may need to rely on notes or reminders. **MILD ALZHEIMER'S:** This is the early stage of Alzheimer's and patients may start to experience increased forgetfulness (i.e. forgetting recent conversations or even events). Additionally, they may start having trouble finding the right words, problem-solving, and even start to show mood/personality changes. **MODERATE ALZHEIMER'S:** Also known as the middle stage of Alzheimer's, the patient will start to show more pronounced memory loss (i.e. forgetting important dates/personal history). Additionally, they will start to have trouble with daily tasks and may start demonstrating behavioral changes, such as wandering and suspiciousness. Patients, at this stage, require significant support and cannot live independently. **Severe Alzheimer's:** This is the late stage of Alzheimer's. Patients need help with all daily activities and may start to lose the ability to communicate coherently. Furthermore, patients will start to have difficulty walking, swallowing, and controlling bowel movements. For prognosis, the survival after diagnosis is on average 4 to 8 years. However, depending on the patient's age, health, and sex (women are typically more affected due to a longer life expectancy) they may live up to 15 to 20 years. The progression of Alzheimer's is typically faster in older patients compared to those who are younger. And some patients may die from complications, such as pneumonia, infections, and falls/injuries. Chat GPT Prompt: What are disease progression and prognosis for Alzheimer's and Link:

<https://www.alz.org/alzheimers-dementia/stages#:~:text=Being%20forgetful%20of%20events%20or,care%20in%20a%20safe%20environment>.

- Continuum of care providers: Common care of providers for Alzheimer's are neurologists, radiologists, primary, geriatricians, therapists, and psychologist/psychiatrist Link: <https://www.alzheimers.gov/professionals/health-care-providers>
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology): **MOLECULAR PHYSIOLOGY:** Alzheimer's is caused by accumulation of two proteins, Tau and Beta-Amyloid. An accumulation of Tau causes tangles to form inside the neuron, whereas beta-amyloid clump into plaque and build up between neurons. This in turn leads to synaptic dysfunction and cognitive decline. **ANATOMY:** The two most affected parts of the brain are the cerebral cortex and the hippocampus. The cerebral cortex faces atrophy of the frontal, temporal, and parietal lobes, which is what causes memory loss and problems with decision making. The hippocampus is the earliest region of the brain that is affected and the damage done to this part causes early short-term memory loss. **ORGAN PHYSIOLOGY:** Due to reduced neuronal communication patients suffer cognitive decline. Additionally, due to synaptic dysfunction there is impaired long-term potentiation (when the synaptic connections are no longer becoming stronger/not as strong after repeated stimulation). Furthermore, there is also a deficit of acetylcholine and a reduction of cerebral blood

flow. CELLULAR PHYSIOLOGY: There is neural and synaptic loss, with synaptic loss dealing with the loss of dendritic spines and synaptic terminals. Additionally, there is glial-mediated inflammation due to the astrocytes becoming more reactive (contributes to neuroinflammation) and microglia being overactivated (leads to chronic inflammation and neuronal damage) Chat GPT Prompt: What are biological mechanisms (anatomy, organ physiology, cell & molecular physiology) that cause/are present with Alzheimer's? Link: /youtu.be/hEw1Yq_4PaA?si=xUKp5hlK8FRUGAWq

- Clinical Trials/next-gen therapies: *Some next-gen therapies are hoping to focus on early intervention by using treatments that will clear the beta-amyloid plaque (lecanemab and donanemab). Additionally, researchers are hoping to use gene therapy and stem cells* <https://www.alzheimers.org.uk/research/our-research/dementia-research-news/researching-new-drugs-alzheimers-disease#:~:text=One%20trial%20is%20focused%20on,called%20TRAILRUNNER%20DALZ%201>.
- Further background on CERAD scores: CERAD scores are an assessment of cognitive function based on the amount of neuritic plaques. Patients are scored in either absent, sparse, moderate, or frequent ([https://pmc.ncbi.nlm.nih.gov/articles/PMC6250207/#:~:text=Second%2C%20the%20CERAD%20score%20reflects,Mental%20State%20Examination%20\(MMSE\).](https://pmc.ncbi.nlm.nih.gov/articles/PMC6250207/#:~:text=Second%2C%20the%20CERAD%20score%20reflects,Mental%20State%20Examination%20(MMSE).)). It was hypothesized that while both ABeta42 and pTAU would have a strong correlation with these scores that pTAU would have a stronger correlation with CERAD scores as it has been found that pTAU is more closely associated with severity of cognitive function (<https://pubmed.ncbi.nlm.nih.gov/19260027/#:~:text=Methods:%20We%20investigated%20the%20relation,with%20clinical%20onset%20and%20progression.>).

Data-Set:

- Describe the two data sets: *The Meta data is a more comprehensive set of data that gives information on each patient's cultural background, age, gender, and cognitive testing (i.e. CERAD score, brain pH, and age of onset cognitive symptoms). The Luminex data set gives the information of each patient/donor's levels of ABeta40, ABeta42, tTAU, and pTAU*
- Cite the sources of the data: *Published by nature neuroscience journal, Gabitto, M.I., Travaglini, K.J., Rachleff, V.M. et al. Integrated multimodal cell atlas of Alzheimer's disease. Nat Neurosci 27, 2366–2383 (2024).* <https://doi.org/10.1038/s41593-024-01774-5>
- Describe how the data was collected/what techniques were used: *The study was conducted with a longitudinal cohort. The study combined qualitative data with data collected from studies of the brain. Brain tissue was collected from postmortem donors, and the middle temporal gyrus was studied. Techniques such as RNA sequencing and the use of antibodies were used to analyze tau and beta-amyloid levels.*

- What units are the data measured in? *All units that the Luminex data are measured in are in pg/ug*
- When was the data published? *Published October 14 2024*
- What was the study design/what were the methodological techniques used? *This study was conducted using a longitudinal cohort, combining qualitative data with quantitative Abeta and tau data collected from brain studies. Specifically, brain tissue was collected from postmortem donors. Methods such as RNA sequencing and antibody use were used for analysis and obtaining data.*
- What subjects was the data obtained from **The data was obtained from patients, with brain samples collected postmortem. Most donors were older, however, it was found that tissue degeneration did not affect the results significantly.*
- Was there bias in how the data was collected? *Yes, there is a cohort bias for female donors when the researchers decided to include sex as another variable in their models. Additionally, there is some bias in the models for ABeta42 levels due to several outliers in the data.*
- What were the limitations/assumptions in the data set? *Some limitations were that researchers decided to exclude the single nucleus -omics data that was generated from 2 donors due to a low RIN and brain pH, which may have limited other data types. Additionally, for assumptions, the researchers assumed the following for the data distribution: Poisson for quantitative neuropathological data, zero-inflated negative binomial for gene expression data, Bernoulli for chromatin accessibility data, and negative binomial for spatial transcriptomic data.*
- Code for learning how to work with a data set is shown below:

```
## Project Notes (can be pasted to VS Code)
## Class 9/11: we wrote code to find the headers, rows, and number of
rows for the MetaData and Luminex files
## Code for UpdatedMetaData.csv headers written with help of ChatGPT
prompt "I have an empty file where I am trying to write code for the
column heads of another csv file called UpdatedMetaData.csv. I am
using python in visual studio code. give me the steps and code that I
need"
## Code for UpdatedMetaData.csv rows written with help of ChatGPT
prompt "Great, that worked. Now I want to figure out the names of the
rows."
## Code for UpdatedMetaData.csv row numbers written with help of
ChatGPT prompt "Now I want the count of rows."
## Code for LuminexData.csv headers, rows, and number of rows was
written with "how should you access and organize data for a csv named
UpdatedLuminex.csv, can you write the full line of code that should do
what was expected and how to fix it, can you find out how many rows of
data and columns are there and add this to the rest of the coding"
```

```

## Code for UpdatedMetaData.csv:
import csv
from statistics import LinearRegression
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

# CSV column info
# Characterizing MetaData csv
file_path_1 = "UpdatedMetaData.csv"

with open(file_path_1, newline="") as f:
    reader = csv.reader(f) ##reads metadata file, opens data set in
python

    # Get headers from MetaData
    headers_1 = next(reader)
    print("Column headers for MetaData:") ## gives a section heading
    for i, h in enumerate(headers_1):
        print(f"{i}: {h}") ##iterates through columns and prints
column names

    # Initialize a list to store first column (row names)
    first_column_1 = []

    # Read remaining rows
    for row in reader:
        first_column_1.append(row[0]) ##adds next rows to the list

    # Print row names
    print("\nRow names (from the first column for MetaData):")
    for name in first_column_1:
        print(name) ##loops through the list and prints the row names

    # Print row count
    print(f"\nNumber of rows (excluding header):
{len(first_column_1)}")

# Characterizing Luminex CSV
file_path_2 = "UpdatedLuminex.csv"

with open(file_path_2, newline="") as f: ##opens file in python
    reader = csv.reader(f)

    # Get headers
    headers_2 = next(reader)
    print("Column headers for Luminex:") ##prints section header
    for i, h in enumerate(headers_2): ##loops through list to print

```

```

headers
    print(f"{i}: {h}")

# Initialize a list to store first column (row names)
first_column_2 = []

# Read remaining rows
for row in reader:
    first_column_2.append(row[0]) ##adds to the list storing first
column row names

# Print row names
print("\nRow names (from the first column for Luminex):")
for name in first_column_2:
    print(name)

# Print row count
print(f"\nNumber of rows (excluding header):
{len(first_column_2)}")

#Code for patient class was written with the help of ChatGPT: "can you
explain how to make a new patient object class by pulling data from
two separate csvs?"
class Patient: ##creates a new patient object definition with metadata
and luminex as dictionaries
    def __init__(self, donor_id, metadata, luminex):
        self.donor_id = donor_id #matches data sets on donor id
        # pull important metadata values relevant for our analysis
        self.age = metadata.get("Age") ##pulls whatever relevant data
type from the csv to add to new file
        self.sex = metadata.get("Sex")
        self.cerad = metadata.get("CERAD score")
        # pull luminex values values relevant for our analysis
        self.abeta42 = luminex.get("ABeta42 pg/ug")
        self.ptau = luminex.get("pTAU pg/ug")

    def __repr__(self):
        return (f"Patient({self.donor_id}, Age={self.age},
Sex={self.sex}, "
                f"CERAD={self.cerad}, ABeta42={self.abeta42},
pTAU={self.ptau})") ##returns all relevant information when called,
can be used as a debugger

metadata_df = pd.read_csv("UpdatedMetaData.csv")
luminex_df = pd.read_csv("UpdatedLuminex.csv")

## Add all donor ids to the patient object, when printed gives all of
the relevant info for each patient based on donor id
patients = []

```

```

for donor_id in set(metadata_df["Donor ID"]) & set(luminex_df["Donor ID"]):
    meta_row = metadata_df[metadata_df["Donor ID"] ==
donor_id].iloc[0].to_dict()
    lum_row = luminex_df[luminex_df["Donor ID"] ==
donor_id].iloc[0].to_dict()
    patients.append(Patient(donor_id, meta_row, lum_row)) ##actually
adds all of the patient information to a new patient object

```

Column headers for MetaData:

```

0: Donor ID
1: Primary Study Name
2: Secondary Study Name
3: Age at Death
4: Sex
5: Race (choice=White)
6: Race (choice=Black/ African American)
7: Race (choice=Asian)
8: Race (choice=American Indian/ Alaska Native)
9: Race (choice=Native Hawaiian or Pacific Islander)
10: Race (choice=Unknown or unreported)
11: Race (choice=Other)
12: specify other race
13: Hispanic/Latino
14: Highest level of education
15: Years of education
16: APOE Genotype
17: Cognitive Status
18: Age of onset cognitive symptoms
19: Age of Dementia diagnosis
20: Known head injury
21: Have they had neuroimaging
22: Consensus Clinical Dx (choice=Alzheimers disease)
23: Consensus Clinical Dx (choice=Alzheimers Possible/ Probable)
24: Consensus Clinical Dx (choice=Ataxia)
25: Consensus Clinical Dx (choice=Corticobasal Degeneration)
26: Consensus Clinical Dx (choice=Control)
27: Consensus Clinical Dx (choice=Dementia with Lewy Bodies/ Lewy Body
Disease)
28: Consensus Clinical Dx (choice=Frontotemporal lobar degeneration)
29: Consensus Clinical Dx (choice=Huntingtons disease)
30: Consensus Clinical Dx (choice=Motor Neuron disease)
31: Consensus Clinical Dx (choice=Multiple System Atrophy)
32: Consensus Clinical Dx (choice=Parkinsons disease)
33: Consensus Clinical Dx (choice=Parkinsons Cognitive Impairment - no
dementia)
34: Consensus Clinical Dx (choice=Parkinsons Disease Dementia)
35: Consensus Clinical Dx (choice=Prion)
36: Consensus Clinical Dx (choice=Progressive Supranuclear Palsy)
37: Consensus Clinical Dx (choice=Taupathy)

```

38: Consensus Clinical Dx (choice=Vascular Dementia)
39: Consensus Clinical Dx (choice=Unknown)
40: Consensus Clinical Dx (choice=Other)
41: If other Consensus dx, describe
42: Last CASI Score
43: Interval from last CASI in months
44: Last MMSE Score
45: Interval from last MMSE in months
46: Last MOCA Score
47: Interval from last MOCA in months
48: PMI
49: Rapid Frozen Tissue Type
50: Ex Vivo Imaging
51: Fresh Brain Weight
52: Brain pH
53: Overall AD neuropathological Change
54: Thal
55: Braak
56: CERAD score
57: Overall CAA Score
58: Highest Lewy Body Disease
59: Total Microinfarcts (not observed grossly)
60: Total microinfarcts in screening sections
61: Atherosclerosis
62: Arteriolosclerosis
63: LATE
64: RIN
65: Severely Affected Donor

Row names (from the first column for MetaData):

H19.33.004
H20.33.001
H20.33.002
H20.33.004
H20.33.005
H20.33.008
H20.33.011
H20.33.012
H20.33.013
H20.33.014
H20.33.015
H20.33.016
H20.33.017
H20.33.018
H20.33.019
H20.33.020
H20.33.024
H20.33.025
H20.33.026

H20.33.027
H20.33.028
H20.33.029
H20.33.030
H20.33.031
H20.33.032
H20.33.033
H20.33.034
H20.33.035
H20.33.036
H20.33.037
H20.33.038
H20.33.039
H20.33.040
H20.33.041
H20.33.043
H20.33.044
H20.33.045
H20.33.046
H21.33.001
H21.33.002
H21.33.003
H21.33.004
H21.33.005
H21.33.006
H21.33.007
H21.33.008
H21.33.009
H21.33.010
H21.33.011
H21.33.012
H21.33.013
H21.33.014
H21.33.015
H21.33.016
H21.33.017
H21.33.018
H21.33.019
H21.33.020
H21.33.021
H21.33.022
H21.33.023
H21.33.025
H21.33.026
H21.33.027
H21.33.028
H21.33.029
H21.33.030
H21.33.031

H21.33.032
H21.33.033
H21.33.034
H21.33.035
H21.33.036
H21.33.037
H21.33.038
H21.33.039
H21.33.040
H21.33.041
H21.33.042
H21.33.043
H21.33.044
H21.33.045
H21.33.046
H21.33.047

Number of rows (excluding header): 84

Column headers for Luminex:

0: Donor ID
1: ABeta40 pg/ug
2: ABeta42 pg/ug
3: tTAU pg/ug
4: pTAU pg/ug

Row names (from the first column for Luminex):

H20.33.045
H20.33.044
H21.33.045
H20.33.046
H20.33.014
H21.33.046
H21.33.047
H20.33.011
H19.33.004
H21.33.005
H21.33.001
H20.33.024
H21.33.007
H20.33.012
H20.33.025
H20.33.004
H20.33.017
H20.33.013
H20.33.015
H20.33.018
H20.33.008
H20.33.005
H20.33.026

H20.33.041
H20.33.027
H21.33.008
H20.33.019
H20.33.001
H20.33.002
H20.33.028
H21.33.006
H20.33.029
H20.33.030
H20.33.031
H20.33.020
H20.33.032
H20.33.033
H20.33.034
H20.33.035
H20.33.036
H20.33.037
H20.33.043
H20.33.016
H21.33.012
H21.33.011
H20.33.038
H21.33.010
H20.33.039
H21.33.009
H21.33.017
H21.33.016
H21.33.015
H20.33.040
H21.33.002
H21.33.003
H21.33.014
H21.33.013
H21.33.021
H21.33.020
H21.33.004
H21.33.022
H21.33.019
H21.33.018
H21.33.023
H21.33.025
H21.33.044
H21.33.026
H21.33.027
H21.33.028
H21.33.029
H21.33.030
H21.33.031

H21.33.032
H21.33.033
H21.33.034
H21.33.035
H21.33.036
H21.33.037
H21.33.038
H21.33.039
H21.33.040
H21.33.041
H21.33.042
H21.33.043

Number of rows (excluding header): 84

Data Analysis:

- For the first part of our data analysis, we generated a bar graph with two bars and performed a T-test. Due to the fact that our overarching question produces a bar graph with no variance and simply shows an output of two values (meaning that a T-test would not provide any useful data analysis), we were advised to use a slightly different question for this portion of the data analysis. Instead, we asked the question: "Is there a difference between mean pTau levels in males and females?" To do this, we produced a bar graph with two bars displaying the mean pTau levels for both males and females. Then, we ran a two-tailed T-test. The t-test gave a t-value of -0.0256 and a p-value of 0.97855, so the null hypothesis was rejected and the results were not statistically significant. The code and prompts are given in the below kernel.
- For the second part of our data analysis, we created a scatter plot with a linear regression and r^2 value to determine whether ABeta42 or pTau had a greater correlation with CERAD scores. To do this, we first had to map CERAD scores onto numerical values. Then, when graphed against ABeta42 or pTau, we were able to compare the r^2 values. While pTau did have a slightly higher r^2 value of 0.106 (compared to 0.070 for ABeta42), normally indicating a stronger correlation with CERAD scores, both of these correlation coefficients were incredibly low, showing little correlation between either ABeta42 or pTAU with CERAD scores. This is surprising considering ABeta42 and pTAU have been linked to Alzheimer's disease, however neither had a strong correlation with CERAD scores, a measure of cognitive function. Further analysis is discussed in "Verify your results."

```
## Bar graph and t-test
# ChatGPT prompt for bar graph: "Can you write and explain code in python for a bar graph with two bars comparing sex and their average ptau levels? there should be error bars on the graph, and ideally, the graph should not have a negative y axis."
# ChatGPT prompt for t-test: "can you write and explain code for a two-tailed t test comparing the means of the female and male ptau levels?"
# ChatGPT prompt for scatter plot and regression: "Can you write and explain code to generally create a scatter plot with linear regression
```

and r^2 value. Also, please write and explain code for creating a separate csv for the data from this scatter plot."
All code was altered and edited from prompts after being sufficiently understood in order to make it function for the project's purposes

```
import csv
from statistics import LinearRegression
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

class Patient: ##creates a new patient object with metadata and
luminex as dictionaries
    def __init__(self, donor_id, metadata, luminex):
        self.donor_id = donor_id #matches data sets on donor id
        # pull important metadata values relevant for our analysis
        self.age = metadata.get("Age")
        self.sex = metadata.get("Sex")
        self.cerad = metadata.get("CERAD score")
        # pull luminex values values relevant for our analysis
        self.abeta42 = luminex.get("ABeta42 pg/ug")
        self.ptau = luminex.get("pTAU pg/ug")
```

```
    def __repr__(self):
        return (f"Patient({self.donor_id}, Age={self.age},
Sex={self.sex}, "
                f"CERAD={self.cerad}, ABeta42={self.abeta42},
pTAU={self.ptau})")
```

```
metadata_df = pd.read_csv("UpdatedMetaData.csv")
luminex_df = pd.read_csv("UpdatedLuminex.csv")
```

Add all donor ids to the patient object, when printed gives all of the relevant info for each patient based on donor id

```
patients = []
for donor_id in set(metadata_df["Donor ID"]) & set(luminex_df["Donor ID"]):
    meta_row = metadata_df[metadata_df["Donor ID"] ==
donor_id].iloc[0].to_dict()
    lum_row = luminex_df[luminex_df["Donor ID"] ==
donor_id].iloc[0].to_dict()
    patients.append(Patient(donor_id, meta_row, lum_row))
```

```
import matplotlib.pyplot as plt
from scipy import stats
import pandas as pd
```

```
# Build a DataFrame from patient objects
# Extract donor_id, sex, and pTau values from each Patient object,
```

```

which are the relevant values for the bar graph
patient_data = pd.DataFrame([
    {"Donor ID": p.donor_id, "Sex": p.sex, "pTAU pg/ug": p.ptau}
    for p in patients
])

# Drop rows where pTau or Sex is missing from the data frame
patient_data = patient_data.dropna(subset=["pTAU pg/ug", "Sex"])

# Group data by sex and calculate mean for the bars + SEM (standard
error of the mean)
grouped = patient_data.groupby("Sex")["pTAU pg/ug"].agg(["mean",
"sem"]).reset_index()

# Plot bar graph
plt.figure(figsize=(6, 5))
bars = plt.bar(
    grouped["Sex"],
    grouped["mean"],
    yerr=grouped["sem"], #creates error bars to show uncertainty
    capsize=5,
    color=["lightcoral", "skyblue"]
)

plt.ylabel("Average pTau Levels (pg/ug)")
plt.title("Average pTau Levels by Sex")
plt.ylim(bottom=0) #ensures y axis starts at 0
plt.show()

# t-test
#inputs the patient_data dataframe which has columns sex and pTAU
pg/ug
male_ptau = patient_data.loc[patient_data["Sex"] == "Male", "pTAU
pg/ug"]
female_ptau = patient_data.loc[patient_data["Sex"] == "Female", "pTAU
pg/ug"]

print(f"\nMale samples: {len(male_ptau)}")
print(f"Female samples: {len(female_ptau)}")

if len(male_ptau) > 0 and len(female_ptau) > 0:
    t_stat, p_value = stats.ttest_ind(male_ptau, female_ptau,
equal_var=True) #finds the t value, p value, and assumes equal
variance
    print(f"T-value: {t_stat:.4f}") #prints t-value with 4 decimal
points
    print(f"P-value: {p_value:.5f}")

    #to determine if result is statistically significant
    alpha = 0.05

```

```

    if p_value < alpha:
        print(f"Since  $p < \{\alpha\}$ , the difference is statistically
significant.")
    else:
        print(f"Since  $p \geq \{\alpha\}$ , the difference is NOT statistically
significant.")
    else:
        print("Not enough data to perform t-test.")

# scatter plot
#Scatter plots from Patient objects (debug-friendly)
from sklearn.linear_model import LinearRegression

cerad_mapping = {"absent": 0, "sparse": 1, "moderate": 2, "frequent":
3} #maps cerad scores onto numeric values

def make_scatter_from_patients(patients, dep_attr, dep_label, out_csv,
color): #creates a reusable function for the two scatter plots
    data = []
    for p in patients:
        raw_cerad = str(p.cerad).strip().lower() if p.cerad is not
None else None #works as a fallback in case the CERAD score is blank
        dep_val = getattr(p, dep_attr) #reads CERAD score and
dependent variable

        # Debug print the first few entries
        #print(f"Donor {p.donor_id}: CERAD={raw_cerad},
{dep_attr}={dep_val}")

        if raw_cerad in cerad_mapping and dep_val not in (None, "",
"nan"):
            try:
                dep_val = float(dep_val)
                data.append((cerad_mapping[raw_cerad], dep_val,
p.donor_id))
            except ValueError:
                pass

    if not data:
        print(f"No valid data for CERAD vs {dep_label}") #works as a
debugger incase there is no valid data in the data set when comparing
CERAD and the dependent variable
        return

    # Convert to DataFrame
    df = pd.DataFrame(data, columns=["CERAD Score (numeric)",
dep_label, "Donor ID"])
    df.to_csv(out_csv, index=False)
    print(f"CSV file '{out_csv}' created with {len(df)} rows.")
#creates csv for the scatter plot

```

```

# Regression
x = df["CERAD Score (numeric)"].values.reshape(-1, 1) #maps the
CERAD values onto the x axis
y = df[dep_label].values #maps the dependent variable values onto
the y coordinates/y axis
model = LinearRegression().fit(x, y) #creates a model based on the
coordinates

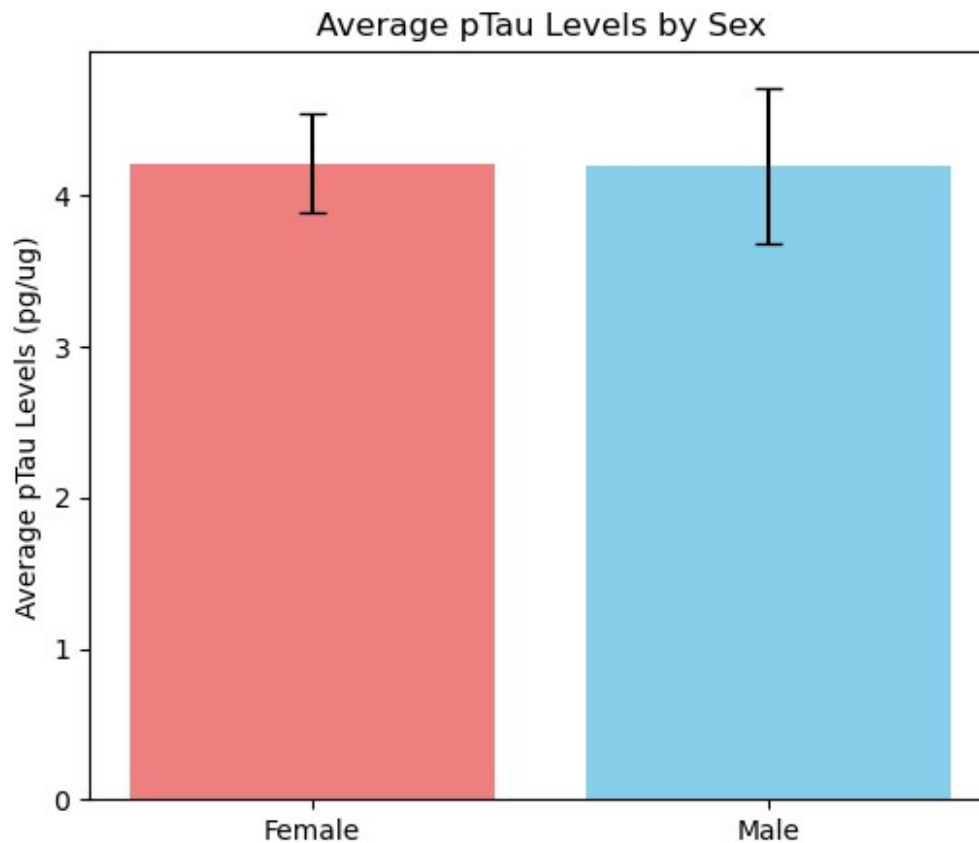
slope, intercept = model.coef_[0], model.intercept_ #based on the
model, finds slope and y-intercept
r2 = model.score(x, y) #finds r^2 value based on the model

# Plot
plt.scatter(x, y, color=color, alpha=0.7, label="Data")
plt.plot(x, model.predict(x), color="red", linewidth=2,
label="Fit")
plt.xlabel("CERAD Score (numeric)")
plt.ylabel(dep_label)
plt.title(f"CERAD Score vs {dep_label}")

equation = f"y = {slope:.2f}x + {intercept:.2f}\nR2 = {r2:.3f}"
#writes the regression equation
plt.text(
    0.05, 0.95, equation,
    transform=plt.gca().transAxes,
    fontsize=10, verticalalignment="top",
    bbox=dict(boxstyle="round", facecolor="white", alpha=0.7)
)
plt.legend()
plt.show()

#Run plots
make_scatter_from_patients(patients, "abeta42", "ABeta42 pg/ug",
"CERAD_vs_ABeta42.csv", "blue")
make_scatter_from_patients(patients, "ptau", "pTAU pg/ug",
"CERAD_vs_pTAU.csv", "green")

```



Male samples: 33

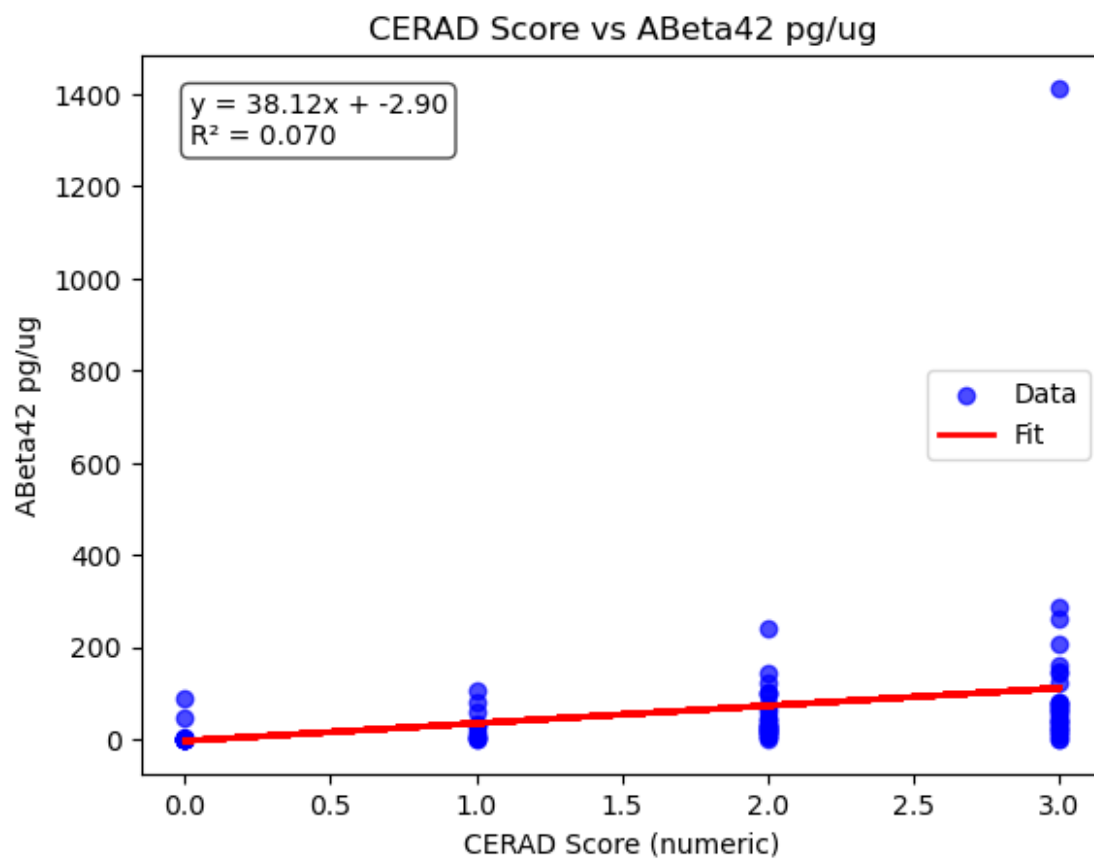
Female samples: 51

T-value: -0.0270

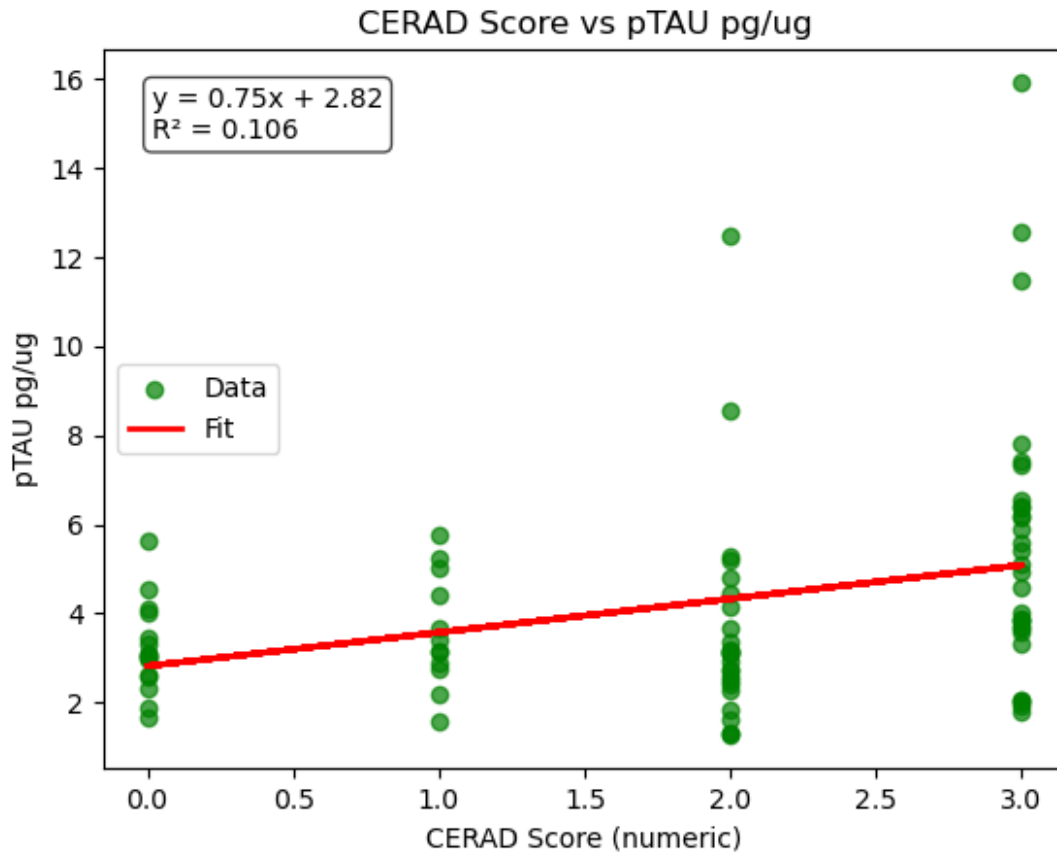
P-value: 0.97855

Since $p \geq 0.05$, the difference is NOT statistically significant.

CSV file 'CERAD_vs_ABeta42.csv' created with 84 rows.



CSV file 'CERAD_vs_pTAU.csv' created with 84 rows.



Verify and validate your analysis:

- The main question proposed was the effects on "Does ABeta42 or does pTau have a bigger contributing effect on Alzheimer's CERAD scores, or is it even?" To interpret and analyze the data, we created a scatterplot that plotted ABeta42 and pTau against CERAD scores. From the data, we gathered that ABeta42 against CERAD score has an R^2 of 0.070, while pTau against CERAD score had an R^2 value of 0.106. By looking at both of the data, the R^2 for pTau had a slightly higher correlation compared to ABeta42; however, since both R^2 values are low, it demonstrates there is an overall weak correlation to CERAD scores. In all, this demonstrates that while pTau has a higher effect on CERAD scores due to its R^2 explaining 10.6% of the variance compared to ABeta42's 7%, both variables aren't stand-alone predictors for CERAD scores and suggest there are other factors that cause the variation.
- Furthermore, the data from the scatterplot of ABeta42 and pTau plotted against CERAD scores can be further supported by a journal article from Oxford University titled Amyloid- β levels and cognitive trajectories in non-demented pTau181-positive subjects without amyloidopathy (<https://pubmed.ncbi.nlm.nih.gov/35973034/>). In the journal article, the cognitive health of 285 non-demented patients with various ratios of pTau, tTau, ABeta42, and ABeta40 were observed over the span of 5 years. Researchers found that patients who had elevated levels of pTau, but normal tTau, ABeta42, and ABeta40 showed no significant onset of dementia. Similarly, patients who had elevated levels of ABeta42 and ABeta40, but normal tTau and pTau, also showed no significant onset of dementia. However, patients who had elevated levels of both ABeta42, ABeta40, and

pTau, but normal levels of tTau, had a “significantly higher risk of dementia.” These results from the study prove that neither pTau nor ABeta42/ABeta40 individually, on their own, have a significant effect on Alzheimer’s and cognitive function.

- Overall, our findings show that while pTau demonstrates a slightly stronger correlation with CERAD scores compared to ABeta42, both pTau and ABeta42 have a weak correlation and predictive value, accounting for only 10.6% and 7% of the variance, respectively. These results align with research conducted by Oxford University, which demonstrates that neither pTau nor ABeta42/ABeta40 alone significantly increases dementia risk or cognitive decline, but instead it is when both biomarkers are abnormally elevated that risk rises substantially. In all, this highlights that for Alzheimer’s and CERAD scores, no single biomarker can serve as a definitive predictor.
- When it comes to ethics, this raises important considerations in diagnosis and research. For diagnosis, we recommend not to rely too heavily on these biomarkers individually could lead to misdiagnosis of a patient since ABeta42 and pTau both demonstrate a weak correlation to CERAD scores. For example, a doctor may diagnose a patient with early-onset Alzheimer’s due to abnormally elevated levels of pTau compared to ABeta42, even though individually these two biomarkers do not have an effect on the progression of cognitive function or Alzheimer’s. Furthermore, in research, overlooking the lack of correlation between these two biomarkers individually on CERAD scores could lead to a waste of taxpayer money, time, resources, and potential harm to patients. An example of this would be if researchers disregarded the lack of correlation and proceeded with trials using grant funding to test medicine on patients, even though it would never lead to improvement in Alzheimer’s. Ultimately, while pTau and ABeta42 are important biomarkers for Alzheimer’s, individually they present a weak correlation to CERAD scores, which underscores the need for caution and ethical consideration when using them in both diagnosis and research.
- Lastly, to further our work, in the future we can recommend using a larger data set and removing outliers, such as the one patient who had ABeta42 levels of 1412 pg/ug, to see if this would have an effect on our results. Furthermore, while it was found that individually ABeta42 and pTau alone don’t have a significant correlation to CERAD scores, we could try looking into what ratio of pTau and ABeta42 can lead to an increase in CERAD scores or if it has any effects. On top of looking into the ratios of pTau and ABeta42 together, we can introduce subgroups, such as sex or education, to see if changes between males and females or varying levels of education can lead to higher ratios of pTau and ABeta42 and their effect on CERAD scores. Altogether, these approaches may lead to a more comprehensive understanding of how ABeta42 and pTau interactions and demographic factors may influence CERAD scores and Alzheimer’s risk.
- Works cited: Oberstein, Timo Jan et al. “Amyloid- β levels and cognitive trajectories in non-demented pTau181-positive subjects without amyloidopathy.” *Brain : a journal of neurology* vol. 145,11 (2022): 4032-4041. [doi:10.1093/brain/awac297](https://doi.org/10.1093/brain/awac297)

Project Notes

- In class 9/11: Worked on understanding Alzheimer’s, writing code to characterize the data (headers, number of rows, etc.)
- In class 9/16 (Addison absent due to illness): Worked on generating a bar graph to answer the big question. This graph showed the correlation between both ABeta42 and pTau with CERAD scores, based on linear regressions.

- In class 9/18: Presented our bar graph, directed to change it as we would be unable to perform a t-test. Learned about t-tests and ANOVA tests
- In class 9/23 (Marielle absent due to illness): Worked on generate scatter plot with linear regression
- In class 9/25: Worked on verifying the analysis with background research and solidifying scatter plot and linear regression.

Questions for our TA:

- Do you know more about how CERAD scores and how they are developed?
- Is it still not expected to have a collaborative Jupyter notebook, or just a file we share back and forth?