Investigating Cardiovascular Disease Risk Factors Through Dataset Analysis and Statistical Learning

Akhil Pratyush Simhambhatla, Jason Mao, Jose Cordova, Koushik Tripurari

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for approximately 17.9 million deaths annually (World Health Organization, 2021). Modifiable risk factors such as unhealthy diets, tobacco use, harmful alcohol consumption, and physical inactivity contribute significantly to the development of CVDs (Yusuf et al., 2004; Danaei et al., 2009). The current study aims to enhance our understanding of the relationships between these factors and cardiovascular health. Previous research has identified key factors associated with CVDs, including age, gender, BMI, blood pressure, cholesterol levels, and lifestyle habits (Kannel et al., 1976; Stamler et al., 1993). However, further investigation into their interplay is needed. This study analyzes a comprehensive dataset containing 12 objective, examination, and subjective features related to cardiovascular health, including demographic characteristics, clinical measurements, and health habit characteristics to predict the presence of CVDs. Machine learning and statistical learning methods have previously been shown to successfully improve CVD risk prediction compared to traditional rules-based risk assessment scales, utilizing routinely collected healthcare and EHR data (Quesada et al., 2019). A literature review guided the choice of machine learning techniques, including tree-based methods, regularized regression, and other forms of non-linear modeling, to identify potential risk factors and patterns contributing to the development of CVDs (Smith et al., 2012; Patel et al., 2018). The findings will inform targeted prevention and intervention strategies to improve cardiovascular health outcomes.

Methods

Dataset

The Cardiovascular Disease dataset (Ulianova, 2019) was obtained from Kaggle and contains 70000 rows. Each row represents an individual subject and provides several features that describe the subject. The dataset includes a wide range of input features classified into three categories: objective, examination, and subjective features. Objective features consist of demographic and physical features including age, height (cm), weight (kg), and gender. Examination features are derived from medical examinations and include the clinical characteristics of systolic blood pressure (SBP), diastolic blood pressure (DBP), cholesterol levels (with categories: normal, above normal, and well above normal), and blood glucose levels (with categories: normal, above normal, and well above normal). Subjective features are based on behavioral information provided by the patients and cover smoking habits, alcohol intake, and physical activity. These are coded as binary variables indicating whether subjects partake in such activities.

We derived individuals’ body mass index (BMI) using their height and weight information. Because height and weight are highly correlated features, BMI is useful as a variable that incorporates information from both. Clinically, though it is an imperfect measure, BMI is used to classify obesity. Obesity and obesity-related risk factors have consistently been found to be strongly related to the risk of developing CVD (Ortega et al., 2016). Similarly, we calculate mean arterial pressure (MAP) from the SBP and DBP as useful summary measures. SBP and DBP are strongly associated with the risk of CVD, but MAP has been studied less extensively. It is unclear what combination of blood pressure measures is best used to predict CVD (Kengne et al., 2009). Individuals with highly abnormal blood pressure measures and physical measurements were excluded from the study as likely data errors. Individuals with SBP < 50, SBP > 400, DBP < 20, DBP > 400, DBP > SBP, and BMI > 50 were excluded. We preprocessed the data using centering and scaling and set a tune length of 5 when appropriate or required for modeling.

Analysis

The aim of this project is to predict the presence of CVD in patients based on different demographic, physical, clinical, and lifestyle factors. To do this, we started our model building with the most straightforward regression methodology for classification, logistic regression, which is commonly used for its ease of implementation and interpretability. Then, we advanced to other models including ridge regression, tree-based methods, and non-linear classification methods. In the following section, we will briefly discuss each model created and our model-building process. In each model, we use the features available to us to predict the presence of CVD in our subjects. We internally validate the predictive performance of our models by splitting our data into 80/20 training and training datasets. Models are fit to the training data and then used to generate predictions in the testing data. The predicted and observed values are compared to produce predictive performance measures including accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The predictive performance of our models will be summarized in the results section. Analyses were conducted in R version 4.2.3.

Logistic Regression

Logistic regression (LR) is a generalization of linear regression that models regressors against the log odds ratio of an event occurring, or in other words, the logit function of a probability *p*. This is done to bound the value of *p* between 0 and 1, which allows us to use logistic regression to model binary classification problems and make inferences about regressors (Witten et al., 2021). We utilized best subset selection to build our model and used deviance residuals and AIC values to assess model fit, finding that the model fits the data well.

Regularized Regression

To improve the performance of the regression technique, we implemented ridge logistic regression models. Ridge regression is a form of regularization model. These techniques seek to regularize, or in other words, constrain, the coefficient estimates of regression models towards zero. The purpose of this is to improve model fit and reduce variance at the cost of increased bias (Witten et al., 2021). In addition to regularizing the model, we also utilized cross-validation (CV) for this model to improve accuracy and prevent overfitting. This also helps us tune the hyperparameters of our model by automatically selecting the optimal value of lambda. After training the model and testing it with a test dataset, we can see from the sparse matrix that the model's coefficients of the predictors all shrunk towards zero compared to logistic regression, as expected for ridge regression. In addition, we fit a least absolute shrinkage and selection operator (LASSO) logistic regression model. This is another form of regularized regression model. However, unlike ridge regression, which retains all input variables into the model, LASSO regression will fully shrink coefficients to zero, fully dropping variables from a model. While this is not necessarily superior to ridge regression for prediction, this technique is useful for when inference is a goal of the model to limit the number of variables required for interpretation (Witten et al., 2021). We utilize similar techniques of CV to select an optimal lambda to fit our model. Our regularized models were fit to the same variable set as the logistic regression model for coefficient comparability.

Random Forests

Random forest (RF) is a machine learning algorithm that creates a collection of decision trees and combines their predictions to produce a more accurate result. Decision trees are used to make decisions based on the input data, and each tree within the forest is created using a random subset of the training data. The algorithm then aggregates the predictions from all the trees and selects the most common class label or the average prediction for regression tasks. By leveraging the power of multiple decision trees, random forests can effectively reduce both bias and variance, leading to a robust and accurate model. Furthermore, the algorithm provides an estimate of feature importance, enabling users to identify the most influential variables in their datasets. We utilized a grid search to find the best parameters for our random forest.

Support Vector Machines

Support vector machines (SVM) are algorithms that are frequently used to perform regression and classification tasks, though they are more popular for the latter. They are an extension of maximal margin and support vector classifiers. In essence, SVMs create hyperplanes, which, simply put, are dividing lines in *p*-dimensional space, in order to separate data points and predict their value by determining which side of the line the data point falls into. Support vector machines extend the classifiers by allowing for non-linear boundaries between classes using kernels to define the feature space (Witten et al., 2021). However, one notable weakness of SVMs is the exponential growth in time it takes to analyze large datasets. As such, in our model, we do not use a kernel based approach, but a L2-regularized linear method. We utilized CV to find the optimal cost to the margin function.

Bagging Model

The bagging (bootstrap aggregating) model is a popular ensemble learning technique that leverages multiple decision tree models to improve the accuracy and reduce the variance of the final model. This approach involves creating several random samples with replacement from the training dataset and training a decision tree model on each sample. The final prediction of the bagging model is then computed by averaging the predictions of all the individual decision trees. Bagging is an effective technique for enhancing the performance of decision tree models because it mitigates the risk of overfitting and provides better generalization performance. Moreover, bagging is robust to noisy and incomplete data. However, one potential weakness of bagging models is that they may not effectively address bias in the data. In our implementation, we used the caret package in R to train the bagging model.We trained our bagging model using the "treebag" method with 5-fold CV.

Gradient Boosting

Gradient boosting (GB) is another popular ensemble learning technique that involves combining multiple weak models, typically decision trees, to create a strong and accurate predictive model. The key difference between bagging and GB is that in GB, the individual decision trees are trained sequentially, with each new tree trying to correct the errors of the previous tree. Gradient boosting iteratively fits the decision trees to the residuals of the previous trees until the model achieves satisfactory performance. Gradient boosting is particularly effective in reducing bias and improving accuracy on complex datasets with many features. In this implementation, we used the caret package in R to train the gradient boosting model using the “gbm” method.

Table 1. Summary of Demographic, Clinical, and Behavioral Features

Variable	**Total, N = 68459**	**No CVD, N = 34624 (50.6)**	**CVD, N = 33835 (49.4)**	**p-value**
Age (years)	53.3 (6.8)	51.7 (6.8)	54.9 (6.3)	<0.001
Gender				0.046
Female	44565 (65.1)	22664 (65.5)	21901 (64.7)	
Male	23894 (34.9)	11960 (34.5)	11934 (35.3)	
Height (cm)	164.4 (7.9)	164.5 (7.8)	164.4 (7.9)	0.004
Weight (kg)	74.0 (13.9)	71.5 (13.0)	76.5 (14.4)	<0.001
BMI (kg/m^2)	27.4 (5.0)	26.4 (4.7)	28.4 (5.2)	<0.001
Diastolic Blood Pressure (mmHg)	81.3 (9.5)	78.1 (8.2)	84.5 (9.6)	<0.001
Systolic Blood Pressure (mmHg)	126.6 (16.7)	119.6 (12.5)	133.9 (17.3)	<0.001
Mean Arterial Pressure	96.4 (11.1)	91.9 (8.9)	101.0 (11.2)	<0.001
Cholesterol				<0.001
Normal	51366 (75.0)	29023 (83.8)	22343 (66.0)	
Above Normal	9266 (13.5)	3742 (10.8)	5524 (16.3)	
Well Above	7827 (11.4)	1859 (5.4)	5968 (17.6)	
Blood Glucose				<0.001
Normal	58241 (85.1)	30558 (88.3)	27683 (81.8)	
Above Normal	5038 (7.4)	2080 (6.0)	2958 (8.7)	
Well Above	5180 (7.6)	1986 (5.7)	3194 (9.4)	
Smokes?				<0.001
No	62431 (91.2)	31417 (90.7)	31014 (91.7)	
Yes	6028 (8.8)	3207 (9.3)	2821 (8.3)	
Drinks Alcohol?				0.028
No	64807 (94.7)	32712 (94.5)	32095 (94.9)	
Yes	3652 (5.3)	1912 (5.5)	1740 (5.1)	
Regular Physical Activity?				<0.001
No	13459 (19.7)	6296 (18.2)	7163 (21.2)	
Yes	55000 (80.3)	28328 (81.8)	26672 (78.8)	

Note: Categorical features expressed as n (%) and analyzed with chi-squared tests. Numerical features expressed as mean (standard deviation) and analyzed with t-tests.

KNN Model

The k-nearest neighbors (KNN) model is a non-parametric machine learning algorithm that is commonly used for classification and regression tasks. In the KNN model, the prediction for a new observation is made based on the class or value of its k nearest neighbors in the training data. One major weakness of the KNN model is that it can be sensitive to the choice of k. Choosing a value that is too small may result in overfitting, whereas a value that is too large may result in underfitting. We tested several values of k and chose one that minimized error rate in the training set.

Results

Exploratory Analysis

We removed approximately 3% of our data due to data issues. The remaining data is balanced in the CVD outcome. All of our explanatory variables were found to be significantly associated with cardiovascular disease in our univariate analyses (Table 1). Older age, obesity (high BMI), high blood pressure, high cholesterol, high blood glucose, smoking, and drinking appear to be associated with higher risk of CVD. Physical activity is associated with lower risk. This is consistent with established knowledge and suggests we should consider all these features in our analytical models to predict CVD.

Additional exploratory analyses examined the correlations between variables (Figure 1). We found that the blood pressures were correlated, but height and weight were not, in part affirming our decisions to create the BMI and MAP variables. We used statistical plots to provide visual insights into the data. Figure 2 shows that people with CVD tend to have higher MAP, with the distribution being right skewed in both groups. BMI presents a similar pattern, though the distribution is flatter than that of MAP. Figure 3 shows that the relative proportions of people who have CVD when stratified and grouped by glucose and cholesterol levels is different across strata. This suggests a possible interaction of the two variables. Alcohol and smoking habits do not appear to interact. In boxplots of the continuous variables (Figure 3), we see that many of the continuous variables are right skewed with many outliers. This suggests that models may benefit from nonlinear modeling or data transformation.

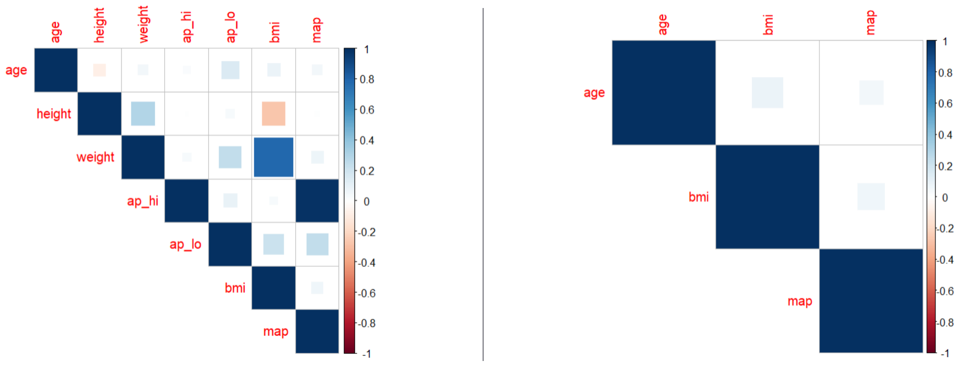

Figure 1. Correlation Plot between different variables without and with grouping

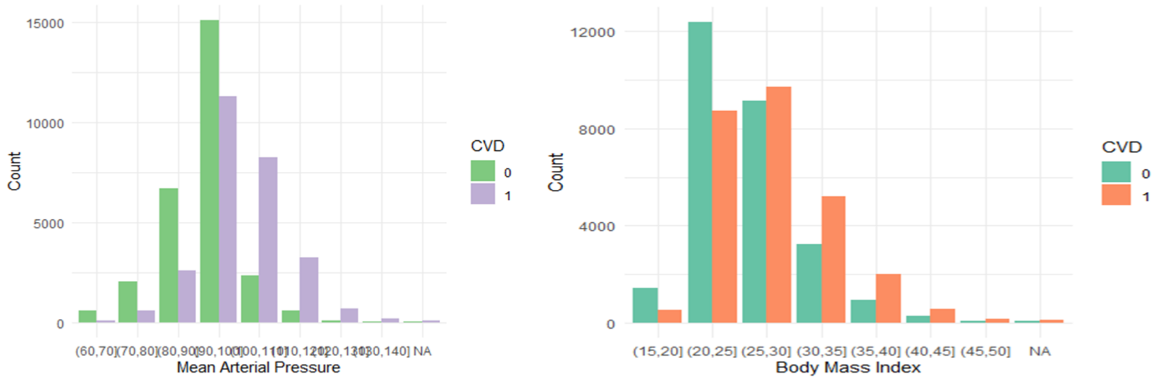

Figure 2: Distribution of MAP and BMI

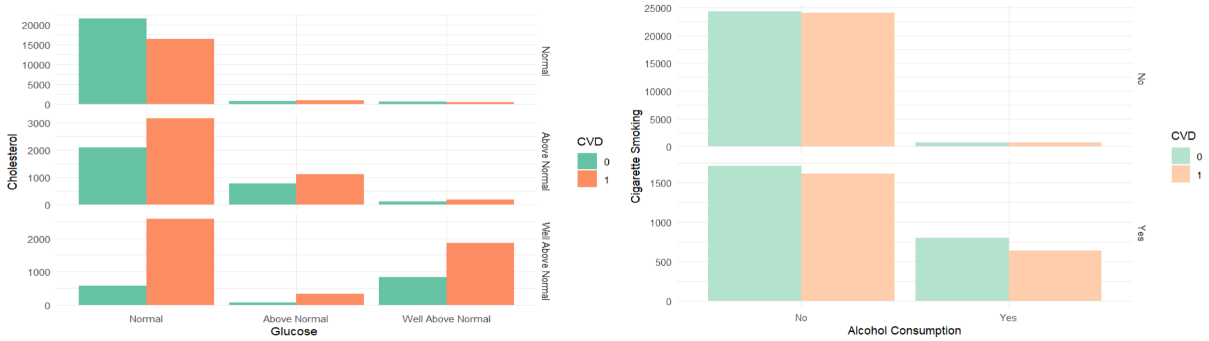

Figure 2: Glucose Vs Cholesterol Levels and Alcohol vs Smoking Status

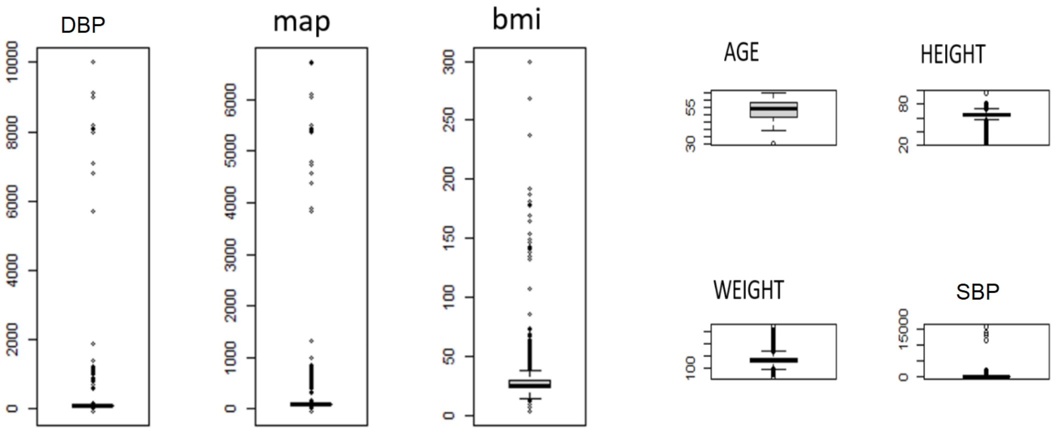

Fig 3: Box Plot for continuous columns in the dataset

Model Results

We compared the coefficients found by our three regression based models (Table 2). The coefficients suggest that increased age, cholesterol levels, physical activity, BMI, and MAP are associated. Interestingly, high blood glucose is not significantly different from normal blood glucose, and very high blood glucose is associated with decreased odds of CVD. Age also appears to have been constrained to zero in the LASSO model, although high blood glucose was not.

Table 2. Coefficients of Regression Models

Variable	**Logistic Regression**	**Ridge**	**LASSO**
Age	0.06	0.05	NA
High Cholesterol	0.41	0.39	0.44
Very High Cholesterol	1.12	1.18	1.21
High Blood Glucose	0.00	0.00	0.06
Very High Blood Glucose	-0.33	-0.21	-0.22
Physical Activity	-0.22	-0.20	-0.20
BMI	0.03	0.03	0.04
MAP	0.07	0.07	0.03

Note: Cholesterol and blood glucose variables use normal levels as a comparison group.

We then compared the predictive performance of our models (Table 3). All of our models performed moderately well with respect to each measure of predictive performance, indicating our models were able to achieve good predictions. Logistic regression had the lowest sensitivity while LASSO had the lowest specificity and accuracy. However, given that the variable sets used were essentially identical and led to very similar coefficients, it is not surprising that differences were minor. RF and SVM performed better overall, perhaps indicating there are non-linear relationships or interactions that need to be accounted for that were not adequately explained by the regression models. However, SVM had the lowest PPV, indicating a lower test predictive performance, though the sensitivity is high.

Table 3. Measures of Model Predictive Performance

Model	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
LR	0.720	0.68	0.670	0.705	0.738
Ridge	0.715	0.797	0.630	0.689	0.751
LASSO	0.696	0.787	0.603	0.671	0.734
RF	0.731	0.773	0.689	0.713	0.751
SVM	0.728	0.753	0.709	0.660	0.793
Bagging	0.6897	0.7305	0.648	0.678	0.702
GB	0.7298	0.758	0.701	0.720	0.740
KNN	0.7105	0.738	0.681	0.702	0.719

Discussion

In this report, we examine how various demographic, physical, clinical, and lifestyle factors may be associated with CVD, as well as examining whether machine learning methods are suitable for predicting CVD with the given data. For the most part, variables that are indicators of poor health, e.g. high BMI associated with obesity, are associated with increased risk of developing CVD. In the presence of strong predictors, it appears that some variables, such as gender and smoking habits were not retained in the linear models. Interestingly, increased levels of blood glucose were negatively associated with risk of CVD in our models. This is contrary to previously established findings (Borg et al., 2011). This may be because there are underlying confounders or interactions that we are not accounting for in our models or may be impossible to evaluate with our limited set of predictors.

Our algorithms for predicting CVD in our population performed very similarly. While the other algorithms may have benefits over logistic regression, increases in performance were minor. If the goal in creating these predictive models is for inference regarding risk factors for CVD or to create an interpretable algorithm for use in healthcare, it may be beneficial to rely on regression based methods instead of more advanced methods such as RF or SVM. In addition, limitations with our data may be restricting our ability to take full advantage of the benefits of using more complex methods. In the presence of a much broader dataset, such as electronic medical record data that includes a much larger number of different potential predictors for CVD, more advanced and non-linear methods may perform significantly better than regression methods for prediction.

The limited number of variables and simplicity of the information, e.g. simplifying cholesterol into a categorical variable, is a limitation to this study. This constrains our ability to deeply examine the variable relationships in our data or to investigate the functional form of how we include data in our regression models. This could be important towards checking the assumptions required for adequate regression modeling. In addition, we have a very large dataset relative to our number of possible predictors. Another factor to consider is that the distribution of the CVD to no CVD subjects in our data is approximately even. This is a much higher rate of CVD than is present in the general population. This could be benefiting the predictive ability of our models, but would likely result in our model fitting poorly to external data collected elsewhere.

Conclusion

Our results provide valuable insights into the relationships between various demographic, physical, clinical, and lifestyle factors and the risk of CVD. These findings could be useful for healthcare professionals in developing personalized prevention strategies and targeted interventions for at-risk individuals. Moreover, the study also contributes to the ongoing discussion on the suitability of machine learning methods in predicting CVD and offers a comparison between different algorithms, thus providing a foundation for future research in this area. As more comprehensive and diverse datasets become available, the performance of these methods may further improve. The incorporation of additional data sources, such as genetic information, imaging data, and environmental factors, could also lead to the development of more accurate and holistic predictive models.However, when met with limited datasets, it may be beneficial to utilize simple and interpretable regression models before attempting to use more advanced methods to create prediction models.

One potential application of our work is the development of decision support tools for clinicians, enabling them to make more informed decisions regarding patient care. By leveraging the predictive power of machine learning models, these tools can help identify individuals who may benefit from early intervention or closer monitoring, ultimately improving patient outcomes and reducing the burden of CVD on the healthcare system.

Our report supports findings that machine learning methods can be applied to health data to accurately predict chronic conditions such as CVD. Our analysis revealed several key findings, including a higher prevalence of cardiovascular disease in older individuals, as well as in those who have high blood pressure, and are overweight or obese.Our analysis also found that there were differences in the prevalence of CVD across different demographic groups. These findings can help to inform targeted prevention and intervention strategies that are tailored to the specific needs of different groups. Healthcare providers should encourage regular blood pressure screenings, particularly for individuals over the age of 50 or those who are overweight or obese. Blood pressure control can significantly reduce the risk of CVD, and identifying individuals who have high blood pressure but are not aware of it can help to ensure that they receive appropriate treatment. There is a need to develop targeted weight loss and exercise programs for individuals who are overweight or obese. Our analysis found that weight was a significant predictor of CVD, and interventions that promote weight loss and physical activity can help to reduce the risk of the disease.

References

Danaei, G., Ding, E. L., Mozaffarian, D., Taylor, B., Rehm, J., Murray, C. J. L., & Ezzati, M. (2009). The preventable causes of death in the United States: Comparative risk assessment of dietary, lifestyle, and metabolic risk factors. PLoS Medicine, 6(4), e1000058.

Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N., & Stokes, J. (1976). Factors of risk in the development of coronary heart disease—six-year follow-up experience. Annals of Internal Medicine, 55(1), 33-50.

Quesada, J. I. P., Lopez-Pineda, A., Gil-Guillén, V. F., Durazo-Arvizu, R., Orozco-Beltrán, D., Lopez-Domenech, A., & Carratalá-Munuera, C. (2019). Machine learning to predict cardiovascular risk. International Journal of Clinical Practice, 73(10). https://doi.org/10.1111/ijcp.13389

Patel, R. S., Ghasemzadeh, N., Eapen, D. J., Sher, S., Arshad, S., Ko, Y. A., ... & Quyyumi, A. A. (2018). Novel biomarker of oxidative stress is associated with risk of death in patients with coronary artery disease. Circulation, 131(4), 323-333.

Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., & Burton, P. R. (2012). Genetic epidemiology and public health: hope, hype, and future prospects. The Lancet, 360(9345), 1489-1498.

Stamler, J., Stamler, R., & Neaton, J. D. (1993). Blood pressure, systolic and diastolic, and cardiovascular risks. Archives of Internal Medicine, 153(5), 598-615.

World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., ... & INTERHEART Study Investigators. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. The Lancet, 364(9438), 937-952.

Ulianova, S. (2019). Cardiovascular Disease dataset. Kaggle.com. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Ortega, F. B., Lavie, C. J., & Blair, S. N. (2016). Obesity and Cardiovascular Disease. Circulation Research, 118(11), 1752–1770. https://doi.org/10.1161/circresaha.115.306883

Kengne, A. P., Czernichow, S., Huxley, R. R., Grobbee, D. E., Woodward, M., Neal, B., Zoungas, S., Cooper, M. E., Glasziou, P., Hamet, P., Harrap, S. B., Mancia, G., Poulter, N. R., Williams, B., & Chalmers, J. (2009). Blood Pressure Variables and Cardiovascular Risk. Hypertension, 54(2), 399–404. https://doi.org/10.1161/hypertensionaha.109.133041

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R. Springer.

Borg, R., Kuenen, J., Carstensen, B., Zheng, H., Nathan, D. G., Heine, R., Nerup, J., Borch-Johnsen, K., & Witte, D. R. (2011). https://doi.org/10.1007/s00125-010-1918-2