The International Journal of Engineering and Science (IJES)
|| Volume || 9 || Issue || 09 || Series I || Pages || PP 01-06 || 2020 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

Survival Analysis for Breast Cancer Patients: Case Study at Public Hospital in Malaysia

AzmeKhamis& Siew Ya Bing
Department of Mathematics and Statistics,
UniversitiTun Hussein Onn Malaysia, BatuPahat, Malaysia

---***ABSTRACT***---
Breast cancer is the most frequently occurring cancer as well as the most fatal form of malignancy among Malaysian women. This study aimed to highlight the survival rate of Malaysian breast cancer patients in hope by knowing the pattern of survival and related prognostic factors. This study was conducted using secondary data obtained from a general hospital in Johor Bahru. All 38 cases of breast cancer diagnosed from January 12, 2006 to February 15, 2017 were selected. Age, ethnicity and treatment received were collected to determine prognostic factors. The Chi-square test of independence determines if two variables are independent of one another. Patients' age (p-value = 0.602) and ethnicity (p-value = 0.284) have no association with the survival status of patients while treatment received (p-value <0.001) has association with the patients' survival status. Kaplan-Meier method was used to estimate the survival curves of the breast cancer patients. Patients under 40 years old had a poor survival compared to the patients aged between 41-60 years old and patients above 60 years old. Malay patients tend to have a poorer survival after breast cancer compared to Chinese, Indian and Others ethnic. Patients who received both local and systemic therapy treatment experience death more quickly during the period of study. Log-rank test was used to test the significant difference between the survival experiences of the patients. Treatment group was found to have a significant difference in the survival experience; whereas age group and ethnic group of the breast cancer patients do not have a significant difference in the survival experience.

KEYWORDS*: Survival Analysis, Chi-square Test of Independence, Kaplan-Meier Method, Log-rank test*

Date of Submission: 30-09-2020 Date of Acceptance: 13-10-2020

I. INTRODUCTION

Breast cancer is a malignant tumor arising from the cells of the breast. It predominantly occurs in women, but in rare cases men can get breast cancer too [1]. Breast cancer is highly curable if diagnosed at an early stage. There are multiple factors that influenced the survival of breast cancer patients including age at diagnosis, ethnicity, cancer staging at diagnosis, lymph nodes status, treatment received, immunohistochemistry subtype, nuclear grade, histological grade, access to care and environmental factors [2]. There are multiple factors that influenced the survival of breast cancer patients including age at diagnosis, ethnicity, cancer staging at diagnosis, lymph nodes status, treatment received, immunohistochemistry subtype, nuclear grade, histological grade, access to care and environmental factors.

As stated by World Health Organization [3], breast cancer is the most commonly diagnosed cancer among women, impacting 2.1 million women each year. The prevalence of breast cancer was reported as increasing in most of the Asian countries [4], [5] and [6]. Breast cancer is the most frequently occurring cancer as well as the most fatal form of malignancy among Malaysian women [7]. According to the GLOBOCAN 2018 released by the International Agency of Research on Cancer (IARC), among the South-Eastern Asia countries, Malaysia has the fourth highest estimated age-standardized incidence rate of breast cancer which accounts for 47.5 deaths per 100,000 population and second highest estimated age-standardized mortality rate of breast cancer which accounts for 18.4 deaths per 100,000 population [8].

The survival analysis, an important branch of statistics, commonly used in medical studies to develop and validate prognostic index for mortality or disease recurrence, and to measure the outcome of treatment [9].Survival analysis is generally defined as a set of methods for data analysis, for which the outcome variable of interest is time until an event occurs [10]. An event may be either cure or death. Clearly define both the beginning of the period of time and the time of the event is important. The time between the two is known as survival time, even when the event which ends it is not death [11].

Survival analysis is concerned with time-to-event or survival data. Generally, it dealt with death as the event, but it can handle any event occurring over a period of time, and this need not be always adverse in nature. When the outcome of a study is the time to an event, it is often not possible to wait until the event in question

has happened to all the subjects, for example, until all are dead [12]. This phenomenon is known as censoring due to information is incomplete for these subjects. Censoring may occur when patients have not yet experienced the event such as relapse or death before the study ends or they lost to follow-up during the study period or they experience a different event that make further follow-up impossible such as died from causes unrelated to the disease [13]. In the presence of censoring, the true time to event is underestimated. Visualising the survival process of an individual as a time-line, their event is beyond the end of the follow-up period. This situation is often called right censoring. Traditional regression methods are not equipped to handle censoring and hence survival analysis methods are the only techniques capable of handling censored observations without treating them as missing data [12].

The survival analysis has been widely used in various fields, including sociology for "event-history analysis" [14], engineering for "failure-time analysis" [15] and medical for "patients-survival time analysis" [16] , 17] and [18]. In medical studies, most of survival analysis use Kaplan-Meier plots to visualize survival curves [19], log-rank test to compare the survival times between two or more treatment groups [20] and Cox proportional hazards regression to investigate the relationship of predictors and the time-to-event through the hazard function [21].

In this study, survival analysis is used in the medical field for analysing the survival time of breast cancer patients in Johor. The dataset used in this study comes from Hospital Sultan Ismail, Johor Bahru. Hospital Sultan Ismail (HSI) is a 704 bed tertiary specialist hospital located in urban Johor Bahru.

II. RESERCH METHODOLOGY

Current study is based on 38 patients who admitted in Hospital Sultan Ismail, Johor Bahru during the period of January 12, 2006 to February 15, 2017. The data is a right-censored data since some of the patients were still alive at the end of the study.

There are five variables of interest were included in this study which are age, ethnicity, treatment received, survival status and survival time. Patients were organized into three groups: younger than 40 years old (<40), 41-60 years old and older than 60 years old (>60). Patients' ethnicity was categorized into four groups i.e. Malay, Chinese, Indian and others. Patients' treatments included local therapy treatment (radiation therapy), systemic therapy treatment (chemotherapy, hormone therapy, chemoradiotherapy and surgery) and both local and systemic therapy treatment. Survival status is indicated in '1' where the patients have experienced death from breast cancer and '0' where the patients is still alive.

Chi-square Test of Independence

Chi-square test of independence is an analysis that investigate the independency of the variables included in the population. To test the independence of two categorical variables with Chi-square test of independence, one calculates the frequency of cases in each of a contingency table that is expected assuming the variables are independent and summarize the degree to which the obtained frequency counts in cells of the table depart from the expected values.

As with parametric tests, the non-parametric tests, including the χ^2 assume the data were obtained through random selection. However, it is not uncommon to find inferential statistics used when data are from convenience samples rather than random samples. To have confidence in the results when the random sampling assumption is violated, several replication studies should be performed with essentially the same result obtained.[22] The formula of the Pearson Chi-square test is shown in equation (3.1):

$$\sum {\chi^2}_{i-j} = \frac{(O-E)^2}{E} \tag{3.1}$$

Same with other statistical tests, the Chi-square test assumes a null hypothesis and an alternate hypothesis. The null of the Chi-square test is the two variables are independent and the alternate hypothesis is that they dependent. If the *p*-value that comes out in the result is less than the significance level, which is 0.05 usually, then the null hypothesis is rejected.

Survival Analysis

Survival analysis is generally defined as a set of methods for analysing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage or divorce. The survival time can be measured in days, weeks and years [23].

To predict the time to event with survival data, the best and most convenient way is to identify the probability density function of time to event, $f(t)$. The formula is given in equation (3.2):

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}, \quad \text{for } t > 0. \tag{3.2}$$

Survival function, $S(t)$ gives the probability that a person survives longer than some specified time t. It gives the probability that the random variable T exceeds the specified time t. The formula is shown in equation (3.3):

$$\begin{aligned} S(t) &= \int_t^{\infty} f(u)du \\ &= P(T \geq t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t), \qquad \text{for } t > 0. \end{aligned} \tag{3.3}$$

Hazard function, $h(t)$ gives the instantaneous potential per unit time for the event to occur, given the individual has survived up to time t. It is the probability of failure in an infinitesimally small time period between t and $t + \Delta t$ given that the subject has survived up till time t. The formula is given in equation (3.4):

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \frac{\lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{3.4}$$

Censoring

Medical and epidemiological studies are mostly conducted with an interest in measuring the occurrence of an outcome event. When the outcome of a study is the time to an event, it is often not possible to wait until the event has happened to all the subjects [24]. Hence, censoring is common in medical studies.

Kaplan-Meier Method

Kaplan-Meier Method is a non-parametric statistic. It is the most popular method used for survival analysis. The Kaplan-Meier estimator consists of the product of a number of conditional probabilities resulting in an estimated survival function in the form of a step function [25]. The survival probability at any particular time is calculated by the formula given in equation (3.5):

$$S(t) = \frac{\text{Number of subject living at the start - Number of subject died}}{\text{Number of subjects living at the start}} \tag{3.5}$$

The Kaplan-Meier product-limit estimator is given in equation (3.6):

$$\hat{S}(t) = \begin{cases} 1, & t < t_1 \\ \prod_{i:\, t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), & t \geq t_1 \end{cases} \tag{3.6}$$

The graph of s(t) against t is called the survival curve. The Kaplan–Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution [26]. The cumulative surviving is given in equation (3.7):

$$S(p) = K_1 \times K_2 \times K_3 \times \ldots \times K_p \tag{3.7}$$

The proportion surviving period *i* having survived up to period *i* is shown in equation (3.8):

$$K = \frac{r_i - d_i}{r_i} \tag{3.8}$$

Log-rank Test

The log-rank test is used to compare the survival distribution of samples. It is used to determine whether or not differences exist in the survival experiences of two groups. The log-rank test analyses the null hypothesis. The null hypothesis of the log-rank test is there is no difference in survival between two independent groupsand the alternate hypothesis is that there is difference in survival between two independent groups. If the *p*-value that comes out in the result is less than the significance level, which is 0.05 usually, then the null hypothesis is rejected [25]. The log-rank statistic is calculated by the formula shown in equation (3.9):

$$\text{Log} - \text{rank statistic} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{3.9}$$

III. RESULTS AND DISCUSSIONS

The Chi-square test of independence determines if two variables are independent of one another. In this study, age group, ethnicity and treatment receive served as the independent variables. The dependent variable was survival status of patients. The null of the Chi-square test is the two variables are independent and the alternate hypothesis is that they dependent.

H_0 = The two variables are independent with each other
H_1 = The two variables are dependent with each other

Table 1: Chi-square test of independence

Variable	Pearson Chi-square Value	*p*-value
Age	1.015	0.602
Ethnicity	3.798	0.284
Treatment received	18.552	0.000

Based on Table 1, Chi-square test of independence resulted in $P(\chi^2 > 1.015)$ with *p*-value of 0.602 for patients' age vs survival status and $P(\chi^2 > 3.798)$ with *p*-value of 0.284 for patients' ethnicity vs survival status. Since both of them have *p*-value greater than 0.05, we do not reject the null hypothesis. We can conclude that patients' age and survival status as well as patients' ethnicity and survival status are independent with each other. As for treatment received and survival status, the Chi-square test statistic is found by $P(\chi^2 > 18.552)$ with *p*-value of 0.000094. Since the *p*-value is less than 0.05, we reject the null hypothesis. It provides strong evidence to suggest that treatment received and patients' survival status are dependent.

The Kaplan–Meier estimator is a non–parametric estimator, used to estimate the survival distribution function from survival data. The plot of survival curves is an important part of survival analysis for each group of interest.

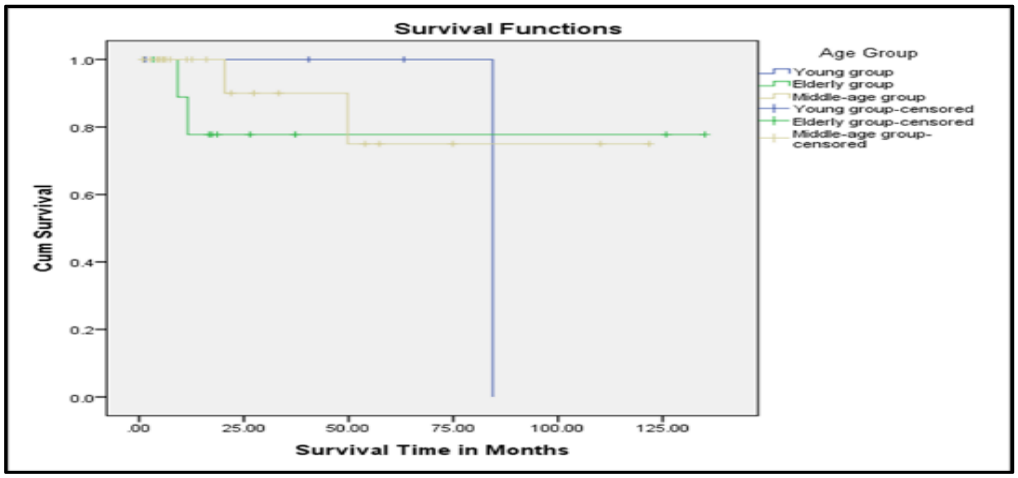

Figure 1: Comparison of Kaplan–Meier survival curve by age at diagnosis

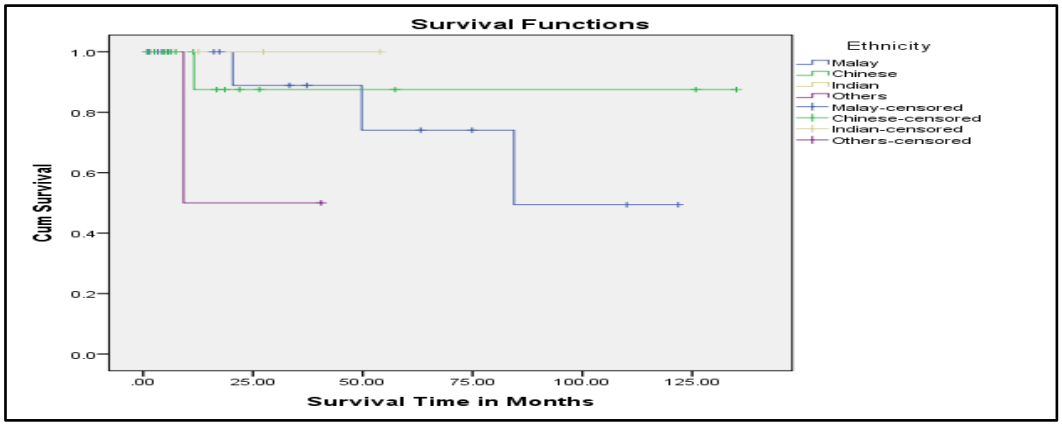

Figure 2: Comparison of Kaplan–Meier survival curve by ethnicity

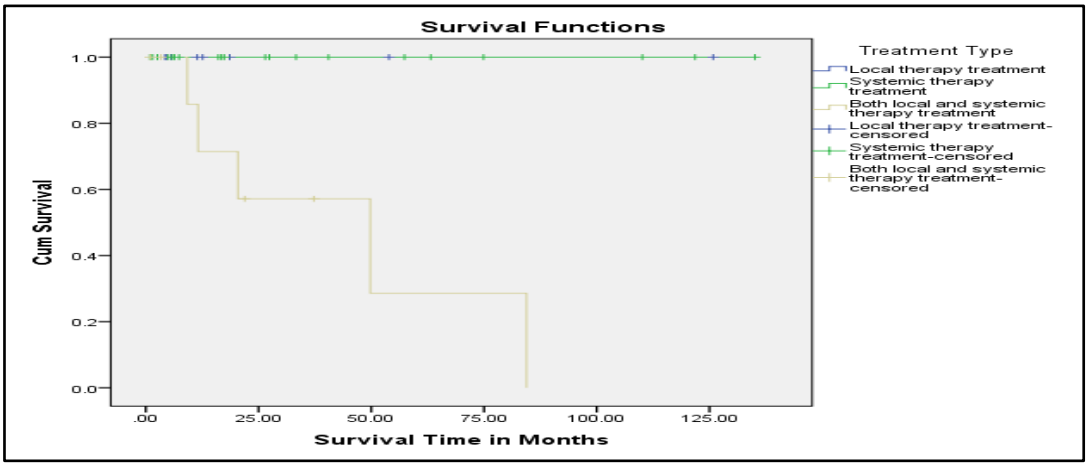

Figure 3: Comparison of Kaplan–Meier survival curve by treatment received

Based on Figure 1, patients in the young group experience death more quickly in the course of time as indicated by the quick successive drops of the line graph. From Figure 2, Malay patients with breast cancer were associated with substantially increased risk of death as compared to other ethnic groups. Based on Figure 3, patients who received both local and systemic therapy treatment experience death more quickly in the course of time.

The log-rank test is used to determine whether or not differences exist in the survival experiences of two groups. The null of the log-rank test is there is no difference in survival between two independent groups and the alternate hypothesis is that there is difference in survival between two independent groups.

H_0 = There is no significant difference in survival between groups

H_1 = There is a significant difference in survival between groups

Table 2: Log-rank test

	Log Rank (Mantel-Cox)	*p*-value
Age group	0.330	0.848
Ethnic group	4.534	0.209
Treatment group	15.374	0.000

Table 2 displays the log-rank statistic. The log-rank statistic for age group is 0.330, and the *p*-value = 0.848, this indicates that the null hypothesis is not rejected. The breast cancer patients in the three age groups have no significant different survival curves at 5% level of significance.

The log-rank statistic for ethnic group is 4.534, and the *p*-value = 0.209, this indicates that the null hypothesis is not rejected. It indicates that the breast cancer patients in the four ethnic groups have no significant different survival curves at 5% level of significance.

The log-rank statistic for treatment group is 15.374, and the *p*-value =0.000459, this indicates that the null hypothesis is rejected. The breast cancer patients in the three treatment groups have significant different survival curves at 5% level of significance.

IV. CONCLUSION

In this study, the age and ethnicity of breast cancer patients give no significant effect on the survival status of the patients. Meanwhile, the treatment given have significantly affect towards the survival status of breast cancer patients. Women under 40 years of age had a poor survival compared to the other two age groups. Malay patients tend to have a poorer survival after breast cancer compared to the other three ethnic groups. Patients who received either local therapy treatment or systemic therapy treatment had higher survival rather than those who received both local and systemic therapy treatments. Treatment group was found to have a significant difference in the survival experience; whereas age group and ethnic group of the breast cancer patients do not have a significant difference in the survival experience.

REFERENCES

[1]. "Breast Cancer," *MedicineNet*, 2018. [Online]. Available: https://www.medicinenet.com/breast_cancer_facts_stages/article.htm#breast_cancer_facts. [Accessed: 10-Oct-2018].

[2]. N. Nordin, N. M. Yaacob, N. H. Abdullah, and S. M. Hairon, "Survival Time and Prognostic Factors for Breast Cancer among Women in North-East Peninsular Malaysia," *Asian Pacific J. Cancer Prev.*, vol. 19, no. 2, pp. 497–502, 2018.

[3]. "Cancer: Breast cancer," *World Health Organization*, 2018. [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. [Accessed: 10-Oct-2018].

[4]. N. B. Pathy *et al.*, "Ethnic differences in survival after breast cancer in South East Asia," *PLoS One*, vol. 7, no. 2, pp. 1–6, 2012.
[5]. V. M. Medina, A. Laudico, M. R. Mirasol-Lumague, H. Brenner, and M. T. Redaniel, "Cumulative incidence trends of selected cancer sites in a Philippine population from 1983 to 2002: A joinpoint analysis," *Br. J. Cancer*, vol. 102, no. 9, pp. 1411–1414, 2010.
[6]. S. K. Park, Y. Kim, D. Kang, E. J. Jung, and K. Y. Yoo, "Risk factors and control strategies for the rapidly rising rate of breast cancer in Korea," *J. Breast Cancer*, vol. 14, no. 2, pp. 79–87, 2011.
[7]. N. A. Abdullah *et al.*, "Survival Rate of Breast Cancer Patients In Malaysia: A Population-based Study," *Asian Pacific J. Cancer Prev.*, vol. 14, no. 8, pp. 4591–4594, 2013.
[8]. "Estimated age-standardized incidence rates and mortality rates (World) in 2018, worldwide, both sexes, ages 0-74," *IARC*, 2018. [Online]. Available: http://gco.iarc.fr/today. [Accessed: 15-Oct-2018].
[9]. C.T.C. Arsene &P.J.G. Lisboa "*Artificial Neural Networks Used in the Survival Analysis of Breast Cancer Patients. A Node-Negative Study.Outcome Prediction in Cancer"*, pp. 191-239, 2007.
[10]. R. Singh &K. Mukhopadhyay. "Survival analysis in clinical trials: Basics and must know areas". *Perspectives in Clinical Research*, *2*(4), pp. 145–148, 2011.
[11]. I. Zwiener, M.Blettner &G. Hommel. Survival analysis: part 15 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *108*(10), pp. 163–169, 2011.
[12]. A. Hazra &N. Gogtay."Biostatistics Series Module 9: Survival Analysis". *Indian Journal of Dermatology*, *62*(3), pp. 251–257, 2017.
[13]. K. J. Jager, P. C. Van Dijk, C. Zoccali &F.W. Dekker. "The analysis of survival data: The Kaplan-Meier method". *Kidney International*, *74*(5), pp. 560–565, 2008.
[14]. R. P. Larson &A.M. Lindner. "Professionalization through attrition? An event history analysis of mortalities in citizen journalism". *Information, Communication and Society*, *21*(5), pp. 746–760, 2018.
[15]. P. Rajbongshi &S. Thongram. "Survival Analysis of Fatigue and Rutting Failures in Asphalt Pavements". *Journal of Engineering*, *2016*, pp. 1–7, 2016.
[16]. O. B. Ajagbe, Z. Kabair &T. O'Connor. "Survival analysis of adult tuberculosis disease". *PLoS ONE*, *9*(11), pp. 1–10, 2014.
[17]. M. Montaseri, J. Y.Charati &F. Espahbodi."Application of Parametric Models to a Survival Analysis of Hemodialysis Patients". *Nephro-Urology Monthly*, *8*(6), pp. 1–6, 2016.
[18]. M. T. Redaniel, R. M. Martin, D. Gillatt, J.Wade &M. Jeffreys."Time from diagnosis to surgery and prostate cancer survival: A retrospective cohort study". *BMC Cancer*, *13*(1), pp. 1–6, 2013.
[19]. M. K. Goel, P. Khanna &J. Kishore. "Understanding survival analysis: Kaplan-Meier estimate". *International Journal of Ayurveda Research*, *1*(4), pp. 212–216, 2010.
[20]. P. Sedgwick. "The log rank test". *BMJ (Online)*, *341*(7765), pp. 1–2, 2010.
[21]. B. George, S. Seals &I. Aban. "Survival analysis and regression models". *Journal of Nuclear Cardiology*, *21*(4), pp. 686–694, 2014.
[22]. M. L. McHugh, "The Chi-square test of independence," *Biochem. Medica*, vol. 23, no. 2, pp. 143–149, 2013.
[23]. R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," *Perspect. Clin. Res.*, vol. 2, no. 4, pp. 145–148, 2011.
[24]. A. Hazra and N. Gogtay, "Biostatistics Series Module 9: Survival Analysis," *Indian J. Dermatol.*, vol. 62, no. 3, pp. 251–257, 2017.
[25]. M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate," *Int. J. Ayurveda Res.*, vol. 1, no. 4, pp. 212–216, 2010.
[26]. M. Usman, H. G. Dikko, S. Bala, and S. U. Gulumbe, "An application of kaplan-meier survival analysis using breast cancer data," *Sub-Saharan African J. Med.*, vol. 1, no. 3, pp. 132–137, 2014.

AzmeKhamis, et. al. "Survival Analysis for Breast Cancer Patients: Case Study at Public Hospital in Malaysia." *The International Journal of Engineering and Science (IJES)*, 9(9), (2020): pp. 01-06.