

Analyzing Cardiometabolic Risk Factors: Insights from Mutual Information

STA 160

Instructor: Fushing Hsieh

Written By: Lynn Waiyan Kyaw, Zhe Jiang, Carl Ge

1 Abstract

Cardiometabolic diseases, including heart disease, diabetes, and stroke, are leading causes of morbidity and mortality globally. This study leverages data from the Behavioral Risk Factor Surveillance System (BRFSS) to explore the relationships between various health indicators and the risk of developing these diseases. Utilizing statistical measures such as mutual information, entropy, and odds ratios, we analyzed the influence of predictors on cardiometabolic outcomes across different subpopulations. Our findings indicate that factors such as body mass index (BMI), general health (GenHlth), and high blood pressure (HighBP) consistently show strong associations with diabetes, heart disease, and stroke.

Our methodology involved defining conditional subpopulations based on the presence of one cardiometabolic disease of the two to evaluate the existence of the third disease. This approach allowed for a focused analysis on how specific predictors influence the risk of diabetes, heart disease, and stroke within these subpopulations. After the implementation, we found that general health (GenHlth) is the top factor across all conditions and diseases, indicating a potential public health intervention to mitigate the impact of cardiometabolic diseases.

2 Introduction

Cardiometabolic diseases such as heart disease remain a formidable health challenge globally, with the United States experiencing a significant impact each year. According to the Centers for Disease Control and Prevention (CDC), the U.S. consistently ranks among the countries most affected by heart diseases, reflecting a health concern that affects millions of individuals in the country. Annually, heart disease claims a staggering number of lives in the U.S., with approximately 695,000 deaths attributed to it in 2021, making it the leading cause of mortality nationwide for men, women, and people of most racial and ethnic groups [1]. In terms of other cardiometabolic diseases, recent projections from a comprehensive study published in the Journal of the American College of Cardiology shed light on the escalating trajectory of cardiovascular diseases (CVD) in the United States. The study, conducted using data from the 2020 U.S. Census Bureau and the National Health and Nutrition Examination Survey, offers an outlook on the future risks of CVDs across different demographic groups. By analyzing trends from 2025 to 2060, the study highlights a trend of rising cardiovascular risk factors and diseases. Among the general population, rates of key risk factors such as diabetes, hypertension, dyslipidemia, obesity, strokes are projected to surge significantly [2]. The primary goal of analyzing both the overall population and specific subpopulations (introduced in later section) is to understand how disease dynamics or risk factor distributions differ among varied demographic or clinical groups. This approach enables us to analyze both the overall population and specific subpopulations, providing insights into how disease dynamics or risk factor

distributions vary among different demographic or clinical groups. Additionally, we will compare rankings of mutual information to those of the overall population if there is a specific pattern for each disease.

3 Dataset Description

The heart disease indicators dataset that was utilized in this study was obtained from the Kaggle library: <https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data> [3]. To summarize, this dataset is derived from the Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the CDC since 1984. The dataset for the year 2015 comprises responses from 441,455 individuals with 330 features, including questions directly asked of participants and calculated variables. After cleaning, the dataset contains 253,680 survey responses, with a notable class imbalance: 229,787 respondents do not have or have not had heart disease, while 23,893 have reported heart disease. In terms of the variables included in the dataset, there is one binary target variable: HeartDiseaseorAttack and 21 feature variables (e.g. Education, Income, BMI, PhysActivity) that are either binary or ordinal.

3.1 Dataset Cleaning and Preparation

Before working on the heart disease indicators dataset, we performed a data cleaning process to increase the accuracy and reliability of the dataset. First, we searched the dataset for any missing values since they can skew the results of the correlation analysis and may lead to inaccurate interpretations. After checking the dataset for missing values, we found that there were no missing values. Then, there are 253680 total observations for this dataset.

We then also divided the variable types into binary and non-binary categories. Binary predictor variables are: highBP, HighChol, CholCheck, DiffWalk, Sex, Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, Stroke, HeartDiseaseorAttack, Diabetes (we combine 1 and 2 to 1, so finally it has 2 values: 0 and 1). The non-binary ones are: GenHlth, MentHlth, PhysHlth, Age, Education, Income, BMI. The stats summary of each variable are shown below:

Table 1: Summary Statistics of Health Indicators

Variable	Mean	Standard Deviation
Heart Disease or Attack	0.094186	0.292087
High BP	0.429001	0.494934
High Chol	0.424121	0.494210
Chol Check	0.962670	0.189571
BMI	28.382364	6.608694
Smoker	0.443169	0.496761
Stroke	0.040571	0.197294
Diabetes	0.296921	0.698160
Physical Activity	0.756544	0.429169
Fruits	0.634256	0.481639
Veggies	0.811420	0.391175
Heavy Alcohol Consumption	0.056197	0.230302
Any Healthcare	0.951053	0.215759
No Doc because Cost	0.084177	0.277654
General Health	2.511392	1.068477
Mental Health	3.184772	7.412847
Physical Health	4.242081	8.717951
Difficulty Walking	0.168224	0.374066
Sex	0.440342	0.496429
Age	8.032119	3.054220
Education	5.050434	0.985774
Income	6.053875	2.071148

Next, we conducted a correlation analysis by producing a correlation matrix for our variables. A correlation matrix is a useful tool in data analysis as it summarizes the relationships between variables within a dataset. This information is invaluable for identifying patterns, assessing the degree of independence or dependence among variables, selecting relevant features for analysis, and prioritizing variables based on their correlations with the target outcome (diabetes, stroke, and heart disease).

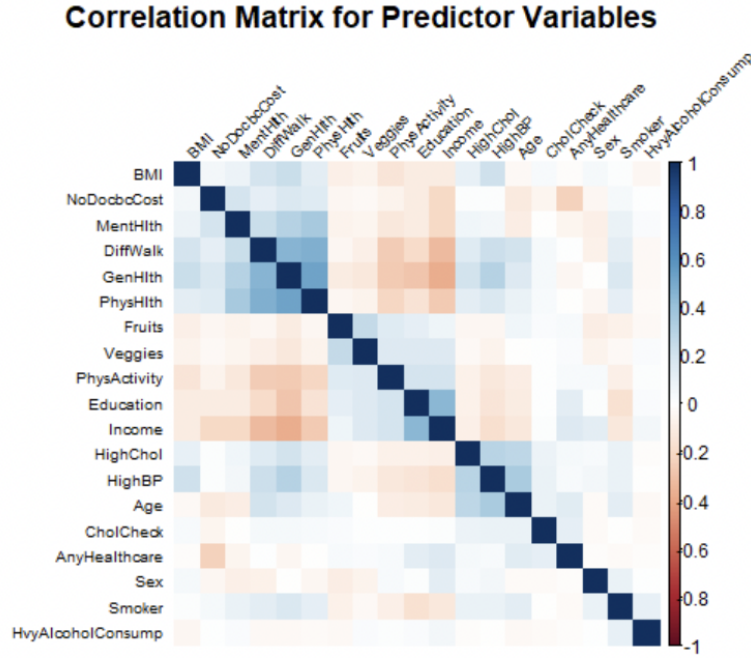


Figure 1: Predictors Correlation Matrix

In this predictors correlation matrix, we can see that blue shading indicates a positive correlation, suggesting a moderate to high linear relationship where increases in one variable are associated with corresponding increases in another. Light orange areas represent moderate negative correlations, indicating a more moderate but still noticeable relationship between variables. On the other hand, white shading signifies no correlation or very weak correlations, suggesting little to no linear relationship between the variables. Using the information from the correlation matrix however, we can see that only a few predictors exhibit strong positive correlations (darker blue areas) as they indicate potential dependencies, and most predictors have areas with no or weak correlations (white areas) suggest potential mutual independence between these variables.

After that we divided the responses (Stroke, HeartDiseaseorAttack, Diabetes) into 8 categories using 3 numbers, following the order above) to represent the response feature. For example, 1-0-1 represents the subpopulation has Stroke and Diabetes, but do not have HeartDisease or Attack.

4 Methodology

Entropy: we used entropy to measure uncertainty or randomness in our data. By calculating entropy for the response variables (diabetes, stroke, and heart disease),

we can quantify the amount of information or predictability associated with each variable. The formula that we used for calculating entropy is:

$$H(X) = - \sum_{i=1}^n P(X_i) \log_2 P(X_i)$$

Conditional Entropy: we leveraged conditional entropy (CE) to evaluate the relationship between the response variables (diabetes, stroke, and heart disease) and predictor in the dataset. The formula for the conditional entropy is given by:

$$H(Y | X) = - \sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2 \frac{P(x, y)}{P(x)}$$

Where Y represents the response variables and X is the fused variable represented as $X = (X_1, X_2)$.

We calculated $CE[Y | X]$ to quantify the uncertainty in each response variable given the knowledge of the predictor X . Additionally, if X is a fused variable, such as $X = (X_1, X_2)$, we compared $CE[Y] - CE[Y | X]$ with $CE[Y] - CE[Y | X_1] + CE[Y] - CE[Y | X_2]$ to determine the presence of interacting effects. Lower values of conditional entropy indicate stronger associations or dependencies between the response variables and the fused variable, highlighting their mutual influence and the potential for joint effects in our analysis.

Mutual Information: We calculated the mutual information between X and Y by:

$$I[X, Y] = CE[Y] - CE[Y | X] = CE[X] - CE[X | Y]$$

We calculated the mutual information between fused variable (X_1, X_2) and the response variable Y using the formula:

$$I[(X_1, X_2), Y] = CE[Y] - CE[Y | X_1] + CE[Y] - CE[Y | X_2] + I[(X_1, X_2) | Y] - I[X_1, X_2]$$

Where $I[(X_1, X_2), Y]$ represents the mutual information between non-fused covariates X_1 and X_2 with respect to Y , $CE[Y]$ is the conditional entropy of Y , and $CE[Y | (X_1, X_2)]$ is the conditional entropy of Y given the knowledge of both X_1 and X_2 .

The ecological effect ($I[(X_1, X_2), Y]$) indicates the combined impact of X_1 and X_2 on Y , beyond their individual effects. This formula allows us to evaluate the relationships between features and the response variable, providing insights into feature importance and selection for our analysis of mutual independence or dependence between diabetes, stroke, and heart disease. Utilizing this methodology aids in identifying influential predictors and enhances our understanding of predictive factors in our cardiometabolic dataset.

Odds Ratio: We utilized odds ratios as a statistical measure to assess the relationship between our response variables and our predictor variables, aiding in the assessment of their mutual independence or dependence. The formula of the odds ratio is given by:

$$\text{Odds Ratio} = \frac{A/C}{B/D}$$

Table 2: Odds Ratio Table

	Outcome Yes	Outcome No
Predictor Yes	A	B
Predictor No	C	D

It is important to note that odds ratios are applicable only to binary variables, and we computed them based on contingency tables, which helped quantify the likelihood of an event (such as disease occurrence) given the presence or absence of another condition (e.g., health behavior).

5 Results & Discussion

Firstly, we calculate the odds ratio for all the binary predictors with respect to each individual response.

Table 3: Odds Ratio Between Health Indicators and Diseases

Predictor	Stroke	Heart Disease or Attack	Diabetes
HighBP	4.02	4.59	4.78
HighChol	2.58	3.59	3.24
CholCheck	2.60	3.64	5.87
Smoker	1.86	2.20	1.41
PhysActivity	0.49	0.54	0.50
Fruits	0.87	0.87	0.79
Veggies	0.62	0.73	0.68
HvyAlcoholConsump	0.64	0.59	0.41
AnyHealthcare	1.25	1.40	1.21
NoDocbcCost	1.69	1.41	1.41
DiffWalk	5.24	4.27	3.70
Sex	1.03	1.80	1.18

When examining the overall odds ratios, we observed notable differences compared to the results obtained from mutual information analysis (figure 2, 4, 7) for the overall population. Due to these inconsistencies, we have opted to rely solely on

mutual information for analysis within the subpopulation. The decision to prioritize mutual information over odds ratios stems from specific observations, particularly highlighted in the analysis of the effect of CholCheck on Diabetes within the subpopulation. Despite the high odds ratio associated with cholesterol tests among diabetic patients, signifying a frequent occurrence of cholesterol tests among this group, the corresponding mutual information value is low. This discrepancy suggests that undergoing a cholesterol test may not significantly influence the development of diabetes in patients. Additionally, high conditional entropy (1.92815) and low mutual information (0.00686) between CholCheck and diabetes suggest that while cholesterol tests (CholCheck) are widely performed among diabetic patients, they have little direct impact on the development of diabetes. Therefore, we decided to mainly use mutual information only for the rest of the analysis since mutual information covers a more complex association compared to the odds ratio(only linear).

Conditioned Subpopulation:

We structured our analysis by defining subpopulations based on the exclusive presence of one of two diseases, either Stroke or Heart Disease, specifically excluding cases where both diseases occur simultaneously. This method helps us focus on how the presence of one condition influences the development of another condition, considered as target. For example, when studying the influence of predictors on Diabetes, we consider subpopulations that include individuals diagnosed with either Stroke or Heart Disease alone. Specifically, these subpopulations are categorized as follows: having Stroke but not Heart Disease or having Heart Disease but not Stroke. This division allows us to isolate the effects of other diseases(response) on targeted disease(response), allowing us to explore the association between predictors and response in a more concise way.

Diabetes:

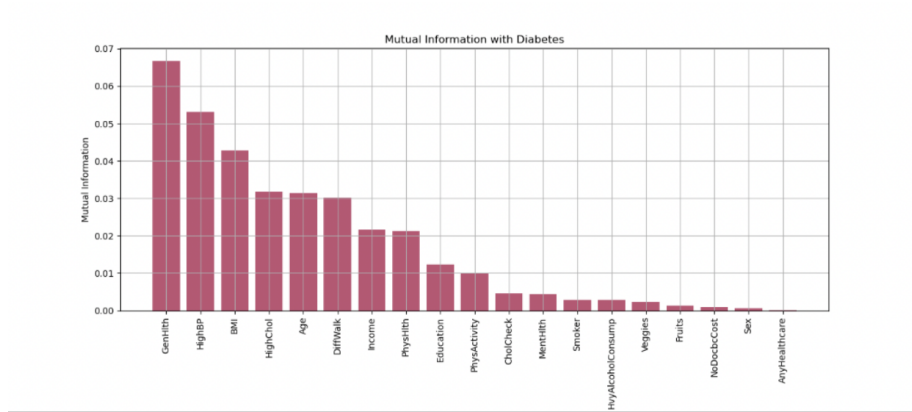


Figure 2: Mutual Information Histogram for Diabetes (Overall)

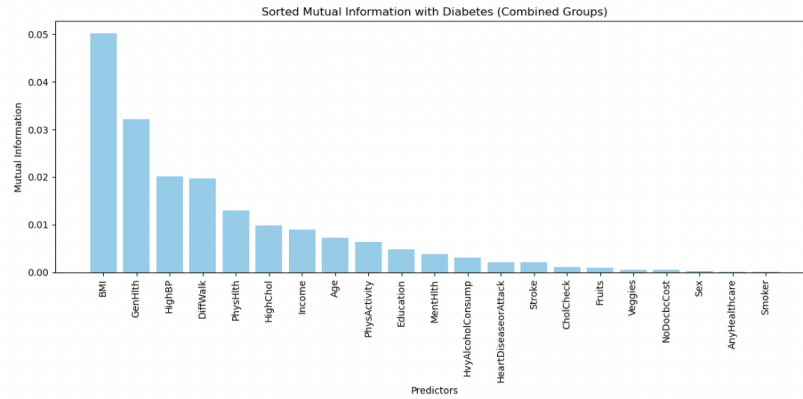


Figure 3: Mutual Information Histogram for Diabetes (Subpopulation)

Starting with our first subdivision group, target on diabetes, we noticed that the top features of diabetes in both the overall population and the subpopulations are the patient's body mass index (BMI), general health (GenHlth), and high blood pressure (HighBP) according to Figure 2 and Figure 3. The consistency across groups suggests that these factors are robust features of diabetes risk, irrespective of other cardiovascular conditions. Moreover, the mutual information for BMI in the subpopulation has been significantly increased compared to the overall dataset, indicating that BMI plays a more important role in diabetes risk than individuals who are already suffering from either stroke or heart disease.

Heart Disease or Attack:

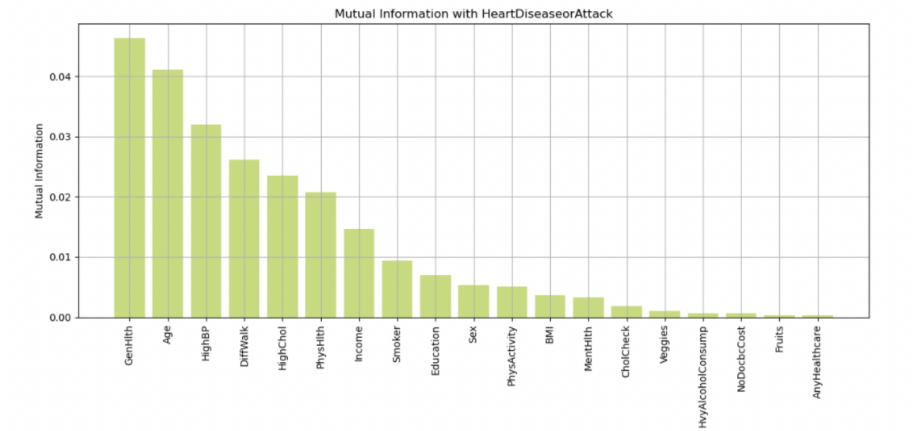


Figure 4: Mutual Information Histogram for Heart Disease or Attack (Overall)

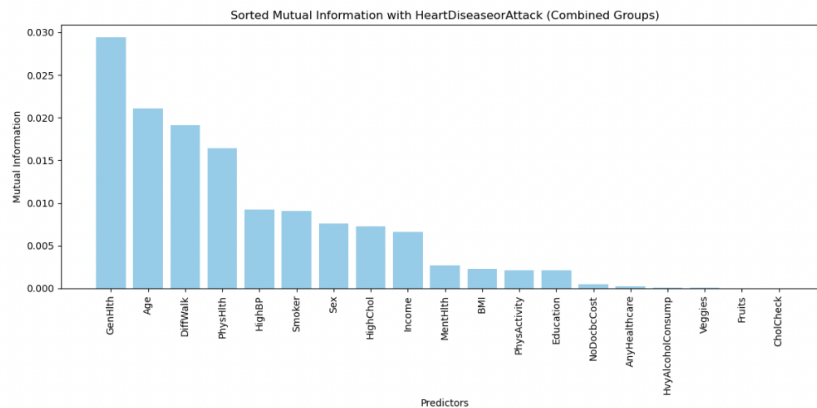


Figure 5: Mutual Information Histogram for Heart Disease or Attack (Subpopulation)

Onto the next subdivision group targeting on heart disease, we can discover that general health (GenHlth), difficulty-with-walking (DiffWalk), and patient age (Age) emerge as the top recurring highest features in both the overall and subpopulation plot. This indicates that these factors are influential across all conditions.

However, we noticed something counter-intuitive where the mutual information score for HighBP were high in the overall population in Figure 4, but it is relatively low in the subpopulation plot in Figure 5 because in terms of medical context, high blood should both have positive associations with heart disease or attack. To study this

unusual strong occurrence, we decided to investigate the interaction effect of other factor with highBP to heart disease or attack.

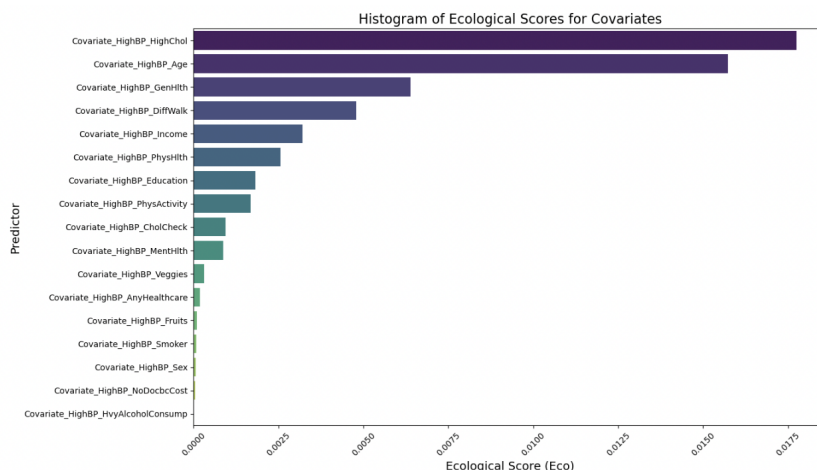


Figure 6: Ecological Scores Histogram

The investigation of interaction between HighBP and HighChol revealed interesting findings, concerning the unusual occurrence from the plots displayed earlier. Initially, the individual Mutual Information (MI) values for high cholesterol and high blood pressure were observed to fall within the middle range of the overall MI values concerning heart disease response. However, when hypertension (HighBP) and high cholesterol (HighChol) were considered as fused variables, the resulting ecological scores surpassed the individual MI values, indicating a strong interaction effect. This interaction effect suggests that the presence of one condition (high cholesterol) amplifies the impact of the other (high blood pressure) on the risk of heart disease. Consequently, when high blood pressure is coupled with high cholesterol, the risk of developing heart disease escalates significantly. This particular investigation explores how high blood pressure (HighBP) and high cholesterol (HighChol) interact to amplify the risk of heart disease, when considered together.

Stroke

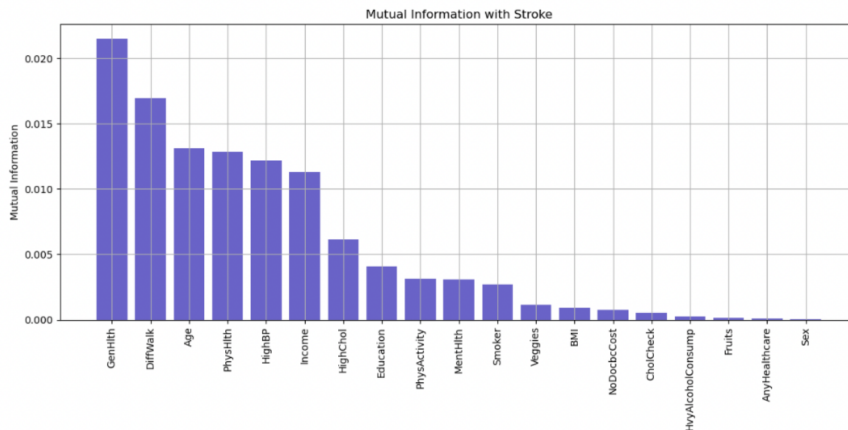


Figure 7: Mutual Information Histogram for Stroke (Overall)

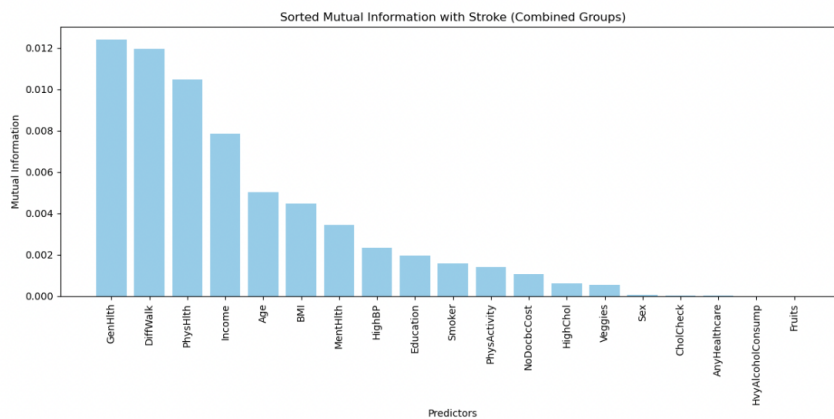


Figure 8: Mutual Information Histogram for Stroke (Subpopulation)

For our last subdivision targeting on stroke, we can see on Figure 7 and Figure 8 that general health (GenHlth), difficulty-with-walking (DiffWalk), and physical health (PhysHlth) are the common highest features in both the overall and subpopulation plot. Although the ranks of the features varies, but in general, the ranks are not too far away to their original positions.

6 Conclusion

To summarize our findings, the consistent patterns observed in the mutual information rankings across subpopulations indicate a robust association between certain

predictors and cardiometabolic diseases, regardless of the presence of other conditions.

Similar patterns of mutual information across these subpopulations indicate that certain predictors are robustly associated with a cardiometabolic disease irrespective of the presence of one of the other two diseases. In this case, general health’s consistent ranking as a top predictor in both the overall population and subpopulations suggests it holds a central role in predicting the risk of these major health conditions. This implies a strong association between overall perceived health status and the risk of developing stroke, diabetes, and heart disease, suggesting that interventions targeting general health are likely to be effective in managing diabetes risk universally. Based on that, we can come up with some strategies based on GenHlth for preventing cardiometabolic disease. These strategies could include promoting healthy lifestyle changes, such as increased physical activity and balanced diets, which are universally beneficial for good general health, resulting in a good intervention of cardiometabolic disease.

7 References

1. Centers for Disease Control and Prevention. (2024). Heart disease facts. Centers for Disease Control and Prevention. <https://www.cdc.gov/heartdisease/facts.html>
2. Roth, S. (2022, August 1). New US population study projects steep rise in cardiovascular diseases by 2060. American College of Cardiology. <https://www.acc.org/About-ACC/Press-Releases/2022/08/01/16/37/New-US-Population-Study-Projects-Steep-Rise-in-Cardiovascular-Diseases-by-2060>
3. Teboul, A. Heart disease health indicators dataset. Kaggle. <https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data>