A Statistical Analysis of CVD using Binary Logistic Regression

Yue Yu*

College of Letters and Science, University of California, Santa Barbara

*Corresponding author: yu72@ucsb.edu

Keywords: non-communicable diseases, cardiovascular disease, logistic regression, binary logistic regression

Abstract: Background: As a chronic disease, cardiovascular disease (CVD) has greatly affected people's quality of life. Limited studies have investigated the relationship between many risk factors and infecting CVD, and few of them use a large sample size to conduct the study. We investigated the relationship between several different risk factors, such as blood pressure and cholesterol, and CVD. **Methods:** A dataset which contains 70000 samples is obtained from Kaggle, and binary logistic regression analysis was used to investigate the impact of clinical and lifestyle factors on the prevalence of CVD. Variable selection, such as forward selection and backward selection, was performed to build the final model. **Results:** The result of forward-backward selection and backward selection show that all features, except gender, are significant under 0.1% level, and the AIC of them are both 77216. The percentage of accuracy of the model is 0.7267. The proportion of actual positive samples identified correctly (recall/sensitivity) is 0.663, the proportion of predict positive samples identified correctly (precision) is 0.755, and the proportion of actual negative samples identified correctly (specificity) is 0.789. **Conclusion:** Our study suggested that age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol intake, and physical activity are the most important factor in determining CVD. The findings from our study have important public health implications and call for future studies to explore the potential mechanism of these findings. **CCS Concept**•Mathematics of computing → Probability and statistics → Statistical paradigms → Regression analysis

1. Introduction

The epidemiological shift in the 20th century was accompanied by a decline in deaths and disabilities caused by communicable diseases and an increase in noncommunicable diseases (NCDs). In NCDs, cardiovascular disease (CVD) is currently the main cause of global mortality and morbidity [1]. The burden of cardiovascular disease has been increasing for decades in almost all countries except high-income countries [2]. Due to its impact on quality of life and mortality, effective prediction measures are supportive of experts and doctors in early diagnosis to avoid developing heart failure. Over the past decade, numerous studies have investigated the relationship between many risk factors and CVD. Toshiaki Ohkuma et al. conducted a meta-analysis of individual participant data to examine the correlation between brachial-ankle pulse wave velocity (baPWV) and the risk of CVD [3]. In addition, Mary J. Roman et al. compared the ability of measuring intima-media thickness and atherosclerotic plaque alone in predicting cardiovascular disease [4]. Besides that, A.C. Carlsson et al. studied four predictors of cardiovascular disease: waist-hip ratio (WHR), waist circumference (WC), sagittal abdominal diameter (SAD), and waist-hip height ratio (whhr) in men and women [5]. Also, Omayma Alshaarawy et al. estimated the association between levels of polycyclic aromatic hydrocarbons (potent atmospheric pollutants) biomarkers and CVD [6], and Billy A. Caceres et al. investigated sexual orientation differences in cardiovascular disease risk and cardiovascular disease [7]. Finally, David H. Chae et al. examined associations between racial discrimination, mood disorders, and cardiovascular disease (CVD) among Black Americans [8]. All of these previous epidemiological studies have associated whether have CVD with several risk factors. However, they fail to consider more potential risk factors which are believed to be closely related to CVD, such as blood pressure, cholesterol, and blood sugar [1].

Admittedly, some studies have investigated the relationship between the risk factors mentioned above and getting CVD. Douglas G. Manuel MD et al. developed a predictive algorithm named Cardiovascular Disease Population Risk Tool to estimate the risk of getting cardiovascular disease [9], and Chipo Mutyambizi et al. investigates the prevalence and associated factors of diabetes and cardiovascular comorbidity in South Africa [10]. Both studies' predictors included body mass index, hypertension, diabetes, obesity, smoking, alcohol, and demographic risk factors. However, Douglas G. Manuel MD et al. pay too much attention to the prediction and ignore the importance of interpretations for the relationship between CVD and the risk factors. In addition, Chipo Mutyambizi et al. have used multinomial logistic regression models to analyze the relationship between diabetes - cardiovascular disease comorbidity and several predictors, while the relationship between diabetes and CVD, has not been analyzed.

In light of the limited studies investigating the relationship between more risk factors and CVD, and the lack of studies investigating the interpretation of the relationship between factors and CVD in prediction, we analyzed data from Kaggle to investigate the relationship between risk factors and CVD to find a prediction model that is both helpful and interpretable.

2. Method

2.1 Data Source

The dataset is obtained from Kaggle. Kaggle is an online community composed of data scientists and machine learning practitioners. It provides cutting-edge data science, which allows users to find and publish data sets, explore and build models in a web-based data science environment.

2.2 Data Description

The dataset contains 70000 samples, 11 features of the human body, and whether they have cardiovascular disease. Five of them are continuous variables and the rest six are category variables. The dataset is downloaded from Kaggle (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset). All of the dataset values were collected at the moment of medical examination. The dataset did not mention how it was collected.

There are 3 types of input features: objective feature, such as age and gender, meaning it is factual information; examination feature, such as cholesterol and glucose, meaning it is the result of medical examinations; and subjective feature, such as smoking and alcohol intake, meaning that the information is given by the patient.

2.3 Research variable

Cardio, a target variable, is a category variable, representing the presence or absence of cardiovascular disease. Cardio has two values, 0 and 1, where 0 means the patient doesn't have cardiovascular disease and 1 means the patient has cardiovascular disease.

Cholesterol, an examination feature, is a category variable, representing the content of cholesterol. Cholesterol has three values: 1, 2, and 3, where 1 means the content of cholesterol is at a normal level, 2 means the content of cholesterol is above the normal level, and 3 means the content of cholesterol is well above the normal level.

Physical activity, a subjective feature, is a category variable, representing whether the patient has physical activity. Physical activity has two values: 1 and 0, where 1 means the patient does have physical activity, while 0 means the patient doesn't have physical activity.

Gender, an objective feature, is a category variable, representing the gender of the patient. Gender has two values: 1 and 2, where 1 means female while 2 means male.

AP_HIGH, an examination feature, is a continuous variable, representing the systolic blood pressure of the patient. The range of AP_HIGH is 60-240 mmHg.

AP_LOW, and examination feature, is a continuous variable, representing the diastolic blood pressure of the patient. The range of AP_LOW is 40-190 mmHg.

2.4 Statistical method

Logistic regression (LR) is a multivariable method designed for dichotomous outcomes. It is especially suitable for models involving disease status (illness / Health) and decision-making (yes / no). Hence, it is widely used in health science research. In LR, the logarithm of the probability of obtaining a positive result (where "positive" is defined by the code of the result variable, i.e. Y = 1); a simple algebraic operation converts it into the probability of the result.

Binary logistic regression analysis, a non-linear regression technique, is applied to situations in which the response variable is dichotomous (0,1). The expected probability of a binary outcome is:

$$P(Y = 1) = \frac{1}{1+e^{\beta_1 X_1+\cdots+\beta_n X_n}} \tag{1}$$

where Xn are independent variables with numeric values (if they are binary, 0 usually represents control and 1 represents case) and the βn are the regression coefficients that quantify their contribution to the probability.

Univariate binary logistic regression analysis was used to determine the impact of clinical and lifestyle factors on the prevalence of CVD. Multiple logistic regression analysis was also used to establish the final model. The variable selection of multivariate logistic regression model adopts a bidirectional stepwise process, forward selection and backward selection, based on Akaike information criterion (AIC).

The dataset has 70000 samples. The original data set has been analyzed with Excel, removing missing values and correcting negative values. (The resulting sample is 68783). We first did exploratory data analysis. We calculated the mean and standard deviation of each variable, and figure out the range of each variable. Later, the chi-square test and t-test were used to find the correlation between each variable and dependent variable Y, Cardiovascular disease. To visualize X and Y, we make density plots for continuous variables and pie charts for categorical variables separately. To analyze the relation of X and Y, we plot boxplots for continuous variables and bar plots for categorical variables separately. And We fit a logical model to observe the correlation between X and Y. Finally, the confusion matrix is calculated to evaluate the performance of the data.

3. Result

Our CVD dataset was obtained from the Kaggle website. The dataset comprises 11 features of 70000 patients, divided into 5 numerical features and 6 nominal features. Features include age, height, weight, systolic blood pressure, diastolic blood pressure, gender, cholesterol, glucose, smoking, alcohol intake, and physical activity. The diagnostic class contains two values: 0 and 1, where 0 means no CVD and 1 means CVD exist.

Table 1 Distribution of features of the study population

Characteristics	N (68783)	Percent(%)
Gender		
Women (1)	44795	65.13
Men (2)	23988	34.87
Cholesterol		
Normal (1)	51582	74.99
above normal (2)	9315	13.54
well above normal (3)	7886	11.47
Glucose (gluc)		
Normal (1)	58474	85.01
above normal (2)	5074	7.38
well above normal (3)	5235	7.61
Smoking (smoke)		
Yes (1)	6053	8.8
No (0)	62730	91.2
Alcohol intake (alco)		
Yes (1)	3689	5.36
No (0)	65094	94.64
Physical activity (active)		
Yes (1)	55258	80.34
No (0)	13525	19.66
Cardiovascular disease (cardio)		
Presence (1)	34041	49.49
Absence (0)	34742	50.51
Continuous variable	Mean (SD)	Range
Age (years)	53.33 (6.77)	30-65
Height (cm)	164.4 (8.18)	55-250
Weight (kg)	74.12 (14.33)	11-200
Systolic blood pressure (ap hi)	126.6 (16.76)	60-240
Diastolic blood pressure (ap lo)	81.38 (9.68)	40-190

Table 1 shows the distributions of all characteristics of the study population. Most patients' cholesterol and glucose content are at the normal level. And more than 80% of patients do not drink or smoke, and do physical activities. The proportion of having CVD is approximately 50%.

Table 2 shows the results of baseline logistic regression. All features, except gender and height, are significant under 0.1% level, and the AIC of it is 77217. Table 3 shows the results of the logistic regression model with forward and backward selection. A risk factor, gender, has been dropped. Other than that, all features are significant under 0.1% level. The AIC of it is 77216. Table 4 shows the result of the logistic regression model with forward selection only. All features, except gender and height, are significant under 0.1% level, and the AIC of it is 77217. Table 5 shows the result of the logistic regression model with backward selection only. Gender also has been dropped and the AIC of the model is 77216.

Table 6 calculates the confusion matrix for the bestlam model to evaluate the accuracy of the modeling. The percent correct is 0.7267. Table 7 calculates some evaluation criteria.

Table 2 The results of baseline logistic regression model, where Pr(>|z|) stands for p-value (AIC: 77217)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.13E+01	2.31E-01	-48.884	< 2e-16
AGE	5.13E-02	1.35E-03	37.965	< 2e-16
GENDER	-1.82E-02	2.19E-02	-0.832	0.4053
HEIGHT	-3.65E-03	1.28E-03	-2.848	0.0044
WEIGHT	1.09E-02	6.87E-04	15.847	< 2e-16
AP_HIGH	5.32E-02	8.71E-04	61.012	< 2e-16
AP_LOW	1.67E-02	1.36E-03	12.296	< 2e-16
CHOLESTEROL	4.99E-01	1.56E-02	32.025	< 2e-16
GLUCOSE	-1.19E-01	1.77E-02	-6.737	1.62E-11
SMOKE	-1.43E-01	3.48E-02	-4.123	3.75E-05
ALCOHOL	-2.10E-01	4.22E-02	-4.959	7.08E-07
PHYSICAL_ACTIVITY	-2.27E-01	2.19E-02	-10.388	< 2e-16

Table 3 The result of logistic regression model with forward and backward selection (AIC: 77216)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.13E+01	2.21E-01	-51.002	< 2e-16
AGE	5.13E-02	1.35E-03	37.958	< 2e-16
HEIGHT	-4.12E-03	1.15E-03	-3.581	0.000342
WEIGHT	1.09E-02	6.87E-04	15.846	< 2e-16
AP_HIGH	5.31E-02	8.71E-04	61.011	< 2e-16
AP_LOW	1.67E-02	1.36E-03	12.284	< 2e-16
CHOLESTEROL	5.00E-01	1.56E-02	32.057	< 2e-16
GLUCOSE	-1.19E-01	1.77E-02	-6.732	1.67E-11
SMOKE	-1.51E-01	3.36E-02	-4.484	7.33E-06
ALCOHOL	-2.11E-01	4.22E-02	-5.007	5.52E-07
PHYSICAL_ACTIVITY	-2.27E-01	2.19E-02	-10.389	< 2e-16

Table 4 The result of logistic regression model with forward selection only (AIC: 77217)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.13E+01	2.31E-01	-48.884	< 2e-16
AGE	5.13E-02	1.35E-03	37.965	< 2e-16
GENDER	-1.82E-02	2.19E-02	-0.832	0.4053
HEIGHT	-3.65E-03	1.28E-03	-2.848	0.0044
WEIGHT	1.09E-02	6.87E-04	15.847	< 2e-16
AP_HIGH	5.32E-02	8.71E-04	61.012	< 2e-16
AP_LOW	1.67E-02	1.36E-03	12.296	< 2e-16
CHOLESTEROL	4.99E-01	1.56E-02	32.025	< 2e-16
GLUCOSE	-1.19E-01	1.77E-02	-6.737	1.62E-11
SMOKE	-1.43E-01	3.48E-02	-4.123	3.75E-05
ALCOHOL	-2.10E-01	4.22E-02	-4.959	7.08E-07
PHYSICAL_ACTIVITY	-2.27E-01	2.19E-02	-10.388	< 2e-16

Table 5 The result of logistic regression model with backward selection only (AIC: 77216)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.13E+01	2.21E-01	-51.002	< 2e-16
AGE	5.13E-02	1.35E-03	37.958	< 2e-16
HEIGHT	-4.12E-03	1.15E-03	-3.581	0.000342
WEIGHT	1.09E-02	6.87E-04	15.846	< 2e-16
AP_HIGH	5.31E-02	8.71E-04	61.011	< 2e-16
AP_LOW	1.67E-02	1.36E-03	12.284	< 2e-16
CHOLESTEROL	5.00E-01	1.56E-02	32.057	< 2e-16
GLUCOSE	-1.19E-01	1.77E-02	-6.732	1.67E-11
SMOKE	-1.51E-01	3.36E-02	-4.484	7.33E-06
ALCOHOL	-2.11E-01	4.22E-02	-5.007	5.52E-07
PHYSICAL_ACTIVITY	-2.27E-01	2.19E-02	-10.389	< 2e-16

Table 6 The confusion matrix (Percent correct: 0.7267)

	TRUE		
Predicted	0	1	Total
0	13701	5740	19441
1	3659	11292	14951
Total	17360	17032	34392

Table 7 Evaluation of the model

accuracy	0.7267
recall_sensitivity	0.663
precision	0.7553
specificity	0.7892
E/(1+E)	0.6991761

4. Discussion

This study applied logistic regression to construct a model which can predict whether a patient has CVD. In the baseline model and the forward selection model, the results show that age, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol intake, and physical activity are the most important factor in determining CVD since the p-value is much less than 0.05. However, the result of backward selection and forward-backward selection suggest that all the risk factors that we investigated, except for gender, have a significant association with CVD.

Logistic regression was employed to make a statistical analysis of the relationship between age, gender, weight, height, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity, and cardiovascular disease. It can be found that, if all other predictor variables are held constant, the odds of getting CVD occurring increased by 4.88% (95% CI [1.049, 1.055]) for a one-unit increase in age. It was found that holding all other independent variables constant, the odds of getting CVD occurring decreased by 0.4% (95% CI [1.002, 1.006]) for a one-unit increase in height. It was found that holding all other independent variables constant, the odds of getting CVD occurring increased by 1.07% (95% CI [1.01, 1.012]) for a one-unit increase in weight. It was found that holding all other independent variables constant, the odds of getting CVD occurring increased by 5.04% (95% CI [1.053, 1.056]) for a one-unit increase in systolic blood pressure. It was found that holding all other independent variables constant, the odds of getting CVD occurring increased by 1.6% (95% CI [1.014, 1.02]) for a one-unit increase in diastolic blood pressure. It was found that holding all other independent variables constant, the odds of getting CVD occurring increased by 33.3% (95% CI [1.618, 1.679]) for a one-unit increase in cholesterol. It was found that holding all other independent variables constant, the odds of getting CVD occurring decreased by

10.6% (95% CI [1.092, 1.161]) for a one-unit increase in glucose. It was found that holding all other independent variables constant, the odds of getting CVD occurring decreased by 13% (95% CI [1.097, 1.229]) for a one-unit increase in smoke. It was found that holding all other independent variables constant, the odds of getting CVD occurring decreased by 17.4% (95% CI [1.152, 1.318]) for a one-unit increase in alcohol. It was found that holding all other independent variables constant, the odds of getting CVD occurring or decreased by 18.5% (95% CI [1.212, 1.298]) for a one-unit increase in physical activity.

After using the cv method to select lambda, the confusion matrix for the bestlam model has been calculated. The percentage of accuracy is 0.727, which is acceptable. Other than that, more evaluation criteria have been calculated. The proportion of correctly identify samples (accuracy) is 0.727, the proportion of actual positive samples identified correctly (recall/sensitivity) is 0.663, the proportion of predict positive samples identified correctly (precision) is 0.755, The proportion of actual negative samples identified correctly (specificity) is 0.789. In addition, we predict the probability of a new sample being positive. And according to this model, the probability of this sample being 1 is about 0.699, which means that the probability of a patient having CVD is 0.699.

Ingrida Grabauskyte et al. had compared the decision tree induction with binary logistic regression for the prediction of the risk of cardiovascular disease. And the result shows that systolic blood pressure, age, weight, and cholesterol et al. are associated with getting CVD [11], which is highly consistent with our study. In addition, Grabauskyte also found that alcohol intake was associated with a lower probability of getting CVD, which is unexpected but also consistent with our study. A possible explanation is that healthy people tend to drink more alcohol than people that have illnesses.

There are some limitations of this study that need to be noticed. First, the subject feature, such as alcohol intake and physical activities, was collected through self-report, which may cause recall bias. However, the subject feature only focuses on whether the patient has similar behavior. When the question involves the specific frequency of alcohol intake or the amount of alcohol intake in liters, these kinds of questions will have a huge recall bias. Our dataset only concerns whether the patients have consumed alcohol, which means that the deviation caused by recall bias will be much smaller. In addition, the lack of specific definitions of many features greatly reduces the practicability of prediction. Through our study, we can know that physical activity is helpful to prevent getting cardiovascular disease. However, due to the lack of specific definitions of physical activities, we cannot obtain enough information to give patients suggestions on appropriate exercise volume in order to achieve the purpose of preventing getting CVD. Therefore, future studies should have a detailed definition of human features, such as alcohol intake and physical exercise, in order to better understand the relationship between them and getting CVD.

5. Conclusion

Logistic regression has been used to analyze the relationship between multiple risk factors, such as blood pressure and cholesterol, and getting cardiovascular disease. The result has shown that age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity have a high correlation with getting CVD. The implication is that doctors can infer whether the patient will suffer from CVD through various indicators of physical examination, so as to give suggestions and reduce the probability of getting CVD. Future work can investigate the relationship between the specific content of cholesterol, blood glucose, alcohol intake, and the probability of getting CVD. In addition, the relationship between the frequency of smoking and physical activity and cardiovascular disease is also worth studying.

References

[1] Joseph P, Leong D, McKee M, Anand SS, Schwalm JD, Teo K, Mente A, and Yusuf S. Reducing the Global Burden of Cardiovascular Disease, Part 1 The Epidemiology and Risk Factors. Circulation Research. 2017;121:677–694. https://doi.org/10.1161/CIRCRESAHA.117.308903.

[2] Roth GA, Mensah GA et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study, Journal of the American College of Cardiology, Volume 76, Issue 25, 2020, Pages 2982-3021, ISSN 0735-1097, https://doi.org/10.1016/j.jacc.2020.11.010.

[3] Ohkuma T, Ninomiya T et al. Brachial-Ankle Pulse Wave Velocity and the Risk Prediction of Cardiovascular Disease An Individual Participant Data Meta-Analysis. Hypertension. 2017;69:1045–1052. https://doi.org/10.1161/HYPERTENSIONAHA.117.09097.

[4] Roman MJ, Kizer JR, Best LG, Lee ET, Howard BV, Shara NM, and Devereux RB. Hypertension. 2012;59:29–35 https://doi.org/10.1161/HYPERTENSIONAHA.111.181925.

[5] Carlsson AC, Riserus U, Ärnlöv J, Borné Y, Leander K, Gigante B, Hellénius ML, Bottai M, Faire Ud. Prediction of cardiovascular disease by abdominal obesity measures is dependent on body weight and sex – Results from two community based cohort studies. Nutrition, Metabolism and Cardiovascular Diseases, Volume 24, Issue 8, 2014, Pages 891-899, ISSN 0939-4753. https://doi.org/10.1016/j.numecd.2014.02.001.

[6] Alshaarawy O, Elbaz HA, Andrew ME. The association of urinary polycyclic aromatic hydrocarbon biomarkers and cardiovascular disease in the US population. Environment International, Volumes 89–90, 2016, Pages 174-178, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2016.02.006.

[7] Caceres BA, Makarem N, Hickey KT, & Hughes TL (2019). Cardiovascular Disease Disparities in Sexual Minority Adults: An Examination of the Behavioral Risk Factor Surveillance System (2014-2016). American Journal of Health Promotion, 33(4), 576–585. https://doi.org/10.1177/0890117118810246.

[8] Chae DH, Nuru-Jeter AM, Lincoln KD, Arriola KRJ. Racial Discrimination, Mood Disorders, and Cardiovascular Disease Among Black Americans. Annals of Epidemiology, Volume 22, Issue 2, 2012, Pages 104-111, ISSN 1047-2797, https://doi.org/10.1016/j.annepidem.2011.10.009.

[9] Manuel, DG, Tuna M, Bennett C, Hennessy D, Rosella L, Sanmartin C, Tu JV, Perez R, Fisher S, & Taljaard M (2018). Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the cardiovascular Disease Population Risk Tool (cVDPoRT). CMAJ: Canadian Medical Association Journal, 190(29), E871+. https://link.gale.com/apps/doc/A547075813/CWI?u=ucsantabarbara&sid=bookmark-CWI&xid=5211286f.

[10] Mutyambizi C, Chola L, Groot W et al. The extent and determinants of diabetes and cardiovascular disease comorbidity in South Africa – results from the South African National Health and Nutrition Examination Survey (SANHANES-1). BMC Public Health 17, 745 (2017). https://doi.org/10.1186/s12889-017-4792-8.

[11] GRABAUSKYTĖ I, TAMOŠIŪNAS A, KAVALIAUSKAS M, RADIŠAUSKAS R, BERNOTIENĖ G, & JANILIONIS V (2018). A Comparison of Decision Tree Induction with Binary Logistic Regression for the Prediction of the Risk of Cardiovascular Diseases in Adult Men. Informatica, 29(4), 675–692. https://doi-org.proxy.library.ucsb.edu:9443/10.15388/Informatica.2018.187.