Analysis of factors associated with Heart Disease / Stroke

Gabriel Josh Balingit, Hengyuan Liu, Ka Wai Sit, Anne Celine Nygaard Weiseth

Abstract - **This paper explores some of the factors associated with heart disease and a fused response representing both heart disease and stroke. After regenerating 100 different samples and calculating several mutual information values, we see that the most notable factors are diabetes, old age, poor health, high cholesterol, and high blood pressure. Furthermore, cluster maps suggest that a combination of these factors, particularly those involving poor health, can significantly elevate risk, and in some cases, make it more likely to suffer at least one of these cardiovascular diseases. We also employed resampling techniques and applied separate random forest algorithms. When examining the results of these models, we observe accuracy levels around 94% to 96% along with patterns consistent with our findings.**

I. INTRODUCTION

As heart diseases unfortunately are very common worldwide, it is crucial to study the factors behind them. This report investigates the dataset "heart_diseas_health_indicators_BRFSS2015" to seek answers to what factors are associated with heart disease and which group has the highest risk. Furthermore, the analysis includes detailed examinations of factors associated with heart disease in combination with stroke and which groups have the highest risk of suffering from this cardiovascular disease. After briefly describing our data, the paper goes into the method for determining which are the associated variables by using Shannon's entropy. The method is applied to one-way, two-way and three-way interactions for selected features. We used cluster maps to visually point out trends in the two and three-way interactions. Additionally, as machine learning methods in health research have become popular, random forest was chosen as the classification method for predicting heart disease and for predicting heart disease and stroke.

II. DATA DESCRIPTION

The Heart Disease Health Indicators Dataset by Alex Teboul from Kaggle was used in our project. It was consolidated from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset[18]. The original dataset comprises responses from 441,455 individuals and features 330 variables. These features correspond to either questions directly asked of participants or are calculated variables based on their responses. We use the consolidated dataset, which contains 253,680 survey responses and 22 features with no missing data, mainly used for the binary classification of heart disease. The modified version of the dataset has provided documentation detailing the dataset's creation and the definitions of its features. Nonetheless, Teboul recognizes the potential for misconceptions and provides the BRFSS 2015 dataset codebook as a point of reference.

The dataset consists of 14 binary features, which include HeartDiseaseorAttack, HighBP, HighChol, CholCheck, Smoker, Stroke, PhyActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, and Sex. These features are encoded as either 0 (Without Condition) or 1 (With Condition). The dataset also contains eight non-binary features: BMI, Diabetes, GenHlth, MentHlth, PhysHlth, Age, Education, and Income. These features are encoded using different schemes and varying numbers of groups. Specifically, BMI is encoded on a scale from 12 to 98, where each number represents a different level of Body Mass Index. Diabetes is encoded as 0 (No Diabetes), 1 (Pre-Diabetes), or 2 (With Diabetes). GenHlth is ranked in descending order of quality from 1 (Excellent General Health) to 5 (Poor General Health). MentHlth is ranked in descending order of quality from 0 (0 days in the last 30 days felt mental health was not good) to 30 (30 days in the last 30 days felt mental health was not good). PhysHlth is ranked in descending order similar to MentHlth. Age is ranked in ascending order from 1 (age group 18 years old - 24 years old) to 13 (age 80 or older) in 5-year increments. Education is ranked in ascending order of quality from 1 (never attended school or up to kindergarten) to 6 (being in college for 4 years or higher education level). Income is ranked in ascending order from 1 (income less than 10,000) to 8 (income 75,000 or higher).

In our dataset, we noticed that some variables such as BMI (12-98), PhysHlth(0-30), and MentHlth(0-30) have multiple groups with very low sample sizes. This could potentially create issues during the data analysis process. To address this, we decided to group each variable individually into smaller groups.

For BMI, we followed the official guidelines from the CDC and WHO to group the BMI index into 4 categories: 0 represents a BMI index range up to 19, 1 represents a BMI index range higher than 19 and lower than 25, 2 represents a BMI index range of at least 25 to less than 30, and 3 represents a BMI index range of at least 30 or higher. To enhance the clarity of the four groups, we assigned names to each of them: 0 (Underweight), 1 (Healthy), 2 (Overweight), and 3 (Obese). By combining the BMI index into these four groups, each sample had a sufficient sample size, which facilitated the data analysis process and resulted in more accurate results while preserving the same relationship for BMI. Our new variable for BMI was labeled as BMI_cat.

Upon analysis of the PhysHlth variable, we found that there is no standard time unit for medical data since the treatment and procedures for different diseases vary. Therefore, we decided to group the data by weekly intervals to achieve a smaller group size while preserving the same relationship of PhysHlth without losing too much information. We created a new variable called PhysHlth_cat, which we categorized into 5 groups: 0 represents no bad physical health days in the last month, 1 represents at most one week of bad physical health days in the last month, 2 represents 1 to 2 weeks of bad physical health days in the last month, 3 represents 2 to 3 weeks of bad physical health days in the last month, and 4 represents 3 weeks to 1 month of bad physical health days in the last month. By combining the PhysHlth days into these five groups, each sample had a sufficient sample size, which facilitated the data analysis process and resulted in more accurate results. We faced a similar challenge with MentHlth as with PhysHlth, so we applied the same grouping strategy for the same reason.

It is important to note that most features in the dataset suffer from data imbalance which can pose challenges as we begin our analysis. The only features that are balanced are HighBP, HighChol, Smoker, and Sex. The remaining features exhibit varying degrees of imbalance, with binary features being imbalanced at ratios of mostly 90-10 or 80-20. All non-binary features are also imbalanced, with some groups having significantly higher proportions than others within the same feature. We suspect that the imbalance issue may be due to a lack of sample size, and we have checked the frequency count and proportions of the groups to address this issue.

We found that all binary features have a sufficient sample size for both groups. However, nonbinary features such as BMI, MentHlth, PhysHlth, and Education exhibited some groups with very few samples or some groups with less than one percent proportion for that feature. Due to this imbalance issue, we will conduct further exploratory analysis to determine the significance of these features and their relationship with heart disease and stroke.

III. INDIVIDUAL FEATURES VS HEART DISEASE

In order for us to determine the variables associated with heart disease, we started by calculating the entropy of heart disease. By denoting heart disease as Y and the different possible states of heart disease as b_1,... b_m, we can express this entropy as

$$CE(Y) = -\sum_{y\in\{b_1,\ldots,b_m\}} P(Y=y)\cdot\log(P(Y=y)) \tag{1}$$

Next, we calculated the conditional entropy of heart disease given some covariate for every other feature in our data. By denoting a given covariate as X and the different possible values of X as a_1,...a_k, we can express this conditional entropy as

$$CE(Y|X) = \sum_{x\in\{a_1,\ldots,a_k\}} CE(Y|X=x)\cdot P(X=x) \tag{2}$$

where

$$CE(Y|X=x) = -\sum_{y\in\{b_1,\ldots,b_m\}} P(Y=y|X=x)\cdot\log(P(Y=y|X=x)) \tag{3}$$

Using these two values, we then calculated the mutual information shared between heart disease and a given covariate for every feature. Given the notation provided above, we can express this mutual information as

$$Info(X,Y) = CE(Y) - CE(Y|X) \tag{4}$$

In a general sense, entropy can be interpreted as “the level of uncertainty involved in predicting some response”. And so, if mutual information is small, there exists a considerable amount of uncertainty in predicting Y when given knowledge of X. Likewise, if mutual information is large, the uncertainty in predicting Y given X is quite low. While these situations can be explained by random chance, it is also possible that definite predictions may be facilitated by some relationship between X and Y. As a result, our team sorted features based on the amount of mutual information they share with heart disease.

However, to ensure that these values were reliable, we constructed contingency tables with Y and some given X and then regenerated count data using the multinomial distribution. When it comes to the number of samples, our team ultimately settled on 100- as using more lead to longer running times. In regards to implementation, we started by regenerating the counts in every category of a given X. If we let N be the total number of observations in our data, we can express this resampling process as

$$(A_1, \ldots, A_k) \sim Multinomial(N,\ P(X = a_1), \ldots, P(X = a_k)) \tag{5}$$

We then regenerated the counts in every category of Y given that X is a_i and repeated this across all possible values of X. If we let n_i be the number of observations in our data such that X is a_i, we can express this sampling process as

$$(B_1, \ldots, B_m) \sim Multinomial(n_i,\ P(Y = b_1 | X = a_i), \ldots, P(Y = b_m | X = a_i)) \tag{6}$$

With new count data for 100 samples, we calculated Info(X, Y) 100 times and summarized them all through an average. We then repeated this process for every other feature, leaving us with 21 mutual information values.

TABLE I
Top Ten mutual information with heart disease

Variable	Mutual Information
GenHlth	0.032154
Age	0.028316
HighBP	0.021955
DiffWalk	0.017856
HighChol	0.016097
PhysHlth_cat	0.013326
Diabetes	0.013101
Stroke	0.012715
Income	0.010053
Smoker	0.006340

Among these values, we see that the top ten variables which share the most information with heart disease include GenHlth, Age, HighBP, DiffWalk, HighChol, PhysHlth_cat, Diabetes, Stroke, Income, and Smoker. Based on the dataset, we observed a clear trend indicating that as general health declines, the likelihood of heart disease increases. Furthermore, advancing age, high blood pressure, difficulty walking, deteriorating physical health, history of stroke, lower income levels, and smoking habits are all associated with an increased risk of heart disease.

IV. TWO-WAY INTERACTIONS VS HEART DISEASE

Although the risk of heart disease changes across different groups, it is important to report that it consistently remains below 50%. In other words, varying knowledge of one's general health, age, or sex alone will always show that heart disease is less likely to occur. However, this begins to change when analyzing two-way interactions between different variables. Similar to what was done with individual features, we generated 100 samples for reliability, calculated various mutual information values, and examined the interactions which shared the most information with heart disease. If we let (X_1, X_2) be the interaction between two different variables and S_2 be the set of all values that (X_1, X_2) can take, we can express mutual information as

$$Info((X_1,\ X_2), Y) = CE(Y) - CE(Y|X_1, X_2) \tag{7}$$

where

$$CE(Y|X_1, X_2) = \sum_{x_1, x_2 \in S_2} CE(Y|X_1 = x_1, X_2 = x_2) \cdot P(X_1 = x_1, X_2 = x_2) \tag{8}$$

and

$$CE(Y|X_1 = x_1, X_2 = x_2) = -\sum_{y \in \{b_1, \ldots, b_m\}} P(Y = y|X_1 = x_1, X_2 = x_2) \cdot \log(P(Y = y|X_1 = x_1, X_2 = x_2)) \tag{9}$$

After finding all of the two-way interactions which shared the largest amount of information with heart disease, we calculated the conditional probabilities of heart disease given each interaction. With these probabilities, we then made various cluster maps using a mean hierarchical clustering algorithm. We also regenerated 100 different samples using the multinomial distribution to better support any findings associated with a particular cluster. Ultimately, the most notable two-way interactions were general health with age and general health with stroke.

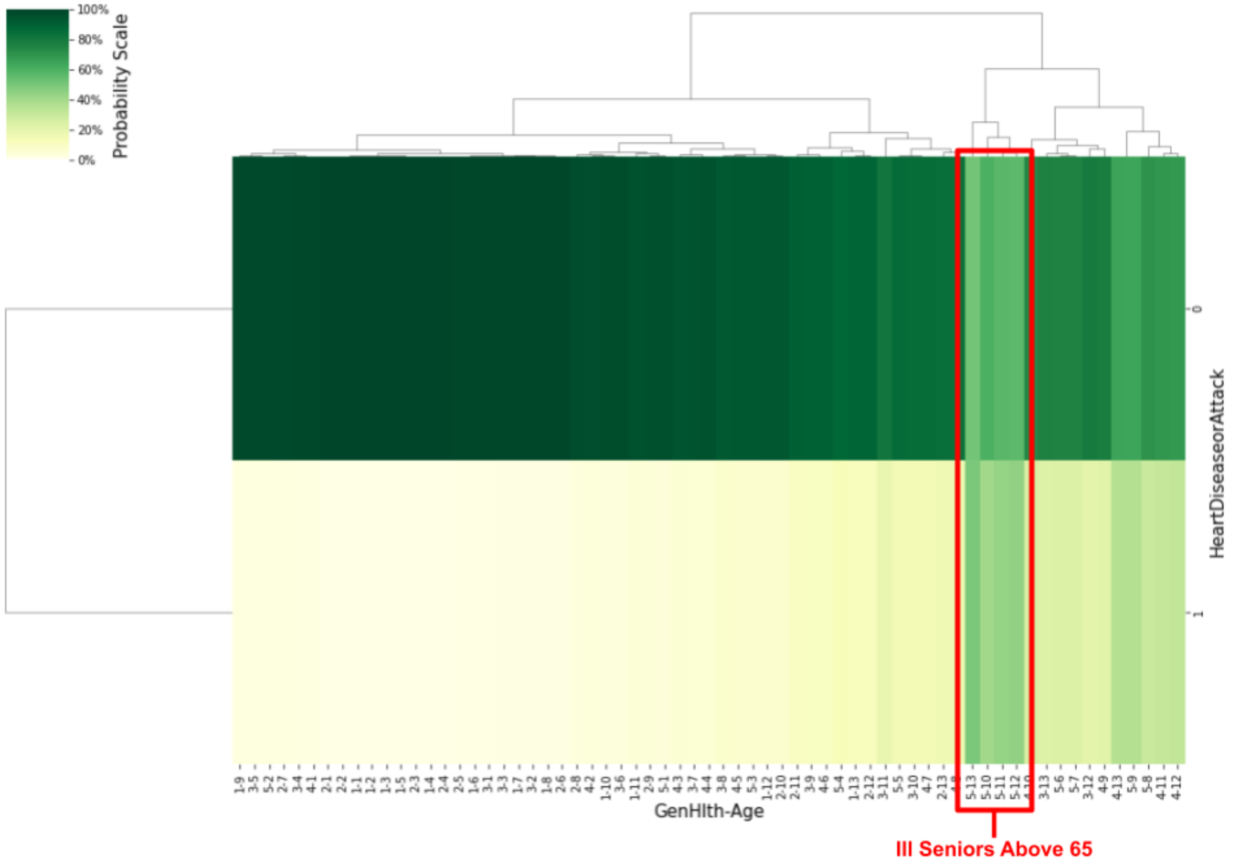


Figure 1: Clustermap of P(HeartDiseaseorAttack|GenHlth-Age)

When examining the two-way interaction between general health and age, we see several interesting patterns. For example, it is evident from the cluster map above that seniors above 65 who exhibit the worst general health can be identified as a distinct group. Upon further inspection, we see across 100 other samples that they are the group which possesses the highest risk of heart disease, at around 44%. This is especially notable when looking back at our data and seeing that the risk of heart disease for this cluster is higher than those with the worst general health (34%) and seniors older than 65 years old (18%). This ultimately highlights how having information on one's general health and age can expose significantly higher risks than those seen with information on only either. Moreover, these numbers suggest that people afflicted with serious illnesses should improve their wellness before growing old and that seniors should avoid worsening health- as doing otherwise could drastically increase the likelihood of heart disease.

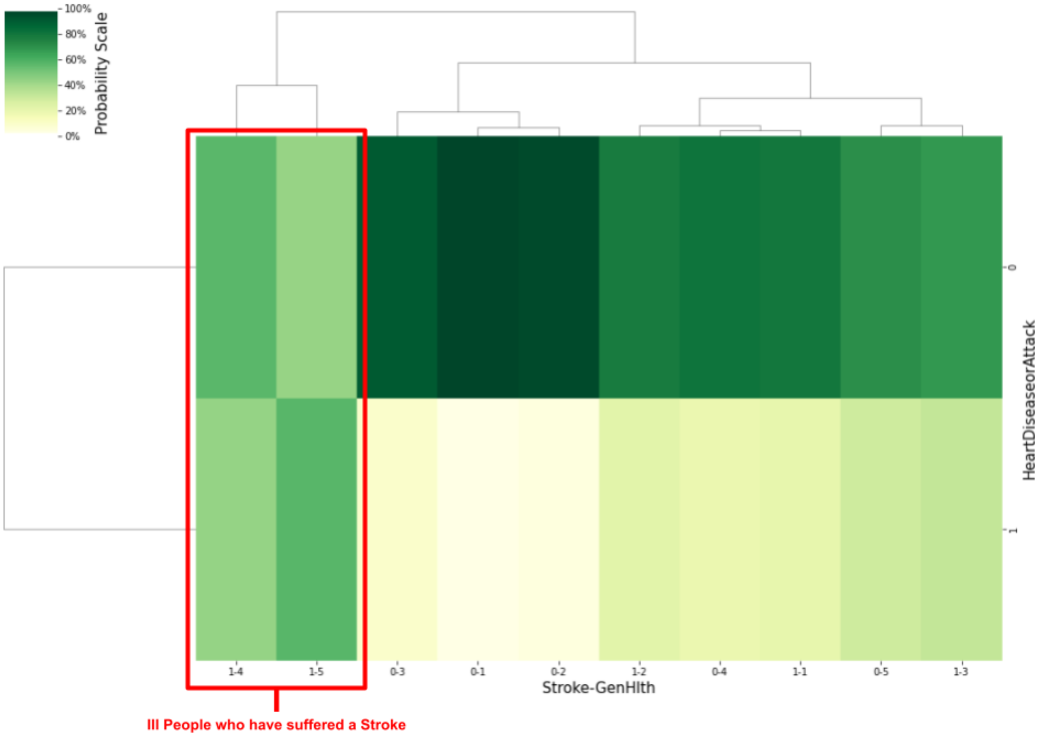
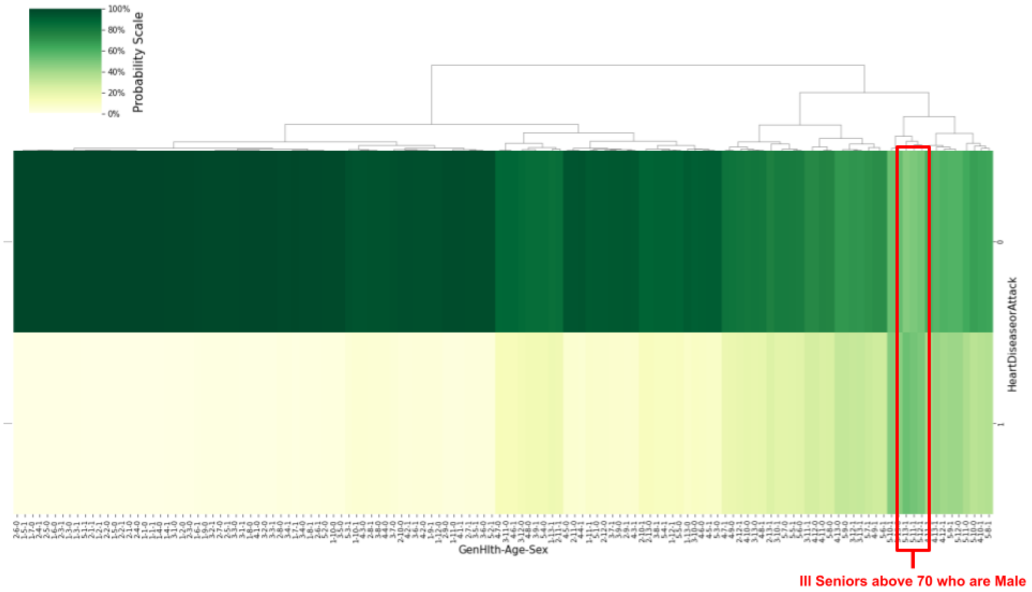


Figure 2: Clustermap of P(HeartDiseaseorAttack|Stroke-GenHlth)

We also noticed interesting patterns with the two-way interaction between stroke and general health. Based on the cluster map above and data from 100 other samples, we see that individuals who have experienced a stroke and exhibit poor general health possess the highest risk of heart disease, at around 50%. This observation is quite significant as it illustrates how a random coin toss is the ultimate difference between well-functioning respiration and debilitating chest pain. Another interesting finding is the subcluster held within this group, specifically those who have suffered a stroke and display the worst general health. Across all 100 of our regenerated samples, we see that this subcluster is the only one which is more likely to have heart disease, at around 57%. This is especially remarkable considering how much lower the risk is for those who have suffered a stroke (38%) and those with the worst general health (34%). Like with the two-way interaction between general health and age, we see that significant differences can occur when knowing about one's general health and stroke history. However, the differences here are not only significant but completely life-altering. Namely, those who have suffered a stroke should avoid worsening their general health, and those with the worst general health should avoid activities that increase the risk of stroke- as doing otherwise would not only increase the risk of heart disease but can possibly make it more likely to happen.

V. THREE-WAY INTERACTIONS VS HEART DISEASE

After doing this, we proceeded to analyze three-way interactions. However, due to computational limitations, we were unable to compile a list of mutual information values for every three-way interaction. And so, we ultimately decided to focus on all of the three-way interactions involving general health and age. Among this list, the most notable were the ones that additionally had sex, high blood pressure, high cholesterol, and stroke.



Ill Seniors above 70 who are Male

Figure 3: Clustermap of P(HeartDiseaseorAttack|GenHlth-Age-Sex)

By examining the cluster map for general health, age, and sex, it is evident that male seniors older than 70 with the worst general health are a distinct group. When further investigating this across 91 out of 100 samples, we see that this cluster possesses the highest risk of heart disease, at around 51%. And so, like with those who have suffered a stroke with poor health, a random coin toss is the ultimate difference between healthy breathing and severe chest discomfort. This is especially notable when we recall that seniors above 70 with the worst health have a 45% risk of heart disease. We also see from our data that men who exhibit the worst health have a 40% risk and that male seniors older than 70 have a 27% risk. Based on these numbers, it is apparent that having knowledge about one's general health, age, and sex can expose significantly different risks than those seen with knowledge about only two of the three variables. Moreover, these numbers suggest that unhealthy males should strive to improve their well-being as they age and that elderly men should avoid activities that deteriorate their overall health. By doing otherwise and allowing themselves to have the worst health past 70 years, the risks will not only change but can increase near the point where heart disease is more likely to happen.

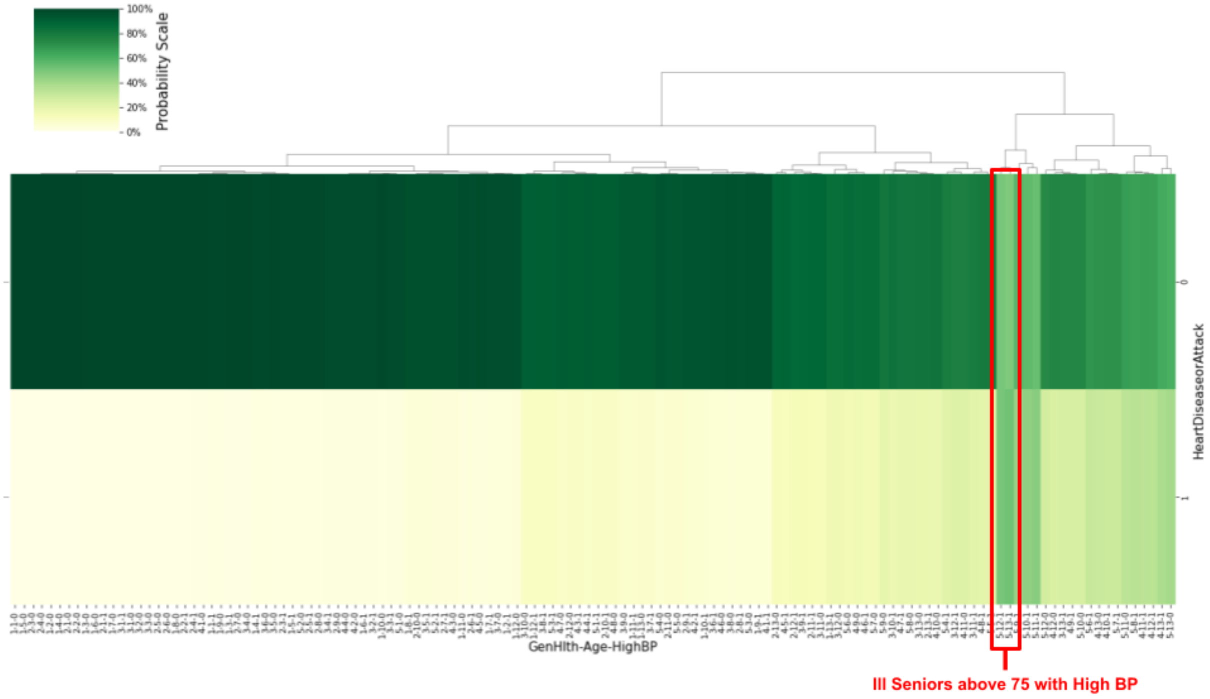


Figure 4: Clustermap of P(HeartDiseaseorAttack|GenHlth-Age-HighBP)

The three-way interaction between general health, age, and high blood pressure also provides a fair amount of interesting insight. When examining the cluster map above along with data from 99 regenerated samples, it is evident that ailing seniors above 75 who have high blood pressure represent a distinct group with the highest risk of heart disease, at around 51%. Again, we see that difference between experiencing smooth respiration and intense chest pain is a simple flip of a coin. This finding is quite alarming as our data shows that the risk of heart of disease is only 46% for ill seniors older than 75, 40% for unhealthy individuals who possess high blood pressure, and 25% for seniors above 75 with high blood pressure. Like with the previous three-way interaction, we see that being aware of one's general health, age, and blood pressure history can uncover significantly higher risks than those seen with information on only two. Furthermore, these statistics suggest that unhealthy individuals with high blood pressure should better their wellness and/or lower their blood pressure levels before growing too old and that elderly people with either issue should do the same. Failing to do so can ultimately result in an elevated risk of heart disease becoming more likely to occur.

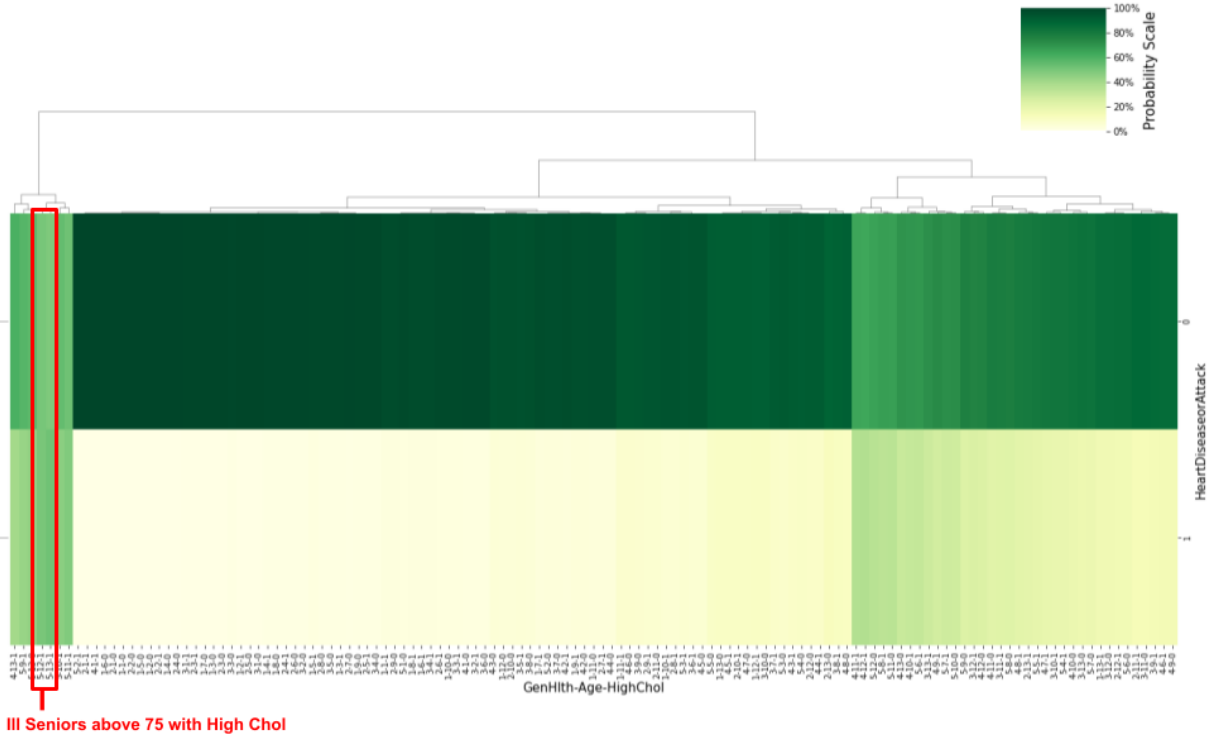


Figure 5: Clustermap of P(HeartDiseaseorAttack|GenHlth-Age-HighChol)

Similarly notable results can be observed with a deeper analysis of the three-way interaction between general health, age, and high cholesterol. For instance, the cluster map above along with supporting data from 95 other samples show that seniors above the age of 75 with poor health and high cholesterol exhibit the same heightened risk of heart disease (51%) as ill seniors above 75 with high cholesterol. Like before, we also see significant disparities between this finding and other risk values. As mentioned earlier, the risk of heart disease for ill seniors older than 75 is only 46%. Our data also shows that unhealthy individuals who possess high cholesterol levels have only a 40% risk and that seniors above 75 with high cholesterol only have a 27% risk. Again, we see that having information on three features, such as general health, age, and cholesterol history, can unveil risks significantly higher than those seen with information on only two. These numbers also indicate that people with severe health ailments and high cholesterol should work towards improving their well-being and/or decreasing their cholesterol levels as they age and that seniors with either issue should do the same. Based on our analysis, neglecting these actions will not only increase the risk of heart disease but can make it more likely to develop.

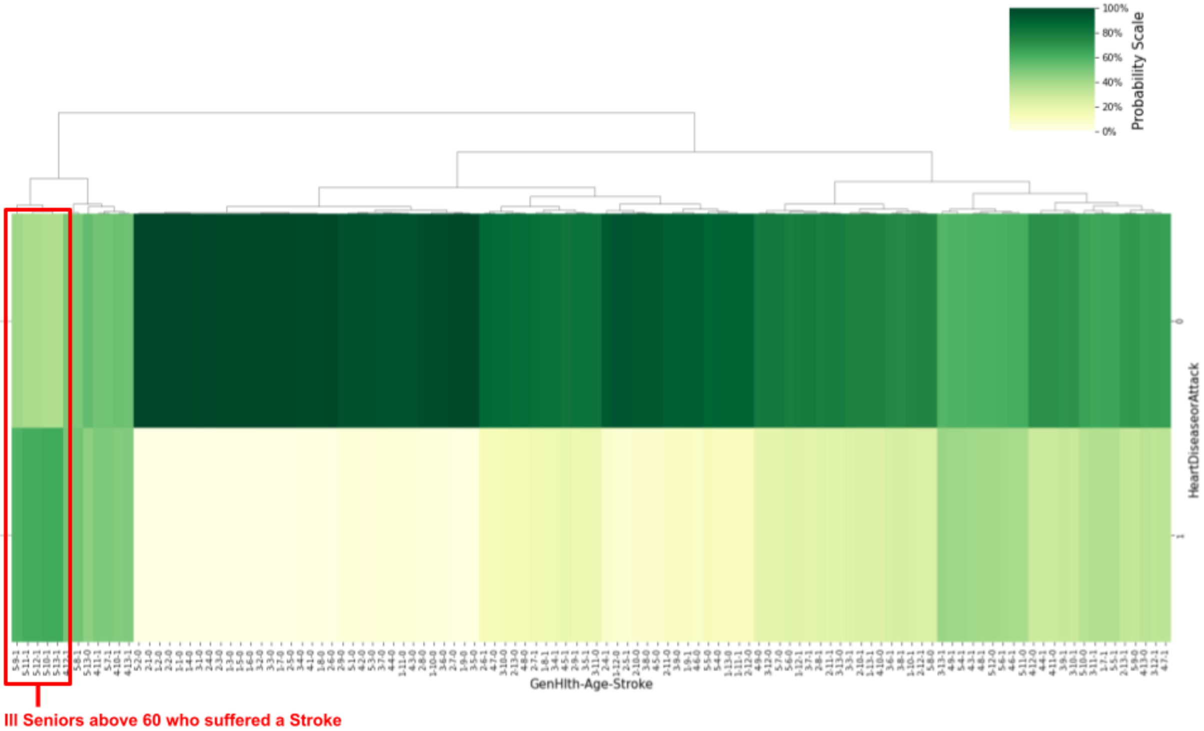


Figure 6: Clustermap of P(HeartDiseaseorAttack|GenHlth-Age-Stroke)

We also see notable patterns when examining the three-way interaction between general health, age, and stroke. Taking into account the cluster map above and the results from 95 other samples, we constantly see that unhealthy seniors above 60 with a history of stroke exhibit the highest risk of heart disease, at around 62%. This observation is quite significant as it highlights one of the rare instances, in our data, where heart disease is more likely to occur. It is even more remarkable when considering other risk values, particularly those given with knowledge of only two out of the three variables. For instance, our data shows that the risk of heart disease is only 57% for those who have suffered a stroke with the worst general health, 42% for seniors above 60 with the worst general health, and 40% for seniors above 60 who have suffered a stroke. As with other three-way interactions, it is evident that having information about an individual's general health, age, and stroke history can reveal even higher risks than those seen with information on only two out of the three. These numbers ultimately suggest that heart disease can remain less likely to occur if unhealthy individuals with a history of stroke improve their wellness as they age. Similarly, seniors who have suffered a stroke or exhibit poor health can better prevent heart disease by avoiding activities that increase their risk of stroke and/or worsen their well-being.

VI. RANDOM FOREST FOR HEART DISEASE

After examining interactions, random forest was chosen as the machine learning approach to predict HD and no HD. Random forest, a supervised method that uses decision trees, turned out to be very accurate at predicting our response variable. Oversampling on HD was necessary to eliminate bias towards the larger group, which, in this case, was no HD. The significant imbalance could have resulted in a much lower recall. Metrics such as precision, recall, and accuracy [4] are defined below:

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \tag{10}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \tag{11}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive\ +\ True\ Negative + False\ Positive + False\ Negative} \tag{12}$$

TABLE II
CLASSIFICATION REPORT

	Precision	Recall
0 - No HD	0.99	0.93
1 - HD	0.93	0.99

With an impressive high precision at 99 percent, the model predicts a person does not have HD when the truth is that the person does not have HD 99 percent of the time. One reason for such a high score can be the fact that most observations fall in under not having HD. For predicting HD, the precision was a little lower, and the model predicted HD correctly 93 percent of the time. Furthermore, to see the proportion of true positives that was identified correctly, the recall for the model was 93% for no HD and 99% for HD. The precision and recall can often collide as improving one means reducing the other. After doing k-fold cross-validation with k= 5, the optimal model contains 30 decision trees. The total portion of predictions the model correctly predicted, accuracy was 96%. Given a substantial number of nodes, it was difficult to visualize it in an organized way. However, when examining some of the upper nodes, we see patterns consistent with our previous findings.

VII. DISCUSSION FOR HEART DISEASE

Under the guidelines from the CDC, WHO, and other reputable institution, our analysis revealed the following associations: declining general health is associated with suppressed immune system function[7], aging is linked to compromised physiological resilience[6], high blood pressure hinders blood flow[8], difficulty of walking decreases physical activity[5], high cholesterol disrupts lipid metabolism[9], worsening physical health impairs oxygen delivery to tissues[10], diabetes elevates blood sugar levels[11], stroke causes damage to cerebral blood vessels[12], low income is associated with damage to cerebral blood vessels[13], and smoking reduces oxygen supply[14].

Our findings are consistent with the evidence suggesting that all of these factors are linked to an increased risk of heart disease. Additionally, we have identified several other variables that exhibit unique relationships with heart disease, which will be discussed separately in the following sections.

In our analysis, we hypothesize that the group with the lowest income (group 1) is likely to consist of individuals in occupations such as travel artists or start-up entrepreneurs, earning up to $10,000 annually. Despite this income level, they seem to have access to better resources. This could be attributed to potential economic assistance from their jobs, for things like food and housing. On the other hand, we hypothesize that Group 2 contains individuals earning between $10,000 and $15,000 per year (or $7.25 hourly, the federal minimum wage). These individuals often work in minimum-wage jobs, such as fast food, which offer limited career advancement opportunities and leave little extra funds for retirement or investments, resulting in Group 2 members having fewer resources and more financial constraints compared to Group 1. Group 2 may be more inclined to allocate their limited income towards instant gratification, such as unhealthy foods or smoking, as a means of stress reduction, further impacting their financial situation.

For the highest income group, they generally have better access to nutritious food and possess a greater awareness of health concepts. These factors may explain why they exhibit a lower likelihood of heart disease compared to the other income groups.

In our analysis, we have observed a pattern indicating that individuals in the lowest education level group have a higher likelihood of avoiding heart disease compared to groups 2, 3, and 4. Our speculation is that individuals in the lowest education group may be employed in physically active occupations, which could account for their lower incidence of heart disease and lower educational attainment.

Conversely, individuals in education groups 2, 3, and 4 are likely to have less physically demanding jobs but greater opportunities for career advancement. However, lower education levels generally correspond to lower income levels. While their income may still be modest, this limited financial capacity hampers their ability to make substantial investments. Notably, engaging in physical activity in the workplace is generally associated with positive effects on overall health. Similar to the low income groups, low education groups can lean towards unhealthy food choices or develop detrimental habits, which in turn increase the risk of heart disease. Higher-income groups typically occupy occupations with less physical activity, thereby contributing to an elevated risk of heart disease[2].

For the educated groups 5 and 6, characterized by higher income and a greater awareness of health concepts, their lower likelihood of heart disease can be attributed to these factors.

The code book for the original dataset establishes heavy alcohol consumption as consuming 7 or more drinks per week for females and 14 or more drinks per week for males. Interestingly, our dataset revealed a counterintuitive finding: individuals classified as heavy alcohol consumers had lower likelihood of heart disease compared to those who did not engage in heavy alcohol consumption. It is important to note that the definition of heavy alcohol consumption lacks a universally agreed-upon standard, as some define it as consuming less than 3 drinks a day for females and no more than 4 drinks a day for males. In this context, it is worth considering the potential benefits of moderate red wine consumption[1], given its popularity as the most prevalent wine in America. Consequently, it is plausible that individuals categorized as heavy alcohol consumers and those who consume alcohol within recommended limits, may possess a reduced risk of heart disease.

The likelihood of older individuals having healthcare is higher compared to younger people. This can be attributed to the increased prevalence of health issues requiring medical attention as people age, along with more frequent check-ups and screenings to monitor their well-being. Our findings reveal that individuals with access to healthcare exhibit a higher incidence of heart disease, which can be attributed to their higher age. Age serves as a significant risk factor for heart disease, with the probability of developing the condition escalating as individuals grow older. Our findings are consistent with research from “Retooling for an Aging America: Building the Health Care Workforce” by the National Library of Medicine[3].

VIII. INDIVIDUAL FEATURES VS HEART DISEASE AND STROKE

As we transition our focus from heart disease to a fused response involving both heart disease and stroke, it is important to note that we applied similar methods and formulas when finding associated features. In other words, we generated 100 different samples using the multinomial distribution and calculated numerous mutual information values in order to obtain reliable statistics.

TABLE III
Top Ten mutual information with heart disease and stroke

Variable	Mutual Information
GenHlth	0.040668
Age	0.034098
HighBP	0.026936
DiffWalk	0.025083
PhysHlth_cat	0.018397
HighChol	0.018318
Diabetes	0.015460
Income	0.014932
Smoker	0.007229
Education	0.006288

After doing so, we found that the top ten variables that share the most information with heart disease and stroke were GenHlth, Age, HighBP, DiffWalk, PhysHlth_cat, HighChol, Diabetes, Income, Smoker, and Education. According to our dataset, we found a clear trend indicating that as general health declines, the probability of getting heart disease and stroke increases. In addition, some factors like age, high blood pressure, difficulty walking, deteriorating physical health, high cholesterol, history of diabetes, lower income levels, smoking habits, and lower education levels are all associated with an increased risk of heart disease.

Due to changes in any one or more factors, the likelihood of neither heart disease nor stroke decreases. This is as general health declines, age increases, high blood pressure and cholesterol are present, physical health declines, walking becomes more difficult, diabetes is present, income is lower, smoking status is positive and education level is lower. On the other hand, the probability of having a stroke alone, heart disease alone, or both increases with the same factors. These include declining general health, aging, having high blood pressure and cholesterol, physical health decline, difficulty walking, the presence of diabetes, lower income level, smoking status, and lower education level. These results highlight the importance of early intervention and management of modifiable risk factors to prevent the development of heart disease and stroke.

IX. TWO-WAY INTERACTION VS HEART DISEASE AND STROKE

As with our analysis on heart disease, significant changes began emerging when we examined two-way interactions. Like before, we generated 100 samples and calculated the mutual information values for every possible interaction. The most remarkable in this comprehensive list ultimately include general health with age and general health with diabetes.

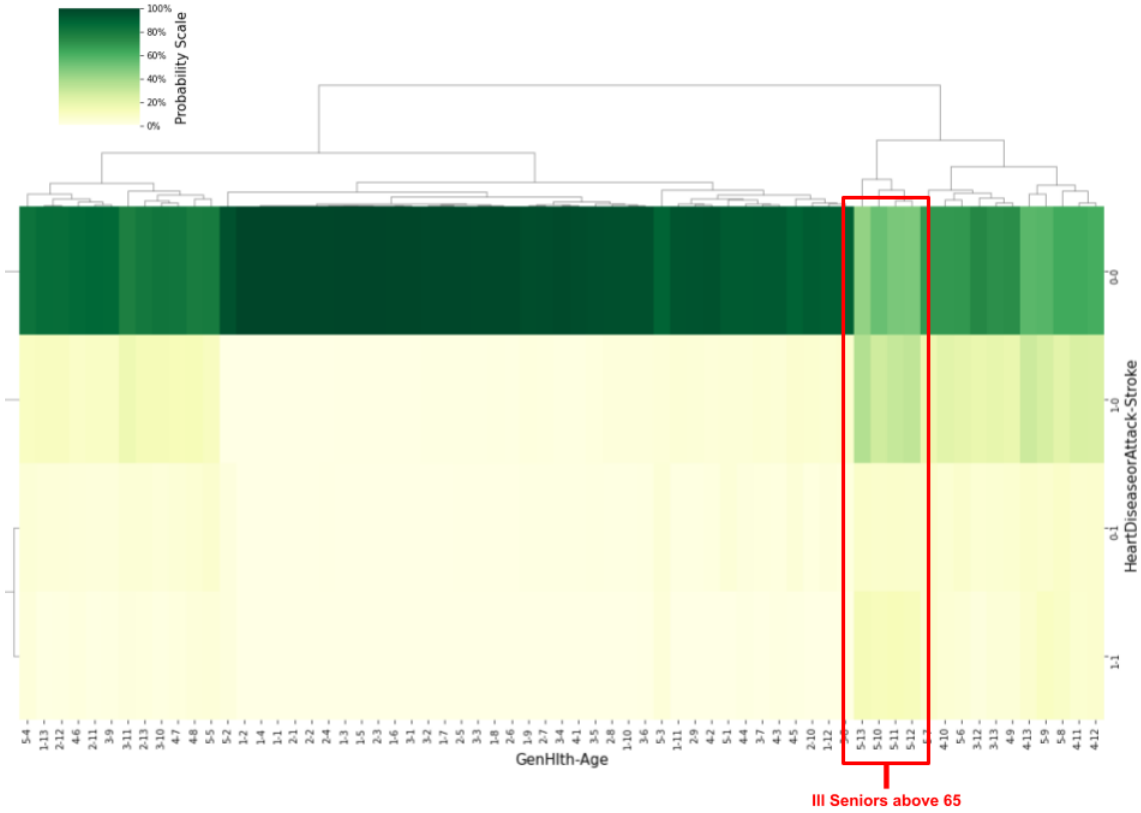


Figure 7: Clustermap of P(HeartDiseaseorAttack-Stroke|GenHlth-Age)

A subset of our cluster map for general health and age shows that individuals above 65 with severe health ailments are a distinguishable group. A closer analysis across 100 other samples reveals that this category of people has the greatest risk of suffering from both cardiovascular diseases, at around 12%, and heart disease alone, at around 32%. We also see interesting patterns for a subset of people within this cluster, specifically ill seniors above the age of 80. Results from 99 other samples ultimately show that this subcluster is the only one which is more likely to suffer at least one cardiovascular disease, at around 57%. This finding is made even more significant when our data shows that the risk of suffering at least one cardiovascular disease is only 30% for seniors older than 80 and only 41% for those who exhibit the worst overall health. Like before, we see that having information on one's general health and age can expose remarkably higher risks than those observed with knowledge on only either. However, the risks here are not only higher but also hold the potential to profoundly change one's life. Ultimately, our findings suggest that individuals with poor health should engage in healthy activities before growing old and that elderly people should avoid worsening their well-being- as doing otherwise can lead to the development of at least one cardiovascular disease.

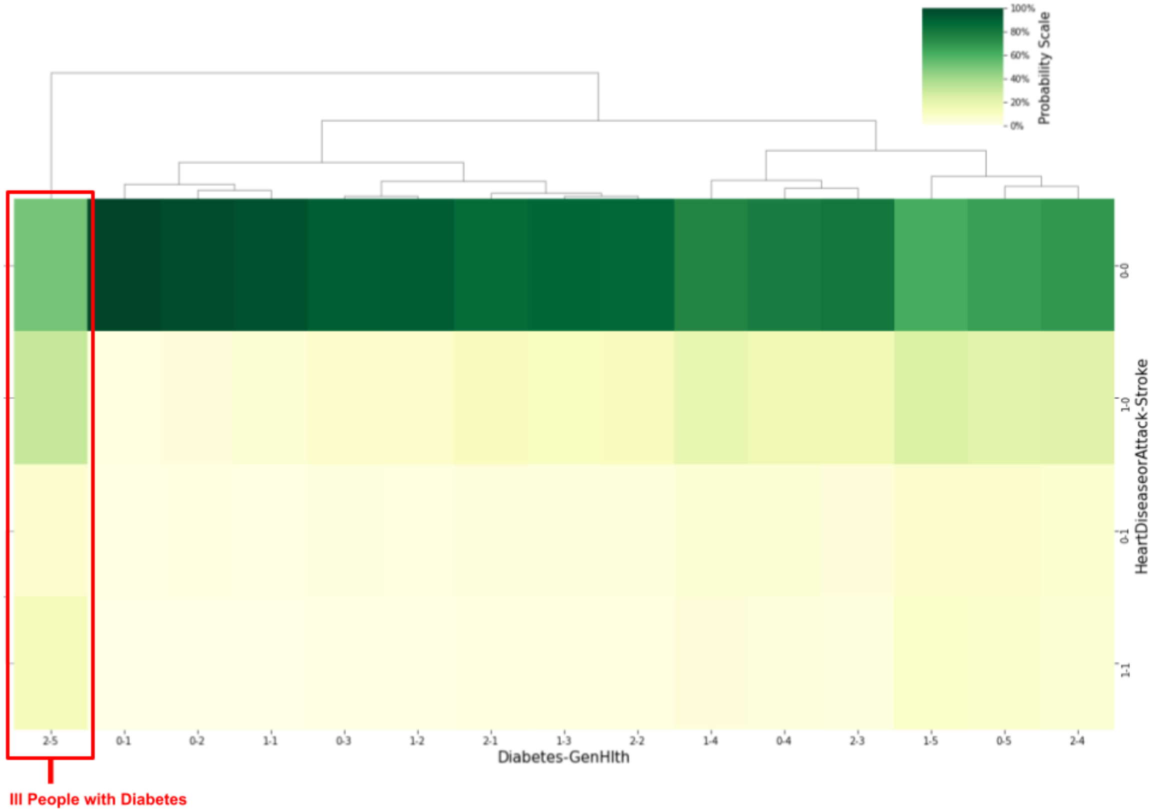
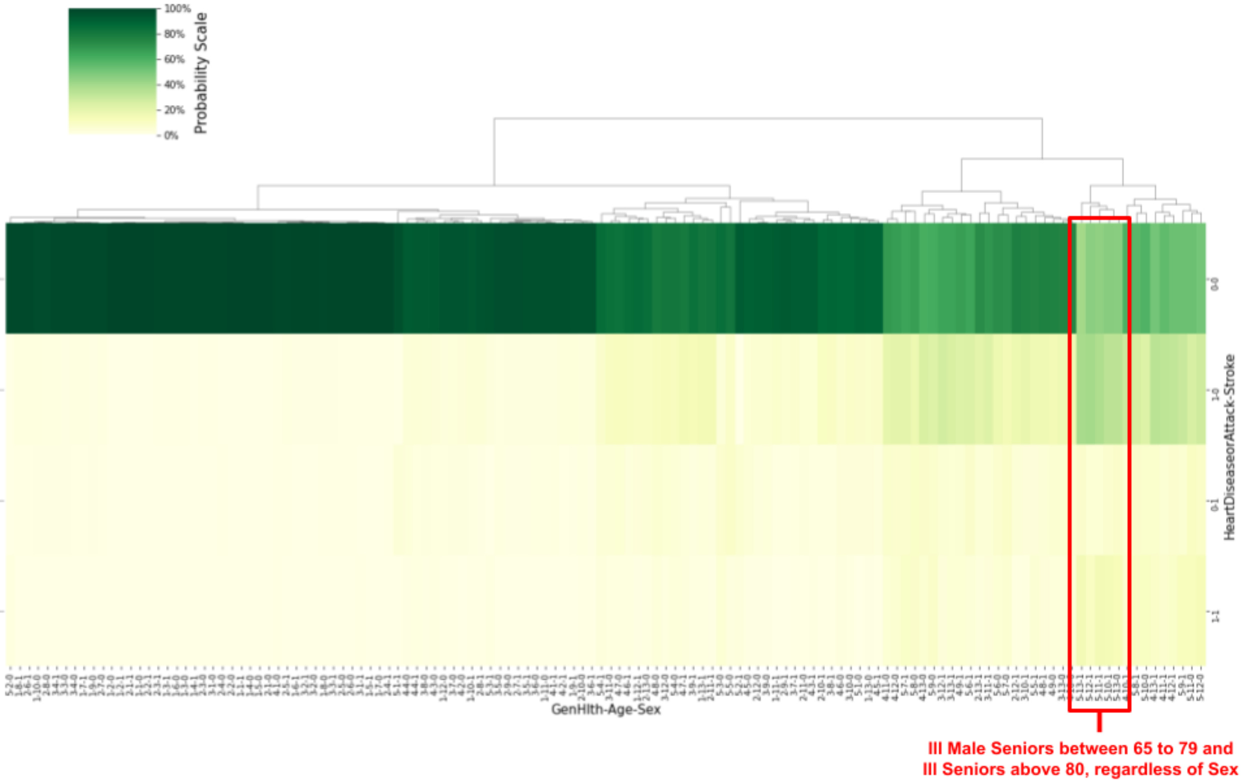


Figure 8: Clustermap of P(HeartDiseaseorAttack-Stroke|Diabetes-GenHlth)

There are also findings worth noting for the two-way interaction between diabetes and general health. Firstly, when examining the cluster map shown above, it is evident that diabetics who exhibit the worst general health are different from every other group. Across all 100 of our regenerated samples, we consistently see that this cluster possesses the highest risk of suffering from both cardiovascular diseases, at around 13%, and heart disease alone, at around 30%. Upon further inspection, it is also the cluster with the lowest chance of having neither stroke or heart disease, at around 50%. In other words, unhealthy diabetics are just one coin toss away from suffering at least one disease. This finding is even more remarkable when considering how our data shows that the risk of suffering at least one disease is only 27% for diabetics and only 41% for people with the worst general health. Like with the two-way interaction between general health and age, it is apparent that possessing knowledge about a person's general health and diabetic history can reveal significantly heightened risks. Ultimately, these numbers imply that diabetics should focus on maintaining their wellness and that unhealthy individuals should avoid activities that lead to diabetes- as doing otherwise can nearly make suffering at least one cardiovascular disease more likely to happen.

X. THREE-WAY INTERACTION WITH HEART DISEASE AND STROKE

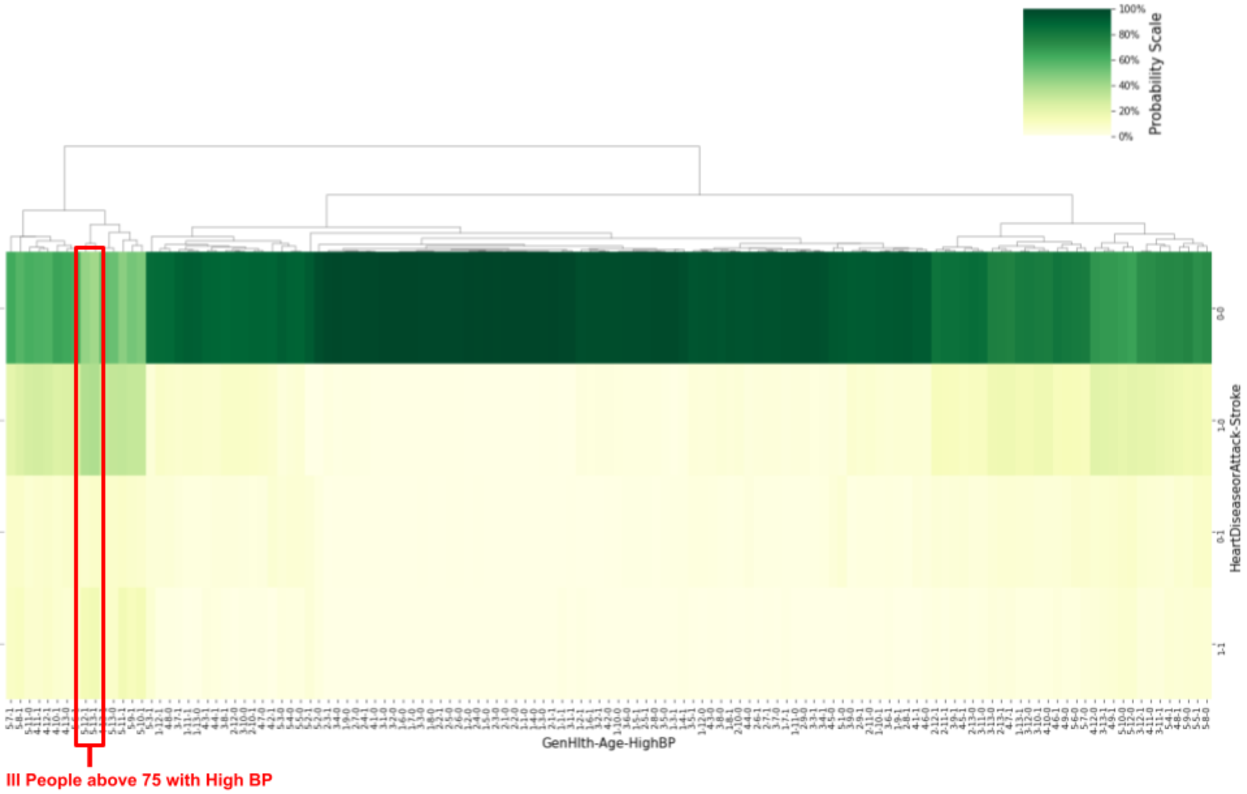
Like before, we furthered our analysis by closely examining all of the three-way interactions involving general health and age. By applying similar methods as previously established, we found that the most notable three-way interactions additionally involved sex, high blood pressure, and high cholesterol.



Ill Male Seniors between 65 to 79 and
Ill Seniors above 80, regardless of Sex

Figure 9: Clustermap of P(HeartDiseaseorAttack-Stroke|GenHlth-Age-Sex)

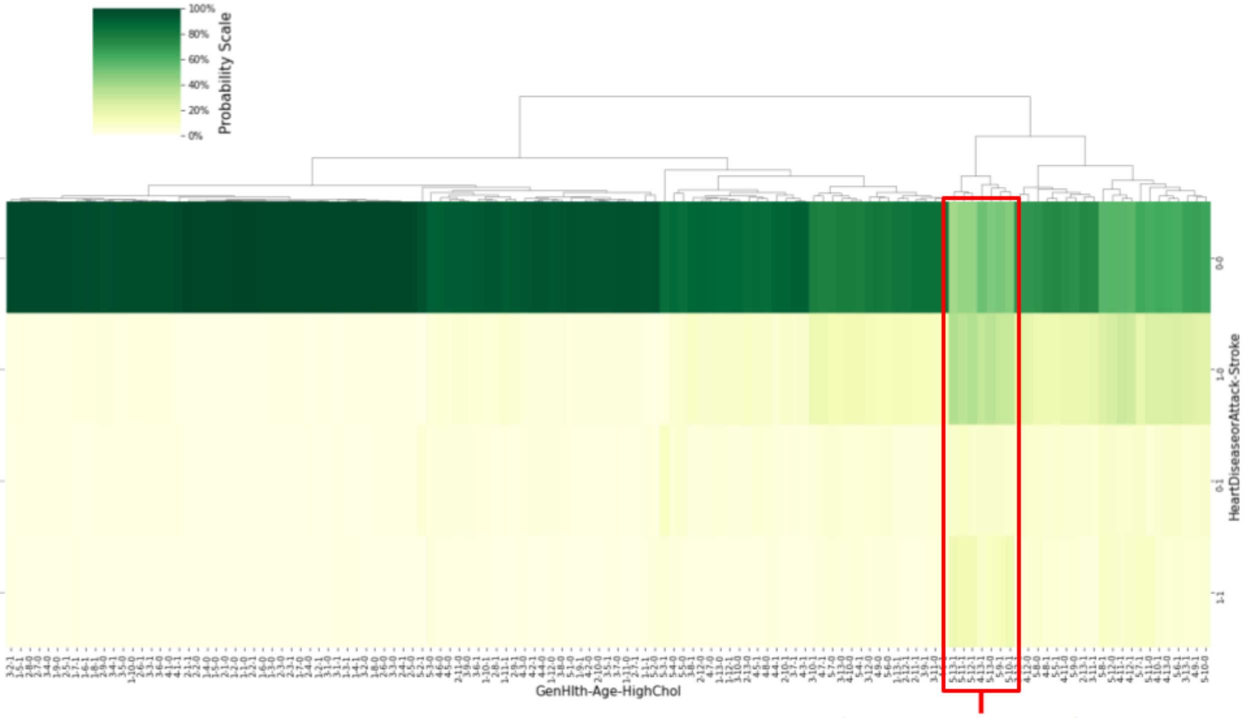
When examining the cluster map for general health, age, and sex, we see a discerning group comprising unhealthy male seniors between 65 to 79 and ill seniors above 80, regardless of sex. However, data from regenerated samples show that this initial grouping is reliable only under certain conditions. For instance, 98 samples show that ailing seniors older than 60, regardless of sex, have the greatest risk of suffering both heart disease and stroke, at around 12%. Meanwhile, 93 samples show that our initial cluster exhibits the highest risk of heart disease alone, at around 36%. 99 samples also show that this group is one of the only few more likely to suffer at least one of the cardiovascular diseases, with a risk of around 56%. This is especially significant considering how our data shows that the risk of suffering at least one of the cardiovascular diseases is 51% for ill seniors above 65, 41% for people with poor health, and 28% for male seniors between 65 to 79 and senior above 80, regardless of sex. Like what we have seen before, these numbers highlight how having information on one's general health, age, and sex can expose significantly elevated risks than those seen with information on only two. Moreover, they suggest that those who suffer severe ailments should improve their wellness as they age. They also suggest that male seniors between 65 to 79 and seniors above 80, regardless of sex, should avoid worsening their health. If these groups choose to do otherwise, their risk of suffering heart disease and/or stroke will not only change but increase to a point where it is more likely to occur.



Ill People above 75 with High BP

Figure 10: Clustermap of P(HeartDiseaseorAttack-Stroke|GenHlth-Age-HighBP)

There are also several notable patterns when it comes to the three-way interaction between general health, age, and high blood pressure. As shown by the cluster map above, we see that unhealthy seniors above 75 with high blood pressure are a distinct group holistically. However, as with the previous three-way interaction, a closer examination reveals different groupings for certain conditions. For instance, 98 different samples show that seniors above 65 who possess poor health and high blood pressure represent an exclusive cluster with the greatest risk of both heart disease and stroke, at around 14%. On the other hand, we do see across 94 samples that unhealthy seniors above 75 with high blood pressure exhibit the highest risk of heart disease alone, at around 36%. Meanwhile, 100 samples reveal that seniors older than 70 afflicted by severe health issues and high blood pressure are more likely to suffer at least one of the cardiovascular diseases, at around 57%. This is especially remarkable when reexamining our data and seeing that likelihood of suffering at least one of the cardiovascular diseases is 53% for ill seniors above 70 years old, 48% for unhealthy individuals who exhibit high blood pressure, and 30% for seniors above 70 with high blood pressure. Like before, we see that possessing knowledge about a person's general health, age, and blood pressure history can reveal risks higher than those observed with knowledge about only two. These numbers also suggest that those with poor health and high blood pressure should improve their well-being and/or lower their blood pressure as they age and that elderly people with either issue should do the same. By doing otherwise, the risk of suffering heart disease and/or stroke will not only increase but, in some cases, reach a point where it is more likely to happen.



Ill Seniors above 60 with High Chol and
Ill Seniors above 80 without High Chol

Figure 11: Clustermap of P(HeartDiseaseorAttack-Stroke|GenHlth-Age-HighChol)

In addition to this, there are several interesting patterns with the three-way interaction between general health, age, and high cholesterol. As indicated by the cluster map above, there is a distinguishable group of unhealthy seniors above 60 with high cholesterol and unhealthy seniors above 80 without high cholesterol. However, further analysis reveals varying groups, as seen with previous three-way interactions. For example, 99 different samples show that ill seniors between 60 and 79 with high cholesterol and ill seniors above 80, regardless of cholesterol, have the greatest risk of both heart disease and stroke, at around 14%. On the other hand, we see across 97 samples that unhealthy seniors between 65 and 79 with high cholesterol and ill seniors above 80, regardless of cholesterol, exhibit the highest risk of heart disease alone, at around 34%. 100 samples also reveal this group is one of the only few more likely to suffer at least one of the cardiovascular diseases, at around 55%. This is especially remarkable when reexamining our data and seeing that the likelihood of suffering at least one cardiovascular disease is 51% for seniors above 65 with poor health, 39% for unhealthy individuals regardless of high cholesterol, and 27% for seniors between 65 and 79 with high cholesterol and seniors above 80, regardless of cholesterol. Similar to our previous findings, understanding a person's overall health, age, and cholesterol history can uncover greater risks than those discovered with information on only two. These numbers also suggest that individuals with poor health[7] and high cholesterol[9] should take steps to improve their well-being and/or lower their cholesterol as they age and that elderly people with either affliction should do the same. Neglecting these actions will only increase the risk of suffering heart disease and/or stroke but, in some cases, make it more likely to happen.

XI. RANDOM FOREST FOR HEART DISEASE AND STROKE

After examining three-way interactions, we then did a random forest to predict the four outcomes no HD nor stroke (0), stroke (1), HD (2) and HD and stroke (3). Similarly to the random forest on HD alone, the data was resampled due to the significant imbalance. It was done by dividing the total number of observations, 250 k, by 4 so that they would have a uniform distribution. Here, we also used 80 percent as training and 20 percent as testing data. As this model classified four, and not only two, the classification report were slightly less than the first one.

TABLE IV
Random Forest result

	Precision	Recall
0 - No HD nor stroke	0.95	0.83
1 - Stroke	0.96	0.99
2 - HD	0.87	0.94
3 - HD and stroke	0.98	1.00

The highest precision of the four classes were for HD and stroke at 98 percent. This can mean that people with both HD and stroke have many other factors in common which makes it easier to classify. For predicting HD, the precision was a little lower with the model predicting correctly 87 percent of the time. Next, the highest recall was 100% for HD and stroke. It shows that the model successfully indemnified all instances of HD and stroke. Similarly to the first random forest model, the amount of nodes was also too large, nevertheless the trees support our previous findings.

XII. DISCUSSION FOR HEART DISEASE AND STROKE

Like before, our team conducted further research to provide possible explanations for these patterns. Based on the CDC, WHO, the American Heart Association, and the National Institute of Neurological Disorders and Stroke, some of the reasons these factors affect the risk of heart disease and stroke are that they can cause a weakened immune system[7], an aging body[6], damage to blood vessels[8], disturbance to parts of the brain[5], development of other heart disease when one had occurred, an increased risk of atherosclerosis[11], and an increased risk of triglycerides[14]. Since all ten variables are about health, it was not surprising that general health gave the most information. Age and blood pressure were also among the most informative. Income, smoking, and education levels were at the bottom of the top ten and the reason for this can be that these are more complex as they indicate information about social status, not only health status. However, it is worth noting that variables like education and income. These have a less intuitive effect on the risk of heart disease and/or stroke.

Education and income both affect each other as well as influence heart disease and stroke. Lower levels of education can cause a higher risk of heart disease and stroke due to low awareness of health which includes unhealthy diets, low income, and limited access to health care[2]. This can again be due to the health education American high schools provide[16]. In addition, limited access to healthy food options is more common among those with lower education levels partly because healthier food options tend to be more expensive, and individuals with lower education levels may have less income to afford them[13]. This brings in income as health care is heavily dependent on income. Lack of access to healthcare can increase the risk of heart disease and stroke as people may not receive timely and adequate medical attention, preventive care, and treatment for health conditions such as high blood pressure, high cholesterol, diabetes, and obesity[17].

XIII. CONCLUSION

In summary, mutual information values show that diabetes, old age, poor health, high cholesterol, and high blood pressure are some of the factors associated with a heightened risk of heart disease. Cluster maps also indicate that possessing a combination of these factors- particularly those involving poor health- can further increase risk and, in some cases, make heart disease more likely to occur. In addition to this, we applied resampling techniques and fitted a random forest algorithm. Upon evaluating the results of this model, we observe a final accuracy rate around 94% to 96% along with trends that support our previous findings. When employing these analytical techniques for a fused response involving heart disease and stroke, many of the same patterns persist. Although we are able to provide possible explanations for our findings and cite reputable sources such as the CDC and the WHO, it is important to emphasize that we do not claim any causal relationship. We are not health professionals, and our assessment should not be used as primary medical advice. Instead, our research aims to provide general insight surrounding factors associated with certain cardiovascular diseases- as understanding the biological mechanisms involved would require additional data and further analysis.

References

[1] "The truth about red wine and heart health," Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/red-wine/art-20048281#:~:text=Drink%20in%20moderatio%20%E2%80%94%20or%20not,lower%20risk%20of%20heart%20disease. (accessed May 12, 2023).

[2] M. Merschel, "More school, less heart disease? researchers keep finding evidence," www.heart.org, https://www.heart.org/en/news/2020/08/12/more-school-less-heart-disease-researchers-keep-finding-evidence#:~:text=Even%20in%20people%20without%20existing,study%20in%20JAMA%20Internal%20Medicine. (accessed May 12, 2023).

[3] Health Status and health care service utilization - retooling for an ..., https://www.ncbi.nlm.nih.gov/books/NBK215400/ (accessed May 12, 2023).

[4] Precision and Recall, https://en.wikipedia.org/wiki/Precision_and_recall (accessed May13, 2023)

[5] "Chronic disease fact sheet: Physical inactivity," Centers for Disease Control and Prevention, https://www.cdc.gov/chronicdisease/resources/publications/factsheets/physical-activity.htm (accessed May 12, 2023).

[6] "Heart health and aging," National Institute on Aging, https://www.nia.nih.gov/health/heart-health-and-aging#:~:text=Changes%20that%20happen%20with%20age,of%20arteries%20over%20many%20years. (accessed May 12, 2023).

[7] "Know your risk for heart disease," Centers for Disease Control and Prevention, https://www.cdc.gov/heartdisease/risk_factors.htm (accessed May 12, 2023).

[8] "High blood pressure symptoms and causes," Centers for Disease Control and Prevention, https://www.cdc.gov/bloodpressure/about.htm#:~:text=High%20blood%20pressure%20can%20damage%20your%20arteries%20by%20making%20them,Chest%20pain%2C%20also%20called%20angina. (accessed May 12, 2023).

[9] "Know your risk for high cholesterol," Centers for Disease Control and Prevention, https://www.cdc.gov/cholesterol/risk_factors.htm (accessed May 12, 2023).

[10] "Understand your risks to prevent a heart attack," www.heart.org, https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack (accessed May 12, 2023).

[11] "Diabetes and your heart," Centers for Disease Control and Prevention, https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html#:~:text=If%20you%20have%20diabetes%2C%20you,are%20to%20have%20heart%20disease. (accessed May 12, 2023).

[12] "Heart complications after a stroke increase the risk of future cardiovascular events," American Heart Association, https://newsroom.heart.org/news/heart-complications-after-a-stroke-increase-the-risk-of-future-cardiovascular-events#:~:text=%E2%80%9CWe%20know%20heart%20disease%20and,said%20Benjamin%20J.R.%20Buckley%2C%20Ph. (accessed May 12, 2023).

[13] C. Scott, "Study attributes 60-70% of excess heart disease among low-income Americans to poverty rather than traditional risk factors," Epidemiology & Biostatistics, https://epibiostat.ucsf.edu/news/study-attributes-60-70-excess-heart-disease-among-low-income-americans-poverty-rather (accessed May 12, 2023).

[14] C. for T. Products, "How smoking affects heart health," U.S. Food and Drug Administration, https://www.fda.gov/tobacco-products/health-effects-tobacco-use/how-smoking-affects-heart-health (accessed May 12, 2023).

[15] "Low income and work stress contribute to link between education, heart disease and stroke," European Society of Cardiology, https://www.escardio.org/The-ESC/Press-Office/Press-releases/Low-income-and-work-stress-contribute-to-link-between-education-heart-disease-and-stroke (accessed May 13, 2023).

[16] Centers for Disease Control and Prevention. "Health Education in Schools | CDC." *www.cdc.gov*, 21 Apr. 2021, www.cdc.gov/healthyyouth/health-education/index.htm.

[17] Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Edited by Brian D. Smedley et al. *PubMed*, Washington (DC), National Academies Press (US), 2003, www.ncbi.nlm.nih.gov/books/NBK220358/.

[18] A. Teboul, "Heart disease health indicators dataset," Kaggle, https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset (accessed May 13, 2023).