Module_1: *Alzheimer's Research Project*

Team Members:

Gabby Holohan & Jack Hancock

Project Title:

The Effect of the APOE genotype on the Age of Onset of Alzheimers

Project Goal:

This project seeks to investigate the impact of the APOE genotype on the age of onset of Alzheimer's symptoms for patients treated in the US.

Disease Background:

Fill in information about 11 bullets:

- Prevalence & incidence
    - Prevalence: In 2025, about 7.2 million Americans age 65 and older are living with Alzheimer’s dementia. That is about 1 in 9 people age 65+. Throughout the world, this number is 55 million (2024).
    - Incidence: For ages 65-69: about 2.8 per 1,000 person-years https://www.alz.org/getmedia/ef8f48f9-ad36-48ea-87f9-b74034635c1e/alzheimers-facts-and-figures.pdf
- Economic burden
    - estimated at over $360 billion annually in the U.S. alone (2024), including direct health care costs, long-term care, and unpaid caregiving. The out-of-pocket expenses for patients and caregivers is estimated at 91 billion https://curealz.org/the-basics-of-alzheimers-disease/statistics-and-costs/?
- Risk factors (genetic, lifestyle):
    - Less than 1% purely genetic (disease almost guaranteed)-->develop middle age
    - environmental risk factors: older age, increased chances if 1st degree relative had disease --> familial genetic factors not understood yet, apolipoprotein E (APOE) gene -- APOE e4 increases risk - 25-30% of population carries APOE e4 but not all develop disease - having two copies AOE e4 increases risk more than 1, some rare changes in genes that almost ensure Alzheimer's onset but very rare <1%, down syndrome more likely (3 copies chromosome 21 - gene involved in producing protein leading to creation of Beta-amyloid) --> symptoms occur 10-20yrs earlier w/ down syndrome, overall more women w/ Alz bc they tend to live longer, mild cognitive impairment (MCI) can be at higher risk for Alz, traumatic brain injurgy (TBI) for 50+ --> more risk w/ more severe or multiple, air pollution

exposure, heavy alcohol use --> also early onset, poor sleep patterns, lifestyle risks:

- lack of exercise,
- obesity,
- smoking/secondhand exposure https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447

- Societal determinants
    - lower education level may coorelate to poorer brain health
    - for 45+ reporting symptoms of decining memory and frequent confusion: 2x as many w/o hs degree & lowest for college grads
    - potentially due to cognitive reserve
    - access to health care: early dx for other health issues, prevent chronic disease
    - built environment: unsafe or unhealthy environments can impact brain health, can influence behavior (ex: running, eating healthy, sidewalks, etc)
    - loneliness increases risk of premature death (for everything) & people socially isolated or lonely at higher risk for dementia
    - https://www.cdc.gov/alzheimers-dementia/php/sdoh/index.html
- Symptoms
    - repeating statements/questions, forgetting things, misplacing items, getting lost in familiar places, forgetting names of people/ objects, having trouble conversating trouble concentrating/thinking (math, multiple tasks, etc) --> eventually may not recognize numbers
    - difficulty decision making
    - trouble planning or doing simple tasks (getting dressed)
    - changes in personality/behavior (depressed, withdrawn, distrusting, anger, sleeping habits changing, delusions, wandering, losing inhibitions)
    - Some preserved skills --> things people can do even late-stage https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447
- Diagnosis
    - neurologist or doctor reviews symptoms, medical & medicine history
    - physical exam (past strokes, prakinson's, depression, sleep apnea, etc)
    - mental status testing --> tests cognitive ability
    - lab tests: check for other conditions or cerebrospinal fluid test (measuring amyloid beta and tau in fluid)
    - MRI, CT, PET (can recognize plaques or tau tangles, but mostly in research setting)
    - new blood tests being developed measuring amyloid & tau content
    - https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048075
- Standard of care treatments (& reimbursement)

  - Cholinesterase inhibitors (donepezil, rivastigmine, galatamine) are used for mild to moderate disease
  - Memantine is used for moderate to severe disease
  - Lecanemab and donanemab are disease-modifying monoclonal antibodies for early Alzheimer's
  - Non-drug interventions include physical activity, cognitive stimulation, occupational therapy, and caregiver support
  - Medicare covers FDA-approved anti-amyloid monoclonal antibodies under registry-based evidence collection
  - https://www.cms.gov medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=N&ncaid=305
- Disease progression & prognosis
  - Preclinical stage: mild memory loss, no significant impairment, pathological changes in entorhinal cortex 1st then hippocampus
  - Mild Alzheimer's disease: forgetting new information and appointments, impaired problem solving and judgement, mood swings and personality changes, pathological changes reach cerebral cortex
  - Moderate Alzheimer's disease: pathological changes further affect cerebral cortex (language, reasoning and sensory processing), more severe symptoms, behavioral issues and social isolation, language disorder
  - Severe Alzheimer's disease: lose independence in simple tasks/daily activities, pathological damage covers all cortex areas, cognitive abilities reach lowest state, difficulty performing motor tasks, sleep disturbance, extrapyramidal motor signs
  - No cure, continue to decline until death
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC7815481/
- Continuum of care providers
  - Medical: Primary care, neurologist/geriatrician, psychiatrist (behavioral symptoms), nurse practicioner/dementia clinic teams
  - Cognitive and rehab: neuropsychology, occupational therapy, physical therapy, speech-language pathology (communicattion/swallow)
  - Psycho-social and navigation: social worker/care manager, community programs (Alzheimer's Association 24/7 Helpline, support groups, safety planning)
  - Home and long-term care: home health aides, adult day programs, assisted living, memory care, hospice in late stage -https://www.alz.org/professionals/health-systems-medical-professionals/health-systems/dementia-care-delivery-accreditation
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)
  - common features among subjects include synaptic loss, accumulation of plaques, and tau tangles
  - neurotic plaques --> amyloid beta peptide comes from precursor protein APP, cleaved by alpha, beta, and gamma secretase
  - people w/o AD: APP cleaved by alpha secretase --> cannot form amyloid beta peptide

  - people w/ AD: APP cleaved by beta secretase then gamma secretase resulting in amyloid beta 40, 42, C99 amino acid peptides
  - increased AB42 results in oligomers that cluster around meningeal and cerebral vessels and gray matter, forming plaques
  - neurofibrillary tangles --> formed by tau
  - tau normally binds to microtubules and has phosphates attached
  - phosphorylation of tau is altered in AD, increasing phosphorylation (mechanics not yet understood)
  - causes tau to detach from microtubules and be hyper-phosphorylated --> filamentous structures called paired helical filaments which are insoluble
  - tangles start in transentorhinal cortex spread to hippocampus then cover cerebral cortex
  - hippocampal pyramidal cells: granulo-vacuolar degeneration
  - decrease in synaptic buttons associated with vascular degeneration
  - increases the risk and degree/progression of dementia, but mechanisms not yet understood https://pmc.ncbi.nlm.nih.gov/articles/PMC7815481/
- Clinical Trials/next-gen therapies
  - Additional anti-amyloid (non-antibody) approaches: ALZ-801 - oral agent targeting AB oligomers; Phase 3 APOLLOE4 reported promising topline clinical and biomarker signals in 2025
  - multiple tau mAbs (e.g., semorinemab, tilavonemab) show biomarker effects with mixed clinical outcomes; field remains active (ongoing Phase 2/3 work; some negative readouts like bepranemab Phase 2). Tau aggregation inhibitors (e.g., HMTM/LMTX) continue under study with new analyses presented in 2025
  - Metabolic/inflammation targets (GLP-1s): oral semaglutide in EVOKE / EVOKE+ Phase 3 for early AD—main phase completion targeted for Sept 2025; observational data suggest reduced AD risk but RCT results pending.
  - Diagnostics enabling trials/earlier access: blood-based biomarkers (e.g., p-tau217 assays) and evolving FDA-cleared blood tests are entering practice/research pathways, complementing PET/CSF
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC11712804/
  - https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.14346? https://alzres.biomedcentral.com/articles/10.1186/s13195-024-01666-7?

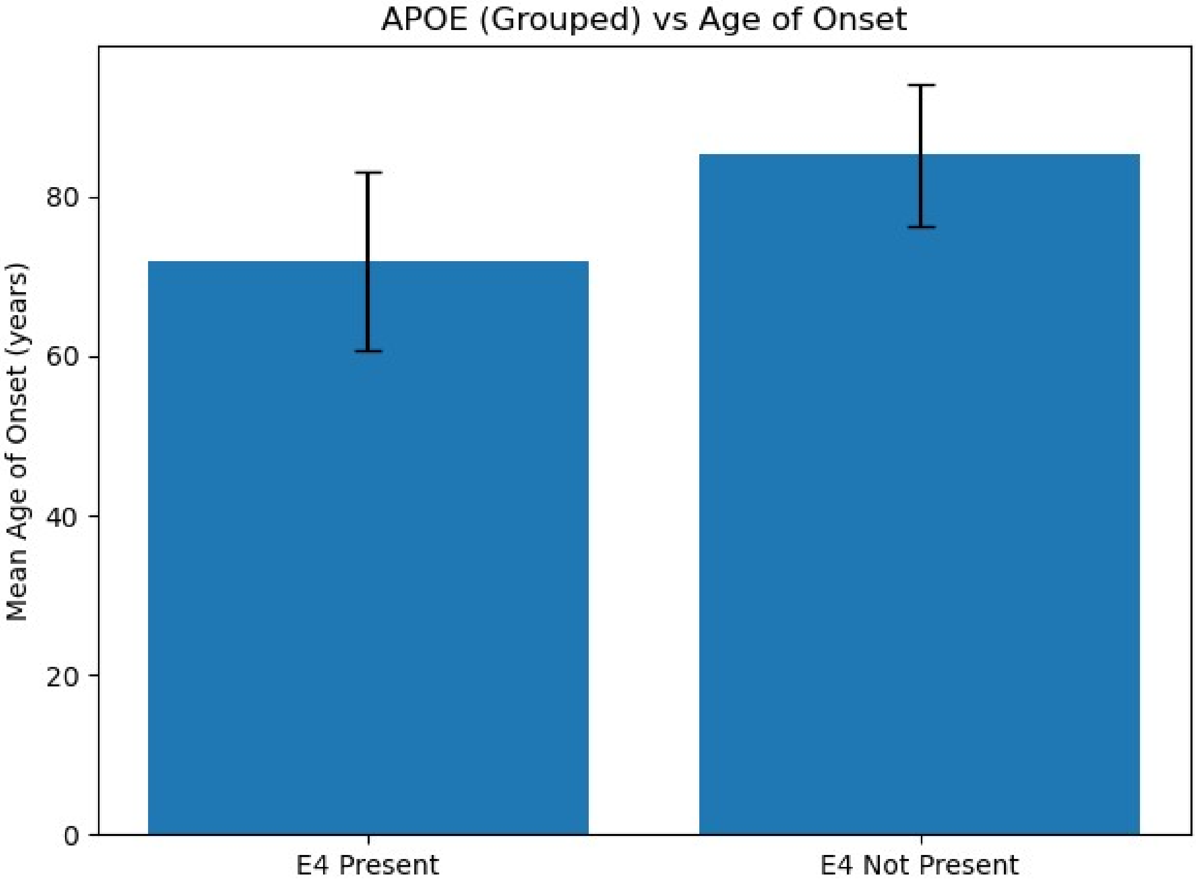
Data-Set:

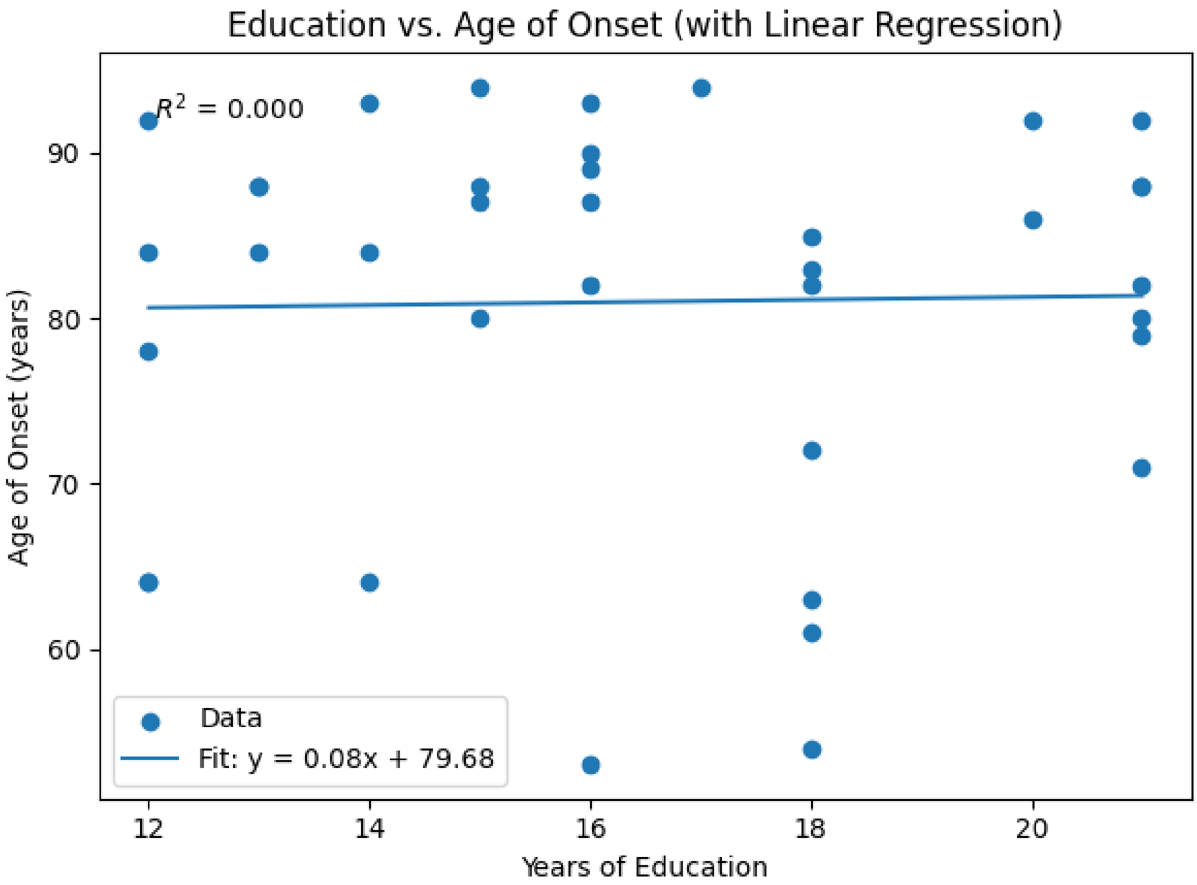
Data from the Allen Institute was used to determine the relationship between the APOE genotype and the age of onset of Alzheimer's. The Allen Institute's "Integrated multimodal cell atlas of Alzheimer's disease" used raw sequencing reads from ten unique public datasets, hosted on Synapse and the SRA, focused on the prefrontal cortex. Additionally, the Allen Institute's Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) team applied snRNA-seq, snATAC-seq, snMultiome, and MERFISH methods to study cells and nuclei from post-mortem human brain tissue provided by the Adult Changes in Thought (ACT) study and the University of Washington's Alzheimer's Disease Research Center (ADRC). Data from SEA-AD and the public databases include demographic information, diagnoses, and genotypes. Molecular data were expressed as gene expression counts, chromatin peak accessibility, and spatial coordinates. All data are from US donors.

Data Analyis:

Python was used to generate a bar graph comparing the age of symptomatic onset of subjects with a present APOE E4 allele and subjects without the APOE E4 allele. The bar graph shows that the average age of onset for subjects with the APOE E4 allele is lower than the subjects without the APOE E4 allele. A Students T-test was performed, producing a t-value of -3.6149 and a p-value of 0.002009. The p-value indicates a statistically significant difference between the groups is present on a 95% confidence interval.

-





- The linear regressions shows that there is not significant correlation between age of onset of symptoms and years of education. The R^2 value of 0.000 makes the lack of a linear relationship apparent.

```
import csv
import re
import matplotlib.pyplot as plt
from scipy import stats
import numpy as np

# ---------- Data model ----------
class Patient:
    all_patients = []  # collect every created Patient

    def __init__(self, donor_id, apoe, age_onset, age_diag, sex,
education):
        self.donor_id = donor_id
        self.apoe = apoe
        self.age_onset = age_onset
        self.age_diag = age_diag
        self.sex = sex
        self.years_education = education
        Patient.all_patients.append(self)

    def __repr__(self):
        return (f"Patient(ID={self.donor_id}, APOE={self.apoe}, "
                f"Onset={self.age_onset}, Diagnosis={self.age_diag}, "
                f"Sex={self.sex}, Education={self.years_education})")


# ---------- Helpers ----------
def _norm_header(s: str) -> str:
    """Normalize a header for resilient matching (lowercase, alnum
only)."""
    return re.sub(r'[^0-9a-z]+', '', (s or '').lower())

def _resolve(headers, candidates_norm_keys):
    """
    Resolve an actual header name from a list of candidate normalized
keys.
    Returns the matched real header or None.
    """
    norm_map = {_norm_header(h): h for h in headers if h is not None}
    for key in candidates_norm_keys:
        if key in norm_map:
            return norm_map[key]
    return None

def _clean_missing(val):
    """Map common missing markers to None; strip whitespace."""
    v = (val or "").strip()
    return None if v.lower() in {"", "na", "n/a", "nan", "none",
"null", "."} else v

def _clean_number_like(val):
    """
    Parse numbers: '81.0' -> 81.0 (float). Returns float or None.
    """
    v = _clean_missing(val)
    if v is None:
        return None
    try:
        return float(v)
    except:
        return None


# ---------- Loader (skips rows with NO onset-of-symptoms value)
----------
def load_patients_from_csv(filepath="NO DATE GENOTYPE Metadata
(1).csv"):
    """
    Create Patient objects from the CSV, skipping rows with NO onset-
of-symptoms value.
    """
    # Clear any prior loaded patients if you re-run
    Patient.all_patients.clear()

    with open(filepath, newline='', encoding="utf-8") as csvfile:
        reader = csv.DictReader(csvfile)
        headers = reader.fieldnames or []

        # Resolve columns robustly (tolerate spacing/case variations)
        donor_col = _resolve(headers, ["donorid", "donor_id", "id",
"donoridnumber"])
        apoe_col  = _resolve(headers, ["apoegenotype", "apoe",
"apoe_genotype", "apoe(genotype)"])
        onset_col = _resolve(headers, ["ageofonsetcognitivesymptoms",
"ageofonset", "onsetage", "ageatcognitivesymptomonset"])
        diag_col  = _resolve(headers, ["ageofdementiadiagnosis",
"ageatdiagnosis", "diagnosisage"])
        sex_col   = _resolve(headers, ["sex", "gender"])
        education_col = _resolve(headers, ["education",
"yearsofeducation", "yoe", "education(years)"])

        for row in reader:
            donor_id  = _clean_missing(row.get(donor_col, "") if
donor_col else None)
            apoe_raw  = _clean_missing(row.get(apoe_col,  "") if
apoe_col  else None)
            age_onset = _clean_number_like(row.get(onset_col, "") if
onset_col else None)
            age_diag  = _clean_number_like(row.get(diag_col,  "") if
diag_col  else None)
            sex       = _clean_missing(row.get(sex_col,   "") if
sex_col   else None)
            education = _clean_number_like(row.get(education_col, "")
if education_col else None)

            # Skip entries with NO onset-of-symptoms value
            if age_onset is None:
                continue

            # Normalize APOE string formatting (e.g., strip spaces)
            apoe = apoe_raw.replace(" ", "") if apoe_raw else None

            Patient(donor_id, apoe, age_onset, age_diag, sex,
education)


# ---------- Grouping, Printing, T-test, and Plot ----------
def apoe_group_label(apoe: str) -> str:
    """
    Per instructions:
      - Group '3/4' and '4/4' as 'E4 Present'
      - All other genotypes -> 'E4 Not Present'
    """
    if not apoe:
        return "E4 Not Present"
    apoe_norm = apoe.strip()
    return "E4 Present" if apoe_norm in {"3/4", "4/4"} else "E4 Not
Present"

def print_transformed_data():
    """
    Print the transformed data: Donor ID, E4-group label, Age of
Onset.
    """
    print("DonorID\tGroup\t\tAgeOnset")
    for p in Patient.all_patients:
        group = apoe_group_label(p.apoe)
        # Align group label for readability
        pad = "\t" if group == "E4 Present" else "\t"
        age_disp = int(p.age_onset) if (p.age_onset is not None and
float(p.age_onset).is_integer()) else p.age_onset
        print(f"{p.donor_id or ''}\t{group}{pad}{age_disp}")

def analyze_and_plot():
    """
    Create two groups (E4 Present vs E4 Not Present), run Student's t-
test,
    and make a 2-bar plot of mean Age of Onset for each group.
    """
    e4_present = []
    e4_not_present = []

    for p in Patient.all_patients:
        group = apoe_group_label(p.apoe)
        onset = p.age_onset
        if onset is None:
            continue
        if group == "E4 Present":
            e4_present.append(onset)
        else:
            e4_not_present.append(onset)

    # Basic safety checks
    if len(e4_present) == 0 or len(e4_not_present) == 0:
        print("Not enough data to compare groups.")
        print(f"E4 Present count: {len(e4_present)}, E4 Not Present
count: {len(e4_not_present)}")
        return
    # Means
    mean_present = sum(e4_present) / len(e4_present)
    mean_not = sum(e4_not_present) / len(e4_not_present)

# Standard deviation (instead of SEM)
    sd_present = np.std(e4_present, ddof=1)
    sd_not = np.std(e4_not_present, ddof=1)

# --- Bar plot with two bars (no specific colors) ---
    labels = ["E4 Present", "E4 Not Present"]
    means = [mean_present, mean_not]
    errors = [sd_present, sd_not]  # use SD

    plt.figure()
    plt.bar(labels, means, yerr=errors, capsize=5)
    plt.ylabel("Mean Age of Onset (years)")
    plt.title("APOE (Grouped) vs Age of Onset")
    plt.tight_layout()
    plt.show()

    # --- Student's t-test (Welch) ---
    tval, pval = stats.ttest_ind(e4_present, e4_not_present,
equal_var=False)

    print("\n--- Statistical Comparison (Student's t-test, Welch)
---")
    print(f"Group sizes: E4 Present n={len(e4_present)}, E4 Not
Present n={len(e4_not_present)}")
    print(f"Means: E4 Present = {mean_present:.2f}, E4 Not Present =
{mean_not:.2f}")
    print(f"t-value = {tval:.4f}")
    print(f"p-value = {pval:.6f}")
    if pval < 0.05:
        print("Result: Statistically significant difference at α =
0.05")
    else:
        print("Result: Not statistically significant at α = 0.05")


# ---------- Scatter + Linear Regression ----------
def plot_education_vs_onset():
    """
    Scatter of Years of Education vs. Age of Onset with a linear
regression line.
    Prints slope, intercept, R^2, p-value, and std error.
    """
    x = []  # years of education
    y = []  # age of onset
    for p in Patient.all_patients:
        if p.years_education is not None and p.age_onset is not None:
            x.append(p.years_education)
            y.append(p.age_onset)

    if len(x) == 0:
        print("No data available for education vs. onset.")
        return

    # Need at least 2 points for a regression
    if len(x) < 2:
        print("Not enough points for regression (need at least 2).")
        return

    x = np.array(x, dtype=float)
    y = np.array(y, dtype=float)

    # --- Linear regression ---
    slope, intercept, r_value, p_value, std_err = stats.linregress(x,
y)
    # Use a smooth line across the span of x for nicer display
    x_line = np.linspace(x.min(), x.max(), 200)
    y_line = slope * x_line + intercept
    r2 = r_value ** 2

    # --- Plot scatter + regression line ---
    plt.figure()
    plt.scatter(x, y, label="Data")
    plt.plot(x_line, y_line, label=f"Fit: y = {slope:.2f}x +
{intercept:.2f}")
    plt.xlabel("Years of Education")
    plt.ylabel("Age of Onset (years)")
    plt.title("Education vs. Age of Onset (with Linear Regression)")
    plt.legend()
    # Put R^2 on the plot
    plt.text(0.05, 0.95, f"$R^2$ = {r2:.3f}",
transform=plt.gca().transAxes,
             ha="left", va="top")
    plt.tight_layout()
    plt.show()

    # --- Report stats ---
    print("\n--- Linear Regression: Education vs. Age of Onset ---")
    print(f"Slope = {slope:.4f}")
    print(f"Intercept = {intercept:.4f}")
    print(f"R-squared = {r2:.4f}")


# ---------- Run everything ----------
if __name__ == "__main__":
    # 1) Load patients from the provided CSV, skipping rows with no
onset age
    load_patients_from_csv("NO DATE GENOTYPE Metadata (1).csv")

    # 2) Print transformed data (APOE collapsed into two groups; print
age of onset)
    print_transformed_data()

    # 3) Make two-bar plot and run t-test
    analyze_and_plot()

    # 4) Scatter + regression with R^2
    plot_education_vs_onset()
```

Verify and validate your analysis:

Our data analysis compares presence of the ε4 allele, often considered a genetic risk factor for Alzheimer's Disease (AD), with age of onset of AD symptoms. After running a Student's T-test, we found a t-value of -3.6149 and a p-value of .002009. This is statistically siginificant for a 95% confidence interval. Therefore, we concluded presence of the ε4 allele is correlated to a decrease in age of onset of symptoms. A 1997 study by Blacker et al. described a significant decrease in age of onset of Alzheimer's between subjects with two copies of the ε4 allele (average age of onset at 66.4 years old) and subjects with one or no copies of the ε4 allele (average age of onset at 72 years old). Similarly, a review by Liu et al. (2013) further asserts that the ε4 allele can accelerate AD conversion and progression. In conclusion, our results coincide with previous studies in indicating that the presence of the ε4 allele decreases the age of onset of symptoms of AD.

- https://pubmed.ncbi.nlm.nih.gov/9008509/#full-view-affiliation-1
- https://pmc.ncbi.nlm.nih.gov/articles/PMC3726719/?utm_source=chatgpt.com

Conclusions and Ethical Implications:

We found that the presence of the ε4 allele significantly descreased the age of onset of Alzheimer's across our data set. This conclusion suggests that genetic testing for the ε4 allele could provide useful information about patients' risks of early onset. Genetic testing has several ethical concerns including access to care, possible discrimination cases, and emotional distress in patients. Populations with the ε4 allele who also have access to genetic testing may be able to apply more preventative measures, such as implementing lifestyle changes associated with preventing AD, causing a greater discrepancy in age of onset across socioeconomic statuses. However, knowledge of the impact of the ε4 allele on onset of AD could impact insurance policies and employment benefits for those with the ε4 allele. Disclosure of this finding may require updated protections for those with the ε4 allele. Additionally, disclosure of the impact of the ε4 allele on onset age may cause patients emotional distress as there is no current cure to AD. On the contrary, some patients may take aforementioned steps to prevent early onset of AD symptoms.

Limitations and Future Work:

The data set used to perform this analysis of the effect of APOE genotype on the age of onset of AD only included subjects from the United States. Therefore, a future analysis using more diverse datasets is necessary to provide a comprehensive understanding of the impact of the presence of the ε4 allele on the age of onset of AD across demographics. Additionally, this analysis focused on the presence of the ε4 allele rather than the specific genotype. Further analysis should include comparison of age of onset between different genotypes to discuss whether number of ε4 alleles has a compounding effect on the age of onset.

Team Notes:

- 9/09/25: Began research on Alzheimer's Disease
- 9/11/25: Chose a research question (APOE genotype's effect on age of onset), continued research, discussed team communication
- 9/16/25: Began developing code to separate E4 carriers and non-carriers
- 9/17/25: Went to office hours to troubleshoot code, finished bar graph
- 9/18/25: Presented current findings and performed Students T-test
- 9/23/25: Created scatterplot data of age of onset vs years of education
- 9/25/25: Performed a linear regression on the scatterplot

Questions for TAs:

We have no questions at this point in time.