In [4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn as skl
import seaborn as sb
```

Use the dataset framingham, the data set is used to predict the 10 year risk of coronary heart disease CHD. The following is description of each feature.

• Sex: male or female(Nominal)
• Age: Age of the patient;
Behavioral
• Current Smoker: whether or not the patient is a current smoker
• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.
Medical( history)
• BP Meds: whether or not the patient was on blood pressure medication
• Prevalent Stroke: whether or not the patient had previously had a stroke
• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
• Diabetes: whether or not the patient had diabetes
Medical(current)
• Tot Chol: total cholesterol level
• Sys BP: systolic blood pressure
• Dia BP: diastolic blood pressure
• BMI: Body Mass Index
• Heart Rate: heart rate
• Glucose: glucose level
Predict variable (desired target)
• 10 year risk of coronary heart disease CHD (“1”, means “Yes”, “0” means “No”)

In [12]:
```
HD_df = pd.read_csv(filepath_or_buffer='framingham.csv')
HD_df
```

Out[12]:

| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentH |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 39 | 4.0 | 0 | 0.0 | 0.0 | 0 | |
| **1** | 0 | 46 | 2.0 | 0 | 0.0 | 0.0 | 0 | |
| **2** | 1 | 48 | 1.0 | 1 | 20.0 | 0.0 | 0 | |
| **3** | 0 | 61 | 3.0 | 1 | 30.0 | 0.0 | 0 | |
| **4** | 0 | 46 | 3.0 | 1 | 23.0 | 0.0 | 0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **4233** | 1 | 50 | 1.0 | 1 | 1.0 | 0.0 | 0 | |
| **4234** | 1 | 51 | 3.0 | 1 | 43.0 | 0.0 | 0 | |
| **4235** | 0 | 48 | 2.0 | 1 | 20.0 | NaN | 0 | |
| **4236** | 0 | 44 | 1.0 | 1 | 15.0 | 0.0 | 0 | |
| **4237** | 0 | 52 | 2.0 | 0 | 0.0 | 0.0 | 0 | |

4238 rows × 16 columns

1. Identify what are the data types of each column. (10)

Write your answer here
• Sex: (Nominal)
• Age: (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
• Education: Ordinal
Behavioral
• Current Smoker: (Nominal)
• Cigs Per Day: (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
Medical( history)
• BP Meds: (Nominal)
• Prevalent Stroke: (Nominal)
• Prevalent Hyp: (Nominal)
• Diabetes: (Nominal)
Medical(current)
• Tot Chol: (Continuous)
• Sys BP: (Continuous)
• Dia BP: (Continuous)
• BMI: (Continuous)
• Heart Rate: (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
• Glucose: (Continuous)
Predict variable (desired target)
• 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

2. Create a correlation heatmap of the data (10)

In [2]:
```
#Enter your code here.
```

3. According to the heatmap, do you think some of the feature should not be count in the logistic regression? Which ones? (10)

```
Enter your answer here
```

4. How many empty values are there in each risk factor?(10)

In [1]:
```
#Enter your code here.
```

5. Show how you will handle these null values for each risk factor? Why? (10)

```
Enter your answer here
```

6. Handle these null values. (10)

In [3]:
```
#Enter Your code here
```

7. Show the histogram of each factors(10)

In [4]:
```
#Enter your code here
```

8. Split the data set into X_train, X_test, y_train, y_test. (10)

In [20]:
```
#Enter your code here
```

9. Train the value and show the predict result of first 10 data points. (10)

In [5]:
```
#Enter your code here

```

10. Evaluate the model.(10) (You can use any score or method to evaluate the model, but you need to explain detail about your result)

In [6]:
```
#Enter your code here
```