International Journal of Engineering & Technology, 7 (4.30) (2018) 355-358

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Study on Significant Predictors for Prediction of Undiagnosed T2DM Using Binary Logistic Regression Model

S. S. N. Zainal[1]*, M. J. Masnan[1], A. Ahmed[1], N. A. M. Amin[1] and M. I. Omar @Ye Htut[2]

[1]*Institute of Engineering Mathematics, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau,Perlis, Malaysia*
[2]*University Health Centre, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia.*
**Corresponding author Email: sitisalsabilahnabilah@yahoo.com*

Abstract

Type 2 Diabetes Mellitus (T2DM) is a chronic disease that can cause premature deaths worldwide. Malaysia is one of the many countries that facing this serious epidemic. The World Health Organization (WHO) has also estimated that Malaysia would have 2.8 million people having T2DM disease in 2030. This study aims to identify significant predictors for prediction of undiagnosed T2DM patients in one of the highest prevalence states of T2DM. Binary logistic regression model proposed to predict the presence of T2DM among undiagnosed respondents. The selection of significant predictors using univariate, multivariate and backward stepwise selection was implemented in this study. The study concludes that four predictors were found significant for prediction of undiagnosed T2DM patients.

Keywords: *Binary Logistic Regression model; Significant predictors; undiagnosed T2DM*

1. Introduction

The prevalence of Type 2 Diabetes Mellitus (T2DM) is rapidly increasing and can affect the health of people worldwide [1][2]. According to [3], there are 2 major types of diabetes mellitus which are Type 1 diabetes mellitus (T1DM) and T2DM disease. T1DM disease is defined as insulin-dependent where the body does not enough produce insulin while T2DM is defined as non-insulin dependent where the body ineffectively uses insulin. Besides, diabetes mellitus disease can occur among the pregnancy woman that called as gestational diabetes mellitus (GDM) disease. GDM disease can be characterized as a temporary condition that occurs during pregnancy and can put at risk as T2DM disease for a long term. However, T2DM disease is one of the main health care problem that threaten to reach pandemic by 2030.

Latest finding from the National Health & Morbidity Survey (NHMS II) in 2015, the prevalence of total diabetes mellitus has risen to 17.5% where 8.3% of prevalence were among individual with known diabetes mellitus and 9.2% of prevalence among individual with undiagnosed diabetes mellitus. The prevalence of diabetes mellitus is referred as the percentage of the number of diabetes mellitus cases in a population. Figure 1 shows the trend of total prevalence for diabetes mellitus aged 18 years and above from the NHMS report for five alternate years [4][5][6]. Those report revealed that the prevalence of diabetes mellitus has increased at 11.6% in 2006 and 17.5% in 2015. Undiagnosed diabetes mellitus become alarming cases where it started to increase slowly at 0.8% difference from known diabetes mellitus in 2011 and continually increase in 2015. According to the newspaper report [7], 50% from 3.6 million Malaysians i.e. about 1.8 million people are believed to have the T2DM disease where they are not yet diagnosed and never have health screening procedure.

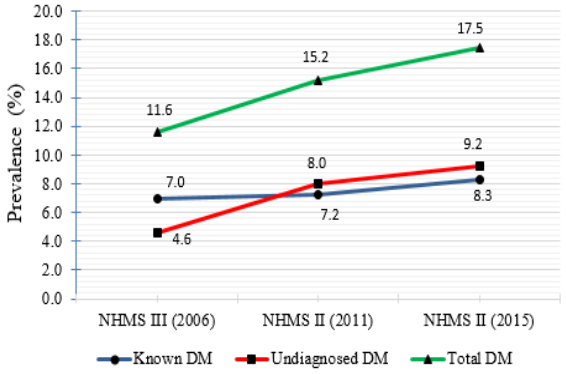

Fig. 1: Trend of total prevalence for diabetes mellitus aged 18 years and above [4][5][6].

In primary clinical care, binary logistic regression model has been applied to investigate related predictors associated with diabetes mellitus. There are twelve predictors have been identified by [8] in predicting the diabetes or prediabetes mellitus which include age, family history of diabetes, marital status, educational level, work stress, duration of sleep, physical activity, gender, eating fish, drinking coffee, preference for salty food, and body mass index (BMI). While, some significant predictors suggested by [9] might be indicated to the chances for having diabetes mellitus disease were age, BMI, hypertension, dyslipidemia, impaired fasting glucose and impaired glucose tolerance. Another researcher found that the significant predictor of BMI category (i.e. who are obese about 1.5 to 5 times higher than individuals have normal BMI) is strongly associated to the risk of T2DM disease [10]. Thus, this research aims to identify the significant predictors in predicting undiagnosed T2DM patients using binary logistic regression model.

2. Study Design

This study used the secondary data from a survey by [6]. This report was a household survey conducted by Institute of Public Health (IPH) once for every 5 years. Since the focus of this study is Perlis, only data collected from that state was applied in this research. Among the collected data in the NHMS survey were some sociodemographic variables, clinical assessment and lifestyle risk factors. The area of study that involved in this research is divided into two parts; rural and urban. Kangar area is categorized as urban while other part of Perlis is considered rural areas.

The target population is referred to NHMS survey [6] that involved all individuals residing in the non-institutional living quarters (LQs). Institutional population such as those staying in hotel, hospitals, hostel, etc. were excluded from the survey. A total of 1814 respondent in Perlis state is included in this study. Based on the sampling frame, the two-stage stratified sampling proportionate to the population size was applied in the data collection throughout the national level which includes all Federal Territory and states in Malaysia. The geographical areas in Malaysia were divided into Enumeration Blocks (EBs). The NHMS survey covered both urban and rural areas for every state in Malaysia. A total of 869 EBs sample by state is selected from the total EBs in Malaysia, where 536 EBs urban and 333 EBs rural area. A total of 50 EBs sample for Perlis state is divided in 25 EBs for urban and 25 EBs for rural was randomly selected.

3. Binary Logistic Regression Model

Binary logistic regression model is widely used to analyze data as statistical method for binary response variables (0 and 1) with one or more independent variables. This model is less restrictive than other techniques [11] and increasingly applied in various field such as medical, health, social sciences and education research. The potential of logistic regression as predictive model is inevitable for research in detecting, screening and predicting [1], [12], [13] the undiagnosed T2DM patients based on the collected information. In statistics, binary logistic regression is usually used to find the best model fitting and to describe the association between the dependent and independent variables [14].

Moreover, binary logistic regression model has been applied by [15] to investigate the association between dichotomous dependent variables and one or more independent variables in the nursing domain. This model was also applied by [15] to estimate the influence of some accident factors on severity as the dependent variable with two categories of fatal or injury. In this paper, it considered the general binary logistic regression model with multiple of explanatory variables. Let k predictors for a binary response variables Y denote by $X_1, X_2, \ldots X_k$. Thus, $\pi(x)$ represents the conditional probability that $P(Y=1|x)$ and $1-\pi(x)$ represent the conditional probability that $P(Y=0|x)$ These probabilities are written in the following equations [14]:

$$\pi(x) = P(Y=1|X_1, X_2, \ldots X_k) \quad (1)$$

$$1-\pi(x) = P(Y=0|X_1, X_2, \ldots X_k) \quad (2)$$

The model for the log odds is:

$$\text{logit}(\pi(x)) = \ln\left(\frac{P(Y=1|X_1, X_2, \ldots X_k)}{P(Y=0|X_1, X_2, \ldots X_k)}\right) \quad (3)$$

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1, + \beta_2 x_2, \ldots \beta_n x_k + \varepsilon \quad (4)$$

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon \quad (5)$$

$$\pi(x) = P(Y=1|X_1, X_2, \ldots X_k) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon}} \quad (6)$$

The parameter β_j refers to the effect of X_j on the log odds that Y=1, controlling the other predictors. For example, exp (β_j) is the multiplicative effect on the odds of a one-unit increase in X_j, at fixed levels of the other predictors.

3.1. Predictors Associated with Undiagnosed T2DM

The NHMS survey consists of more than twenty parts that related to health and morbidity of respondents. However, only three parts which are sociodemographic, clinical assessments and lifestyle risk factors were included for this study. From these, about twelve different predictors were used for identifying which significant predictor that can be selected for prediction of undiagnosed T2DM. Based on Table 1, T2DM disease (dependent variable) is measured based on the presence or absence of the T2DM among the respondents. Hence, let Y = 1 represent for 'presence of T2DM' and Y = 0 represent for 'absence of T2DM'.

Table 1: Description of variables involved in the study.

Variables	Description	Type of variables
Dependent: T2DM disease	Undiagnosed T2DM disease in population of Perlis state	Binary 1: Presence of T2DM 0: Absence of T2DM
Independent:		
Gender	Male and female	Nominal
Age	Between 18 until 75 years and above	Scale
Ethnicity	Malay, Chinese, Indian, other Bumiputra and other	Nominal
Marital Status	Never married, married and divorcee	Nominal
Citizen status	Malaysian, non-Malaysian	Nominal
Education level	No formal education, primary, secondary, unclassified	Nominal
Occupation status	Government/semi government, private employee, self-employed, unpaid worker, retiree	Nominal
Physical activity	Active, not active	Nominal
Smoking status	Daily, occasional, former, never smoker	Nominal
Current drinker	Yes, no	Nominal
Height	Height of respondent	Scale
Weight	Weight of respondent	Scale

The binary logistic regression model is the most popular prediction model that has been used in field of health science [17]. A predictive equation using logistic regression analysis that incorporated patients' medical history, blood pressure, blood test, urine test and some demographic characteristics such as age, sex, BMI and ethnicity as independent variables for prediction of undiagnosed diabetes was applied by [13][18]. For this study, the predictive equation using binary logistic regression with multiple different predictors are formed.

$$P(Y=1) = \frac{e^Z}{1+e^Z} \quad (7)$$

where;

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...\beta_k X_k + \varepsilon \quad (8)$$

Thus, considering all twelve different predictors included in this study, the estimated full model of binary logistic regression model is as follows,

$$\begin{aligned}\hat{Z}(1/0) &= \hat{\beta}_0 + \hat{\beta}_1 \text{gender} + \hat{\beta}_2 age + \hat{\beta}_3 ethnicity \\ &+ \hat{\beta}_4 citizen + \hat{\beta}_5 marital\_status + \hat{\beta}_6 education + \hat{\beta}_7 occupation \\ &+ \hat{\beta}_8 physical\_activity + \hat{\beta}_9 smoking\_status + \hat{\beta}_{10} current\_drin\text{ker} \\ &+ \hat{\beta}_{11} height + \hat{\beta}_{12} weight\end{aligned} \quad (9)$$

3.2. Selection of Predictors and Statistical Data Analysis

Before performing the binary logistic regression analysis, the multicollinearity of the predictor is tested by checking the variance inflation factor (VIF) for more than 10 and tolerance value either less or equal to 0.1 and. The multicollinearity is checked to fulfill the assumption of the binary logistic regression model. Selection of significant predictor is implemented in two phases. The first phase is the univariate selection and followed by the multivariate selection for the second phase.

For the univariate selection, each predictor will be used to develop a univariate binary logistic regression model using twelve different predictors. The significant value of each predictor will be evaluated. Any predictors with significant value less than 0.05 are considered significant ($p < 0.05$). Once all the significant predictors are identified, the multivariate selection is carried out. Then, all the significant predictors will be included in the full model. However, all the significant predictors included may not necessarily significantly multivariate. Thus, another predictor is implemented in backward stepwise selection.

In this case, backward stepwise predictor selection is applied that begins with a full model and gradually eliminate any predictor that is not significant to the model. Any predictor that has been eliminated before, can be reselected based on the contribution to the model. The best-reduced model can be obtained based on the significant value of less than 0.05 ($p < 0.05$). The final binary logistic regression model by backward stepwise selection could be minimized the multicollinearity problem and to resolve the overfitting.

Apart from that, the -2 log likelihoods, Cox and Snell R^2 and Negelkerke R^2 are also applied in the binary logistic regression model to assess and evaluate the overall model. For goodness of fit, the Hosmer and Lemeshow's test also is used to check the fitness of the model and the contribution of the predictor to the model. The analysis was implemented using the IBM Statistical Package for the Social Science (SPSS) Version 20.

4. Result and Discussion

The result from univariate selection of undiagnosed T2DM is shown in Table 2. The significant predictors are selected based on $p < 0.05$ should be included in the model and its contribution to the predictive ability of the model. According to Table 2, it is obvious that seven significant predictors are significant. The seven significant predictors are gender, age, marital status, education, occupation, smoking status and weight. After performing the univariate and multivariate selection, all the seven significant predictors were tested again in backward stepwise selection to identify the best significant predictors in predicting the presence or absence of T2DM among the undiagnosed respondents.

Table 3 shows the percentage of correct predicted value for dependent variable based on the full model of binary logistic regression. About 830 undiagnosed T2DM respondents observed that absence of T2DM are correctly predicted to be 'No' while 127 undiagnosed T2DM respondents observed absence of T2DM but predicted to be '1' are not correctly predicted. By the full model, the total percentage of undiagnosed T2DM cases are correctly predicted shows the accuracy of 86.7% compare to the null model.

Table 2: Univariate selection of undiagnosed T2DM.

Variables	Percentage correct of predicted value (%)	β_0	βi	$p < .05$
X_1 – gender	87.4	-1.382	-0.367	0.028
X_2 – age	87.4	-2.849	0.018	0.000
X_3 – ethnicity	87.4	-1.742	-0.150	0.148
X_4 – citizen	87.4	-0.660	-1.259	0.219
X_5 – marital status	87.4	-2.635	0.347	0.029
X_6 – education	87.4	-1.192	-0.280	0.008
X_7 – occupation	87.0	-2.751	0.281	0.001
X_8 – physical activity	87.3	-1.973	0.034	0.853
X_9 – smoking status	87.4	-1.384	-0.169	0.008
X_{10} – current drinker	80.0	-21.203	19.986	0.999
X_{11} – height	87.2	-3.890	0.012	1.012
X_{12} – weight	87.2	-2.881	0.014	0.007

Table 3: Classification table: the percentage correct predicted value of dependent variable for full model.

Observed		Predicted		Percentage correct (%)
		No	Yes	
Undiagnosed T2DM	No	830	0	100
	Yes	127	0	0
Total percentage (%)		-	-	86.7

Table 4 presents the statistical test of individual predictors using backward stepwise selection. These four significant predictors were identified in backward stepwise selection with the significant value ($p < 0.05$). The four identified significant predictors would be included in the full model of binary logistic regression for prediction of undiagnosed T2DM.

Table 5 shows the overall evaluation and goodness of fit test statistics for the full model using the four selected significant predictors. The full model is predicted the outcome perfectly with the Cox and Snell R^2 is less than 1. Moreover, the Negelkerke R^2 is preferred to use which suggested that the model explains 6% of the variation in the outcome.

Table 4: Statistical test of individual predictors using backward stepwise selection.

Variables	β_0	β_i	$p < .05$
X_2 – age	-3.868	0.023	0.003
X_7 – occupation		0.202	0.041
X_1 – gender		-0.408	0.048
X_{12} – weight		0.013	0.050

Table 5: Overall evaluation and goodness of fit statistics.

Model summary		
-2 Log likelihoods	Cox and Snell R^2	Negelkerke R^2
717.464	0.033	0.060
Goodness of fit: Hosmer & Lemeshow test		
χ^2	df	$p > .05$
9.843	8	0.276

For assessing the goodness of fit of the model, the number of likelihood less than 1 is explained how good the model fits to the data. If the model fits perfectly, the likelihood = 1 and -2 loglikelihood (-2LL) = 0. In our model, -2LL = 717.464 shows the value is far from zero that it is difficult to make a statement. However, the Hosmer and Lemeshow's test were checked and indicated as a

good fit to the model ($p > 0.05$). Thus, the result from the Hosmer and Lemeshow's test shows that the full model was fit well to the data with $\chi^2(8) = 9.843$ ($p > 0.05$).

5. Conclusion

This study intended to identify the significant predictors in predicting of undiagnosed T2DM in one of the highest prevalence of diagnosed T2DM. Moreover, this study is very important for the healthcare authorities to overcome the problem of undiagnosed T2DM and help in providing care for the people who newly diagnosed. As a result, four significant predictors (age, gender, occupation and weight) are identified and useful for prediction of undiagnosed T2DM patients. The four significant predictors were identified are associated with undiagnosed T2DM will be recommended for the future research.

Acknowledgement

The author would like to acknowledge the support from the Fundamental Research Grant Scheme (FRGS) under a grant number of FRGS/1/2016/SKK06/UNIMAP/02/1 from the Ministry of Higher Education Malaysia. The author also would like to acknowledge the support from the Science fund under a grant number of 9003-00592 from the Ministry of Science, Technology & Innovation.
The author also shows a gratitude to the the Director of Institute of Public Health, Ministry of Health (MOH) and Medical Research and Ethics Committee (MREC), Institute for Health Management, Kuala Lumpur (reference no. KKM/NIHSEC/ P16-1856) who provided the approval and permit for collecting data.

References

[1] Bagheri N et al (2014), Undiagnosed diabetes from cross-sectional GP practice data: an approach to identify communities with high likelihood of undiagnosed diabetes. *BMJ Open*, vol. 4, no. 7, pp. 1–10.
[2] Zainal SSN, Masnan MJ, Amin NAM & Mohamed N (2017), Spatial analysis for prevalence of type 2 diabetes mellitus - A state investigation, *AIP Conference Proceedings*, vol. 1905, pp. 1–6.
[3] World Health Organization (WHO), *Global Report on Diabetes*, International Standard Book Number, ISBN, France:WHO (2016).
[4] *The Third National Health and Morbidity Survey 2006 (NHMS III): Diabetes Mellitus*, Malaysia: Institute for Public Health (2008).
[5] *National Health and Morbidty Survey 2011 (NHMS 2011): Non-Communicable Diseases*, Malaysia: Institute for Public Health (2011).
[6] *National Health and Morbidity Survey 2015 (NHMS 2015): Communicable Diseases, Risk Factors & Other Health Problems*, Malaysia: Institute for Public Health (2015).
[7] "3.6 juta rakyat Malaysia hidap diabetes," *Berita Harian Online*, para. 2, November 15, 2017. [Online]. Available: https://www.bharian.com.my. [Accessed Dec. 20, 2017].
[8] Meng XH, Huang YX, Rao DP, Zhang Q & Liu Q (2013), Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J. Med. Sci.*, vol. 29, no. 2, pp. 93–99.
[9] Bonora E *et al.* (2004), Population-Based Incidence Rates and Risk Factors for Type 2 Diabetes in White Individuals: The Bruneck Study," vol. 53, no. 7, pp. 1782-1789.
[10] Ganz ML, Wintfeld N, Li Q, Alas V, Langer J & Hammer M (2014), The association of body mass index with the risk of type 2 diabetes: a case–control study nested in an electronic health records system in the United States. *Diabetol. Metab. Syndr.*, vol. 6, no. 1, p. 50.
[11] Tabachnick BG & Fidell LS, *Using Multivariate Statistics*, 5th ed. United States of America: Person Education Inc (2007).
[12] Dugee O *et al.* (2015), Adapting existing diabetes risk scores for an Asian population: a risk score for detecting undiagnosed diabetes in the Mongolian population., *BMC Public Health*, vol. 15, no. 1, pp. 1–9.
[13] Tabaei BP & Herman WH (2002), A Multivariate Logistic Regression Equation to Screen for Diabetes, *Diabetes Care*, vol. 25, no. 11, pp. 1–5.
[14] Hosmer DW & Lemeshow S (2000), *Applied Logistic Regression*, 2nd ed., no. 1, United States of America: John Wiley & Sons Inc.
[15] Park HA (2013), An introduction to Logistic Regression: from basic concepts to interpretation with particular attention to nursing domain. *J. Korean Acad. Nurs.*, vol. 43, no. 2, pp. 154.
[16] Al-Ghamdi AS (1997), Using Logistic Regression for estimating the influence of some accident factors on severity," *Accid. Anal. Prev.*, vol. 34, pp. 729–741.
[17] Pramono LA *et al.* (2010), Prevalence and predictors of undiagnosed diabetes mellitus in Indonesia. *Acta Med. Indones.*, vol. 42, no. 4, pp. 216–223.
[18] Tetrault JM, Sauler M, Wells CK & Concato J (2008), Reporting of multivariable methods in the medical literature. *J. Investig. Med.*, vol. 56, no. 7, pp. 954–957.