Modelling of Cancer Treatment with Activity-Meter Records Using Linear and Binary Logistic Regressions

Heri Kuswanto, Taufik Afif Maldini

Abstract*: Cancer is a term for diseases in which cells grow abnormally without control and can attack surrounding tissue. In the recent years, cancer has become the leading cause of death in the world. The most commonly used cancer treatment is chemotherapy with a series of 5-Fluorouracil (5-FU) based therapies. This study aims to determine the factors which influence the success of chemotherapy. The statistical method used in this study is lineaar regression analysis and binary logistic regression. The results of linear regression revealed that that the dose of 5-FU 1 compound and the patient's body mass index had a significant effect on the level change of White Blood Cells (WBC) with a significant level of 5%, while the dose of Irinotecan compounds had a significant effect at a significance level of 20%. The change in the percentage of neutrophil of patients is affected by the age of the patient. Moreover, the average hours of deep sleep is affected by the age of the patient and the dose of compound P (Panumumab). The logistic regression analysis showed that the patient's age, the average number of patient's footsteps and the dose of 5 FU-1 given during chemotherapy had a significant effect on WBC changes after chemotherapy with a significance level of 10%. The logistic regression model is able to correctly predict the WBC change with 75.44% AUC.

***Index Terms*: chemotheraphy, AUC, logistic.**

I. INTRODUCTION

Cancer is a term for diseases in which cells grow abnormally without control and can attack the surrounding tissue. This process is called metastasis, which is the main cause of cancer deaths. In recent years, cancer is a leading cause of death rates and mortality rates worldwide. According to the Worlh Health Organization (WHO), cancer is the second leading cause of death in the world and accounted for 8.8 million deaths in 2015 [1].

One of the methods that is commonly used for cancer treatment is chemotherapy. Chemotherapy is known as a treatment to kill cancer cells [2]. Chemotherapy is done by injecting an anti-cancer drug into a blood vessel in the hand. This therapy relies on the ability of special drugs to destroy cancer cells that attack the body. These drugs work by slowing down or stopping the growth of cancer cells. The drug component that is widely used in the treatment of cancer is 5-Fluorouracil [3].

A research on the case of using 5-FU in chemotherapy was carried out by Tamoki [4]. The study discussed the extraction of side effects given after 5-FU-based chemotherapy. This research is a relatively new study in the field of chemotherapy. In this study further analysis on the effects of chemotherapy combined with activity meters that are attached to the patient's body after chemotherapy is conducted. For 14 days after undergoing chemotherapy, activity meter was used to measure the number of footsteps and hours of sleep in patients. In contrast to the research conducted by [4] which used machine learning methods, the present study is conducted to determine factors influencing the success of chemotherapy with a series of 5-FU-based therapies, using standard statsitical models.

Based on the description above, this study will analyze the factors that are considered to influence the change in the percentage of neutrophils contained in White Blood Cells (WBC), and the average hours of deep sleep after chemotherapy using the linear regression method. Furthermore, in this study the changes in White Blood Cells (WBC) after chemotherapy is classified into two classes. Logistic regression method is applied to find out the factors that influence the results of the classification.

II. THEORY AND METHODOLOGY

A. Linear Regression

Linear regression is a method to model the relationship between response variable (y) and predictors ($x_1,x_2,x_3,\ldots,x_p$). The linear regression model with p predictors can be written as [5].

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \quad (5)$$

where : $i = 1,2,3,\ldots,n$

y_i = value of response variable on i-th observation

X_{ik}= value of k-th predictor variable on i-th observation, with $k = 1,2,\ldots,p$

β_0= intercept of the model

β_k= coefficient regression of the k-th variable

ε_i = *error* on i-th observation assuming to be identic, independent and normally distributed with zero mean and constant varians.

The regression parameters are estimated by using Ordinary Least Square (OLS). The OLS method minimizes the sum of error square, and resulted on the following estimator [5]:

Revised Manuscript Received on June 13, 2019

Heri Kuswanto, Department of Statstistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

Taufik Afif Maldini, Department of Statstistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (6)$$

where

$\hat{\beta}$ = vector of the estimated parameters with the size of $(p + 1) \times 1$

X = matrix of predictors with size of $n \times (p + 1)$

Y = vector of response variable with size of $n \times 1$

Tesing the parameters significant

1) Simultaneous test

The simultaneous test is a test of model significance in order to test whether the predictors simultaneously influence the response. The hypothesis is

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : at least one $\beta_k \neq 0, k = 1,2,\ldots,p$

The statistical test is.

$$F_{test} = \frac{MS_{regression}}{MS_{error}} \quad (7)$$

and the null hypothesis H_0 is rejected if $F_{test} > F_{\alpha\,;\,p,(n-p-1)}$.

2) Partial test

Partial test is applied to each predictor individually. The hypothesis of partial test is

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0, k = 1,2,\ldots,p$

and the statistical test is

$$t_{test} = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \quad (8)$$

where $SE(\hat{\beta}_k)$ is the standard error of $\hat{\beta}_k$.

The test rejects the H_0 if $|t_{test}| > t_{\alpha/2;(n-p-1)}$, meaning that the predictor has significant influence to the response [5, 6].

B. Binary logistic regression

In the logistic regression, the relationship between predictor variables and response can be written as

$$\pi_i(x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}$$

$$\pi_i(x_i) = \frac{exp({x_i}^T \beta)}{1 + exp({x_i}^T \beta)}$$

The parameter estimation in logistic regression model is conducted with Maximum Likelihood Estimation (MLE). The MLE estimates $\widehat{\boldsymbol{\beta}}$ whick maximizes the likelihood, resulting on the following estmator

$$\hat{\beta} = (X^T \mathrm{W} X)^{-1} X^T \mathrm{W} z \quad (9)$$

In logistic regression model, the retalitionship between predictor and response is not linear.

Testing the significant of the parameters

The parameters are tested with the same procedure as linear regression i.e. involving of simlutaneous test and partial (individual test). The simultaneoaus test is done with Likelihood Ratio test, while the partial test is done with Wald test.

1) Simultaneoaus test

The simultaneous test in logistic regression is used to test the following hypothesis:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_i = 0,$

H_1 :at least $\beta_i \neq 0 ; i = 1,2,..p$

The statistical test is :

$$G = -2\,ln \frac{\left(\frac{n_1}{n}\right)^{n_i} \left(\frac{n_1}{n}\right)^{n_0}}{\sum_{i=1}^{n} \widehat{\pi_i}^{y_i} (1-\widehat{\pi_i})^{(1-y_i)}} \quad (10)$$

where, $n_i = \sum_{i=1}^{n} y_i$; $n_0 = \sum_{i=1}^{n} (1 - y_i)$; $n = n_1 + n_0$

The statistic G is Likelihood Ratio Test where G follows Chi-Square distribution so that H_0 is rejected if $G > \chi^{\alpha}_{(x,\alpha)}$ with v is the degree of freedom without β_0 [7].

2) Partial test

The partial test is done with Wald Test for the following hypothesis:

$H_0 : \beta_j = 0, j = 1,2,\ldots,p$

$H_1 : \beta_j \neq 0$

Statistical test : $$W^2 = \frac{\hat{\beta}_j^2}{SE(\hat{\beta}_j)^2} \quad (11)$$

with $SE(\hat{\beta}_j)^2 = \sqrt{var(\hat{\beta}_j)}$

The statistics above follows chi-squre distribution, and hence, the H_0 if rejected if $W^2 > \chi^2_{(v,\alpha)}$ with v is degrees of freedom [7].

C. Data and Variable

The data used in this study are secondary data about records on patients undergoing 5-FU-based chemotherapy, obtained from a private hospital. In these data, clinical studies were conducted on 28 patients who underwent chemotherapy with 5-FU. After chemotherapy, an activity meter placed on the patient's body to record the number of patient's footsteps and the patient's sleep hours used as variables in this study, as well as other patient's characteristics.

The variables used in this study consist of two variables, namely the response variable (Y) and predictor variable (X). In the linear regression analysis, three response variables were used, namely level changes in White Blood Cells (Y1), Neutrophil changes (Y2) and the average hours of deep sleep without sounding the patient (Y3). Whereas in the binary logistic regression, the analysis uses changes on total White Blood Cells count (Y4) as the response.

The models use 12 predictor variables consisting of categorical variables namely patient gender (X1) and patient body mass index (BMI) (X3). The numerical predictor variables are patient age (X2), average number of footsteps per day (X4), compound 5-FU1 (X5), 5-FU2 (X6), levamisole (X7), oxaliplatin (X8), irinotecan (X9), bevacizumab (X10), capecetabine (X11) and general committee (X12).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. RESULTS

A. Characteristic of Chemoteraphy Patients

This section describes the patients' characteristics regading the variables used in the analysis. Table 1 shows the descriptive statistics of three responses i.e. change on WBC and Neutrophil levels as well as deep sleep of patient after chemotheraphy.

Table 1. Descriptive statistics of patients' characteristics

Variable		Mean	Standard Deviation	Minimum	Maximum
WBC	increase	869	528	100	2000
	decrease	-933	600	-2200	-100
Neutrophil	Increase	730	401	148	1210
	decrease	-698.2	420.6	-1479.3	-39.4
Deep sleep	No	224.1	74.20	104.50	346.90
	Yes	0			

Furthermore, the descriptive statistics of several predictors are presented in Table 2.

Table 2. Descriptive statistics for predictors

Variable	Mean	Standard deviation	Minimum	Maximum
X_2	64.96	11.06	39.00	84.00
X_4	5333.00	2243.00	1753.00	10726.00
X_5	641.60	80.90	450.00	750.00
X_6	3823.20	480.80	2800.00	4650.00
X_7	327.50	31.02	250.00	380.00

The proportion of patients based on their gender is given in the pie chart on Fig. 1 showing that the majority (79%) of the patients are male (green color) and only 21% of them are female (blue color).

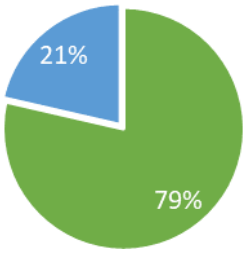

Fig 1. Proportion of the patients based on gender

B. Data Analysis with Multiple Linear Regression

The multiple linear regression is fitted for each predictor separately. Simultaneous test using F-test is carried out prior to examining univariate test. The results of simultaneous test is presented in Table 3.

Table 3 Significant test of linear regression model

Response variable	F	p-value
Y_1	3.68	0.014
Y_2	3.65	0.067
Y_3	3.18	0.059

We see that the P-value for each regression is significant with significant level of 0.1. Therefore, we conclude that there is at least one predictor significantly influence the change level of the White Blood Cells (Y_1), change of neutrophil (Y_2) and average deep sleep duration of the patients (Y_3). Furthermore, partial test is conducted to test the significant influence of each predictor to the responses. Table 4 shows the results of partial test for each response as the results of applying stepwise procedure.

Table 4 Estimation and hypothhesis testing of linear regression

	Predictor	Coef	T	P	VIF
Y_1	Constant	2079	1.37	0.184	
	5-FU 1	-5.47	-2.20	0.038	1.42
	BMI(2)	1185	2.07	0.050	2.96
	BMI(3)	2669	3.65	0.001	2.87
	BMI(4)	1663	2.46	0.022	2.44
	IRI (Irinotecan)	-4.12	-1.53	0.139	1.09
Y_2	Constant	-553	-2.30	0.029	
	Irinotecan	4.13	1.91	0.067	1.00
Y_3	Constant	88.6	1.09	0.288	
	Age	2.30	1.82	0.081	1.12
	P (panitumumab)	-0.2199	-2.24	0.034	1.12

From the table, we see that the 5-FU1 and Body Mass Index (BMI) significantly influence the level of WBC change as the P-values are less than or equal to 0.05 (with significant level of 5%), while IRI (Irinotecan) did not significantly influence the level of WBC change. The VIF values confirm that there is no multicollinearity among predictors. The R-square for this model is 45.56%.

The Irinetocan dose has influence to the change of neutrophil (Y_2) with significant level of 10%, and very low coefficient of determination (12.32%). Meanwhile the deep sleep (side effect) is influenced by the age of patient and the panitumumab dose, with also a very low R-square (20.28%). The regression models for level change of White Blood Cells (Y_1), change of neutrophil (Y_2) and average deep sleep duration of the patients (Y_3) can be written as follows, respectively.

$$Y_1 = 2079 + 1185X_{3(2)} + 2669\ X_{3(3)} + 1663\ X_{3(4)} - 5.47\ X_5 - 4.12\ X_9$$
$$Y_2 = -553 + 4.13\ X_8$$
$$Y_3 = -1489 + 2.30\ X_2 - 0.2199\ X_{12}$$

C. Data Analysis with Logistic Regression Model

The logistic regression is focused on investigating the change of WBC level which is categorized into two classes i.e. decreasing and increasing. The proportion of each class can be seen in Fig. 2 below.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

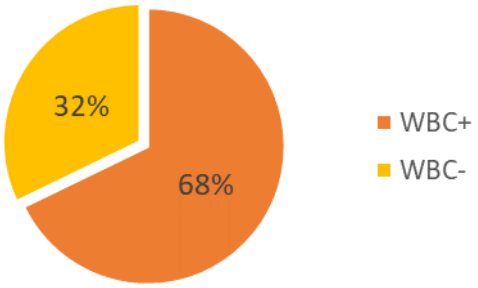

Fig 2. Proportion of classes of WBC change

The figure clearly shows that the proportion of the class response is unbalance i.e. 32% for decreasing WBC, and 68% for icreasing WBC). Furthermore, the backward-stepwise (Wald) procedure is applied to investigate the significant variables. The best model resulted from Backward Stepwise (Wald) is given in Table 5.

Table 5. Best regression model

Variable	B	P-*value*	OR	95% C.I for OR	
				Lower	Upper
Age	-0.275	0.059	0.759	0.570	1.011
Step	-0.001	0.034	0.999	0.997	1.000
5-FU 1	-0.026	0.073	0.974	0.947	1.002
Constant	44.242	0.046	$1.637x10^{16}$		

From the table, we see that the increasing or decreasing of WBC level is influenced by the age of patients, footsteps and the dose of 5-FU1. The age of patients has negative influence to WBC change, meaning that the older the patients, the WBC level of the patients will tend to decrease. The same results is obtained for step and 5-FU1 dose. The logistic regression model can be written as follow:

$$\hat{g}(x) = 44.242 - 0.026x_2 - 0.001x_4 - 0.275x_5$$

Classification Results Using Logistic regression

The performance of the logistic regression model can be seen from the classification accuracy through the AUC value [8]. The confusion matrix showing the observed and predicted class can be seen in Table 6.

Table 6. Confusion matrix

observation		Prediction	
		Change of WBC	
		Decreasing	Increasing
Change of WBC	Decreasing	6	3
	Increasing	3	16

From the table, the AUC value is 75.44%, menaing that the model can predict or classify the change of WBC correctly with the rate of about 75%.

IV. CONCLUSION

The univariate analysis using stepwise multiple linear regression found that the dose of compound 5-FU1 and BMI of the patients are significant in the model at a significance level of 5%, while the dose of irinotecan is significant at a significance level of 20%. The Neutrophil changes after chemotherapy is influenced by the dose of Irinotecan (IRI) compound with a significance level of 10%. While the hours of deep sleep is affected by the dose of the compound Panitumumab (P) with a significance level of 5% and the age of the patient at a significance level of 10%. The indicator that has a significant effect on changes in White Blood Cells (WBC) is the age of the patient, the average number of steps and the compound dose of 5-FU1, with the accuracy of the classification of changes in White Blood Cells by 75.44%.

The results of the analysis in this study suggest that to suppress the effects of chemotherapy and the hours of deep sleep are advised to use the compound Panitumumab (P). Chemotherapy based on 5-fluorouracil (5-FU) compounds is recommended to be combined with irinotecan (IRI) to reduce the number of white blood cells and the number of neutrophils in cancer patients. Moreover, in order to decrease the number of White Blood Cells (WBC), patients are also advised to intensify walking because the number of footsteps can affect the decrease in the number of White Blood Cells (WBC).

Considering the fact that the number of observations in the analysis is too small, further analysis by using more advance methods such as logistic regression ensemble [9,10], or bayesian approach [11,12] can be considered in order to improve the prediction performance.

REFERENCES

1. World Health Organization. ”Cancer: Fact Sheets” (February 10th, 2018) [online]. Availabe: httpp://www.who.int/mediacentre/factsheets/fs297/en/index.html
2. National Cancer Institute. “Chemotherapy” (February 17th, 2015) [online]. Available https://www.cancer.gov/about-cancer/treatment/types/chemotherapy
3. Longley, D.B. & Johnston, P.G. “5-Fluorouracil Molecular Mechanisms of Cell Death” in Srivastava R., Apoptosis, Cell Signaling, and Human Diseases, Humana Press, 2007.
4. Tamoki, M. “Extraction of side effect characteristics given by 5-FU using machine learning”, unpublished report, 2017.
5. Johnson, R.A. & Wichern, D.W. Applied Multivariate Statistical Analysis, 6th edition. New Jersey: Printice Hall, 2007.
6. Drapper, N. R. & Smith, H. Applied Linear Regression. John Wiley and Shon, New York, 1998.
7. Hosmer, D.W., Lemeshow, S. & Sturdivant, X.R. Applied Logistic Regression 3rd Edition. New Jersey: John Wiley & Sons, 2013.
8. Bekkar, M., Djemaa, H. K. & Alitouche T. A. “Evaluation Measures for Models Assessment over Imbalanced Data Sets”, J. Inf. Eng. Appl. 3, 2013, pp. 27–38.
9. Kuswanto, H., Asfihani, A., Sarumaha, Y. & Ohwada, H. “Logistic regression ensemble for predicting customer defection with very large sample size”, Procedia Computer Science 71, 2015, pp. 86-93.
10. Kuswanto H., Werdhana, W. “Classification of Alzheimer related genes using LORENS with important and significant features”, Internetworking Indonesia Journal 10(1), 2018, pp. 29-34.
11. Astuti, A.B., Iriawan, N., Irhamah, Kuswanto, H. “An algorithm for determining the number of mixture components on the bayesian mixture model averaging for microarray data”, Journal of Mathematics and Statistics 11(2), 2015, pp. 45-51.
12. Marty, R., Fortin, V., Kuswanto, H., Favre, A.-C., Parent, E. “Combining the bayesian processor of output with bayesian model averaging for reliable ensemble forecasting”, Journal of the Royal Statistical Society. Series C: Applied Statistics 64(1), pp. 75-92