Summary Paper: Investigating Indicators of Heart Disease

**Statistics for Data Science Course Group 11**

Jennifer Attema, Paras Bhatia, John Ing, Rafick Jungul, Rosy Zhou

**Project Overview**

Cardiovascular disease (CVD), commonly known as heart disease, remains a leading cause of mortality worldwide. Understanding the factors contributing to CVD is crucial for prevention and intervention strategies. This report investigates potential indicators of heart disease based on an annual survey of over 400,000 adults conducted by the CDC in 2020.

Ultimately, this investigation seeks to answer the following key research questions:

1. Does the dataset support the common assumptions that the selected indicators contribute to heart disease?
2. What is the relationship between general health and heart disease and how might other interconnected factors could be influencing this result?

For all indicators, we start with the null hypothesis assuming that there is no relationship and determine the appropriate test(s) to identify if a significant difference or relationship exists between the indicators and the presence of heart disease.

**Analysis and Conclusions**

**Smoking and Heart Disease Association:**

* A **Chi-Square Test** revealed a highly significant association between smoking and heart disease (χ² = 3713.0331, p < 0.0001). This indicates that smoking individuals in this dataset have a significantly higher likelihood of heart disease.

**Alcohol Consumption and Heart Disease Association:**

* The **Mann-Whitney U test** revealed a significant difference in heart disease occurrence between alcohol drinkers and non-drinkers. Alcohol drinkers showed a lower incidence of heart disease compared to non-drinkers, even without assuming normality in the data.
* The **Chi-Square test** confirmed a significant association between alcohol consumption and heart disease. It indicated that alcohol drinkers are less likely to have heart disease, suggesting a potential link between alcohol consumption and a reduced risk of heart disease.

**Sleep Time and Heart Disease Association:**

* **Spearman's Correlation** was very small (0.0075), indicating that the sleep time and heart disease have a very weak positive monotonic relationship, suggesting that as sleep time increases, heart disease also tends to increase. The small p-value (2.11e-05) indicates the observed correlation is unlikely due to random chance.

**Kidney Disease and Heart Disease Association:**

* The **Chi-Square Test** results indicated that there is an association between kidney disease and heart disease. Since all of the expected frequencies are greater than 5, the Chi-Square Test results can be trusted.

**BMI and Heart Disease Association:**

* Both the **Spearman’s Coefficient** & **Mann-Whitney U Test** revealed extremely low p-values (3.7 e-231 and 9.0 e-231, respectively) suggesting the null hypothesis can be rejected as there is a significant difference between the means of the two groups.

**Mental Health and Heart Disease Association:**

* A **Mann-Whitney U Test,** used due to the non-normal distribution of mental health data, showed borderline evidence of an association between mental health and heart disease (p = 0.0514).

**General Health:**

* Our analysis explored the relationship between self-reported general health and heart disease, testing the hypothesis that poorer general health is significantly associated with a higher prevalence of heart disease and other health conditions such as diabetes, walking difficulties, and kidney disease. Additionally, we sought to better understand the factors driving individuals' self-perception of their general health.
* A **Chi-Square Test** of independence confirmed this hypothesis, showing a strong association between general health ratings and heart disease (χ² = 21,542.177, df = 4, p < 0.001).
* Additional chi-square and **Kruskal-Wallis** tests demonstrated significant relationships between general health and other categorical (e.g., diabetes, age category) and numeric variables (e.g., BMI, sleep duration).
* A **logistic regression analysis** further highlighted key predictors of self-reported poor general health, identifying difficulty walking, diabetes, heart disease, and kidney disease as strong risk factors (odds ratios ranging from approximately 1.5 to 2.38), while physical activity and female sex emerged as protective factors.
* Ultimately, our findings underscore the complexity and subjectivity of general health as a metric, reflecting both the influence of underlying risk factors on health perception and the bidirectional relationship between poor self-reported health and heart disease.

**Overall Conclusions:**

This investigation suggests a number of relationships may exist between heart disease and other health metrics based on the CDC 2020 dataset. In all cases, it must be emphasized that these relationships, many of which were significant, only show association and do not imply causation. However, the fact that these relationships exist does warrant further investigation into causal relationships and whether improving lifestyle habits in a large population over time can actually reduce the prevalence of heart disease or whether the condition is largely genetic. Further reading of existing studies using additional datasets is recommended before drawing any major conclusions from the investigations of this report alone.