

Research on Risk Prediction for Cardiovascular Diseases

Fanqi Gao

Shenyang No.2 High School, Shenyang, 110000, China

y101577@outlook.com

Abstract. Cardiovascular disease is a common disease that seriously threatens human health, and occupies the first place in the death caused by various diseases. Despite the fact that studies have documented many factors that lead to cardiovascular disease in medical development, there are still many unknown factors that have not been explored. To provide a comprehensive summary and explore various indicators for certain patients with cardiovascular diseases, a binary logistic regression model was used in this study. It has been concluded that age, Trest BPS, Cholestrol, FBS, Rest ECG, and slope have limited impact on the indicators, but gender, Exang, old peak, ca, and thal have some negative effects on them. It can be observed from the table above that the research model has an overall prediction accuracy of 85.15%, and its fit is acceptable. The precise and reliable materials given by these data can be utilized for future medical research, specifically in cardiovascular disease research, and they also give direction for subsequent research.

Keywords: Medical development, cardiovascular disease, binary logistic regression model.

1. Introduction

Cardiovascular disease is a common disease that seriously threatens human health, and occupies the first place in the death caused by various diseases [1]. Premature death is still primarily caused by it in Europe, even though CVD is also a factor (cardipvascular), Over the past few decades, mortality rates have significantly declined in several European countries. Over 80% of all CVD mortality is believed to occur in developing countries [2]. The estimate of disability-adjusted life years (DALYs) is projected to increase from 85 million DALYs in 1990 to approximately 150 million DALYs globally by 2020, which will remain the most common cause of productivity loss [3]. As a result, personalized risk assessment has become one of the long-term goals of cardiovascular medicine, and the European Society of Cardiology has particularly highlighted the significance of models that predict challenges of advancing healthcare and preventing a population-wide problem. This article aims to study the prediction of cardiovascular disease risk.

There are many factors which can lead to the cardiovascular disease and what most noticeable is people's lifestyles. According to the World Health Organization, lifestyle is highly linked to CVD, particularly tobacco use, unhealthy diet habits, physical inactivity, and psychosocial stress. Over three-quarters of all CVD mortality can be avoided with sufficient lifestyle changes [4]. Preventing CVDs and their related disability is an important challenge faced by everyday individuals, politicians, and especially healthcare workers. CVD prevention is a strategy that aims to eliminate or eliminate the impact of CVDs and their related disability, utilizing a coordinated set of actions at both the public and

individual levels. Prevention's foundation is established in cardiovascular epidemiology and evidence-based medicine [5]. This paper will study people's life factors and other factors that lead to cardiovascular disease, such as family health history, social environment and psychological factors.

The UK Biobank CVD Risk Prediction (UKCRP) was trained and validated by researchers from Fudan University using machine learning methods, with comprehensive phenotypic and follow-up information from a large perspective UK biobank cohort in a similar direction. They set out to generate a system that could predict the risk of myocardial infarction (MI) for the next decade, ischemic stroke (IS), and hemorrhagic stroke (HS) at the individual level. UKCRP's performance was assessed against established CVD risk scores to examine whether the proposed model is superior [6]. Machine learning (ML)-based methods have exceptional advantages in processing large amounts of time and event data with multidimensional space, which make them an outstanding alternative to traditional prediction methods. Nonetheless, they frequently have restricted follow-up time, include excessive covariates, or utilize pre-determined variable domains or participants. The need to build new cardiovascular predictions is highlighted by these limitations, which will enhance the ability to predict and stratify risks in real-world clinical practice. Furthermore, Grading of Recommendations Assessment, Development, and Evaluation (GRADE) was introduced by the JTF because assessing the quality of evidence is achieved through a process that is both transparent and rigorous, with an emphasis on is there a possibility that further research would or would not affect confidence in the estimate of intervention effects or diagnostic accuracy [7]. Quality indicators that pertain to studies are limited, inconsistent findings, indirect evidence, and imprecision.

In the judgement of the guideline group, these are applied to every outcome that has a critical importance for decision-making. Based on these indicators, judgments are made to assess the quality of evidence from high, to moderate, low, and very low. This assessment takes into account the quality of evidence matters when it comes to critical outcomes, not those that are not crucial for decision-making. By adopting this new approach, it is possible to identify systematic reviews or randomized control trials (RCTs) that have bias, inconsistent, or imprecise, which can result in a decrease in the quality of evidence from high to moderate or low-quality. Likewise, enhancing observational data from cohort or case-control studies is possible from moderate to low. To understand the changes that have taken place in CVD mortality and incidence rates. it's possible to use models like the IMPACT mortality model [8]. The predicted correlation between age and gender on CHD mortality is estimated by using information about modifications that have an effect on coronary risk factors and treatment obtained from RCTs regarding the effectiveness of different treatment modes. In various countries, this model has been utilized. It has been discovered that reducing major risk factors, particularly smoking, BP, and cholesterol, was found to be beneficial, accounting for more than half of the decrease in CHD deaths, despite an increase in obesity and type 2 diabetes [9]. The National Institute for Health and Clinical Excellence (NICE) recently stated that the human and economic arguments in favor of CVD prevention are highly favorable, and several committees from other countries agree with them [10].

In summary, after consideration and optimization, this paper will use UK biobank CVD risk prediction model (UKCRP) which is made by ML method to have further research on the risk prediction of cardiovascular disease.

2. Methods

2.1. Data Source

The data for this literature is collected from the Kaggle website in 2021. The factors which can affect the probability of occur of the cardiovascular diseases include age, gender, CP levels, Trest BPS levels, Cholestrol levels, FBS levels, Resting ECG levels, Thalach levels, Exang levels and old peak history.

2.2. Variable Selection

The data contains 10 variables (age, gender, CP levels, Trest BPS levels, Cholestrol levels, FBS levels, Resting ECG levels, Thalach levels, Exang levels and Oled Peak History Recorded). All of these factors

can influence the occur rate of cardiovascular disease. Table 1 shows the number distribution of patients with cardiovascular disease with different values in different types of indicators.

Table 1. Logogram and range of the 10 factors.

Elements	Logogram	Range
occurrence rate of CVD	Y	-
age	X1	33-77
gender	X2	-
CP levels	X3	0.0-3.00
Trest BPS levels	X4	94.00-20-.00
Cholestrol levels	X5	126.00-432.60
FBS levels	X6	0.00-1
Resting ECG levels	X7	0.00-1.20
Thalach levels	X8	84.10-202.0-
Exang levels	X9	0.00-1.00
oled peak history	X10	0.00-4.34

2.3. Research Protocol

This research will use logistic regression to model the independent variable's statistical relationship with dependent variables. The relationship between X and Y will be given by SPSSAU. Binary Logistic regression is a statistical technique that examines the association between binary dependent variables (values 0 and 1) and multiple independent variables. When applying logistic regression, te author usually focuses on the probability of the prediction. If the probability is above a certain threshold (usually 0.5), the predicted observation belongs to the positive class. Otherwise, it belongs to the negative class. This approach allows logistic regression not only to classify, but also to quantify the uncertainty of predictions, providing more information for research.

3. Results and Discussion

3.1. Model Results

As can be seen from table 2, no level ge, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, that is the independent variable, target is used as the dependent variable used in binary Logit regression analysis. From the table above, it is clear that This means that age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, that can explain the 0.49 change in target. It can be seen from the above table that the formula of the model is: $\ln\left(\frac{p}{1-p}\right) = 3.450 - 0.005 * age + \dots + 0.579 * slope - 0.773 * ca - 0.900 * thal$.

The probability of target being 1 is represented by p, while the probability of target being 0 is represented by 1-p. The final concrete analysis shows that: CP, Thalach will have a significant positive impact on target, and gender, Exang, old peak, ca, that will have a significant negative impact on target. However, age, Trest BPS, Cholestrol, FBS, Rest ECG, slope do not affect target.

Table 2. Binary Logit regression prediction coefficients.

elements	coefficient	SE	Z	Wald χ^2	p	OR
age	-0.005	0.023	-0.212	0.045	0.832	0.995
gender	-1.758	0.469	-3.751	14.067	0.000	0.172
CP	0.860	0.185	4.638	21.510	0.000	2.363
trest BPS	-0.019	0.010	-1.884	3.549	0.060	0.981
Cholestrol	-0.005	0.004	-1.224	1.499	0.221	0.995
FBS	0.035	0.529	0.066	0.004	0.947	1.036

Table 2. (continued).

Rest ECG	0.466	0.348	1.339	1.793	0.181	1.594
Thalach	0.023	0.010	2.219	4.924	0.026	1.023
Exang	-0.980	0.410	-2.391	5.719	0.017	0.375
old peak	-0.540	0.214	-2.526	6.383	0.012	0.583
slope	0.579	0.350	1.656	2.742	0.098	1.785
ca	-0.773	0.191	-4.051	16.414	0.000	0.461
thal	-0.900	0.290	-3.104	9.634	0.002	0.406
intercept	3.450	2.571	1.342	1.800	0.180	31.51

3.2. Model Test

The model's fit quality can be measured by its prediction accuracy (Table 3). It can be observed from the table above that the research model has an overall prediction accuracy of 85.15%, and its fit is acceptable. 76.81% is the prediction accuracy when the true value is 0, which is a good result. Additionally, when the true value is 1, the prediction accuracy is 92.22%.

Table 3. Binary Logit regression prediction accuracy summary.

	Predicted value			Accuracy	Error rate
		0	1		
True value	0	106	32	76.81%	23.19%
	1	13	152	92.12%	7.88%
Summarizing				85.15%	14.85%

4. Conclusion

The study selected a variety of data to focus on factors that may be involved in the development of cardiovascular disease. In conclusion, it is undeniable that due to the limited amount of data in the sample, the model used in this paper may have some errors except for factors. In addition, the sample does not cover all ages and races, which may also lead to some deviations in the results, thus affecting the accuracy of the results. Still, the study has a lot of value and merit. Firstly, the method in this paper adopts the graphical method to show and analyze the relationship between the occurrence probability of cardiovascular disease and various influencing factors more intuitively. This allows the experiment to be visualized, making the results clearer and more intuitive. In addition, this experiment did not have a single sample like many previous experiments, but chose multiple linear regression to make it more comprehensive. Furthermore, it has a certain positive effect on the timely intervention and treatment of cardiovascular diseases. Apart from CP.

In addition to the known factors related to cardiovascular disease such as level and FBS level, there are more factors that may be related to cardiovascular disease worthy of attention and consideration, such as living habits and family history of disease. Whether these factors are related to the emergence of cardiovascular disease requires further medical investigation, which means that these findings will point the way for further related research in the future. The discovery of new causative factors in addition to those already identified could help people detect and treat cardiovascular disease earlier and with less delay. Despite the shortcomings, enough evidence exists to back up increased public health and individual prevention efforts. There is still a lot of room for improvement in managing risk factors, even in individuals who are at high risk. It is important to use more precise models to assess risk factors to improve patient survival and quality of life.

References

- [1] Hu, J. and Yang, S.L. (2020) Research status of evaluation of cardiac diastolic function by ultrasonography. *Journal of Molecular Imaging*, 43(4), 593-596.
- [2] Perk, J., De Backer, G., Gohlke, H. et al. (2012) European Guidelines on Cardiovascular Disease Prevention in Clinical Practice. *Int.J. Behav. Med.* 19, 403-488).
- [3] World Health Organization. (2002) Joint WHO/FAO Expert Consultation on Diet, Nutrition and the Prevention of Chronic Diseases. World Health Organization Report, 916.
- [4] Last, J.M. (2001) A dictionary of epidemiology. 4th ed. New York: Oxford University Press.
- [5] Guyatt, G.H., Oxman, A.D., Kunz, R., Vist, G.E., Falck-Ytter, Y. and Schunemann, H.J. (2008) What is 'quality of evidence' and why is it important to clinicians? *BMJ*, 336, 995–998.
- [6] Ford, E.S., Ajani, U.A., Croft, J.B., et al. (2007) Explaining the decrease in U.S. deaths from coronary disease, 1980–2000. *N Engl J Med*, 356, 2388-2398.
- [7] Capewell, S. and O'Flaherty, M. (2011) Rapid mortality falls after risk-factor changes in populations. *Lancet*, 378, 752-753.
- [8] National Institute for Health and Clinical Excellence. (2010) Prevention of Cardiovascular Disease: Costing Report. Nice Public Health Guidance, 25. <http://www.nice.org.uk/nicemedia/live/13024/49325/49325.pdf>
- [9] Kotseva, K., Wood, D., De Backer, G., De Bacquer, D., Pyorala, K, . and Keil, U. (2009) EUROASPIRE III: a survey on the lifestyle, risk factors and use of cardioprotective drug therapies in coronary patients from 22 European countries. *Eur J Cardiovasc Prev Rehabil*, 16, 121-137.
- [10] Cuenca-García, M., et al. (2012) Combined influence of healthy diet and active lifestyle on cardiovascular disease risk factors in adolescents. *Scandinavian Journal of Medicine & Science in Sports*, 3, 553-562.