Dean&Francis

ISSN 2959-6157

The Research on Factors that Possibly lead to Diabetes

Rubin Li[1, *]

[1]School of Mathematics, Hunan University, Changsha, 410006, China

*Corresponding author: Lijm@xaut.edu.cn

Abstract:

Although prior research has indicated that the occurrence of diabetes is associated with genetic inheritance, overweight conditions and dietary habits, there are still many other unknown factors worth studying. This research employed the Probit model approach to handle the data from the BRFSS 2015 dataset on Kaggle. The dataset was published and updated in 2021 and encompassed information on 253,680 individuals. The conclusion is that the onset of diabetes is strongly associated with Sex, Age, BMI, Physical activity, Eating habits, Smoking, Mental health and stroke, and is particularly closely related to Difficulty walking, Heart disease, High cholesterol, and High blood pressure. Many factors that may be closely associated with diabetes have not been found in prior studies, such as Mental health, Difficulty walking, Heart disease, High cholesterol, and High blood pressure. This provides some new ideas for studying the pathogenesis of diabetes and the treatment of diabetes, and points the way for future research.

Keywords: Diabetes; causative factors; probit regression.

1. Introduction

Diabetes has emerged as one of the most highly regarded chronic diseases. According to a report by the World Health Organization (WHO), Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. Buse et al. analyzed the data on Carotid artery intima-media thickness (CIMT), which is independently predictive for risk of cardiovascular events and concluded that the risk of cardiovascular disease in people with diabetes is two to three times that of people without diabetes [1]. Cummings et al. also concluded that individuals with diabetes were more likely to report stress and/or depressive symptoms [2]. In 2021, it is estimated that 537 million people have diabetes, and this number is projected to reach 643 million by 2030. It is also reckoned that over 6.7 million people aged 20-79 will die due to diabetes-related causes in 2021.The number of children and adolescents (i.e., up to 19 years old) having diabetes is on the rise year by year. In 2021, over 1.2 million children and adolescents suffer from type 1 diabetes. In recent decades, the prevalence of diabetes among adults in China had been on the rise [3]. Therefore, understanding the causative factors of diabetes is of great significance for early detection and treatment of diabetes. This article is intended to explore the diverse potential factors that could give rise to diabetes, assist people in evaluating the risk of developing diabetes, and enable them to take meaningful preventive actions in accordance with the risk level.

The factors implicated in the occurrence of diabetes were extremely complex and varied, involving multiple factors such as genetics, environment and lifestyle. Yuan et al. had found that disturbances in gut microbial function and metabolism are closely related to type 1 diabetes [4]. Scholars at home and abroad had discovered a certain correlation between diabetes and obesity [5], living environment [6], and eating habits [7]. In addition, Sanghera and Blackett discovered that genetic factors also exerted some influence on diabetes [8]. However, these documents have investigated relatively fewer factors related to diabetes and lack systematic and comprehensive statistical data. Therefore, this article centers on the following twelve factors (Sex, Age, BMI, Physical activity, Eating habits, Smoking, Mental health, Difficulty walking, Heart disease, Stroke, High cholesterol, High blood pressure) to study whether they will have an impact on the etiology of diabetes, and further construct a suitable model for examining the relationship between these factors and diabetes.

In a comparable direction, Xu et al. utilized logistic regression analysis to investigate the association between diabetes prevalence and regional variations. fitting different models such as age, gender and physical activity with inland provinces as the reference group and comparing them with coastal provinces [9]. However, the data selection in the literature is biased, with women and urban residents being oversampled. In addition, the sample data is incomplete and insufficient to provide a multivariate model. Yao et al. used a model of Meta-Analysis to fit the data and evaluate the dose-response relationship between dietary fiber intake and diabetes risk [10]. However, dietary fiber intake is usually assessed through food frequency questionnaires, which may have measurement errors, leading to inaccurate estimates of dietary fiber intake. Bao et al. used Cox proportional hazards regression models to evaluate the relationship between physical activity and sedentary behavior on the risk of developing type 2 diabetes (T2D) in gestational diabetes (GDM), and adjusted for other possible confounding factors, such as pre-pregnancy body mass index, gestational weight gain, and time since CDM diagnosis [11]. However, the model was limited in sample size and lacked sufficient statistical power to detect significant differences.

In conclusion, this article will employ the Probit regression model to analyze the influence of the factors on diabetes, i.e., whether they are causes of diabetes.

2. Methods

2.1 Data Source

This literature's data is sourced from the Kaggle website. It was compiled by Alex Teboul based on the existing BRFSS 2015 dataset on Kaggle, and was released and updated in 2021 for 253,680 people.

2.2 Indicator Selection

The data utilized in this paper encompasses a total of 253680 individuals, comprising those with and without diabetes of whom 111706 are male and 141974 are female. The ages of the patients span from 18 to 80 years and above. The data consists of 12 variables (Sex, Age, BMI, Physical activity, Eating habits, Smoking, Mental health, Difficulty walking, Heart disease, Stroke, High cholesterol, High blood pressure).

Table 1. Logogram and numbers of the 12 factors

Variables	Logogram	Number1	Diabetes1
Sex	x_1	253680	35346
Age	x_2	253680	35346
BMI	x_3	119286	25039
Physical activity	x_4	191920	22287
Eating habits	x_5	142712	17357
Smoking	x_6	112423	18317
Mental health	x_7	23370	4955
Difficulty walking	x_8	42675	13121
Heart disease	x_9	23893	7878

Stroke	x_{10}	10292	3268
High cholesterol	x_{11}	107591	23686
High blood pressure	x_{12}	108829	26604

For some indicators in the above table, the following is an explanation. Number 1 represents the quantity of people having the condition. Diabetes1 is the number of diabetes patients with the condition. BMI is the number of people with a BMI greater than 27. Physical activity is the number of people has physical activity in the past 30 days. Eating habits is the number of people consumes fruit and vegetables 1 or more times per day. Mental health is the number of people has poor mental health.

Table 1 presents the number of individuals and diabetes patients who have the condition. Table 1 illustrates the logogram for each factor as presented above. The dataset comprises 253,680 individuals, of whom 35,346 have diabetes.

Table 2. Gender distribution across age demographics

Age	[21-30]	[31-40]	[41-50]	[51-60]	[61-70]	[71-80]	[81-90]
Number2	13298	24946	35976	57146	65438	39513	17363
Female	6736	13787	20064	32274	36014	22577	10522
Male	6562	11159	15912	24872	29424	16936	6841
Diabetes2	218	940	2793	7351	12291	8544	3209

*Number2: The quantity of individuals within each age demographic.

**Diabetes2: The quantity of individuals within each age demographic diagnosed with diabetes.

Table 2 illustrates the number of females and males, and diabetes patients in different age groups. The bulk of participants in the data were aged 51-60, 61-70 and 71-80 years old.

2.3 Research Proposal

This article employs the Probit regression model, utilizing the presence of diabetes as the dependent variable (Y) and 12 components as independent variables (X), where 0 indicates absence and 1 indicates presence. This article utilizes SPSSPRO to investigate the association between the influence of X on Y, with the aim of finding the impact of 12 variables on the incidence of diabetes.

2.4 Model Principle

Probit regression is a statistical analysis method that is comparable to logistic regression and is employed to analyze data with categorical dependent variables. Deforming the Logistic mode, it is evident that:

$$P = \frac{e^{X\beta}}{1+e^{X\beta}} \tag{1}$$

The right-hand side of the aforementioned equation $(\frac{e^{X\beta}}{1+e^{X\beta}})$ is identical to the probability distribution function of the conventional growth distribution (also known as the logistic distribution). The Probit model implies that the probability distribution function on the right side is identical to the ordinary normal distribution:

$$P = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X\beta)^2}{2}} \tag{2}$$

2.5 Model Testing

The result of the likelihood ratio chi-square test of the model reveals that the significance level of the P value is 1%, which is significant at the level and the null hypothesis is rejected, so the model is valid.

Table 3. Classification evaluation indicators

Accuracy	Recall	Precision	F1	AUC
0.862	0.862	0.824	0.819	0.799

ISSN 2959-6157

It can be seen from Table 3 that in terms of classification evaluation indicators, the rates of accuracy and recall are both 0.862, indicating that the model performs well in predicting correct samples and predicting positive samples. The rate of precision is 0.824, which means that the proportion of positive samples predicted as positive samples is relatively high. The F1 is the harmonic mean of precision and recall, and its value is 0.819. AUC value is 0.799, which is close to 1, indicating good classification effect.

3. Results and Discussion

3.1 Data Processing

The study in this article demonstrates that there are numerous variables Influencing diabetes. As could be seen in Figure 1, the factors related to diabetes are BMI, Physical activity, Eating habits, Smoking, Mental health, Difficulty walking, Heart disease, Stroke, High cholesterol, and High blood pressure. There are 253,680 samples in total. Determine the proportion of diabetic samples in each group of samples, and categorize them according to various contributing variables to assess the likelihood of each risk leading to diabetes.

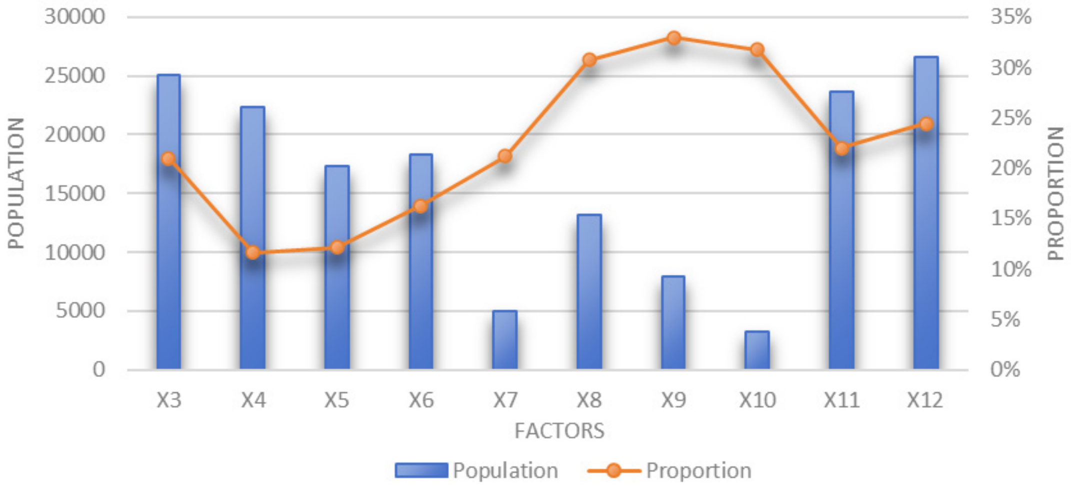

Fig. 1 Factors in the percentage of persons with diabetes

As could be seen in Figure 1, it pertains to the number among people suffering from diabetes with different conditions and the percentage of diabetes sufferers among the persons with these conditions. It can be seen that heart disease accounts for the highest proportion, up to 32.97%. Physical activity has the lowest percentage at 11.61%. Therefore, according to the data in the figure, heart disease has the largest proportion, leading to the biggest concealed hazards associated with diabetes.

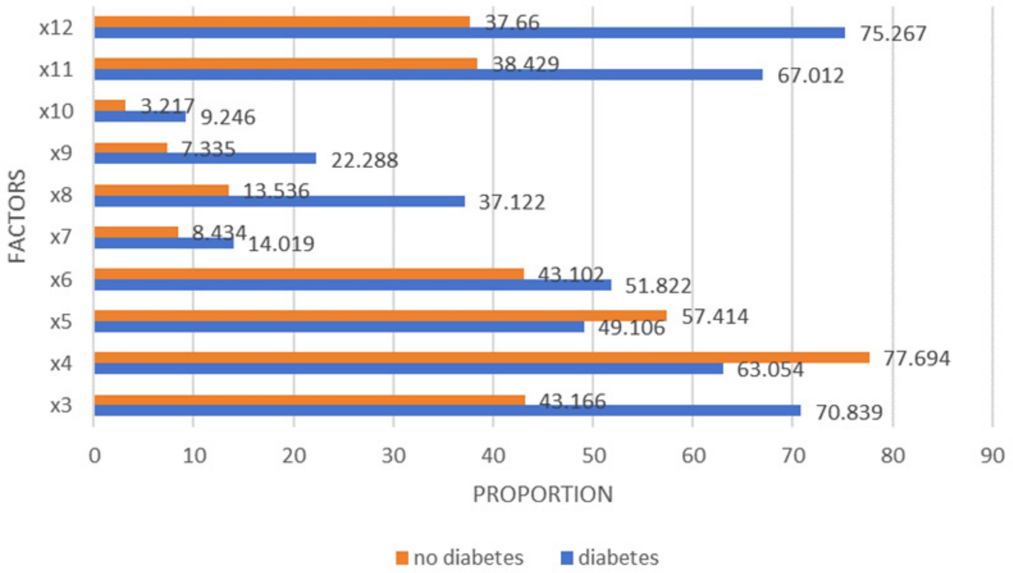

Fig. 2 Comparison between diabetic versus non-diabetes under various variables

In the figure 2, the comparison figure displays the difference among the percentage of obtaining diabetes and the percentage of no diabetes on the parameters x3-x12. The biggest difference is the factor of high blood pressure, which is nearly 40%. In addition, the remaining factors are also worth analyzing because the proportion of people with diabetes and those without diabetes is quite different among these factors, which leads to the greatest chance of developing diabetes.

3.2 Model Evaluation

In Figure 3, it can be seen that numerous elements that are potentially associated to the onset of diabetes into the model, including Sex, Age, BMI, Physical activity, Eating habits, Smoking, Mental health, Difficulty walking, Heart disease, Stroke, High cholesterol, and High blood pressure.

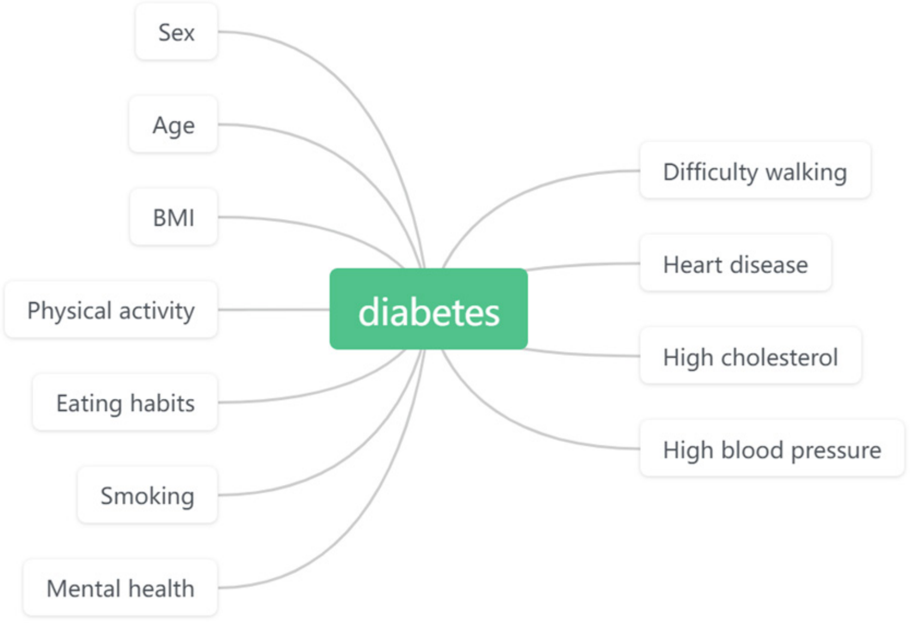

Fig. 3 Schematic diagram of related variables

Through calculation, this research obtained the final linear regression equation:

$$Probit(p) = -3.27 + 0.103x1 + \ldots + 0.474x12 \quad (3)$$

Where p denotes the likelihood of diabetes equal 1.

Table 4. Model results

Variables	coefficient	SE	z value	p-value	marginal effect	95% CI
x_1	0.103	0.007	14.557	0.000***	0.019	0.089~0.117
x_2	0.068	0.001	47.872	0.000***	0.012	0.065~0.071
x_3	0.037	0	76.047	0.000***	0.007	0.036~0.038
x_4	-0.105	0.008	-13.428	0.000***	-0.02	-0.12~-0.09
x_5	-0.083	0.007	-11.85	0.000***	-0.015	-0.097~-0.069
x_6	0.018	0.007	2.543	0.011**	0.003	0.004~0.032
x_7	0.114	0.011	10.038	0.000***	0.022	0.091~0.135
x_8	0.312	0.009	36.123	0.000***	0.063	0.295~0.329
x_9	0.265	0.01	26.227	0.000***	0.053	0.246~0.285
x_{10}	0.175	0.014	12.134	0.000***	0.034	0.147~0.204
x_{11}	0.338	0.007	47.08	0.000***	0.063	0.324~0.353
x_{12}	0.474	0.008	62.402	0.000***	0.088	0.459~0.489
Constant	-3.27	-	-	-	-	-

***: 1%, **: 5%, *: 10% significance level.

As can be seen from Table 4, the conclusion of the model is that since the p-values of the model are all less than 0.05, this research might further deduce that these variables are likely to be associated to diabetes. On the one hand, the p-value of Smoking is less than 0.05 but greater than 0.01, indicating that this factor has a certain impact on diabetes, but the impact is not very strong. On the other hand, the p values of Sex, Age, BMI, Physical activity, Eating habits, Mental health, Difficulty walking, Heart disease, Stroke, High cholesterol and High blood pressure are less than 0.01, which demonstrate that these characteristics have a particularly significant link with diabetes.

when it comes to impacting relationships, marginal effect value combined with regression coefficient value show major contributing factors: the change in diabetes will be 1.884%, 1.241%, 0.678%, -1.963%, -1.524%, 0.326%, 2.158%, 3.348% with one unit increase of Sex, Age, BMI, Physical activity, Eating habits, Smoking, Mental health and stroke separately. These factors are associated with the onset of diabetes, but the association is not particularly significant. In contrast, there may be other factors that play a more prominent role in the onset of diabetes. Additionally, the change in diabetes will be 6.284%, 5.339%, 6.27%, 8.809% with one unit increase of Difficulty walking, Heart disease, High cholesterol, High blood pressure. It is undeniable that there is a very strong link between these factors and the likelihood of developing diabetes, which means they have some reference and research value.

In contrast to past research, they preferred to employ single-factor specific analyses, confined to recognized potential variables for diabetes, for example The onset of diabetes may be related to human genes. In contrast, this study does not involve only one variable to determine whether it is related to the onset of diabetes, hence it can successfully prevent the mistake caused by not managing a specific variable. In addition, this experiment may broaden the concepts of future diabetes research, enable medical personnel recognize diabetes in a timely way, and decide new treatment possibilities.

4. Conclusion

This research collected diverse data and focused on variables that may be associated with the onset of diabetes. It is determined that the occurrence of diabetes is connected to Sex, Age, BMI, Smoking, Mental health, stroke, Difficulty walking, Heart disease, High cholesterol and High blood pressure. However, Physical activity and Eating habits can reduce the risk of diabetes.

It is clear that owing to the low quantity of data, the model may contain errors in addition to variables, and the sample does not represent people of every race and age. And because the information of people is collected through surveys during sampling, the sample itself may have errors, which may also reduce the accuracy of the results, but the research still has great advantages. On the one hand, this experiment chose Probit regression in multivariate linear regression, instead of using univariate analysis like many previous experiments, so that a multi-faceted analysis can be performed, making the results more comprehensive. On the other hand, a graphical method was used to visualize the differences in the proportions of various factors between diabetes and non-diabetes populations. This allows the experimental results to be visualized and the differences to be expressed more clearly and intuitively. Secondly, it has a certain positive effect on the treatment of diabetes. In addition to the known factors related to diabetes, there may be many such issues associated to diabetes that demand notice and discussion, such as mental health, difficulty walking, and other chronic diseases. Whether these factors are related to the occurrence of diabetes needs additional medical study, which implies that these results will indicate the way for future relevant research, which may assist with diagnose diabetes early and treat it as quickly as feasible, improving the quality of life of patients.

References

[1] Buse J B, Ginsberg H N, Bakris G L, Clark N G, Costa F, Eckel R, Fonseca V, Gerstein H C, Grundy S, Nesto R W, Pignone M P, Plutzky J, Porte D, Redberg R, Stitzel K F, Stone N J. Primary Prevention of Cardiovascular Diseases in People With Diabetes Mellitus: A Scientific Statement From the American Heart Association and the American Diabetes Association. Circulation, 2007, 115: 114-126.

[2] Doyle M. Cummings, Kari Kirian, George Howard, Virginia Howard, Ya Yuan, Paul Muntner, Brett Kissela, Nicole Redmond, Suzanne E. Judd, Monika M. Safford; Consequences of Comorbidity of Elevated Stress and/or Depressive Symptoms and Incident Cardiovascular Outcomes in Diabetes: Results From the REasons for Geographic And Racial Differences in Stroke (REGARDS) Study. Diabetes Care, 2016, 39 (1): 101-109.

[3] Wang L, Peng W, Zhao Z, et al. Prevalence and Treatment of Diabetes in China, 2013-2018. JAMA. 2021, 326(24): 2498-2506.

[4] Yuan X, Wang R, Han B, et al. Functional and metabolic alterations of gut microbiota in children with new-onset type 1 diabetes. Nature Communications, 2022, 13(1): 6356.

[5] Marie N, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study. The Lancet, 2013, 384: 766-781.

[6] Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Med, 2017, 15: 131.

[7] Mozaffarian D, Hao T, Rimm E B, Willett W C, Hu F B. Changes in Diet and Lifestyle and Long-Term Weight Gain in Women and Men. N Engl J Med, 2011, 364(25): 2392-2404.

[8] Sanghera D K, Blackett P R. Type 2 Diabetes Genetics: Beyond GWAS. J Diabetes Metab, 2012, 23(198): 6948.

[9] Xu S, Ming J, Xing Y, et al. Regional differences in diabetes prevalence and awareness between coastal and interior provinces in China: a population-based cross-sectional study. BMC Public Health, 2013, 13: 299.

[10] Yao B., Fang H., Xu W. et al. Dietary fiber intake and risk of type 2 diabetes: a dose–response analysis of prospective studies. Eur J Epidemiol, 2014, 29: 79-88.

[11] Bao W, Tobias D K, Bowers K, et al. Physical Activity and Sedentary Behaviors Associated With Risk of Progression From Gestational Diabetes Mellitus to Type 2 Diabetes Mellitus: A Prospective Cohort Study. JAMA Intern Med, 2014, 174(7): 1047–1055.