DOI Number: 10.5958/0976-5506.2019.00218.3

Analysis of Ordinal Logistic Regression Model on Breast Cancer Diagnosis by Birads Mammography

M. Nadjib Bustan[1], M. Arif Tiro[1], Suwardi Annas[1], Adiatma[1]

[1]Department of Statistics, Faculty of Mathematic and Natural Science, Universitas Negeri Makassar, Indonesia

ABSTRACT

The right diagnosis is needed for appropriate therapy. The diagnosis of breast cancer is quite ambiguous and requires high accuracy. Mammography is a method of diagnosing breast cancer using BIRADS (Breast Imaging-Reporting and Data System) assessment. This study aimed to assess the accuracy of BIRADS classification in the diagnosis of breast cancer and predictors that influence it through a logistic regression model test. The research method was cross sectional study by collecting data from the results of mammography examinations obtained from Medical Record documents, SIRS (Hospital Information Systems), and the radiologist's expertise of mammography. The data came from 47 hospital breast cancer patients that contained information on potential predictors of breast cancer namely tumor location, metastases, age, weight, and education. Logistic regression model analysis was performed to find the best statistical test model for breast cancer diagnosis classification based on BIRADS assessment. The diagnosis classification of BIRADS was consisting of normal, benign, and malignant grades. For this reason, hypothesis testing was conducted with G test for simultaneous model testing. Then, a development of an appropriate logit model by using a partial test. Followed by conducting a suitability and feasibility test model with the Goodness of Fit using the Hosmer-Lemeshow Test. The results of the analysis revealed that the ordinal logistic regression was the best model of BIRADS classification diagnosis with an accuracy value of 52.5%. The result of ordinal logistic regression model for malignant breast cancer:

$$\hat{\pi}_1(x) = \frac{\exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}{1 + \exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}$$

The result for benign cancer:

$$\hat{\pi}_2(x) = \frac{\exp(-17,696 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}{1 + \exp(-17,696 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})} - \frac{\exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation Pekerjaan} + 2,549\,\text{location})}{1 + \exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}$$

A significant predictor factors were the location of the tumor, age, education, and the work of cancer patients. The conclusion of the diagnosis classification of breast cancer using BIRADS of mammography is quite accurate and assessment of diagnosis classification BIRADS should pay attention to tumor location factors, age, education, and work of breast cancer patients.

***Keywords**: Ordinal logistic regression, BIRADS, mammography, breast cancer diagnosis*

INTRODUCTION

Every disease requires an accurate diagnosis so that doctors can provide appropriate treatment. The diagnosis of breast cancer is quite ambiguous but still requires a high accuracy of diagnosis.[1], [2]

Diagnosis of breast cancer requires several types of testing, namely physical or clinical examination,

Corresponding Author:
M. Nadjib Bustan
Department of Statistics,
Faculty of Mathematic and Natural Science,
Universitas Negeri Makassar, Indonesia
Email: mnbustan@unm.ac.id

radiological examination, histopathological examination, genetic examination, and immunology.[3]

Radiological examinations for the diagnosis of breast cancer using mammography were assessed for the malignancy levels by using BIRADS (Breast Imaging-Reporting and Data System) developed by the American College of Radiology (ACR) and carried out by radiologists. BIRADS assessment is scoring from 1 to 6 with the meaning that 1 is negative, 2 is benign, 3 is probably benign, 4 is suspicious for malignancy, 5 is highly suggestive of malignancy, 6 = known biopsy malignancy.[4],[5]

For developing the model and assessing the accuracy of the mammography examination results, a statistical approach could applying the Ordinal Regression Logistics.[6]

The results of the model analysis will find the best model, the accuracy of the selected model, and determine the predictor factors that influence the presence of breast cancer.[7]

Ordinal Regression Logistics is one of the statistical methods for analyzing ordinal scale of response variables consisting of three or more categories. Predictor variables used in this model in the form of category data or quantitative data.[8]

Ordinal Regression Logistic Model for ordinal data response variables are often referred to as cumulative logistic models. The response variable in the cumulative logistic regression model is in the form of multilevel data represented by numbers 1, 2, 3, ..., k. With k is the number of categorical response variables. The cumulative logistic regression model will compare cumulative opportunities, ie opportunities less than or equal to the jth response category on p predictor variables expressed in vector of the x_i. $\mathrm{P}(\mathrm{Y} \le j|x_i)$, with opportunities greater than the response category j, x_i, P (Y> j | x_i).[9]

Cumulative opportunity forms are defined as follows:

$$\pi_k(x_k) = \mathrm{P}(\mathrm{Y} \le j|x_i) = \frac{\exp[g_j(x_k)]}{1+\exp[g_j(x_k)]}$$

$$= \left(\frac{\exp\left(\beta_{0j}+\sum_{k=1}^{r}\beta_k x_k\right)}{1+\exp\left(\beta_{0j}+\sum_{k=1}^{r}\beta_k x_k\right)}\right);$$

dengan $k = 1, 2, \ldots, j, \ldots, r$

$$\pi_k(x_k) = \mathrm{P}(\mathrm{Y} \le j|x_i) = p_1 + p_2 + \ldots + \pi_r$$

The formula for general logistic distribution function is: $\mathrm{F}(x) = \dfrac{1}{1+e^{-x}} = \dfrac{1}{1+e^{x}}$

If $P(Y \le \mathrm{j})$ is compared with the probability of a respons variabel on category ($j+1$) until category r, the result is:

$$\frac{P(Y\le j)}{P(Y>j)} = \frac{P(Y\le j)}{1-P(Y\le j)} = \frac{\dfrac{\exp\left(\beta_{0j}+\sum_{k=1}^{r}\beta_k x_k\right)}{1+\exp\left(\beta_{0j}+\sum_{k=1}^{r}\beta_k x_k\right)}}{\dfrac{1}{1+\exp\left(\beta_{0j}+\sum_{k=1}^{r}\beta_k x_k\right)}}$$

$$= \exp\left(\beta_{0j}+\sum_{k=1}^{r}\beta_k x_k\right)$$

$$\frac{P(Y\le j)}{P(Y>j)} = \frac{P(Y\le j)}{1-P(Y\le j)} = \frac{\pi_1+\pi_2+\ldots+\pi_j}{\pi_{j+1}+\pi_{j+2}+\ldots+\pi_r}$$

Next, execute logistic transformation to be logit model of ordinal regression logistic:

$$\text{Logit } [P(Y \le j)] = \log\frac{P(Y\le j)}{1-P(Y\le j)}$$

$$= \log\frac{\pi_1+\pi_2+\ldots+\pi_j}{\pi_{j+1}+\pi_{j+2}+\ldots+\pi_r}$$

$$\text{Logit } [P(Y \le j)] = \beta_{0j} + \sum_{k=1}^{r}\beta_k x_k$$

with the value of β_k, for k = 1, 2,…,r to each of ordinal regression logistic is the same.

MATERIALS AND METHOD

To conduct a model analysis, data on breast cancer patients was needed. The source of data collection came from medical records documents, SIRS (Hospital Information System), and mammography images. Data containing information about patient identity and potential determinants in the term of age, tumor location, metastases, education, employment, and supplemented by the results of reading mammography expertise. The BIRADS assessment results are converted to ordinal data where 1 was normal, 2-3 were benign, 4-5-6 were malignant. The study design was a cross sectional study that collected data on breast cancer patients who were treated and registered at one Makassar hospital, Indonesia. The collected data was analyzed to find the best model. The analytical steps taken include: - estimating parameters; - testing logit model parameters with simultaneous testing, partial test and logistic analysis; - and testing the suitability and accuracy of the model with the Goodness of Fit (GOF) test and the Hosmer-Lemeshow test.[9]

RESULTS

Parameter estimation was conducted by using simulataneus test of ordinal regression logistic by G test with the formula $G = -2\, ln \dfrac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{\,y_i}\left(1-\hat{\pi}_i\right)^{(1-y_i)}}$

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

Rejection area Ho if G > $\chi^2_{(v,\alpha)}$ with *v* degree of freedom is equal with the number of parameters in the model without β_0.

Simultaneous testing obtained a calculated value of -2log-*likelihood* model of 61,146. Because p-value is equal 0,001 is smaller than α = 0.01, then Ho is rejected, which means that at least one predictive variable has a significant effect on the classification of breast cancer BIRADS. In this case the variables of age, education, occupation, and tumor location significantly influence the classification of breast cancer BIRADS.

Table 1: Statistical Output of G Test of Ordinal Regression Logistic

Model	-2 Log Likelihood	Chi-Square	df	P-Value
Intercept Only	89,529			
Final	61,146	28,382	9	0,001

To identify the role of each variable, parietal test is conducted with the result as the follows:

Table 2: Statistical Outputs of Partial Test of Ordinal Regression Logistic

	Estimation	Std. Error	Wald	P-Value
BIRADS Malignant	-19,436	4,222	21,192	0,000
BIRADS Benign	-17,696	4,230	17,505	0,000
Age	1,538	0,740	4,326	0,038
Education	-0,069	0,022	10,152	0,001
Occupation	-16,313	1,516	115,797	0,000
Location of tumor	2,549	1,086	5,510	0,019

From the results of the parameter estimation above, it is found that there are four predictor variables that have a significant effect on the variable level of malignancy BIRADS namely, age, tumor location, education and occupation.

The logit model of its statistical outputs are:

g_1 (malignant) = − 19,436 + 1,538 age − 5,725 education − 16,313 occupation + 2,549 location

g_2 (benign) = − 17,696 + 1,538 age − 5,725 education − 16,313 occupation + 2,549 location

Based on the three logit functions above, logit 1 is a logit function for malignant BIRADS and logit 2 is a logit function for benign BIRADS. Furthermore, from the two logit functions, the probability function of each category is obtained.

The logit model formula could be used to calculate the probability formulation for each response variable. The probability formula for malignant breast cancer BIRADS is as follows.

$$\hat{\pi}_1\ (\text{ganas}) = \frac{\exp(g_1(x))}{1+\exp(g_1(x))}$$

Formulation of probability BIRADS benign breast cancer is:

$$\hat{\pi}_2\,(x) = \frac{\exp(g_2(x))}{1+\exp(g_2(x))} - \frac{\exp(g_1(x))}{1+\exp(g_1(x))}$$

So:

For Y=1 (BIRADS malignant)

$$\hat{\pi}_1(x) = \frac{\exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}{1 + \exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}$$

For Y=2 (BIRADS benign)

$$\hat{\pi}_2(x) = \frac{\exp(-17,696 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}{1 + \exp(-17,696 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})} - \frac{\exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}{1 + \exp(-19,436 + 1,538\,\text{age} - 5,725\,\text{education} - 16,313\,\text{occupation} + 2,549\,\text{location})}$$

To determine the model formed from the above predictor variables is appropriate or not in accordance with the data, the suitability model of Goodness of Fit is used by using Hosmer-Lemeshow test:

$$\hat{C}(\text{Hosmer} - \text{Lemeshow}) = \sum_{k=1}^{g} \frac{\left(o_k - n_k{}'\bar{\pi}_k\right)^2}{n_k{}'\bar{\pi}_k\left(1 - \bar{\pi}_k\right)}$$

Rejection area H_0: $\hat{C} > \chi^2_{(\alpha,g-2)}$ or *p-value* $<\alpha = 0.01$

Table 3: Goodness of Fit Test of Ordinal Regression Logistic Model

	Chi-Square	Df	Sig.
Pearson	58,427	73	0,893
Deviance	56,987	73	0,916

The p-value results for Pearson and Deviance are more than α> 0.01, with values of 0.893 and 0.916, respectively. Ho is not rejected, which means that the model obtained is in accordance with the data or there is no significant difference between the results of the observation with the possible predictions of the model.

Thus, variables that significantly influence the increase in breast cancer BIRADS are variables of age, education, occupation, and location of the tumor.

To find out the model that is formed is feasible, it can be seen from the R2 value.

Table 4: Pseudo R-Square of Ordinal Regression Logistic Model

Cox and Snell	0,453
Nagelkerke	0,525
McFadden	0,303

Based on the table "Pseudo R-Square" the value of Nagelkerke R Square is 0.525. In other words, the resulting model with five predictor variables, the variables of age, education, occupation and location that have a significant effect while body weight variables did not significantly influence the increase risk in breast cancer BIRADS. In addition, the model was also able to explain the variation of breast cancer BIRADS classification.

In addition, the model was also able to explain the variation of breast cancer BIRADS classification amounting of 30.3%.

DISCUSSION

The accuracy test of diagnosis can be done using logistic regression models. There are three main types of logistic regression known, namely binary logistic regression, multi-nominal logistic regression and ordinal logistic regression.[10]

The selection of logistic regression types depends on the measurement scale of dependent variable data. Because the diagnosis of breast cancer BIRADS is categorical and ordinal (normal, benign, malignant), ordinal logistic regression is chosen.[11,12] The results of the model analysis show that the Ordinal Regression Logistic model along with six predictors are only able to explain the variation of breast cancer BIRADS classification by 30.3

This happens because this model data still requires some important potential predictors such as marital status, age of menarche, menopausal status, and others.[13,14]

CONCLUSION

The conclusion of the diagnosis of classification of breast cancer using BIRADS of Mammography is quite accurate, and assessment of classification diagnosis BIRADS should pay attention to tumor location factor, age, education, and work of breast cancer patients.

Ethical Clearance: Obtained from the university committee

Conflict of Interest: None

Source of Funding: Universitas Negeri Makassar, Indonesia

REFERENCES

1. John Hopkins University. Staging &amp; Grade - Breast Cancer | Johns Hopkins Pathology [Internet]. JHU Medicine. 2018 [cited 2018 Sep 20]. Available from: https://pathology.jhu.edu/breast/my-results/staging-grade
2. Ministry Health. Guide to Management of Breast Cancer. Jakarta: National Cancer Mitigation Committee; 2015
3. Johns Hopkins. Overview of the Breast - Breast Cancer | Johns Hopkins Pathology [Internet]. JH Medicine Pathology. 2018 [cited 2018 Oct 14]. Available from: https://pathology.jhu.edu/breast/basics/overview
4. ACR. ACR BI-RADS® ATLAS — MAMMOGRAPHY.
5. Balleyguier C, Ayadi S, Van Nguyen K, Vanel D, Dromain C, Sigal R. BIRADS^{TM} classification in mammography. Eur J Radiol. 2007 Feb;61(2):192–4.
6. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression.
7. Agresti A. Categorical data analysis. Wiley-Interscience; 2013. 714 p.
8. Sharma B, Abhimanyu A, Anuradha A, Gigras Y. Logistic Regression for Breast Cancer Analysis. Data Min Knowl Eng. 2017;9(6):109–13.
9. Hosmer DW, Lemeshow S. Applied logistic regression. Wiley; 2000. 373 p.
10. Kitbumrungrat K. Comparison Logistic Regression and Discriminant Analysis in classification groups for Breast Cancer. IJCSNS. 2012;12(5).
11. Yusuff H, Mohamad N, Ngah UK, Yahaya AS. BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION. IJRRAS. 2012;10(1).
12. Chhatwal J, Alagoz O, Lindstrom MJ, Kahn CE, Shaffer KA, Burnside ES. A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis. Am J Roentgenol. 2009 Apr;192(4):1117–27.
13. Bustan MN, Coker AL, Addy CL, Macera CA, Greene F, Sampoerno D. Oral contraceptive use and breast cancer in Indonesia. Contraception. 1993 Mar 1;47(3):241–9.
14. Kamińska M, Ciszewski T, Łopacka-Szatan K, Miotła P, Starosławska E. Breast cancer risk factors. Prz menopauzalny . Menopause Rev. 2015 Sep;14(3):196–202.