Module_1: *(Template)*

Team Members:

Mahsood Ali and Haley Cossman

Project Title:

The Effect of Education Level and Tau Pathology on Alzheimer's Development

Project Goal:

This project seeks to computationally and statistically analyze real patient data to gain a deeper understanding about the effect of education and Tau on Alzheimer's disease and to suggest opportunities for interventions that could be developed to prevent/treat/cure this disease within the US.

Disease Background:

- Prevalence & incidence:
  7.2 million Americans aged 65 and older (https://pmc.ncbi.nlm.nih.gov/articles/PMC12040760/). prevalence = 7.2/342 x 100% = 2.11%, incidence = 0.9/342 x 100% = 0.263%

- Economic burden: In the United States, the care for Alzheimer's is projected to cost $384 billion dollars in 2025. https://www.alz.org/alzheimers-dementia/facts-figures

- Risk factors (genetic, lifestyle): age, family history, genetics, physical activity, diet, smoking, drinking, amount of sleep, social isolation (https://www.nia.nih.gov/health/alzheimers-causes-and-risk-factors/what-causes-alzheimers-disease)

- Societal determinants: Education, sex, environment, and race are all factors that play into Alzheimer's disease. Adults aged 45 and older with worsening memory are more than twice more likely to be people without a high school diploma, this rate is the lowest for college graduates. Additionally, if someone is exposed to unsafe or unhealthy environments/housing this can negatively impact brain health which can increase the risk for Alzheimer's. Other risk factors can be based on discrimination as by 2060 the rise in Alzheimer's disease is led by minority populations being affected the most. Cases in Hispanics will increase seven times over today's estimate and Non-Hispanic Blacks

will increase four times over today's estimates. Additionally, women are almost two times more likely to be affected by Alzheimer's than men.

https://www.cdc.gov/alzheimers-dementia/php/sdoh/index.html
https://brainhealth.dc.gov/page/social-determinants-health-and-dementia

https://www.cdc.gov/alzheimers-dementia/php/sdoh/index.html
https://brainhealth.dc.gov/page/social-determinants-health-and-dementia

- Symptoms: memory loss, difficulty solving problems or completing tasks, confusion, vision changes, problems with speaking, poor judgement, mood/personality changes (https://www.alz.org/alzheimers-dementia/10_signs)

- Diagnosis: A healthcare professional or a doctor trained in brain conditions (neurologist/geriatrician) will review symptoms and medical history. The healthcare professional may interview a close family/friend and conduct several tests and a physical exam.

During an appointment, a healthcare professional looks at: Whether an individual has impaired memory or thinking skills, also known as cognitive skills. Whether an individual exhibits changes in personality or behaviors. The degree of an individual's memory or thinking impairment or changes. How thinking impairment affects an individual's ability to function in daily life. The cause of an invidual's symptoms. The healthacre professional may need lab tests, brain-imaging tests or detailed memory testing. These tests can provide useful information for a diagnosis. They also can rule out other conditions that cause similar symptoms. https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048075

- Standard of care treatments (& reimbursement): Cholinesterase inhibitors (increase cell to cell communicaiton and preserve chemical messanger depleted by Alzheimer's), Memantine (slows progression of symptoms), Leqembi/Kisunla (IV infusion, prevents clumping of amyloid plaques in the brain). Most of the pill options cost aorund $150/month, the IV infusions cost around$27,000-32,0000/year. (https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/drc-20350453)

- Disease progression & prognosis:

Early Stage: Individuals may still live independently, but begin to experience memory lapses, such as forgetting names, words, or objects, difficulty organizing, and noticeable challenges in work or social settings.

Middle Stage: The longest stage and charaterized by more pronounced dementia symptoms. Individuals may struggle with personal history, daily tasks, orientation, mood, and behavior. Assistance during this stage becomes crucial for clothing, hygiene, and safety as wandering, sleep changes, and personality shifts emerge.

Late Stage: Severe cognitive and physical decline. Individuals lose the ability to communicate, move independently, or recognize surroundings, requiring constant care. They become highly vulnerable to infections and other complications.

https://www.alz.org/alzheimers-dementia/stages

- Continuum of care providers: Neurologists, neuropsychologists, psychiatrists, and radiologists (https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/doctors-departments/ddc-20350457)

- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology):

Anatomical & Organ Changes: Alzheimer's primarily affects the hippocampus and cortex, leading to synapse loss, neuronal death, brain atrophy, and impaired blood–brain barrier function. Neuroinflammation from overactive glial cells further accelerates damage.

Cellular & Molecular Mechanisms: Abnormal buildup of amyloid-β plaques and tau tangles disrupts synaptic signaling, axonal transport, and communication between neurons. This drives excitotoxicity, mitochondrial dysfunction, oxidative stress, and loss of proteostasis.

Genetic & Immune Factors: Certain genes increase the risk of Alzheimer's by affecting how proteins are made, broken down, and cleared from the brain. Problems with microglia and astrocytes reduce the brain's ability to remove toxic proteins and trigger long-lasting inflammation, which further damages neurons.

https://pmc.ncbi.nlm.nih.gov/articles/PMC5290713/

- Clinical Trials/next-gen therapies: amyloid-targeting antibodies (Kisunla), therapies that target tau protein tangles and vaccines to reduce buildup, focused ultrasound to open the blood-brain barrier and reduce amyloid plaques, gene therapy to preserve cognitive function. (Google Gemini)

Data-Set:

We opened the dataset in Python and found that each row represents a single donor/patient. The columns have 60+ variables describing demographic info (sex, race, age at death), education (highest level, years of education), genetic factors (APOE), clinical features (cognitive status, age of onset, diagnosis), and neuropathology (Braak stage, CERAD score, etc.).

The data is based on postmortem brain donation data in which Donors had varying clinical histories, cognitive testing, and neuropathological staging. Most of the subjects in this study were older adults who consented to brain donation. Biases/limitations with the given data set: Donor-based datasets tend to overrepresent White, highly educated participants.

Education measured as years or highest level doesn't capture true cognitive ability of an individual.

Questions:

Q1 (ANOVA): Does the level of tau pathology (Braak stage) differ significantly between groups with different levels of education?

Q2 (Scatter): Is there a relationship between Braak stage and age of onset of cognitive symptoms?

Q3 (Cognitive reserve hypothesis): Do people with more years of education show later onset of dementia symptoms even at higher tau burden?

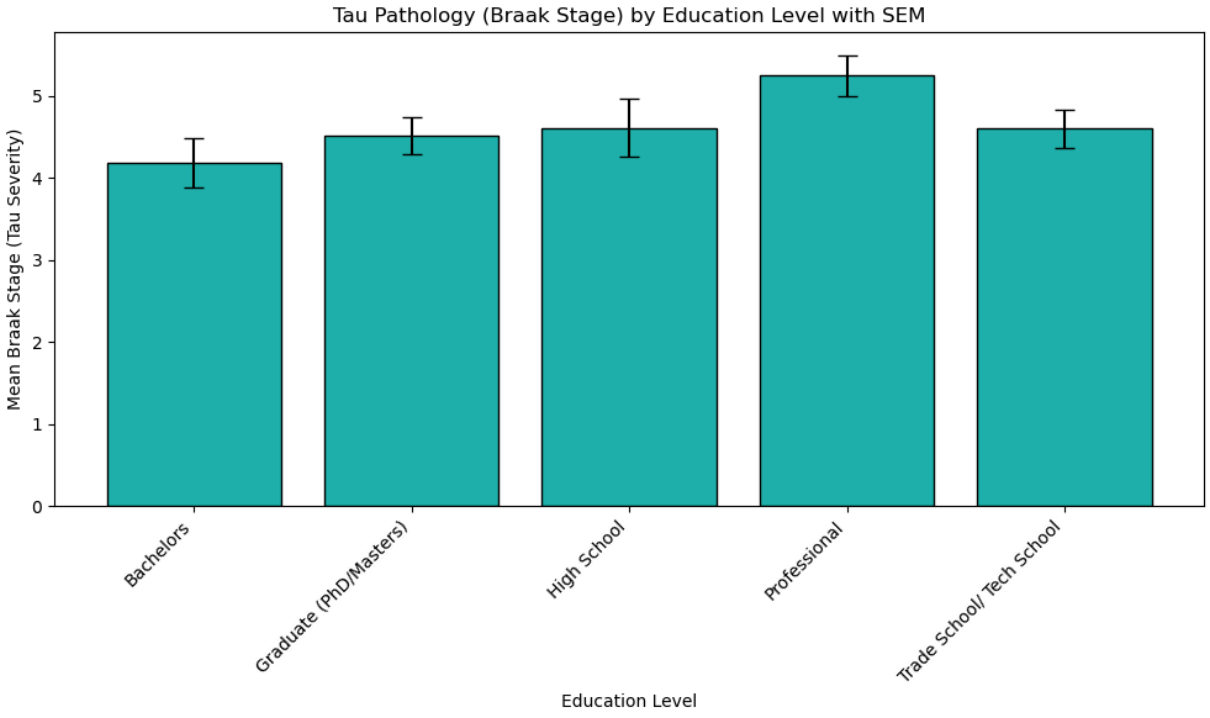
Data Analyis:

We analyzed the data to evaluate the relationship of education level, Tau pathology, and the age of Alzheimer's onset using graphical comparisons, linear regression, and statistical tests.

In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt

# load data set
file_path = "UpdatedMetaData.csv"
df = pd.read_csv(file_path)

# make columns for education and tau
edu_col = "Highest level of education"
tau_col = "Braak"

# drop rows with missing info
subset = df[[edu_col, tau_col]].dropna()

# convert roman numerial Braak stages to integers
roman_to_num = {
    "0": 0, "I": 1, "II": 2, "III": 3, "IV": 4, "V": 5, "VI": 6
}

# remove "Braak" and strip spaces
subset[tau_col] = subset[tau_col].astype(str).str.replace("Braak", "", regex=False)

# map roman numerials to integers
subset[tau_col] = subset[tau_col].map(roman_to_num)

subset = subset.dropna(subset=[tau_col])

# convert to float
subset[tau_col] = subset[tau_col].astype(float)

# group by education and calculate tau mean, count, and standard deviation
group_stats = subset.groupby(edu_col)[tau_col].agg(['mean', 'count', 'std'])
```

```
group_stats['sem'] = group_stats['std'] / group_stats['count']**0.5

# print stats
print("Group Statistics for Braak (Tau pathology) by Education:")
print(group_stats)

# plot bar graph error bars
plt.figure(figsize=(10,6))
plt.bar(group_stats.index, group_stats['mean'],
        yerr=group_stats['sem'], capsize=6,
        color="lightseagreen", edgecolor="black")

# add labels and title
plt.ylabel("Mean Braak Stage (Tau Severity)")
plt.xlabel("Education Level")
plt.title("Tau Pathology (Braak Stage) by Education Level with SEM")
plt.xticks(rotation=45, ha="right")
plt.tight_layout()
plt.show()
```

```
Group Statistics for Braak (Tau pathology) by Education:
                                mean  count       std       sem
Highest level of education
Bachelors                   4.181818     22  1.401916  0.298889
Graduate (PhD/Masters)      4.520000     25  1.122497  0.224499
High School                 4.611111     18  1.500545  0.353682
Professional                5.250000      4  0.500000  0.250000
Trade School/ Tech School   4.600000     15  0.910259  0.235028
```



In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

file_path = "UpdatedMetaData.csv"
df = pd.read_csv(file_path)
```

```
edu_col = "Highest level of education"
tau_col = "Braak"

subset = df[[edu_col, tau_col]].dropna()

roman_to_num = {
    "0": 0, "I": 1, "II": 2, "III": 3, "IV": 4, "V": 5, "VI": 6
}
subset[tau_col] = subset[tau_col].astype(str).str.replace("Braak", "", regex=False)
subset[tau_col] = subset[tau_col].map(roman_to_num)
subset = subset.dropna(subset=[tau_col])
subset[tau_col] = subset[tau_col].astype(float)

# group by education
groups = [group[tau_col].values for name, group in subset.groupby(edu_col)]

# one-way ANOVA
f_stat, p_val = stats.f_oneway(*groups)

# print results
print("ANOVA Results:")
print(f"F-statistic = {f_stat:.3f}, p-value = {p_val:.4f}")

# group by education and calculate tau mean, count, and standard deviation
group_stats = subset.groupby(edu_col)[tau_col].agg(['mean', 'count', 'std'])
group_stats['sem'] = group_stats['std'] / group_stats['count']**0.5
print("\nGroup statistics:")
print(group_stats)
```

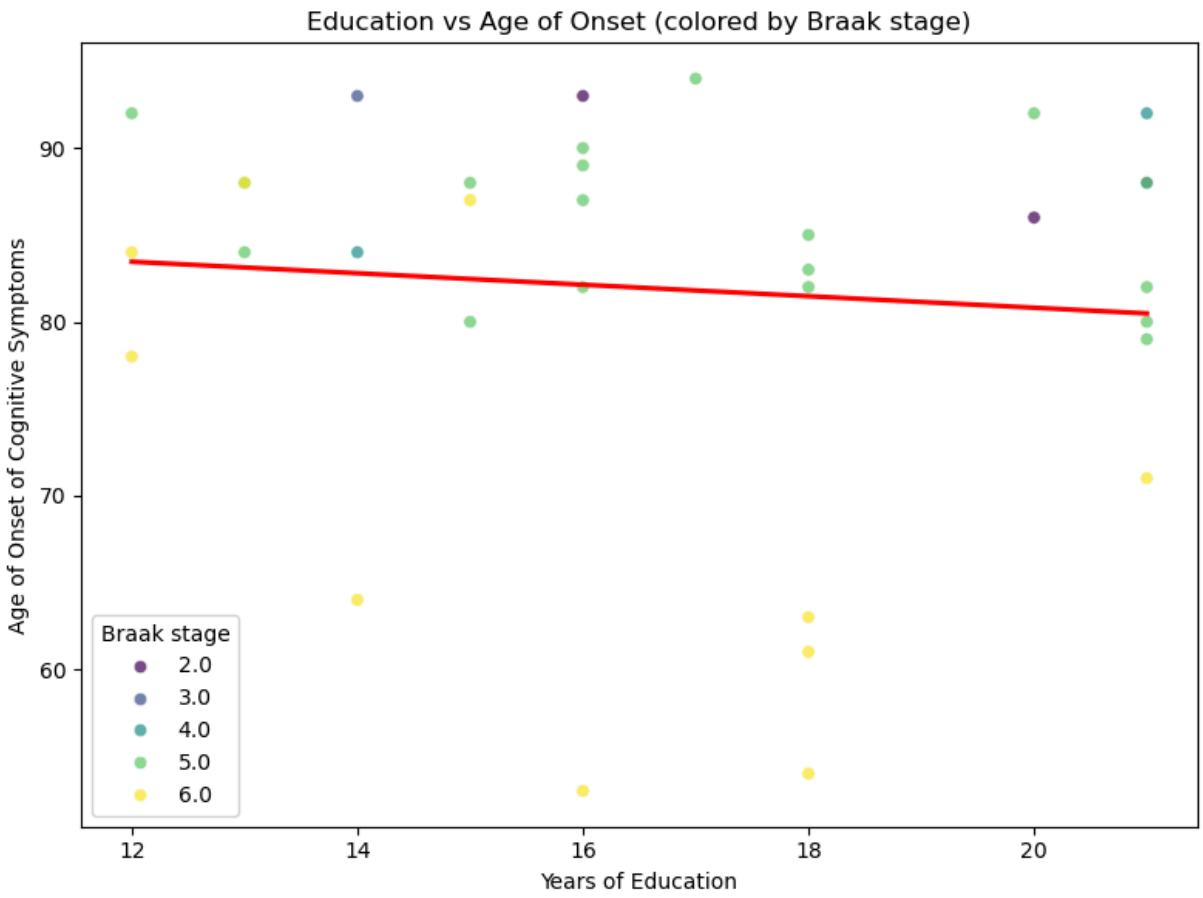
```
ANOVA Results:
F-statistic = 0.786, p-value = 0.5378

Group statistics:
                                mean  count       std       sem
Highest level of education
Bachelors                   4.181818     22  1.401916  0.298889
Graduate (PhD/Masters)      4.520000     25  1.122497  0.224499
High School                 4.611111     18  1.500545  0.353682
Professional                5.250000      4  0.500000  0.250000
Trade School/ Tech School   4.600000     15  0.910259  0.235028
```

In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file_path = "UpdatedMetaData.csv"
df = pd.read_csv(file_path)

# make columns
edu_years = "Years of education"
age_onset = "Age of onset cognitive symptoms"
age_dx = "Age of Dementia diagnosis"
tau_col = "Braak"

# drop empty rows
```

```
subset = df[[edu_years, age_onset, age_dx, tau_col]].dropna()

# map roman numerial Braak stage to integers
roman_to_num = {"0":0, "I":1, "II":2, "III":3, "IV":4, "V":5, "VI":6,
                "Braak 0":0, "Braak I":1, "Braak II":2, "Braak III":3,
                "Braak IV":4, "Braak V":5, "Braak VI":6}

# convert to float
subset[tau_col] = subset[tau_col].replace(roman_to_num).astype(float)

# make scatter plot
plt.figure(figsize=(8,6))
# points colored by tau level
sns.scatterplot(data=subset, x=edu_years, y=age_onset, hue=tau_col, palette="viridi

# add regression line
sns.regplot(data=subset, x=edu_years, y=age_onset, scatter=False, color="red", ci=N

# label axes
plt.xlabel("Years of Education")
plt.ylabel("Age of Onset of Cognitive Symptoms")
plt.title("Education vs Age of Onset (colored by Braak stage)")
plt.legend(title="Braak stage")
plt.tight_layout()
plt.show()
```

```
/var/folders/gk/wz_l2ywd5776gql6mzblr68c0000gn/T/ipykernel_44780/3741633008.py:22: F
utureWarning: Downcasting behavior in `replace` is deprecated and will be removed in
a future version. To retain the old behavior, explicitly call `result.infer_objects
(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silen
t_downcasting', True)`
  subset[tau_col] = subset[tau_col].replace(roman_to_num).astype(float)
```

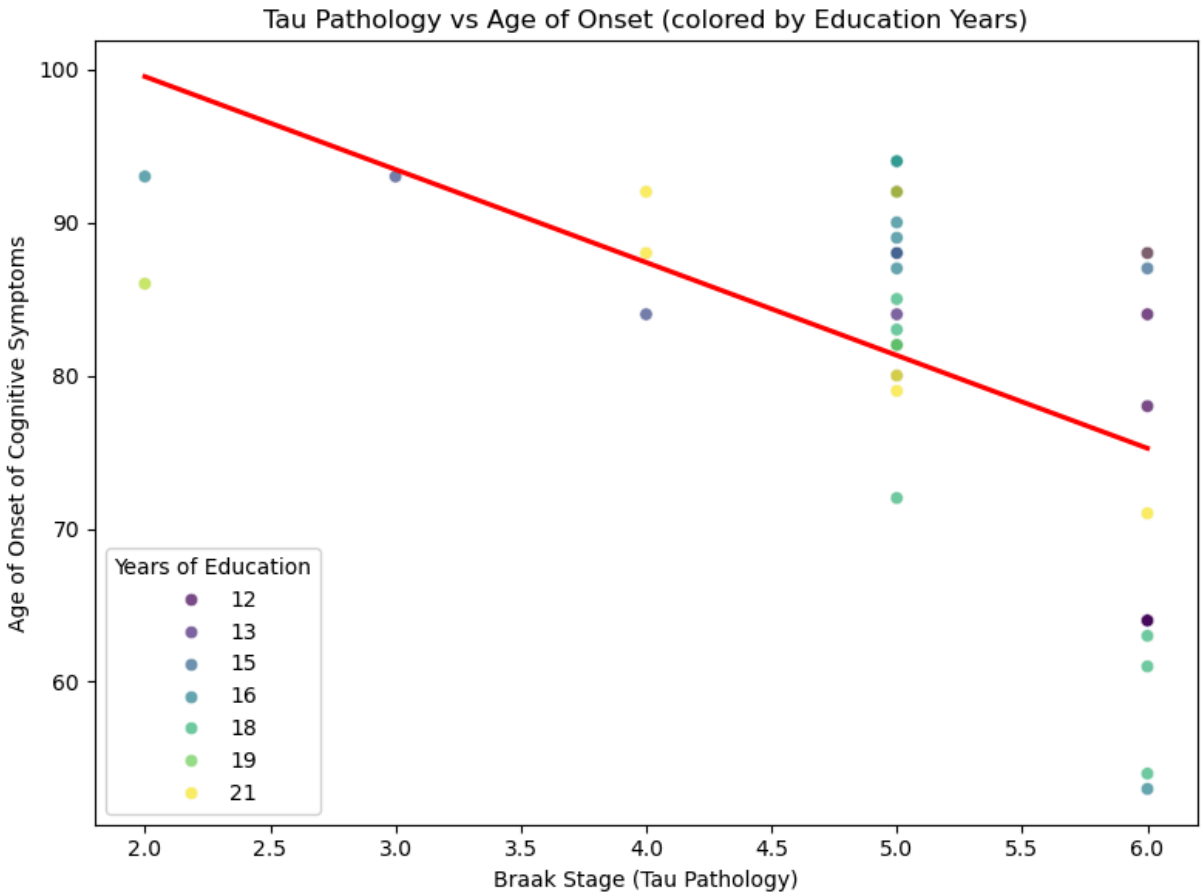


In [ ]:

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from scipy import stats

file_path = "UpdatedMetaData.csv"
df = pd.read_csv(file_path)

edu_years = "Years of education"
age_onset = "Age of onset cognitive symptoms"
tau_col = "Braak"

subset = df[[edu_years, age_onset, tau_col]].dropna()

roman_to_num = {
    "0": 0, "I": 1, "II": 2, "III": 3, "IV": 4, "V": 5, "VI": 6,
    "Braak 0": 0, "Braak I": 1, "Braak II": 2, "Braak III": 3,
    "Braak IV": 4, "Braak V": 5, "Braak VI": 6
}
subset[tau_col] = subset[tau_col].replace(roman_to_num).astype(float)

# make scatter plot
plt.figure(figsize=(8,6))
# points colored by education level
sns.scatterplot(
    data=subset, x=tau_col, y=age_onset,
    hue=edu_years, palette="viridis", alpha=0.7
```

```
)

# add regression line
sns.regplot(
    data=subset, x=tau_col, y=age_onset,
    scatter=False, color="red", ci=None
)

# label axes
plt.xlabel("Braak Stage (Tau Pathology)")
plt.ylabel("Age of Onset of Cognitive Symptoms")
plt.title("Tau Pathology vs Age of Onset (colored by Education Years)")
plt.legend(title="Years of Education")
plt.tight_layout()
plt.show()

# get x and y values
x = subset[tau_col].values
y = subset[age_onset].values

# fit linear regression model
model = LinearRegression()
model.fit(x.reshape(-1,1), y)
slope = model.coef_[0]
intercept = model.intercept_
r2 = model.score(x.reshape(-1,1), y)

# determine strength of relationship
corr, p_val = stats.pearsonr(x, y)

n = len(subset)

# print results
print("----- Regression & Correlation Results -----")
print(f"Sample size (n): {n}")
print(f"Slope: {slope:.3f}")
print(f"Intercept: {intercept:.3f}")
print(f"R²: {r2:.3f}")
print(f"Pearson correlation (r): {corr:.3f}")
print(f"P-value: {p_val:.4f}")
```

```
/var/folders/gk/wz_l2ywd5776gql6mzblr68c0000gn/T/ipykernel_92186/190824080.py:21: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
  subset[tau_col] = subset[tau_col].replace(roman_to_num).astype(float)
```




```
----- Regression & Correlation Results -----
Sample size (n): 38
Slope: -6.069
Intercept: 111.667
R²: 0.289
Pearson correlation (r): -0.538
P-value: 0.0005
```

Verify and validate your analysis:

According to research done by Fumihiko Yasuno, Hiroyuki Minami, and Hideyuki Hattori on the effect of education levels on tau development in Alzheimer's patients, higher education levels have a correlation to higher tau accumulation. People with higher education levels were able to tolerate more tau buildup before symptoms developed. These results are similar to our data because we found higher tau levels in more educated individuals, but roughly the same age of onset across all education levels. Both this research and our data support the cognitive reserve theory, which says that more educated people are able to sustain tau without cognitive decline for longer than less educated people. https://pubmed.ncbi.nlm.nih.gov/32285577/

Another study done by Yue Cai, Lili Fang, and Jie Yang found that higher education levels slowed tau buildup in early stages, but accelerated tau buildup in later stages (when amyloid-B appears). This suggests that higher education can initially help delay symptoms and Alzheimer's development, but has no effect as the disease progresses. These results

somewhat align with ours because we also found that more education can help hide initial symptoms, but overall has little to no effect in preventing Alzheimer's development. https://jamanetwork.com/journals/jamaneurology/article-abstract/2835855

Conclusions and Ethical Implications:

Figure 1: Bar Plot — Tau Pathology (Braak Stage) by Education Level

With this graph the observation we made was that the mean braak stage is lowest in Bachelors (~4.18) and Graduate (Masters/PhD, ~4.52) and Highest in the Professional group (5.25), but that group has only 4 people, which is a small sample compared to the other groups. Based on our analysis of this graph, there is no clear linear trend of education to lower tau pathology. Even with bachelors/Graduate groups showing the lowest means, the standard errors overlap heavily which is likely not statistically significant.

Figure 2: Scatterplot — Education (Years) vs Age of Onset (colored by Braak Stage)

In this figure the regression line slopes slightly downward which implies that the more years of education, the slightly earlier onset an individual may have. Additionally the figure shows a wide scatter which correlates to a weak relationship overall. Based on our analysis, there is no strong protective effect of education on onset age. Although this can show a possible cognitive reserve effect which is when highly educated people may mask symptoms longer, so once symptoms emerge, they emerge suddenly at higher pathology levels.

Figure 3: Scatterplot — Tau Pathology vs Age of Onset (colored by Education Years)

In this scatterplot we observed a stronger negative slope which means that the more tau accumulation an inivdiual has, the earlier onset of Alzheimer's for these individuals. With educationn included in the scatterplot, education years are spread across all pathology levels. Based on our analysis, Tau pathology is a strong predictor of onset age, but education years do not clearly hinder tau’s effect in this dataset.

Overall Analysis:

Tau pathology (Braak stage) is the strongest predictor for the age of onset in Alzheimer's and our data set showed that the higher tau burden, the earlier symptoms may appear for the individual. Education effects on Tau accumulation and the age of onset for Alzheimer's show a weak correlation and possibly shows that more years of education are linked with earlier onset, which could be due to cognitive reserve masking.

Ethical:

Although we studyed various differences in education, this does not mean there should be a difference in medical treatment for Alzheimer's based on education levels. Each paitent should cared for with equal effort and be made a priority as Alzheimer's is a serious

condition. Additionally, based on this data set, continuing research would help get a better understanding of the effect of education on tau accumulation, although going into clincial research for this topic does not beg a great importance. Lastly, with the implications of this research, education does not enable or protect an inidviudal from Alzheimer's and, as a result, inidviduals should not be shamed for either having or not having education.

Team Notes:

During the last week, we made two graphs comparing education levels on Alzheimer's development, one by development percent and another by number of cases. After making these graphs, we realized we would not be able to perform data analysis because we could not make error bars for our graphs. Due to this, we decided to change our graph to compare education levels on Tau severity. We then ran an ANOVA statistical test to determine whether there is a significant difference between at least one of the groups. The test resulted in a p-value of 0.5378, which is larger than the alpha value of 0.05, meaning we fail to reject the null hypothesis because the difference between groups is insignificant. We then ran two linear regression tests to see if there was a correlation between Braak state and age of onset or age of diagnosis of Alzheimer's, and we found that higher tau levels caused an earlier age of onset.

Limitations and Future Work:

Our analysis was limited by our unequal and somewhat small data set, which made it difficult to draw significant conclusions. In the professional education group, there were only four participants, causing skewed results that are easily influenced by outliers. Additionally, having data on tau accumulation over time, rather than a single value, would help us determine whether the cognitive reserve theory applies and clarify the effect of education levels on the onset of Alzheimer's.

Future work should focus on tracking long-term tau development over a larger sample size with equal representation from all education levels. This could help researchers continue to study the effect of education on Alzheimer's onset and look more closely at whether higher education helps delay symptom progression to a certain point. This research could also lead to the development of treatment plans or interventions for patients based on education levels.

Ethically, our conclusions suggest that higher education doesn't prevent Alzheimer's progression, but it can delay the appearance of symptoms, showing the importance of early Alzheimer's screening across all education levels. If more educated individuals are less likely to be diagnosed until later progression, they are susceptible to more sudden cognitive decline and lose the opportunity to benefit from treatments that are most effective during early stages. Public health messaging must avoid overstating the role of education in

preventing Alzheimer's because this misinformation could mislead individuals into delaying testing and increase their risk of late diagnosis.