Module_1: Neurodegeneration

Team Members:

Noah Edouard, Kai Huang

Project Title:

Tau Pathology and Disease Duration in Alzheimer’s Disease: Investigating the Relationship Between Clinical Progression and Protein Biomarkers

Project Goal:

This project seeks to examine whether longer disease duration (calculated as age at death minus age at symptom onset) is associated with higher levels of tau pathology (tTau and pTau) in Alzheimer’s disease. By integrating laboratory measurements of brain proteins with patient clinical information, the analysis aims to uncover whether disease progression length is correlated with biomarker accumulation. Identifying such relationships could provide important insights into the dynamics of Alzheimer’s pathology over time, helping to clarify whether extended disease duration reflects greater tau burden and thus potentially more severe neurodegeneration.

Disease Background:

- Prevalence & incidence Global Info:

- Global Prevalence: 0.7% (Chatgpt)- Indicates that nearly 1 in every 140 people worldwide are living with Alzheimer’s disease at any given time.
- Global Incidence: 0.13% (Chatgpt)- Reflects the proportion of new cases per year, showing a steady increase as populations age USA Info:
- In the U.S., about 7.2 million Americans aged 65 and older are living with Alzheimer’s in 2025() Alzheimer’s Association). -Among U.S. older adults, approximately 11% of people age 65+ have Alzheimer’s disease. (Alzheimer’s Association)
- Deaths from Alzheimer’s have more than doubled between 2000 and 2021 in the U.S. Alzheimer’s went from being a less frequent cause of death to a leading cause among older adults(Chatgpt)
- Globally, there were over 55 million people with dementia in 2020 (of which Alzheimer’s is the most common form), and that is projected to nearly double every 20 years: ~78 million by 2030, ~139 million by 2050. (Alzheimer's Disease International) In the U.S., there is substantial variation by region and population: For example, in some U.S. counties, the prevalence of Alzheimer’s dementia among those 65+ reaches ~15-16%.

- Economic burden -Global Info

- Based on 2025's current reports, the economic burden is projected to be roughly $1.8-2.0 trillion(Chatgpt). Includes direct healthcare costs, caregiving, and indirect costs like productivity loss.
- By 2030, global costs are projected to be more than $2.8trillion(Chatgpt).Thisrisingburdenemphasizestheurgencyofdevelopingeffectiveinterventions.USAInfo:-TheannualindirectcostintheU.S.ofAlzheimer'sdisease(includingunpaidcaregivingandproductivitylosses)isestima$ 832 billion. This includes ~$599billioninunpaidcaregivingand$ 233 billion from productivity losses(PubMed).
- In 2025, the total cost of dementia in the U.S. is estimated at $781billion.Ofthat,medicalandlong-termcarecostsare$ 232 billion, and billions more are associated with caregiving hours, loss of income, and reduced quality of life(USC Schaeffer).
- A large share of the burden is “hidden” costs borne by unpaid caregivers. For instance, in 2023 in the U.S., unpaid caregivers provided ~18.4 billion hours of care valued at approx $346.6 billion.

- Risk factors (genetic, lifestyle)- via Chatgpt

- Genetic Factors
- early-onset genes = make more amyloid; late-onset genes = impair clearance, immune response, or brain resilience.
- Lifestyle Factors Negatively Affecting Alzheimers(Chatgpt):
- High blood pressure
- Type 2 Diabetes
- High cholestrol and obesity
- Smoking/vaping
- Alcohol consumption
- Poor mental health
- Lifestyle Factors Positively Affecting Alzheimers(Chatgpt):
- Physical activity
- Adequate sleep
- Mentally challenging activities(puzzles)
- Lifelong learning

- Societal determinants- via Chatgpt

- Access to Healthcare: Limited access delays diagnosis and treatment.
- Social Isolation: Lack of social support is linked to faster cognitive decline.
- Socioeconomic Status: Poverty reduces access to healthy diets, education, and care.
- Urbanization (Air Pollution): Environmental stressors may increase neurodegenerative risk.

- Symptoms- via Chatgpt

- memory loss
- executive function
- language impairment
- disorientation(time,place,people)

- behavior and psychological problem: anxiety, depression, agitation
- Progressive loss of independence

- Diagnosis

- Clinical History & Cognitive Testing → Doctors assess memory, thinking, and daily functioning using tools like MMSE or MoCA.
- Neurological & Physical Exam → Rules out other conditions (stroke, Parkinson's, thyroid problems) that can mimic dementia.
- Brain Imaging (MRI/CT, PET scans) → Detects brain shrinkage, reduced metabolism, or amyloid/tau buildup.
- Biomarker Testing (CSF or blood tests) → Measures amyloid, tau, and other proteins linked to Alzheimer's pathology.
- Diagnostic Criteria (NIA-AA, DSM-5) → Provides standardized definitions to confirm Alzheimer's and distinguish it from other dementias.

- Standard of care treatments (& reimbursement)

1. therapies that may last a long time: Cholinesterase inhibitors: Donepezil, Rivastigmine, Galantamine - used to improve cognition and daily function by increasing acetylcholine NMDA receptor antagonist : memantine used to reduce excitotoxicity(overreaction by glutamate) - mainly used in combination for modern and severe stage of the disease
2. disease modifying therapies(newly approved): Aducanumab approved in 2021: limited intakes, monoclonal antibody against β amyloid Lecanemab approved in 2023: – reduces amyloid plaques,slow cognitive decline by ~27% in 18 months under specific trials Donanemab (expected approval ~2025) – targets amyloid protofibrils, also slows decline. Reimbursement(U.S and globally): limited reimbursement of Aducanumab due to weak evidence. With Lecanemab fully approved, Medicare covers under registry participation. Similar reimbursement debates ongoing in Europe & Japan, balancing cost vs benefit. Annual costs are $26,500$–28,000 per patient, raising access & equity issues.

- Disease progression & prognosis 6 main stages:

1. Preclinical stage: Amyloid plaques, tau tangles begin decades before symptoms. Patients are cognitively normal but may show changes on biomarkers (PET scans, CSF testing).

2. Mild Cognitive Impairment: Subtle memory loss, difficulty with word-finding or problem-solving. Daily function mostly intact.

3. Mild AD dementia: Noticeable memory lapses, confusion about time/place, impaired planning. Patients may withdraw socially.

4. Moderate AD dementia: Greater dependence on caregivers. Problems with language, wandering, personality/behavioral changes (agitation, depression, delusions).

5. Severe AD dementia: Loss of independence, inability to communicate, motor decline, difficulty swallowing.

6. Prognosis: Median survival is typically 8–12 years after diagnosis, though it varies (as short as 3 years or as long as 20). The leading causes of death are complications such as infections (e.g., pneumonia) or immobility-related issues.

- Continuum of care providers -Primary care physicians: Early detection, referral, comorbidity management. -Neurologists / Geriatricians: Confirm diagnosis, manage disease-modifying and symptomatic drugs. -Neuropsychologists: Cognitive testing and monitoring. -Psychiatrists / Psychologists: Address depression, anxiety, agitation. -Nurses / Nurse Practitioners: Care coordination, patient/family education. -Social workers & case managers: Resource navigation, planning for long-term care. -Speech, occupational, and physical therapists: Support independence and safety. -Home health aides and long-term care facilities: Daily living assistance in later stages. -Caregivers (family/friends): Critical for daily functioning, often requiring emotional and financial support.

- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology) Main cause: abnormal protein accumulation and neuronal dysfunction Anatomy & Organ Physiology: Hippocampus (early atrophy → memory loss). Cortical regions (frontal and temporal lobes → language, reasoning deficits). Cell & Molecular Physiology: Amyloid-β plaques: Extracellular deposits that disrupt signaling, cause inflammation. Tau tangles: Hyperphosphorylated tau accumulates inside neurons → destabilizes microtubules. Neuroinflammation: Activated microglia and astrocytes cause chronic inflammation. Synaptic dysfunction: Loss of neurotransmitters, especially acetylcholine, key for memory. Mitochondrial dysfunction & oxidative stress: Further neuronal injury. Vascular contributions: Cerebral hypoperfusion and blood–brain barrier breakdown exacerbate pathology.

- Clinical Trials/next-gen therapies Tau-directed therapies Tau-lowering antisense oligonucleotides (ASOs) (e.g., BIIB080) Anti-tau antibodies and small-molecule tau aggregation inhibitors Active immunization (vaccines) Vaccines that train the immune system against Aβ (e.g., ACI-24) or tau Oral / non-infusion anti-amyloid options Aβ oligomer blockers (e.g., ALZ-801) and subcutaneous/long-acting antibody formats Neuroinflammation & microglia modulation Targets like TREM2, complement, inflammasome; aim to rebalance immune responses Synaptic/neuronal resilience Boosting neurotrophic signaling (e.g., TrkB), restoring plasticity, myelin/oligodendrocyte support Vascular & metabolic approaches Cerebrovascular health, insulin signaling, mitochondrial function, oxidative stress reduction Gene/RNA therapies ASOs/siRNA or gene edits aimed at APP processing, APOE, or tau (MAPT) expression Advanced delivery Blood–brain barrier strategies (intranasal delivery, focused ultrasound) to improve CNS drug uptake

Data-Set:

Data information retrieved from https://canvas.its.virginia.edu/courses/153653/files/15906851?wrap=1

Questions for Data Set Exploration: Describe how the data was collected? What techniques were used? What units are the data measured in? By whom and when was the data manufactured? How were the individuals the data studied, chosen?

What is included in the data set?

The Meta data set includes demographic and clinical information such as age at death, sex, race (with options for White, Black/African American, Asian, American Indian/Alaska Native, Native Hawaiian or Pacific Islander, Unknown or unreported, or Other, with a field to specify another race if applicable), and Hispanic/Latino ethnicity. It also records highest level of education, years of education, APOE genotype, cognitive status, age of onset of cognitive symptoms, and age at dementia diagnosis. Medical history variables include whether the individual had a known head injury, whether they underwent neuroimaging, and the consensus clinical diagnosis (choices include Alzheimer's disease, Alzheimer's possible/probable, ataxia, and others). Neuropathological measures included in the dataset are arteriolosclerosis, LATE, RIN, and whether the case was categorized as a severely affected donor

The Luminex dataset offers highly quantitative measurements of hyperphosphorylated tau and beta-amyloid (Aβ) levels in formalin-fixed postmortem brain tissue, using bead-based multiplex immunoassays to overcome limitations of traditional histology. In studies applying this technique, researchers found that cases with high Alzheimer's neuropathologic change showed elevated Aβ in the frontal cortex and striatum, and increased tau in the frontal cortex and hippocampus—mirroring the spatial patterns of neurodegeneration. This aligns with the surface-level findings in the SEA-AD paper, where tau and Aβ burden were quantified via immunohistochemistry and integrated into a continuous pseudoprogression score. While SEA-AD emphasizes cell-type vulnerability and transcriptomic shifts, Luminex complements this by offering precise protein-level data that reinforces the pathological staging and regional specificity of tau and Aβ accumulation in Alzheimer's disease.

How was data collected:

1. Quantitative Neuropathology Technique: Immunohistochemistry (IHC) on formalin-fixed, paraffin-embedded brain tissue Markers Used: pTau (AT8), Aβ (6E10), pTDP-43, α-synuclein, NeuN, GFAP, IBA1 Units Measured In: Number of stained objects per mm² (e.g., plaques, tangles) Percent area stained (%) Cell counts per cortical layer
2. Single-Nucleus Genomics Techniques: snRNA-seq: Measures gene expression at single-nucleus resolution snATAC-seq: Profiles chromatin accessibility snMultiome: Simultaneously captures RNA and ATAC data from the same nucleus Units Measured In: RNA: Unique Molecular Identifiers (UMIs) per gene per nucleus ATAC: Number of accessible chromatin peaks per nucleus
3. Spatial Transcriptomics Technique: MERFISH (Multiplexed Error-Robust Fluorescence In Situ Hybridization) Units Measured In: Transcript counts per cell Spatial coordinates within tissue sections
4. Cell Type Annotation & Integration Technique: Mapping to reference taxonomy using BICAN standards Units Measured In: Cell identity labels (e.g., supertypes) Pseudoprogression scores derived from Bayesian modeling of pathology metrics
5. External Dataset Harmonization Technique: Reprocessing and integration of 10 publicly available snRNA-seq datasets Units Measured In: Harmonized gene expression profiles Cell type assignments mapped to reference taxonomy Who Manufactured the Data Sets: Lead Institutions: Allen Institute for Brain Science (Seattle, WA) University of Washington Alzheimer's Disease Research Center (UW ADRC) Kaiser Permanente Washington Health Research Institute (ACT Study) Consortium Name: Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) Principal Investigators: Dr. Ed S. Lein Dr. C. Dirk Keene Dr. Michael Hawrylycz When Was the Dataset Manufactured? Donor Tissue Collection Period: The brain tissue samples were collected from 84 donors through rapid autopsy protocols, with a mean postmortem interval of 7.0 hours. Data Generation Timeline: Neuropathology, snRNA-seq, snATAC-seq, snMultiome, and MERFISH profiling were conducted in the years leading up to the

paper's publication. The paper was: Received: April 24, 2024 Accepted: August 28, 2024 Published Online: October 14, 2024 This places the manufacturing of the dataset primarily between 2022 and 2024, with final integration and analysis completed in early 2024.

What criteria were used to select the donor cohort, and how representative is it of the broader Alzheimer's population? Selection Criteria: Donors were recruited from two well-established longitudinal studies: Adult Changes in Thought (ACT) study University of Washington Alzheimer's Disease Research Center (UW ADRC) Inclusion required: Death within a specific time window to allow rapid autopsy (mean postmortem interval: 7.0 hours) High-quality tissue suitable for single-nucleus and spatial profiling No diagnosis of confounding neurodegenerative disorders (e.g., ALS, Down syndrome, FTLD) Inclusion of common comorbidities like Lewy body disease, vascular pathology, and LATE Demographics: Total Donors: 84 Age Range: 65–102 years (mean: 88) Sex: 51 female, 33 male AD Pathology Spectrum: 9 with no AD 12 low AD neuropathological change (ADNC) 21 intermediate ADNC 42 high ADNC Genetic Risk Representation: Nearly half of high ADNC donors carried the APOE4 allele, a known genetic risk factor for AD Clinical Data: -Longitudinal cognitive testing across four domains: memory, executive function, language, visuospatial -Dementia status and comorbidities were recorded and integrated into analysis This cohort was intentionally designed to reflect the full spectrum of AD pathology rather than a simple case-control model, making it highly representative of real-world AD progression and heterogeneity.

Code used to inspect headers of Meta File: import csv as csv

```
with open ("UpdatedMeta.csv", newline="") as f: reader = csv.reader(f) headers = next(reader) for h in headers: print(h)
```

Acquisition of Tau and Beta-Amyloid Data from Luminex Data Set Technique Used: Immunohistochemistry (IHC) on formalin-fixed, paraffin-embedded (FFPE) brain tissue sections. Markers and Antibodies: Tau (pTau): Detected using the AT8 antibody, which targets phosphorylated tau at Ser202/Thr205. Beta-Amyloid (Aβ): Detected using the 6E10 antibody, which binds to amino acids 1–16 of the Aβ peptide. Staining Protocol: Tissue sections were cut at 5 μm thickness. Heat-induced epitope retrieval was performed using Diva Decloaker solution at 110 °C for 15 minutes. Chromogenic staining was visualized using HRP-mediated oxidation of DAB (brown precipitate). Duplex staining was also performed (e.g., AT8/pTDP-43, 6E10/IBA1) using alkaline phosphatase-based detection with Ferangi Blue chromogen (blue precipitate). Quantification Method: Whole-slide imaging at 20x magnification. Cortical layers were annotated and segmented using HALO image analysis software. Quantitative metrics included: Number of pTau+ neurons per unit area Number and diameter of Aβ plaques per unit area All values were normalized and converted to z-scores for modeling Integration into Disease Modeling: These measurements were used to construct a Continuous Pseudoprogression Score (CPS) via Bayesian modeling. CPS captured the exponential accumulation of tau tangles and amyloid plaques across disease stages.

Code used for Luminex Data Set header extraction:

```
import csv as csv
```

```
with open ("UpdatedLuminex.csv", newline="") as f: reader = csv.reader(f) headers = next(reader) for h in headers: print(h)
```

Data Analyis:

We analyzed the data by first calculating each patient’s disease duration as the difference between age at death and age of symptom onset, then excluding patients with missing values. To explore the relationship between duration and tau pathology, we created scatter plots of disease duration versus tTau and pTau and calculated Pearson correlation coefficients with p-values to assess statistical significance. Patients were also grouped into duration bins, where we plotted the mean tau levels with error bars (SEM) to visualize variability across groups. Finally, we performed independent Student’s t-tests comparing patients with shorter versus longer disease duration (split at the median) to test for significant differences in tau levels, reporting the t-values and p-values, and determining significance at α = 0.05. This approach combines visualization, correlation, and hypothesis testing to evaluate whether longer disease duration is associated with higher tau accumulation.

Concise Steps of Data Analysis Step 1: In Python, we created a class of “patient” objects(DonorID, tTau, pTau,age_symp_on, death_age)Step 2: We populated Patient objects by reading two CSV files, Luminex for protein data and Metadata for clinical infoStep 3: We sorted the data by Donor ID so each patients' protein measurements and clinical info lined upStep 4:From the merged tables we computed disease duration for each donor duration, kept only valid/positive durations, and dropped rows with missing in Duration, pTau, or tTau to build the final analysis DataFrame.Step 5: We visualized the association between disease duration and tau markers by generating scatter plots of Duration vs tTau and pTau and fitting a simple linear regression line.Step 6: We used a t-test and linear regression to determine if as disease duration increased, tau levels also increased

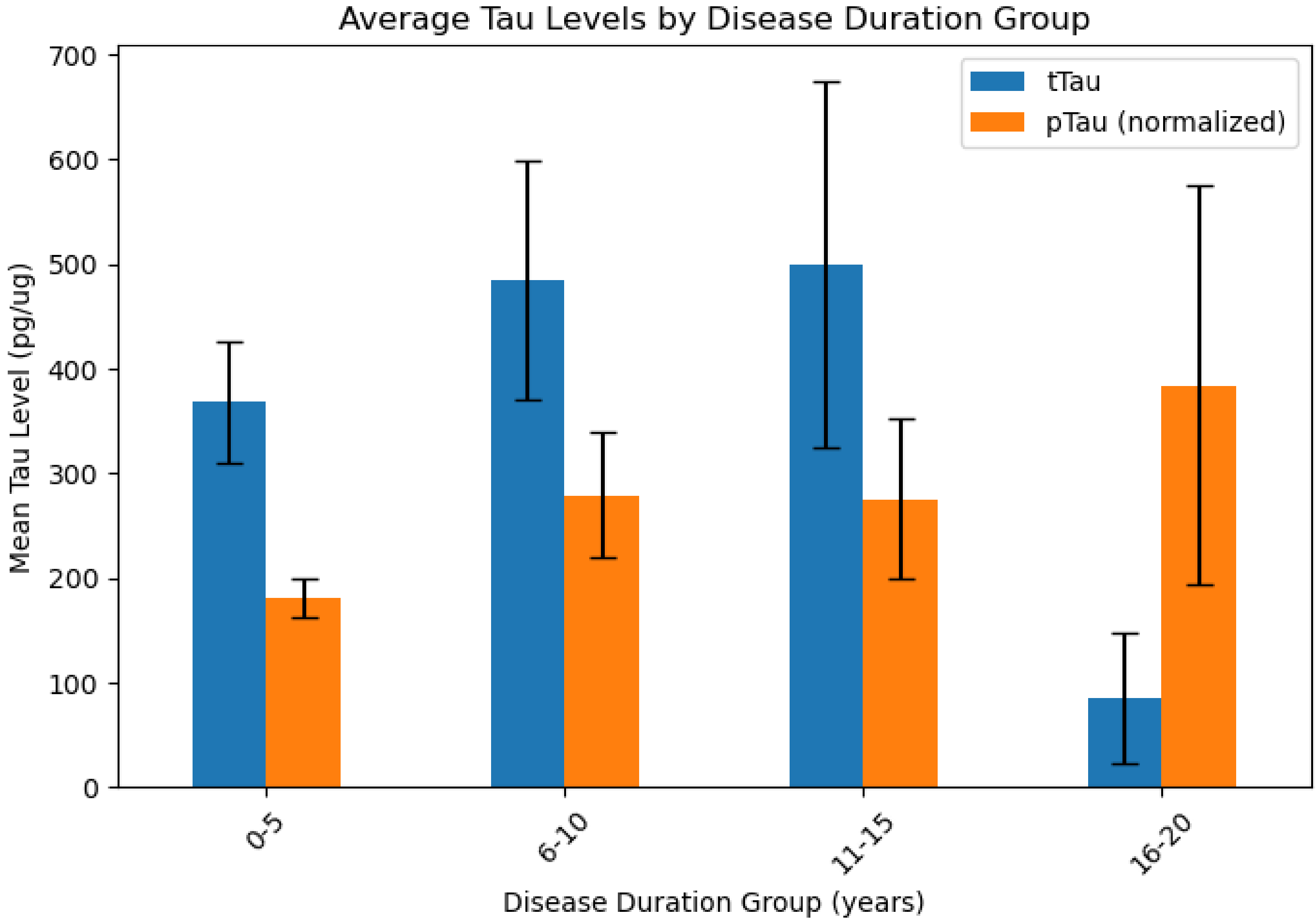
In [4]:

```python
import csv
import warnings
import matplotlib.pyplot as plt

class Patient:
    all_patients = []

    def __init__(self, DonorID, tTau: float, pTau: float):
        self.DonorID = DonorID
        self.tTau = tTau
        self.pTau = pTau
        self.death_age = None
        self.age_symp_on = None
        Patient.all_patients.append(self)

    def __repr__(self):
        return (f"{self.DonorID} | tTau {self.tTau} | pTau {self.pTau} | "
                f"Death Age {self.death_age} | Symptom Onset {self.age_symp_on}")

    def summary(self):
        return {
            "DonorID": self.DonorID,
            "Age at Death": self.death_age,
            "Age of Symptom Onset": self.age_symp_on,
            "tTau pg/ug": self.tTau,
            "pTau pg/ug": self.pTau
        }

    @classmethod
    def combine_data(cls, filename: str):
        with open(filename, encoding="utf8") as f:
            reader = csv.DictReader(f)
            rows = list(reader)
            for i in range(len(rows)):
                if Patient.all_patients[i].DonorID == rows[i]["Donor ID"]:
                    if rows[i]["Age at Death"]:
                        Patient.all_patients[i].death_age = int(rows[i]["Age at Death"])
                    if rows[i]["Age of onset cognitive symptoms"]:
                        Patient.all_patients[i].age_symp_on = int(rows[i]["Age of onset
                else:
                    warnings.warn("IDs do not match.")

    @classmethod
    def instantiate_from_csv(cls, protein_file: str, clinical_file: str):
        with open(protein_file, encoding="utf8") as f:
            reader = csv.DictReader(f)
            rows = list(reader)
            for row in rows:
                Patient(
                    DonorID=row['Donor ID'],
                    tTau=float(row['tTAU pg/ug']),
                    pTau=float(row['pTAU pg/ug'])
                )
        Patient.all_patients.sort(key=lambda p: p.DonorID)
        Patient.combine_data(clinical_file)




Patient.instantiate_from_csv("UpdatedLuminex.csv", "UpdatedMetaData.csv")

# Print out patients
for p in Patient.all_patients[:30]:  # print first 30 as example
    print(p)


import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from scipy.stats import pearsonr, ttest_ind, sem

def analyze_tau_correlation_with_bars_and_ttest():
    durations = []
    ttaus = []
    ptaus = []

    # Collect data from patients
    for p in Patient.all_patients:
        if p.death_age and p.age_symp_on and p.tTau and p.pTau:
            duration = p.death_age - p.age_symp_on
            if duration >= 0:
                durations.append(duration)
                ttaus.append(p.tTau)
                ptaus.append(p.pTau * 50)  # Scale pTau values by 50

    if not durations:
        print("No patients with complete data for analysis.")
        return
    # Create dataframe
    df = pd.DataFrame({
        "Duration": durations,
        "tTau": ttaus,
        "pTau": ptaus
    })

    # --- Bar Graphs with Error Bars (duration bins) ---
    bins = [0, 5, 10, 15, 20]  # bins for grouping disease duration
    labels = ["0-5", "6-10", "11-15", "16-20"]

    df["DurationGroup"] = pd.cut(df["Duration"], bins=bins, labels=labels, right=True)
    grouped = df.groupby("DurationGroup")[["tTau", "pTau"]]

    mean_values = grouped.mean()
    sem_values = grouped.sem()  # Standard Error of the Mean

    mean_values.plot(
        kind="bar",
        yerr=sem_values,
        capsize=5,
        figsize=(8, 5),
        legend=True
    )
    plt.title("Average Tau Levels by Disease Duration Group")
    plt.xlabel("Disease Duration Group (years)")
    plt.ylabel("Mean Tau Level (pg/ug)")
    plt.xticks(rotation=45)
    plt.legend(["tTau", "pTau (normalized)"])
    plt.show()

    # --- Correlation values ---
    corr_tTau, pval_tTau = pearsonr(df["Duration"], df["tTau"])
    corr_pTau, pval_pTau = pearsonr(df["Duration"], df["pTau"])
    print(f"Correlation between Disease Duration and tTau: r = {corr_tTau:.3f}, p = {pva
    print(f"Correlation between Disease Duration and pTau (scaled ×50): r = {corr_pTau:.

    # --- T-tests (short vs. long duration) ---
    median_duration = df["Duration"].median()
    short_group = df[df["Duration"] <= median_duration]
    long_group = df[df["Duration"] > median_duration]

    t_ttau, p_ttau = ttest_ind(short_group["tTau"], long_group["tTau"], equal_var=False)
    t_ptau, p_ptau = ttest_ind(short_group["pTau"], long_group["pTau"], equal_var=False)

    print("\nT-Test Results (Short vs Long Duration):")
    print(f"tTau: t = {t_ttau:.3f}, p = {p_ttau:.3f} --> {'Significant' if p_ttau < 0.05
    print(f"pTau (normalized): t = {t_ptau:.3f}, p = {p_ptau:.3f} --> {'Significant' if

# ---- Run it ----
analyze_tau_correlation_with_bars_and_ttest()

import csv
import warnings
from pathlib import Path


import numpy as np
import pandas as pd
```

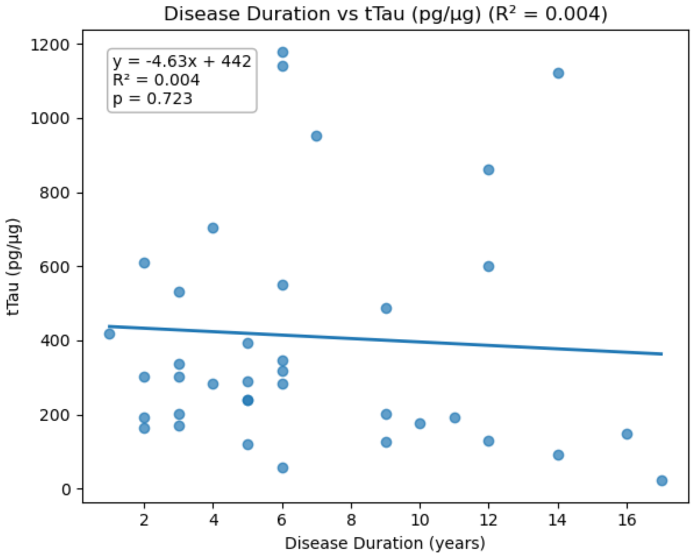
```python
import matplotlib.pyplot as plt
from scipy.stats import linregress, pearsonr, ttest_ind, sem




class Patient:
   all_patients = []


   def __init__(self, DonorID, tTau: float, pTau: float):
       self.DonorID = DonorID
       self.tTau = tTau
       self.pTau = pTau
       self.death_age = None
       self.age_symp_on = None
       Patient.all_patients.append(self)


   def __repr__(self):
       return (f"{self.DonorID} | tTau {self.tTau} | pTau {self.pTau} | "
               f"Death Age {self.death_age} | Symptom Onset {self.age_symp_on}")


   @classmethod
   def instantiate_from_csv(cls, protein_file: str):
       with open(protein_file, encoding="utf8") as f:
           for row in csv.DictReader(f):
               # Skip rows with missing tau values
               if not row.get("Donor ID"):
                   continue
               try:
                   t = float(row["tTAU pg/ug"])
                   p = float(row["pTAU pg/ug"])
               except (TypeError, ValueError):
                   continue
               Patient(DonorID=row["Donor ID"], tTau=t, pTau=p)


       # Make lookup by ID for later enrichment
       cls._by_id = {p.DonorID: p for p in cls.all_patients}


   @classmethod
   def enrich_with_clinical(cls, clinical_file: str):
       """Match by Donor ID (not by row index)."""
       missing = 0
       with open(clinical_file, encoding="utf8") as f:
           for row in csv.DictReader(f):
               did = row.get("Donor ID")
               if not did:
                   continue
               p = cls._by_id.get(did)
               if p is None:
                   missing += 1
                   continue
               # Safely parse ages
               def _to_int(x):
                   try:
                       return int(x)
                   except (TypeError, ValueError):
                       return None
               p.death_age   = _to_int(row.get("Age at Death"))
               p.age_symp_on = _to_int(row.get("Age of onset cognitive symptoms"))


       if missing:
           warnings.warn(f"{missing} clinical rows had Donor IDs not present in protein


   @classmethod
   def to_dataframe(cls) -> pd.DataFrame:
       rows = []
       for p in cls.all_patients:
           if p.death_age is None or p.age_symp_on is None:
               continue
           duration = p.death_age - p.age_symp_on
           if duration is None or duration < 0:
               continue
           rows.append({"Duration": duration, "tTau": p.tTau, "pTau": p.pTau})


       df = pd.DataFrame(rows).replace([np.inf, -np.inf], np.nan).dropna()
       return df




def scatter_with_regression(ax, x, y, y_label):
   ax.scatter(x, y, alpha=0.7)
   slope, intercept, r, p, stderr = linregress(x, y)
   xs = np.linspace(np.nanmin(x), np.nanmax(x), 200)
   ax.plot(xs, slope * xs + intercept, linewidth=2)
   ax.set_xlabel("Disease Duration (years)")
   ax.set_ylabel(y_label)
   ax.set_title(f"Disease Duration vs {y_label} (R² = {r**2:.3f})")
   ax.text(
       0.05, 0.95,
       f"y = {slope:.3g}x + {intercept:.3g}\nR² = {r**2:.3f}\np = {p:.3g}",
       transform=ax.transAxes,
       va="top", ha="left",
       bbox=dict(boxstyle="round,pad=0.3", fc="white", ec="0.7")
   )
   return dict(slope=slope, intercept=intercept, r=r, r2=r**2, p=p, stderr=stderr)




def analyze_tau_correlation_with_bars_and_ttest(df: pd.DataFrame):
   if df.empty:
       print("No patients with complete data for analysis.")
       return None


   # --- Scatter plots ---
   plt.figure(figsize=(12, 5))
   ax1 = plt.subplot(1, 2, 1)
   stats_t = scatter_with_regression(ax1, df["Duration"].to_numpy(), df["tTau"].to_numpy


   ax2 = plt.subplot(1, 2, 2)
   stats_p = scatter_with_regression(ax2, df["Duration"].to_numpy(), df["pTau"].to_numpy


   plt.tight_layout()
   plt.show()


   return {
       "scatter_tTau": stats_t,
       "scatter_pTau": stats_p,
   }


PROTEIN = Path("UpdatedLuminex.csv")
CLINIC  = Path("UpdatedMetaData.csv")


Patient.instantiate_from_csv(PROTEIN)
Patient.enrich_with_clinical(CLINIC)
df = Patient.to_dataframe()


print(f"Built dataframe with {len(df)} patients.")
results = analyze_tau_correlation_with_bars_and_ttest(df)
```

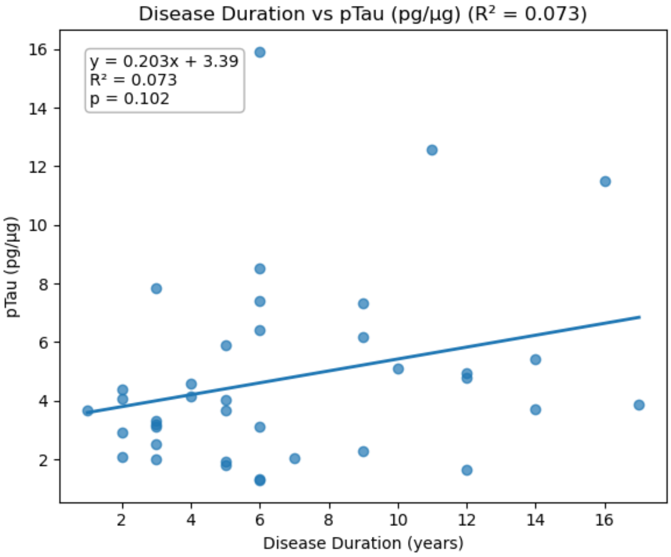
```
H19.33.004 | tTau 1552.414737 | pTau 1.901052632 | Death Age 80 | Symptom Onset None
H20.33.001 | tTau 756.0905263 | pTau 2.737894737 | Death Age 82 | Symptom Onset None
H20.33.002 | tTau 313.5252632 | pTau 2.615789474 | Death Age 97 | Symptom Onset None
H20.33.004 | tTau 318.5284211 | pTau 7.412631579 | Death Age 86 | Symptom Onset 80
H20.33.005 | tTau 107.3484211 | pTau 1.327368421 | Death Age 99 | Symptom Onset None
H20.33.008 | tTau 125.9336842 | pTau 2.569473684 | Death Age 92 | Symptom Onset None
H20.33.011 | tTau 1141.492355 | pTau 8.536842105 | Death Age 93 | Symptom Onset 87
H20.33.012 | tTau 950.7410526 | pTau 4.545263158 | Death Age 91 | Symptom Onset None
H20.33.013 | tTau 272.5084211 | pTau 3.106315789 | Death Age 94 | Symptom Onset None
H20.33.014 | tTau 258.6242105 | pTau 3.398947368 | Death Age 82 | Symptom Onset None
H20.33.015 | tTau 393.1831579 | pTau 1.827368421 | Death Age 88 | Symptom Onset 83
H20.33.016 | tTau 488.8989474 | pTau 2.282105263 | Death Age 93 | Symptom Onset 84
H20.33.017 | tTau 239.3778947 | pTau 5.881052632 | Death Age 69 | Symptom Onset 64
H20.33.018 | tTau 177.5663158 | pTau 5.110526316 | Death Age 81 | Symptom Onset 71
H20.33.019 | tTau 312.7442105 | pTau 2.884210526 | Death Age 87 | Symptom Onset None
H20.33.020 | tTau 21.71894737 | pTau 3.873684211 | Death Age 81 | Symptom Onset 64
H20.33.024 | tTau 309.08 | pTau 5.222105263 | Death Age 90 | Symptom Onset None
H20.33.025 | tTau 384.84 | pTau 3.691578947 | Death Age 94 | Symptom Onset None
H20.33.026 | tTau 191.0505263 | pTau 12.56736842 | Death Age 75 | Symptom Onset 64
H20.33.027 | tTau 224.2431579 | pTau 3.365263158 | Death Age 99 | Symptom Onset None
H20.33.028 | tTau 192.0284211 | pTau 2.927368421 | Death Age 94 | Symptom Onset 92
H20.33.029 | tTau 302.2315789 | pTau 3.191578947 | Death Age 91 | Symptom Onset 88
H20.33.030 | tTau 114.6231579 | pTau 6.56 | Death Age 86 | Symptom Onset None
H20.33.031 | tTau 335.7452632 | pTau 7.827368421 | Death Age 87 | Symptom Onset 84
H20.33.032 | tTau 156.6284211 | pTau 12.47052632 | Death Age 98 | Symptom Onset None
H20.33.033 | tTau 92.80210526 | pTau 3.712631579 | Death Age 68 | Symptom Onset 54
H20.33.034 | tTau 569.2336842 | pTau 2.593684211 | Death Age 85 | Symptom Onset None
H20.33.035 | tTau 533.5926316 | pTau 4.036842105 | Death Age 99 | Symptom Onset None
```

```
H20.33.036 | tTau 345.8894737 | pTau 1.28 | Death Age 100 | Symptom Onset 94
H20.33.037 | tTau 283.24 | pTau 4.569473684 | Death Age 96 | Symptom Onset 92
C:\Users\lafor\AppData\Local\Temp\ipykernel_28452\891557506.py:102: FutureWarning: The d
efault of observed=False is deprecated and will be changed to True in a future version o
f pandas. Pass observed=False to retain current behavior or observed=True to adopt the f
uture default and silence this warning.
  grouped = df.groupby("DurationGroup")[["tTau", "pTau"]]
```



```
Correlation between Disease Duration and tTau: r = -0.059, p = 0.723
Correlation between Disease Duration and pTau (scaled ×50): r = 0.269, p = 0.102

T-Test Results (Short vs Long Duration):
tTau: t = 0.223, p = 0.826 --> Not Significant
pTau (normalized): t = -1.005, p = 0.325 --> Not Significant
Built dataframe with 38 patients.
```





Verify and validate your analysis:

*The correlations between disease duration and tau levels (tTau: r = –0.059, p = 0.723; pTau: r = 0.269, p = 0.102) were weak and not statistically significant.Similarly, the Student's T-tests comparing short vs. long disease duration groups showed no significant differences for either tTau (p = 0.826) or pTau (p = 0.325). Consistent with visual inspection of the scatter plots and bar graphs, which did not show clear trends or group differences.

CSF Results for the controls and AD patients(Validation Source): Wallin, A_K, et al. "CSF Biomarkers for Alzheimer's Disease: Levels of -Amyloid, Tau, Phosphorylated Tau Relate to Clinical Symptoms and Survival." Dementia and Geriatric Cognitive Disorders, vol. 21, no. 3, 1 Jan. 2006, pp. 131-138, karger.com/dem/article-abstract/21/3/131/97685/CSF-Biomarkers-for-Alzheimer-s-Disease-Levels-of, https://doi.org/10.1159/000090631. Accessed 30 Sept. 2025.

Sensitivity = The ability of a test to correctly identify people with ADHigher sensitivity means fewer missing cases"In many studies P-tau is more AD-specific; in this table all three markers show the same specificity (88%) because of the chosen cut-offs.""Because tau (esp. T-tau) shows high diagnostic sensitivity, we ask whether tau also predicts disease duration/progression rate."

Conclusions and Ethical Implications:

Conclusions: Shorter disease duration does not significantly correlate with higher tau levels (tTau or pTau). Tau burden alone is not a strong predictor of survival after sympton onset Other genetic, and environmental factors likely play a larger role in disease progression. Ethical Implications: Clinicians should avoid using tau levels alone to predict survival, to prevent anxiety or false reassurance for patients/families. Hospitals and doctors need a holistic approach that integrates multiple biomarkers and clinical factors.

Limitations and Future Work:

Future research should analyze additional biomarkers such as amyloid, inflammation markers, and APOE genotype to gain a more comprehensive view of the factors that shape Alzheimer's progression. Subgroup differences, particularly between early-onset and late-onset Alzheimer’s, warrant close investigation to determine whether the biological drivers of disease duration differ by onset type. In addition, longitudinal studies that track tau changes within the same individuals over time could provide valuable insight into how tau dynamics relate to disease trajectory, offering a more nuanced understanding than cross-sectional data alone.

Notes and Questions for TA's:

Session 1 – Research Agenda: Research disease background Activities: Noah: Wrote the section on disease background, covering prevalence through diagnosis. Kai: Wrote the section on standards of care, from current treatments to clinical trials and next-generation therapies. Notes: Established a foundational background for the project. Each section complements the other for a complete literature review.

Session 2 – Dataset Extraction & Inspection Agenda Write a code to view the headers of each file Create leading questions for focus the project on Complete research Complete notebook check in Activities: Both: Collaborated to write code for dataset inspection. Noah: Focused on the metadata set, analyzed and documented findings. Kai: Focused on the Luminex dataset, analyzed and documented findings. Both: Took detailed notes describing the dataset’s structure and contents. Decisions: Based on joint notes and inspections, we conferred and selected a research question to guide the next phase of work. Project Questions Considered: How does age at death correlate with cellular and molecular markers in severely affected donors? Does APOE Genotype influence cognitive status and age of dementia diagnosis in relation to cellular vulnerability? Investigate how different APOE alleles (e.g., ε3/ε4, ε4/ε4) interact with cognitive decline and age of diagnosis, and whether they correlate with specific cell-type changes. Are there differences in disease progression or cellular pathology across race (White, Black/African-American, Latino, Asian) donors? How does the accumulation of tau pathology (AT8+ cells per area) vary by sex and how does this factor interact with cognitive status and age of dementia diagnosis? Is beta-amyloid plaque size associated with years of education? Session 3- Statistics and Linear Regression Agenda Finalize leading question Add more information to research Begin object oriented programming Make sure github is set up so we can work on code together Use the data we extract to create a bar graph Tasks Both: Inspect data set to create better research question Both: Add information about difference and tTau and pTau Group Discussion: We decided that we want to consider age at death in our research We decided to combine age of death and age of onset cognitive symptoms as our independent variable, disease duration Questions Considered: “Does longer disease duration (death_age – age_symp_on) correlate with higher tau levels (tTau, pTau)?” “Do tau levels rise faster than amyloid during the disease course (death_age – age_symp_on)?” “Among patients stratified by age of symptom onset (early vs late), do those who die younger have higher tau levels?” Chosen Question: Does longer disease duration (death_age – age_symp_on) correlate with higher tau levels (tTau, pTau)? Homework Kaiwen: create pTau graph Noah: create tTau graph

Session 4: Tasks: Decide what type of graph would be most successful in representing our question Finish second notebook update Include better citations Add to data set exploration Include Data-Analysis

Session 5 Tasks: Noah: Normalize ptau levels on bar graph Noah: Adjust graph to just include necessary disease duration Kai: Create linear regression line Kai: obtain r^2 value

Session 6: In class tasks: Noah:Normalize the graph Kai: Create linear regression Both: Make conclusions from the data Both: Create presentation Finalize background information After class tasks: Presentation Planning: Background information Explanation how and why we chose our question explain how we analyzed the data(code and graphs) *Mention t-test, p-value/r-value, Describe conclusions from data Describe limitations and future work

Questions for our TA: N/a