Heart Disease Prediction Using Statistical Models

Jaya Ren, Felix Su, Xingchen Wang, Simon Zhang
University of California, Los Angeles

Dec 11th, 2021

1 Abstract

According to the CDC, heart disease is the leading cause of death in the United States. Hence, finding ways to effectively predict heart disease would be greatly beneficial, especially because heart disease is much easier to treat when discovered early. This report is an attempt to predict heart disease using statistical models. Data used were collected ethically from several hospitals, totalling 4220 patients. An abundance of classification models in R software were used to identify potentially significant risk factors towards the prediction of heart disease. We selected the most important variables for modeling and identified linear discriminant analysis (LDA) and support vector machine (SVM) with linear kernel to be our best models.

2 Introduction

Heart disease is the leading cause of death in the United States, causing about 1 in 4 deaths. There are many risk factors contributing to the development of heart disease, such as high blood pressure, diabetes, and cholesterol levels. As the disease is easier to treat when detected early, an early diagnosis of heart disease is a crucial task for many healthcare providers. We explore a large dataset of 4220 patients to address this problem and develop a predictive categorical model that can determine with maximum accuracy whether a patient is at risk of developing heart disease or not. This was achieved first by cleaning our data and employing data visualization to do an initial screening of which variables were significant. Then, using techniques such as backward and forward stepwise regression, t-tests, and chi-squared tests, we did our initial variable selection. Finally, through some trial and error, we developed our most optimal model, which currently has a predictive accuracy of about 81.4%. More data is needed to verify the applicability of our model, and more research is needed to determine whether there are other variables that play a role in developing heart disease.

3 Exploratory Data Analysis and Data Cleansing

There are 4220 observations and 20 variables in the training dataset, with 12 categorical predictors, 7 numerical predictors, and response variable HeartDisease. The number of observations diagnosed with heart disease is about the same as the number of observations without heart disease, so we have a balanced training set.

3.1 Numerical Variables

By examining the summary of training data, we noticed zero values in numerical predictors RestingBP and cholesterol. However, it's impossible to have 0 blood pressure or cholesterol, so we have considered different ways of imputing such invalid values. First, we used both the Hmisc and mice packages to impute these 0s with mean and median. Then, we also tried to impute 0s by the mean of its corresponding age group as we considered that people in certain age groups are more likely to have higher blood pressure or cholesterol. The results of our imputation will be further discussed in the Models and Methods section.

Further, by creating density plots using the ggplot2 package in Figure 1, we found that MaxHR, Age, Oldpeak, and avg_glucose_level were good candidates among numerical predictors. We found, however, that Age and bmi [4] were more significant as categorical variables, so we recoded these two predictors to reflect that. For example, people who are older, i.e. age>60, are more likely to have heart disease. [6]

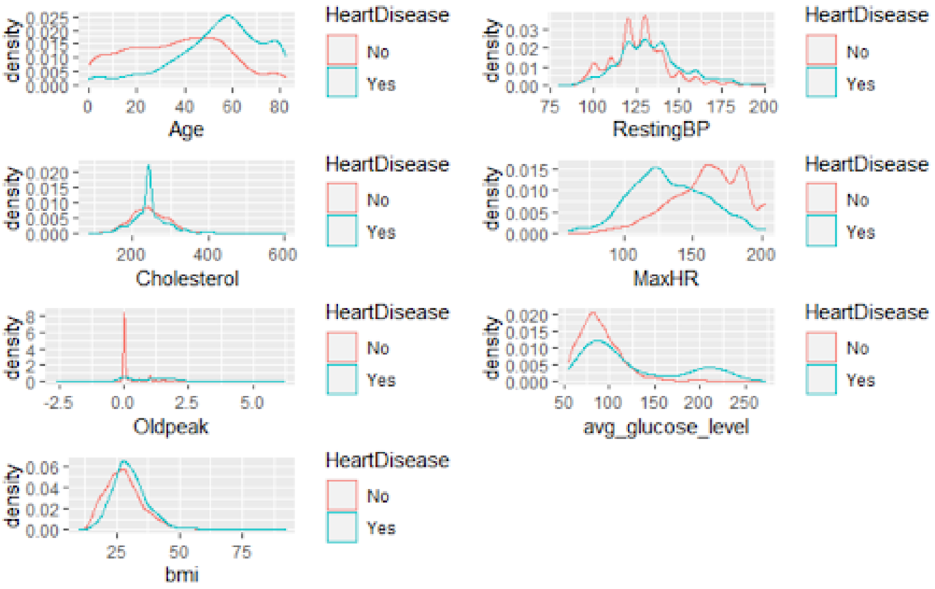


Figure 1: Density Plots for All Numerical Variables

Finally, we looked into the relationships among numerical predictors. Shown in the correlation map in Figure 2, we notice that Age and MaxHR, as well as avg_glucose_level and cholesterol [1] were highly correlated. Principal component analysis (PCA) was employed to resolve this collinearity problem, which we explore in the following sections.

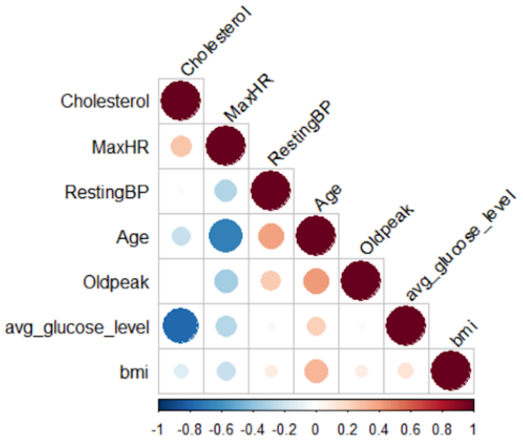


Figure 2: Correlation Map for Numerical Predictors

3.2 Categorical Variables

We noticed that there are five levels in the Sex column: F, M, Female, Male, other. Since there's only one sample labelled "other", we decided to drop that observation. We then converted F and M to Female and Male respectively so that our Sex predictor only had two levels. There are four variables that have NA values: ever_married, work_type, Residence_type, and smoking_status. They each have 631 NAs and all NAs belong to the same observations. We can either drop these NAs or impute them. However, it's also possible that these predictors don't have significant effects on our prediction of heart disease. Hence, we wanted to first examine the relationships between categorical predictors and the diagnosis of heart disease.

By creating stacked bar charts for the four variables with NAs in Figure 3, we noticed that the Residence_type and smoking_status didn't seem to be highly correlated with heart disease, as the ratios of heart disease and no heart disease in each bar are about the same for each plot. Although the work type "children" has a obviously lower ratio of heart disease, we later found out that children work type is equivalent to people who are too young to work, as shown in Figure 4, which is not really a type of work. As the information about age and heart disease is already covered in the Age column, work_type won't be very important. Lastly, we can definitely see differences in ratios across bars for marital status. However, we didn't find any research indicating the relationship between marital status and heart disease and it sounds unreasonable to assume married people have a higher chance of having heart disease, so we would like to think of this as a coincidence. Hence, we chose to remove these four predictors with NAs as it might not be worth the loss of information by dropping the rows with NAs. More detailed imputation is discussed in Section 3.3.

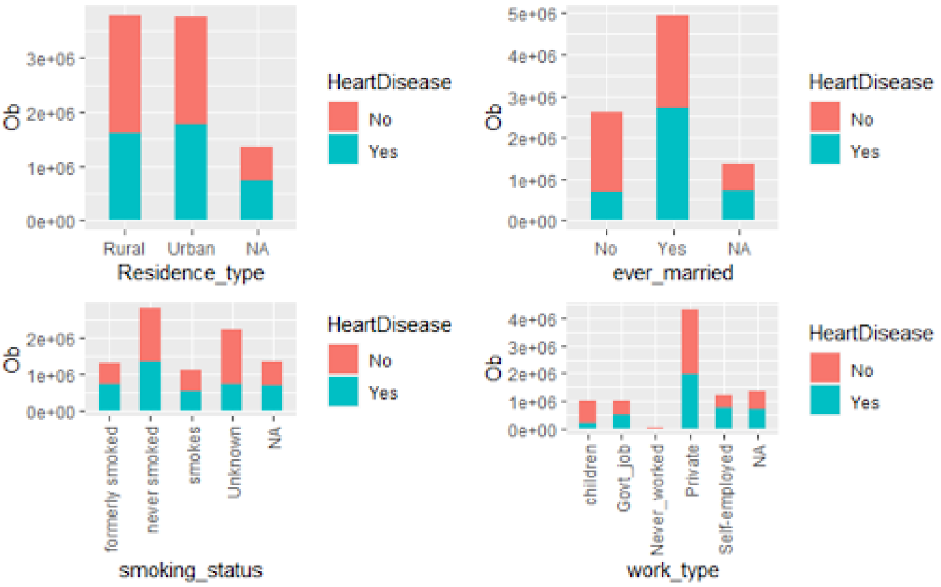


Figure 3: Stacked Bar Charts for Variables with Missing Values

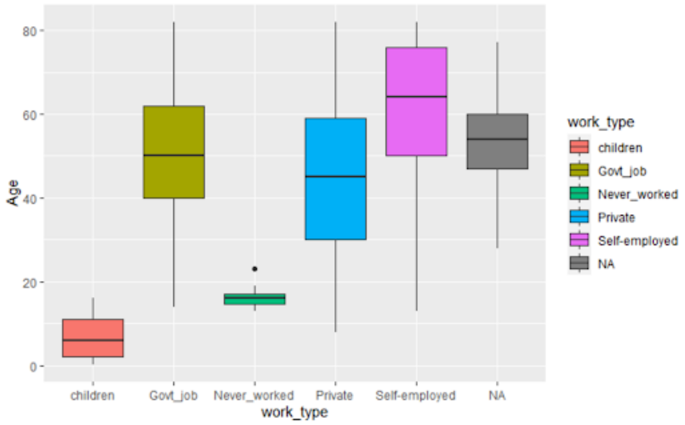


Figure 4: Age Distribution for Different Work Types

Next, we looked at the stacked bar charts for other categorical predictors. As shown in Figure 5, the diagnosis of heart disease differs greatly in stroke, hypertension, FastingBS, and ST_Slope. There are noticeable differences in the diagnosis of heart disease for different Sex and ExerciseAngina as well. Therefore, these predictors are good candidates for categorical predictors.

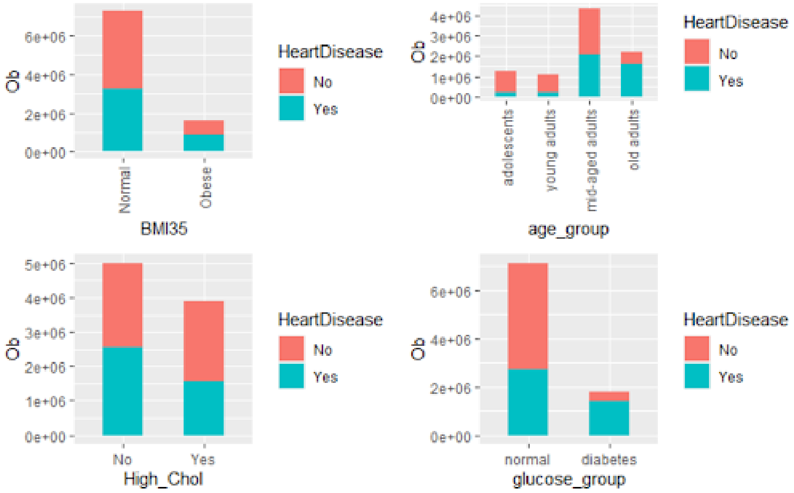


Figure 6: Stacked Bar Charts for Newly Defined Categorical Variables

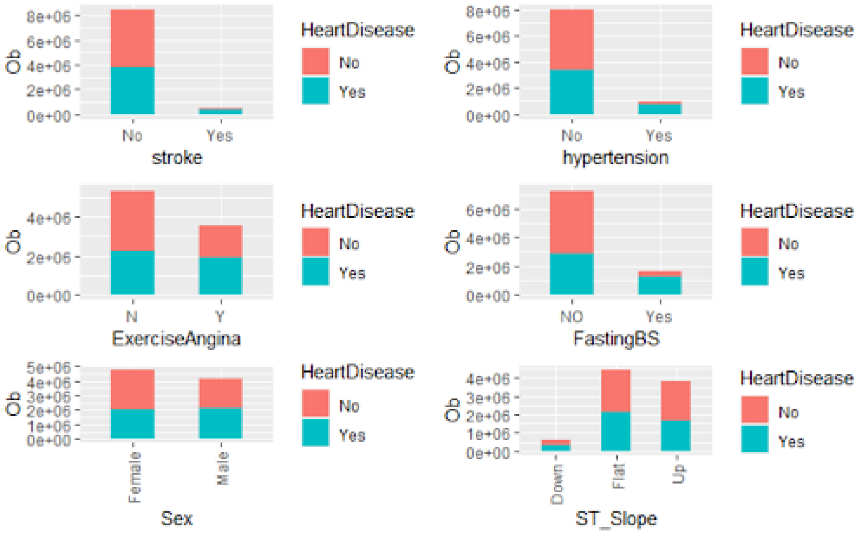


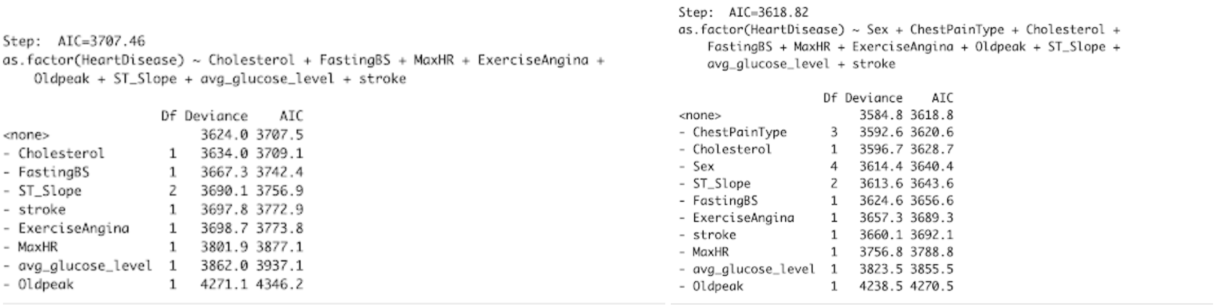
Figure 5: Stacked Bar Charts for Important Categorical Predictors

Lastly, as mentioned in the Section 3.1, BMI and Age might have more significant meanings for predicting heart disease as categories rather than continuous numbers. Therefore, we considered several ways to convert these two variables. For BMI, we considered two, three, and four subintervals. For Age, we considered two, four, and six subintervals. Age was found to have the best results when splitting into four subintervals: adolescents (0-17), young adults (17-30), mid-aged adults (30-60), and old adults($>$ 60), while BMI had the best results when splitting into two subintervals with the cutoff value of 35.

Similarly, it is well documented that cholesterol and glucose level are good predictors of heart disease [1]. Therefore, we also tried to create categorical predictors to indicate high cholesterol level and high glucose level. By creating stacked bar charts in Figure 6, we found all four categorical predictors we created have shown great differences in the diagnosis in heart disease, indicating they are good candidates for our models. However, since cholesterol and glucose level are also good candidates as numerical predictors, further investigation is needed to determine if it is better to convert them to categorical predictors instead.

3.3 Data Cleaning

Although we dropped those columns with NA eventually, we still spent a lot of time investigating them. We first used the traditional impute function in the Hmisc package to try to impute them as the mean, median or as the proportion of previous recorded values, but this method did not work very well. Then



```
Step:  AIC=3707.46
as.factor(HeartDisease) ~ Cholesterol + FastingBS + MaxHR + ExerciseAngina +
    Oldpeak + ST_Slope + avg_glucose_level + stroke

                  Df Deviance    AIC
<none>                 3624.0 3707.5
- Cholesterol      1   3634.0 3709.1
- FastingBS        1   3667.3 3742.4
- ST_Slope         2   3690.1 3756.9
- stroke           1   3697.8 3772.9
- ExerciseAngina   1   3698.7 3773.8
- MaxHR            1   3801.9 3877.1
- avg_glucose_level 1  3862.0 3937.1
- Oldpeak          1   4271.1 4346.2
```

(a) AIC

```
Step:  AIC=3618.82
as.factor(HeartDisease) ~ Sex + ChestPainType + Cholesterol +
    FastingBS + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
    avg_glucose_level + stroke

                  Df Deviance    AIC
<none>                 3584.8 3618.8
- ChestPainType    3   3592.6 3620.6
- Cholesterol      1   3596.7 3628.7
- Sex              4   3614.4 3640.4
- ST_Slope         2   3613.6 3643.6
- FastingBS        1   3624.6 3656.6
- ExerciseAngina   1   3657.3 3689.3
- stroke           1   3660.1 3692.1
- MaxHR            1   3756.8 3788.8
- avg_glucose_level 1  3823.5 3855.5
- Oldpeak          1   4238.5 4270.5
```

(b) BIC

Figure 7: Stepwise Feature Selection

we attempted to use aregimpute function since it has multiple imputation using additive regression, bootstrapping, and predictive mean matching. However, the training and testing accuracy did not have a difference as before. Finally, we tried to use the MICE package to impute those NAs, since it has specific methods for different variables. We imputed Residence_type and ever_married by method logistic regression because these two predictors all have two categories, for Residence_type, it only has Rural and Urban; for ever_married, the answer will only be yes or no. We then imputed work_type and smoking_status with polynomial regression, since they all have more than two categories. However, in our final model, we chose not to include any columns from these four original NA columns because of the lower accuracy.

After looking at the distribution value of Cholesterol, we found that those with zero values tend to have higher risk of Heart Disease. We found that 85% of the people with zero Cholesterol had heart disease. We changed these zero values to NAs and tried to impute them by the previous three methods, but the test accuracy did not improve. We also tried to replace the zero values with the median or third quartile. Neither has increased our model accuracy. Thus, we chose to leave Cholesterol as is.

4 Models and Methods

After exploratory analysis, we moved on to the next step to select important features for the modeling later.

4.1 Feature Selection

By conducting chi-square tests over all categorical predictors against heart disease, we found many predictors were correlated with heart disease, such as the age_group category we created (here we used the one with four categories for demonstration as it later showed to be the best one). By using t-tests, we found Cholesterol, MaxHR, Oldpeak, and avg_glucose_level to be important predictors.

We used backward selection based on both AIC (Figure 7a) and BIC (Figure 7b) to choose the most important predictors, and reduce the number of predictors to around 10. We tried to use exhaustive selection, but our computers could not run it successfully. We mainly use BIC as our criterion, since all the predictors included in BIC are in AIC as well. From this, we did further research on Sex [3], glucose level, blood sugar [5], ExerciseAngina, stroke [7], ST Slope [2], and several other variables that were significant. It turns out they are good predictors for heart disease according to many researches.

4.2 Modeling Fitting

We split our training dataset into a new training dataset with 70% of the original dataset and a new testing dataset with 30% of the original dataset. We ran our model on our new training dataset first and if the accuracy enhanced, we then tested our new testing dataset and submitted on kaggle to see the true testing accuracy.

4.2.1 Principal Component Analysis

Multicollinearity was detected after conducting a t-test on the numerical variables, which we fixed by using Principal Component Analysis (PCA). We chose four numerical variables, MaxHR, Oldpeak, avg_glucose_level, and Cholesterol, since they were significant in the backward selection model.

4.2.2 Logistic Regression

Logistic regression seemed to be an efficient choice, given our response variable only had two categories. We build a logistic regression model using the selected predictors, giving us a training accuracy of 0.815. However, our test accuracy was lower at 0.80632, which indicated that our logistic regression model perhaps suffered from overfitting. We believed that this model might not be the best fit, and we decided to look for a better solution.

4.2.3 Linear Discriminant Analysis

Linear discriminant Analysis (LDA) seemed like a good choice because our dataset exhibited a linear boundary. Our best model using LDA gave us a training accuracy of 0.8137 and a testing accuracy of 0.81422. As the testing accuracy was greater than the training accuracy, we believed this model to be better than logistic regression, and was perhaps a better fit for the data.

4.2.4 Random Forest

Random Forest is an efficient technique that can be applied to both regression and classification problems. We first tuned the hyperparameters mtry and number of trees, and then put all the predictors in to reduce the dimensions, but our accuracies were lower than our LDA model.

4.2.5 Support Vector Machine

We also used the support vector machine models (SVM) with both linear and radial kernels with their best parameters after tuning function, but the testing accuracy did not break through 0.813. SVM with linear kernel has similar training accuracy as LDA. Its testing accuracy is slightly lower than LDA, but it's higher than our logistic regression model. There may exist some overfitting problem for support vector machine with radial kernel since we got training accuracy over 85%.

4.2.6 Other Models Attempted

Other models like K-Nearest Neighbors, Lasso Regression, Ridge Regression and Quadratic Discriminant Analysis, did not give us a better accuracy than LDA, as shown in Figure 8.

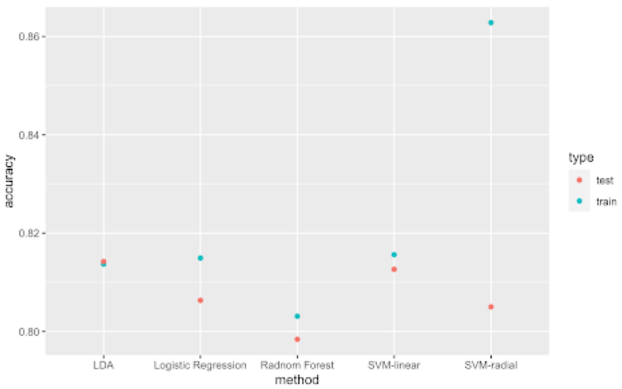


Figure 8: Testing and Training Accuracy Comparison

5 Discussion and Limitations

Each model method has its own strengths and weaknesses. For example, logistic regression and LDA perform better with linear boundaries, while QDA outperforms both for nonlinear boundaries. If the relationship between the target variable and predictors is complex, a different model or machine learning method will be much better. KNN does not work well in large datasets with high dimensions, and it is difficult to find the best K parameter to fit a predictive model.

Because the size of observations in the training sample (4220) is small relative to the total population of those with heart disease, our result may not be representative. Further, there are many types of heart disease- coronary artery disease (CAD), heart arrhythmias, and pericardial disease to name a few- several of which may have different risk factors from others. Thus, our model would suffer from an inability to generalize to all types of heart disease.

We chose not to include the columns with NA because our imputations didn't generate better accuracies, but there might be a better way to impute our missing values. We also did not remove outliers in the process of data analysis, which could have led to higher variability in the dataset.

6 Conclusion and Recommendations

By developing and testing several different classification models, this study clearly established the viability of predictive diagnosis of heart disease. By employing several statistical techniques, we identified several significant risk factors towards the development of heart disease. While not perfect, our final model does much better than a random classification, and can serve as a starting point for future study.

Due to time constraints, we failed to explore more potential transformations of the variables to develop a more precise model. Further directions of study include applying more advanced techniques from Machine Learning and Deep Learning- for example; Bayes rule or Markov Models- to build a more dynamically and self-developed model.

This was a great opportunity for us to dive deep into this interesting topic. The whole process made us try different models, compare the trade off between them, investigate the potential interaction and correlation between different predictors, deepen our statistical model understanding from the lecture, and apply this technology to this real-world project.

7 Acknowledgement

This report is part of the course project for Stats 101C at UCLA. We want to express our appreciation for Prof. Akram Almohalwas for his guidance and help throughout the course.

References

[1] Jo Ann S Carson, Alice H Lichtenstein, Cheryl AM Anderson, Lawrence J Appel, Penny M Kris-Etherton, Katie A Meyer, Kristina Petersen, Tamar Polonsky, and Linda Van Horn. Dietary cholesterol and cardiovascular risk: a science advisory from the american heart association. *Circulation*, 141(3):e39–e53, 2020.

[2] MS Elamin, DASG Mary, DR Smith, and RJ Linden. Prediction of severity of coronary artery disease using slope of submaximal st segment/heart rate relationship. *Cardiovascular research*, 14(12):681–691, 1980.

[3] Zujie Gao, Zengsheng Chen, Anqiang Sun, and Xiaoyan Deng. Gender differences in cardiovascular disease. *Medicine in Novel Technology and Devices*, 4:100025, 2019.

[4] Sadiya S Khan, Hongyan Ning, John T Wilkins, Norrina Allen, Mercedes Carnethon, Jarett D Berry, Ranya N Sweis, and Donald M Lloyd-Jones. Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. *JAMA cardiology*, 3(4):280–287, 2018.

[5] Chanshin Park, Eliseo Guallar, John A Linton, Duk-Chul Lee, Yangsoo Jang, Dong Koog Son, Eun-Jeong Han, Soo Jin Baek, Young Duk Yun, Sun Ha Jee, et al. Fasting glucose level and the risk of incident atherosclerotic cardiovascular diseases. *Diabetes care*, 36(7):1988–1993, 2013.

[6] Jennifer L Rodgers, Jarrod Jones, Samuel I Bolleddu, Sahit Vanthenapalli, Lydia E Rodgers, Kinjal Shah, Krishna Karia, and Siva K Panguluri. Cardiovascular risks associated with gender and aging. *Journal of cardiovascular development and disease*, 6(2):19, 2019.

[7] Ye Ruan, Yanfei Guo, Yang Zheng, Zhezhou Huang, Shuangyuan Sun, Paul Kowal, Yan Shi, and Fan Wu. Cardiovascular disease (cvd) and associated risk factors among older adults in six low-and middle-income countries: results from sage wave 1. *BMC public health*, 18(1):1–13, 2018.