BINARーLOGISTIC MODELING BACKWARD METHODS TO DETERMINE THE RISK FACTORS THAT AFFECT CERVICAL CANCER CASES

Febbi Yustitia Aksari¹*, Windhu Purnomo², Diah Indriani³

¹Department of Biostatistics and Population Faculty of Public Health Airlangga University
²Department of Biostatistics and Population Faculty of Public Health Airlangga University
³Department of Biostatistics and Population Faculty of Public Health Airlangga University

*Correspondence author: Febbi Yustitia Aksari, febbi1994@gmail.com

https://doi.org/10.32827/ijphcs.5.6.84

ABSTRACT

Background: Cervical cancer is second place as a killer of women in the world. Cervical cancer still be a problem of women's reproductive health. The incidence of both cervical cancer cases and deaths showed a trend that is increasing. Several factors are considered to influence the occurrence of cervical cancer are age, first marriage age, number of children, type of contraception, and length of contraception usage. This study aimed to analyze the influence of risk factors for cervical cancer to the cervical cancer cases in Wisnuwardhana’s Cancer Foundation in 2017 use binary logistic regression method.

Materials and Methods: Based on the calculation of the sample size obtained case and control samples each of 40, so the total sample in this study as many as 80 samples. The method of determining and sampling is done by using simple random sampling method. The results were analysed using multivariate binary logistic regression modelling backward method.

Result: The results of multivariate binary logistic regression showed five predictor variables were suspected of having links with cervical cancer cases, there are three variables, they are age (P_value = 0.010), first marriage age (P_value = 0.044) and type of contraception (P_value = 0.000) that significantly affect the cases of cervical cancer.

Conclusion: The study revealed that age, first marriage age, and type of contraception are the significant risk factors associated with cervical cancer at Wisnuwardhana Cancer Foundation Surabaya.

Keywords: binary logistic regression, cervical cancer, risk factors.
1.0 Introduction

Cancer is one of the leading causes of morbidity and mortality worldwide with about 14 million new cases and 8.2 million cancer-related deaths in 2012 (Ministry of Health of Republic Indonesia, 2013). Cervical cancer is also the second ranked cancer that mostly happened in women, estimating as many as 530,000 new cases by 2012.

WHO explained that the incidence of cancer increased from 12.7 million cases in 2008 to 14.1 million cases in 2012, in addition the number of deaths increased from 7.6 million people in 2008 to 8.2 million in 2012. Cancer is the 13th leading cause of death in the world by 13% after cardiovascular disease, estimated in 2030 the incidence of cancer can reach up to 26 million people and 17 million of them die from cancer, especially in some poor and developing countries, it will happen much faster (WHO, 2018).

Prevalence of cancer in Indonesia is also quite high seen from the data of Basic Health Research (Riskesdas) in 2013, prevalence of malignant tumors / cancer in Indonesia is 1.4 per 1000 population, or about 330,000 people (Ministry of Health of Republic Indonesia, 2013). Breast cancer and cervical cancer (cervix) is the highest type of cancer that occurs in women in Indonesia. The Ministry of Health of the Republic of Indonesia reveals of the many types of cancer that attacks the population of Indonesia, breast cancer and cervical cancer (cervix) highest case in the entire Hospital (RS) (Ministry of Health of Republic Indonesia, 2013).

Based on data of Hospital Information System 2010, 12,014 cases of breast cancer cases (28.7%), cervical cancer 5349 cases (12.8%).

One of the provinces in Indonesia that has a quite high case of cervical cancer is East Java province (Ministry of Health of Republic Indonesia, 2013). Data on cases and deaths reported by Sentinel Hospital in East Java in 2009-2014 showed that there was an increase in both the number of cases and the number of deaths from cervical cancer. The largest percentage of increase occurred in 2013, which increased 71.5% of cases from 2012. Deaths from cervical cancer also increased every year, the largest percentage increase in the number of deaths occurred in 2013, an increase of 81% from 2012 (East Java Provincial Health Office, 2013).

The incidence of cervical cancer and the magnitude of the impact caused a need to make efforts to anticipate both the promotive and preventive against the occurrence of cervical cancer both in local and national scale. Cervical cancer is the easiest type of cancer to be prevented and treated, but many patients find it late in early detection, so it is found in an advanced stage and at high risk of death. Early detection can use pap smear examination, but this method is still too expensive for most community members in developing countries.

Regression is an appropriate statistical approach for this case but for most categorical data the use of logistic regression is more optimal and effective because it produces an unbiased estimator (Hosmer & Lemeshow, 2000). Logistic regression itself is divided into several types including binary logistic regression, multinomial, and ordinal. Binary logistic regression model is one logistic regression model used to analyze the relationship between one response variable and several predictor variables. The response variable consists of the dichotomous qualitative data that is worth one to express the existence of a characteristic and a zero value to express the non-existence of a characteristic. The advantages of this regression is to have
an odds ratio that shows how much influence the predictor variable of a reference category on a response variable.

In this study, the response variable is a diagnosis of cervical cancer with a value of one for cervical cancer (class V) and the value of zero for normal results (class I). So, this study applied binary logistic regression to determine the risk factors that affect the case of cervical cancer. This research is expected to provide information that is very important for the public in general about the factors that affect the incidence of cervical cancer so early prevention can be done as early as possible.

2.0 Materials and Methods

This research is included in analytic observational research which aims to give description and explanation about influence between variables. Researchers only make measurements of research variables contained in the medical record and patient status cards for further analysis. Researchers do not provide treatment or intervention to patients who are being the research target while the research design used is a case control approach (case control). The case-control approach is an epidemiological design that studies the relationship between exposure (research variables) and disease by way of comparison between case groups and controls based on exposure status.

This study began by determining a group of diseased people (cases) and a group of people with no diseases (control) then retrospectively studied risk factors (research variables). The case control design is chosen because it is appropriate to examine a number of exposures (risk factors) to a disease, deliver quick results, and it is relatively cheap and easy compared to other analytical studies.

The population study for case and control group were all married status women who followed the normal (class I) papsmear examination as control population at Wisnuwardhana Surabaya Cancer Foundation in January-December of 2017, amounting to 97 patients. Results of examination of patients showing abnormal or cancerous condition (class V) at Cancer Foundation Wisnuwardhana Surabaya in January-December 2017 as a case population of 72 patients.

Based on the calculation of the sample size obtained case and control samples each of 40, so the total sample in this study as many as 80 samples. The method of determining and sampling is done by using simple random sampling method, ie each member of the population has the opportunity to be selected as a sample (Azwar, 2016).

Simple random sampling is one of the probability sampling techniques. This method was chosen because it was suitable for relatively homogeneous research populations, so that the selected sample was representative enough to represent the study population. This method is also selected because it is easy to do and a list of sample units is available from medical records.
Research variable used is Y which is variable response (disease) and X as predictor variable (exposure / risk factor). The predictor variable is the variable that influences or causes the change or the emergence of the response variable. Predictor variables (X) are patient age, first marriage age, number of children, type of contraception, and length of contraception usage. Response variable is variable that influenced or become result because of predictor variable, in this research used as response variable (Y) is diagnosis of cervical cancer based on pap smear examination at Wisnuwardhana Surabaya Cancer Foundation 2017.

The risk factors of cervical cancer in this study based on medical record and patient status card at Wisnuwardhana Surabaya Cancer Foundation presented according to Hendrik L. Blum health determinant theory as follows:

2.1 Environmental factor

Environment has the greatest influence compared to other factors. Number of children, HPV infection, economic level, ethnic minority and education level belong to environmental factors.

2.2 Behavior Factor

Behavior is the second factor affecting the health status of society because healthy or unhealthy environment of health of individual, family and society depends on human behavior itself. First marriage age, type of contraception, length of contraception usage, number of sexual partners, sexually transmitted diseases and smoking included into behavioral factors that affect the occurrence or absence of cervical cancer.

2.3 Health Service Factors

Health services are the third factor affecting the health status of the community because the presence of health facilities is crucial in health recovery services, prevention of disease, treatment and nursing as well as groups and communities that require health services.

2.4 Genetic Factors

Heredity (genetic) is a factor that has been present in the human being brought at birth, the age variable of the patient and the history of cervical cancer in the family included in genetic factors.

The research steps in identifying significant influences between several predictor variables or factors that are thought to affect the case of cervical cancer are as follows:

- Perform secondary data collection using data collection sheets.
- Analyzing descriptive continuous variables based on mean, median and standard deviation as well as categorical variables based on contingency tables.
- Independence test.
- Testing of univariable parameter significance.
- Testing of multivariable parameter significance.
f. The establishment of a binary logistic regression model using the Backward Selection method.
g. The classification results of cervical cancer using Backward Selection method.

3.0 Result

3.1 The Description of the Patient Characteristics in Wisnuwardhana’s Cancer Foundation

Early detection of cervical cancer using pap smears test at Wisnuwardhana’s Cancer Foundation shows the characteristics of patients with class I (normal) and class V (abnormal) results are as follows:

Table 1: Descriptive Test Results of Continuous Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pap smear Result</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>Normal</td>
<td>29</td>
<td>70</td>
<td>42.03</td>
<td>8.09</td>
</tr>
<tr>
<td></td>
<td>Abnormal</td>
<td>32</td>
<td>76</td>
<td>50.5</td>
<td>10.46</td>
</tr>
<tr>
<td>First Marriage Age (years)</td>
<td>Normal</td>
<td>19</td>
<td>39</td>
<td>24.83</td>
<td>4.19</td>
</tr>
<tr>
<td></td>
<td>Abnormal</td>
<td>10</td>
<td>33</td>
<td>20.88</td>
<td>4.55</td>
</tr>
<tr>
<td>Number of children</td>
<td>Normal</td>
<td>0</td>
<td>6</td>
<td>1.95</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>Abnormal</td>
<td>0</td>
<td>8</td>
<td>2.75</td>
<td>2.75</td>
</tr>
<tr>
<td>Length of Contraception Usage</td>
<td>Normal</td>
<td>1</td>
<td>25</td>
<td>5.8</td>
<td>4.97</td>
</tr>
<tr>
<td></td>
<td>Abnormal</td>
<td>0</td>
<td>30</td>
<td>11</td>
<td>7.31</td>
</tr>
</tbody>
</table>

Based on Table 1, mean of age with abnormal results indicates an older age compared with the mean of patient age with normal results. Standard deviations in both patients with normal and abnormal results indicate a heterogeneous age variable. The first marriage age of abnormal outcome tend to be younger than the normal outcome, both of standard deviation showing the homogeneous first marriage age. In the number of children variable, mean of number of children are 1 to 2 children and standard deviation show that both of number of children variable are homogeneous. In the length of contraception usage variable, majority of patients with abnormal outcome had been using contraceptive for approximately 11 years, then standard deviation show that the length of contraception usage variable are heterogeneous.

Table 2: Descriptive Test Results of Categorical Variable

<table>
<thead>
<tr>
<th>Pap smear Result</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>Hormonal</td>
<td>10 (20.4%)</td>
</tr>
<tr>
<td>Non Hormonal</td>
<td>30 (96.8%)</td>
</tr>
<tr>
<td>Total</td>
<td>40 (50.0%)</td>
</tr>
</tbody>
</table>
Based on Table 2, the type of contraception variables, majority of hormonal contraception users get abnormal pap smear results about 79.6% and majority of non hormonal contraception users get normal pap smear results about 96.8%.

### 3.2 The Relationship Between Each Predictor Variable to The Response Variable

**Table 3:** Bivariate Test Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Chi Square Value</th>
<th>P_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>42.686</td>
<td>0.120 (not correlated)</td>
</tr>
<tr>
<td>First Marriage Age</td>
<td>23.400</td>
<td>0.220 (not correlated)</td>
</tr>
<tr>
<td>Number of Children</td>
<td>15.513</td>
<td>0.050 (not correlated)</td>
</tr>
<tr>
<td>Type of Contraception</td>
<td>41.290</td>
<td>0.000 (correlated)</td>
</tr>
<tr>
<td>Length of Contraception Usage</td>
<td>37.568</td>
<td>0.004 (correlated)</td>
</tr>
</tbody>
</table>

Based on Table 3 it is known that the variables related to the results of pap smears (normal and abnormal classes) are type of contraception and length of contraception usage. Furthermore, the results of the influence analysis using binary logistic regression begins by performing a simultaneous test with the choice of Backward Wald procedure. The following are the results of the significance test simultaneously.

### 3.3 Logistic Regression Analysis

Results from the steps in identifying significant influences between several predictor variables that are thought to affect the case of cervical cancer are as follows:

**Table 4:** Omnibus Tests of Model Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Chi-Square</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>-3.185</td>
<td>1</td>
<td>0.074</td>
</tr>
<tr>
<td>Model</td>
<td>73.811</td>
<td>3</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Based on Table 4 it can be seen that G2 (73.811) > χ² (0.05,3) = 17.730 or P_value = 0.000 < α = 0.05 then the decision is reject H₀ which means there is at least one predictor variable that has significant effect.

**Table 5:** Univariable Significance Test Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE</th>
<th>Wald Chi-Sq</th>
<th>P_value</th>
<th>Exp (β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.093</td>
<td>0.028</td>
<td>10.969</td>
<td>0.001</td>
<td>0.911</td>
</tr>
<tr>
<td>First Marriage Age</td>
<td>0.231</td>
<td>0.069</td>
<td>11.089</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>Number of children</td>
<td>-0.327</td>
<td>0.156</td>
<td>4.397</td>
<td>0.036</td>
<td>0.721</td>
</tr>
<tr>
<td>Type of Contraception</td>
<td>-4.762</td>
<td>1.077</td>
<td>19.568</td>
<td>0.000</td>
<td>0.009</td>
</tr>
<tr>
<td>Length of Contraception Usage</td>
<td>-0.142</td>
<td>0.045</td>
<td>10.101</td>
<td>0.001</td>
<td>0.868</td>
</tr>
</tbody>
</table>
Based on Table 5 it can be seen that through the significance test of each predictor variable to the response variable shows that the age variable, the first marriage age, the number of children, the type of contraception and the length of contraception usage, each has a P-value less than $\alpha = 0.05$.

**Table 6: Multivariable Significance Test Results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE</th>
<th>Wald Chi-Sq</th>
<th>P-value</th>
<th>Exp ($\beta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.136</td>
<td>0.053</td>
<td>6.622</td>
<td>0.010</td>
<td>1.146</td>
</tr>
<tr>
<td>First Marriage Age</td>
<td>-0.218</td>
<td>0.108</td>
<td>4.068</td>
<td>0.044</td>
<td>0.804</td>
</tr>
<tr>
<td>Type of Contraception</td>
<td>6.191</td>
<td>1.607</td>
<td>14.839</td>
<td>0.000</td>
<td>488.484</td>
</tr>
</tbody>
</table>

Based on Table 6 it can be seen that the age variable, the first marriage age, and the type of contraception have the value of $P_{\text{value}}$ less than $\alpha = 0.05$ then the decision is reject $H_0$. Therefore, the following three variables, age, first marriage age, and type of contraception significantly influence the diagnosis of cervical cancer cases.

Establishment of binary logistic regression model using predictor variable that has significant both independence test, partial test and simultaneous test. The regression model are as follows:

$$
\pi(x) = \frac{1}{1 + e^{-(4.976+0.139(\text{age})-0.283(\text{first marriage age})-0.474(\text{number of children})+7.044(\text{type of contraception})+0.173(\text{length of contraception usage})}}
$$

Based on the results of the analysis of independence test, partial test and simultaneous test showed that predictor variables that significantly influence the response variables (diagnosis of cervical cancer cases) are age variable, first marriage age, and type of contraception. The best regression model based on the variables that significantly influence the cases of cervical cancer or which can represent cases of cervical cancer at the Wisnuwardhana Cancer Foundation are as follows:

$$
\pi(x) = \frac{1}{1 + e^{-(0.136(\text{age})-0.218(\text{first marriage age})+6.191(\text{type of contraception})}}
$$

The age variable shows the Odds Ratio value of $\exp (\beta) = 1.146$. This indicates that every 1 year increase in the variable age increases the risk of 1.146 times cervical cancer. The first marriage age variable shows the Odds Ratio value of $\exp (\beta) = 0.804$. It indicates that every addition of 1 year old married first can cause abnormal result of diagnostic as many as 0.804 times. It will cause normal result of cervical cancer diagnostic equal to 0.804 times. Variable type of contraception show value of Odds Ratio equal to $\exp (\beta) = 488.484$. This indicates that patients who use hormonal contraception increase the risk for cervical cancer by 488.484 times.
The following is a table of classification results of cervical cancer cases which is given the logistics model.

**Table 7: Cross Tabulation Classification Results**

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cervical cancer</td>
<td>Normal</td>
</tr>
<tr>
<td>Cervical Cancer</td>
<td>37</td>
<td>3</td>
</tr>
<tr>
<td>Normal</td>
<td>5</td>
<td>35</td>
</tr>
</tbody>
</table>

Accuracy of classification 90%

Sensitivity 92.5%

Specificity 87.5%

APER 10%

The actual number of cervical cancer and predicted positive cervical cancer is predicted as much as 37 whereas predicted cervical cancer is not 3. The actual number with normal results and predicted positive cervical cancer is 5, predicted normal by 35. Logistic regression model that has formed classification accuracy of 90%, sensitivity of 92.5% and 87.5% specificity for cervical cancer cases at Wisnuwardhana Cancer Foundation Surabaya, then at 10% APER value means there is a 10% chance of error in classifying objects using this model.

**4.0 Discussion**

The result of data processing using binary logistic regression to cervical cancer case data at Wisnuwardhana Surabaya Cancer Foundation 2017 with computer application yield 3 significant variables that influence from 5 predictor variables tested. Any insignificant variable will be excluded from the model automatically so that the best model can be obtained. Predictable variables are age, age of first married and type of contraception have value of Sig Wald <0.05 so reject H0 or that mean give a significant partial influence to the incidence of cervical cancer. The magnitude of influence is shown by the value of exp (β) or also called Odds Ratio (OR). The age variable with exp (β) = 1.146 explains that every 1 year increase in the variable age of the patient then the risk increases 1.146 times cervical cancer. Based on the consensus of cervical cancer NIH in Rasjidi (2009) mentioned that the incidence of cervical cancer higher at age more than 65 years and the average age of cervical cancer patients is 52 years (Rasjidi, 2009). Many studies have found that incidence of cervical cancer in old age increases and tumors look more aggressive, indicated in the study that every elderly woman has a greater risk of lymph node metastasis.

Among the factors that aggravate the incidence of cervical cancer in women aged ≥35 years have a risk of 5.86 times to experience the incidence of cervical pre-cancer lesions than those <35 years old, it also indicates an increase in the age of a person can increase the greater risk (Pradya, 2015). This study is also in accordance with the theory that the older age of a person...
will experience the process of decline, so that at a later age may cause the possibility of more frequent sickness and susceptible to infection (Sukaca, 2009).

The first marriage age variable with exp (β) = 0.804 means that every 1 year increment of first marriage age will cause no cervical cancer 0.804 times. The younger the age of a woman to have sex then the increased risk for cervical cancer two times greater than women who have sex after the age of 20 years (Wijaya, 2010). Based on other studies, women who married at age ≤ 20 years are five times more likely to develop cervical cancer (Setyarini, 2012). Moreover, in the study of Iraqi women indicated that very young married age is strongly associated with abnormal pap smear examination results (Khalaf, Rasheed, & Hussain, 2015).

Variable types of contraception with exp (β) = 488.484 explains that patients using hormonal contraception devices will be at risk of cervical cancer by 488.484 times greater than patients who use non hormonal contraception. The results of data analysis of hormonal contraception use by patients in this study most experienced cervical cancer. The results above also relate to research Mutia Rissa Pratiwi about "There is Influence of Contraceptive Use of Esterogen and Progesterone With Cervical Cancer Occurrence". The likelihood of cervical cancer for patients with a history of combined hormonal contraception usage was 17.9 times compared with patients not taking combined hormonal contraceptives (Pratiwi, 2010).

Research on the use of contraception with cervical cancer to get the proportion of cases of cervical cancer more in the case of cases using hormonal contraception. Nurwijaya (2010) also hypothesized that the viscosity of mucus in the cervical area caused by the use of birth control pills can promote cervical cancer (Nurwijaya, 2010). The viscosity of the mucus that has the property will prolong the existence of a carcinogenic agent in the cervix carried over through sexual intercourse such as Human Papilloma Virus.

Variable number of children and the length of contraception usage did not get results that significantly affect the incidence of cervical cancer shown according to the results of pap smear examination. Sukaca (2009) mentions that dangerous parity is having more children than 2 people or the labor distance is too close (Sukaca, 2009). Based on the results of research that has been done, the distribution of children > 2 people as many as 21 people (72.4%) included into class V or cervical cancer but a large percentage is not enough to explain the effect of cervical cancer. The number of children with a frequency of > 2 people do not necessarily give birth with a distance that is too close. Nubia Munoz et al. in Dwianggimawati (2010) mentioned that the number of children was not statistically significant for the incidence of cervical cancer (Dwianggimawati, 2010).

The length of contraception usage based on binary logistic regression testing also showed results that did not significantly influence. Groups affected by cervical cancer use an average contraceptive device within 11 years with a minimum period of <1 year and a maximum of 30 years while respondents who do not get cervical cancer on average using contraceptives within 5 years. Oktariningtias (2010) states that the use of birth control pills risks increasing the occurrence of cervical cancer if used for 5 years or more because the results of pap smears showed significant changes in cervical cells (Oktariningtias, 2010).

The reliability of a model that is formed, in addition to passing in simultaneous and partial tests for each parameter coefficient is the accuracy of the model in predicting to be close to
the actual value. Performance of binary logistic regression method based on predictor variables that have significant effect on the sensitivity, specificity and accuracy are 92.5%; 87.5% and 90%. The level of sensitivity, specificity and accuracy produced using this method is quite high. The sensitivity value of 92.5% indicates when a patient suffering from cervical cancer, then diagnosed by medical personnel and 92.5% classification results will show the patient is positive cervical cancer. While the value of the specificity of 87.5% then when patients who do not suffer from cervical cancer diagnosed by medical personnel, there is an 87.5% chance that the diagnosis will be negative.

5.0 Conclusion and recommendation

Age, first marriage age, and type of contraception variables have a significant influence on cases of cervical cancer at Wisnuwardhana Cancer Foundation Surabaya. Every 1-year increase in the age variable then the risk increased 1.146 times affected by cervical cancer. Each addition of 1 year of first marriage, it will cause no occurrence of cervical cancer of 0.804 times later in patients using hormonal contraception devices increased the risk for cervical cancer by 488.484 times compared to patients who use non hormonal contraception. Subsequent studies need to expand the research variables such as the risk factors of multiple sexual partners and history of cancer.

Acknowledgement

I am highly thankful to all the people who have helped this research and publication process

Declaration

Author(s) declare that all works are original and this manuscript has not been published in any other journals. There was no financial support of any organization for this work.

References


