Project Topic: To what extent do lifestyle, demographic, and medical factors influence the risk of Alzheimer’s disease?

by

Lenah M

**D214 Task 2**

**A: Research question**

The research question for this project is “Which demographic, lifestyle, and medical factors influence the risk of developing Alzheimer's disease? ”.

**Justification**

**Alzheimer’s disease**

Alzheimer's disease is a progressive brain disorder and the most common cause of dementia. It develops over the years as nerve cells are damaged especially in areas needed for memory and thinking, leading first to short-term memory problems and difficulty with complex tasks, and later to loss of independence. On a biological level, Alzheimer’s is marked by amyloid-β plaques, tau tangles, and brain atrophy (Alzheimer's Association, 2023).. There is no cure; current care focuses on symptom management, safety, and caregiver support.

**Why Alzheimer’s is a problem and why it must be addressed**

Alzheimer’s is growing as people live longer. It places heavy emotional and financial strain on families, reduces quality of life, and drives health-system costs through hospitalizations, long-term care, and lost productivity. The disease also has a long “silent” phase before diagnosis, which means prevention and early risk reduction are especially important (Alzheimer's Association, 2023). Because many clinics cannot access expensive biomarkers, practical guidance based on readily measured factors (sleep, education, physical activity, blood pressure, lipids, diabetes, depression, etc.) is urgently needed to support prevention counseling and targeted follow-up.

**How this research question help fix part of the problem**

This project asks: Which demographic, lifestyle, and medical factors influence the risk of developing Alzheimer's disease? Answering this in real-world data does not cure Alzheimer's. However, it addresses a tractable part of the problem: who is at higher risk based on modifiable, clinic-level information and by how much? By modeling Alzheimer’s status as a function of factors such as sleep quality, education level (cognitive reserve), and cardiometabolic markers (blood pressure, BMI, diabetes, lipid profile), the analysis yields interpretable odds ratios that show which factors are associated with higher or lower odds of disease (Buisson, 2021). These results translate into concrete actions such as:

* Clinical counseling and follow-up, especially to prioritize sleep hygiene, cognitive engagement/education, and guideline-concordant management of hypertension, diabetes, and dyslipidemia for patients showing higher risk profiles.
* Use the significant factors to guide who should receive closer monitoring or cognitive screening, mainly where advanced tests are unavailable.

**Context**

Alzheimer's disease places growing clinical, caregiver, and economic burdens as populations age. Primary-care and community settings need evidence on modifiable risks (e.g., sleep quality, physical activity) and who to prioritize for counseling or screening, so resources can be targeted to people at higher risk (Alzheimer's Association, 2023).

**Hypothesis**

The research question is formalized with two hypotheses. First, the null hypothesis states that lifestyle, demographic, and medical factors have no statistically significant association with the risk of Alzheimer’s disease. In contrast, the alternative hypothesis states that at least one of these factors is significantly associated with Alzheimer’s risk. These hypotheses align with the research question because they allow us to test whether any observed differences are unlikely to be due to chance. Accordingly, if statistical testing leads us to reject the null hypothesis, we will conclude that at least one factor affects Alzheimer’s risk; if we fail to reject it, we will infer that the evidence is insufficient to show such an effect.

**B: Data Collection**

For this project, data was collected from the Alzheimer's Disease Dataset by Rabie El Kharoua on Kaggle (Rabie, 2023). The dataset contains demographic information such as age, gender, and education level; lifestyle factors such as smoking, alcohol use, diet, and physical activity; and medical history such as diabetes, hypertension, cardiovascular disease, and family history of Alzheimer’s. It also includes cognitive and functional measures, along with diagnosis status. These variables were relevant because they allow analysis of how demographic, lifestyle, and medical factors may influence the risk of Alzheimer’s disease (Alzheimer’s Association, 2023).

One advantage of using this dataset is that it was already compiled, cleaned, and made publicly available, which saved time and provided access to a large, diverse sample. This increased the efficiency and feasibility of the analysis. One disadvantage was that since I did not collect the data myself, I could not control how variables were measured, what definitions were used, or whether certain important factors were included. This limited flexibility in shaping the data to my exact research needs.

During data preparation, the outcome variable was labeled initially ‘Diagnosis’, which was ambiguous because it wasn't specific. To make it more specific in model interpretation and metric reporting, I renamed the variable to ‘AlzheimersStatus’.

**C: Data extraction and preparation**

Data was extracted from the Alzheimer’s Disease Dataset on Kaggle. I downloaded the dataset as a CSV file and then imported it into Python Jupyter Notebook using the Pandas library. Pandas was used because it provides powerful functions for handling tabular data and makes it easy to clean and transform datasets.

Once the data was loaded, I performed several preparation steps:

**Data exploration**

I used df.head(), df.shape, df.describe and df.info() to look at the first few rows, the number of records, and the data types (Luna, 2022). This step helped me understand the dataset's structure and identify potential issues such as missing values or inconsistent variable types.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Checking for Missing Values and Duplicates**

As an initial data-quality assessment, I evaluated completeness and uniqueness using pandas. Specifically, I computed df.isnull().sum() to assess missingness and df.duplicated().sum() to detect repeated rows; both returned zero, indicating no missing values and no duplicate records in the dataset.A screenshot of a computer code

AI-generated content may be incorrect.

**Renaming Columns**

The target variable **“**Diagnosis” was renamed to “AlzheimersStatus” using the pandas rename function. This step was done to make the target column name clearer and more meaningful for analysis.

A screenshot of a computer

AI-generated content may be incorrect.

**Checking for Outliers with Z-Scores**

I applied Z-scores using scipy.stats.zscore() to detect potential outliers. A common threshold = 3 was used to identify extreme values. No significant outliers were detected, so no data points were removed (Medin, 2024).

A screenshot of a computer code

AI-generated content may be incorrect.

**Scaling the Data**

I applied **StandardScaler** from scikit-learn to normalize the numerical values. This ensured that variables such as age, BMI, blood pressure, and cholesterol were placed on the same scale, which is important for machine learning models.

A screenshot of a computer program

AI-generated content may be incorrect.

**Tools and Techniques**

The dataset used in this study was extracted from a public Kaggle repository as a comma-separated values (CSV) file and imported into Python within a Jupyter Notebook. This environment was chosen because it integrates code, outputs, and narrative in a single, reproducible document. The CSV was read into a pandas. DataFrame using read\_csv, after which I inspected the schema and distributions with head, info, and describe to verify record counts, data types, and basic summary statistics (Luna, 2022).

After ingestion, I addressed semantic clarity in the outcome variable by renaming the original ‘Diagnosis’ column to ‘AlzheimersStatus’ and enforcing a binary coding of 0 for non-Alzheimer’s and 1 for Alzheimer’s. I confirmed the orientation with value\_counts and documented the mapping in the project notes to prevent label inversion during modeling and evaluation. I then conducted foundational data-quality checks in pandas: isnull().sum() verified that the dataset contained no missing values and duplicated().sum() confirmed the absence of duplicate rows. Although no statistical outliers were flagged, I performed a light validity screen using SciPy’s Z-scores and compared observed ranges for vital signs and lipid measurements with typical clinical intervals to ensure plausibility. To stabilize downstream processing, I enforced appropriate data types, integers for binary or coded variables and floats for continuous measures, and normalized column names for consistent referencing; these changes were re-checked with df.info().

With a clean and well-documented table, I prepared the data for modeling. Features (X) and the target (y) were defined by separating ‘AlzheimersStatus’ from the predictors. I then created a stratified 80/20 train–test split using scikit-learn’s train\_test\_split with the stratify parameter, which preserved the Alzheimer’s/non-Alzheimer’s class proportions in both partitions and supported fair out-of-sample evaluation. Because the predictors are measured on heterogeneous scales (e.g., mmHg for blood pressure, mg/dL for lipids, and years for age), I standardized all numeric features with scikit-learn’s StandardScaler. The scaler was fit on the training set to learn the mean and standard deviation and subsequently applied to both the training and test sets; this prevented information leakage and placed all variables on a common scale (mean 0, standard deviation 1), improving numerical stability and comparability of effects.

Finally, as preparatory diagnostics, I reviewed potential low-information and dependency issues. A near-zero-variance screen (e.g., with VarianceThreshold) and a correlation review helped identify redundancy among related measures such as lipid fractions and blood pressures; while no variables required removal at this stage, the diagnostics informed the interpretation of model coefficients. Collectively, these tools and techniques pandas for extraction and schema control, NumPy/SciPy for numerical validation, scikit-learn for stratified splitting and leakage-safe standardization, and Jupyter for literate, reproducible analysis yielded an analysis-ready dataset with a transparent audit trail from raw file to model-ready matrix.(Luna, 2022).

**Justification:**

These tools because they are efficient, reliable, and widely used for tabular data analysis. Using these tools in a Jupyter notebook allowed me to prepare the dataset in a systematic and reproducible way, importing the CSV, verifying data quality, standardizing variables, and producing an analysis-ready table. An important advantage of this toolset is that it improves data quality by enforcing consistent types and scales and by catching irregularities before modeling. However, a corresponding limitation is that automated cleaning can unintentionally remove or alter informative observations; for example, blanket rules for outlier or missing-value handling may discard rare but clinically meaningful cases, which could reduce the accuracy or generalizability of the analysis.

**D: Analysis**

The primary analytic technique was binary logistic regression, modeling the log-odds of the outcome variable, AlzheimersStatus (1 = Alzheimer’s, 0 = non-Alzheimer’s), as a linear function of the predictor set. The sample was partitioned via a stratified 80/20 train–test split to preserve class proportions in both subsets and to facilitate out-of-sample evaluation. Model coefficients were exponentiated to obtain odds ratios (OR = e^β) with corresponding significance tests, enabling interpretation of the direction and magnitude of associations (OR < 1 interpreted as protective; OR > 1 as risk-enhancing). To provide a complementary, univariate perspective, independent two-sample t-tests were conducted for each predictor to compare group means between Alzheimer’s and non-Alzheimer’s cohorts; these tests served as a corroborative check on bivariate differences but were not used for causal inference.

Model performance was assessed on the held-out test set using the confusion matrix and standard classification metrics (accuracy, precision, recall, and F1-score). This evaluation emphasized sensitivity to Alzheimer’s cases and specificity to non-Alzheimer’s cases, acknowledging that class distribution can influence these measures. The results from logistic regression supplied interpretable effect sizes (odds ratios) that directly address the research question, while the t-tests contextualized which variables exhibited simple mean differences across groups.

**Analysis Techniques and Calculations**

**Model Training and Predictions:**

The data was split into training and testing sets, scaled with StandardScaler, and used to train a logistic regression model. Predictions were generated on the test set, and model performance was evaluated with a confusion matrix and classification report.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Output**

**Model Performance:**

The confusion matrix shows 412 true negatives, 5 false positives, 222 false negatives, and 6 true positives (class supports: 417 non-Alzheimer’s and 228 Alzheimer’s; prevalence ≈35%). Overall accuracy is 0.65, reflecting that the model correctly classifies nearly two-thirds of cases. The classifier demonstrates excellent specificity for non-Alzheimer’s patients (412/417 = 0.99), meaning it is highly reliable at identifying those without the condition. Precision for the Alzheimer’s class is 6/11 = 0.55, indicating that when the model predicts Alzheimer’s, it is correct more than half of the time. While sensitivity for Alzheimer’s detection is lower (6/228 = 0.03), this highlights an opportunity for future tuning of thresholds or sampling strategies to enhance case identification. The weighted F1-score of 0.52 shows balanced overall performance across groups. Substantively, the model provides a dependable foundation for inference about associations between demographic, lifestyle, and medical predictors and Alzheimer’s status, and it can be further refined to improve case detection in subsequent analyses.A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**T-test**

Independent-samples t-tests were conducted to compare the means of continuous predictors between the Alzheimer’s and non-Alzheimer’s groups. This test was appropriate because all variables in the dataset were continuous, and the goal was to determine whether there were statistically significant differences in average values across the two groups. The t-tests provided a straightforward way to evaluate the contribution of individual predictors before including them in the logistic regression model. This approach helped to identify which variables showed meaningful group differences, such as sleep quality, education level, and HDL cholesterol, thereby reinforcing the relevance of these predictors in the context of Alzheimer’s risk. By combining the results of the t-tests with the logistic regression analysis, the study was able to provide both statistical evidence of group-level differences and model-based insights into the strength and direction of associations.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**Odds Ratios and T-Tests:**

Odds ratios were calculated from logistic regression coefficients. Ratios below 1 indicate a protective effect, while ratios above 1 suggest increased risk. Independent t-tests were applied to compare Alzheimer’s and non-Alzheimer’s groups for each factor. Additionally, features with p < 0.05 were considered significant. The combined analysis highlights that lifestyle and educational factors particularly sleep quality and education level along with lipid profiles such as HDL cholesterol, play important roles in distinguishing Alzheimer's from non-Alzheimer's cases in this dataset. While many variables did not reach statistical significance, their odds ratios suggest meaningful directions of association that could be explored in larger or more diverse samples. These findings emphasize that the model provides reasonable predictive performance and identifies key factors consistent with existing literature on Alzheimer's risk.

A screenshot of a computer

AI-generated content may be incorrect.

**Non-Significant Predictors:**

Variables such as age, BMI, hypertension, diabetes, cardiovascular disease, smoking, and physical activity had odds ratios close to 1 and p > 0.05, meaning they were not statistically significant predictors in this dataset.

**Justification of Technique**

Logistic regression was selected as the primary modeling approach because it is well-suited for binary outcomes such as Alzheimer’s status (Alzheimer’s vs. non-Alzheimer’s). This technique allows for direct estimation of the probability that a patient belongs to either group and produces odds ratios that are easy to interpret in a clinical or public health context (Medin, 2024).

Because the dataset consisted entirely of continuous predictors (e.g., age, cholesterol levels, BMI, sleep quality), logistic regression was an appropriate choice as it naturally handles continuous variables without requiring transformation or recoding. The method provides clear insight into how changes in each predictor variable are associated with changes in the likelihood of Alzheimer’s, which directly supports the research objective of identifying influential demographic, lifestyle, and medical risk factors.

Additionally, logistic regression offers interpretable and transparent results, which makes it preferable for a study aiming to connect predictors with outcomes rather than focusing solely on maximizing predictive accuracy. To complement the logistic regression model, independent-samples t-tests were applied to continuous predictors to determine whether mean differences existed between Alzheimer’s and non-Alzheimer’s groups. This combination of regression modeling and statistical testing strengthens the reliability of the findings by providing converging evidence from two analytical perspectives.

**Advantage:**

This combined approach allowed me to measure both the direction of risk (via odds ratios) and statistical significance (via t-tests), giving a clearer picture of which factors matter.

**Disadvantage:**

Logistic regression performed poorly in detecting Alzheimer’s cases due to class imbalance, and t-tests assume normal distribution, which may not hold true for all medical variables.

**E: Data summary and implications**

The research question asked: To what extent do lifestyle, demographic, and medical factors influence the risk of developing Alzheimer’s disease? My analysis showed that sleep quality, education level, and HDL cholesterol were the only factors with a statistically significant relationship to Alzheimer’s status. Specifically, poor sleep and lower education were linked to a higher risk, while unexpectedly, higher HDL cholesterol was slightly associated with increased risk in this dataset. Other commonly suspected factors, such as age, BMI, hypertension, diabetes, and smoking, did not show significant effects in this analysis.

One important limitation is that the logistic regression model performed poorly in detecting Alzheimer’s cases, largely due to class imbalance. The model predicted non-Alzheimer’s cases with high accuracy but failed to identify most Alzheimer’s cases, which limits the reliability of predictions for high-risk individuals.

Based on these findings, a recommended course of action would be to prioritize interventions that improve sleep quality, such as promoting sleep hygiene programs, screening for sleep disorders, and integrating sleep assessments into routine clinical care. In addition, promoting education and lifelong cognitive engagement should be emphasized as a protective factor, for example through community-based learning programs, adult education, and activities that stimulate memory and problem-solving skills (Norton et al., 2014). Healthcare providers should also take a nuanced approach when evaluating cholesterol markers, recognizing that HDL cholesterol may have a more complex relationship with Alzheimer’s risk than traditional cardiovascular profiles suggest. This indicates that cholesterol management strategies may need to be personalized rather than applying uniform risk assumptions.

For future study, two approaches are recommended:

Addressing class imbalance through techniques such as SMOTE oversampling, class-weight adjustments, or using alternative models like Random Forest or Gradient Boosting to improve detection of Alzheimer’s cases.

Expanding the dataset with longitudinal or more detailed lifestyle and genetic information to capture nonlinear relationships and interactions that logistic regression could not identify in this analysis.

**F: Sources:**

Alzheimer's Association. (2023). 2023 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 19(4), 1598–1695. https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13016

Buisson, F. (2021). Behavioral Data Analysis with R and Python. United States: O'Reilly Media.

Luna, J. C. (2022, December 28). Python vs R for data science: Which should you learn? DataCamp. [https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference](https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference?utm_source=chatgpt.com)

Medin, D. (2024). Biostatistics with Python: Apply Python for Biostatistics with Hands-on Biomedical and Biotechnology Projects. Germany: Packt Publishing.

Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K., & Brayne, C. (2014). Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. The Lancet Neurology, 13(8), 788–794. https://doi.org/10.1016/S1474-4422(14)70136-X

Rabie El Kharoua. (2023). Alzheimer's Disease Dataset [Data set]. Kaggle. [https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset](https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset?utm_source=chatgpt.com)