# Predicting Smoking and E-Cigarette Habits from Personal Health history

**DATA110 Final Project - Heart Disease**

Team BARBZ: Brianna, Roaya, Audrija, Ariel

**Group Meetings**

- 10/31 at Alumni Hall: Brianna, Roaya, Audrija
- 11/07 at Alumni Hall: Brianna, Roaya, Audrija
- 11/14 at Alumni Hall: Brianna, Roaya, Ariel
- 11/25 on Zoom: Brianna, Audrija, Roaya, Ariel
- 11/30 on Zoom: Brianna, Audrija, Roaya, Ariel
- 12/1 in Gardner Hall: Brianna, Audrija, Roaya, Ariel

## 7 stages of Data Science

1. **Defining the Problem:** Does cigarette or e-cigarette use affect the risk of heart attack, stroke, or asthma?
    a. Null hypothesis: Cigarette and e-cigarette usage does not change the probability of a patient experiencing heart disease, stroke, or asthma.
    b. Alternative hypothesis: Cigarette and e-cigarette usage does affect the probability of a patient experiencing heart disease, stroke, or asthma.
2. **Data collection:** Our data is coming from the CDC after they collected information on 445,132 patients. The samples are the patients and features are the columns that categorize them. There are 40 features that include things like state, sex, weight, number of physical days, and whether or not they have experienced specific health conditions. Due to a large sample size and small differences in the data points due to reassigning categorical to numerical, we have arrived at a falsely inflated F statistic and false deflated p-value. We also found that there was an uneven number of patients who had one disease alone, specifically there were more patients with asthma alone, leading to potential skewness.

```
data.shape
```

```
(445132, 40)
```

3. **Data Preparation:** In order to clean data, we decided to drop all columns other than SmokerStatus, ECigaretteUsage, HadHeartAttack, HadStroke, and HadAsthma. This was to simplify the data and make it easier to use because we were trying to decipher a relationship between the two. We did this by using data.filter[] and inserting the names of the columns we wanted to keep.

```
fn = 'heart_2022_with_nans'
data = pd.read_csv(f'~/110-F25/project/C-HeartDisease/{fn}.csv')
df = data.filter(['SmokerStatus','ECigaretteUsage','HadHeartAttack','HadStroke','HadAsthma'])
df.head()
```

| | SmokerStatus | ECigaretteUsage | HadHeartAttack | HadStroke | HadAsthma |
|---|---|---|---|---|---|
| **0** | Never smoked | Not at all (right now) | No | No | No |
| **1** | Never smoked | Never used e-cigarettes in my entire life | No | No | No |
| **2** | Never smoked | Never used e-cigarettes in my entire life | No | No | No |
| **3** | Current smoker - now smokes some days | Never used e-cigarettes in my entire life | No | No | Yes |
| **4** | Never smoked | Never used e-cigarettes in my entire life | No | No | No |

We then wanted to simplify the range of answers in the SmokeStatus and ECigaretteUsage columns into numbers in order to calculate functions on them. We first started by assigning the possible categorical values in a corresponding numerical descent. This includes ‘Never Smoked’ to the value 0, ‘Former Smoker’ to the value 1, ‘Current smoker - now smokes some days’ to the value 2, and ‘Current smoker - now smokes every day’ to the value 3. We did the same with e-cigarette usage.

```
smoker_lookup = {
    "Never smoked": 0,
    "Former smoker": 1,
    "Current smoker - now smokes some days": 2,
    "Current smoker - now smokes every day": 3
}

ecig_lookup = {
    "Never used e-cigarettes in my entire life": 0,
    "Not at all (right now)": 1,
    "Some days": 2,
    "Every day": 3
}
```

Then, in order to change the values in the DataSeries, we created smoker_nums and ecig_nums, both lists for the two columns to make it easier to replace the columns. We looped over each row using a for loop and got the categorical value for SmokerStatus and ECigaretteUsage in the i row. Using .get(), the code is able to look for the corresponding numerical value based on the categorical string and appends the value to their respective lists. Finally, we reassigned the SmokerStatus column with the smoker_nums list and ECigaretteUsage with ecig_nums.

```
smoker_nums = []
ecig_nums = []

# looping through the columns to assign the numerical value
for i in range(len(df)):
    smoker_text = df.loc[i, "SmokerStatus"]
    ecig_text = df.loc[i, "ECigaretteUsage"]

    smoker_nums.append(smoker_lookup.get(smoker_text))
    ecig_nums.append(ecig_lookup.get(ecig_text))

df["SmokerStatus"] = smoker_nums
df["ECigaretteUsage"] = ecig_nums
```

4. **Data Exploration**: In terms of data exploration for our project, we created two new columns based on the data given to us that the rest of our project is based on. The first is the 'Both' column, which returns the mean of the numerical value of SmokeStatus and ECigaretteUsage in each row. This gives us a combined measure of the two behaviors, which will help us determine if one or both behaviors lead to an increased chance of heart attack, stroke, or asthma. The combined measure simplifies analysis because we now have one feature that summarizes the risk while also somewhat capturing overlap of users who both smoke and use e-cigarettes, making it useful for modeling.

```
# adding both column - the mean of SmokerStatus and ECigaretteUsage
df["Both"] = (df["SmokerStatus"] + df["ECigaretteUsage"]) / 2

df.head()
```

| | SmokerStatus | ECigaretteUsage | HadHeartAttack | HadStroke | HadAsthma | Both |
|---|---|---|---|---|---|---|
| **0** | 0.0 | 1.0 | No | No | No | 0.5 |
| **1** | 0.0 | 0.0 | No | No | No | 0.0 |
| **2** | 0.0 | 0.0 | No | No | No | 0.0 |
| **3** | 2.0 | 0.0 | No | No | Yes | 1.0 |
| **4** | 0.0 | 0.0 | No | No | No | 0.0 |

Our second new column was the 'Disease' column. Similar to SmokerStatus and ECigaretteUsage, this column assigned the specific disease with a numerical value. We first started by assigning 'Yes' to the value 0 and 'No' to the value 1. Since all three columns (HadHeartAttack, HadStroke, and HadAsthma) only had yes or no values, we were able to generalize this code to all three diseases.

```
disease_lookup = {
    "No": 0,
    "Yes": 1,
}
```

We then followed the same code as with SmokeStatus and ECigaretteUsage to convert the categorical data into numerical data. We created lists for the three diseases and looped through each row to append 0s and 1s to the corresponding list if that $i^{th}$ patient had the respective disease. I.e. if the $i^{th}$ patient had experienced only a heart attack, HA_nums[i] would return 1. We created new columns ('HA', 'ST', 'AS') that now contain 0 and 1s depending on if they experienced that disease.

```
HA_nums = []
ST_nums = []
AS_nums = []

for i in range(len(df)):
    HA_text = df.loc[i, 'HadHeartAttack']
    ST_text = df.loc[i, 'HadStroke']
    AS_text = df.loc[i, 'HadAsthma']

    HA_nums.append(disease_lookup.get(HA_text))
    ST_nums.append(disease_lookup.get(ST_text))
    AS_nums.append(disease_lookup.get(AS_text))

df['HA'] = HA_nums
df['ST'] = ST_nums
df['AS'] = AS_nums
```

We then created a new list which would be more comprehensive and will eventually hold a list of numbers 0-4, which will indicate their disease history. We looped through each patient again to determine if they **only** had a heart attack, **only** had a stroke, or **only** had asthma. For example, if the $i^{th}$ patient had HA = 0, ST = 0, AS = 0, they would have not experienced any of the 3 diseases and would return 0. If the $i^{th}$ patient had HA = 1, ST = 0, AS = 0, they would have experienced only a heart attack and would return 1. If the $i^{th}$ patient had HA = 0, ST = 1, AS = 0, they would have experienced only a stroke and would return 2. If the $i^{th}$ patient had HA = 0, ST = 0, AS = 1, they would have experienced only asthma and would return 3. If the $i^{th}$ patient had experienced a combination of more than one of the diseases, they were assigned the value 4.

```
disease_nums = []

for i in range(len(df)):
    HA = df.loc[i, 'HA']
    ST = df.loc[i, 'ST']
    AS = df.loc[i, 'AS']

    if HA + ST + AS == 0:
        # no disease
        disease_nums.append(0)
    elif HA == 1 and ST == 0 and AS == 0:
        # heart attack only
        disease_nums.append(1)
    elif HA == 0 and ST == 1 and AS == 0:
        # stroke only
        disease_nums.append(2)
    elif HA == 0 and ST == 0 and AS == 1:
        # asthma only
        disease_nums.append(3)
    else:
        # multiple diseases
        disease_nums.append(4)

df['Disease'] = disease_nums
```

| | 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 |

5. **Model Building:**

In order to test our hypothesis and evaluate whether cigarette use, e-cigarette use, or combined nicotine exposure differ across health-history groups, we utilized one-way ANOVA as our primary statistical model. Although decision trees or logistic regression are more common methods of classification, ANOVA was more in line with the structure of our data because we are investigating the statistical significance of the differences in means of a continuous variable (usage score) across multiple categorical groups (disease types). Prior to modeling, we created three quantitative variables that could serve as the response variable in our analysis. The first variable, SmokerStatus, encoded cigarette use on a four-point scale ranging from 0 (Never) to 3 (Smokes every day). The second variable, ECigaretteUsage was also in line with this structure, evaluating e-cigarette use on the same scale. Based on these two variables, we were able to explore the traditional and contemporary forms of nicotine exposure separately. However, because many individuals may smoke cigarettes, e-cigarettes, or engage in both behaviors, we also created the third measure “Both” that represents the numerical average of smoking and vaping values. Based on this combined score, we observed nicotine exposure better.

Our aim was not in predicting an individual’s nicotine habit but rather in determining whether meaningful differences exist between people who have experienced different diseases: heart attack only, stroke only, or asthma only. These three groups were selected from the larger dataset because it would allow us to analyze clear and mutually exclusive disease categories without complications that could be introduced by individuals experiencing multiple conditions. To be more detailed, our independent variable (Disease) contains three non-overlapping categories: heart attack only, stroke only, and asthma only. These groups are each treated mutually exclusive in order to isolate potential relationships between each disease experience and nicotine behavior. Each of the three nicotine related variables, SmokerStatus, ECigaretteUsage, and Both are analyzed separately, resulting in three ANOVA models. Because the size of the dataset is extremely large and exceeds 445,000 participants, ANOVA’s stability to uneven group sizes and large samples make ANOVA the best way to organize the data with. In addition, our data showed clear distributional patterns that visually suggest differences between groups, especially in cigarette use, reinforcing ANOVA.

The statistical model treated nicotine use as the outcome and disease category factor, so for each model, we extracted the subset of participants that belong to the three disease groups and compared their mean smoking, vaping, and combined usage scores. The ANOVA implementation in Python calculated F-statistic and p-value for each outcome variable and returned group means for interpretation. By structuring our analysis this way, we ensured that our conclusions regarding group differences were not only based on visual inspections from histograms but also on formal statistical evidence capable of detecting whether any observed differences were unlikely to have occurred by chance.

6. **Model Evaluation:**

After constructing the ANOVA models, we evaluated the results in order to determine if nicotine use truly differs across individuals with different disease histories. The output of each ANOVA provided an F-statistic and p-value, which indicated whether the variation in nicotine use across groups was greater than the natural variation that was expected within each group. As our dataset was extraordinarily large, even small numerical differences could produce extremely small p-values, so our we had to both examine statistical significance and understand the practical magnitude of group differences for our interpretation

| Test Variable | F-Statistic | P-Value | Result |
|---|---|---|---|
| **SmokerStatus** | **436.97** | **0.000000** | **Significant** |
| **ECigaretteUsage** | **102.75** | **0.000000** | **Significant** |
| **Both** | **230.45** | **0.000000** | **Significant** |

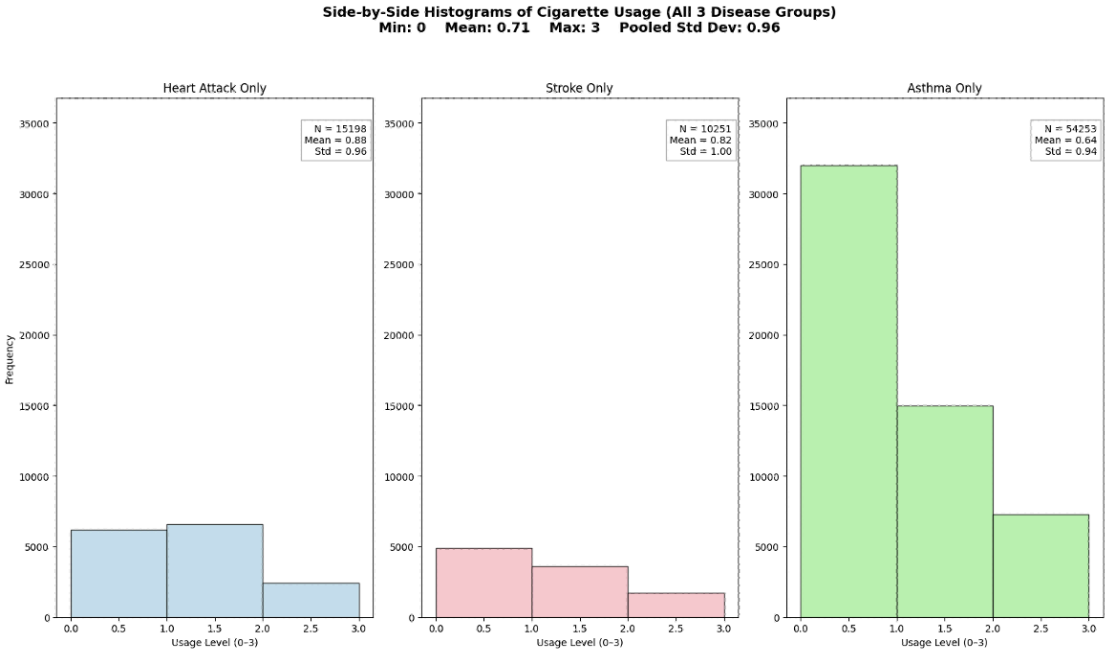
### 6.1 Cigarette Use (SmokerStatus)

The first model that was considered was the traditional cigarette model. The ANOVA for SmokerStatus has a very large F-statistic and an exceptionally small p-value that was below 0.05, which shows the statistical significant difference between the heart attack, stroke, and asthma groups. Based on the means, the Heart Attack Only group exhibited

the highest average cigarette use, followed by the Stroke Only group and Asthma only group being lowest. These numerical trends could also be observed through our histograms, where heart=attack patients showed a wider range of moderate to heavy smoking behavior, while asthma patients were mostly clustered around 0. This indicates that people with a history of major cardiovascular events generally report higher smoking rates than those that only have respiratory conditions.

```
======================================
ANOVA: SmokerStatus by Disease
======================================
F-statistic: 436.9704
P-value:     0.000000

Group Means:
Disease 1 (Heart Attack Only): 0.8776
Disease 2 (Stroke Only):       0.8166
Disease 3 (Asthma Only):       0.6421
--------------------------------------
```



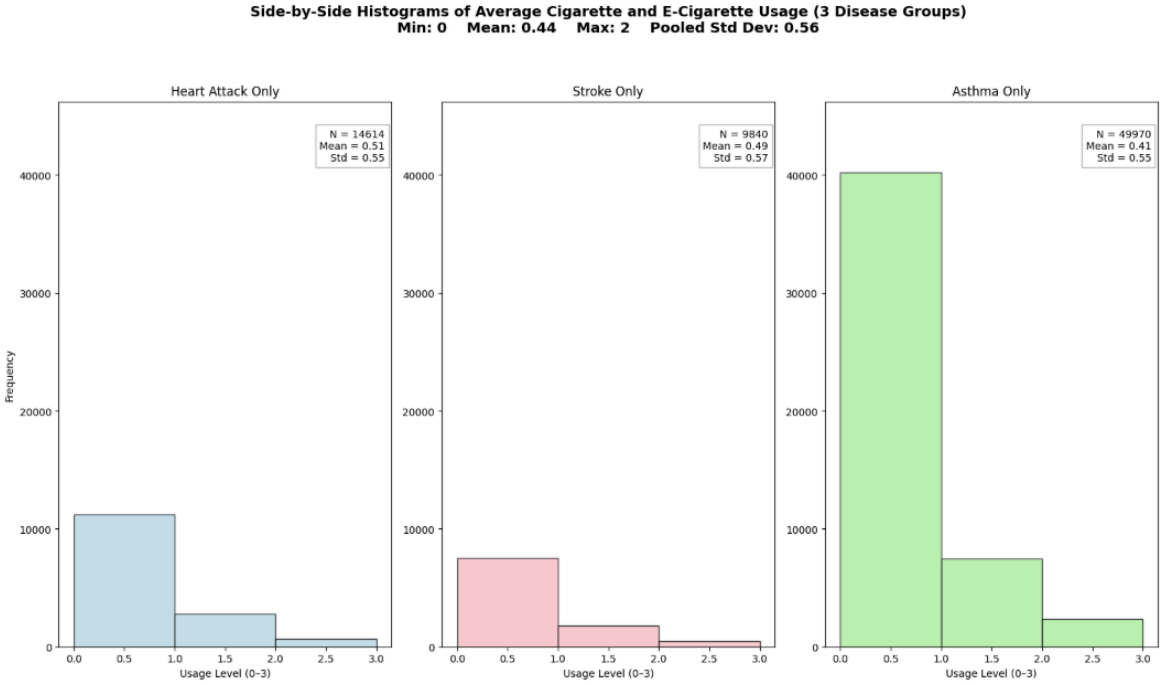

### 6.2 E-Cigarette Use (ECigaretteUsage)

The second model that was considered was the e-cigarette use. Although the ANOVA again returned a statistically significant result of about 102, this significance is largely attributable to the large sample size rather than meaningful behavioral differences. As it can also be observed through the histograms, the actual mean of e-cigarette usage for all 3 disease groups remains close to 0 which indicates that vaping is relatively uncommon among the individuals that are in this dataset regardless of their health records. Although the p-value that is below 0.05 close to 0 rejects the null hypothesis, the practical implications remain limited, and the results show that e-cigarette behavior

only plays a small role in distinguishing nicotine use across the health categories we analyzed.

```
========================================
ANOVA: ECigaretteUsage by Disease
========================================
F-statistic: 102.7483
P-value:     0.000000

Group Means:
Disease 1 (Heart Attack Only): 0.1795
Disease 2 (Stroke Only):       0.1867
Disease 3 (Asthma Only):       0.2278
----------------------------------------
```

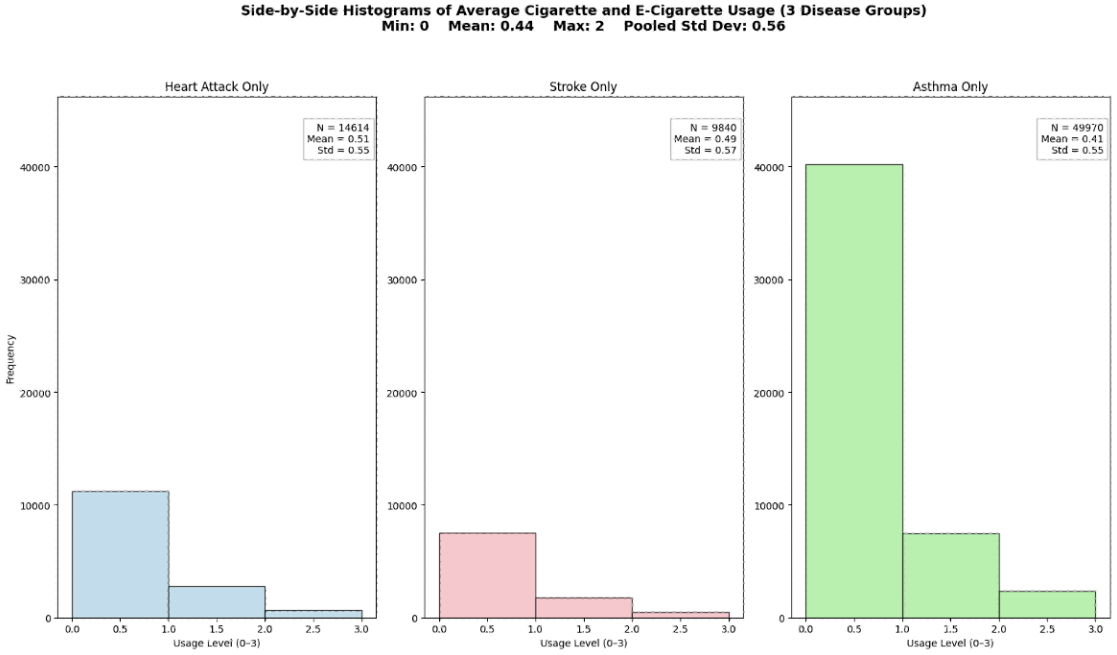



### 6.3 Combined Nicotine Exposure (Both)

The last model evaluated the combined nicotine exposure measure with both cigarette and e-cigarette use combined together into a single value. The ANOVA had a large f-statistic of about 230 with p value below 0.05 close to 0, displaying a strong statistical significance that is similar to the cigarette=use model as it could also be seen through the histograms. This outcome was somewhat expected as traditional smoking data had more variability and weight in the combined score compared to vaping. As a result, the group differences also mirror the findings of SmokerStatus, where heart-attack patients showed the highest combination of nicotine exposure while asthma patients reported the least. This also contributes to the finding that the primary driver of nicotine-related differences across disease groups is the conventional cigarette use compared to e-cigarette consumption.

```
========================================
ANOVA: Both by Disease
========================================
F-statistic: 230.4538
P-value:     0.000000

Group Means:
Disease 1 (Heart Attack Only): 0.5148
Disease 2 (Stroke Only):       0.4860
Disease 3 (Asthma Only):       0.4116
----------------------------------------
```




### 6.3 Interpretation Relative to Hypothesis

All three ANOVAs show results that support the rejection of the null hypothesis, indicating that nicotine use varies across disease categories. However, it was also found that the degree of practical significance differs depending on the type of nicotine behavior that was examined. Traditional smoking shows a more meaningful pattern with usage following the order of Heart Attack > Stroke > Asthma. Contrastingly, the difference in e-cigarette use was not as meaningful even though it was statistically significant because all 3 of the group means were too close to 0. The combined exposure measure follows the same pattern as the traditional cigarette use, and largely mirrors the differences that were observed in smoking behavior.

### 6.4 Limitations

Although all three ANOVA models show a statistical significance, there are several limitations that must be acknowledged. To begin with, the large sample size amplifies sensitivity, and causes the models to detect and interpret the smallest differences to be

statistically significant. Group imbalance with asthma-only groups having a larger group compared to others also further inflates f-statistics. Furthermore, the dataset is also based on self-reported values, which means that there could be a variation in honesty, and memory, and social desirability could influence the responses. Lastly, ANOVA's assumption of independence and roughly equal variance across groups may not be held perfectly for a dataset with such diversity in demographic and geographic populations. Despite these limitations, the results still consistently demonstrate that health history is associated with patterns of nicotine use, even though the strength of the relationships could vary considerably depending on the type of nicotine behavior that was examined.

7. **Model Deployment:**

Although our primary goal was to investigate whether nicotine use differs across disease categories rather than creating a predictive system, it is also critical to take into consideration how our results analysis could be applied in real-world contexts. All attempts to use a model or analytical framework based on this data must be approached carefully, with an understanding of both the potential uses and the limitations.

In the case of public health, the insights of our analysis could be valuable in identifying populations that could benefit most from targeted clinical intervention to treat tobacco use. The findings that individuals with a history of heart attack reports a higher level of cigarette use is in line with the established clinical knowledge about smoking being a major risk factor for cardiovascular disease, and therefore, public health agencies could utilize the results to strengthen outreach campaigns such as encouraging post-cardiac-event patients to enroll in clinical intervention programs or increasing awareness programs about the actual experienced heightened dangers of continued nicotine use for those with cardiovascular histories. Clinicians could also take use of these findings in patient conversations, utilizing the patterns we observed in order to contextualize the importance of reducing nicotine use after a major health event.

However, as we had acknowledged in the section above, the model comes with some limitations, so deploying it could come with challenges, especially for e-cigarette use even though the results show statistically significant differences between groups. Therefore, it must be considered carefully prior to usage. For instance, if the model is attempting to predict smoking behavior from disease history alone, it would perform poorly because disease history does not explain much in regards to the variability in nicotine behavior. Furthermore, relying on the self-reported data could also create uncertainty, and using such models for individualized decision-making could lead to inaccurate assumptions about the patients. Moreover, ethical considerations must also be made as creating a model that classifies individuals based on their disease history could raise ethical concerns of reinforcing harmful stereotypes or introducing bias in contexts of insurance, employment, or healthcare prioritization.

Therefore due to these concerns, our results are most suitable for usage in population-level insights rather than individual-level predictions. They highlight broad

behavioral patterns that could contribute to public health, but should not be used to make determination claims about any specific individuals. In order to consider future deployment, it would be critical to incorporate additional variables such as age, socioeconomic status, stress level, or mental health indicators in order to improve the predictive accuracy while maintaining ethical transparency. Overall, the most responsible use of our model is to support public health education, and guide further research rather than contribute to automated decision making in regards to one's nicotine related behaviors.