DOI: 10.54254/2753-8818/52/2024CH0102

Research on the factors affecting diabetes mellitus based on logistic regression

Junxi Li

School of Applied Statistics, Northeastern University, Qinhuangdao, 066000, China

202119298@stu.neuq.edu.cn

Abstract. In recent years, more and more diabetic patients have appeared all over the world, and people have begun to pay more and more attention to this kind of health problems, and it is a necessary thing to understand the influencing factors related to diabetes mellitus. This study explores the key factors that influence the occurrence of diabetes using multivariate logistic regression analysis and can be used to predict diabetes in individuals. The data for the study was obtained from the Kaggle website and various factors affecting diabetes were analyzed and a multivariate logistic regression model was developed to assess the impact of different factors on the risk of developing diabetes. The study found that good lifestyle habits and better basic personal circumstances the lower the risk of developing diabetes. These findings emphasize the importance of individuals focusing on their daily habits and improving their quality of life, which can help individuals reduce their risk of diabetes, and for those who are potentially at risk of developing diabetes, personal information can be used to make predictions and provide appropriate advice to help them change their bad habits.

Keywords: Diabetes, multiple logistic regression, risk factor, predictive model.

1. Introduction

Diabetes is one of the most prevalent chronic diseases globally, affecting a large population each year and imposing a significant financial burden on the economy. Ten years ago, experts predicted that by 2030, the global population suffering from diabetes would exceed 366 million, with China reaching 42.3 million. However, data from 2008 show that the number of diabetic patients had already reached 94.5 million, far exceeding the previously projected value for twenty years later [1]. This alarming increase in incidence underscores the urgency of understanding the factors contributing to diabetes. Identifying these factors is crucial for several reasons. Firstly, understanding the risk factors can aid in the prevention and control of diabetes, facilitating the development and implementation of effective health policies [2]. Secondly, recognizing these factors allows for the personalization of treatment regimens, which can improve treatment outcomes and enhance the quality of life for patients.

Many factors influencing diabetes are closely related to lifestyle, such as diet, physical activity, and smoking. However, diabetic patients often struggle with self-management, showing the best compliance with medication but the worst with blood glucose monitoring. Age, gender, self-efficacy, and diabetes-related knowledge are significant contributors to diabetes [3]. Therefore, managing diabetic patients requires not only personal effort but also substantial public healthcare resources, support from society, government, and hospitals, and increased awareness-raising efforts [4]. Therefore, all people need to pay attention to this issue, not only individuals need to understand the factors that cause diabetes, but also the relevant authorities should take certain measures. Healthcare professionals should prioritize regular assessment of the debilitating conditions of diabetic patients and develop professional, multifaceted, and safe intervention programs. These programs should consider the specific conditions of elderly patients with diabetes and aim to standardize and improve overall care [5]. Regular phone and SMS follow-ups for type 2 diabetes patients have been shown to improve lifestyle and behavioral habits, contributing to better disease control and rehabilitation [6].

For the data selection in this study, factors that may lead to diabetes were chosen, including lifestyle habits such as smoking, physical activity, daily consumption of vegetables, mental health, physical health and fruits and some personalized basic information such as age, education level and income. Personalized diet and exercise interventions for pregnant women, for instance, can effectively control weight gain during pregnancy, reduce the incidence of gestational diabetes mellitus (GDM), and ensure the health of pregnant women [7]. Wang highlighted that people who remain sedentary for long periods, do not engage in physical activity, and consume large amounts of alcohol are more likely to develop diabetes [8]. While moderate alcohol consumption may reduce diabetes prevalence, the overall trend shows an increase in diabetes prevalence with higher alcohol intake.

This study will analyze the impact of these factors on diabetes, providing insights that can help individuals reduce harmful habits in their daily lives. Additionally, the financial burden of diabetes on patients, families, and society is enormous. By studying the influencing factors of diabetes, valuable references can be provided for reducing the burden of disease and improving the efficiency of medical resource utilization.

2. Methodology

2.1. Data source and description

The data is derived from the Diabetes Health Indicators Dataset, a pure set of CDC's (Centers for Disease Control and Prevention) BRFSS2015 survey responses. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual CDC survey on health-related issues. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. A CSV of the data set available on Kaggle for the year 2015 has been used for this project. This dataset consists of 441,455 individuals and has 330 features. In it I screened out 44,009 data and studied 12 of them. These are either questions that the participants are asked directly, or they are calculated on the basis of the responses of the participants.

2.2. Index selection and description

The target variable Diabetes_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There are also a number of independent variable indicators. As shown in Table 1.

In the table it is possible to see the names of the variables that influence the factors of diabetes, as well as each variable their type: binary and continuous. In the last column is a description of each variable, explaining in detail what each variable corresponds to.

Table 1. Variable interpretation.

Variable	Type	Meaning
HighCoal	Binary	0 for no high cholesterol 1 for high cholesterol
Smoker	Binary	Respondents smoked at least 100 cigarettes in their entire lifeRespondents smoked at least 100 cigarettes in their entire life.0 for no and 1 for yes
HeartDiseasorAttack	Binary	coronary heart disease (CHD) or myocardial infarction (MI). 0 for no and 1 for yes
PhysActivity	Binary	physical activity in past 30 days (not including job). 0 for no and 1 for yes
Fruits	Binary	Consuming fruit 1 or more times per day 0 for no and 1 for yes
Veggies	Binary	Consuming fruit 1 or more times per day 0 for no and 1 for yes
PhysHlth	Continuous	How many days during the past 30 days when physical health not good
MentHlth	Continuous	How many days during the past 30 days when mental health not good
Sex	Binary	0 for female and 1 for male
Education	Continuous	Education level ranging from 1 to 6
Income	Continuous	Income scale ranging from 1 to 8 1 for less than $10,000, 5 for less than $35,000 and 8 for $75,000 or more

2.3. Method introduction

In this study, Multiple Logistic Regression Analysis was employed to investigate the factors influencing diabetes [9, 10]. This method is suitable for the research as it allows people to assess the relationship between multiple independent variables and a categorical dependent variable. The first step is to clean the data, processing and removing outliers to ensure that the retained data is viable for this analysis. The basic information of the independent variables was then analyzed with descriptive statistics, examined the mean and distribution of each variable, roughly analyzed the relationship between each variable and the presence or absence of diabetes through scatter plots. After that, the logistic regression model was specified with diabetes status as the dependent variable and the selected independent variables as predictors.

The logistic regression model was fitted using maximum likelihood estimation. This process involved iteratively adjusting the model parameters to maximize the likelihood of observing the given data.

3. Results and discussion

3.1. Descriptive analysis

Analyzed through descriptive statistics, it can be seen that the dependent variable diabetes_012 is a categorical variable with three values, while the dependent variables HighChol, smoker, PhysActivity, HeartDiseasorAttacker, Fruits, and Vegies are binary variables that take values between 0 and 1 between 0 and 1, while the other variables are continuous variables. The highest frequency of no diabetes was found in the sample with 37,107 or 84.32% of the total, while 6,001 or 13.64% of the total were diabetic. Slightly more people did not have high cholesterol than those with high cholesterol, 57.87% and 42.13% respectively. Similarly, non-smokers outnumbered smokers by about 10%. Significantly more respondents did not have heart disease than those with heart disease, who accounted for less than 10 percent of the total. The study also found that more than 77% of the respondents exercised daily and that the number of people who ate vegetables and fruits daily was also higher than the number of people who did not eat fruits daily.

The scatterplot (Figure 1) roughly show that cholesterol, smoking and heart disease are positively correlated with the prevalence of diabetes, while physical activity, healthy eating habits, education level and income are negatively correlated with the prevalence of diabetes. The results of these preliminary analyses suggest that healthy lifestyles and diets are more prevalent in the sample diabetes.

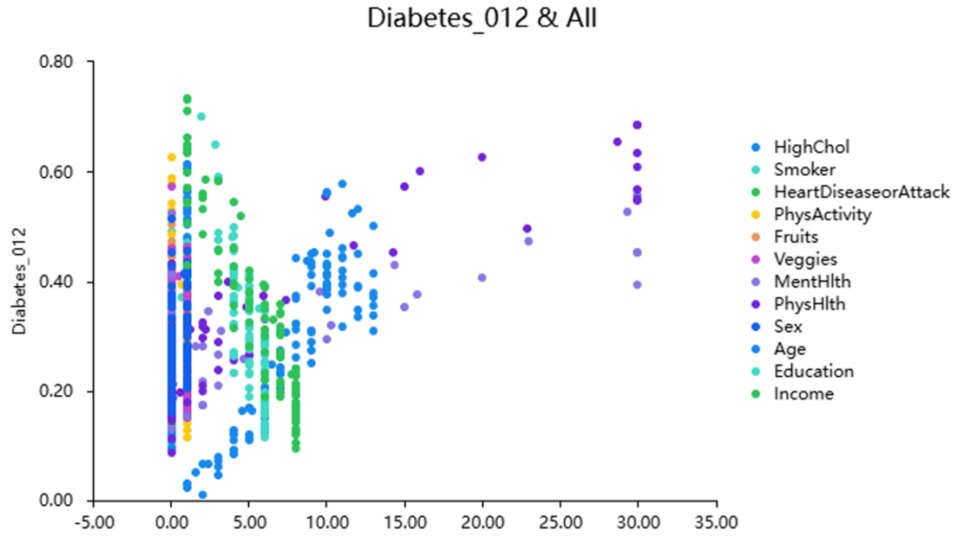

Figure 1. Scatterplot of variables

3.2. Model results

Multiple logistic regression (Table 2) analyses were acted to further validate the hypotheses of the model and to explore the effect of the respective variables on diabetes. The model used no diabetes (diabetes_012 equal to 0) as the control group. The first equation 1.0/0.0 means that, relative to 0.0 (no diabetes), with 1.0 (prediabetes). The second equation 2.0/0.0 means that, relative to 0.0 (no diabetes), at 2.0 (with diabetes).

The table 2 leads to the equation of multiple logistic regression:

$$\ln\left(\frac{1.0}{0.0}\right) = -3.346 + 0.779 \times \text{Highchol} - 0.103 \times \text{Smoker} + \cdots - 0.199 \times \text{Education} \tag{1}$$

$$\ln\left(\frac{2.0}{0.0}\right) = -1.825 + 0.858 \times \text{Highchol} - 0.067 \times \text{Smoker} + \cdots - 0.169 \times \text{Education} \tag{2}$$

The multivariate logistic regression model allows people to predict diabetes and to detect the risk of diabetes in the population. This paper can also analyze the influencing factors, relative to 0.0, the regression coefficient of HighChol is 0.779 with a p-value of less than 0.01 at 1.0, implying that HighChol has a significant positive effect on Diabetes, with an OR of 2.179, implying that the magnitude of the change in HighChol when increased by one unit of HighChol was 2.179 times (change from 0.0 to 1.0).Similar significant positive relationships with p less than 0.01 are also seen in MentlHlth and PhysHlth and Age. In contrast, Income and Education are included in the significant negative effects with p-values less than 0.01. As opposed to 0.0, in the context of 2.0. In the other group, as opposed to 0.0, in the case of 2.0. Except for Smoke and MentHlth, all other influences are significant.

Multiple logistic regression analysis also revealed that HighChole had a significant positive effect on diabetes. The relatively small effect of smoking suggests that the effect of smoking on diabetes is more complex and may be moderated by other factors. HeartDiseaseAttack significantly increased the risk of diabetes, emphasizing the importance of cardiovascular health in the management of diabetes. Physical activity had a large effect on diabetes, and regular physical activity is an important measure to prevent diabetes. Daily fruit and vegetable consumption habits had negative regression coefficients in both models, suggesting that good dietary habits can help reduce the risk of diabetes. Mental health and physical health had smaller regression coefficients in both models, suggesting that they have a smaller direct effect on diabetes, but further research is needed to investigate their potential indirect effects. Sex and age showed a significant positive relationship in both models, suggesting that males and those who are older have a higher risk of developing diabetes. Income and education had negative regression coefficients in both models, suggesting that higher income and education levels are effective in reducing the incidence of diabetes.

Table 2. Model results.

1.0	Cofficients	p	OR
Highchol	0.779	0.000	2.179
Somker	-0.103	0.141	0.902
HeartDiseasorAttack	0.030	0.786	1.030
PhysActivity	-0.163	0.039	0.850
Fruits	0.052	0.476	0.949
Veggies	-0.105	0.226	0.900
MentlHlth	0.013	0.002	1.013
PhysHlth	0.017	0.000	1.017
Sex	0.093	0.192	1.097
Age	0.086	0.000	1.090
Income	-0.064	0.000	0.938
Education	-0.199	0.000	0.820
Constant	-3.346	0.000	0.035
2.0	Cofficients	p	OR
Highchol	0.858	0.000	2.358
Somker	-0.067	0.028	0.935
HeartDiseasorAttack	0.524	0.000	1.689
PhysActivity	-0.323	0.000	0.724
Fruits	-0.091	0.004	0.913
Veggies	-0.137	0.000	0.872
MentlHlth	0.003	0.091	1.003
PhysHlth	0.023	0.000	1.024
Sex	0.244	0.000	1.276
Age	0.132	0.000	1.141
Income	-0.109	0.000	0.896
Education	-0.169	0.000	0.844
Constant	-1.825	0.000	0.161

4. Conclusion

This study provides an insight into the factors influencing diabetes mellitus through multivariate logistic regression analysis. It was found that high cholesterol, heart disease, smoking, less physical activity, unhealthy dietary habits, and lower income and education levels were significant diabetes risk factors. Conversely, regular physical activity, good dietary habits, and higher levels of income and education helped prevent the development of diabetes. These findings provide an important reference for the development of targeted diabetes prevention and management strategies, which can help to reduce the enormous burden of the disease on individuals, families and society, and optimize the efficient use of healthcare resources. Future studies can explore the indirect effects of mental health, social support and other factors on diabetes to improve the overall effectiveness of disease management and patients' quality of life.

References

[1] Yang W, et al. 2010 Prevalence of diabetes among men and women in China. New England Journal of Medicine, 362(12), 1090-1101.

[2] Musleh S, et al. 2020 Identification of potential risk factors of diabetes for the Qatari population. In 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies, IEEE, 243-246.

[3] Oyebode O and Orji R 2019 Detecting factors responsible for diabetes prevalence in Nigeria using social media and machine learning. In 2019 15th International Conference on Network and Service Management (CNSM).

[4] Sun S N, et al. 2011 The current status and influential factors of self-management in diabetic patients. Chinese Journal of Nursing, 3, 229-233.

[5] Li J, et al. 2014 Psychosocial factors influencing self-management behaviors in patients with diabetes: A review. Chinese Journal of Nursing, 2, 207-211.

[6] Jia W, et al. 2019 The current situation and influential factors of frailty in elderly patients with diabetes. Chinese Journal of Nursing, 2, 188-193.

[7] Guo X, Liu Z and Li H 2002 Type 2 diabetes mellitus induced by diets and its features of renal involvement in rat. Chinese Journal of Diabetes, 5, 35-39.

[8] Wang J H., et al. 2014 The effects of follow-up by telephone with SMS on lifestyle and behavior habits of patients with type 2 diabetes. Journal of Nursing Administration, 10, 701-703.

[9] Wang Z 2017 The relationship between sedentary behavior, alcohol consumption and diabetes mellitus in the prevalence of middle-aged and elderly people. Journal of Nursing Administration.

[10] Wu X, et al. 2015 Effect of dietary and exercise intervention on the incidence of GDM in women with pre-pregnancy obesity. Modern Clinical Nursing, 9, 24-27.