OPEN ACCESS Freely available online

IJCMB
International Journal of Clinical Medicine and Bioengineering

ISSN 2737-534X
Volume 2, Issue 2
https://www.iikii.com.sg/journal/IJCMB
International Journal of Clinical Medicine and Bioengineering

Article

Building Statistical Model for Predicting Risk of Diabetes

Te-Jen Su [1], Feng-Chun Lee [2] and Shih-Ming Wang [3,*]

[1] Department of Electronic Engineering, National Kaohsiung University of Sciences and Technology, Kaohsiung, Taiwan; sutj@nkust.edu.tw
[2] Department of Electronic Engineering, National Kaohsiung University of Sciences and Technology, Kaohsiung, Taiwan; alicelee0712@gmail.com
[3] Cheng Shiu University, Kaohsiung, Taiwan, Department of Computer Science and Information Engineering, Kaohsiung, Taiwan
* Correspondence: K1115@gcloud.csu.edu.tw

Received: Apr 19, 2022; **Accepted:** May 19, 2022; **Published:** Jun 30, 2022

Abstract: In recent years, diabetes has become one of the most common human diseases in the world, and is even the main cause of high mortality and economic losses, while timely diagnosis and prediction provide patients with appropriate methods for prevention and treatment. By using a logistic regression model, we tried to predict type 2 diabetes. The statistical analysis was conducted with SPSS for descriptive analysis of data, a chi-square test, and logistic regression analysis to predict the risk factor of diabetes. As the result, five main predictive factors were identified: waist circumference, family history, hypertension, cardiovascular disease, and age. The overall prediction rate of the logistic regression model for predicting diabetes was 80%. The research results help prevent the occurrence of diabetes or facilitate early treatment, reduce misdiagnosis and avoid wasting health care resources.

Keywords: Type 2 Diabetes, Risk factors, Logistic Regression

1. Introduction

With the improvement of modern living standards, the rate of diabetes is gradually increasing. The number of young patients with diabetes also increase, so diabetes has become an important global public health problem. In 2017, it was estimated that about 450 million people worldwide were diagnosed with diabetes, and about 1.37 million deaths were due to diabetes [1]. More than 100 million American adults suffer from diabetes. Diabetes was the seventh leading cause of death in the United States in 2020 [2]. According to estimates, the number of people suffering from diabetes in the world in 2025 will double that in 1995, with a prevalence rate of 5.4% [3], which shows the necessity of the disease prediction model for diabetes.

2. Methodology

The research mechanism was designed to construct a logistic regression model, establish the correlation between the outpatient data information and characteristics of diabetes, select the relevant variables based on the literature and follow-up statistical results, and perform a logistic regression analysis. Through the research, the real diabetes risk factors are found to establish a diabetes prediction model. The prediction model can screen pre-diabetes risk factors [4–6]. The patients of the hospitals in southern Taiwan from 2016 to 2017 were taken as the research objects, and the hospital visit data were collected for data mining and analysis to define the possible important risk factors related to diabetes for early discovery of the disease to enable subsequent treatment. The statistical analysis was made with SPSS to find the risk factor and predict their occurrence to cause diabetes [7]. Related factors with explanatory power were also identified for the assessment of an individual's risk of diabetes. The result provides a reference basis for doctors' diagnosis and decision-making as well as assistance for personal health management.

3. Logestic Regression

The data on diabetes-related factors were collected and sorted by frequency distribution analysis. The main variables included 14 items such as cardiovascular disease, waist circumference, age, family history, fasting blood glucose, taking antihypertensive drugs, and having suffered from kidney disease. 97.9% of patients had cardiovascular diseases, and 94.5% had excessive waist circumferences (Table 1).

Table 1. Diabetes-related variables and frequency allocation.

No.	Related Variables	Percentage
1	cardiovascular disease	97.9%
2	waist too thick	94.5%
3	aged over 45	91.9%
4	family history of diabetes	89.5%
5	Fasting blood sugar exceeds 125mgl	78.7%
6	blood pressure lowering drugs	68.4%
7	kidney disease	39.2%
8	high blood cholesterol	35.7%
9	asthma	33.7%
10	dialysis	25.9%
11	stroke	12.5%
12	paralysis of hands and feet	9.3%
13	can't see clearly	6.7%
14	slurred speech	2.9%

According to the analysis results, important references for diabetes included cardiovascular disease, excessive waist circumference, age, family history of diabetes, taking blood pressure drugs, kidney disease, high cholesterol, asthma, kidney dialysis, stroke, short-term hands and feet symptoms such as numbness or weakness, invisibility or blurred eyes, and slurred speech. The 14 symptoms were analyzed by chi-square test to find out the possible factors related to diabetes in Table 2.

Table 2. Correlation coefficients and significance.

Correlation Coefficient Range (Absolute Value)	Variable Correlation
0.80 to 1.00	very relevant
0.60 to 0.79	high correlation
0.40 to 0.59	moderately relevant
0.20 to 0.39	low correlation
0.01 to 0.19	very low correlation

Logistic regression was used to analyze variables one by one for their predictabilities of diabetes. The analysis results of the above variables are as follows. The overall prediction rate of the logistic regression model for predicting diabetes with waist circumference was 64%. The individual test value of excessive waist circumference was Wald = 44.32 at a significance of $p = 0.000$, indicating that there was a significant relationship between excessive waist circumference and diabetes. The patients with excessive waist circumference had the odds of developing diabetes Odds = Exp(−0.142 + 1.187). If the waist is too thick, the predictability of developing diabetes was 2.843, while that was 0.868 with a shorter waist circumference. The odds ratio of diabetes was 3.278, that is, the risk of developing diabetes was 3.27 times higher for patients with longer waist circumference (Table 3).

Table 3. Logistic regression analysis of waist too thick predicting diabetes.

Variable	B	S.E	Wald	Significance	Odds Ratio	Odds Ratio as a 95% CI	
						Lower Limit	Upper Limit
waist too thick	1.187	0.178	44.320	0.000	3.278***	2.311	4.65
constant	−0.142	0.119	1.416	0.234	0.868		

The overall prediction rate of the logistic regression model of family history for predicting diabetes was 64.5%, with a sensitivity of 63.4% and a certainty of 66.2% (Table 4).

Table 4. Family history predicts diabetes predictor rate.

	Observed Value		Predictive Value		
			Diabetes		Correct Percentage
			No	Yes	
Step 1	Diabetes	No	151	77	66.2
		Yes	128	222	63.4
	Overall Percentage				64.5

The individual test value of family history Wald = 46.935 (p = 0.000), indicating that family history was significantly related to diabetes. The odds of the patient with a family history was 2.883, and 0.848 for those without a family history. Thus, the odds ratio was 3.401, meaning that the patients with a family history of diabetes were 3.4 times more likely to develop diabetes than those without a family history (Table 5).

Table 5. Logistic regression analysis of family history predicting diabetes.

Variable	B	S.E	Wald	Significance	Odds Ratio	Odds Ratio as a 95% CI	
						Lower limit	Upper limit
family history	1.224	0.179	46.935	0.000	3.401***	2.396	4.827
constant	−0.165	0.120	1.892	0.169	0.848		

The overall prediction rate of the logistic regression model for predicting diabetes with hypertension was 70.1%, with a sensitivity of 64.4% and a certainty of 78.9% in Table 6.

Table 6. Hypertension predicts diabetes predictor rate.

	Observed Value		Predictive Value		
			Diabetes		Correct Percentage
			No	Yes	
Step 1	Diabetes	No	179	48	78.9
		Yes	124	224	64.4
	Overall Percentage				70.1

The high blood pressure was significantly related to diabetes (Wald = 93.424, p = 0.000), and the odds of hypertensive patients suffering from diabetes was 4.669, while non-hypertensive patients had 0.693. Thus, the odds ratio of the hypertensive patient to the non-hypertensive patient was 6.737, implying that the risk of developing diabetes was 6.73 times higher for the hypertensive patient (Table 7).

Table 7. Logistic regression analysis of hypertension predicting diabetes.

Variable	B	S.E	Wald	Significance	Odds Ratio	Odds Ratio as a 95% CI	
						Lower limit	Upper limit
hypertension	1.908	0.197	93.424	0.000	6.737***	4.576	9.918
constant	−0.367	0.117	9.872	0.002	0.693		

The overall prediction rate for the patient with cardiovascular disease was 65.8, the sensitivity was 55.9%, and the specificity was 81.2%. Cardiovascular disease was significantly related to diabetes (Wald = 70.3, p = 0.000), and the odds of the patient with cardiovascular disease was 4.618, while that of the patient without cardiovascular disease was 0.845. The odds ratio was 5.46, that is, the patient with cardiovascular disease had 5.46 times higher risk to develop diabetes than those without cardiovascular disease (Table 8).

Table 8. Logistic regression analysis of Cardiovascular disease predicting diabetes.

Variable	B	S.E	Wald	Significance	Odds Ratio	Odds Ratio as a 95% CI	
						Lower limit	Upper limit
Cardiovascular disease	1.698	0.203	70.300	0.000	5.464***	3.674	8.127
constant	−0.168	0.110	2.342	0.126	0.845		

Table 9 shows that the variables had significant correlations with diabetes, and hypertension showed the highest odds ratio between hypertensive and non-hypertensive patients.

Table 9. Logistic regression analysis of predicting variables for diabetes.

Variable	B	S.E	Wald	Significance	OR Value	Odds Ratio as a 95% CI	
						Lower limit	Upper limit
Waist Circumference	1.19	0.178	44.32	0.000	3.27***	2.31	4.65
Family History	1.22	0.179	46.94	0.000	3.40***	2.39	4.83
Hypertension	1.9	0.197	93.42	0.000	6.74***	4.58	9.99
Cardiovascular Disease	1.70	0.203	70.30	0.000	5.46***	3.67	8.13
Age	1.39	0.189	54.11	0.000	4.00***	2.77	5.79

In this study, logistic regression was used to establish a prediction model of diabetes risk assessment with the consideration of AUC accuracy, sensitivity, and specificity. The area under the ROC curve was used to evaluate the identification capabilities of the various models (Swets et al). When the area is under the curve of (AUC) ≥ 0.7, the model has the diagnostic ability. When the variables were analyzed, the AUC (areas under the ROC curve) of the model was 0.655, and the overall prediction rate of the logistic regression model for diabetes was 80% (Fig. 1, Table 10).

Table 10. Predictive rates of all variables predicting diabetes.

Observed Value		Predictive Value		
		Diabetes		Correct Percentage
		No	Yes	
Diabetes	No	153	66	69.9%
	Yes	46	296	86.5%
Overall Percentage				80%

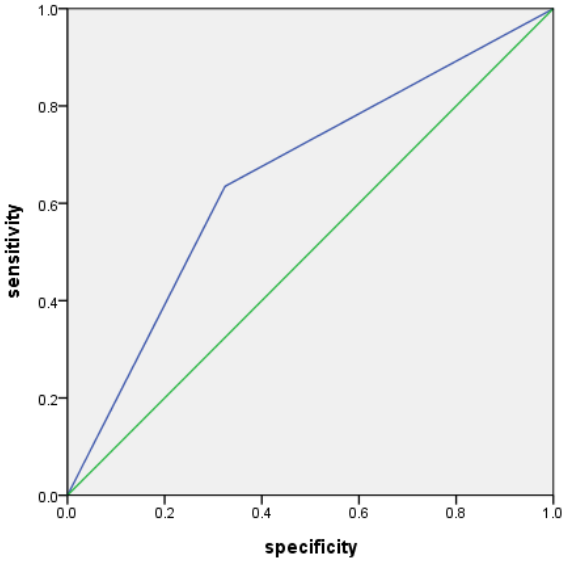

Fig. 1. ROC curve of logistic regression.

4. Conclusion

Different predictive models have also been developed to predict the occurrence of diseases by using logistic regression. For example, Kazemnejad (2010) used neural networks and logistic regression to detect diabetes blood glucose tolerance and pointed out that the three basic data of age, gender, and BMI were all related to the risk factors of diabetes. When including age, hypertension, cardiovascular disease, family history, and other variables, the logistic regression prediction rate reached 80%. The research results of this study help people understand whether they have the possibility of having diabetes in the future, check their health status, and correct bad habits. A reference to professionals and doctors in diagnoses and analyses is also provided from the result to prevent the occurrence of diabetes or detect it as soon as possible for early treatments or tracking.

Author Contributions: Contributions are listed as follows: T.-J.S., Conceptualization, Methodology, Supervision; F.-C.L., Data curation, Formal analysis; S.-M.W., Data curation, Writing - original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.D.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract*. **2018**, *138*, 271–281.
2. CDC. Centers for Disease Control and Prevention and Others. Centers for Disease Control and Prevention, US Department of Health and Human Services; National Diabetes Statistics Report 12–15, Atlanta, GA, USA, 2020.
3. Chang, C.J. Reflection on the prevalence of diabetes in Taiwan and its related factors Newsletter. *Diabetes Care Foundation* **2002**, *3*, 4.
4. Schwarz, P.E.; Li, J.; Lindstrom, J.; Tuomilehto, J. Tools for predicting the risk of type 2 diabetes in daily practice. *Horm. Metab. Res*., **2009**, *41*, 86–97.
5. Yu, W.; Liu, T.; Valdez, R.; Gwinn, M.; Khoury, M.J. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Med. Inform. Decis. Mak*. **2010**, *10*, 16.
6. Naz, H.; Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J. Diabetes Metab. Disord*. **2020**, *19*, 391–403.
7. Swets, J.A. Measuring the Accuracy of Diagnostic Systems. *Science* **1998**, *240*, 1285–1293.

Publisher's Note: IIKII stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 The Author(s). Published with license by IIKII, Singapore. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.