Introduction

Heart disease is a leading cause of death worldwide. Due to the potential for heart disease to cause negative effects, such as death, it may be important to investigate ways to prevent heart disease.

One way to prevent heart disease is to examine the risk factors for heart disease. Understanding risk factors may help prevent heart disease development because, for example, it may allow medical practitioners to identify individuals who are susceptible to developing heart disease. After identification, early preventative interventions could be applied to this high-risk group, reducing the likelihood of heart disease.

Research Questions and Hypotheses

Therefore, the aim of the current investigation is to investigate the factors that contribute to the development of heart disease. More specifically, we investigate the three following questions:

1. What lifestyle variables are related to the diagnosis of heart disease?
2. What variables are most important to the diagnosis of heart disease?
3. What groups of people are most likely to develop heart disease? #3) Does the effect of the most important predictor variable differ across levels of other predictor variables?

We hypothesize that, based on previous research (Ryo, Cho & Kim, 2012), that a number of variables (e.g. body mass index, smoking habits) will be related to heart disease. Furthermore, we will investigate the questions of identifying importance of coefficients, and identification of high risk groups for heart disease in an exploratory manner.

Methods and Dataset

To investigate our research questions, we obtain a dataset on Kaggle about heart disease, a subset of data collected by the Center for Disease Control and Prevention (CDC) in 2022. The CDC conducted telephone interviews for 401,958 residents of the United States across all 50 states. Participants were interviewed about a number of (that is, 279) general lifestyle factors (e.g. smoking habits) and chronic health diseases.

The dataset on Kaggle was filtered to contain only variables related to heart disease, resulting in 18 heart-disease related factors, and cleaned (e.g. missing values were removed), resulting in 319,795 responses with no missing values. Amongst the dataset of 319,795 responses we obtained on Kaggle, 18,078 duplicate responses were found and removed, resulting in a final sample size of 301,717 observations.

Figure 1 provides a full description of variables in the dataset, including the questionnaire presented to interviewees, response type, observed distribution. In total, the dataset contains 17 lifestyle predictor factors (3 numeric, 14 categorical) to investigate on self-reported heart disease.

Figure 1. Variables in the Dataset

Variable Name	Questionnaire	Data Type	Observed distribution
HeartDisease	Have you ever had coronary heart disease or myocardial infection in your life?	Binary (yes, no)	Imbalanced binary (9% yes, 91% no)
BMI	What is your body mass index?	Continuous	Normal (mean = 28.3, variance = 40)
Smoking	Have you smoked at least 100 cigarettes in your entire life?	Binary (Yes, no)	Binary (59% no, 41% yes)
AlcoholDrinking	Do you drink heavily? (Men: more than 14 drinks a week, women: more than 7 drinks a week)	Binary (Yes, no)	Imbalanced binary (7% yes, 93% no)
Stroke	Have you ever had a stroke?	Binary (Yes, no)	Imbalanced binary (4% yes, 96% no)

Variable Name	Questionnaire	Data Type	Observed distribution
PhysicalHealth	Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?	Discrete (0-30 days)	Poisson (lambda = 3.3)
MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?	Discrete (0-30 days)	Poisson (lambda = 3.9)
DiffWalking	Do you have serious difficulty walking or climbing stairs?	Binary (Yes, no)	Imbalanced binary (14% yes, 86% no)
Sex	Are you male or female?	Binary (Yes, no)	Binary (48% Male, 52% Female)
AgeCategory	What is your age group?	Ordinal categorical (e.g. 18-24)	Normal
Race	What is your race? (recoded to binary variable)	Binary (White, non-white)	Imbalanced binary (77% white, 23% non-white)
Diabetic	Have you ever had diabetes?	Nominal categorical (No, borderline diabetes, yes, yes (during pregnancy))	Positive skew (86% no, 14% non-no)
PhysicalActivity	Have you had physical activity or exercise during the past 30 days other than their regular job	Binary (Yes, no)	Imbalanced binary (78% yes, 22% no)
GenHealth	Would you say your health in general is	Ordinal categorical (Poor, fair, good, very good)	Negative skew (poor 4%, fair 10%)
SleepTime	On average, how many hours of sleep do you get in a 24-hour period?	Continuous	Normal (mean = 7, variance = 2)
Asthma	Have you ever had asthma?	Binary (Yes, no)	Imbalanced binary (14% yes, 86% no)
KidneyDisease	Have you ever had kidney disease (not including kidney stones, bladder infections)?	Binary (Yes, no)	Imbalanced binary (4% yes, 96% no)
SkinCancer	Have you ever had skin cancer?	Binary (Yes, no)	Imbalanced binary (9% yes, 91% no

Analysis

Question 1: What lifestyle variables are related to heart disease?

```
To examine what lifestyle variables are related to heart disease, we chose to run a logistic regression
To run the logistic regression model, we followed two broad steps: first, we aimed to create the best f
```

Question 2: What variables are most important to the development of heart disease?

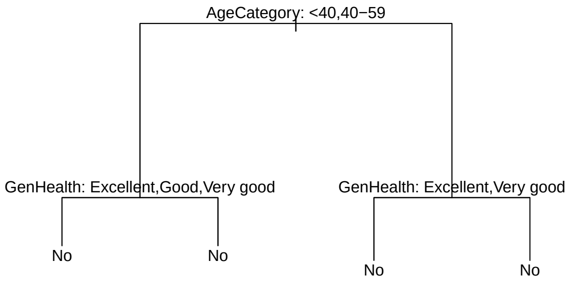
```
xxx
```

Question 3: In the real world, what groups of people are most likely to develop heart disease?

```
To examine what groups of people are most likely to develop heart disease, we ran a decision tree class
To run the decision tree models, we followed the broad steps of:
1. Fit decision tree regression.
2. Prune the tree with cross validation by deviance.
3. Filter data of vulnerable group based on the produced boundaries.
4. Repeat the above step with the filtered data to identify people who has higher risk, until the proba

We run this process for three times and the tree is show below.
```

Figure #3.1 Decision Tree on the whole dataset

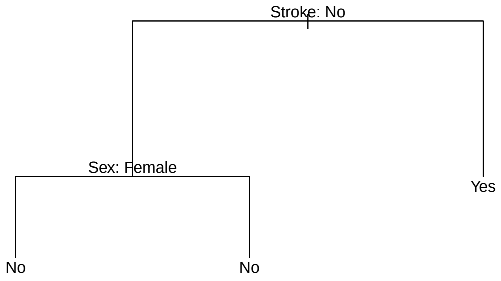

AgeCategory: <40,40−59
GenHealth: Excellent,Good,Very good
GenHealth: Excellent,Very good
No
No
No
No

We can see from the results of the decision tree that the most certain thing about the decision tree is that if you are under 59 years of age and in good overall health, then the decision tree will assume that you will not develop heart disease. Other factors such as mental health, race, physical activity, etc., have little impact on whether or not you will develop heart disease if you meet the age and overall health criteria.

Now we have identified the `low risky group`: they are **under 59 years** and their overall **health is good or above**. So let's go a step further and identify those who are not in this range, as `higher risk group`. If removing those who are not in the low risk group, we can select total sample of 26945 (only 8.93% of the total sample). This people are at a higher risk of getting heart disease: The proportion of having heart disease increases from 9.03% in the whole dataset to 31.30%

If we repeat fitting decision tree model to the `high risky group`, we can get another boundary to distinguish if people have heart disease.

Figure #3.2 Decision Tree observations of higher risk



Taking the screened `higher riksy group` observations to the next step of decision tree regression, we identified another important signal: **whether or not the person had a stroke**. If there was no stroke the decision tree would assume that the person would not have had a heart attack. In fact if we pick out the people who are already in our risk population who also had a stroke, we can see that the risk of having heart disease if they had a stroke increased from 31.10% to 50.67%. This is also a significant increase. Such a result is not difficult to explain. Stroke is often associated with hardening and blockage of blood vessels, and this often indicates that the patient has a worse blood circulation, which is also an indicator of heart disease. Now that we have greatly identified our `extremely risky group` by age, overall health status, and whether or not we have had a stroke. Beside, sex now seems also become a important factor. In fact, for people falls into `extremely risky group`, men are at higher risk than women (56.54 for men and 43.72% for women).

Table 2: Vulnerable Groups Identification

Groups	Condition	Probability of Having Heart Disease
Whole population benchmark	NA	9.03%
Low risky group	Age below 59 or in general good health condition	6.85%
High risky group	Age above 60 or in health condition below good	31.10%
Extremely risky group	In High risky group and had stroke history	50.67%

Discussion

The present investigation aimed to develop a better understanding of risk factors that contribute to heart disease by asking three broad questions. First, we investigated what variables are related to the development of heart disease. Our investigation identified 15 relevant variables (body mass index, smoking habits, alcohol habits, prior history of stroke, mental health, difficulty walking, gender, age, race, diabetes diagnosis, general health, sleep time, asthma diagnosis, kidney disease diagnosis, skin cancer diagnosis) that are positively linked to the development of heart disease. Second, we investigated what variables are most important to the development of heart disease. Our investigation showed X, Y, Z, as the most important risk factors to heart disease, with x increasing probability of obtaining heart disease by ___-. Third, we investigated what groups of people are most likely to develop heart disease. Our investigation showed that men over the age of 45, with prior history of stroke, and poor health in general, were most likely to experience heart disease.

Taken together, the findings from the current investigation identified, in a large-scale and representative sample of residents in the United States, risk factors that may contribute to the development of heart disease, findings of which may have implications for the real world, such as informing medical practitioners.

Limitations and Future Directions

While the present investigation yielded several insights into the risk factors related to heart disease, it should be noted that there were also a number of limitations.

One limitation is that, although our investigation yielded a large number of significant risk factors linked to heart disease, thefactors within this investigation is only able to explain a weak proportion of the variation in heart disease, with an adjusted R-squared of **0.2** (though see Ryo et al., 2012 for similar R-squared values). In other words, though risk factors were identified, these risk factors may be less relevant because they may not be the main driving factors behind heart disease onset.

Why may there be a low R-squared value? There may be a variety of reasons. For one, it is possible that the methods of the current investigation may contribute to a low R-squared value. For example, it is possible that that the current investigation fails to capture important predictor variables (e.g. biological or genetic markers) that are relevant to heart disease onset, to explain for the low variation explained in heart disease. Furthermore, it is also possible that the subjective nature of the methodology (e.g. telephone survey, self-reported heart disease, retrospective self-reports of exercise history) may hinder the extent to which explanatory variables can be related to response variables, compared to more objective measures of heart disease (e.g. biological measurements of heart disease).

Second, it is also possible that the current modelling techniques may be the result of low R-squared values. For example, xxx.

Future work should improve upon study methodology and statistical modelling techniques to provide risk factors that have stronger explanatory power on heart disease onset, because xyz.

References

Ryoo, J.-H., Cho, S. H., & Kim, S.-W. (2012). Prediction of risk factors for coronary heart disease using framingham risk score in Korean men. PLoS ONE, 7(9). https://doi.org/10.1371/journal.pone.0045030