Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n1p75

Modelling the change of white blood cell on colorectal cancer treatment using probit regression
By Kuswanto et al.

Published: 02 May 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 01, May 2020, 75-85
DOI: 10.1285/i20705948v13n1p75

Modelling the change of white blood cell on colorectal cancer treatment using probit regression

Heri Kuswanto*[a], Nesia Balqis[a], Hayato Ohwada[b], and Setiawan Toha[a]

[a] *Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS Sukolilo, Surabaya 60111 Indonesia*
[b] *Department of Industrial Administration, Faculty of Science and Technology, Tokyo Universty of Science, Chiba-Japan*

Published: 02 May 2020

Colorectal cancer has become one of the cancer types with high incidence rate all over the world. Various efforts have been carried out to find a way to decrease the risk of cancer. Chemotherapy using 5-Fluorouracil (5-FU) is one of the common cancer treatments that is expected to drive the white blood cell (WBC) level into the normal level. This research investigates the factors influencing the change of WBC level in cancer patients treated with 5-FU combined with physical treatment in the form of footsteps. By focusing on the change of WBC level, i.e. decreasing or increasing the WBC level as the response, probit regression was applied to the data measured from 28 cancer patients who have undergone 14 days of treatment. The probit regression found that age of the patient, average number of daily footsteps and the dose of 5-FU signficantly influence the change of WBC. The regression is able to classify the case with a satisfactory results, i.e. 85.71% classification accuracy. This finding can be a guideline to better treat the colorectal cancer patient to reach a normal WBC.

keywords: classification, footstep, WBC, chemotherapy.

*Corresponding author: heri_k@statistika.its.ac.id

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

Cancer is one of the chronic diseases which gained a highly increasing incidence rate over years. Cancer is used to describe a disease in the form of abnormal cells in the body that exceed the limit. These cells can attack other parts of the body. Cancer has become a serious health problem in both developed and developing countries. According to the statistics reported by the World Health Organization (WHO), cancer has been the second leading cause of death globally, causing about 9.6 million deaths by 2018. Lung, prostate, colorectal, stomach and liver cancer have been identified as the most common types of cancer in men. Meanwhile, women commonly suffer from breast, colorectal, lung, cervix and thyroid cancer. Colorectal cancer is a cancer with the third largest incidence in the world.

The medical treatment for a cancer case is commonly done through chemotherapy, i.e. a type of cancer treatment that uses certain drugs to kill the cancer cells. Chemotherapy works by stopping or slowing the growth of cancer cells that usually grow and spread rapidly. For some patients, chemotherapy can be the only method that will be carried out in handling the cancer case. In case of colorectal cancer, the drug component that is widely used in the treatment is 5-Fluorouracil (see Longley et al. (2003) for more detail about this). An experiment has been conducted by a team in Japan in 2017 by placing an activity meter on the patient's body to measure several outputs of interest related to the patient's performance, such as the footsteps, hours of deep and shallow sleep, blood pressure, etc. The experiment is a relatively new study in the chemotherapy field. The experiment was conducted within 14 successive days and intended to investigate the impact of footsteps on the cancer treatment performance, beside the 5-FU. Moreover, the experiment was intended to measure the side effects of the treatment, as measured by deep and shallow sleep. This present study focuses on looking at the main effect of the treatment.

There have been a lot of studies conducted related to cancer detection. Panetta (1995) applied a logistic model of periodic chemotherapy using a differential equation. The time-varying periodic parameter was used to model the growth of cells, in particular cancer cells, in the presence of chemotherapeutic drugs. A study by Zhou et al. (2004) proposed a logistic regression to classify the cancer-related genes selected using a Bayesian method. Gibbs sampling and Markov chain Monte Carlo (MCMC) methods were used to discover the most important genes. Zangmo and Tiensuwan (2018) applied logistic regression to model factors influencing the survival of cancer patients in Bhutan. Madhu et al. (2014) used multinomial logistic regression to study the influence of residence and socio-economic status on breast cancer incidence in Southern Karnataka. Logistic regression has also been used by Seddik and Shawky (2015) for diagnosing breast cancer based on a set of input variables that describe some characteristics of tumor images. The most recent work of Chang et al. (2018) uses Bayesian logistic regression to predict breast cancer cases. Meanwhile, probit regression has been applied in many health studies such as Ruspriyanty and Sofro (2018), Gitto et al. (2015) among others. More recent development on logistic regression model can be found in Algamal (2017) and Algamal and Lee (2019) who discuss adaptive penalized logistic regression and two-stage sparse

logistic regression. Both models have been successfully applied to classify gene expression microarray data.

Most of the previous studies above were mainly focused on detecting the cancer. This present study differs signficantly from the previous research in some areas, i.e. this study investigates the impact of footsteps as one of the predictors and uses probit regression as the tool to model the colorectal cancer case. Moreover, this study focuses on investigating the change of WBC level. The hypothesis of measuring the impact of footsteps on cancer patients has also been investigated by several researchers. Wilson et al. (2005) investigated the impact of walking intervention as part of an exercise for breast cancer patients in African American women. They found that increasing walking for exercise can improve body mass index and anthropometric measures which are associated with reduced risk of cancer recurrence. By using the Cox proportional hazards model, Wiliam (2013) tested prospectively whether post-diagnosis running and walking differ significantly in their association with breast cancer mortality. The study found that walking and running exercise can significantly reduce the risk of cancer mortality. Frensham et al. (2018) investigated the effect of a 12-week online walking intervention on health and quality of life in cancer survivors. The study revealed that this intervention improves the health quality of the cancer patients such as mental health, physical fitness, systolic blood pressure, etc. Therefore, it is important to investigate the impact of footsteps to WBC change.

The structure of the paper is as follows. The next section provides a brief description about Probit regression. The data and variables are described in section 3. In section 4, we perform the results of the analysis. Section 5 concludes the paper.

2 Probit Regression

Probit regression was firstly introduced by Chester Itner Bliss in 1934 in toxicology field (Casella and Berger, 2002). Probit regression is a form of logistic regression where the response variable used is the categorical data type. In probit regression, the link function used is a standard normal cumulative inverse function (or probit). The use of this method has several advantages. For some multivariate cases, the goodness in models with random effects will increase with the use of probit regression rather than logistic regression. Likewise, if there is overdispersion, the use of probit regression will improve the goodness of the model (Hahn and Soyer, 2007). The estimation of parameters in the probit model is derived from the normal cumulative distribution function (CDF)(Gujarati and Porter, 2003).

Suppose there is a response Y assumed to be derived from a variable Y^* with the equation (1) below

$$y^* = \beta^{\mathbf{T}}\mathbf{x} + \epsilon \tag{1}$$

where $\beta = [\beta_0 \quad \beta_1 \quad \beta_2 \quad ... \quad \beta_q]^T$, $\mathbf{x} = [1 \quad x_1 \quad x_2 \quad ... \quad x_q]^T$ and q is the number of predictor variable x, the β has the size of $(q+1)$x1 and ϵ is assumed to be standard

normal distribution. The formation of categories for the response variable in bivariate probit model involves of determining a certain threshold for Y^* for instance γ such that:

$$Y = 0 \text{ if } y^* \leq \gamma$$
$$Y = 1 \text{ if } y^* > \gamma.$$

The probability p_i of choosing any alternative over not choosing it can be expressed as in (2) where ψ represents the cumulative distribution of a standard normal random variable:

$$p_i = P[Y = 1|X] = \int_{-\infty}^{x_i'\beta} (2\pi)^{-1/2} \exp\left(\frac{-t^2}{2}\right) dt = \Phi(x_i'\beta) \tag{2}$$

The significance of the β parameters in the probit model can be tested by using the Likelihood Ratio test with the following hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0$, $H_1 :$ there is at least $\beta_i \neq 0$. The likelihood ratio test compare sthe maximized loglikelihood values for the restricted and unrestricted coefficient estimate of β. Furthermore, to test the significance of the individual β can be done with the Wald statistic which computes hypothesis tests using only the unrestricted probit coefficient estimates β.

One of the important features of probit regression is marginal effect, indicating the relationship between a specific predictor variable and the response. This marginal effect accounts partial change of the outcome (response) in the probability. The marginal effect of a continuous predictor x_k on the probability $P[Y_i = 1|X]$ can be expressed as

$$\frac{\partial p_i}{\partial x_{ik}} = \phi(x_i'\beta)\beta_k$$

where ϕ represents the probability density function of normal standard distribution. The marginal effect above is interpreted by holding other variables constant. Meanwhile, the marginal effect on dummy variables indicates the discrete change in the predicted probabilities, which can be written as

$$\Delta = \Phi(\bar{x}\beta, d = 1) - \Phi(\bar{x}\beta, d = 0)$$

where d is the dummy with its corresponding indicator.

The goodness of fit of the probit regression model can be assessed through the Akaike Information Criteria (AIC) value calculated from the model. Moreover, the Likelihood Ratio (LR) test can also be used to test the fitness statistically. The Likelihood ratio test basically compares the log-likelihood function of the unrestricted model ($\ln L_U$) with the restricted model ($\ln L_R$) as follow:

$$LR = -2(\ln L_R - \ln L_U)$$

where the $LR \backsim \chi^2$ with degree of freedom equals to the number of restrictions.

3 Data and Variables

The data used in this study are secondary data obtained from research conducted by Mehata (2017), which are the data of patients undergoing 5-FU-based chemotherapy from a specialist of gastrointestinal surgery at one of the private universities in Tokyo, Japan. The data consists of 51 observations from patients undergoing chemotherapy and activity meter monitoring within 14 days. However, several patients terminated the treatment by stopping to activate the activity meter, and finally only 28 patients completed the experiment. Table 1 presents the variables observed from the patients used in this study.

Table 1: Variables

Symbol	Variable	Scale
y	Change of WBC (Decreasing or Increasing)	Nominal
x_1	Gender	Nominal
x_2	Age	Ratio
x_3	Average daily foot Steps	Ratio
x_4	Height	Ratio
x_6	Weight	Ratio
x_7	Dose of FU1	Ratio
x_8	Dose of FU2	Ratio

4 Results and Discussion

4.1 Descriptive Statistic

This part describes the colorectal patients' characteristics, including the white blood cell (WBC) change. The data used in this study is data from colorectal cancer patients who underwent chemotherapy treatment using the 5-Fluorouracil compound, combined with physical treatment in the form of footsteps. The description of the patient characteristics is important to gain general information about the patient condition and the variables used to determine its effect on changes in WBC amount in the patient's blood. Gender, as one of the predictor variables in this study, is claimed to affect changes in the amount of WBC. Figure 1 depicts the response variable, i.e. the number of patients with decreasing or increasing WBC differentiated by the patient's gender.

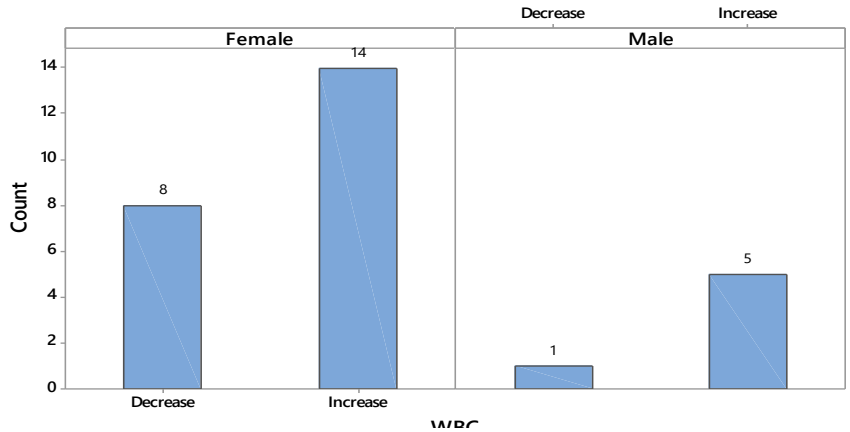

Figure 1: Characteristic of WBC change by gender

WBC changes differentiated by gender showed that the majority of patients (both male and female) experienced an increase in WBC. However, the chart does not provide a clear guideline about whether the change in WBC is significantly influenced by gender. This study also investigates the influence of average number of patient footsteps on the change of WBC level. Figure 2 provides information about the descriptive statistic of the average number of footsteps distinguished by the increase and decrease in the amount of WBC after chemotherapy.

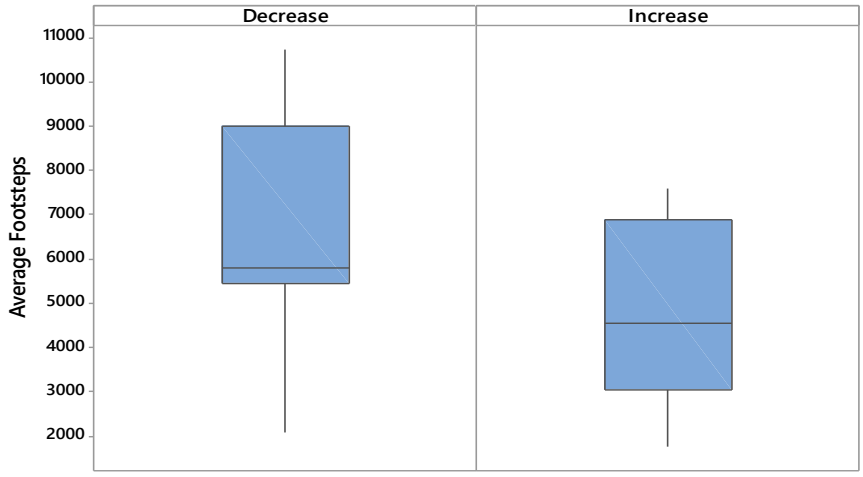

Figure 2: Average number of footsteps

Based on the boxplots presented in Figure 2, we see that patients with higher average

number of footsteps are more likely to experience a decrease in WBC, while patients with lower average number of footsteps tend to have increasing WBC after the therapy. It is interesting to note that the distribution of patients with decreasing WBC is asymmetric with a very long tail, meaning that most of the patients with decreasing WBC did footsteps of around 5000 to 6000 steps a day. Meanwhile, the distribution of steps carried out by the patients experiencing increasing WBC seems to be more symmetric, centred at about 4500 steps.

4.2 Probit Model

Prior to applying probit regression, identification of the relationship pattern between the response and predictor variables needs to be done to have an initial guess on the nature of the relationship. A scatterplot will be presented to show the pattern of relationships between predictor variables and the raw data of the response variables, i.e. the level of WBC change measured in cell/mm3 units, as shown in Figure 3.

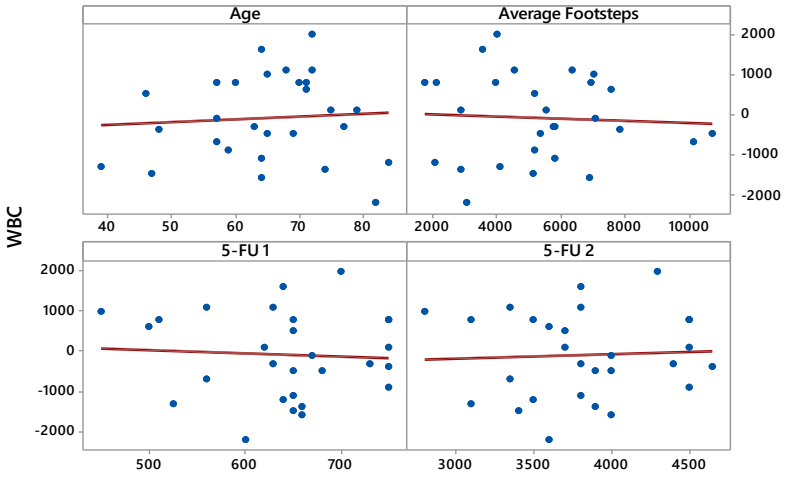

Figure 3: Scatterplot between predictors and response

The scatterplots presented in Figure 3 show the relationship between changes in WBC level in the patient's blood with several predictor variables, namely the patient's age, average number of steps, 5-FU 1 dose and 5-FU 2 doses. The age of patient is likely to have positive relation with the WBC change, while the footsteps shows a negative pattern. This means that older patients tend to have higher change in WBC, while a higher number of steps tends to decrease the WBC level. The impact of the 5-FU dose on the WBC change is unclear from the scatterplot, and hence the probit regression is expected to statistically justify this. Table 2 provides the summary of the parameters of probit model.

Table 2: Parameters of probit regression

Variable	Coef	Std. Err.	z	P-value
Gender	1.9369	3.4848	0.56	0.578
Age	-0.2930	0.1439	-2.04	0.042
Foot steps	-0.0013	0.0006	-2.08	0.037
Height	-0.1452	0.1847	-0.79	0.432
Weight	0.0535	0.0456	1.17	0.241
FU1	-0.0303	0.0224	-1.35	0.177
FU2	0.0008	0.0028	0.30	0.767

The likelihood ratio (LR) statistic = 18.10 (P-value = 0.0115), the log-likelihood = -8.534, and the AIC value = 33.068. The P-value of the likelihood ratio statistic is less than the significant level 0.05, indicating that there is at least one predictor variable significant in the model. The values in Table 2 indicate that the change of WBC is significantly influenced by the age of the patient and the number of daily steps done by the patient. Meanwhile, gender, body weight and the FU dose did not influence the change of WBC level, as the P-values are greater than 0.05. Nevertheless, the results above did not necessarily identify the best model since there might be multicollinearity among the predictors. Therefore, further analysis is conducted by applying stepwise regression. The summary the stepwise process is listed in Table 3.

Table 3: Parameters of probit regression

Step	Predictor	Log-likelihood	AIC
1	Gender, Age, Steps, Height, Weight, FU1	-8.579	31.158
2	Age, Steps, Height, Weight, FU1	-8.846	29.692
3	Age, Steps, Weight, FU1	-9.311	28.623
4	Age, Steps, FU1	-10.725	29.450
5	Age, Steps	-13.468	33.379
6	Steps	-15.465	34.931

From the table, we see that the best probit model is the one involving only four predictors, i.e. age, steps, weight and FU1, as the AIC is the smallest among other models. The probit regression coefficients for the best model are given in Table 4.

Table 4: Parameters of probit regression

WBC	Coef	Std. Err.	z	P-value
Age	-0.2376	0.1189	-2.00	0.046
Steps	-0.0012	0.0005	-2.11	0.035
Weight	0.0605	0.0405	1.50	0.135
FU1	-0.0270	0.0135	-2.00	0.046

We see that the best probit regression has three variables that significantly influence the change of WBC, i.e. age, steps and FU1. The age influences the WBC change negatively, which means that the younger the patient, the more likely the given treatment will increase the WBC. The number of steps also significantly influences the WBC change, with negative sign. This means that increasing steps will tend to decrease the WBC. The dose of FU1 needs to be increased if the patient has to decrease the WBC level. Meanwhile, weight of the patient does not influence the change of WBC. Among these variables, the influence of age dominates the WBC change. From the table, the probit regression model can be written as in (3)

$$\hat{y}^* = 37.0756 - 0.2376Age - 0.0012Steps + 0.0605Weight - 0.0270FU1 \tag{3}$$

The coefficients of the model also indicate the marginal effect of the variable on the change of WBC. This can be interpreted as follows: Patients 1 year older will tend to have decreasing WBC by a probability of 23.76%. Every 1 step increase of daily footsteps will decrease the probability of increasing WBC by 0.12%. Increasing 1 kg of the patient weight will increase the probability of increasing WBC by about 6.05%. Furthermore, increasing the dose of FU1 tends to decrease the probability of increasing WBC by 2.70%. The best probit model is used to calculate the probability of failure (WBC decrease) and success (WBC increase). As an illustration, given a patient 63 years old doing average footsteps of 5750.57 steps daily, with weight 67 kg and FU1 dose given of 730, the predicted y is thus -1.0187. The probability value for the predicted y is 0.2171. This means that the treatment given to the patient will tend to decrease the WBC with the probability of 0.2171.

From the results obtained through the best model in univariate probit modelling, classification can be done on each response variable with the variable predictor present in the model. The measure of the goodness of classification used is the value of accuracy or success of the model to correctly predict the response variable. The obtained prediction results are presented in Table 5.

Table 5: Parameters of probit regression

		Actual		Total
		y_0	y_1	
Prediction	$\hat{y}_0$	6	1	7
	$\hat{y}_1$	3	18	21
Total		9	19	28

Based on the results of the predictions presented at Table 5, an accuracy value of 85.71% is obtained, which can also be stated as that the best model obtained can correctly predict the actual data by 85.71%, or with misclassification of 14.29%.

5 Conclusion

The characteristics of colorectal cancer patients who underwent chemotherapy treatment using 5-Fluorouracil were dominated by female patients, and from a total of 28 patients it was found that 19 patients had increasing level of WBC in blood after having 5-Fluorouracil-based chemotherapy. The best model to predict the change in WBC revealed that the change of WBC level is influenced by the age of patient, average number of footsteps, patient body weight and 5-FU1 compound dose, where the resulting model is able to classify the response category with an accuracy of 85.71%. Depending on the initial level of WBC of the patient, daily footsteps can be recommended as one of the ways to decrease the WBC level.

Acknowledgement

The authors thank to Tomoki Mehata from the Tokyo University of Science-Japan for providing the dataset used in this research.

References

Algamal, Z. (2017). Classification of gene expression autism data based on adaptive penalized logistic regression. *Electronic Journal of Applied Statistical Analysis*, 10(2):561–571.

Algamal, Z. Y. and Lee, M. H. (2019). A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. *Advances in Data Analysis and Classification*, 13:753–771.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Press.

Chang, M., Dalpatadu, R. J., Phanord, D., and Singh, A. K. (2018). Breast cancer prediction using bayesian logistic regression. *Biostatistics and Bioinformatics*, 2(3):1–5.

Frensham, L. J., Parfitt, G., and Dollman, J. (2018). Effect of a 12-week online walking intervention on health and quality of life in cancer survivors: A quasi-randomized controlled trial. *International Journal of Environmental research and Public Health*, 15(2081):1–7.

Gitto, L., Noh, G. H., and Andrez, A. R. (2015). An instrumental variable probit (ivp) analysis on depressed mood in korea: the impact of gender differences and other socio-economic factors. *Int J Health Policy Manag.*, 4(8):523–530.

Gujarati, D. T. and Porter, D. C. (2003). *Basic econometrics*. McGraw-Hill.

Hahn, E. D. and Soyer, R. (2007). Probit and logit models: Differences in a multivariate realm. *Available at: http://home.gwu.edu/ soyer/mv1h.pdf*, pages 1–14.

Longley, D., Harkin, D., and Johnson, P. J. (2003). 5-fluorouracil: mechanism of action abd clinical strategies. *Nature*, 2:330–338.

Madhu, B., Ashok, N., and Balasubramanian, S. (2014). A multinomial logistic regression analysis to study the influence of residence and socio-economic status on breast cancer incidences in southern karnataka. *nternational Journal of Mathematics and Statistics Invention*, 2(5):1–8.

Mehata, T. (2017). Extraction of side effect characteristics given by 5-fu using machine learning. *Unpublished Report Tokyo University of Japan*.

Panetta, J. (1995). A logistic model of periodic chemotheraphy. *Applied Mathematics Letters*, 8(4):83–86.

Ruspriyanty, D. and Sofro, A. (2018). Analysis of hypertension disease using logistic and probit regression. *J. of Phys: Conf. Ser.*, 1028(012054):1–6.

Seddik, A. F. and Shawky, D. M. (2015). Logistic regression odel for breast cancer automatic detection. In *SAI Intelligent Systems Conference 2015*, pages 150–154. IEEE.

Wiliam, P. T. (2013). Breast cancer mortality vs. exercise and breast size in runners and walkers. *PLOS One*, 8(12):1–6.

Wilson, D. B., Porter, J. S., Parker, G., and Kilpatrick, J. (2005). Anthropometric changes using a walking intervention in african american breast cancer survivors: a pilot study. *Prev. Chronic. Dis.*, 2(2):1–7.

Zangmo, C. and Tiensuwan, M. (2018). Application of logistic regression models to cancer patients: a case study of data from jigme dorji wangchuck national refrerral hospital (jdwnrh) in bhutan. *J. of Phys: Conf. Series*, 1039(012031):1–10.

Zhou, X., Liu, K. Y., and Wong, S. (2004). Cancer classification and prdiction using logistic regression with bayesian gene selection. *Journal of Biomedical Informatics*, 37(4):249–259.