Conference Proceedings

3rd INTERNATIONAL CONFERENCE ON RESEARCH, IMPLEMENTATION AND EDUCATION OF MATHEMATICS AND SCIENCE (3rd ICRIEMS)

Yogyakarta, 16 – 17 May 2016

ISBN 978-602-74529-0-9

The Global Challenges on The Development and
The Education of Mathematics and Science

Faculty of Mathematics and Science
Yogyakarta State University

3rd ICRIEMS : The Global Challenges on The Development and The Education of Mathematics and Science

- Mathematics & Mathematics Education
- Physics & Physics Education
- Chemistry & Chemistry Education
- Biology & Biology Education
- Science Education

Published by:
Faculty of Mathematics and Science
Yogyakarta State University
Karangmalang, Yogyakarta 55281
Telp. (0274)550227, Fax. (0274)548203

Board of Reviewer

Preface

Bless upon God Almighty such that this proceeding on 3rd International Conference on Research, Implementation, and Education of Mathematics and Sciences (ICRIEMS) may be compiled according to the schedule provided by the organizing committee. All of the articles in this proceeding are obtained by selection process by the reviewer team and have already been presented in the Conference on 16 – 17 May 2016 in the Faculty of Mathematics and Natural Sciences, Yogyakarta State University. This proceeding comprises 9 fields, that is mathematics, mathematics education, physics, physics education, chemistry, chemistry education, biology, biology education, and science education.

The theme of this 3rd ICRIEMS is '*The Global Challenges on The Development and The Education of Mathematics and Science*'. The main articles in this conference are given by six keynote speakers, which are Prof. Allen Price, Ph.D (Emmanuel College Boston USA), Ana R. Otero, Ph.D (Emmanuel College Boston USA), Dr. Michiel Doorman (Utrecht University, Netherlands), Prof. Dr. Marsigit, M.A (Yogyakarta State University), Asst. Prof. Dr. Warakorn Limbut (Prince of Songkla University, Thailand), and Prof. Dr. Rosly Jaafar (Universiti Pendidikan Sutan Idris, Malaysia). Besides the keynote and invited speakers, there are also parallel articles that presented the latest research results in the field of mathematics and sciences, and the education. These parallel session speakers come from researchers from Indonesia and abroad.

Hopefully, this proceeding may contribute in disseminating research results and studies in the field of Mathematics and Sciences and the Education such that they are accessible by many people and useful for the Nation Building.

Yogyakarta, May 2016

The Editor Team

Forewords From The Head Of Committee

Assalamu'alaikum warahmatullahi wabarakatuh

May peace and God's blessings be upon us all

First of all, allow me to thanks to God, Allah SWT, who has been giving us blessing and mercies so we can join this conference. Ladies and Gentlemen, it is my great honor to welcome you to Indonesia, a unique country which has more than 17,000 islands, more than 1,300 ethnic groups, and more than 700 local languages, and I am also very happy to welcome you to Yogyakarta, the city of education, culture, tourism, and a miniature of Indonesia. We wish you be happy and comfortable in attending the conference in this city.

The third International Conference on Research, Implementation, and Education of Mathematics and Science (ICRIEMS 3rd) 2016 is organized by the Faculty of Mathematics and Science, State University of Yogyakarta. In this year, theme of the conference is : The Global Challenges on The Development and The Education of Mathematics and Science. This conference are dedicated to the 52nd anniversary of Yogyakarta State University and to face challenges of Asean Economic Community in 2016.

This conference facilitates academics, researchers and educators to publish and disseminate their research in the fields of pure, application and education of Science and Mathematics. Furthermore, the purposes of the conference are to establish interaction, communication, and cooperation among academics, researchers and educators at an international level.

On behalf of the committee of this conference, I would like to express our highest appreciation and gratitude to the keynote speakers, including:

1. Allen Price, Ph.D. (Associate Professor of Emmanuel College, Boston USA)
2. Ana R. Otero, Ph.D. (Emmanuel College, Boston USA)
3. Dr. L.M. (Michiel) Doorman (Associate Professor of Utrecht University, Netherland)
4. Prof. Dr. Marsigit, MA. (FMIPA, Universitas Negeri Yogyakarta)
5. Asst. Prof. Dr. Warakorn Limbut (Faculty of Science, Prince of Songkla University, Thailand)
6. Prof. Dr. Rosly Jaafar (Faculty of Physics, Universiti Pendidikan Sultan Idris, Malaysia)

Furthermore, we inform you that the papers presented in this conference are about 200 papers from 302 applicants, who come from various countries and various provinces throughout Indonesia. Therefore, I would like to give my appreciation and many thanks to the presenters and participants who have been actively involved in this seminar.

Finally, I would like to thank the committee members who have been working very hard since half a year ago to ensure the success of the conference. However, if you find any shortcomings and inconveniences in this conference, please forgive us. We would very

happy to receive your suggestions for improvement in the next conference. Thank you very much.

Wassalamu'alaikum warohmatullahi wabarakatuh.

Yogyakarta, May 2016

Dr. Warsono, M.Si.

Forewords From The Dean Of Faculty Of Mathematics And Sciences, Yogyakarta State University

Assalamu'alaikum warahmatullahi wabarakatuh. My greetings for all of you. May peace and God's blessings be upon us all.

On behalf of the Organizing Committee, first of all allow me to extend my warmest greeting and welcome to the International Conference on Research, Implementation, and Education of Mathematics and Sciences, the third to be held by the Faculty of Mathematics and Science, State University of Yogyakarta, one of the excellent and qualified education universities in Indonesia. This conference is also celebrate the 52th Anniversary of State University of Yogyakarta.

This conference proudly presents keynote speeches by six excellent academics, these are: Allen Price, Ph.D., Ana R. Otero, Ph.D., Dr. Michiel Doorman, Prof. Dr. Marsigit, MA., Asst. Prof. Dr. Warakorn Limbut, and Prof. Dr. Rosly Jaafar, and around 200 regular speakers.

The advancement of a nation will be achieved if education becomes a priority and firmly supported by the development of technology. Furthermore, the development of technology could be obtained if it is supported by the improvement of basic knowledge such as mathematics, physics, chemistry, and biology. The empowerment of this fundamental knowledge may be achieved by conducting research which is then implemented in developing the technology and the learning process in schools and universities.

This international conference is aimed to gather researchers, educators, policy makers, and practitioners to share their critical thinking and research outcomes. Moreover, through this conference it is expected that we keep updated with new knowledge upon recent innovative issues and findings on the development and the education of mathematics and science, which is in accord with the theme of the conference this year. All material of the conference which are compiled in the abstract book and proceedings can be useful for our reference in the near future.

This conference will be far from success and could not be accomplished without the support from various parties. So let me extend my deepest gratitude and highest appreciation to all committee members who have done an excellent job in organizing this conference. I would also like to thank each of the participants for attending our conference and bringing with you your expertise to our gathering. Should you find any inconveniences and shortcomings, please accept our sincere apologies.

To conclude, let me wish you fruitful discussion and a very pleasant stay in Yogyakarta.

Wa'alaikumsalam warahmatullahi wabarakatuh

Yogyakarta, May 2016
Dean Faculty of Mathematics and Science
Yogyakarta State University

Dr. Hartono, M.Si.

Regular Papers:

Mathematics and Mathematics Education

M – 07

Cox Proportional Hazard Model with Multivariate Adaptive Regresion Spline

Hendra Dukalang[1], B. W. Otok[2], Ismaini Zain[2], Herlina Yusuf[3],
[1]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[2]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[3]Dept. of Public Health, Universitas Negeri Gorontalo
hendra37.hd@gmail.com

Abstract— Events related to the survival time always happens in everyday life, one of which is time duration that need to recover from illness. Time that we need until the event happened is called survival data. Generally, not all of survival data can be observed and it is called data censored. One of statistical method that can be used to analyze and determine the survival rate of survival data is the cox proportional hazard models. In its development, the residuals of the cox proportional hazard (Cox PH) model can be used as response variable for regression function. The relationship between response variable and predictor variables often is not known the function of regression. So we are needed nonparametric regression. One of method nonparametric regression that can be used is Multivariate Regression Adaptive Spline (MARS). In this study, survival analysis is focused on the patients of HIV/AIDS which is a deadly disease. To determine survival rate of HIV/AIDS patients is used a hazard function and survival function with time duration patient stayed as variable. To know the other factors of the survival of HIV/AIDS patient is used Cox PH Models with MARS approach. The results showed that gender is one factor in the survival of HIV/AIDS patients, and treatment compliance, employment status, CD4 count, age and educational level.

Keywords: *Survival Analysis, Cox PH, MARS, HIV/AIDS*

I. INTRODUCTION

Survival analysis is a statistical analysis that is specifically used to analyze the data or cases related to the time duration until the event happened and there are data censored [1]. At first time, studied of survival is focused on the probability predictions of response, survival, average life expectancy and comparing the treatment of survival illustration experiment in humans. But survival analysis developed in the identification of risk factors and prognostic factors associated with the development of the disease [2]. One method of analysis that can be used for survival data are cox proportional hazards regression (Cox PH). Cox PH regression modeling can also be used to determine which combination of independent variables that influence in the model. In its development, Cox PH regression modeling can include relationships between predictor variables with the model function multivariate regression adaptive spline (MARS).

MARS is one of nonparametric regression method that does not depend on the assumption of a certain curve shape so it has flexibility in high dimensional data and modeling involves a lot of interaction with a few variables [3]. The variable responses in MARS modeling can use the residuals of the modeling Cox PH, so the survival modeling of MARS can be interpreted as MARS modeling the response variable is the residual result of modeling Cox PH [4]. The Previous research has been done to use of survival analysis with MARS approach in DBD cases, where the response variable of MARS models use *martingale residual* for uncensored data [5]. Then Cox proportional hazard and MARS used to analyze product sales with a electronic media system [6]. Previously, they had done research on survival analysis using MARS approach for the case of survival of heart patients in Germany, and show that the MARS method give better results than Cox PH regression [4].

In this study, the Regression Cox PH using MARS approach is used to determine the factors that influence survival of HIV/AIDS patients. Human Immunodeficiency Virus (HIV) is a virus that decrease the body's immune system so that the people affected by this virus will be susceptible to various infections and then causes *Acquired Immune Deficiency Syndrome* (AIDS). Research on HIV/AIDS in Indonesia is more emphasis on efforts to reduce the incidence of HIV/AIDS and how the healing response of

HIV/AIDS. One of them is a mixture survival modelling for HIV/AIDS cases in Semarang [7]. To determine the factors that affect the survival of HIV/AIDS.

II. LITERATURE REVIEW

A. Survival Analysis

Survival analysis is a statistical method that can be used to analyze data that related to start time (time origin) or start point until the specific event happened (end point) or failure event [8]

To determining the survival time, there are three factors required:

1. *Time origin* (starting point), is time to record and analyze an incident when the patients were first declared HIV/AIDS.
2. *Ending event of interest* (recent events) is the expired recording time. This time is useful to know the status of censored or not censored patient to be able to do analysis. Recent events in this study is the time when the HIV/AIDS patients were declared dead.
3. *Measurement scale for the passage of time* as a limit of the time of incident from the beginning to the end. The scale is measured in days, weeks, months, or years. In this study measuring scale used the time duration when the patients were suffering HIV/AIDS in months.

In survival analysis, there is difficulty data observing that is the possibility of some individual observations who cannot be observed from the start point to the end point, this situation is called the censored data [1]. In this study, there are three causes of censored data.

1. *Loss to follow up*, occurs when the patient decides to move another hospital or refuse to observe.
2. *Drop Out*, occurs the patient chooses to go home.
3. *Termination of Study*, occurs when the research period was ended while the patient has not reached the failure event.

B. Hazard Function and Survival Function

In survival analysis, there are two main functions that is survival function and hazard function [8]. Survival function is the basis of survival analysis, because it includes the probability of survival from the time varying provide important information about the survival data. Survival Function is an individual opportunity who can survive over time *t* [2], and usually denoted by.

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)\,du \tag{1}$$

$$S(t) = \exp\left[-\int_0^t \lambda(u)\,du \right] \tag{2}$$

Hazard Function is an individual probability to reach specific incidents at time interval $(t, \Delta t)$ with individual assuming to stay on at this time interval. And usually denoted by $\lambda(t)$. This function is used to express the *hazard rate* or the rate of cure and survival up to time-*t*.

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{3}$$

Where $f(t)$ is probability density function (PDF) on the distribution of the estimated survival data, and it is known that:

$$\int_0^t f(u)\,du = 1 - S(t) \tag{4}$$

So generally, the relationship of survival function and cumulative hazard function based on that equation is as follows:

$$\Lambda(t) = -\ln\ \mathrm{S}(t) \tag{5}$$

C. Distribution Estimates

Estimation of distribution used to the survival data which in this study is duration of suffering HIV/AIDS patients to otherwise experience *failure event*. Estimation of distribution is conducted by

Anderson-Darling test (AD) because it has a strong strength and accurate if we compared with other distribution test [9].

Equation Anderson-Darling test statistic (AD) is as follows:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2_i - 1)\left[\ln F(Y_i) + \ln F(Y_{n+1-i})\right] \quad (6)$$

Where: F = the cumulative distribution function of the conjecture distribution.
Y = survival time data.
n = number of sample

D. Cox Proportional Hazard Model

Regression modeling to determine the factors that influence survival data for uncensored data is called Cox Proportional Hazard Regression models [10]. Cox PH regression is used when the observed outcome was the length of time of an event. This Modeling is a log-linear relationship between X and the general function of hazard on T are as follows:

$$\lambda(t|X - x) = \lambda_0(t)e^{\beta x} \quad (7)$$

For variable X that has covariate, the equation used is as follows:

$$\lambda_i(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} \quad (8)$$

Where:

$\lambda_i(t)$ = hazard function for individual to $-i$

$\lambda_0(t)$ = baseline hazard

$\beta_1, \beta_2, \ldots, \beta_p$ = coefficient regression

$x_1, x_2, \ldots, x_p$ = variable value for individual to $-i$

The most important assumptions that must be met in the regression is Cox Proportional Hazard assumptions which means that the ratio of the hazard function is constant over time or equivalent to the statement that the ratio of the hazard function of an individual against another individual hazard function is proportional. This research will use the approach chart using log minus log survival plots to check the assumptions Proportional Hazard. According to the Cox regression model, the hazard function for failure individual-i for time-t can be written as in Equation (9) is as follows:

$$\lambda_i(t) = \lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j\right) \quad (9)$$

Modelling using Cox Proportional Hazard produces two types of residual, that is Martingale Residual and Deviance Residual that obtained from Cox Null Model. This study used Martingale Residual which serves as the response variable for modeling MARS. Residual Martingale equation is as follows:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds \quad (10)$$

$$M_i(t) = N_i(t) - \hat{\Lambda}_i(t) \quad (11)$$

Where

$M_i(t)$ = Martingale Residual-i at time-t

$N_i(t)$ = The process of counting events (data uncensored given value of 1 and data censored given value of 0) for data-i at time-t

$Y_i(s)$ = Indicators, if subject-i is *under risk immediately* before-t

$\hat{\Lambda}_i(t)$ = Breslow estimator of the cumulative baseline hazard function

E. Multivariate Adaptife Regression Spline

Multivariate Adaptive Regression Splines (MARS) is one of the new flexible method for modeling high-dimensional regression data. MARS is a form of extension of the Basis Splines Functions where the number of basis function is the parameters of the model.

Some terms that need to be considered in the methods and modeling MARS is as follows,

1. *Knots* is the point of a regression line to form a region of a regression function.
2. *Basis Function* (BF) is a collection of some of the functions that are used to describe the relationship between the response variable and the predictor variable.
3. Interaction is a correlation between variables and the maximum number of interaction (MI) 1, 2, and 3.

The general equation MARS models are as follows:

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - \tau_{km} \right) \right] + \varepsilon \tag{12}$$

Estimator model of multivariate adaptive regression splines or MARS [3]:

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - \tau_{km} \right) \right] + \varepsilon \tag{13}$$

Where the first summation covers all the bases for a single variable functions, covering all the bases the second summation function for the interaction between two variables, the third summation includes all the base functionality for the interaction between the three variables and so on [3].

MARS modeling is determined by trial and error for the combination of BF, MI, and MO to get the value of minimum GCV. GCV equation is as follows:

$$GCV(M) = \frac{ASR}{\left[1 - \frac{\tilde{C}(M)}{N} \right]^2} = \frac{\frac{1}{N} \sum_{i=1}^{N} \left[y_i - \hat{f}_M(x_i) \right]^2}{\left[1 - \frac{\tilde{C}(M)}{N} \right]^2} \tag{14}$$

In the case of additive modeling suggested to use a value of d = 2, based on the decline in expectation value of ASR [3]. While suggests conventional value $d = 4$ [1]. The smaller the value of d, the larger models which will produce with more functions of the base, and conversely the greater the value of d, the smaller models which will produce with fewer basis functions.

F. HIV/AIDS

Human Immunodeficiency Virus (HIV) is a virus that decrease the body's immune system so that the people affected by this virus will be susceptible to various infections and then causes *Acquired Immune Deficiency Syndrome* (AIDS). There are about 5 -10 million people living with HIV who do not yet show any symptoms but as a potential source of infection. AIDS is a disease that is very dangerous because it has a case fatality rate of 100% in five years, meaning that within 5 years after diagnosis of AIDS in upholding then all people will die [11]. Factors that affect the survival of people with HIV/AIDS are age, gender, education level, status employment, status marital, history ARV, absolute CD4 count, opportunistic infections, functional status, stage, and treatment compliance.

III. METHODOLOGY RESEARCH

A. Data Source

The data used in this research is secondary data on the medical records of HIV/AIDS patients in one hospital counted 100 data. Variables used in this research are:

Y : Survival Time
X_1 : Age
X_2 : Gender
X_3 : Education level
X_4 : Status of jobs
X_5 : Marital status
X_6 : History ARV

X_7 : absolute CD4 levels
X_8 : Opportunistic Infections
X_9 : Functional Status
X_{10} : Stadium
X_{11} : Compliance therapy

B. Method Analysis

a. Determine the survival data that will be used to eliminate the data censored.
b. Describing the characteristics of patients with HIV/AIDS
c. Predicting survival data distribution using the smallest of Anderson-Darling value
d. Determining the baseline hazard function
e. Estimating the survival function and cumulative hazard function
f. Using the Cox PH models to get Martingale residual,
g. Doing plotting data to know the Martingale residual predictor variables.
h. Modeling Cox PH with MARS approach through the following steps:
    1. Modeling with MARS combined Basis Function (22, 33, 44), Maximum Interaction (1, 2, 3), and the Minimum observation (0, 1, 2, 3)
    2. Getting the best model based on the value of the minimum GCV
    3. Modeling Cox Proportional Hazard with MARS approach
    4. Interpretation models
    5. Determine the level of interest for each of the significant variables in the model
i. Summing up the results of the analysis

IV. ANALYSIS AND DISCUSION

A. Descriptive Statistics

Before the description of the characteristics of patients with HIV/AIDS, then the description of the survival data were used.

TABLE 1. DESCRIPTIVE DATA SURVIVAL

N Total	n censored	n observation
100	51	49

Table 1 shows that of the 100 data obtained, there are 51 data classified in the data censored, where this data must be removed because it cannot be used in the survival analysis. It can be concluded that the survival data in this study there are as many as 49 data.

TABLE 2. DESCRIPTIVE PATIENTS HIV/AIDS

Variable	Characteristics	Number	Variable	Characteristics	Number
Age	Toddlers (0-5 years)	5	Absolute CD4 levels	>350	1
	Children (5-12 years)	1		200-350	5
	Adolescents (12-23 years)	2		<200	43
	Adults (>23 years)	41	Opportunistic Infections	< 2	17
Gender	Female	25		> 2	32
	Male	24	Functional Status	Normal	9
Education	Higher	11		Ambulatory	6
	Primary	21		lying	34
	None	17	Stadium	Stage I	12
Jobs	Working	31		Stage II	8
	Not Working	18		Stage III	15
Marital Status	Married	27		Stage IV	14
	Not Married	22	Compliance therapy	Comply	22
ARV	Ever	19		non-compliant	27

History	Never	30			

Table 2 shows the ingredients HIV/AIDS patients who experience even failure are aged above 23 years of age or older. With a CD4 count of less than 200.

B. Distribution Estimates

Estimation of the distribution is used to determine the distribution of survival data were used. The distribution function was used to estimate the survival function and cumulative hazard function. The distribution function is also used to determine the baseline hazard function which is used in the modeling.

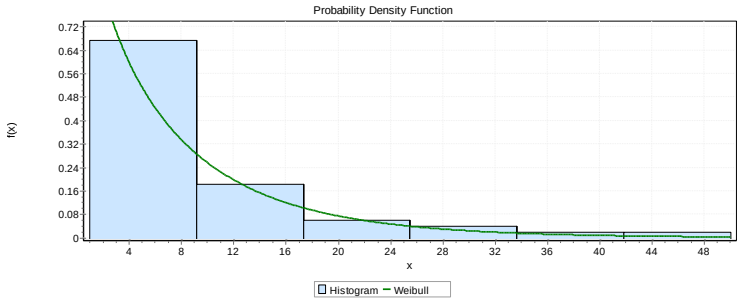


FIGURE 1. HISTOGRAM SURVIVAL DATA USED IN THE STUDY TO ESTIMATE THE DATA DISTRIBUTION.

Based on the estimation of the distribution using Anderson-Darling test, it is known that the smallest Anderson-Darling value is contained in a 2-parameter Weibull distribution in the amount of 1.93 with estimates of the parameters are and Based on estimates of parameters for two parameter Weibull distribution, then the baseline hazard function obtained are as follows :

$$\lambda_0(t|\eta,\gamma) = \frac{\gamma}{\eta}\left(\frac{t}{\eta}\right)^{\gamma-1}$$
$$= \frac{7.149}{0.859}\left(\frac{t}{0.859}\right)^{7.149-1}$$
$$= 8.322\left(\frac{t}{0.859}\right)^{6.149}$$

C. Estimated survival function and hazard function

Survival function is used to determine the probability of the patient's recovery, and cumulative hazard function is used to determine the rate of cure of HIV/AIDS. The estimation results of the survival function and the hazard function is as follows:

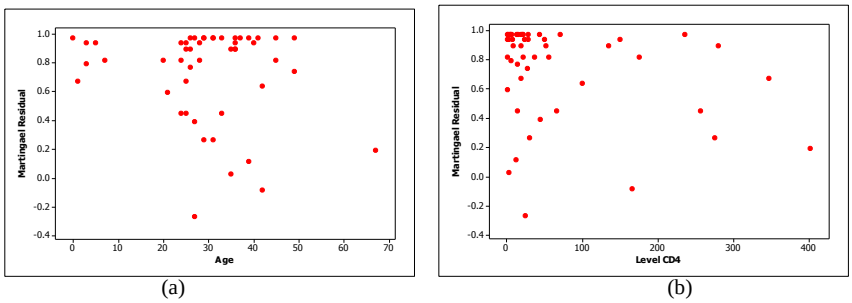
TABLE 3: ESTIMATED SURVIVAL FUNCTION AND CUMULATIVE HAZARD FUNCTION

Survival time	$S(t)$	$\Lambda(t)$	**Survival time**	$S(t)$	$\Lambda(t)$
1	0.969	0.031	14	0.666	0.406
2	0.939	0.063	15	0.575	0.553
3	0.899	0.106	16	0.542	0.612
5	0.831	0.185	18	0.478	0.738
7	0.811	0.209	24	0.445	0.810
8	0.791	0.234	27	0.412	0.887
11	0.769	0.262	28	0.377	0.976
12	0.720	0.328	36	0.337	1.087
13	0.694	0.365	50	0.281	1.269

Table 3 shows that the longer a patient is suffering from HIV/AIDS, the lower probabilities of survival for people is living with HIV/AIDS. On the contrary, the longer a patient suffering from HIV/AIDS, the higher the survival rate of patients with HIV/AIDS. It can be concluded that the probability of survival of patients with HIV/AIDS is inversely related to the survival rate of patients with HIV/AIDS.

D. Cox Proportional Hazard with Multivariate Adaptive Regression Spline

Before modeling with MARS, it is important to know the pattern of the relationship between the predictor variables and the response variable MARS modeling.



(a) (b)

FIGURE 2. SCATTER PLOT MARTINGALE RESIDUAL VS PREDICTOR VARIABLES (a) AGE, (b) THE ABSOLUTE CD4 LEVELS

Figure 2 shows that there is no particular pattern of variable X to variable Y. The pattern of relationships that do not tend to form patterns, showed that it could be used in MARS. MARS modeling done by trial and error for 36 combinations Basis Function (BF), Maximum Interaction (MI), and the Minimum Observation (MO) to get the best model based on the value of the minimum GCV.

Based on the results of trial and error combination BF, MI, and MO, the combination of which produces minimum GCV value is a combination of 22, 3, 1 with a value of GCV = 0.573 with R2 = 0.729. Based on the results of this combination, it is known MARS models produced are as follows:

$$Y = 0.721 + 0.391 * BF3 + 0.200 * BF4 - 0.001 * BF5 - 0.015 * BF7 - 0.012 * BF10$$
$$+ 0112 * BF11 - 2.603 * BF12 + 0.017 * BF14 - 0.088 * BF16 - 0.010 * BF18;$$

Where:
BF2 = (X11 = 2);
BF3 = max (0, X1 - 35,000) * BF2;
BF4 = max (0, 35,000 - X1) * BF2;
BF5 = max (0, X7 - 1000) * BF3;
BF7 = max (0, 275,000 - X7) * BF2;
BF8 = (X2 = 1) * BF2;
BF10 = max (0, X7 - 25,000) * BF8;
BF11 = max (0, 25,000 - X7) * BF8;
BF12 = (X4 = 1) * BF8;
BF14 = max (0, X7 - 257 000) * BF4;
BF16 = (X3 = 3) * BF4;
BF18 = max (0, X7 - 236,000);

Resulting in a model hazard rate or the rate of survival of patients with HIV/AIDS as follows:

$$\lambda(t) = \lambda_0(t)\exp\left(\hat{Y}\right)$$
$$= 8.322\left(\frac{t}{0.859}\right)^{6.149} \cdot \exp\left(\begin{matrix} 0.721 + 0.391 * BF3 + 0.200 * BF4 - 0.001 * BF5 - 0.015 * BF7 - 0.012 * BF10 \\ +0112 * BF11 - 2.603 * BF12 + 0.017 * BF14 - 0.088 * BF16 - 0.010 * BF18; \end{matrix}\right)$$

TABLE 4: INTERACTION ON BASIS FUNCTION

BF	Interactions	Specification
3 and 4	x1 and x11	Age and compliance
5	x7 * x1 and x11	CD4 levels, age and compliance
7	x7 and x11	CD4 levels and compliance
10	x2 and x11	Gender and compliance
11	x7 and x2	CD4 levels and gender
12	x7 * x2 and x11	CD4 levels, gender and adherence

14	x4 * x2 and x11	Employment, gender and adherence
16	x3* x1 and x11	The level of education, age and compliance

The Modeling results show that in general, the variables that affect the survival of patients with HIV/AIDS there are six variables: X1 (Age), X2 (Gender), X3 (Level of Education), X4 (Employment Status), X7 (Kadar CD4) and X11 (Compliance Therapy). The sixth of these variables has a good influence on the model, either individually or when interacting with other variables.

Table 4 shows the interaction of the variables that affect the survival of patients with HIV / AIDS. As for the variables that influence individual is adherence therapy and education level.

TABLE 5. VARIABLE INTEREST RATE

Variable	Importance	GCV
Gender	100	0.147
Therapy adherence	84.01	0.112
Employment Status	79.913	0.104
CD4 levels	78.947	0.102
Age	68.334	0.084
Education	16.149	0.032

Table 5 shows that gender have the largest contribution to the resulting model 100%. Then, the second largest contribution is in the amount of therapy adherence 84.010%. Then the third largest contribution is the employment status 79.913%, then the fourth biggest contribution is the Absolute CD4 cell count of 78.94%, the fifth biggest contribution is the Age of 68.334%, and the sixth biggest contribution is the level of education, amounting to 16 149.

V. CONCLUSION

HIV/AIDS patients who died is the average adult aged 23 years or older (age of majority), with CD4 levels below 200. Based on the modeling results with Cox Proportional Hazard MARS approach, which used a combination Basis Functions, Maximum interaction and minimum His observations are 22, 3, and 1 with a minimum GCV value was 0.028. Variables influencing the survival of patients with HIV/AIDS in individuals is age and compliance, levels of CD4 and compliance, gender and adherence, levels of CD4 and gender, CD4 count, gender and adherence, CD4 count, age and compliance, employment, gender and adherence, education level, age and compliance. Gender have the largest contribution to the resulting model, by 100%. Then, the second largest contribution is in the amount of therapy adherence 84.010%. Then the third largest contribution is the employment status of 79.913%, then the fourth biggest contribution is the Absolute CD4 cell count of 78.947%, the fifth biggest contribution is the Age of 68.334%, and the sixth biggest contribution is the level of education, amounting to 16.149%.

REFERENCES

[1] Kleinbaum. D. G. (2012). *Survival Analisis*, London, Springer

[2] Lee, E.T. (2003). Statistical Method for survival Data Analysis. London John Willey

[3] Friedman, J.H., (1991), "Multivariate Adaptive Regression Spline", *The Annals of Statistics*, Vol. 19, pp 1-141.

[4] Kriner, M. (2007). Survival Analisis with Multivariate Adaptive Regression Splines. Disertasi. Munchen University.

[5] Nisa', F.S. dan Nudiantara (2012). Analisis Survival dengan Pendekatan Multivariate Adaptive Regression Spline pada Kasus Demam Berdarah Dengue (DBD). *Jurnal Sains dan Seni ITS.* Vol. 1, No. 1, 318-323

[6] Irwansyah, E. Nyoman, D.A, dan Bekti R.D. (2014). Cox Proportional Hazard with Multivariate Adaptive Regression Spline to Analyze the product Sales Time in E-Commerce. *Article in International Journal of Applied Mathematics and Statistics*

[7] Saputro. A. S. (2013) pemodelan *mixture survival* untuk kasus HIV/AIDS. Universitas Airlangga. Surabaya

[8] Collect, D. (2003). *Modeling Survival Data in Medical Research*. London: Chapman & Hall/CRC

[9] Purhadi. (2012). Analisis Survival Faktor-faktor yang mempengaruhi Laju kesembuhan pasien Penderita Demam Berdarah Dengue (DBD) di RSU Haji Surabaya dengan Regresi Cox. *Jurnal Sains dan Seni ITS*, Volume I. No. I., 271-267.

[10] Cox, D. R. (1972). Regresion Model and Live Tables (with discussion), *Journal of The Royal Statistical Society*, 34 : 187-220

[11] Wibisono B, (1989). *Epidemiologi AIDS*; petunjuk untuk petugas kesehatan, Departemen Kesehatan RI. Jakarta.

M – 10

Propensity Score Stratification Analysis using Logistic Regression for Observational Studies in Diabetes Mellitus Cases

Ingka Rizkyani Akolo[1], B.W.Otok[2], Santi W. Purnami[2], Rama Hiola[3]
[1]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[2]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[3]Dept. of Public Health, Universitas Negeri Gorontalo
inkarizkyani05@gmail.com / molavecha@gmail.com

***Abstract*—** Observational studies are the basis of epidemiological research to draw the conclusions of the effects or a response treatment. In general, a randomized trial is required in order to meet the assumption of independence to minimize the bias effects. However in an observational study, particularly in medical field, randomization not able to implement because conduces in doubtful treatment effects estimation. Propensity score is the conditional probability to get certain treatments involving the observed covariates. This method is used to reduce bias in the estimation of the impact of treatment on observational data for their confounding factors. If treatment is binary, then the logistic regression model is one estimated of propensity score because of easiness in terms of estimation and interpretation. In the analysis of observational studies, propensity score stratification (PSS) has proven to be one of methods to adjust the unbalanced covariate for the purposes of causal inference. The data used in this study is the medical records of patients DM in X hospital about the factors that influence the type of diabetes mellitus. In this study PSS used in diabetes mellitus cases to reduce bias due to confounding factors, so that can be known the factors affect the type of diabetes mellitus with obesity as confounding factors. The results of PSS analysis is known that the variables directly influence the type of DM are obesity, age, gender and variable does not directly influence the type of DM are genetic variable, sport activities and dietary habit of patients DM.

Keywords: *observational studies, confounding, propensity score stratification, diabetes melitus*

I. INTRODUCTION

The attention of non-communicable diseases is increasing currently. From ten leading causes of death, two of them are non-communicable diseases. Diabetes mellitus (DM) is a non-communicable disease with high prevalence. International Diabetes Federation (IDF) stated that people with diabetes mellitus figure reached 382 million people of the world in 2013. It is estimated as 592 million in 2035. In Indonesia, people with diabetes mellitus has reached 8.4 million in 2000 and is estimated to be approximately 21.3 million in 2030. Because of high number of patients, it makes Indonesia ranks fourth after the United States, India and China [1].

According to the results of Indonesia Basic Health Research (RISKESDAS) in 2013, an increase in the prevalence of Indonesia's diabetes mellitus in 2007 was 1.1% to 2.1% in 2013. The results of the analysis of the Diabetes Mellitus prevalence's picture based on a doctor's diagnosis and symptoms increase with age. It began with age ≥ 65 years old of decline. The prevalence of diabetes in women is 1.7% while men have 1.4%. Based on its territory, the prevalence of urban areas (2.0%) is higher than in rural areas 1.0% [2].

Diabetes mellitus (DM) is a chronic metabolic disorder due to the pancreas does not produce enough insulin or the body can not use the insulin that is produced effectively. Insulin is a hormone that regulates blood glucose levels. Diabetes mellitus is classified into type 1 diabetes, which is known as insulin-dependent or childhood-onset diabetes, characterized by a lack of insulin production. Type 2 diabetes, known as non-insulin-dependent or adult-onset diabetes, caused by the body's inability to use insulin effectively which then lead to overweight and lack of physical activity [3].

Increasing the number of people with diabetes are mostly caused by the interaction between the factors of genetic susceptibility and exposure to the environment, such as changes in lifestyle and physical activity often leading to obesity. It is a risk factor for the onset of DM [4]. Therefore, diabetes mellitus type 2 is often also called diabetic lifestyle for causes not only because of heredity, but also environmental factors include age, obesity, insulin resistance, food, physical activity, and unhealthy play roles in the occurrence of diabetes [5].

Research on the incidence of diabetes mellitus (DM) has been done in large quantities. For example Wicaksono [4] investigated the factors associated with the occurrence of diabetes mellitus (DM) type II using descriptive analysis and logistic regression. Trisnawati et al. [6] studied the risk factors of type 2 DM outpatients using the McNemar test and logistic regression and Indriyani et al. [7] studied the effect of physical exercise to decreased levels of blood sugar of patients with type 2 DM using the t test with the one group pretest-posttest study design.

The above researches mostly used descriptive analysis and logistic regression without considering the possibility of a powerful combination of factors affecting diabetes mellitus (DM). In fact, as explained previously that the combination of these factors led to the existence of confounding variables that lead to obtain inaccurate conclusions.

Some previous studies have tried to discuss confounding factors randomly, but in the case of health sector, it can not be done. But how the confounding variables included in the factors studied. Therefore, we need a method that can handle the effects of bias caused by these confounding factors. One method that can handle confounding is the propensity score method. it was first introduced by Rosenbaum and Rubin in 1983. The propensity score is defined as the conditional probability to receive interventions based on those characteristics before the intervention [8]. This method is a statistical adjustment that can be used to analyze data from non-experimental research design where design giving treatment through randomization to treatment or control group is not possible. Researchers can use the propensity score for statistical balance or equalize the group of research subjects to reduce bias due to the provision of treatment which is not random.

One method of propensity score that is proven to reduce bias due to confounding effects is the propensity score stratification method. This method focuses on the division of classes / strata based on the estimated value of propensity score. The division of classes / strata aims to balance the distribution between treatment and control groups so that estimate of average treatment effect more accurate.

Several studies of the model used to estimate the value of propensity score, they are McCaffrey et al.[9] which used a model of generalized boosted, McCandless et al. [10] used Bayesian, and Littnerova et al.[11] used logistic regression to estimate propensity score. Of all the study, estimated by logistic regression simpler and easier in interpretation, particular to the category data used.

Based on the description above, the aim of research in this study are to get an estimation of average treatment effect and binary logistic regression model based on the propensity score that shows the factors affecting the type of DM in patients treated in X hospital district after being controlled by confounding variables of obese patients' status.

II. THEORY

2.1 Logistic Regression Model

According to Hosmer & Lemeshow [12] binary logistic model is the logarithm of odds ratio of occurrence of success (π) and probability of occurrence of fail (1 - π) . The specific form of the logistic regression model with *p* predictor variables expressed in equation (2.1)

$$\pi(\mathbf{x}) = \frac{\exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m x_m\right)}{1 + \exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m x_m\right)} \tag{2.1}$$

Form of simplification of the equation above, then used a logit transformation of the form below.

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = \mathbf{x}^T \boldsymbol{\beta} \tag{2.2}$$

with $\pi(\mathbf{x})$ is the probability of success, $1-\pi(\mathbf{x})$ is probability of fail event, β_m are the parameters of the linear function with the predictor variables $m = 1,2 , ... , p$.

2.2 Propensity Score

Propensity score analysis introduced by Rosenbaum and Rubin 1983 in the journal entitled "The central role of the propensity score in observational studies for causal effects". Propensity score analysis is a statistical method that rapidly evolving innovative and useful for evaluating treatment effects when using observational data [13]. Rosenbaum and Rubin [8] define the propensity score for observation *i* (i = 1, ..., n) as the conditional probability of a specific treatment (Z_i = 1) versus non-treatment (Z_i = 0) based on the characteristics of the covariates $\mathbf{x}_i$ observed.

According to Guo & Fraser [13] the value of propensity score is defined as follows.

$$e(\mathbf{x}_i) = P(Z_i = 1 | X_i = x_i) \tag{2.3}$$

According to Littnerova et al.[11] propensity score using a logistic regression model, the response variable is a binary where to treatment and to the control unit with the following model.

$$e(\mathbf{x}_i) = \mathrm{P}(Z_i = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})} \tag{2.4}$$

with β_0 is a constant, $\beta_1, \beta_2, \cdots, \beta_p$ the regression coefficients and $x_1, x_2, \cdots, x_p$ are covariate variables.

According to Cochran & Rubin (1973) in the Pan & Bai [14] measures the bias is reduced for each covariate can use equation (2.11)

$$PBR = \frac{B_{\text{before PS}} - B_{\text{after PS}}}{B_{\text{before PS}}} \text{x}100\% \tag{2.5}$$

and

$$B = p_1(x_p) - p_0(x_p) \tag{2.6}$$

with *PBR* is Percent Bias Reduction, *B* is an average difference of the treatment group and the control group for each covariate , $p_1(x_p)$ and $p_0(x_p)$ are proportion of covariates for the treatment group and the control group, $B_{\text{before PS}}$ and $B_{\text{after PS}}$ are represents the difference between the average treatment and control group before propensity score and after propensity score.

2.3 Propensity Score Stratification

Propensity Score Stratification (PSS) is a procedure of classifying subjects into classes based on the estimated propensity score . Subjects are sorted by the estimated propensity score (Austin, 2011). Cochrane (1968) showed that the five sub-class is enough to reduce 90 % of bias with a single covariate [15]. Imbens [16] declared the entire bias under unconfounded associated with the propensity score, it indicates that under the normality used 5 strata change is largely biased with all covariates.

According to Yanovitzky, Zanutto, and Hornik [17] general steps of propensity score analysis are described as follows

1. Choose a covariate as a confounder for the estimation of propensity score. The election process can confounder based on theory and empirical evidence about the relationship between variables.
2. Estimated value of propensity score.
3. Divide the strata based on the propensity score.
4. Check the balance of covariates between the treatment group and the non-treatment.
5. Calculate the effect of confounders.

One way to assess the quality of the propensity score stratification by comparing a variety of statistics such as mean, median, variance, t-test statistics, chi-square test or Kolmogorov-Smirnov (KS) test on each covariate [15]. In this study, KS and chi-square used for testing difference distribution between the treatment group and the control group.

2.4 *Diabetes Mellitus*

Diabetes mellitus is metabolic diseases which is a collection of symptoms that arise in a person because increase in blood glucose levels above normal values . The disease is caused by disorders of the metabolism of glucose due to a deficiency of insulin both absolute and relative terms . There are two types of diabetes mellitus. The first type of DM is type 1, that usually acquired since childhood and results from the pancreas failure to produce enough insulin. The second type of DM is type 2, that

usually acquired an adult and condition in which cells fail to respond to insulin. According Poretsky [18] factors that affect type 1 diabetes is a genetic, autoimmune, age, race and ethnicity, gender, and environmental factors such as viral infections, diet / nutrition, stress. In addition, according Gungor, Hannon, Libman, Bacha, & Arslanian [19] factors affecting the type 2 diabetes are genetic, age, gender and environmental factors such as diet, obesity, sports activities.

III. METHODOLOGY

The method used in this study is propensity score stratification (PSS) method to find the factors that influence the type of diabetes mellitus (DM) with obesity status of patients as a confounding factor. The data used is secondary data from medical records of patients (DM) at Hospital X in 2013. The number of respondents are 497 patients. The Patients consist of patients with type 1 of DM (42 patients) and patients with type 2 DM (455 patients). The response variable is the type of DM and predictor variables are genetic, age, gender, dietary habit, sport activities and obesity. The stages of research process can be seen in Figure 1 below.

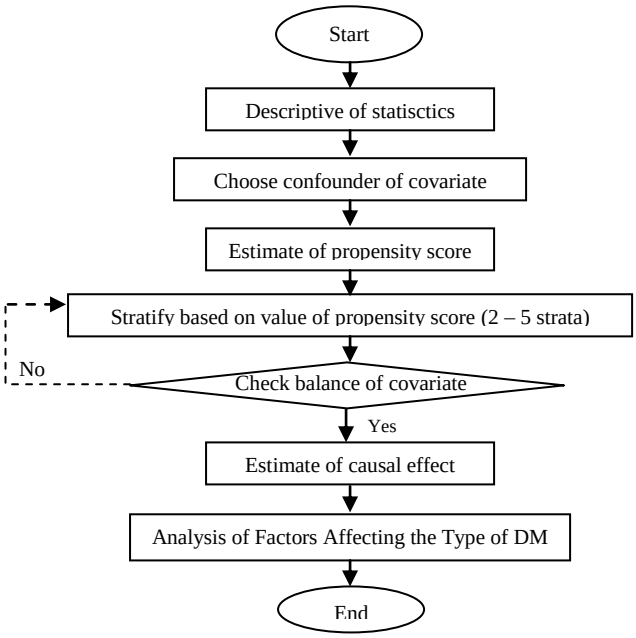


FIGURE 1. STAGES OF RESEARCH PROCESS

IV. RESULT AND DISCUSSION

4.1 Descriptive of Statistics

Descriptive of statistics is an early stage of data exploration to get a general overview of the research data. Characteristics of respondents can be seen from the descriptive of each variables shown in Table 2.

TABLE 1. DESCRIPTIVE ANALYSIS OF COVARIATE

Covariate	Status Obesity		%	Type of DM		%
	Obesity	No Obesity		Type 1	Type 2	
Genetic						
- Have genetic	379	31	82,49	0	410	82,49
- Have not genetic	52	35	27,51	42	45	27,51
Age	431	66	-	42	455	-
Gender						
- Male	192	32	45,07	24	200	45,07
- Female	239	34	54,93	18	255	54,93
Dietary habit						

- Meet	29	63	18,51	25	67	18,51
- No Meet	402	3	81,49	17	388	81,49
Sport Activities						
- Active	29	65	18,91	27	67	18,91
- Less Active	402	1	81,09	15	388	81,09

Based on the table 2 can be shown that to 82.49 % patients have genetics DM, 81.49 % patients have dietary habit (no meet) and 81.09 % patients less active in sports activities. In addition, it was known that the number of female patients (54.93%) are greater than male patients (45.07 %). From table 2 can shown too that the most patients have obesity and type 2 diabetes are genetics diabetes, female gender, dietary habit (no meet) and patients who has less active exercise in sport activities.

4.2 Propensity Score Stratification Analysis

4.2.1 Choose a covariate as a confounder

The first step in the propensity score analysis is to choose covariate as a confounder variable. The determination of confounding variables based on the theory and proven with empirical evidence like the relationship between variables. Testing relationship between variables used chi-square test. Based on research conducted by Betteng, et al.[5] known that obesity has a relationship with genetic factors, dysfunction of the brain, dietary habit is over, less activities of sport, emotional, environmental factors, social factors and lifestyle. Therefore, this relationship will be proven by empirical evidence using chi-square test . Results of testing the correlation between covariates with obesity variables are shown in Table 2.

TABLE 2. TESTING RESULTS CORRELATION BETWEEN COVARIATES WITH OBESITY

Variable	χ^2	**Df**	**P-*value***	**Decision**
$x_4 * x_1$	66,513	1	0,000	Reject H_0
$x_4 * x_2$	2,047	3	0,563	Failed to reject H_0
$x_4 * x_3$	0,358	1	0,549	Failed to reject H_0
$x_4 * x_5$	298,701	1	0,000	Reject H_0
$x_4 * x_6$	314,208	1	0,000	Reject H_0

Based on Table 2 can be shown that genetic, diet and active sports activities has significant influence to obesity variables. Meanwhile age and gender has not significant influence to obesity. Based on those results, so it is a proof that obesity variable is the most variable that associated with other variables. Therefore, obesity variable is selected as confounding variable *Z* with parameter θ.

4.2.2 Estimating the Propensity Scores

In this study the propensity score estimated by logistic regression. There are five variables will be estimated, their variables are genetic, age, gender, dietary habit and sports activities. The result of parameter is shown in Table 3.

TABLE 3. PARAMETER ESTIMATION FOR THE RELATIONSHIP OBESITY (Z) WITH COVARIAT (X)

Covariate	**Parameter** (β)	**SE**	**p-value**	**OR**	**OR (95% CI)**
Intercept	3.8357	1.4479	0.0081	33.9019	1.4069 - 16.8948
Genetic	2.3211	0.6562	0.0004**	10.7902	2.9338 - 39.6853
Age	0.0118	0.0192	0.5397	1.0174	0.9706 - 1.0665
Gender	0.1722	0.4500	0.7020	0.9835	0.3872 - 2.4980
Dietary habit	-1.8721	1.3741	0.1731*	0.1682	0.0114 - 2.49269
Sport Activities	-5.2426	1.6029	0.0011**	0.0057	0.0002 - 0.1328

(*) significant at α = 20%, (*) significant at α = 0,1%,

Based on Table 3 can be shown that the variables have significant influence to obesity at significance level (α = 0.1 %) are variable genetic with *p*-value = 0.000 and sport activities with *p*-value = 0.0011, while dietary habit variable is significance at α = 20%. It is indicates that the status of obesity patients DM was determined by genetic factors, dietary habit, and sports activities of patient DM.

From the estimation parameters are shown in Table 3, it can be obtained the value of propensity score below.

$$e(\mathbf{x}_i) = \frac{\exp\left(3,84 + 2,32\ Gen(1) + 0,01\ Age + 0,17\ Gndr(1) - 1,87\ DH(1) - 5,24\ SA(1)\right)}{1 + \exp\left(3,84 + 2,32\ Gen(1) + 0,01\ Age + 0,17\ Gndr(1) - 1,87\ DH(1) - 5,24\ SA(1)\right)} \quad (2.7)$$

Equation (2.7) illustrates that each age of patients DM is increase one year, so the odds of obesity will increase by 1,017 times. The probability of someone who have genetic DM become obesity is 10.79 times greater than someone who does not have a genetic history of diabetes, the probability of a women having obesity is 0.984 times greater than a men , the probability of someone a healthy diet having obesity is 0,168 times than someone whose diets are not healthy and active sports person's probabilitys having obesity is 0.006 times that of someone who rarely exercise.

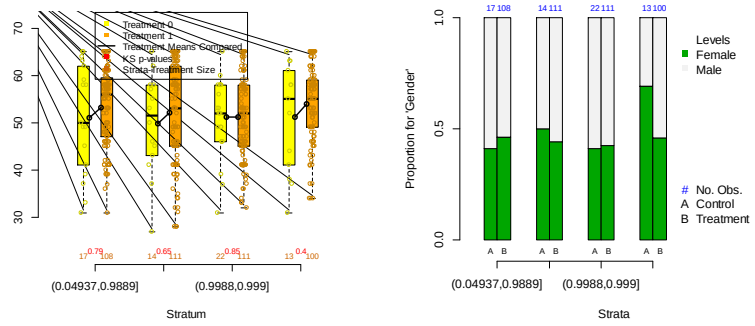
4.2.3 Stratify and Balance the Propensity Scores

After estimating the propensity scores, the next step is subclassified them into different strata. The formation of this stratum aims to balance the treatment and control groups so that estimates of treatment effect is not biased. The number of balanced propensity score strata depends on the number of observations in the data set. Table 4 shows the test of covariate balance after stratification based on the quintiles of the propensity score. Five of the covariates were included in the final propensity score model used for stratification. The initial imbalances were measured by chi-square test for categorical data (genetic, gender, dietary habit and sport activities) and Kolmogorov-Smirnov test for continuous data (age) comparing the obesity and no obesity groups.

TABLE 4. TEST OF STRATA BALANCE

Strata	n	Chi-Square Tests for Balance				KS-Test for Balance
		Genetic	Gender	Dietary Habit	Sport Activities	Age
1	126	0,058	0,800	0,954	0,525	0,790
2	125	0,052	0,780	0,525	1,000	0,650
3	133	0,055	1,000	1,000	0,475	0,850
4	113	0,062	0,150	1,000	1,000	0,400

Based on Table 4 can be shown that after testing using chi-square test for categorical data, their covariates such as genetic, gender, dietary habit and sports activities shows that obesity and no obesity have a balance at all strata. Similarly, for the covariates of age which was tested by Kolmogorov-Smirnov (KS) test. Covariate testing balance is supported by Figure 2. Figure 2 represents a picture which shows a balance between the obesity and no obesity for categorical data (gender) and continuous data (age). So that the analysis can be continued to the next step, the step is estimate average treatment effect or average effect of obesity on the type of DM. Pattern of balance can be seen in Figure 2 below.



(A) (B)

FIGURE 2. COVARIATE OF PROPENSITY SCORE IN BALANCE STRATA (A) AGE, (B) GENDER

4.2.4 Estimating the Causal Effect

Propensity score is an ideal method to see the effect of treatment on observational studies. This method can reduce bias effect because differences distribution of covariate between treatment and control groups. Therefore, before the estimated treatment effects, covariates between the treatment and control groups should be balanced. Because in the previous step has been obtained strata with covariates were balanced, then the next step is estimation of the treatment effect. In this case estimate of the effect of obesity on the type of DM. The estimation results for before and after stratification shown in Table 5.

TABLE 5. RESULT OF ESTIMATION AVERAGE TREATMENT EFFECT (ATE)

ODD RATIO FOR ATE					
BEFORE STRATIFICATION			AFTER STRATIFICATION		
UNADJUSTED	SE UNADJUSTED	95% CI STRATA	ADJUSTED	SE ADJUSTED	95% CI STRATA
16,859	0,3583	8,353 – 34,027	7,065	0,516	2,570 – 19,424

Table 5 shows the result for estimated effect of obesity on the type of DM before and after stratification. From table 5 obtained an average yield effects of obesity on the type of DM before stratification (unadjusted) is 16.859 with a standard error of 0.3585 and after stratification (adjusted) the effect of obesity is 7.065 with the standard error of 0.516. Propensity method also provides estimates of 95% confidence interval between 2.570 and 19.424. This confidence interval shown the difference average between the treatment group and the control of obesity is significant, or in other words, obesity significantly influence the type of DM with the effect is 7.065.

4.3 Analysis of Factors Affecting the Type of DM

After the estimation of treatment effects (obesity) was known then the next step is to determine the relationship of covariates with type of DM.

TABLE 6. PARAMETER ESTIMATION FOR THE RELATIONSHIP BETWEEN TYPE OF DM (Y) WITH COVARIAT (X)

Covariate	Parameter $(\boldsymbol{\beta}^*)$	SE	p-value	OR	OR (95% CI)
Intercept	2.8368	1.4367	0.0483	17.0611	1.0211 – 285,0692
Genetic (1)	21.2478	1375.7492	0.9877	1689671554	-
Age	-0.0448	0.0277	0.1059*	0.9562	0,9056 – 1,0095
Gender(1)	0.7892	0.5213	0.1301*	2.2016	0,7925 – 6,1162
Dietary habit(1)	-0.5024	1.3191	0.7033	0.6051	0,0456 – 8,0288
Sport Activities(1)	-1.3926	1.2916	0.2810	0.2484	0,0198 – 3,1234

(*) significant at $\alpha = 20\%$

Based on Table 6 can be shown that the variables significantly influence to the type of DM at significance level $\alpha = 20\ \%$ are variable age with *p*-value = 0.106 and gender with *p*-value = 0.1301. Based on the table 6 known that the type of DM patients was influenced by the age and gender of patients DM, or age and gender variable are variables that directly influence the type of DM patients.

From the estimation parameters are shown in Table 6, can be obtained logistic regression model covariates significant relationship between the type of DM as below.

$$\pi\left(\mathbf{x}_i\right) = \frac{\exp\left(2,837 - 0,045\,Age + 0,789\,Gender\left(1\right)\right)}{1 + \exp\left(2,837 - 0,045\,Age + 0,789\,Gender\left(1\right)\right)} \tag{2.8}$$

Equation (2.8) illustrates that any increase 1 year of age patients DM, the odds for type of DM decreased by 0.956 times and the probability for women having type 2 of DM is 2,202 times greater than men.

V. CONCLUSSION

Propensity score is a good method to see the effect of treatment on observational studies, particularly data with different background covariates. The different of covariate can make inaccurate conclusions. Propensity score stratification can balance the covariates between the treatment and control groups so that can reduce bias due to confounding effects. Analysis of propensity score stratification shown that the

variables influence obesity are genetic variable, sports activities and dietary habit of patients and the effect of obesity on the type of DM after stratification is amount 7.065 with a standard error of 0.516. In addition, the variables that directly influence the type of DM patients are obesity, age, gender and variable that does not directly affect the type of DM patients are genetic variable, sport activities and dietary habit of patients DM with the obesity as confounding factors if modeled by logistic regression.

REFERENCES

[1] Wild, S., Riglic, G., & Green, A. (2004). Global Prevalence of Diabetes: Estimates for the year 2000 and Projection 2030. Diabetes Care vol 27

[2] Kemenkes. (2013). *Riset Kesehatan Dasar 2013.* Badan Penelitian dan Pengembangan Kesehatan Kemenkes RI

[3] ----------. (2013). InfoDATIN Pusat Data dan Informasi Kemenkes RI: Situasi dan Analisis Diabetes

[4] Wicaksono, R.P. (2011). Faktor-Faktor yang Berhubungan dengan Kejadian DM tipe-II. Karya Ilmiah Kedokteran UNDIP

[5] Betteng, R., Pangemanan., D., & Mayulu, N. (2014) *Analisis Faktor Resiko Penyebab Terjadinya Diabetes Mellitus Tipe 2 pada Wanita Usia Produktif di Puskesmas Wawonasa.* Jurnal e-Biomedik Vol 2 No 2

[6] Trisnawati, S., Widarsa, T., & Suastika, K. (2013). *Faktor Resiko DM Tipe-2 Pasien Rawat Jalan di Puskesmas Wilayah Kec. Denpasar Selatan.* Public Health and Preventive Medicine Archive, Vol 1, No 1

[7] Indriyani, P., Supriyatno, H., & Santoso, A. (2007). *Pengaruh Latihan Fisik; Senam Aerobik terhadap Penurunan Kadar Gula Darah pada Penderita DM Tipe-2 di Wilayah Puskesmas Bukateja Purbalngga.* Jurnal Media Ners vol 1, No.2 pp 49-99

[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] McCaffrey, D.F., Ridgeway, G., & Moral, A.R. (2004). *Propensity Score Estimation with Boosted Regression for Evaluating Causal Effect in Observational Studies.* Pschological Method, 9(4), pp. 403.

[10] McCandless, L.C., Gustafson, P., & Austin, P.C. (2009). *Bayesian propensity score analysis for observational data. Statistics in Medicine*, 28, pp 94-112.

[11] 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.

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

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

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

[15] Mingxiang, L. (2012). *Using the Propensity Score Method to Estimate Causal Effects: A review an Practical Guide.* Organisational Research Methods 00(0) pp. 1-39

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

[17] Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). *Estimating Causal Effects of Public Health Education Campaigns using Propensity Score Methodology*. Journal Elsevier Evaluation and Program Planning 28 (2005) pp. 209–220.

[18] Poretsky, L. (2010). Principle of Diabetes Mellitus, Second Edition. New York: Springer

[19] Gungor, N., Hannon, T., Libman, I., Bacha, F., & Arslanian, S. (2005). *Type 2 Diabetes Mellitus in Youth: The Complete Picture to Date*. Pediatric Clinics of North America

M – 17

Analysis Propensity Score with Structural Equation Model Partial Least Square

Setia Ningsih[1], B. W. Otok[2], Agus Suharsono[2], Reni Hiola[3],
[1]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[2]Dept. of Statistics, Institut Teknologi Sepuluh Nopember
[3]Dept. of Public Health, Universitas Negeri Gorontalo
thya.setianingsih@gmail.com

Abstract— In research of epidemiology, structural equation modeling (SEM) has been become very popular, especially for latent variables. In SEM there are assumptions that must include the data should be normally distributed multivariate and a used large of data. For overcome these problems required the alternative approach of SEM based variance or partial least square (PLS). SEM-PLS does not require an assumption that a lot. In health sector randomization is not possible, because it concerns the lives of humans. So that assumptions independent can't be achieved. This can lead to imbalances covariates and selection bias. Therefore, to overcome these problems applied propensity score (PS). This method is a statistical analysis that can be used to analyze study design Non-Experimental where can't do randomization to treatment groups. Furthermore, as suggested new methods for handling selection bias is a marginal meanweighting through stratification (MMW-S). The analysis result obtained when using MMW-S is powerful because MMWS show strong reduction in of selection bias. The author uses an innovative method by using empirical data HIV/AIDS. Briefly using MMW-S with a predisposition, clinical manifestations, and opportunistic infection. And adherence to antiretroviral (ARV) as a confounding variable. The results showed that the method of MMW-S can removed bias more than 93.5%.

Keywords: *SEM-PLS, Propensity score, MMW-S, HIV/AIDS*

I. INTRODUCTION

Health problems are one of the factors that have an important role in creating quality human resources. In health, SEM has become a very popular method mainly used to examine the Latent variables. Non-Experimental studies or observational studies are empirical investigations of the effects caused by the treatment as randomized experiments Randomized Controlled Trials (RCT) is impossibble [1]. In general, RCT is very required in the research to the independence assumption so that the bias selection can be minimized. However, in the field of health research involving human, RCT is not always practicable. One method is suggested to be used for such problems is propensity score. Once the propensity score has been estimated in a given dataset, a data preprocessing procedure is performed to create comparability between study groups, it is referred to as pre-processing because it is performed before the final treatment effect is estimated, thus replicating the RCT by separating the study design stage from the outcomes analysis [2].

In general, this first entails stratifying the analytic sample into quantiles of the propensity score, and then generating a weight for each individual based on their corresponding stratum and treatment assignment, the stratification reduces bias in the observed covariates used to create the propensity score, and the weighting standardizes each treatment group to the target population [3]. This approach namely marginal mean weighting through stratification (MMWS), can handle a broad array of experimental conditions that researchers will likely encounter in evaluating health care interventions Once generated, the MMWS can then be used within the appropriate outcome model to estimate unbiased treatment effects [3].

II. LITERATURE REVIEW

A. Structural Equation Modeling Partial Least Square (SEM-PLS)

SEM-based variance or based components called as partial least square (PLS) is a method of analysis that is powerful and often referred to as soft modeling because it does not require assumptions such as

data should not normally distributed multivariate, can be used with data of nominal, ordinal, interval and ratio, in addition sample should not be large [4]. SEM-PLS consists of three components are outer model is specifies the relationship between variables latent and indicators or manifest variables (measurement model), inner model is specifies the relationship between the latent variables (structural model), and the weight relation.

PLS is a powerful modeling methods due to not assume the data must be with particular scale of measurement, samples should not be large, not require extremely assumptions. Types of indicators on the PLS are two as follows:

- Reflective indicators tend to be influenced by the latent variables (indicators is a reflection of the latent variables).
- Formative indicators tend to affect the latent variables (indicators are descriptors of latent variables).

Algorithm SEM-PLS as explained by [5], can be illustrated by Figure 1.

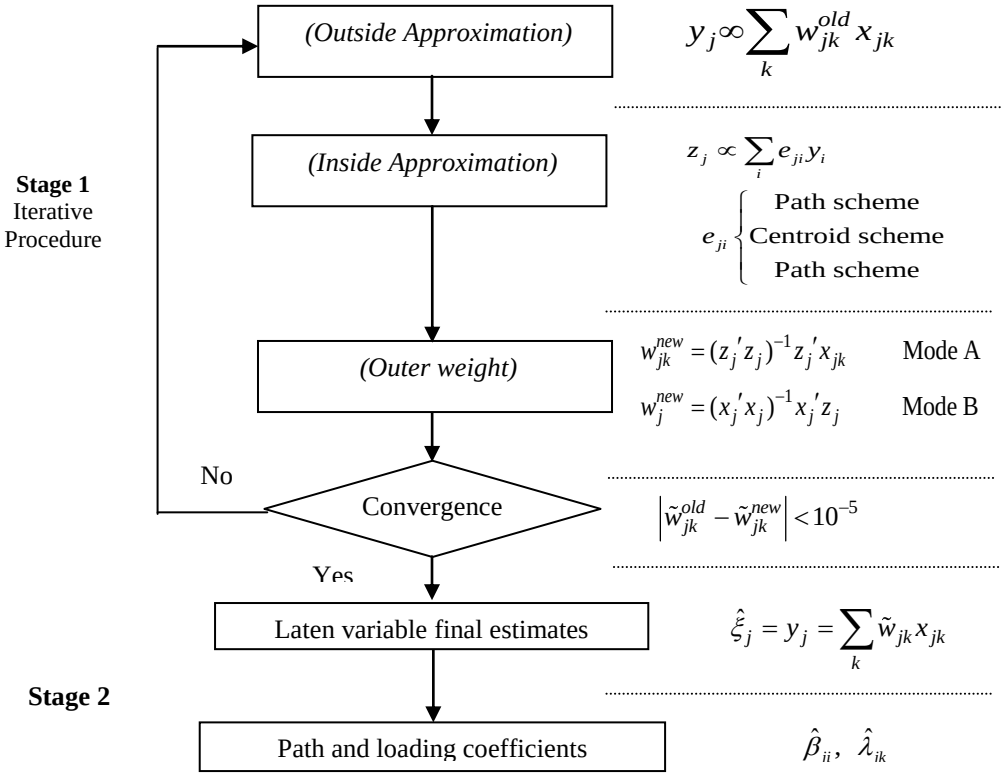


FIGURE 1. PLS ALGORITHM

Evaluation of PLS models is testing the validity and reliability. Validity testing performed to see the value of loading factor produced more than 0.5, if there are indicators has value loading factor below 0.5, then the are excluded from the analysis. Reliability testing show the composite value of each latent variable [6].

B. Propensity Score

The advantage of propensity score in comparison to multivariable adjustment is the separation of confounding factors adjustment and analysis of the treatment effect [7]. If the vector has many covariates were presented in many dimensions, then the propensity score can reduce all the dimensions into one dimension scores [2]. Rosenbaum and Rubin defined the propensity score for $i = 1, 2, \cdots, n$ as a conditional probability of being treated. Indicator of treatment group $(Z_i = 1)$ and control unit $(Z_i = 0)$ based on observed covariates vector (ξ_i). Propensity score can be written mathematically as follows:

$$e(\xi_i) = \mathrm{P}(Z_i = 1 | \xi_i) \quad (1)$$

The goal of propensity scoring is to balance the treated and untreated groups on the confounding factors that affect both the treatment assignment and the outcome, thus it is important to verify that treated and untreated patients with similar propensity score values are balanced on the factors included in the propensity score. Demonstrating that the propensity score achieves balance is more important than showing that the propensity score model has good discrimination [8].

C. Marginal Mean Weigthing through Stratification (MMWS)

Marginal mean weighting through stratification (MMW-S) was introduced as a flexible approach, combining propensity score weighting and propensity score stratification to remove imbalances of pre-intervention characteristics between two or more groups [3]. MMW-S produces more robust analysis than the methods of propensity score matching, propensity score stratification, and propensity score weighting [9].

$$MMWS = \frac{n_q \times \Pr(Z = z)}{n_{z=z,\mathrm{q}}} \tag{2}$$

Where

n_q = Number of individuals in each stratum

$\Pr(Z=z)$ = Probability of the treatment group

$n_{z=z,\mathrm{q}}$ = Number of individuals in each stratum is treated as treated / non-treated

D. HIV/AIDS

Human Immunodeficiency Virus (HIV) is a virus that reduces the body's immune system so that the people affected by this virus will be susceptible to various infections and then causes *Acquired Immune Deficiency Syndrome* (AIDS) [10]. The HIV is decreases gradually the immune system and leads to death as a direct result of one or more opportunistic infections. Opportunistic infection is an infection caused by immune deficiencies as a result of the HIV. Factors that influence the Opportunistic infection is predisposition and clinical manifestation. Predisposition is the internal factors that exist in individuals, families, communities that make easier individuals to behave. Clinical manifestations is presence indication of a disease that is perceived as complaints from patients and has been examined by a doctor or clinic. Predisposition include of age, level of education, work and marital status. And clinical manifestation include of CD4 and clinical stage.

III. METHODOLOGY

A. Source of Data

The data used in this research is secondary data on the medical records of HIV / AIDS patients in one hospital. The number is 91 patients HIV/AIDS. By using several variables as follows:

1. Exogenous Variables
   a Predisposition : age, level of education, work and marital status,
   b. Clinical manifestation : CD4 and clinical stage
2. Endogenous Variables: Opportunistic infection
3. Confounding variables: Adherence therapy ARV

B. Method of Analysis

Based on the research objectives, analysis methods used in this study is

1. Select confounding variables
2. Determine the propensity score approach to SEM
   a. Develop the conceptual model based on the theory
   b. Construct the path diagram
   c. Convert the path diagram into an equation system
   d. Estimate the parameters of model included
   e. Get the path coefficient value
   f. Determine the e propensity score value
3. Divide sample into Q strata based on propensity score and calculate the marginal mean weight

4. Examine the balancing of the covariates
5. Determine percentage bias reduction (PBR)

IV. ANALYSIS AND DISCUSION

A. *Select confounding variable*

Confounding variable according to the epidemiology is a situation where the size of the effect of distorted risk factors because of the correlation between exposure and other factors that influence the results [11]. The actual relationship between exposure factors and impact /disease factors are disappear or covered by other factors, so the influence of confounding factors can increase or decrease the actual relationship. Chi-square test was used to examine the relationship among variables, the following hypotheses [12]:

H_0: There is no significant relationship among variables

H_1: There is a significant relationship among variables

Significance level: $\alpha = 5\%$

Critical region: reject H_0 if $\chi^2 > \chi^2_{1-\alpha}; df = (i-1)(j-1)$ or p-value < α

TABLE 1. RELATIONSHIP BETWEEN ADHERENCE WITH PREDISPOSISI & CLINICAL MANIFESTASI

Variable	χ^2	*P-value*	Decision
ADH*Predisposition	5.315	0,021	reject H_0
ADH*Clinical manifestation	7.662	0,006	reject H_0

Based on the Table 1, can be seen that the adherence has a relationship with the predisposition and clinical manifestation. Therefore variable adherence is confounding variables. Diagram path after the formed variable interactions from compliance with adherence ARV the relationship among predisposition variables and adherence ARV the relationship among clinical manifestations of the opportunistic infection can be seen at figure 2.

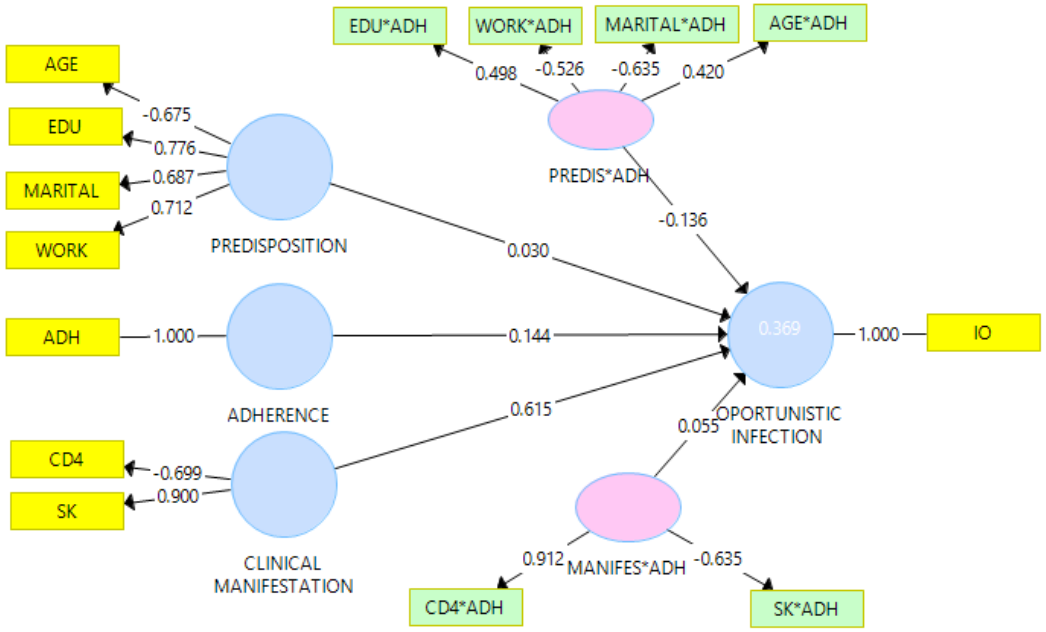


FIGURE 2. DIAGRAM PATH VARIABLE CONFOUNDING

Based on the figure 2 diagram path after putting confounding variables, the structural equation model is:

Opportunistic Infection = 0.030 predisposition + 0.144 clinical manifestation + 0.615 adherence
- 0.136 (Predis*adh) + 0.050 (manifes*adh)

B. Calculating of MMWS

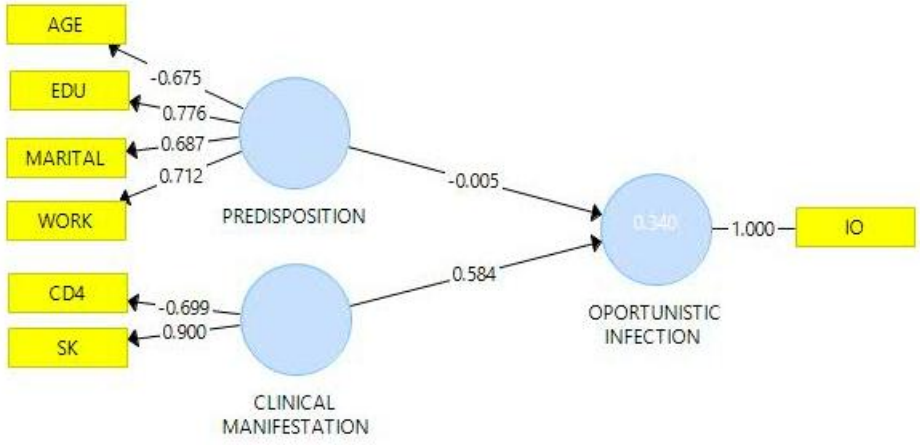


FIGURE 3 : LOADING FACTOR OF EACH LATEN VARIABLE

Based on the figure 2, can be seen that the loading factor of each indicator more than 0.5. Hence can be concluded that a valid indicator to measure the construct predisposition and clinical manifestation. Then calculate the propensity score using SEM-PLS. The propensity score for all respondents are used to divide respondents into five strata. Furthermore, calculating the marginal mean weight to tread groups and untreated groups uses the equation recommended as follows [9].

$$\frac{n_s \times \Pr(Z = z)}{n_{z=z,s}} \tag{3}$$

Where, n_s is the total number of individuals in stratum s, $\Pr(Z = z)$ is assignment probability to treatment groups z. $n_{z=z,s}$ is the total number of individuals in stratum s which is the actual treatment assignment for z.

TABEL 2. THE CALCULATION OF THE MMWS

Stratum	Sample	Unweighted sample		MMWS		Weighted sample	
		Treated	Untreated	z	z'	z	z'
1	19	10	9	0.626	1.42	6	13
2	18	6	12	0.989	1.01	6	12
3	18	4	14	1.484	0.86	6	12
4	18	8	10	0.742	1.21	6	12
5	18	2	16	2.967	0.75	6	12

After calculate MMWS can increase the homogeneity of propensity score between the treatment group and the control group in each stratum. The homogeneity or balance covariates in each stratum using the t-statistic for numerical variables. Insignificant T-values indicate adequate MMWS. The results of cheking balance covariate of each stratum are presented in Table 3.

TABLE 3: *T*-TEST FOR CHECKING BALANCE

Stratum	Predisposition			Clinical Manifestation		
	***T*-value**	$T_{(df,\alpha)}$	**Decision**	***T*-value**	$T_{(df,\alpha)}$	**Decision**
1	0.390	6.314	Not reject H_0	0.911	1.812	Not reject H_0

2	0.298	1.655	Not reject H_0	0.495	1.652	Not reject H_0
3	1.196	6.314	Not reject H_0	0.367	1.703	Not reject H_0
4	1.582	6.314	Not reject H_0	0.747	1.648	Not reject H_0
5	1.116	1.687	Not reject H_0	1.216	1.653	Not reject H_0

Based Table 3 shows that after MMWS there is no difference between the treatment group and control. Furthermore, compute a percentage bias reduction (PBR) on the covariate is another criterion to assess the effectiv of MMWS.

$$PBR = 100 \times \frac{\left(\bar{x}_t - \bar{x}_c\right)_{beforeMMWS} - \left(\bar{x}_t - \bar{x}_c\right)_{AfterMMWS}}{\left(\bar{x}_t - \bar{x}_c\right)_{AfterMMWS}} \quad (4)$$

Based on calculate of percentage bias reduction (PBR) obtained 93.5%. So that MMWS able to reduction bias 93.5%. It this sufficient of the bias reduction based on the examples in Cochran and Rubin PBR value of 80% or higher is satisfactory.

V. CONCLUSION

SEM-PLS can see that loading factor for each indicator on each latent variable is greater than 0.5, so that the indicators (age, level of education, work and marital status) were able to explain the predisposition variable and indicators (CD4, clinical stage) capable explain the clinical manifestation variable. Furthermore, score factor of each of the latent variables used to calculate a propensity score that will be used at this stage of Marginal mean through weighting stratification (MMWS) to reduce bias due to confounding variable. Marginal mean weighting through stratification (MMWS) method is a powerful because MMWS showed a reduction from the selection bias more than 93.5%.

.

REFERENCES

[1] 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 41-55J*.

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

[3] Linden, Ariel. (2014). Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation Clinical Practice 20, 1065-1071.*

[4] Ghozali, Imam. (2011). *Structural Equation Modelling Metode Alternatif dengan Partial Least Square*. Badan Penerbit Universitas Diponegoro, Semarang

[5] Trujillo, G.S. (2009). *PATHMOX Approach: Segmentation Trees in Partial Least Squares Path Modeling*. Universitat Politècnica de Catalunya.

[6] Chin, W.W. (1998). *The Partial Least Squares Approach for Structural Equation Modelling. Modern Method for Business Research*. London: Lawrence Erlbaum Associates.

[7] Littnerova Simona, Jarkovsky Jiri, Parenica Jirib, Pavlik Tomas, Spinar Jindric, Dusek Ladislav. (2003) Why to use propensity score in observational studies? Case study based on data from the Czech clinical data base AHEAD 2006–09. *Original Research Article. Coret Vasa 55. pp. e 383 – e 390*

[8] Crowson Cynthia S. Schenck Louis A., Green Abigail B., Atkinson Elizabeth J., Therneau. (2013). *Terry M The Basics of Propensity Scoring and Marginal Structural Models*. Mayo Clinic, Rochester, Minnesota

[9] Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multi-level data. *Journal of Educational and Behavioral Statistics, 35 (5), 499-531*.

[10] Djauzi. S. dan Djoerban, Z. (2003). *Penatalaksanaan Infeksi HIV di Pelayanan Kesehatan Dasar. Edisi II*. Jakarta: Balai Penerbit FK UI; 2003.

[11] Wunsch, Guillaume. (2007). Confounding and control. *Demografi Research Volume 16. Article 4, page 97-120* Published 06 Februari 2007

[12] Daniel, W. W. (1978). *Statistik Nonparametrik Terapan.* Jakarta: Gramedia.