J. Appl. Environ. Biol. Sci., 7(1)198-202, 2017
© 2017, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental
and Biological Sciences
www.textroad.com

Propensity Score Matching Sports Activity of Genesis Diabetes Mellitus Using Logistic Regression

Samsinar[1*], RR Soenarnatalina [1], Arief Wibowo[1], Bambang Widjanarko Otok[2]

[1] Faculty of Public Health, Airlangga University in Surabaya
[2] Department of Statistic, Sepuluh Nopember Institute of Technology (ITS), Surabaya

Received: September 28, 2016
Accepted: December 3, 2016

ABSTRACT

Propensity score is the conditional probability to get a specific treatment based on the observed kovariat. Method is used to reduce the bias in the estimation of treatment effects on the data is the observation due to confounding factors. If the treatment is binary form, logistic regression model is one of the estimation of the value of the propensity score is exactly because easily in the estimation and interpretation. The purpose of this research is to examine the estimates of the propensity score match based on binary logistic regression in the case of diabetes mellitus (DM). The results of the study showed that the Genesis diabetes mellitus influenced by sports activity, triglyceride, hypertension, and obesity. While the sports activity was influenced by the triglyceride, hypertension, and obesity. So sports activity as a variable counfonding, which shows that the patient was cushioned normal triglyceride levels will sports activity of 9.335 times than that have high triglyceride. Furthermore, a person who is not hypertension will sports activity of 4.531 times than that have hypertension, and a person who does not obesity will sports activity of 14.451 times than that of obesity. The test results showed that mathing propensity score sports activity based covariat (triglycerides, hypertension and obesity) influence the genesis of diabetes mellitus.

KEYWORDS*: Confounding , DM , Propensity Score match, Logistic Regression*

INTRODUCTION

In the last few years of the disease is not transmitted to the attention in the field of health because it is one of the causes of the increasing number of death. One of the disease prevalence transmitted not high enough in the world is Diabetes Mellitus (DM). The International Diabetes Federation (IDF) stated that in 2005 there were 200 million (5.1%) people with diabetes (diabetes) in the world, and allegedly 20 years later namely 2025 will be increased to 333 million (6.3%). Countries such as India and China, the United States, Japan, Indonesia, Pakistan, Bangladesh, Italy, Russia, and Brazil is a big ten country with a total population of diabetes most votes [1].

According to the data from the Central Statistics Agency Indonesia (2003) estimated population of Indonesia over the age of 20 years is USD 133 million. With the prevalence of Diabetes Mellitus in urban areas of 14.7% and the rural area of 7.2%, then expected in 2003 there were people with diabetes some 8.2 million in urban areas and 5.5 million in the rural area. Next, based on the pattern of the increase population, is expected in the year 2030 there will be 194 million people over the age of 20 years and assuming that the prevalence of Diabetes Mellitus on urban (14.7%) and rural (7.2%) it is estimated that there are 12 million people with diabetes in urban areas and 8.1 million in the rural area. A very large number of and is a burden that is difficult to be handled by the specialist doctor/ subspecialty even by all existing health workers. Remember that the DM will give impact to the quality of human resources and increased costs of health large enough [2].

The Results of Health Research (Riskesdas) 2013 shows that the prevalence of Diabetes Mellitus in Indonesia to the age above 15 years at 2.1%., in the rural areas of 2.2%, and in urban areas that reach 1.9%, while in the West Sulawesi Province 2.2 percent [3]. According to [3] of Polewali Mandar is highest in the West Sulawesi Province of 4.0%, and on 2015 based on the report SP2TP of Polewali Mandar number of cases of Diabetes Mellitus as much as 4668 cases, while for the working area of the clinic Gardens sari as much as 245 cases.

Research about the risk factors of Diabetes Mellitus concluded the relationship between peripheral obesity and insulin resistance is statistically significant (p<0.001), where the percentage of insulin resistance is higher in patients are obese than non-therapies is increased (OR = 10.8). The relationship between blood pressure and insulin resistance is statistically significant (p<0.001), where the percentage of insulin resistance is higher on the subject of the non-normotensi compared the subject normotensi (OR = 5.1). The relationship between dyslipidemia and insulin resistance is statistically significant (p<0.001), where the percentage of insulin resistance is higher on the subject to dyslipidemia than non-dislipidaemia (OR = 4.60). Peripheral obesity and dyslipidemia proved instrumental to genesis insulin resistance by 32.6% [4]. Another study concluded the risk factors that can be modified to Diabetes Mellitus shows that all the variables have a meaningful relationship to Diabetes Mellitus with the result of the relationship between IMT against Diabetes Mellitus (P = 0,000 and OR = 5.2), the relationship of hypertension Diabetes Mellitus (P = 0,028 and OR = 2.26), the relationship of physical activity against Diabetes Mellitus (P = 0,024 and OR 2.37), the relationship between carbohydrate of Diabetes Mellitus (p 0,007 and OR = 2.99) and the relationship of fiber against Diabetes Mellitus (P = 0,009 and OR 10.2) while from the results of the logistic regression multivariat compounds that most influential of Diabetes Mellitus is IMT consecutive patients 23 kg/m² result with p = 0,000 [5].

***Corresponding author:** Samsinar, Faculty of Public Health, Airlangga University in Surabaya
emails: sinariqepid@gmail.com

Some research about methods based on propensity score, namely: D'Agostino [6] use PS Match (PSM) and PS Stratification (PSS) to reduce the bias in the comparison of treatment groups and control for the case of drugs, Austin [7] compare 4 score propensity method (community participation, PSS, covariate adjustment of PS and PS Weighting) to reduce the systematic differences between the treatment groups and.control on the case of Smoking obtained the conclusion that the methods of community participation is the best method.

Observation research with logistic regression method without attention to the possibility of a strong combination between the factors that affect the genesis diabetes mellitus, whereas the existence of the interaction between these factors can cause the confounding variable which resulted in the conclusion is inaccurate. Given the higher prevalence of Genesis Diabetes Mellitus, then need to examine further the factors that affect Diabetes Mellitus using propensity score match on binary logistic regression to reduce the bias of the variables confounding sports activity.

LITERATURE REVIEW

Propensity score according to [8] is conditional probability on the subject to get a specific treatment by involving kovariat observed. In the case of the random experiment, treatment status Z_i independent without conditions on the response variable Y_i. For observation data non-random, independent could not be achieved because of the confounding factors, **X** namely kovariat that affect both the treatment and the response variable. As a result, a simple comparison of the results of the average between the treatment and control unit will not generally disclose kausal effects. However, conditional independency of the response variable and the status of the treatment can be determined by adjusting for vectors kovariat **X,** then the estimation of appropriate kausal treatment effects can be obtained. The benefits from the propensity score compared with multivariabel adjustment is the separation of confounding factors and the analysis of the influence of the treatment of [9].

Introduces the propensity score [10], as a conditional probability depends on the treatment unit ($Z_i=1$) compared to control unit ($Z_i=0$) with the observed kovariat vector $\boldsymbol{x_i}$.

$$e(x_i) = \mathrm{P}(Z_i = 1 | X_i = x_i) \qquad 1)$$

Generally, and independent conditional with $e(x_i)$,

$$P(x_i, Z_i | e(x_i)) = \mathrm{P}(x_i | e(x_i)) \, \mathrm{P}(Z_i | x_i). \qquad 2)$$

To prove equality 2), just indicated that $\mathrm{P}(Z_i = 1 | x_i) = \mathrm{P}(Z_i = 1 | e(x_i))$. By Definition $\mathrm{P}(Z_i = 1 | x_i) = e(x_i)$. the assumption that is given to X to Z_i independent:

$$\mathrm{P}(z_1, z_2, \cdots, z_N | x_1, x_2, \cdots, x_N) = \prod_{i=1}^{n} e(x_i)^{z_i} \{1 - e(x_i)\}^{1-z_i} \qquad 3)$$

ATT can be estimated by directly comparing the results between the subject treated and not treated in the sample matches [11]. If the result is a continuous scale, effects of medications can be estimated as the difference between the mean results for the subject with the treatment and the results of the average for the subjects that are not given preferential treatment on the samples matched [8]. If the result is dikotomis, treatment effects can be estimated as the difference between the proportion of the subjects experienced in each of the two groups (equal treatment and control) in samples resolved. Binary results ATT also can be described using relative risk [7][8].

Propensity Score Match done with Matching the treatment unit and controls with the values of nearly a score trend, and other kovariat allows, and ignore all the units are not suitable [12]. This is mainly used to compare the two groups of subjects, but can be applied to the analysis of more than two groups. The development of the method *propensity score match* done by [13] that focus on the selection bias, with emphasis on making the conclusion with non research design and develop random differences in the approach has application to community participation.

According to [14] [15], binary logistic regression model is distric comparison of the likelihood of an event/success (□) and the likelihood of failed events (1-□). Specific form of logistic regression model with *p* variables predictors revealed in the equation

$$\pi(\mathbf{x}) = \frac{\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)} \qquad 4)$$

Similarities (4) may be simplified as follows, $g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = \mathbf{x}^T \boldsymbol{\beta}$ 5)

with $\pi(\mathbf{x})$ is the success probability, $1 - \pi(\mathbf{x})$ is the probability to fail, β_k is the parameters of linier function with a variable predictors $k = 1.2 , ... , p$.

METHODOLOGY

The type of research that is used is *Non Reactive* using secondary data in the document which is in the form of Posbindu Kebun Sari of Polewali Mandar District. Sampling techniques using *simple random sampling* [16]. The response variable, namely genesis DM (Y) and 4 variables predictors i.e. sports activity (X_1), Triglyceride (X_2), hypertension (X_3), and obesity (X_4) [17] [18] [19].

The steps of data analysis in this research is as follows [20][21] .

a. Descriptive statistics on the data based on the variables.
b. Determine confounding variable confounding variable, next notated Z with parameter θ
c. Calculate the value of the estimates of the propensity score for data with MLE method.
d. Test whether the propensity score of the treatment group and control has the same distribution on each covariate. If not balance, then return to step (d).
e. Calculate the value of the estimates of the average treatment effect (ATE).

RESULTS AND DISCUSSION

Descriptive analysis is the early stages of the exploration of the data that is done to get an overview of the research data. Patient characteristics can be seen from the descriptive on each of the variables.

Table 1. Patient characteristics based on the genesis of Diabetes Mellitus (DM)

Covariate (X)	**Kejadian DM**				**Total (%)**
	Tidak Ada	**%**	**Ada**	**%**	
Sports activity (X_1)					
- Enough	46	86.7	33	13.3	55.21
- Less	19	44.1	20	55.9	44.79
Triglyceride (X_2)					
- Normal	47	68.1	35	31.9	71.88
- High	18	66.6	18	33.4	28.12
Hypertension (X_3)					
- No	41	74.5	20	25.5	57.29
- Yes	24	58.5	33	41.5	42.71
Obesity (X_4)					
- No	55	77.4	17	22.6	73.96
- Yes	10	40.0	36	60.0	26.04

Table 1 illustrates that respondents with enough sports that do not have diabetes mellitus of 86.7%, respondents with enough sports experience of diabetes mellitus 13.3%. Sports Reponden less that do not experience diabetes mellitus of 44.1%, respondents with less sports who have diabetes mellitus of 55.9%. Normal triglyceride levels that do not experience diabetes mellitus of 68.1%, respondents with normal triglyceride levels that have diabetes mellitus of 31.9%. Reponden with high triglyceride levels that do not experience diabetes mellitus of 66.6%, respondents with high triglyceride levels are experiencing diabetes mellitus of 33.4%. Respondents with no hypertension who did not have diabetes mellitus of 74.5%, respondents with no hypertension is the experience of diabetes mellitus 25.5%. Reponden with hypertension who did not have diabetes mellitus of 58.5%, respondents with hypertension is the experience of diabetes mellitus 41.5%. Respondents with no obesity who do not experience diabetes mellitus of 77.4%, respondents with no obesity is experiencing diabetes mellitus of 22.6%. Reponden with obesity who do not experience diabetes mellitus of 40.0%, respondents with obesity who experience diabetes mellitus of 60.0%.

Then determine confounding variable. Confounding variable is determined based on the theory and empirical evidence in the form of the relationship between the variables. To prove it, done test dependencies between the variables. Test results the dependencies between the variables are displayed in the following table.

Table 2. Test results dependencies Covariate **X**

Variable	χ^2	**df**	**P-*value***	**Decision**
X_1*X_2	7.239	1	.007	Reject H0
X_1*X_3	10.502	1	.001	Reject H0
X_1*X_4	16.665	1	.000	Reject H0
X_1*Y	11.375	1	.001	Reject H0

Table 2 provides information that sports activity (X1) has a relationship with a variable Triglyceride (X2), hypertension (X3), obesity (X4), and Genesis DM (Y). This shows that the sports activity (X1) related with kovariat (X) and is a risk factor of Genesis DM. So sports activity selected as confounding variables (Z) with Parameter θ to know how big the influence from Genesis DM. After confounder determined, the next step is the estimation of the value of the propensity score. Basically the same propensity value with logistic regression model. Therefore, the propensity value can be known if the parameters from the logistic regression model is obtained. The method used to estimad binary logistic regression model parameter is the MLE method. The results of the estimation of the parameters are displayed in the following table.

Table 3. The estimation with MLE Parameters

Covariate	Parameter (β^*)	SE	Z	p-value	OR
Intercept	-4.1726	0.9070	-4.600	4.22e-06	0.0154
X_2 (1)	2.2338	0.7449	2.999	0.002711	9.3353
X_3 (1)	1.5109	0,753	2.426	0.015266	4.5308
X_4 (1)	2.6708	0,760	3.533	0.000411	14.4515

Based on Table 3 it is known that a significant effect of sports activity (X1) is a person who has no variable Triglyceride X2(1) with p-value = 0.0027, does not have hypertension X3(1) with p-value = 0,0152 and does not have the obesity X4(1) with p-value = 0,0004. From the Table 2 can also formed the model of the value of the propensity score as follows.

$$e\left(\mathbf{x}_i\right)=\frac{\exp\left(-4.173+2,234R\_G_1+1,511HT_1+2,671OBS_1\right)}{1+\exp\left(-4.173+2,234R\_G_1+1,511HT_1+2,671OBS_1\right)}$$

After obtained the value of propensity, next is the subject of the treatment will be matched with a control subject based on the value of the propensity score obtained in the preceding. The matching can be done by comparison of 1:1, which means that 1 the subject of treatment matched with 1 the subject of the control with the following as results.

Table 4. The results of the matching the subject of treatment and control Propensity Score Match

Description	Total
Number of treated obs.	23
Number of matched treated obs.	23
Number of untreated obs.	73
Number of matched untreated obs.	23
Number of total matched obs.	46
Number of not matched obs.	50
Number of matching sets	23
Number of incomplete matching sets	0

Table 4 indicates that the subject of treatment as much as 23 observation and control subject as much as 73 observations. After done fitting, subject treatment and control that matches each as much as 23 observations with total subject that matches is 46 observations. The subject that matches match this will be used on the next steps, while the subject of the control does not match (50 observation) not used again (eliminated from the data set). The next step is an evaluation of the balance kovariat where it is expected that the earlier matching will be able to reduce the effects of confounding bias because that is marked with the balance on all existing kovariat. Testing the balance used test chi-square with the following result.

Table 5. The value of the p-value the results of Balance Kovariat

Covariate	Before the Match	After the Match
Triglyceride	.016	1000
Hypertension	.003	0.757
Obesity	,000	0.737

Table 5 shows that all kovariat has been balanced between treatment groups and control groups after the match, but before match unbalanced. This means that there is no difference enough sports activities and less on triglyceride, hypertension and obesity after the match.

The last step from the community participation is the estimates of the Average Treatment Effect (ATE). The results of the estimation of the value of the Average Treatment Effect (ATE) is shown in table 6 below.

Table 6. Results of the estimation of ATE and Standard Error

Propensity Score	Group	ATE	SE (ATE)	*T-Stat*	*P-value*
Activities Sports	Control Treatment	-0,36587	0,14129	-2.56	0.010

The estimation of treatment effects (ATE) is very important in the propensity, because basically the purpose of propensity is getting estimates of ATE that unbiased and more accurate even though there is a *confounding variable* in the design of the research. Based on Table 5 it is known that that the estimation of ATE of -0,36587 with standard error of 0,14129. Estimates of ATE larger than the standard error gives the test value t that also so that produced the value of p-value < α = 5%. This result indicates that the sports activity based covariat (triglycerides, hypertension and obesity) influence the genesis of diabetes mellitus.

CONCLUSION

Propensity score is a good method used to see the effect of treatment on the study of observation especially on the data involves confounding variable in it. The results of the study showed that the Genesis diabetes mellitus with counfounding sports activity in general influenced by hypertension and obesity. Binary logistics regression test results provide information that a person that triglyceride normal will sports activity of 9.335 times than that have high triglyceride. Furthermore, a person who is not hypertension will sport activity of 4.531 times than that have hypertension, and a person who does not obesity will sport activity of 14.451 times than that of obesity. On the test of balance kovariat shows that all kovariat has been balanced between treatment groups and control groups after the match, but before match unbalanced. Test results propensity score matching through ATE shows that sports activity based on a variable triglyceride, hypertension and obesity affect the genesis of diabetes mellitus.

REFERENCES

[1] Departemen Kesehatan Republik Indonesia. (2008). *Pedoman Teknis Penemuan dan Penatalaksanaan Penyakit Diabetes Mellitus*. Ditjen PTM, Edisi 2: cetakan II, Jakarta.

[2] PERKENI, (2006). *Konsensus Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia,* PB PERKENI.

[3] Kementerian Kesehatan RI. (2014). *Pokok-Pokok Hasil Riskesdas 2013*, Badan Penelitian dan Pengembangan Kesehatan, Jakarta.

[4] Arifuddin, W. (2012). *Analisis Faktor Risiko Diabetes Dengan Kejadian Resistensi Insulin Pada Subyek Pria Dewasa Muda Non-Diabetes,* UNHAS, Makassar.

[5] Mengesha, Addisu Y. (2007). Hypertension and related risk factors in type 2 Diabetes Mellitus (DM) patients in Gaborone City Council (GCC) clinics, *Gaborone Botswana.African Health Sciences*. 7(1):244-245.

[6] D'Agostino, R.B. (1998). *Tutorial in Biostatistics Propensity Score Method for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group*. 17, pp. 2265-2281.

[7] Austin, (2011). A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. *Multivariate Behavioral Research*, 46: pp:119–151.

[8] Rosenbaum, P.R., & Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Journal Biometrika*, vol.70, No.1, pp. 41-55.

[9] Littnerova, S., Jarkovsky, J., Parenica, J., Pavlik, T., Spinar, J., & Dusek, L. (2013). Why to use Propensity Score in Observational Studies? Case Study Based on Data from the Czech Clinical Database AHEAD 2006-09, *cor et Vasa, 55(4)*, pp. 383-390.

[10] Li, H., Graham, D.J., & Majumdar, A. (2013). The Impacts of Speed Cameras on Road Accidents: An Aplication of Propensity Score Matching Methods.*Accident Analysis and Prevention*, 60, 148-57.

[11] Imbens, G. W. (2004). *Nonparametric estimation of average treatment effects under exogeneity: Areview.The Review of Economics and Statistics,* 86, 4–29.

[12] Rubin D. B., (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation (*Health Services & Outcomes Research Methodology*) pp 169-188

[13] Kurth, T., Walker, A. M., Glynn, R. J., Chan, A. K., Gaziano, J. M., Berger, K., And Robins, J. M. (2005). *Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, And Propensity-Based Weighting Under Conditions Of Nonuniform Effect.* Johns Hopkins Bloomberg School of Public Health U.S.A.

[14] Hosmer, D.W, & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley and Sons, Inc.

[15] Agresti, A., (1990), *Categorical Data Analysis*, John Wiley and Sons, Inc, New York.

[16] Levy, P.S., and Stanley, L. (1999). *Sampling of Populations: Methods and Applications*. Third Edition. John Wiley and Sons. Inc. New York.

[17] Ilyas, E. I. (2011). *Olahraga bagi Diabetes*. Jakarta: FK UI

[18] Isara and Okundia. (2015). *The burden of hypertension and diabetes mellitus in rural communities in southern Nigeria, The Pan African Medical Journal. 2015;20:103*

[19] James, et al. (2014). Evidence-Based Guideline for the Management of High Blood Pressure in Adults Report From the Panel Members Appointed to the Eighth Joint National Committee (JNC 8). *Journal of American Medical Association.*

[20] Guo S. and Fraser M. W., (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage Publications.

[21] Pan W. and Bai H., (2015). *Propensity Score Analysis: Fundamental and Developments*. New York: Gulford Press.