Module_1: Alzheimer's

Team Members:

Will R, Trey H

Project Title:

Head Injury Incidence and Amyloid Beta in Alzheimers

Project Goal:

This project seeks to examine the relationship between head injury history and Amyloid Beta 40 and 42 levels in the brain in Alzheimer's Disease. Increased levels of AB-42 and AB-40 in Luminex assays and a decreased ratio of AB42:AB40 has been correlated with a higher risk of Alzheimer's, so we aim to determine whether head injury may represent a potential risk factor for the disease through AB concentrations within our dataset.

Disease Background:

Fill in information about 11 bullets:

- Prevalence & incidence
  - Over 7 million Americans live with Alzheimers (13 million by 2050)
  - 1 in 9 (11%) people aged 65 and older has Alzheimers 110 of every 100,000 Americans (~200,000) aged 30-64 are diagnosed with alzheimers (https://www.alz.org/alzheimers-dementia/facts-figures)
  - Incidence : 500,000 new cases projected to be diagnosed this year (https://www.alzsd.org/resources/facts-stats/)
- Economic burden
  - Alzheimer's-related care costs $412 billion in 2025 and will surpass 1 trillion by 2050, with Medicare and Medicaid covering 75% of expenses, making the disease an urgent economic challenge. (https://nchstats.com/alzheimers-disease-in-the-us/)
  - In 2024, unpaid caregivers provided care valued at over 413 billion (https://www.alz.org/alzheimers-dementia/facts-figures)
- Risk factors (genetic, lifestyle)
  - 2/3 of americans with alzheimers are women

    - Older Black Americans 2x likely to have Alzheimers and other dementiascompared to older whites
    - Older hispanics about 1 1/2x as likely to have Alzheimers or other dementias as older whites
    - https://www.alz.org/alzheimers-dementia/facts-figures
- Societal determinants
    - Neighborhood poverty levels increases dementia incidence
    - Lower income is associated with higher dementia risk
    - Fewer years of formal education are linked to higher Alzheimer's risk.
    - https://www.alz.org/getmedia/bbdf2c43-90c7-4f81-8d50-b49567553592/sdoh-dementia-risk-compiled-report.pdf
- Symptoms
    - Difficulty remembering recent conversations, names or events; apathy; and depression are often early symptoms Communication problems, confusion, poor judgment and behavioral changes may occur next. Difficulty walking, speaking and swallowing are common in the late stages of the disease.(Alzheimer's Association 2025 Facts and Figures Report)
- Diagnosis
    - Past Medical History interview with patient and family (health, meds, daily activities, behavior)
    - Cognitive testing (memory attention, language, problem solving)
    - Lab tests (blood, urine)
    - Psychiatric evaluation
    - CSF Test (spinal tap for Alzheimer's proteins like amyloid and Tau)
    - Brain imaging (CT, MRI, PET)
    - follwo up every 6-12 months
    - https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet#:~:text=Collect%20cerebrospinal%20fluid%20via%20a,other%20possible%20caus
- Standard of care treatments (& reimbursement)
    - Most drugs work best for early-middle stage Alzheimers. Cholinterase inhibitors (ChEIs) - prevent the breakdown of acetylcholine, a brain chemical believed to be important for memory and thinking. Donepezil, Galantamine, Rivastigmine, etc. Memantine (NMDA receptor antagonist) (Moderate-severe Alzheimers) (https://www.nia.nih.gov/health/alzheimers-treatment/how-alzheimers-disease-treated)
- Disease progression & prognosis
    - Average prognosis is 3-11 years, 6th leading cause of death Generally 5 stages: Preclinical Alzheimer's disease. Only identified in research settings Mild cognitive impairment due to Alzheimer's disease. Not enough to affect work or relationships

Trouble with decision making, estimating time to complete a task, etc. Mild dementia due to Alzheimer's disease. Memory loss of recent events, changes in personality, getting lost, etc. Moderate dementia due to Alzheimer's disease. All of previous symptoms magnified Severe dementia due to Alzheimer's disease. Lose the ablity to communicate, daily assistance needed (https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448) Mortality rate: 34.0 per 100,000 (6th leading cause of death) (https://www.cdc.gov/nchs/fastats/alzheimers.htm#:~:text=*%20Number%20of%20dea

- Continuum of care providers
    - Early Stage Primary care physician Neurologist/geriatrician Neurophysiologist Informal caregivers Social workers Middle Stage Home health aides Respite care Occupational Therapist Late Stage Skilled nursing facilities Hospice care (https://www.alz.org/alzheimers-dementia/stages#:~:text=Alzheimer's%20disease%20typically%20progresses%20slowly (https://www.alzheimers.gov/professionals/health-care-providers) (https://cconpalm.com/understanding-the-different-types-of-memory-care-services/)
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)
    - Extracellular accumulation of amyloid-beta plaques and the intracellular buildup of neurofibrillary tangles (NFTs) Brain atrophy (shrinkage) Starts in hippocampus and entorhinal cortex, spreads to cerebral cortex Enlarges ventricles in brain Decreased glucose metabolism
- Clinical Trials/next-gen therapies
    - anti-amyloid monoclonal antibodies like donanemab
    - gene therapies using CRISPR technology
    - oral medications
    - GLP-1 agonists like semaglutide https://www.brightfocus.org/resource/whats-next-for-alzheimers-disease-treatments-a-2024-forecast/#:~:text=Donanemab%20made%20headlines%20in%202023,an%20easily%20a

Data-Set:

This data set comes from a 2024 study, "Integrated multimodal cell atlas of Alzheimer's disease" (Hawrylycz, et al.). The study uses a large, multimodal, atlas of Alzheimer's disease built from human postmortem brain tissue. Donors (n=84; ≥65 years, mean ≈88; 51 female/33 male) were drawn from the ACT study and the UW Alzheimer's Disease Research Center. The code below generates a summary table of the data.

In [7]:
```python
import pandas as pd

# --- Load datasets using with open() ---


with open("NO DATE GENOTYPE Metadata.csv", "r") as f:
    metadata_df = pd.read_csv(f)

with open("UpdatedLuminex.csv", "r") as f:
    luminex_df = pd.read_csv(f)

# --- Step 1: Clean categorical columns (Checked/Unchecked -> 1/0) ---
for col in metadata_df.columns:
    if metadata_df[col].astype(str).str.contains("Checked|Unchecked").any():
        metadata_df[col] = metadata_df[col].map({"Checked": 1, "Unchecked": 0})

# --- Step 2: Merge datasets on Donor ID ---
merged_df = pd.merge(metadata_df, luminex_df, on="Donor ID", how="inner")

# --- Step 3: Convert numeric columns properly ---
for col in merged_df.columns:
    merged_df[col] = pd.to_numeric(merged_df[col], errors="ignore")

# --- Step 4: Summary statistics ---
summary = merged_df.describe(include="all").transpose()

# --- Step 5: Save cleaned dataset using with open() ---
output_path = "Cleaned_Merged_Data.csv"
with open(output_path, "w") as f:
    merged_df.to_csv(f, index=False)

# --- Step 6: Display previews ---
print("Merged dataset preview:")
print(merged_df.head())
print("\nSummary statistics:")
print(summary.head(15))
```

```
Merged dataset preview:
     Donor ID Primary Study Name Secondary Study Name  Age at Death     Sex  \
0  H19.33.004                ACT                  NaN            80  Female
1  H20.33.001                ACT                  NaN            82    Male
2  H20.33.002                ACT                  NaN            97  Female
3  H20.33.004                ACT                  NaN            86    Male
4  H20.33.005                ACT                  NaN            99  Female

   Race (choice=White)  Race (choice=Black/ African American)  \
0                    1                                     0
1                    1                                     0
2                    1                                     0
3                    1                                     0
4                    1                                     0

   Race (choice=Asian)  Race (choice=American Indian/ Alaska Native)  \
0                    0                                            0
1                    0                                            0
2                    0                                            0
3                    0                                            0
4                    0                                            0

   Race (choice=Native Hawaiian or Pacific Islander)  ...  \
0                                                  0  ...
1                                                  0  ...
2                                                  0  ...
3                                                  0  ...
4                                                  0  ...

   Total microinfarcts in screening sections  Atherosclerosis  \
0                                          1             Mild
1                                          0             Mild
2                                          0         Moderate
3                                          0             Mild
4                                          2             Mild

  Arteriolosclerosis            LATE   RIN  Severely Affected Donor  \
0           Moderate  Not Identified  8.33                      NaN
1               Mild    LATE Stage 2  8.60                      NaN
2           Moderate    LATE Stage 2  7.87                      NaN
3             Severe    LATE Stage 1  7.83                      NaN
4           Moderate    LATE Stage 1  8.40                      NaN

  ABeta40 pg/ug ABeta42 pg/ug   tTAU pg/ug  pTAU pg/ug
0      0.019621      0.971579  1552.414737    1.901053
1      0.215789      2.744211   756.090526    2.737895
2      0.000598      0.147158   313.525263    2.615789
3     60.766316     80.266316   318.528421    7.412632
4      5.136842     16.156842   107.348421    1.327368

[5 rows x 70 columns]

Summary statistics:
                                                count unique  \
Donor ID                                           84     84
Primary Study Name                                 84      2
```

Secondary Study Name	6	2
Age at Death	84.0	NaN
Sex	84	2
Race (choice=White)	84.0	NaN
Race (choice=Black/ African American)	84.0	NaN
Race (choice=Asian)	84.0	NaN
Race (choice=American Indian/ Alaska Native)	84.0	NaN
Race (choice=Native Hawaiian or Pacific Islander)	84.0	NaN
Race (choice=Unknown or unreported)	84.0	NaN
Race (choice=Other)	84.0	NaN
specify other race	3	1
Hispanic/Latino	84	3
Highest level of education	84	5

	top	\
Donor ID	H19.33.004	
Primary Study Name	ACT	
Secondary Study Name	ADRC Clinical Core	
Age at Death	NaN	
Sex	Female	
Race (choice=White)	NaN	
Race (choice=Black/ African American)	NaN	
Race (choice=Asian)	NaN	
Race (choice=American Indian/ Alaska Native)	NaN	
Race (choice=Native Hawaiian or Pacific Islander)	NaN	
Race (choice=Unknown or unreported)	NaN	
Race (choice=Other)	NaN	
specify other race	Mixed	
Hispanic/Latino	No	
Highest level of education	Graduate (PhD/Masters)	

	freq	mean	std	\
Donor ID	1	NaN	NaN	
Primary Study Name	69	NaN	NaN	
Secondary Study Name	4	NaN	NaN	
Age at Death	NaN	88.72619	8.02909	
Sex	51	NaN	NaN	
Race (choice=White)	NaN	0.964286	0.186691	
Race (choice=Black/ African American)	NaN	0.0	0.0	
Race (choice=Asian)	NaN	0.035714	0.186691	
Race (choice=American Indian/ Alaska Native)	NaN	0.011905	0.109109	
Race (choice=Native Hawaiian or Pacific Islander)	NaN	0.0	0.0	
Race (choice=Unknown or unreported)	NaN	0.0	0.0	
Race (choice=Other)	NaN	0.035714	0.186691	
specify other race	3	NaN	NaN	
Hispanic/Latino	82	NaN	NaN	
Highest level of education	25	NaN	NaN	

	min	25%	50%	75%	\
Donor ID	NaN	NaN	NaN	NaN	
Primary Study Name	NaN	NaN	NaN	NaN	
Secondary Study Name	NaN	NaN	NaN	NaN	
Age at Death	65.0	83.75	90.0	94.0	
Sex	NaN	NaN	NaN	NaN	
Race (choice=White)	0.0	1.0	1.0	1.0	
Race (choice=Black/ African American)	0.0	0.0	0.0	0.0	

```
Race (choice=Asian)                                  0.0    0.0   0.0   0.0
Race (choice=American Indian/ Alaska Native)         0.0    0.0   0.0   0.0
Race (choice=Native Hawaiian or Pacific Islander)    0.0    0.0   0.0   0.0
Race (choice=Unknown or unreported)                  0.0    0.0   0.0   0.0
Race (choice=Other)                                  0.0    0.0   0.0   0.0
specify other race                                   NaN    NaN   NaN   NaN
Hispanic/Latino                                      NaN    NaN   NaN   NaN
Highest level of education                           NaN    NaN   NaN   NaN

                                                     max
Donor ID                                             NaN
Primary Study Name                                   NaN
Secondary Study Name                                 NaN
Age at Death                                       102.0
Sex                                                  NaN
Race (choice=White)                                  1.0
Race (choice=Black/ African American)                0.0
Race (choice=Asian)                                  1.0
Race (choice=American Indian/ Alaska Native)         1.0
Race (choice=Native Hawaiian or Pacific Islander)    0.0
Race (choice=Unknown or unreported)                  0.0
Race (choice=Other)                                  1.0
specify other race                                   NaN
Hispanic/Latino                                      NaN
Highest level of education                           NaN
```

```
C:\Users\wcrou\AppData\Local\Temp\ipykernel_25708\3493971471.py:22: FutureWarning: e
rrors='ignore' is deprecated and will raise in a future version. Use to_numeric with
out passing `errors` and catch exceptions explicitly instead
  merged_df[col] = pd.to_numeric(merged_df[col], errors="ignore")
```

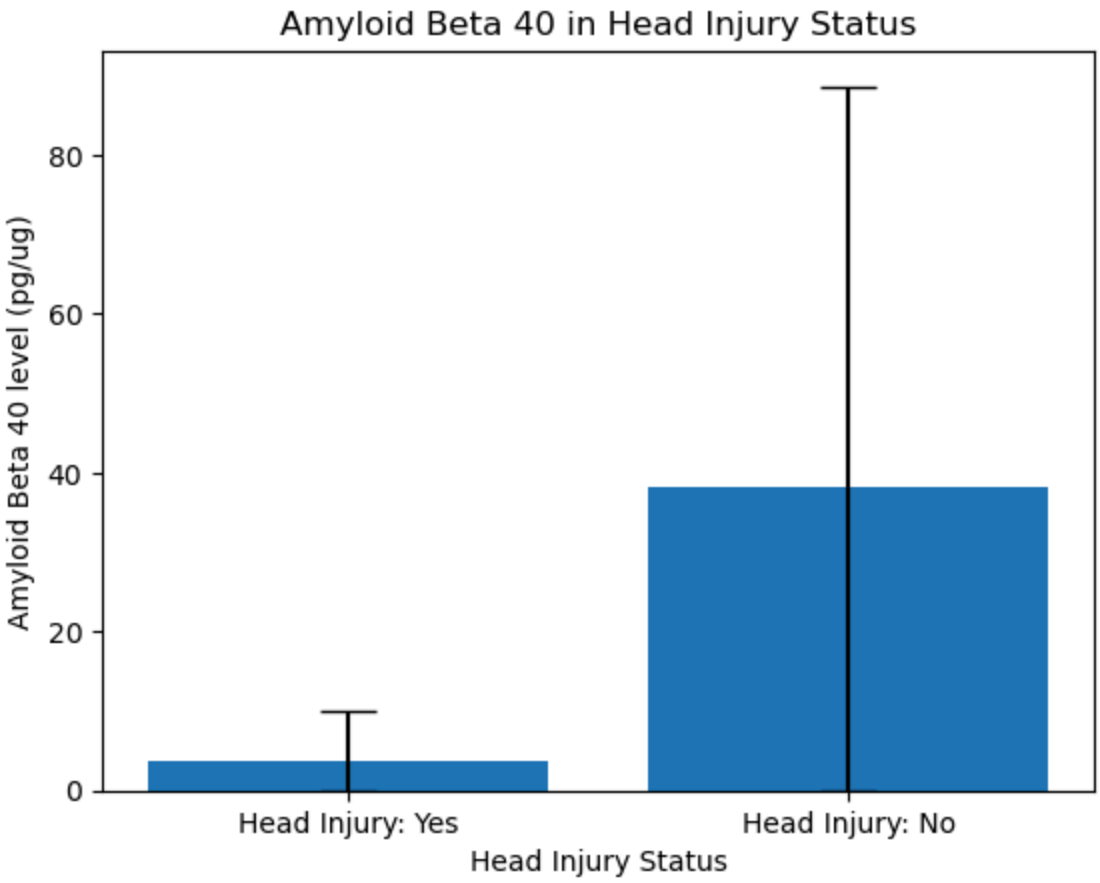
Data Analyis:

This code performs a t-test between known head injury data and amyloid beta 40 levels. First it calculates mean at standard deviations to perform a t-test, and finally plots a bar chart with our data. Though significant, the results are inconclusive because the errors overlap. However, the means disprove our hypothesis that head injuries increase AB-40 levels in Alzheimer's.

In [ ]:
```python
from matplotlib import pyplot as plt
from scipy import stats
import numpy as np
import statistics

# Column names
HI_COL = "Known head injury"
AB40_COL = "ABeta40 pg/ug"

# Create AB40 column
ab40 = pd.to_numeric(merged_df[AB40_COL], errors='coerce')

# Build value lists for each group (Yes / No)
with_injury_vals = merged_df.loc[merged_df[HI_COL] == "Yes", AB40_COL].tolist()
without_injury_vals = merged_df.loc[merged_df[HI_COL] == "No",  AB40_COL].tolist()
```

```
# Means and stdevs
x_with_bar = statistics.mean(with_injury_vals)
x_without_bar = statistics.mean(without_injury_vals)
stdev_with = statistics.stdev(with_injury_vals) if len(with_injury_vals) > 1 else 0
stdev_without = statistics.stdev(without_injury_vals) if len(without_injury_vals) >

labels = ["Head Injury: Yes", "Head Injury: No"]
mean = [x_with_bar, x_without_bar]
stdev = [stdev_with, stdev_without]

yerr = [mean, stdev]

# T-test (independent samples)
t_stat, p_val = stats.ttest_ind(with_injury_vals, without_injury_vals, equal_var=Fa
print("t-statistic:", t_stat)
print("p-value:", p_val)
print("st devs:", stdev)

alpha = 0.05
print("Statistically significant at α = 0.05." if p_val < alpha else "Not statistic

# Bar chart
plt.bar(labels, mean, yerr=yerr, capsize=10)
plt.title("Amyloid Beta 40 in Head Injury Status")
plt.xlabel("Head Injury Status")
plt.ylabel("Amyloid Beta 40 level (pg/ug)")
plt.show()
```

```
t-statistic: -2.7813475068574016
p-value: 0.012890801846914879
st devs: [6.2311247030273815, 50.47182455571414]
Statistically significant at α = 0.05.
```

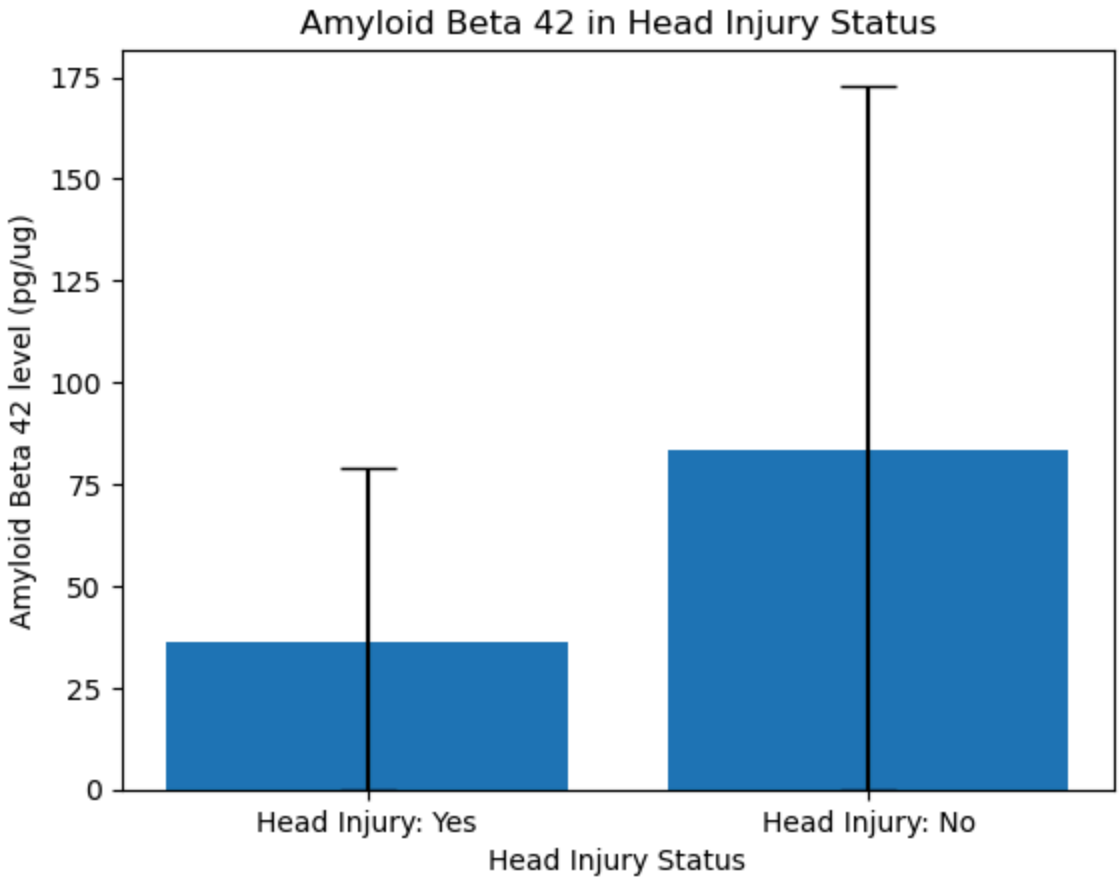


The following code performs a t-test between known head injury data and amyloid beta 42 levels. Though insignificant, the findings also disprove our hypothesis that head injuries increase amyloid beta in the brain by showing that AB-42 is lower in subjects with head injuries than those without.

In [21]:
```
# Column names
HI_COL = "Known head injury"
AB42_COL = "ABeta42 pg/ug"

# Create AB42 columns
ab42 = pd.to_numeric(merged_df[AB42_COL], errors='coerce')

# Build value lists for each group (Yes / No)
with_injury_vals = merged_df.loc[merged_df[HI_COL] == "Yes", AB42_COL].tolist()
without_injury_vals = merged_df.loc[merged_df[HI_COL] == "No",  AB42_COL].tolist()

# Means and stdevs
x_with_bar = statistics.mean(with_injury_vals)
x_without_bar = statistics.mean(without_injury_vals)
stdev_with = statistics.stdev(with_injury_vals) if len(with_injury_vals) > 1 else 0
stdev_without = statistics.stdev(without_injury_vals) if len(without_injury_vals) >

labels = ["Head Injury: Yes", "Head Injury: No"]
mean = [x_with_bar, x_without_bar]
stdev = [stdev_with, stdev_without]

yerr = [mean, stdev]
```

```
# T-test (independent samples)
t_stat, p_val = stats.ttest_ind(with_injury_vals, without_injury_vals, equal_var=Fa
print("t-statistic:", t_stat)
print("p-value:", p_val)
print("st devs:", stdev)

alpha = 0.05
print("Statistically significant at α = 0.05." if p_val < alpha else "Not statistic

# Bar chart
plt.bar(labels, mean, yerr=yerr, capsize=10)
plt.title("Amyloid Beta 42 in Head Injury Status")
plt.xlabel("Head Injury Status")
plt.ylabel("Amyloid Beta 42 level (pg/ug)")
plt.show()
```

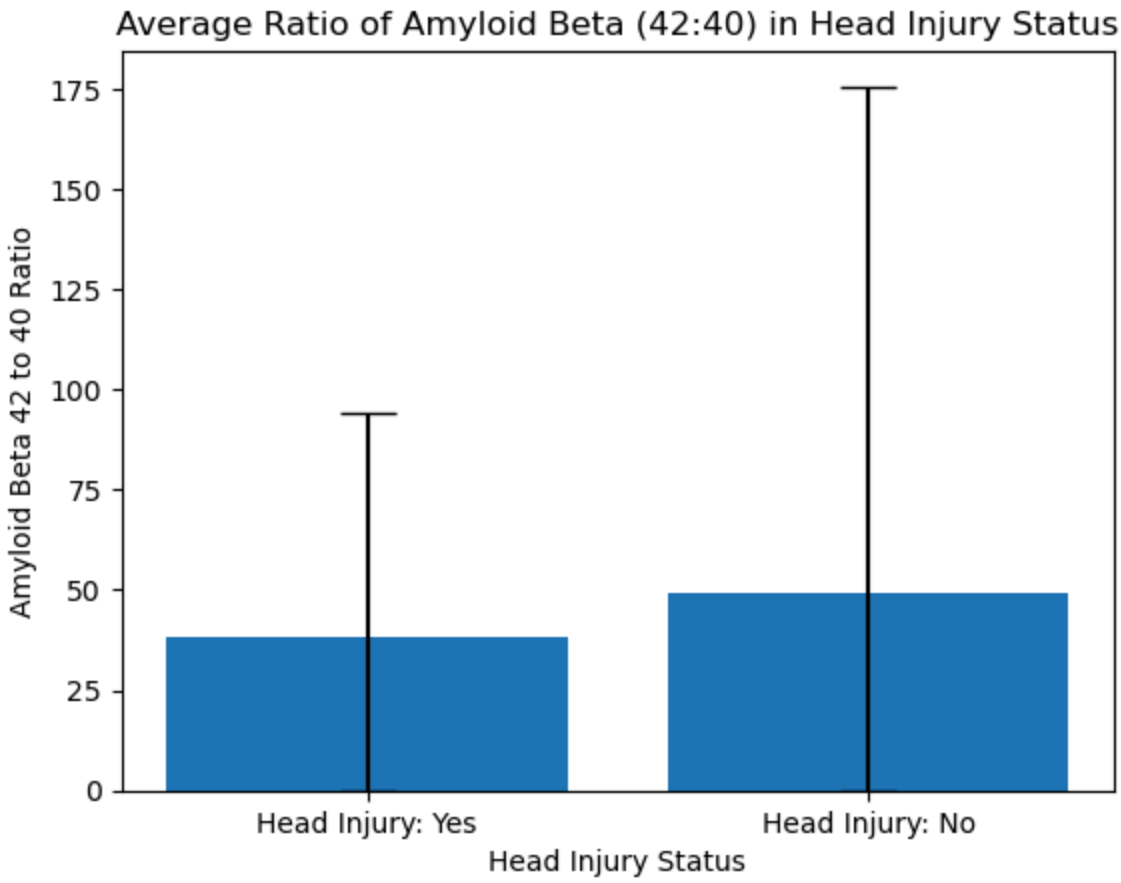
```
t-statistic: -1.8422292185442755
p-value: 0.07767474240814196
st devs: [42.72679311339802, 89.57766241741993]
Not statistically significant at α = 0.05.
```



The following code performs a t-test between known head injury data and the ratio between amyloid beta 42 and 40. The results are not conclusive because the ranges overlap, though they slightly support our hypothesis that d injuries decrease the ration between akyloid beta 42 to 40. These high standard deviations are because of the crazy outliers in the data, where some ratios are 1000:1.

In [22]:
```
# Column names
HI_COL = "Known head injury"
RATIO_COL = "AB42_40"

# Calculate AB42/AB40 ratio and add as new column
merged_df['AB42_40'] = ab42 / ab40.replace({0: np.nan})

# Build value lists for each group (Yes / No)
with_injury_vals = merged_df.loc[merged_df[HI_COL] == "Yes", RATIO_COL].tolist()
without_injury_vals = merged_df.loc[merged_df[HI_COL] == "No",  RATIO_COL].tolist()

# Means and stdevs
x_with_bar = statistics.mean(with_injury_vals)
x_without_bar = statistics.mean(without_injury_vals)
stdev_with = statistics.stdev(with_injury_vals) if len(with_injury_vals) > 1 else 0
stdev_without = statistics.stdev(without_injury_vals) if len(without_injury_vals) >

labels = ["Head Injury: Yes", "Head Injury: No"]
mean = [x_with_bar, x_without_bar]
stdev = [stdev_with, stdev_without]

yerr = [mean, stdev]

# T-test (independent samples)
t_stat, p_val = stats.ttest_ind(with_injury_vals, without_injury_vals, equal_var=Fa
print("t-statistic:", t_stat)
print("p-value:", p_val)
print("st devs:", stdev)

alpha = 0.05
print("Statistically significant at α = 0.05." if p_val < alpha else "Not statistic

# Bar chart
plt.bar(labels, mean, yerr=yerr, capsize=10)
plt.title("Average Ratio of Amyloid Beta (42:40) in Head Injury Status")
plt.xlabel("Head Injury Status")
plt.ylabel("Amyloid Beta 42 to 40 Ratio")
plt.show()
```

```
t-statistic: -0.32055458285916405
p-value: 0.7513505875699449
st devs: [55.8820785165434, 126.13169801820287]
Not statistically significant at α = 0.05.
```

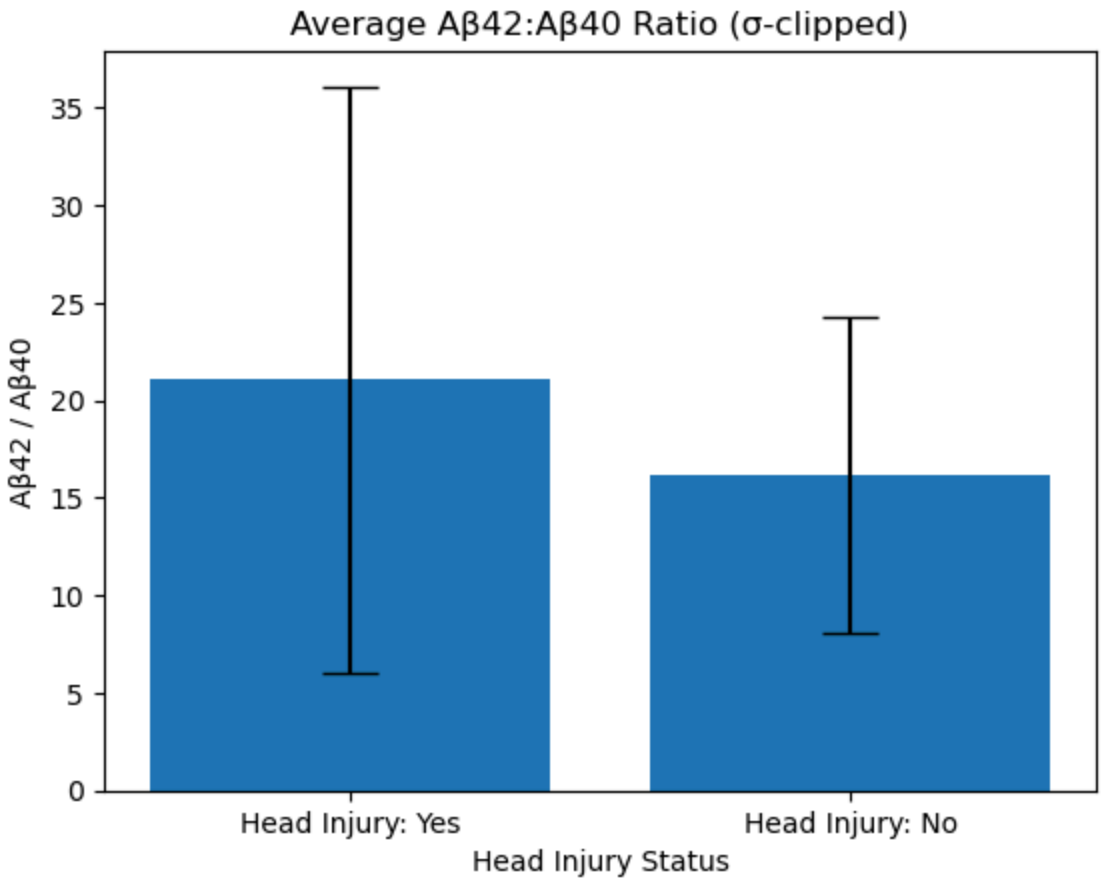


The following code also performs a t-test between known head injury data and the ratio between amyloid beta 42 and 40, but it first removes outliers from the data. ChatGPT was used to learn that the simplest way to remove outliers was to ignoring any data points 3 SDs above and below the mean (sigma clipping). Though the results are still insignificant and inconclusive, the SDs are much lower and the chart does not align with our hypothesis that head injuries decrease the ratio between AB 42 and 40.

In [ ]:
```
import statistics

# Parse to numeric
ab40 = pd.to_numeric(merged_df[AB40_COL], errors='coerce')
ab42 = pd.to_numeric(merged_df[AB42_COL], errors='coerce')

# Sigma clip AB40
_, lo40, hi40 = stats.sigmaclip(ab40.dropna(), low=3.0, high=3.0)
ab40_clip = ab40.where(ab40.between(lo40, hi40))

# Sigma clip AB42
_, lo42, hi42 = stats.sigmaclip(ab42.dropna(), low=3.0, high=3.0)
ab42_clip = ab42.where(ab42.between(lo42, hi42))

# Ratio (42/40), then sigma clip the ratio too
ratio = ab42_clip / ab40_clip.replace({0: np.nan})
_, loR, hiR = stats.sigmaclip(ratio.dropna(), low=3.0, high=3.0)
merged_df[RATIO_COL] = ratio.where(ratio.between(loR, hiR))
```

```
# Build value lists for each group (drop NaNs from clipping)
with_injury_vals    = merged_df.loc[merged_df[HI_COL] == "Yes", RATIO_COL].dropna()
without_injury_vals = merged_df.loc[merged_df[HI_COL] == "No",  RATIO_COL].dropna()

# Means and stdevs
x_with_bar = statistics.mean(with_injury_vals) if with_injury_vals else float("nan"
x_without_bar = statistics.mean(without_injury_vals) if without_injury_vals else fl
stdev_with = statistics.stdev(with_injury_vals) if len(with_injury_vals) > 1 else 0
stdev_without = statistics.stdev(without_injury_vals) if len(without_injury_vals) >

labels = ["Head Injury: Yes", "Head Injury: No"]
means = [x_with_bar, x_without_bar]
stdev = [stdev_with, stdev_without]

yerr = stdev

# T-test (independent samples)
t_stat, p_val = stats.ttest_ind(with_injury_vals, without_injury_vals, equal_var=Fa
print("t-statistic:", t_stat)
print("p-value:", p_val)
print("st devs:", stdev)

# Bar chart
plt.bar(labels, means, yerr=yerr, capsize=10)
plt.title("Average Aβ42:Aβ40 Ratio (σ-clipped)")
plt.xlabel("Head Injury Status")
plt.ylabel("Aβ42 / Aβ40")
plt.show()
```

```
t-statistic: 0.8046243338149376
p-value: 0.4359056955635303
st devs: [15.042473241955296, 8.096966450386093]
```



In [ ]:

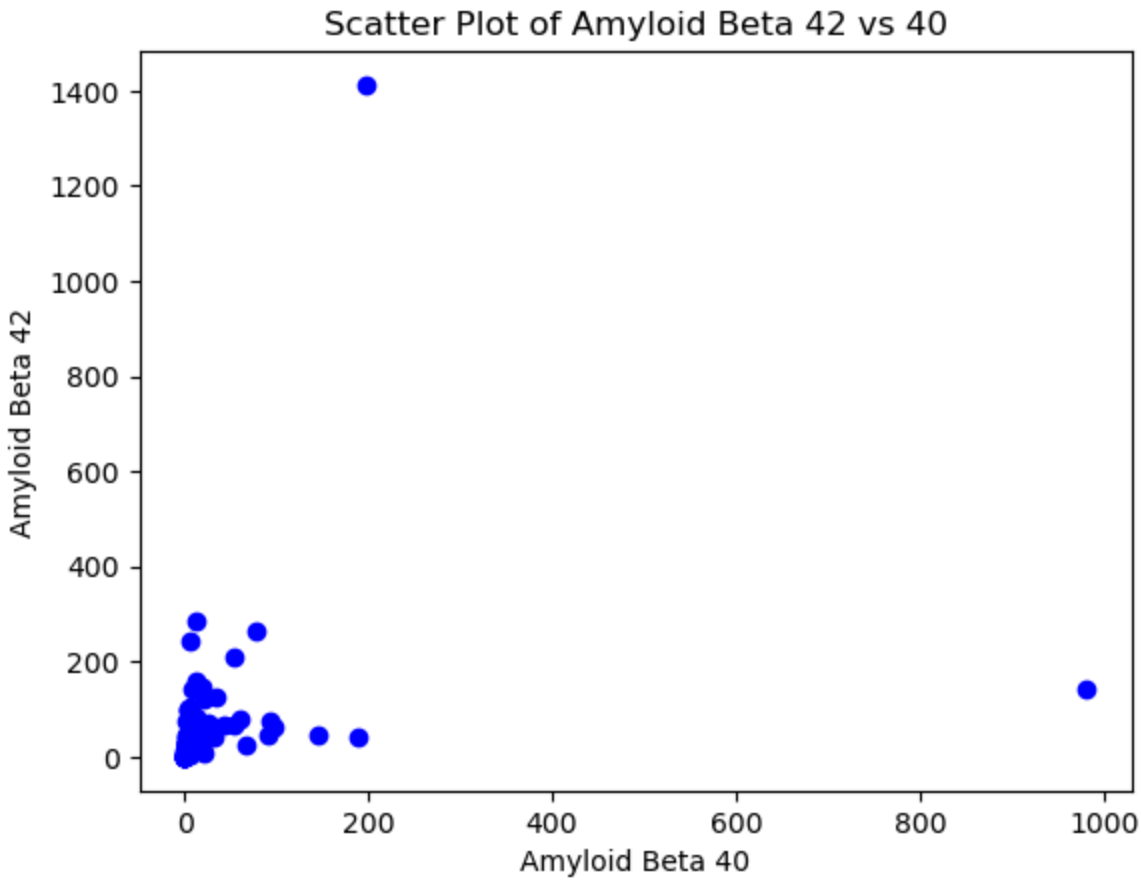
The following code generates a scatterplot. The R^2 value is close to zero, which shows no clear correlation between AB-40 and AB-42 in these patients.

In [ ]:
```
HI_COL_list = merged_df["Known head injury"].tolist()
AB42_COL_list = merged_df["ABeta42 pg/ug"].tolist()
AB40_COL_list = merged_df["ABeta40 pg/ug"].tolist()
RATIO_COL_list = merged_df["AB42_40"].tolist()


X = HI_COL_list  # Independent variable
y1 = AB42_COL_list
y2 = AB40_COL_list
y3 = RATIO_COL_list
                        # Dependent variables

print("Known Head Injury (X):", X)
print("AB42 (y1):", y1)
print("AB40 (y2):", y2)
print("RATIO (y3):", y3)

# Since our question is comparing a binary varaible to a numerical variable, a scat
# However, just for fun, here is a scatterplot of AB42 vs AB40.

plt.scatter(y2, y1, color='blue')
plt.xlabel('Amyloid Beta 40')
```

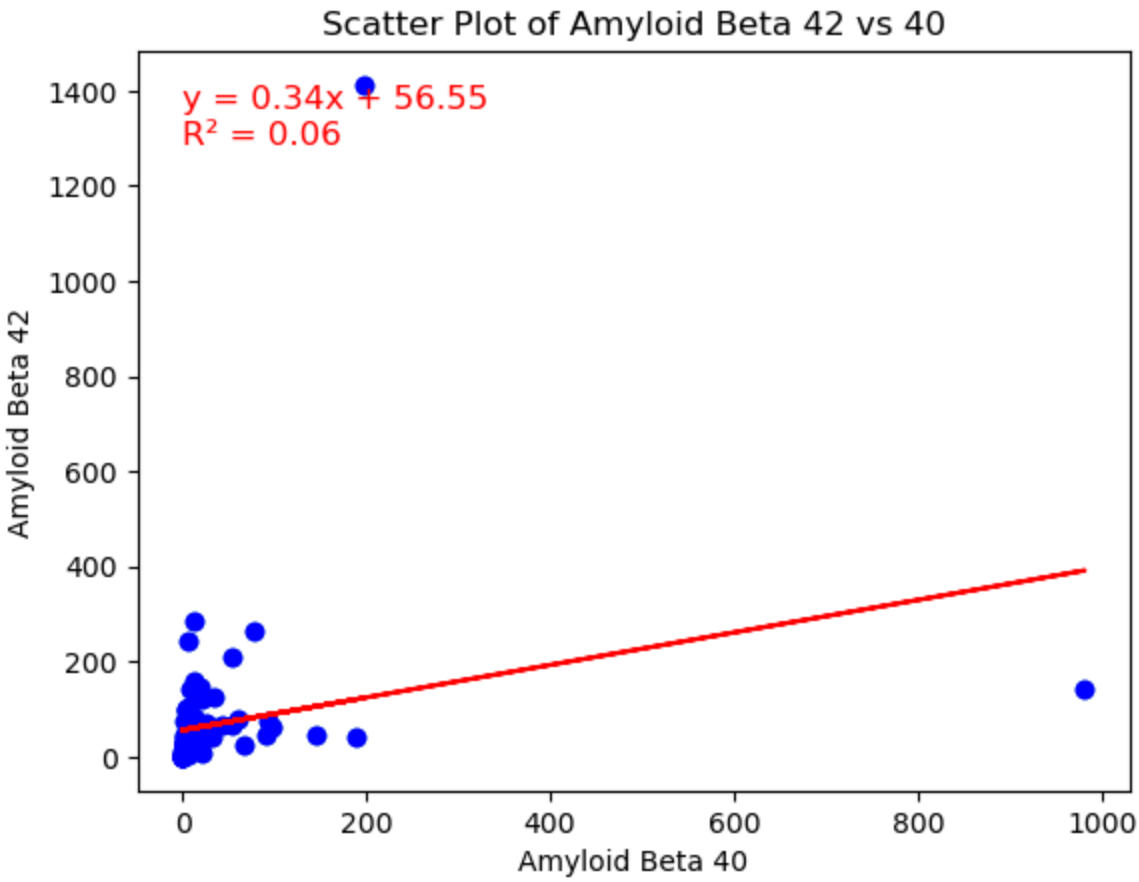
```
plt.ylabel('Amyloid Beta 42')
plt.title('Scatter Plot of Amyloid Beta 42 vs 40')
plt.show()


#11) EXPORT DATA TO A .csv FILE

import pandas as pd

print(AB40_COL_list)
print(AB42_COL_list)

# Create a DataFrame
df = pd.DataFrame({
    'AB40': AB40_COL_list,
    'AB42': AB42_COL_list,
})

# Write to CSV
df.to_csv('patient_data.csv', index=False)

print("CSV file 'patient_data.csv' has been created.")
#12) LOAD LIBRARIES FOR A LINEAR REGRESSION

from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

#13) LOAD DATA SET FOR A LINEAR REGRESSION

df = pd.read_csv("patient_data.csv")


#14) Update these variable names to match EXACTLY your .csv file headers

x = df["AB40"].values.reshape(-1, 1)
y = df["AB42"].values

#15) Perform the linear regression

model = LinearRegression()
model.fit(x, y)

slope = model.coef_[0]
intercept = model.intercept_
r2 = model.score(x, y)

#16) Make scatterplot
plt.scatter(x, y, label="Data")
plt.plot(x, model.predict(x), color="red")

# Annotate equation
equation = f"y = {slope:.2f}x + {intercept:.2f}\nR² = {r2:.2f}"
plt.text(x.min(), y.max(), equation, color="red", fontsize=12, verticalalignment='t

# Annotate scatterplot with labeles and title
plt.scatter(y2, y1, color='blue')
```

```
plt.xlabel('Amyloid Beta 40')
plt.ylabel('Amyloid Beta 42')
plt.title('Scatter Plot of Amyloid Beta 42 vs 40')
plt.show()
```

```
Known Head Injury (X): [nan, nan, nan, 'No', nan, nan, nan, nan, 'No', nan, 'Yes', n
an, 'No', nan, nan, nan, nan, 'Yes', nan, nan, nan, nan, nan, 'Yes', nan, 'Yes', na
n, 'Yes', 'No', nan, nan, 'No', nan, nan, 'No', nan, nan, nan, nan, nan, 'No', nan,
nan, nan, 'No', nan, 'No', 'No', 'No', nan, 'No', nan, nan, 'Yes', 'No', nan, 'Yes',
'Yes', nan, nan, 'Yes', nan, 'No', nan, nan, nan, nan, 'No', 'No', nan, nan, nan, 'N
o', nan, nan, nan, nan, nan, nan, nan, 'Yes', nan, nan, nan]
AB42 (y1): [0.971578947, 2.744210526, 0.147157895, 80.26631579, 16.15684211, 101.830
5263, 60.51157895, 47.70947368, 24.78105263, 16.13789474, 27.60947368, 21.27368421,
209.4347368, 1412.566961, 28.81368421, 45.72842105, 105.1042105, 95.79263158, 63.374
73684, 29.95578947, 18.94736842, 28.85473684, 58.26631579, 42.51368421, 44.25684211,
123.3684211, 4.96, 0.525263158, 102.4557895, 67.65473684, 81.13789474, 27.33473684,
12.69789474, 242.5863158, 60.95368421, 0.245263158, 142.778, 69.98842105, 0.40526315
8, 74.77684211, 0.405263158, 0.670526316, 6.554736842, 82.97263158, 287.412, 18.994,
40.19894737, 24.63789474, 0.137347368, 3.502105263, 68.36631579, 35.27578947, 10.095
78947, 0.525263158, 20.18210526, 10.98842105, 0.019621053, 38.15894737, 2.672631579,
7.666315789, 0.114736842, 8.842105263, 263.5368421, 39.83789474, 0.204385895, 146.86
21053, 18.54736842, 124.4347368, 6.777894737, 31.76315789, 161.0947368, 143.4642105,
75.78947368, 0.449649126, 0.122631579, 76.22631579, 0.490526316, 88.16947368, 47.932
63158, 29.89263158, 33.63789474, 53.87894737, 19.19578947, 0.049052632]
AB40 (y2): [0.019621053, 0.215789474, 0.000597895, 60.76631579, 5.136842105, 3.99157
8947, 11.84526316, 2.529473684, 1.127368421, 0.526168105, 1.944210526, 2.671578947,
52.64210526, 196.732, 1.718947368, 145.2547368, 5.095789474, 3.532631579, 31.5652631
6, 1.843157895, 1.127368421, 1.633684211, 17.22105263, 2.004210526, 91.74842105, 20.
21157895, 4.794736842, 0.030147368, 3.594736842, 53.01263158, 5.176842105, 2.0621052
63, 1.412631579, 5.522105263, 97.8, 0.007088421, 981.444, 25.29578947, 0.000882947,
93.67684211, 0.000804526, 0.001155368, 1.655789474, 12.87684211, 11.41894737, 18.99
4, 189.2905263, 66.77578947, 0.000688421, 0.215789474, 43.23368421, 4.864210526, 0.6
61052632, 0.009426737, 1.412631579, 0.263157895, 0.001077758, 1.547368421, 0.0012610
53, 0.000130411, 0.000597684, 21.20947368, 76.91789474, 31.71157895, 0.072506632, 1
7.82736842, 1.827368421, 34.42947368, 1.375789474, 3.967368421, 12.15368421, 7.49157
8947, 5.302105263, 0.036216842, 0.079678842, 1.450526316, 0.065191789, 5.010526316,
20.53894737, 1.593684211, 7.130526316, 21.42315789, 2.421052632, 0.000981053]
RATIO (y3): [49.517166433422304, 12.71707315065794, nan, nan, 3.145286886329164, 25.
51133966084625, 5.108504398141206, 18.861423220879022, 21.981325863304452, 30.670606
19343318, 14.200866269767332, 7.96296296386408, nan, nan, 16.76240049369563, nan, 2
0.625697163563785, 27.11650774721221, nan, 16.25242718014671, 16.806722688926552, 1
7.662371127610783, 3.3834352081647405, 21.21218487702923, nan, 6.103848759426092, 1.
034467618024906, 17.423184604374086, 28.501610549882916, nan, 15.673241156347762, 1
3.255742725874612, 8.988822654657646, nan, nan, 34.60053487229384, nan, nan, nan, na
n, nan, nan, 3.9586776851318475, 6.443554317992643, nan, 1.0, nan, nan, nan, 16.2292
68268201068, nan, 7.252109932628356, 15.27229297832975, 55.720569906638964, 14.28688
5243133872, 41.755999948244, 18.205434800762323, 24.660544219546278, nan, nan, nan,
0.4168941387422491, nan, nan, 2.8188579356437353, 8.238013701183162, 10.149769584969
752, nan, 4.926549348639659, 8.006102413345795, 13.254806856628043, 19.1500632263176
24, 14.294222751269434, 12.415470294179709, 1.539073308821431, 52.55079825108116, 7.
524357338928067, 17.596848737916098, 2.3337433373052168, 18.75693526589127, 4.717449
070277016, 2.5149862427681526, 7.928695649273271, 49.9999816523674]
```




```
[0.019621053, 0.215789474, 0.000597895, 60.76631579, 5.136842105, 3.991578947, 11.84
526316, 2.529473684, 1.127368421, 0.526168105, 1.944210526, 2.671578947, 52.6421052
6, 196.732, 1.718947368, 145.2547368, 5.095789474, 3.532631579, 31.56526316, 1.84315
7895, 1.127368421, 1.633684211, 17.22105263, 2.004210526, 91.74842105, 20.21157895,
4.794736842, 0.030147368, 3.594736842, 53.01263158, 5.176842105, 2.062105263, 1.4126
31579, 5.522105263, 97.8, 0.007088421, 981.444, 25.29578947, 0.000882947, 93.6768421
1, 0.000804526, 0.001155368, 1.655789474, 12.87684211, 11.41894737, 18.994, 189.2905
263, 66.77578947, 0.000688421, 0.215789474, 43.23368421, 4.864210526, 0.661052632,
0.009426737, 1.412631579, 0.263157895, 0.001077758, 1.547368421, 0.001261053, 0.0001
30411, 0.000597684, 21.20947368, 76.91789474, 31.71157895, 0.072506632, 17.82736842,
1.827368421, 34.42947368, 1.375789474, 3.967368421, 12.15368421, 7.491578947, 5.3021
05263, 0.036216842, 0.079678842, 1.450526316, 0.065191789, 5.010526316, 20.53894737,
1.593684211, 7.130526316, 21.42315789, 2.421052632, 0.000981053]
[0.971578947, 2.744210526, 0.147157895, 80.26631579, 16.15684211, 101.8305263, 60.51
157895, 47.70947368, 24.78105263, 16.13789474, 27.60947368, 21.27368421, 209.434736
8, 1412.566961, 28.81368421, 45.72842105, 105.1042105, 95.79263158, 63.37473684, 29.
95578947, 18.94736842, 28.85473684, 58.26631579, 42.51368421, 44.25684211, 123.36842
11, 4.96, 0.525263158, 102.4557895, 67.65473684, 81.13789474, 27.33473684, 12.697894
74, 242.5863158, 60.95368421, 0.245263158, 142.778, 69.98842105, 0.405263158, 74.776
84211, 0.405263158, 0.670526316, 6.554736842, 82.97263158, 287.412, 18.994, 40.19894
737, 24.63789474, 0.137347368, 3.502105263, 68.36631579, 35.27578947, 10.09578947,
0.525263158, 20.18210526, 10.98842105, 0.019621053, 38.15894737, 2.672631579, 7.6663
15789, 0.114736842, 8.842105263, 263.5368421, 39.83789474, 0.204385895, 146.8621053,
18.54736842, 124.4347368, 6.777894737, 31.76315789, 161.0947368, 143.4642105, 75.789
47368, 0.449649126, 0.122631579, 76.22631579, 0.490526316, 88.16947368, 47.93263158,
29.89263158, 33.63789474, 53.87894737, 19.19578947, 0.049052632]
CSV file 'patient_data.csv' has been created.
```



Verify and validate your analysis:

Findings from the paper "Higher Soluble Amyloid β Concentration in Frontal Cortex of Young Adults than in Normal Elderly or Alzheimer's Disease" (Helmond, et al. 2010) show an increase in amyloid beta 40 and 42 as insoluble plaques in the frontal cortex of Alzheimer's-diagnosed brains versus typical brains.

However, the ratio of AB-42:AB-40 is much less clear in brain tissue, as it can differ between different tissue types (white matter tracts, gray matter, etc.). Multiple studies have show an increased ratio in cerebral spinal fluid (CSF) Though it may be reasonable to assume an increased ratio in CSF would mean a decreased ratio in brain matter, complications in tissue type make that a loaded question to explore, and is certainly not testable with our data. Given this, our outliers and inconclusive results in our ratio analysis make sense.

Additionally, findings from the paper "Long-term risk of dementia among people with traumatic brain injury in Denmark: a population-based observational cohort study" (Fann et al. 2018) show a positive correlation between traumatic brain injury and dementia risk, as did a 2024 meta-analysis "The Association Between Traumatic Brain Injury and the Risk of Cognitive Decline: An Umbrella Systematic Review and Meta-Analysis"(Mavroudis, et al. 2024).

Overall, our only significant findings of a correlation between head injury and lower AB-40 levels disprove our hypothesis and previous findings that AB-40 plaques are more concentrated in the brains of Alzheimer's patients. Outside of that, the complicated relationship of AB-42:AB-40 across brain regions makes it hard to validate the rest of our findings.

Conclusions and Ethical Implications:

In this project, our goal was to test out whether a history of head injury is assoicated with differences in amyloid beta concentrations in brain tissue. Our analysis did not support our original hypothesis that a history of head injury would show increased Aβ42 levels in the brain and an increased Aβ42:Aβ40 ratio. While Aβ40 was signifigantly lower in the head injury group, Aβ42 and the Aβ42:Aβ40 ratio showed no clear or consistent differences. It is important to note that these conclusions are limited by the small sample size available for analysis due to limited data regarding history of head injuries. Ethically, there is a duty to avoid misleading causal narratives or overstating conclusions from a dataset of limited size, as this could misinform clinicians, patients, and the public about the role of a head injury in Alzheimer's pathology. It is particularly important to avoid implying a head injury is protective against amyloid accumulation, as this would contradict the broader literature. It is also important to coniser that the distribution of head injury risk is not uniform across population but is influenced by factors such as occupations, sports participation, and socioeconomic status. Finally, because this data relied on postmortem human tissue, there is a responsibility to respect the privacy of donors and their families.

Limitations and Future Work:

This analysis was subject to several important limitations. The sample size of individuals with both documented head injury status and complete amyloid beta measurements was small, substantially limiting statistical power. In addition, the binary property of the head injury history lacked important information such as severity and timing. Lastly, the data came from postmortem brain tissue, which may not represent the full population of people with Alzheimer's disease.

Future work sould focus on larger datasets that include more detailed information about head injuries. Collecting data on severity and age at injury would allow for more meaningful comparisons. It would also be useful to control for other risk factors, such genetics and lifestyle factors. Finally, future research could expand beyond just amyloid to include other biomarkers, such as Tau content.

Team Notes

9/9

- Started researching Alzheimer's

9/11

- Continued research, learned GitHub and created repository

9/16

- Will out sick
- Analyzed template code
- Made data summary code

9/18

- Trey out sick
- Used template code to make first bar chart and perform t-test, will discuss other questions next week because our findings are very insignificant.

9/23

- Decided to analyze head injuries and amyloid beta relationship instead
- Tailored our code for this and made new bar charts, divided up work between bar chart(s) and scatterplot

9/25

- Continued working on our plots
- Began organizing and annotationg Jupyter notebook for final submission
- Set meeting on Monday to finalize notebook

9/30 - Finished all code (including scatterplots and linear regression) - finished research section - Completed Verification and Validation portion - Completed Conclusions and Ethical Implications section - Completed Limitations / Future Work section

Questions for TA

None Currently