The Research of Influence Factors that lead to Heart Disease

David Xia*

School of Logistics and Electronic Commerce, Zhejiang Wanli University, Ningbo, 315101, China

*Corresponding author: 2020013626@zwu.edu.cn

Abstract. Heart disease is one of the common diseases, and it is the disease with the highest mortality. Although there are many studies in the field of heart disease, including genetic aspects, environmental aspects and so on, most of them are single-factor studies. Therefore, this paper will establish a Probit regression model, based on 11 factors that may induce heart disease, and make a statistical analysis on the health status of 400,000 people collected in 2020. It is concluded that smoking, stroke, different walking, gender, age category, diabetes, Asthma, Kidney disease and skin cancer will have a significant positive impact on Heart Disease, and Physical Activity will have a significant negative impact on Heart Disease. However, Alcohol Drinking does not affect the Heart Disease. The results of this paper can provide a more comprehensive factor analysis direction for preventing heart disease, to do a good job of prevention, reduce the prevalence of diseases and reduce mortality.

Keywords: Heart disease; multiple factors; probit regression model.

1. Introduction

Heart disease has always been a well-known disease, but in the global health assessment report released by the World Health Organization in 2019, the number of people who died of heart disease still ranks first in the world [1]. At the same time, China Cardiovascular Health and Disease Report 2022 also pointed out that the number of people suffering from heart disease is on the rise [2]. Two out of every five patients died of heart disease [2]. Cao et al. also pointed out that since 1990, the incidence and mortality of chronic diseases have gradually increased [3]. Therefore, it is of great significance to study the influencing factors that may lead to heart disease for people to prevent heart disease and maintain healthy physical condition. The purpose of this paper is to analyze the potential impact of heart disease by studying different factors, to help people evaluate their own health index and take preventive measures in time.

Heart disease is divided into congenital heart disease and acquired heart disease. Congenital heart disease is usually caused by environmental factors and genetic factors [4]. In the study of congenital heart disease, Yuan et al. have monitored the incidence of congenital heart disease in 10665 children and analyzed the environmental risk factors [5]. Qiu et al. also based on MTHFR gene, CBS gene and environmental factors, made a comparative analysis of cases of congenital heart disease [6]. However, there is a lack of specific statistical research and analysis on acquired heart disease. Therefore, this paper will focus on the study of acquired heart disease, and at the same time, based on 11 variables (gender, age, smoking, drinking, stroke, different walking styles, diabetes, physical activity, asthma, kidney disease and skin cancer), select a suitable model to analyze and study its influence on the induction of acquired heart disease.

On the research of acquired heart disease, Peking University and Oxford University launched a research based on the health hazards of drinking and found that alcohol can cause health hazards [7]. Du based on Bayesian network model to predict and analyze chronic diseases [8]. Chen et al. also carried out data mining and factor analysis on patients with heart disease [9]. Shi obtained the relationship between diabetes and heart disease through research and analysis [10]. Zhao et al also made a clinical analysis on the relationship between diabetes and heart disease [11]. Zhao concluded that smoking can lead to heart disease through experiments [12]. Based on the multi-level negative binomial regression model, Li analyzed the changing trend of heart disease and mortality level of people over 40 years old [13]. Li also found a new method to predict heart disease, thus reducing the

risk of heart disease [14]. However, the above studies are all single variables, and the influence of many factors on inducing heart disease is not considered.

To sum up, this paper will study and analyze the possible influence of these factors on inducing heart disease based on 11 factors and establishing Probit regression model.

2. Methods

2.1. Data Sources

The data of this document comes from Kaggle website, which is a platform for data modeling and data analysis. Kamil Pytlak collected about 400,000 people's health status about heart disease in 2020 and made statistics and published them.

2.2. Variable Selection

The total number of data involved in this document is 319,795, including 151,990 males and 167,805 females, all of whom are adults over 18 years old. The data contains 11 variables. (Gender, Age, Smoking, Alcohol Drinking, Stroke, Diff Walking, Diabetic, Physical Activity, Asthma, Kidney Disease and Skin Cancer).

Table 1. Logogram and numbers of the 11 factors

Elements	Logogram	*Number1	**Cancer1
Gender	x_1	319795	27373
Age	x_2	319795	27373
Smoking	x_3	131908	16037
Alcohol Drinking	x_4	21777	1141
Stroke	x_5	12069	4389
Diff Walking	x_6	44410	10028
Diabetic	x_7	43361	9065
Physical Activity	x_8	71838	9884
Asthma	x_9	42872	4933
Kidney Disease	x_{10}	11779	3455
Skin Cancer	x_{11}	29819	4980

*Number 1: The number of people suffering from the disease.
**Cancer1: The number of heart disease patients suffering from the disease.

Table 1 shows the number of people suffering from the disease and the number of people suffering from heart disease. The labels of various factors are shown in Table 1. The data sample is 319,795 people, including 27,373 people with heart disease.

Table 2. Gender distribution by age group

Age	[18-29]	[30-39]	[40-49]	[50-59]	[60-69]	[70-79]	[80-90]
*Number2	38019	39303	42797	55139	67837	52547	24153
Female	17390	19588	22390	28822	35620	29279	14716
Male	20629	19715	20407	26317	32217	23268	9137
**Cancer2	263	522	1230	3585	7428	8896	5449

*Number 2: The number of people in each age group.
**Cancer 2: The number of people in each age group who have heart disease.

Table 2 shows the number of women, men and people with heart disease in each age group. There are enough people in each age group to serve as samples, which is convenient for subsequent modeling, calculation and analysis.

2.3. Research Protocol

In this paper, Probit regression model will be adopted, with whether you have heart disease as the dependent variable (Y) and 11 influencing factors as the independent variable (X), where 0 means no and 1 means yes. Then use SPSSAU and SPSSPRO to analyze the influence of X on Y and analyze the relationship between the influence of 11 factors on heart disease.

2.4. Model Principle

Probit regression model used in this paper is often used for statistical analysis of data with dependent variables as classified variables, and its formula is:

$$P = \frac{e^{X\beta}}{1+e^{X\beta}} \quad (1)$$

2.5. Model Testing

Judging the quality of model fitting by the accuracy of model prediction, it can be seen from Table 3 that the overall prediction accuracy of the research model is 91.43%, and the model fitting is acceptable. When the true value is 0, the prediction accuracy is 99.30%; In addition, when the true value is 1, the prediction accuracy is 8.87%.

Table 3. Binary probit regression prediction accuracy summary

		predicted value		forecast accuracy	Prediction error rate
		0	1		
true value	0	45327	320	99.30%	0.70%
	1	3966	386	8.87%	91.13%
Summary				91.43%	8.57%

3. Results and Discussion

Figure 1 shows the proportion of people suffering from heart disease among a series of factors. The factors leading to heart disease are Smoking, Alcohol Drinking, Stroke, Diff Walking, Diabetic, Physical Activity, Asthma, Kidney Disease and Skin Cancer. There are 319795 samples, according to the proportion of people suffering from heart disease in different factors, the probability of heart disease caused by each factor can be determined.

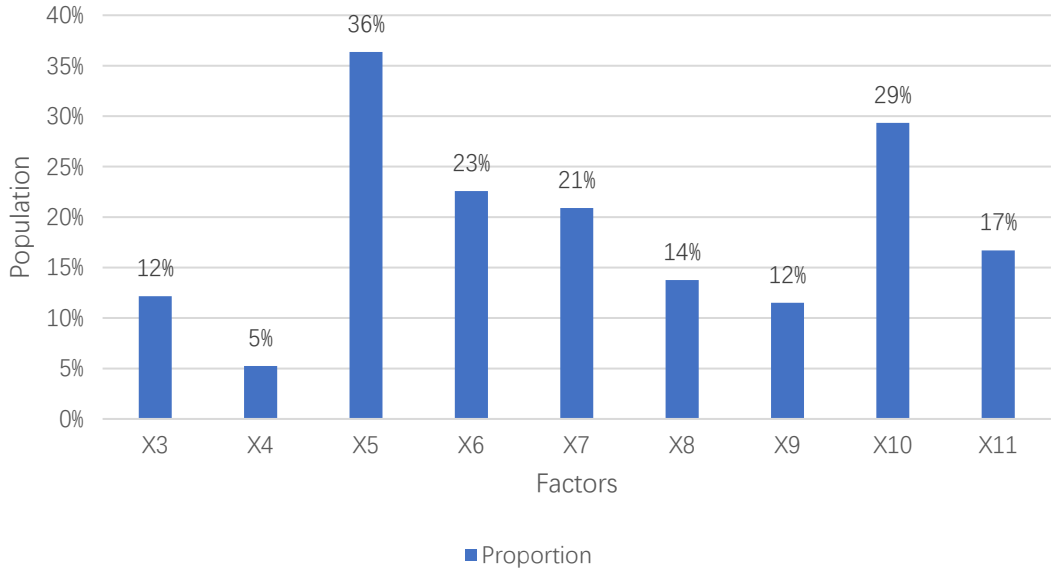


Fig. 1 Factors in the proportion of people with heart disease

It is not difficult to see from Figure 1 that the proportion of stroke is the highest, up to 36%. The proportion of drinking is the lowest, only 5%. Therefore, it can be concluded that stroke is the most likely to cause heart disease, and drinking alcohol is the least likely to cause heart disease.

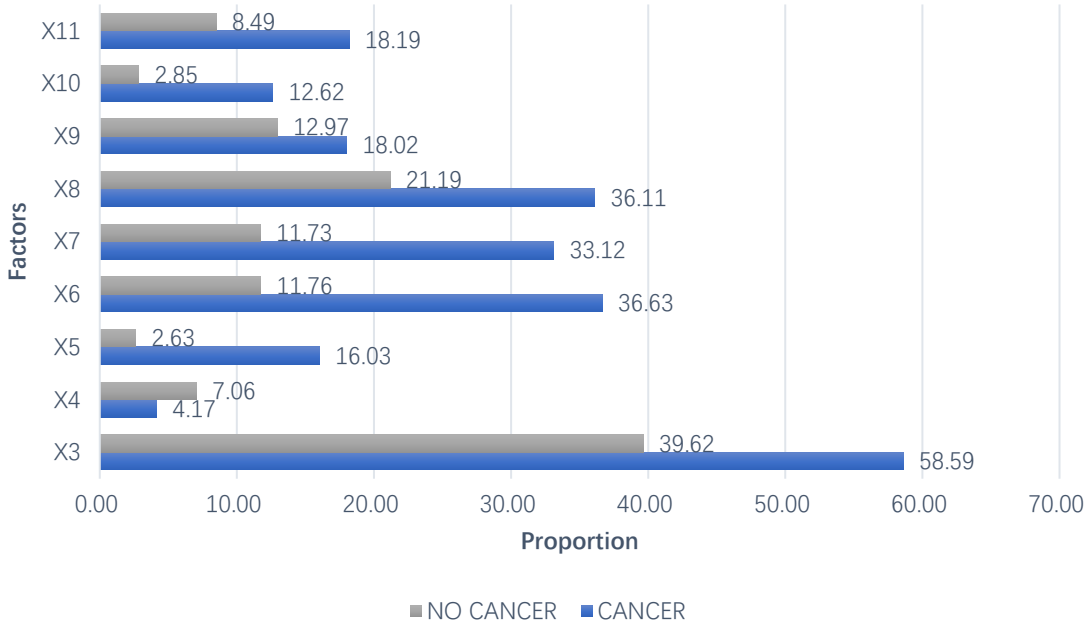


Fig. 2 Comparison of cancer and non-cancer under different factors

As can be seen from Figure 2, the proportion of patients with heart disease and the proportion of patients with heart disease are different in the factor x_3- x_{11}. The greater the difference, the smaller the influence of this factor on heart disease, and the smaller the difference, the greater the influence of this factor on heart disease. Among them, the difference of x_4 is only 3%, which has the greatest impact on heart disease.

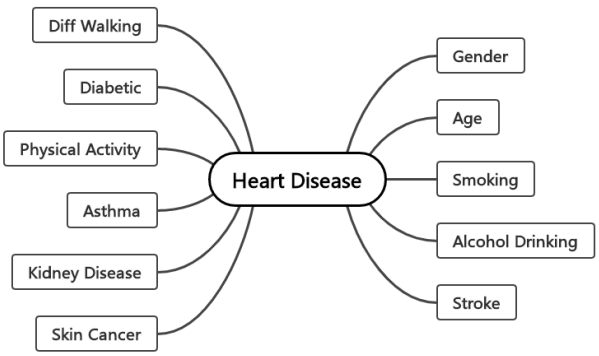


Fig. 3 Variable-related schematics

As can be seen from Figure 3, 11 factors that may induce heart disease are introduced into the model, including gender, age, smoking, drinking, stroke, different walking styles, diabetes, physical activity, asthma, kidney disease and skin cancer. By SPSSAU calculation, the final linear regression equation can be obtained:

$$Probit(p) = -3.004 - 0.354x_1 + 0.128x_2 + \ldots + 0.102x_{11} \quad (2)$$

Where p indicates that the probability of heart disease is 1.

Table 4. Model results

Elements	regression coefficient	standard error	z value	p-value	95% CI	marginal effect
x_1	0.354	0.019	19.104	<0.000	0.317 ~ 0.390	0.046
x_2	0.128	0.003	36.956	<0.000	0.121 ~ 0.135	0.016
x_3	0.227	0.018	12.383	<0.000	0.191 ~ 0.263	0.029
x_4	-0.059	0.039	-1.497	0.134	-0.136~ 0.018	-0.007
x_5	0.737	0.032	23.144	<0.000	0.674 ~ 0.799	0.134
x_6	0.390	0.022	17.489	<0.000	0.347 ~ 0.434	0.057
x_7	0.409	0.022	18.691	<0.000	0.366 ~ 0.451	0.061
x_8	-0.097	0.021	-4.622	<0.000	-0.138~-0.056	-0.013
x_9	0.142	0.025	5.592	<0.000	0.092 ~ 0.192	0.019
x_{10}	0.410	0.035	11.884	<0.000	0.342 ~ 0.478	0.064
x_{11}	0.102	0.024	4.155	<0.000	0.054 ~ 0.150	0.013
intercept	-3.004	0.041	-72.734	<0.000	-3.085 ~-2.923	-

*Dependent variable: Heart Disease

As can be seen from Table 4, 11 items including smoking, alcoholic drinking, stroke, differential walking, gender, age category, diabetes, physical activity, asthma, kidney disease and skin cancer are taken as independent variables. Taking Heart Disease as a dependent variable for binary Probit regression analysis.

The final concrete analysis shows that the regression coefficient of Smoking is 0.227, that of Stroke is 0.737, that of Diff Walking is 0.390, that of Gender is 0.354, that of Age Category is 0.128, and that of Diabetic is 0.409. The regression coefficient values of Asthma, Kidney Disease and Skincare are 0.142, 0.410 and 0.102, respectively, all of which are significant at the level of 0.01 ($p<0.000<0.01$), which means smoking, stroke, diff walking, gender, Age category, diabetes, Asthma, Kidney disease and skin cancer will have a significant positive impact on Heart Disease. As well as the marginal effect, it can be seen that the change (increase) range of Heart Disease is 1229.10%, 1763.66%, 1530.32%, 1898.92%, 3636.52%, 1621.35%, 529.39% and 995.67% respectively when one unit is added.

The regression coefficient of Alcoholic Drinking is -0.059, but it is not significant ($z=-1.497$, $p=0.134>0.05$), which means that alcoholic drinking has no influence on Heart Disease.

The regression coefficient of Physical Activity is -0.097, and it is significant at the level of 0.01 ($z=-4.622$, $p=0.000<0.01$), which means that Physical Activity will have a significant negative impact on Heart Disease. And the marginal effect value is -1.541, which means that when the Physical Activity increases by one unit, the change (decrease) range of Heart Disease is -449.73%.

Compared with other literatures on heart disease, this paper covers several factors that may induce heart disease for statistical research and analysis, rather than studying the influence of a single variable on heart disease. The statistical data analysis in this paper is broader and more comprehensive. It can give a more comprehensive consideration to the cause of heart disease, which can help medical staff to examine the patient's physical health in many directions and prevent heart disease in time.

4. Conclusion

Summary and analysis show that smoking, stroke, different walking, Gender, age category, diabetes, Asthma, Kidney disease and Skin cancer will have a significant positive impact on Heart Disease, and Physical Activity will have a significant negative impact on Heart Disease. However, Alcohol Drinking does not affect the Heart Disease.

Because the data used in this paper comes from a region, the sample data is regional and may be in groups in different regions. For different races and nations, there are differences in different environments, and the conclusions produced by statistical analysis may be wrong with the reality. But this paper still has the use value. First, this paper establishes a Probit regression model, and puts enough sample data into the model to calculate, and the calculated results are highly reliable. At the same time, using multiple linear regression, the conclusion is more comprehensive and specific. Then, in the choice of variables, in addition to the common influencing factors such as Gender, age, smoking and drinking, the factors such as stroke, diff walking, diabetes, physical activity, asthma, kidney disease and skin cancer are also introduced to make a multi-faceted statistical analysis of the pathogenic factors. Although the influence of these factors on inducing heart disease needs further investigation, this literature can help to point out the direction of its related research. In addition, multi-factor analysis is not only applied to heart disease, but also to any other diseases, which can help the medical professionals who prevent diseases to provide a more comprehensive examination plan, find diseases earlier and do a good job in prevention, thus reducing the prevalence and mortality of diseases.

References

[1] World Health Organization. WHO report announced that the top ten causes of death in the world, heart disease is still the "number one killer". Famous doctor, 2020, (17): 30.

[2] Cui Lihong. Data "said" the trend of cardiovascular disease. Journal of Practical Cardiopulmonary Vascular Diseases, 2023, 31(07): 5.

[3] Cao Xinxi, Xu Chenjie, Hou Yabing, et al. Epidemic trend and prediction of high-incidence chronic diseases in China from 1990 to 2025. Prevention and control of chronic diseases in China, 2020, 28 (01): 14-19.

[4] Lu Rongxin, Zhong Qianjin, Feng Yongjian. Meta-analysis of risk factors of congenital heart disease in China population. Journal of third military medical university, 2011, 33 (12): 1265-1268.

[5] Yuan Xue, et al. Monitoring results of congenital heart disease in 10665 children and analysis of environmental risk factors. China Maternal and Child Health, 2006, (06): 781-783.

[6] Qiu Xiaoqiang, et al. A case-control study on MTHFR gene, CBS gene, environmental factors and congenital heart disease. Chinese Journal of Epidemiology, 2006, (03): 260-263.

[7] Zhang Jiaxing. More than 60 diseases are related to drinking. Science and Technology Daily, 2023.

[8] Du Ruixue. Prediction and analysis of chronic diseases such as stroke based on Bayesian network model. Computer programming skills and maintenance, 2023, (06): 47-49.

[9] Chen Yandong, et al. Risk factors of coronary heart disease complicated with heart failure based on data mining. China Journal of Gerontology, 2018, 38(22): 5380-5383.

[10] Shi Manzhu. Diabetes and Heart Disease. Chinese Journal of Internal Medicine, 1995, (09): 581-582.

[11] Zhao Kun, et al. Preliminary study on subclinical diabetic heart disease and related factors. Chinese journal of endocrinology and metabolism, 1998, (06): 24-27.

[12] Zhao Baolu. Smoking, free radicals and heart disease. Journal of Nature, 1989, (08): 610-611.

[13] Li Peiyao. Analysis of major cardiovascular and cerebrovascular diseases deaths and their influencing factors in people over 40 years old in China. China Center for Disease Control and Prevention, 2018.

[14] Li Muzi. A new method for predicting middle-aged people's heart disease comes out. China Science Journal, 2023.