Predicting Smoking and E-Cigarette Habits Using Personal Health History

Brianna Hong
Roaya Elhassani
Audrija Bhattacharya
Ariel Choi
Professor Harlin Lee

UNC | SCHOOL OF DATA SCIENCE AND SOCIETY

MOTIVATION

This project aims to explore whether a person's health history can help predict their cigarette or e-cigarette usage. While nicotine use is widely known to cause serious health issues, we wanted to test the reverse:

Can a person's health history help predict whether they are a smoker or e-cigarette user?

Our goal is to use hypothesis tests to explore whether key health indicators—such as past heart attack, stroke, or asthma—contain meaningful predictive patterns about an individual's nicotine usage.

DATA COLLECTION & PREPARATION

We used the CDC's "Heart Disease Prediction 2022" dataset from Kaggle, containing 445,132 self-reported patient records.

Key Features Selected:

•Health indicators: HadHeartAttack, HadStroke, HadAsthma

•Target labels: SmokerStatus, ECigaretteUsage

To prepare the dataset for analysis, we first filtered the original data to keep only the relevant columns: smoking status, e-cigarette usage, and health condition indicators. Next, we encoded all categorical responses, such as smoking status and disease history, into numerical values for model compatibility. Any missing entries were handled by removing incomplete rows to maintain data consistency. Finally, we created two new features: *Both*, which combines smoking and e-cigarette usage into a single average score, and *Disease*, which encodes whether an individual experienced heart attack, stroke, asthma, multiple conditions, or none at all.

DATA EXPLORATION

Using the newly created features, *Both* and *Disease*, we explored the distributions and relationships in the refined dataset. We observed that the *Both* score revealed a spectrum of nicotine use, with higher scores indicating more frequent or combined smoking and e-cigarette behavior. The *Disease* feature allowed us to efficiently categorize patients by health history. Asthma was the most common single condition in the dataset, while cases involving multiple diseases were less frequent.

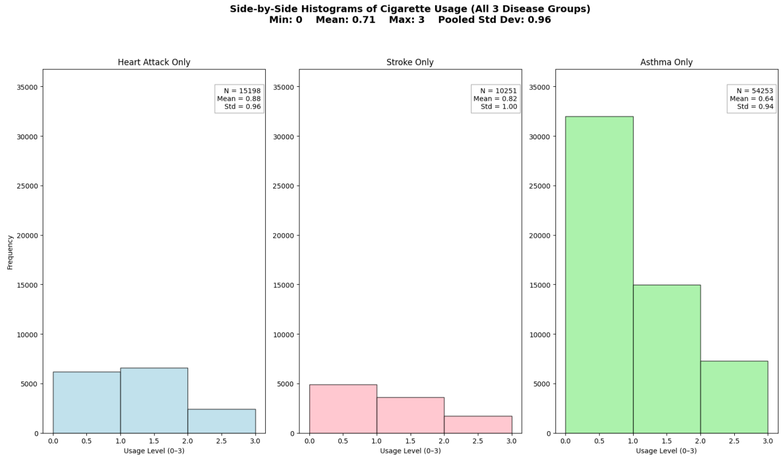
	SmokerStatus	ECigaretteUsage	HadHeartAttack	HadStroke	HadAsthma	Both	HA	ST	AS	Disease
3	2.0	0.0	No	No	Yes	1.0	0.0	0.0	1.0	3
12	1.0	0.0	No	No	Yes	0.5	0.0	0.0	1.0	3
17	1.0	0.0	No	No	Yes	0.5	0.0	0.0	1.0	3
18	0.0	0.0	No	Yes	No	0.0	0.0	1.0	0.0	2
26	2.0	0.0	Yes	No	No	1.0	1.0	0.0	0.0	1

The table above shows a sample of what the data looked like after we built these features. The Both column averages smoking and e-cigarette use into one score, and the Disease column groups health history into clear categories.

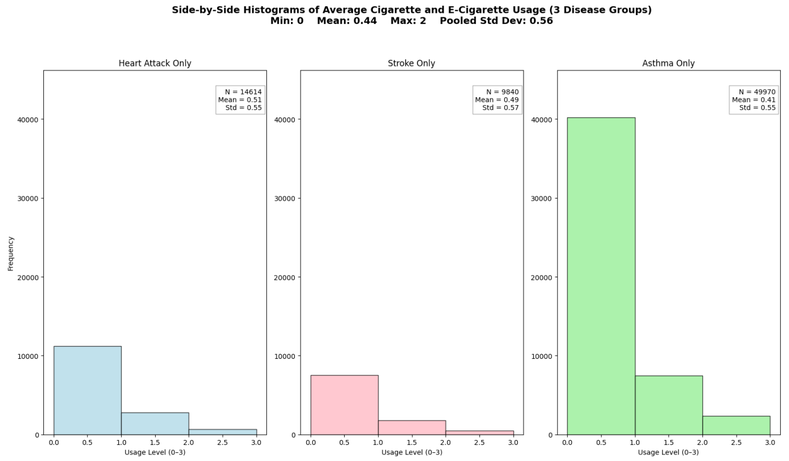
MODEL & EVALUATION

The model built examines nicotine use patterns across three health outcomes: heart attack, stroke, and asthma. Nicotine exposure is represented using three separate measures: traditional cigarette use, e-cigarette use, and a combined average score.

The model compares smoking habits across three diseases and tests whether the differences are real and not due to chance. Each measure was converted to a numeric score reflecting frequency of use (0–3 for cigarettes, 0–1 for e-cigarettes). For each disease group, we calculated the average level of nicotine use and tested whether these group means differed significantly using ANOVA testing to compare averages across the groups.



This set of histograms shows cigarette usage levels among participants with heart attack, stroke, or asthma. Mean usage is highest in the Heart Attack group and lowest in the Asthma group.



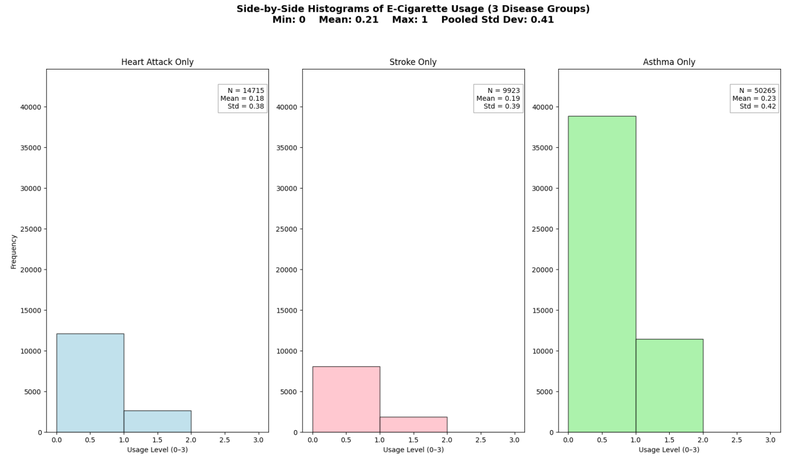
This set of histograms represents the combined average of cigarette and e-cigarette use for each disease group. The Heart Attack group shows slightly higher combined usage than Stroke and Asthma groups.

From these visualization, we can make a few key observations:

- The asthma group has a significantly larger sample size for all three types of nicotine usage
- The distribution of cigarette usage levels is visibly different in those who experienced a heart attack only than a stroke only or asthma only

Some Observations that May Influence Our Statistical Results:

- The numerical values associated with smoking categories have very little variation
- The overall sample size Is very large
- The sample size of each group is not similar
-



These histograms illustrate e-cigarette usage among the three disease groups. Overall usage is low, the majority of participants reporting no e-cigarette use. The Asthma group shows slightly higher average e-cigarette use.

RESULTS

The results of each ANOVA test:

- Cigarettes: F = 436.97, p < 0.00001
- E-Cigarettes: F = 102.75, p < 0,0001
- Both: F = 230.45, p < 0.0001

For all three nicotine categories, under the assumption that the mean usage level is the same in each disease group, there is almost a 0% likelihood of observing the distributions presented in this sample. At a 0.05 significance level, we can safely reject the null assumption, and our results are consistent with there being a difference in means between disease categories for each nicotine usage type.

Visually, we can observe that:

- People with a history of heart attack or stroke reported the highest levels of cigarette smoking
- Asthma patients showed lower overall nicotine use
- E-cigarette use was minimal across all groups and contributed little to overall patterns

While health history does show a clear statistical relationship with smoking behavior, it alone is not strong enough to reliably predict whether an individual smokes. These findings reinforce the already established link between smoking and cardiovascular health. A future model could benefit from including additional lifestyle features to build more a complete predictive model.