Research on the Influence Factors that Possibly Lead to Diabetes

Pengzhou Xu
School of Mathematical Sciences, Inner Mongolia University, Hohhot, 010000, China

Keywords: Diabetes, Pathogenic Factors, Binary Logistic Model.

Abstract: Diabetes has become a serious public health problem worldwide. Previous studies have found that diabetes is related to family genetics, age and high blood pressure, but there are other unknown factors worth investigating. In this study, a Binary Logistic Model was used to process data from the American Behavioral Risk Factor Surveillance System (BRFSS), published in 2015. The data included 30,691 men and women of all income levels and age groups. The study concluded that while diabetes was not associated with Vegetable Consumption, it had significant positive effects on Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke and Difficulty Walking. There were significant negative effects on Exercise, Fruit Consumption, Alcohol and Education. Among the factors closely related to diabetes, Stroke, Age, Difficulty Walking, did not appear in previous studies. The research not only provides some new perspectives for relevant medical personnel to study the pathogenesis of diabetes, but also helps diabetic patients to treat diabetes in a timely manner. At the same time, it also plays a positive role in diabetes prevention.

1 INTRODUCTION

China's economic development has led to changes in Chinese people's living habits, and the incidence of diabetes has increased year by year (Wang et al., 2021). In addition, from 2013 to 2020, the mortality rate of urban diabetic patients in China has increased significantly, and diabetes has become an important public health problem in China (Zhu et al., 2020 & Li et al., 2020). Therefore, understanding the causes of diabetes is of great significance to control the development and treatment of the disease and reduce the mortality. The purpose of this paper is to study the potential factors that lead to diabetes to help people assess their own risk of diabetes, and to take a series of protective and treatment measures.

Diabetes is a metabolic disease, which is usually caused by hereditary and long-term external influences, causing organ lesions, resulting in lower insulin secretion than normal levels (Bai, 2018 & Robinson and Pickering 2024)). The occurrence factors of diabetes are complex, and some scholars have found that diabetes has a certain correlation with age, BMI, overweight and hypertension (Zhang et al., 2022). In addition, Du et al. found that smoking and living conditions were related to diabetes (Du et al., 2022). Based on these research results, this paper will study whether 14 factors (Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, Difficulty Walking, Education, Income) whether these factors are related to diabetes. In a similar direction, Ye et al. used a multi-factor logistic regression model and a mixed graph model (Ye et al.,2024). Logistic regression analysis model is a classic model with high efficiency and simplicity. In the field of medical research, case-control studies are needed to establish multiple paired groups. The literature indicates that unhealthy lifestyle (Smoking, Low Physical Pabor, Alcohol, BMI) is closely related to diabetes, but the screening population in the literature is mainly Chinese elderly, and the possibility of participating in screening in other age groups is not fully considered. Li et al. established a structural equation model (SEM) (Li et al.,2023), but the computational complexity of this model was high, and a large amount of data needed to be collected to ensure the accuracy of the calculation. Moreover, the literature also did not fully consider the possibility of other age groups participating in screening. Duan et al. also adopted a Logistic regression model, but only considered the relationship between sleep time and diabetes, and did not investigate more factors affecting the onset of diabetes (Duan et al., 2019). Deng et al. used the PSO-BP neural network model which combines Particle Swarm Optimization (PSO) algorithm and Backpropagation (BP) neural network, but this model

Xu, P.
Research on the Influence Factors that Possibly Lead to Diabetes.
DOI: 10.5220/0012983100004601

In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA 2024)*, pages 31-37
ISBN: 978-989-758-722-1

is sensitive to parameter setting. Improper Settings will cause the algorithm to fail to converge (Wang et al., 2019).

In short, the research on the influencing factors of diabetes has attracted many scholars. This paper will mainly use the binary Logistic regression model to analyse the influencing factors of diabetes, and put forward suggestions on the prevention of diabetes according to the results.

2 METHODS

2.1 Data Sources

The data in this literature comes from the Kaggle website, and the data is compiled by Alex Teboul based on the American Behavioral Risk Factor Surveillance System (BRFSS) in 2015, with a total of 30691 samples.

2.2 Variable Selection

The data used in this paper count a total of 30691 people, including those who have and do not have diabetes, of whom 13949 are male and 16742 are women. The data contains 14 variables (Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, Difficulty Walking, Education, Income) There are 21 features(Age, Gender, BMI, High Blood Pressure, High Cholesterol, Cholesterol Tests, Heart Disease, Health Insurance, Medical Exam Costs, Health Status, Mental Health, Physical Health, Smoking, Stroke, Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, Difficulty Walking, Education, Income) in the original data, and since this literature does not study the clinical significance in depth, the sample features with strong clinical significance are discarded.

Table 1 shows the number of people with this disease and the number of people with diabetes. As shown in Table 1, the symbols of each factor are shown above for ease of writing. The data sample was 30691 people, of whom 15345 had diabetes.

Table 2 shows the number of females and males, and diabetes patients in each education group. Among the characteristics of education, 1-6 means never receiving an education until graduating from college. The education of people in the data is mainly concentrated in technical school graduation and university graduation.

Table 1: Logogram and numbers of the 14 factors.

Elements	Logogram	Number	Diabetes
Age	x_1	30691	15345
Gender	x_2	30691	15345
BMI	x_3	30691	15345
High Blood Pressure	x_4	17272	11586
High Cholesterol	x_5	16261	10380
Smoking	x_6	14600	7981
Stroke	x_7	1913	1422
Exercise	x_8	21594	9767
Fruit Consumption	x_9	18903	9117
Vegetable Consumption	x_{10}	24252	11625
Alcohol	x_{11}	1325	369
Difficulty Walking	x_{12}	7624	5614
Education	x_{13}	30691	15345
Income	x_{14}	30691	15345

Table 2: Gender distribution by education group.

Education	1	2	3	4	5	6
Number	35	714	1425	8281	8684	11552
Female	21	408	855	4621	4994	5843
Male	14	306	570	3660	3690	5709
Diabetes	21	504	936	4623	4516	4745

Table 3: Gender distribution by income group.

Income	1	2	3	4	5	6	7	8
Number	1528	1891	2303	2888	3454	4435	4967	9225
Female	1075	1269	1507	1781	2030	2373	2506	4201
Male	453	622	796	1107	1424	2062	2461	5024
Diabetes	996	1285	1464	1739	1907	2275	2337	3342

Table 4: Gender distribution by age group.

Age	[18-34]	[35-49]	[50-64]	[65-79]	[80-100]
Number	1929	4769	10992	10594	2362
Female	1043	2693	5990	5679	1337
Male	886	2121	5002	4915	1025
Diabetes	240	1454	5692	6579	1380

Table 5. Binary Logistic regression prediction accuracy.

Accuracy rate	Recall rate	Precision rate	F1	AUC
0.728	0.728	0.729	0.728	0.803

Table 3 shows the number of females and males, and diabetes patients in each income group. Among the income characteristics, 1 to 8 respectively under the annual revenues of 10000 USD to the annual revenues of 75000 USD or above. In the data, people's annual income is mainly distributed in more than 75000 USD.

Table 4 shows the number of females and males, and diabetes patients in each age group. In the data used in this paper, people are mainly distributed in the age range of 50-64 and 65-79 years old.

2.3 Research Protocol

In this paper, the binary Logistic model was used for analysis. Whether diabetes mellitus was a dependent variable (Y), and 14 characteristics were independent variables (X), with 0 representing no and 1 representing yes. Next, this paper uses SPSSAU and SPSSPRO to analyse the relationship between the effect of X on Y, the relationship between the 14 factors on diabetes.

2.4 Model Principle

Binary Logistic regression is a generalized linear regression, which is used to solve the study of influencing factors in binary classification problems. The vector form of the linear regression model is:

$$f(x) = w^T x + b \tag{1}$$

Unit step function is needed to transform the results of regression model into 0 and 1 classification results, and Sigmod function continuity and function value distribution can well replace unit step function. Its formula is as follows:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

The independent variable x in the Sigmod function is replaced by the linear regression equation (1), and the prediction function of the Logistic model is finally obtained as follows:

$$P = \frac{1}{1+e^{-(w^T x+b)}} \tag{3}$$

2.5 Model Testing

Since the quality of model fit is measured by model prediction accuracy, F1 represents the harmonic average of precision and recall, and AUC represents a measure of the ability of the classifier to distinguish classification. It can be seen from Table 5 that the fitting quality of the model is acceptable.

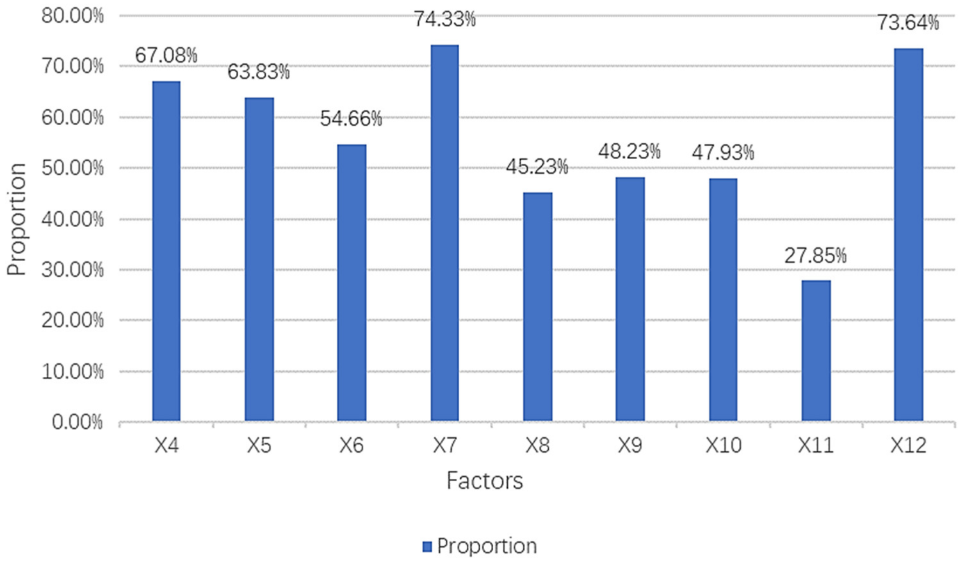

Figure 1: The proportion of factors affecting diabetic patients.

3 RESULTS AND DISCUSSION

3.1 Correlation Analysis

Figure 1 shows the proportion of a range of factors that contribute to diabetes in people with diabetes. Factors that lead to diabetes include High blood pressure, High cholesterol, Smoking, Stroke, No exercise, No eating fruits, No eating vegetables, Alcoholism, Difficulty walking, etc. The proportion of diabetes samples was identified from the samples and classified according to different influencing factors to determine the probability of each risk leading to diabetes.

It can be seen from Figure 1 that stroke has the greatest impact on diabetes at 74.33%, followed by difficulty walking at 73.64%, and alcohol abuse has the smallest impact on diabetes at 27.85%. Therefore, of the characteristics covered in this dataset, stroke is the most risk factor for developing diabetes, with difficulty walking a close second.

Before the regression analysis, the correlation between independent variables and dependent variables was analysed. Table 6 shows the strength of the correlation between diabetes mellitus and 14 factors (Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, Difficulty Walking, Education, Income) using Pearson correlation coefficient. Whether these factors are associated with diabetes can be preliminarily judged by observing whether the p-value is greater than 0.05. The p-value of all 14 factors were less than 0.01, indicating that 14 factors were significantly related to diabetes. In the further study, Pearson correlation coefficient was observed to analyse whether the correlation was positive or negative. Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, Education, Income were significantly negatively correlated with diabetes, while Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Difficulty Walking, were significantly positively correlated with diabetes (Table 6).

Table 6: Pearson correlation analysis.

Factors	p-value	Correlation coefficient
x_1	0.000	0.279
x_2	0.000	0.041
x_3	0.000	0.286
x_4	0.000	0.388
x_5	0.000	0.294
x_6	0.000	0.089
x_7	0.000	0.125
x_8	0.000	-0.147
x_9	0.000	-0.045
x_{10}	0.000	-0.080
x_{11}	0.000	-0.094
x_{12}	0.000	0.272
x_{13}	0.000	-0.153
x_{14}	0.000	-0.208

All in all, these 14 factors are correlated with diabetes. Although the correlation coefficients of Gender, Smoking, Stroke, Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, and

Education are less than 0.2, their p-values are all less than 0.01, so the correlation exists.

3.2 Model Results

As shown in Figure 2, the study introduced various factors that could be related to diabetes into the model, these include Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Exercise, Fruit Consumption, Vegetable Consumption, Alcohol, Difficulty Walking, Education, Income. Through calculation, this study obtained the final linear regression equation:

$$P = -3.874 + 0.147x_1 + 0.257x_2 + \cdots + 0.510x_{12} - 0.057x_{13} - 0.086x_{14} \quad (4)$$

Table 7 is the final result of the model. By observing whether the value of p is greater than 0.05, those factors can be preliminarily judged to have an impact on diabetes. First of all, the p value of Vegetable Consumption is greater than 0.05, which has no effect on diabetes. Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Exercise, Fruit Consumption, Alcohol, Difficulty Walking, Education and Income were all less than 0.01, indicating that these factors had a particularly strong relationship with diabetes. In the further study of the influencing factors, the marginal effect combined with the regression coefficient can specifically reflect the influence of these 13 factors on diabetes. It is reflected in: Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, Difficulty Walking for each additional unit, the changes (increases) of Diabetes were 115.8%, 125.3%, 108.3%, 248.6, 202.3%, 108.1% and 142.3%, respectively. Because these values are large, these factors have a significant impact on developing diabetes.

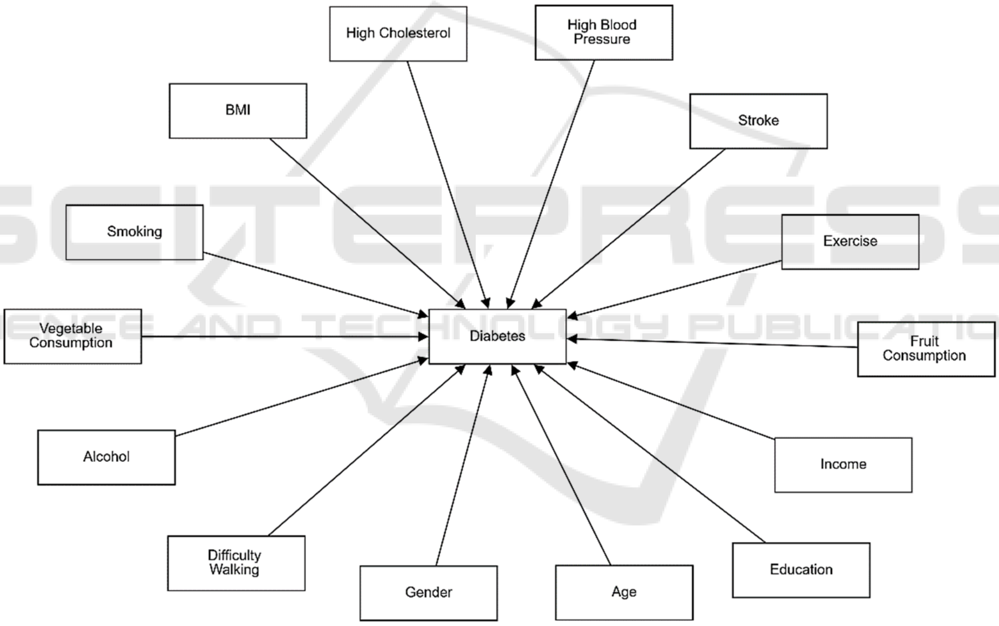

Figure 2: Variable-related schematic.

Table 7: Model results.

Elements	regression coefficient	standard error	z value	p-value	95% CI	marginal effect
x_1	0.147	0.005	26.937	0.000	1.146 ~ 1.171	0.027
x_2	0.257	0.028	9.237	0.000	1.225 ~ 1.366	0.047
x_3	0.079	0.002	34.382	0.000	1.078 ~ 1.088	0.014
x_4	0.910	0.029	31.504	0.000	2.349 ~ 2.630	0.181
x_5	0.704	0.028	25.484	0.000	1.916 ~ 2.135	0.134

x_6	0.078	0.028	2.825	0.005	1.024 ~ 1.142	0.014
x_7	0.352	0.060	5.864	0.000	1.264 ~ 1.600	0.064
x_8	-0.087	0.031	-2.777	0.005	0.862 ~ 0.975	-0.016
x_9	-0.027	0.029	-0.942	0.346	0.920 ~ 1.030	-0.005
x_{10}	-0.096	0.035	-2.768	0.006	0.849 ~ 0.972	-0.017
x_{11}	-0.790	0.072	-11.044	0.000	0.395 ~ 0.522	-0.142
x_{12}	0.510	0.035	14.519	0.000	1.555 ~ 1.785	0.095
x_{13}	-0.057	0.015	-3.822	0.000	0.917 ~ 0.973	-0.010
x_{14}	-0.086	0.007	-11.546	0.000	0.904 ~ 0.931	-0.016
Constant	-3.874	-	-	-	-	-

Compared with other similar studies, they were more likely to look at the effects of age-specific and known factors on diabetes, such as overeating in the elderly, high blood pressure in the elderly, and sedentary adolescents. In contrast, the features contained in this paper are more detailed and comprehensive, which can not only avoid some errors caused by univariate analysis, but also bring some new perspectives to the study of diabetes. Not only will this direction for relevant medical staff pointed out that more treatment, can also be found in a timely manner and prevent diabetes.

4 CONCLUSION

Based on the current study, which focuses on the influencing factors that may contribute to the development of diabetes, the conclusions are: Age, Gender, BMI, High Blood Pressure, High Cholesterol, Smoking, Stroke, and Difficulty Walking may be associated with diabetes. Some of these factors have been overlooked in previous studies.

The research method used in this paper is binary Logistic regression, which can better deal with classification problems than the single factor analysis method used in other studies. However, it is undeniable that the classification accuracy of binary logistic regression may not be high when processing high-dimensional data, which may cause errors in the research results of this paper. At the same time, binary Logistic regression is more sensitive to the selection of features, and may have low precision problems when dealing with data sets with complex relationships. In addition, due to the limited amount of data and the fact that the sample did not cover all age groups and races, the accuracy of the model results may also be affected. However, this study still has great advantages and values. First of all, the selection of data features was screened, and after excluding the features that had little relevance to this study, Pearson correlation analysis was used to ensure that all 14 features were relevant to Diabetes. After ensuring the correlation between Diabetes and 14 characteristics, the graphical method is used to visually show and compare the difference in the proportion of different influencing factors in diabetic patients, which makes the experimental process more intuitive and the experimental results clearer. On the other hand, it has a positive effect on the prevention and treatment of diabetes. In addition to the known factors associated with diabetes such as High blood pressure, High blood lipids and Age, there are many other factors associated with diabetes, such as Stroke and Difficulty walking. Overall, although the relationship between diabetes and these new possible influencing factors needs further study, it is meant to point the way for future research. And it can improve the efficiency of prevention and treatment of diabetes patients, timely detection of diabetes, improve the survival rate of patients.

REFERENCES

Wang L, et al. 2021 Prevalence and Treatment of Diabetes in China, 2013-2018. *JAMA* **326(24)** 2498–2506.

Zhu W, et al. 2019 Mortality trend of diabetes mellitus in China from 2006 to 2020. *Journal of Binzhou Medical College* **46(04)** 299-303+314.

Li Y, et al. 2020 Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross-sectional study. *BMJ.*

Bai Y 2018 Introduction to the cause of diabetes. *Science and technology* **2** 168.

Robinson G N and Pickering R J 2024 Melanocortins and their potential for the treatment prevention and amelioration of complications of diabetes. *Diabetology* **5(1)** 69-84.

Zhang X, et al. 2022 Chinese adults, early onset status and influencing factors of study. *Chinese journal epidemiology.*

Gao Y Gong L and Liang G 2022 Beijing's Daxing district residents diabetes prevalence and influence factors analysis. *Chinese journal of clinical doctors* **50(11)** 1322-1325.

Du J, et al. 2022 Study on influencing factors of diabetes in Chinese elderly people based on health ecology model. *Chronic disease prevention and control in China* **30(6)** 457-460 +464.

Ye T T, et al. 2024 Association between unhealthy lifestyles and hypertension, diabetes and dyslipidemia in old adults in China. *Chinese Journal of Epidemiology* **45(03)** 385-392.

Li X, et al. 2023 Prevalence of diabetes in rural elderly in Yunnan Province and its influencing factors based on structural equation model. *The Chinese journal of disease control* **27(5)** 546-550.

Duan S, et al. 2019 The relationship between nocturnal and daytime sleep patterns and the prevalence of diabetes mellitus in middle-aged and elderly people in China. *Ningxia Medical Journal* **43(8)** 703-707.

Wang S, et al. 2019 Risk factor analysis and screening model construction of type 2 diabetes based on BP neural network based on particle swarm optimization algorithm. *Western Chinese Medicine*.