JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 5, No. 2, October 2021, pp. 340-358

Breast Cancer Survival Analysis Using Cox Proportional Hazard Regression and Kaplan Meier Method

Yuniar Farida[1], Eka Agustina Maulida[2], Latifatun Nadya Desinaini[3], Wika Dianita Utami[4], Dian Yuliati[5]
[1,2,3,4,5]Department of Mathematics, UIN Sunan Ampel Surabaya, Indonesia
yuniar_farida@uinsby.ac.id[1], ekaagustinamaulida@gmail.com[2], nadya.desinaini@gmail.com[3], wikadianita@uinsby.ac.id[4], dian.yuliati@uinsby.ac.id[5]

Article History:

Received : 30-04-2021
Revised : 28-05-2021
Accepted : 30-07-2021
Online : 26-10-2021

Keywords:

Survival Analysis;
Kaplan Meier;
Cox Proportional Hazard;
Breast Cancer.

ABSTRACT

Breast cancer is one of the malignant tumors that begins in the breast cells that develop and attack the surrounding tissues; according to World Health Organization (WHO), breast cancer is globally declared the top five killer cancers. In Indonesia, breast cancer becomes the number one killer cancer. One of the successes in breast cancer treatment is if the cure obtained by cancer patients can be proven to have the same life expectancy as those who do not have breast cancer. This study aims to know the probability of survival of breast cancer patients and know the factors that affect breast cancer patients' survival. The data were consist of 394 medical records of breast cancer patients at Dr. Soetomo Hospital Surabaya in the period January 2018 – December 2019, with variables used, i.e., initial age of infection, clinical stage, tumor size, metastatic to other organs, type of treatment, and patient status (life or death). This study using Kaplan Meier and Cox Proportional Hazard regression methods, and the result showed that the probability of survival of breast cancer patients (with data samples) was 0.737 or 73.7%. The variables that significantly affect breast cancer patients' survival are the initial age of infection, the clinic stage, and the tumor's size. This research provides information and motivation to the community related to life expectancy, especially in breast cancer patients, to stay motivated in the healing process. In addition, this research is also used to add insight to academics, especially the department of statistics, regarding the regression of Cox Proportional Hazard in analyzing the survival of breast cancer patients.

https://doi.org/10.31764/jtam.v5i2.4653

A. INTRODUCTION

Breast cancer is one type of cancer in women and is still a health problem in women worldwide and the second most common malignancy disease that causes death. Breast cancer is a disease derived from malignant cells that form tumors and are then detected in breast tissue. Malignant tumors can develop in breast tissue such as mammary glands, milk ducts, fat tissue, and other connective tissues (Reyna & Lee, 2014) (Sun et al., 2017). Breast cancer affects the breast's glands, ducts, and tissues but does not include its skin. Cancer-detecting breast cells and tissues will make changes in the shape of those cells and tissues abnormal and multiply

uncontrollably (Dewi & Hendrati, 2015). In general, the main complaint in patients with breast cancer is swelling of the breast. Initially, the lump is small, but the longer it grows and attaches to the skin or causes changes to the skin of the breast or nipple (Ghodsi, Salehi, & Hojjatoleslami, 2013).

Various factors are the reason for breast cancer, genetic, family, hormonal, and obesity factors. The World Health Organization (WHO) in 2004 stated that there are five significant cancers in the world, namely lung cancer, breast cancer, colon cancer, gastric cancer, and liver cancer. A survey was conducted by who stated that a person with breast cancer is 8-9% in women. In 2012 there was an increase in the incidence of breast cancer globally; 1.7 million women were diagnosed with breast cancer, and 6.3 million women were diagnosed with breast cancer in the previous five years. Since 2008 there has been an estimated 20% increase in breast cancer incidence, with a 14% increase in mortality, until the occurrence of a specific event called failure event. Every year, over 1.5 million women worldwide (25 percent of all cancer patients) are diagnosed with breast cancer (Kleibl & Kristensen, 2016) (Sun et al., 2017).

Based on Globocan data, in 2018, cancer is the second leading cause of death globally and is responsible for about 9.6 million deaths. While in Indonesia, in the same year amounted to 207,210. The incidence of cancer in Indonesia (136.2/100,000 inhabitants) is 8th in Southeast Asia. While in Asia, it ranks 23rd, Indonesia's highest incidence rate for men is lung cancer, 19.4 per 100,000 inhabitants, with an average mortality of 10.9 per 100,000 inhabitants. The second place is liver cancer of 12.4 per 100,000 inhabitants, with an average mortality of 7.6 per 100,000 inhabitants. At the same time, the highest incidence rate for women is breast cancer at 42.1 per 100,000 inhabitants, with an average mortality of 17 per 100,000 inhabitants. The second place is cervical cancer of 23.4 per 100,000 inhabitants, with an average mortality of 13.9 per 100,000 inhabitants (K. kesehatan RI, 2020)(D. P. K. kesehatan RI, 2019).

High breast cancer cases are caused because breast cancer patients are often unaware of or feel breast cancer symptoms. If breast cancer is detected in the advanced stage, more expensive treatment, more difficult treatment results are not maximal and even accelerate death. One of the successes in breast cancer treatment is if the cure obtained by cancer patients can be proven to have the same life expectancy as the population who do not have breast cancer. The benchmark for successful cancer treatment is the patient's survival rate. One of the most commonly used methods is Kaplan Meier's analysis, followed by a Log Rank test and Cox Proportional Hazard regression.

Survival analysis (endurance analysis) is a statistical method in which the variable that is observed is a variable of time until the occurrence of events (died) or commonly called survival time (Ihwah, 2015). One of the most widely used methods is Kaplan Meier's analysis, followed by a Log Rank test and Cox Proportional Hazard regression. Kaplan Meier's analysis is used to assess survival functions. In contrast, the Log Rank test is used to test whether there is a difference in the survival curve Kaplan Meier. Meanwhile, Cox Proportional Hazard's regression is used to determine the combination of factors that affect his response in survival time.

The key to problem analysis to consider in resiliency analysis is the censored data. The censored data is not discarded but still considered because the minimum up to a certain point can still be seen as having not experienced the event and assuming that sensor events within a specific time occur evenly. Three factors must be considered in determining survival time. First,

the start point cannot be ambiguous, or there are no two or more meanings. Second, the event of the whole affair should be clear. Third, the scale of survival time measurement should be precise (Nurfain & Purnami, 2017) (Ebrahimi et al., 2019).

Some studies on life resilience analysis include Yulianto, Notobroto, & Widodo (2017), conducted the survival analysis of Chronic Kidney Disease (CKD) patients with hemodialysis at Dr. Soetomo Surabaya period 2010-2013 using Kaplan Meier test and Log Rank. The results of the study stated that CKD patients undergoing hemodialysis at Dr. Soetomo Hospital, with the age range of 46-65 years, have a history of hypertension and diabetes mellitus would have a lower average survival compared to patients aged between 26-45 years and have no record of both diseases. Wijaya & Wulandari (2015) also conducted survival analysis in patients with Acute Coronary Syndrome (ACS) at Dr. Soetomo Hospital Surabaya in 2013 using Cox Proportional Hazard Regression. The results of the study stated that on the 5th to 10th day, it was possible that the patient did not experience clinical improvement, and the factors that influenced the rate of clinical improvement of SKA patients were dyslipidemia status, diabetes mellitus, hypertension, and hemodynamic profile. Similar research was also conducted by Nurfain & Purnami (2017), which analyzed Cox Extended regression in leprosy patients in Brondong Subdistrict Lamongan district in 2012-2015 190th day, many patients experienced clinical improvement, and they declared Release From Treatment (RF).

Yadav et al. (2021) conducted a survival analysis between men and women with triple-negative breast cancer (TNBC). In this study, the baseline demographic and cancer characteristics of men and women were compared using the Pearson's Chi-Square test for categorical variables and the Mann-Whitney U test for continuous variables in this paper. A Kaplan Meier and multivariate Cox proportional hazards regression model were applied in this study to compare survival and find prognostic markers. The result of this paper shows that 3-year and 5-year overall survival rates in men were 74.8 percent and 68.8 percent, respectively, while women's rates were 83.2 percent and 74.8 percent. Men had a considerably worse overall survival rate than women (HR: 1.49, 95 percent CI: 1.19-1.86, p =0.01), according to multivariate analysis. In men with TNBC, older age at diagnosis, higher TNM stage, mastectomy, and lack of chemotherapy or radiation were independent negative prognostic markers.

Jawitz et al. (2020) used survival analysis to examine recipient survival under the new system using an updated dataset. In this study, the Kaplan-Meier technique and multivariable Cox proportional hazards regression were used to investigate the relationship between the allocation system and recipient mortality. According to the findings of this study, the short-term survival of recipients listed and receiving a transplant under the old and new allocation processes appears to be equal. The alteration in the allocation system has resulted in several changes in the clinical characteristics of patients undergoing transplants, which will need to be constantly studied in future years.

Han et al. (2021) used the Cox regression model to find characteristics that predicted DNS development. Kaplan–Meier curves were created to quantify the cumulative incidence of DNS. The key predictors of DNS development were identified using a multivariate Cox regression model. According to the findings of this study, the incidence of DNS was 18.8%, with a median onset time of 23.7 days (interquartile range, 14–30 days). A higher cumulative incidence of DNS was related with a blood creatine kinase (CK) level > 175.5 U/L, and an initial Glasgow Coma

Scale (GCS) score of 9 (log-rank test; p = 0.02, respectively). A serum CK level > 175.5 U/L (hazard ratio [HR]: 2.862, 95 percent confidence interval [CI]: 1.491–5.496; p = 0.01) and an initial GCS of 9 (HR: 2.081, 95 percent confidence interval [CI]: 1.048–4.131; p = 0.04) were significant prognostic variables, according to Cox regression analysis.

Based on the research above, the Kaplan Meier method of Log Rank test and Cox Proportional Hazard regression is a powerful and widely used survival analysis approach. The ability to display unadjusted and adjusted HRs (hazard ratios) with their corresponding CIs is the main benefit of Cox regression analysis (confidence interval). Other than that, Cox Proportional Hazard regression does not have assumptions about properties and shapes according to the distribution as assumptions in the other regressions (Stel, Dekker, Tripepi, Zoccali, & Jager, 2011) (Julia, 2012). Because of that, in this study, the Cox Proportional Hazard regression and the Kaplan Meier method were used to analyzing breast cancer patients' survival. This study's results are expected to help determine the probability of survival of breast cancer patients and factors that affect their survival to evaluate whether the treatment is good or not.

B. METHODS

1. Data

This study uses secondary data from the medical record section in Dr. Soetomo Hospital Surabaya from January 2018 until December 2019. The data acquired amounted to 349 patients with details of 201 patients still surviving, 91 patients have died, and 57 patients are missing from observation. Here is a description of the variables used in the study.

Table 1. Research Variables

Variables		Description
T	Survival Time (days)	The time during the patient undergoing hospital treatment 0 = if the patient is missing from the research time and the patient is still surviving 1 = if the patient dies
X_1	Age (Years)	Early age of infection
X_2	Stadium	1 = Early stage (0, I, and II) 2 = Advanced stage (III and IV)
X_3	Tumor Size	1 = ≤5 cm 2 = >5 cm
X_4	*Metastasis*	0 = have not *Metastasis* 1 = have *Metastasis*
X_5	Types of Treatment	1 = Radiotherapy 2 = Chemotherapy

2. Data Analysis

Data analysis techniques are the steps to solve problems from start to finish. Data analysis techniques in this study, namely as follows:

a. Collection of breast cancer data obtained from the medical records section in Dr. Soetomo Hospital Surabaya during January 2018 – December 2019.

b. Describe breast cancer patients' characteristics based on survival time and factors that affect their survival.

1) Survival Function

The $S(t)$ survival function is defined as the probability of an object surviving from a survival time greater than or equal to t. The survival function can also be described as a smooth graph/curve, with $S(t)$ being the column and t being the row. In this case, the chart/curve may decrease from $S(t) = 1$ at $t = 0$ to $S(t) = 0$ on $t = \infty$. In other words, at the time = 0, life chance = 1, and at an infinite time, his life chance = 0. Suppose T is a random variable that symbolizes survival time and has the function of $f(t)$, opportunity distribution, so (Kartsonaki, 2016).

$$S(t) = P(T > t)$$

$$= 1 - f(t) = 1 - P(T \leq t) \quad (1)$$

2) Hazard Function

The Hazard $h(t)$ function defines a momentary failure rate assuming that an object reaches an event at a time interval of t to $(t + \Delta t)$, provided that it has survived until that time (Kartsonaki, 2016). So obtained:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\big(individuals\ in\ t\ experience\ events\ in\ the\ hose\ (t, t+\Delta t)\big)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < (t+\Delta t)\ |T \geq t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < (t+\Delta t))}{P(T \geq t)\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < (t+\Delta t))}{S(t)\Delta t}$$

$$= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < (t+\Delta t))}{\Delta t}$$

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < (t+\Delta t))}{\Delta t} \quad (2)$$

So that it can be stated the relationship between survival function and hazard function is as follows:

$$h(t) = \frac{f(t)}{S(t)} \quad (3)$$

c. Describe the survival curve of breast cancer patients with Kaplan Meier's analysis

Suppose there are n breast cancer individuals observed with long life, $t_1, t_2, \ldots, t_n$ and there is a j individual who dies $(j \leq n)$ in the order of the time of death $t_{(1)} \leq t_{(2)} \leq$

$\cdots \leq t_{(j)}$. Meanwhile $n_{(j)}$ is the number of individuals at risk of dying at $t_{(j)}$ and $d_{(j)}$ are individuals who die at $t_{(j)}$. Thus the estimate of Kaplan Meier $\hat{S}(t)$ is as follows (Zare et al., 2014):

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_{(j)}}{n_{(j)}}\right) \quad (4)$$

The Log Rank test is used in comparing whether there is a difference between Kaplan Meier's survival curves. Here are the hypotheses in the Log Rank test:
H_0: There is no difference in Kaplan Meier's survival curve between different groups.
H_1: There is at least one difference in Kaplan Meier's survival curve between other groups.

With test statistics as follows:

$$\chi^2 = \sum_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i} \quad (5)$$

$$o_i - E_i = \sum_{j-1}^{n} \sum_{i=1}^{G} (m_{ij} - e_{ij}) \text{ and} \quad (6)$$

$$e_{ij} = \left(\frac{n_{ij}}{\sum_{j=1}^{n} \sum_{i=1}^{G} n_{ij}}\right) \left(\sum_{j-1}^{n} \sum_{i=1}^{G} m_{ij}\right) \quad (7)$$

Where G is the number of groups; O_i is individual observation values of the *i;* E_i is individual expectation values of the group to *i;* m_{ij} is the number of subjects who died in the i group at the time of $t_{(j)}$; n_{ij} is the number of subjects at risk of dying in the i group at the time of $t_{(j)}$; and e_{ij} is individual expectation values of the i group at the time of $t_{(j)}$. Decision making for this statistics is *H0* rejected if $\chi^2 > \chi^2_{(\alpha;G-1)}$ or *p-value* < $\alpha = 0.05$ (Kleinbaum & Klein, 2012)

d. Test the differences in breast cancer patients' survival curve based on the results in the second step with the Log Rank test with Equations (5)

e. Test proportional hazard assumptions.
The proportional failure function assumes that the failure ratio function should be constant over time (Dwidayati, 2016). The way to test the hypothesis of proportional failure is by visual test and formal test.
   1) Visual Test
      Determining the assumption of proportional failure on visual tests can use Kaplan Meier's survival curve approach. The survival curve is said not to meet proportional hazard assumptions when the survival lines between groups intersect. The survival curve meets proportional hazard assumptions when the survival lines between groups do not intersect (Selvaraj et al., 2014).

2) Formal Test

Determining the assumption of failure proportional to a formal test can be approached with a Goodness of Fit (GoF) test. There are three steps that must be taken in GoF testing. First, regress survival time with its free variables to obtain Schoenfeld residual values. Second, create time variables that have been sorted from smallest to largest. Third, test the correlation between Schoenfeld residuals and sorted time variables. The hypothesis that uses in this test is:

H_0: ρ = 0 (Assumptions of proportional failure are met);

H_1:ρ≠0 (Assumptions of proportional failure are not met).

With the decision making is H_0 rejected if P-value > α=5% (Zhou, Fine, & Laird, 2013).

f. Create a Cox Proportional Hazard regression model

Modeling survival data using the cox proportional hazard model uses a parametric method to estimate the covariate effect on survival data. Cox's regression is used to determine the influencing factors in survival data for uncensored data (Lee, Moon, & Salamatian, 2012). If X is a vector-sized $p \times 1$ where the elements are covariate $X_1, X_2, \dots, X_p$, then the Cox Proportional Hazard model is

$$h_i(t_j|X) = h_0(t|X)\exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})$$

$$= h_0(t)\exp\sum_{j=1}^{p}\beta_j x_{ji} \quad (8)$$

Where X is $(X_1, X_2, \dots, X_p)$ is an explanatory/predictor variable; $h_0(t)$ is basic failure function; $h_i(t_j|X)$ is individual failure function *i;* x_{ji} is variable value *j* from individual *i*, with $j = 1,2, \dots, p$ and $i = 1,2, \dots, n$; and β_j is *regression coefficient j*, with $j = 1,2, \dots, p$.

g. Test parameters with the Likelihood Ratio test

Parameter testing determines whether independent variables affect dependent variables (Yi & Wang, 2011).

1) Simultaneous Testing (Likelihood Ratio Testing)

Hypothesis:

H₀ : $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (no variables significantly affect)

H₁ : there is at least one $\beta_j \neq 0$, with $j = 1,2, \dots, p$ (there is at least one variable that substantially affects)

Test statistics:

$$\chi^2_{LR} = 2\ logL_v - 2\log L_0 \quad (9)$$

Where L_v is likelihood function value with the independent variable; L_0 is likelihood function value with the independent variable; and p is the number of parameters β. With decision making is H_0 rejected if $\chi^2_{LR} > \chi^2_{p;\alpha}$ or *p-value* < α=5%.

2) Partial Testing (Wald Testing)

Hypothesis:

H₀ : $\beta_j = 0$, with $j = 1,2, \dots, p$ (j – variable has no significant effect)

$H_1 : \beta_j \neq 0$, (j – variable has no significant impact)
Test statistics:

$$\chi^2_{Wald} = \left[\frac{\hat{\beta}_j}{SE\,(\hat{\beta}_j)}\right]^2 \quad (10)$$

$$SE\,(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)} \quad (11)$$

Where $SE\,(\hat{\beta}_j)$ is deviation standard from $\hat{\beta}_j$; and $var(\hat{\beta}_j)$ is variance from $\hat{\beta}_j$. With decision making is H_0 rejected if $\chi^2_{Wald} > \chi^2_{1;\alpha}$ or $p\text{-}value < \alpha$=5%.

h. Calculate hazard ratio
The hazard ratio is the failure of one group of individuals divided by the inability of different individuals failure. Two groups of compared individuals are distinguished by their dependent variables (Uno et al., 2015). Calculating hazard ratios can use standard equations for hazard function, i.e (Lee et al., 2012) (Devarajan & Ebrahimi, 2011).

$$H(t) = H_0(t)e^{y} \quad (12)$$

$$HR = \frac{H(t)^*}{H(t)} \quad (13)$$

Where HR is hazard ratio; $H(t)$ is a hazard at any given time; $H_0(t)$ is the baseline hazard at any given time; e is natural number = 2.714; and $H(t)^*$ is a hazard at any given time for one group of individuals

C. RESULT AND DISCUSSION

1. Kaplan Meier's Survival Curve

This descriptive analysis using Kaplan Meier's survival curve is used to determine the survival picture of breast cancer patients in general. Before drawing the survival curve of Kaplan Meier, calculating the probability of breast cancer patients' survival for two years using Equation (5). After getting the results from estimating the probability of each time with Equation (5), the next Kaplan Meier survival curve will be made, as shown in Figure 1.

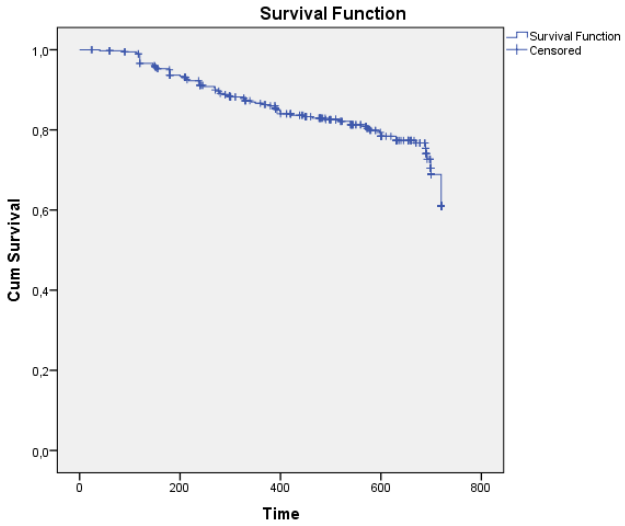

Figure 1. Kaplan Meier's Survival Curve Breast Cancer Patient

Figure 1 obtained survival curve decreased slowly, meaning that the curve explained that a lot of censored data or a lot of data that did not experience the event that is breast cancer patients who died during the study time is two years. It means that there are still many breast cancer patients who still survive in the space of two years. Based on these calculations, the results are that the probability of survival of breast cancer patients over two years is still high at 0.737 or 73.7%.

The following will explain breast cancer patients' characteristics based on suspected factors to affect her using the Kaplan Meier survival curve. To get the probability value in each early age group contracting breast cancer can be calculated using Equation (5). From the research data, the initial age of infection in breast cancer patients was divided into four groups, namely patients with the initial age of contracting 31-40 years, 41-50 years, 50-60 years, and >60 years. After getting the results of the calculation of the probability of each time, the Kaplan Meier survival curve for breast cancer patients will be made based on the initial age of infection, as shown in Figure 2.

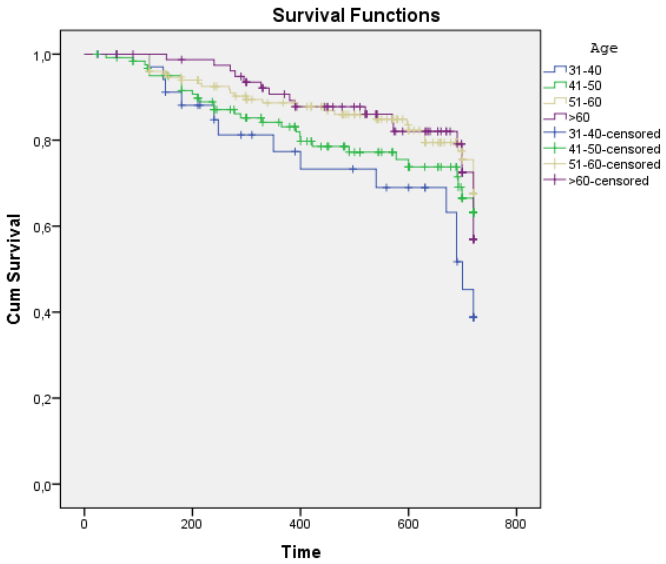

Figure 2. Kaplan Meier's Survival Curve Based on Early Infected Age

Based on Figure 2 graphically, it appears that breast cancer patients with early age 31-40 years have a probability of survival between 0.4 to 1. It can be said that the patient's survival is low. Breast cancer patients with early age 41-50 years, 51-60 years, and >60 years have a probability of survival between 0.7 to 1. It can be said that the patient's survival is relatively high. When viewed from the likelihood of survival of breast cancer patients of early age infected 30-40 years by 0.656 or 65.6%, the initial age of contracting 41-50 years is 0.747 or 74.7%, the initial age of contracting 51-60 years is 0.801 or 80.1% and the initial age of contracting >60 years is 0.778 or 77.8%. Based on this explanation, it is suspected that there are no differences in the survival curves for the four age groups of patients when they first contracted breast cancer.

The Log Rank test can be used to determine whether there is a difference in survival time in the early infected age group. The calculation of the Log Rank test can use the formula in Equation (6). The analysis found that the highest survival of patients is seen in patients whose initial age is more than 50 years old, and the lowest occurs in the early age of the infection 30 – 40 years. Similar to the research conducted by Suganda et al. (2021) showed the highest survival rate occurred at the initial age of disease, more than 60 years, namely 74.1%. Breast cancer is common in women aged 45 and over but has recently shifted, with breast cancer affecting women aged 20 to 30 more (Arshi et al., 2018). Low survival of patients with early age breast cancer under 40 years is associated with hormonal factors that are still active, so the risk of developing breast cancer becomes higher. It can also be linked to the tumor's size when detected, or cancer cells attack much more malignantly. Patients with early age breast cancer over 60 years old are usually associated with the body's condition, weakened cells, or other disease factors (Jobsen et al., 2019).

After the age factor, the next factor that needs to be reviewed is the clinical stage factor. The clinical-stage is one of the factors that affect the survival of breast cancer patients. Clinical staging in breast cancer patients is divided into two groups: early and advanced. To get the probability value in each stage group can be calculated using Equation (5). After getting the results from estimating the probability of each time, the next Kaplan Meier survival curve for breast cancer patients will be made based on the clinical stage, as shown in Figure (3).

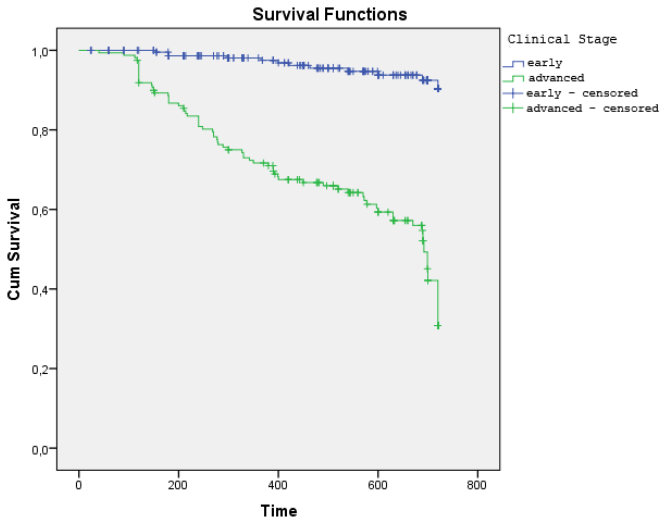

Figure 3. Kaplan Meier's Survival Curve Based on Clinical Stage

Based on Figure 3 graphically, it appears that breast cancer patients with early-stage from early admission up to 720 days have a higher probability of survival between 0.9 to 1. It can be said that the patient's survival is high. Breast cancer patients with advanced stages from the beginning of entry have a decreased curve with a probability of between 0.3 to 1. The patient has low survival, which is viewed from the probability of survival of early-stage breast cancer patients by 0.944 or 94.4%, advanced patients by 0.503 or 50.3%. Based on this explanation, it is suspected that there are differences in the survival curves for the two-stage groups.

To determine whether there is a difference in survival time at the clinical stage, the Log Rank test can be used. Calculation of the Log Rank test can use the formula in Equation (6). The highest patient survival is found in patients with an early stage, and the survival of patients with advanced stages is very low, and there is a difference in the proportion of survival of breast cancer patients with research conducted by Suganda et al. (2021) showing the low survival of advanced-stage patients with the probability of survival is 46.6%. The low survival of advanced patients is due to the advanced stage has involved more life nodes, while the lymph nodes themselves have a role as the body's defense system. Also, in the advanced stages of cancer that attacks the patient has spread to other organs that impact the impaired function of the body organs and the vulnerability of sufferers to infection.

The next factor is the tumor size factor. Tumor size is one of the factors that affect the survival of breast cancer patients. Tumor size in breast cancer patients was divided into two groups, namely 5 cm and > 5 cm. Get the probability value in each group of tumor size can be calculated using Equation (5). After getting the results from estimating the probability of each time, the next Kaplan Meier survival curve for breast cancer patients will be made based on the Tumor size factors as shown in Figure 4.

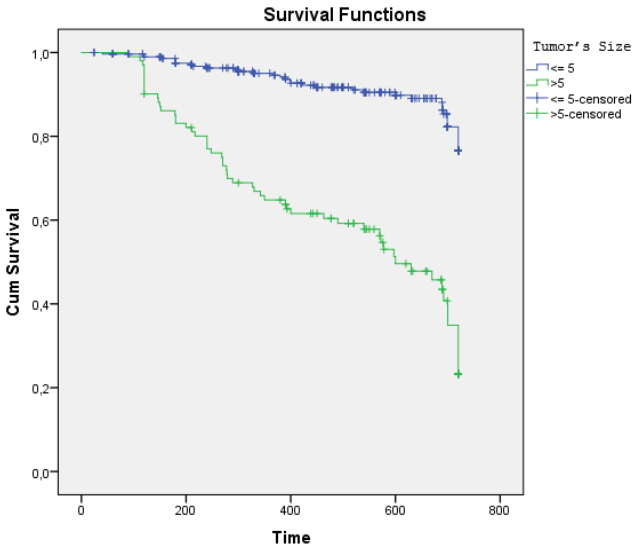

Figure 4. Kaplan Meier's Survival Curve Based on Tumor Size

Based on Figure 4, it is shown that breast cancer patients with tumor size ≤5 cm from initial admission up to 720 days have a higher probability of survival between 0.8 to 1. It can be said that the patient's survival is high. For breast cancer patients with tumor size >5 cm from start to finish decreased with a probability of survival of 0.3 to 1, meaning that the patient has a low survival. Others, if the probability of survival of breast cancer patients with a tumor size of ≤5

cm by 0.877 or 87.7%, and patients with tumor size >5cm by 0.434 or 43.4%. Based on this explanation, it is suspected that there are differences in the survival curves for the two tumor size groups.

To determine whether there is a difference in survival time on tumor size, the Log Rank test can be used. The calculation of the Log Rank test can use the formula in Equation (6). The result shows that low survival of breast cancer patients with tumors of >5 cm is commonly associated with distant lymph nodes. The larger the size of the detected tumor, the more positive node lymph. When the patient receives treatment when the tumor's size has enlarged, the treatment rate becomes lower. There are residual tumors after operative therapy, where the large number of tumors that remain can cause recurrence rates in breast cancer patients (Yao et al., 2020).

The next factor is metastasis. Metastasis is one of the factors that affect the survival of breast cancer patients. Metastases in breast cancer patients were divided into two groups, namely those with metastases and no metastases. To get the probability value in each group of metastases can be calculated using Equation (5). After getting the results from estimating the probability of each time, the next Kaplan Meier survival curve for breast cancer patients will be made based on the metastasis factors, as shown in Figure 5.

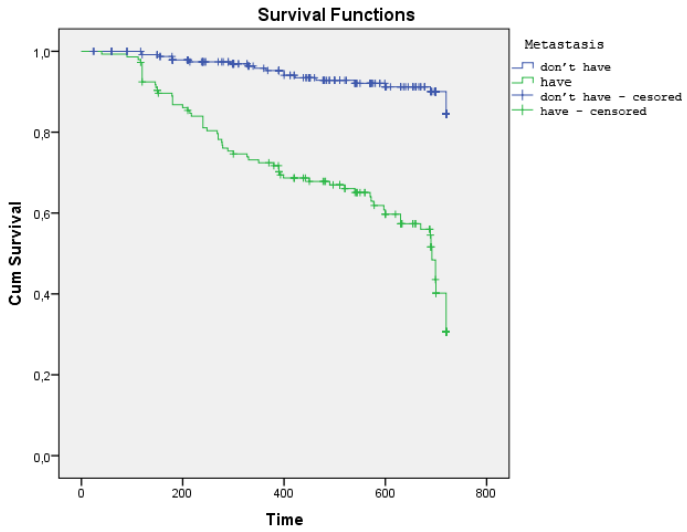

Figure 5. Kaplan Meier's Survival Curve Based on Metastasis

Based on Figure 5, it is graphically seen that breast cancer patients who do not get metastasis have a higher probability of survival between 0.9 to 1. It can be said that the patient's survival is high. However, breast cancer patients with metastatic cancer have a probability of between 0.4 to 1. It means that the patient's survival is low when viewed from the probability of survival of breast cancer patients who do not have metastases by 0.907 or 90.7%, and patients with metastases of 0.527 or 52.7%. Based on this explanation, it is suspected that there are differences in the survival curves for the two metastases.

To find out whether there is a difference in survival time in metastases, the Log Rank test can be used. Calculation of the Log Rank test can use the formula in Equation (6). The result shows that breast cancer patients with cancer cells that have spread have a low survival rate of 57.2%, while patients who do not have metastases have a high survival. Because in patients with metastases, there is often a spread to internal organs such as the lungs, brain, bones, etc.

This causes the organ's malfunction, affecting breast cancer patients' survival (Febriani & Furqon, 2018).

The last factor is the type of treatment, the kind of treatment is one of the factors that affect the survival of breast cancer patients. Types of treatment in breast cancer patients are divided into two groups, namely radiotherapy, and chemotherapy. To get the probability value for each group of the kinds of treatment can be calculated using the formula in Equation (5). After getting the results from calculating the probability of each time, the Kaplan Meier survival curve for breast cancer patients will then be made based on the type of treatment as shown in Figure 6.

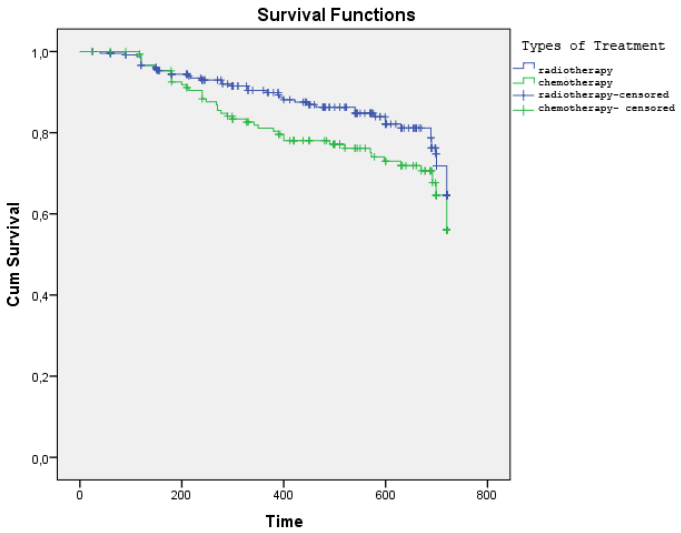

Figure 6. Kaplan Meier's Survival Curve By Treatment Type

Based on Figure 6, it is graphically seen that breast cancer patients undergoing treatment with radiotherapy from early admission to 720 days have a higher probability of survival between 0.7 to 1. It can be said that the patient's survival is high. Breast cancer patients undergoing treatment with chemotherapy had a higher probability of survival between 0.6 to 1. It can be said that the patient's survival is relatively high. The probability of survival of breast cancer patients undergoing treatment with radiotherapy by 0.792 or 79.2%. The probability of patients undergoing treatment with chemotherapy by 0.724 or 72.4%. Based on this explanation, it is suspected that there is no difference in the survival curve for the two treatment groups.

To determine whether there is a difference in survival time on the type of treatment, the Log Rank test can be used. The calculation of the Log Rank test can use the formula in Equation (6). Breast cancer patients treated with radiotherapy have a higher survival rate than patients treated with chemotherapy. This is similar to a study conducted by Wijaya & Wulandari (2015) which showed that the two-year survival of breast cancer patients undergoing chemoradiation was higher than patients undergoing chemotherapy. Low survival of patients undergoing chemotherapy is usually associated with patients already in advanced condition when starting therapy, where the physical and systemic diseases of patients who do not allow surgery or chemotherapy.

2. Log Rank Testing

Furthermore, a log-rank test is used to determine if there is a difference between survival times. The following are the log-rank test results based on suspected factors to affect breast cancer patients' survival.

Table 2. Log Rank Test Results

Variables	df	χ^2 Count
Early Age Infected	3	3.745
Clinical Stadium	1	105.977
Tumor Size	1	96.633
Metastasis	1	82.792
Types of Treatment	1	2.360

Based on the results of the log rank test, it can be known that the survival time of breast cancer patients based on the variable age of the initial infection ($3.745 < \chi2(0.05,3) = 7.815$) and types of treatment ($2.360 < \chi2(0.05,1) = 3.841$) there is no significant difference. Meanwhile, breast cancer patient survival time based on clinical stage variables ($105.977 > \chi2(0.05,1) = 3.841$), tumor size ($96.633 > \chi2(0.05,1) = 3.841$), metastasis ($82.792 > \chi2(0.05,1) = 3.841$) there are significant differences in.

3. Proportional Hazard Assumption Test

Proportional hazard assumption testing is also conducted with the Goodness of Fit test approach. The Goodness of Fit test is performed to obtain more objective decisions. In this test, H0 noted that factors that are thought to affect breast cancer patients' survival meet proportional hazard assumptions. H1 believes that elements that are supposed to affect breast cancer patients' survival do not meet the proportional hazard assumptions. Here is the goodness of fit test for all factors thought to affect breast cancer patients' survival.

Table 3. Goodness of Fit Test Results

	Times		
	Pearson Correlation	**Sig. (2 tailed)**	**N**
Times	1		394
Unstandardized Residual	-0.231	0.000	394
Early Age Infected	0.050	0.320	394
Clinical Stadium	0.063	0.214	394
Tumor Size	0.129	0.051	394
Metastasis	-0.092	0.069	394
Types of Treatment	0.008	0.873	394

Based on Table 3 obtained, the goodness of fit test results is a variable that meets the assumption of proportional hazard in all variables because the p-value of all variables is more significant than the α of 0.05. It is also claimed that it can be directly done modeling using cox proportional hazard regression.

4. Cox Proportional Hazard regression

Next is the creation of a model with cox proportional hazard regression. In this step obtained the results of regression as in Table 4

Table 4. First Cox Proportional Hazard Regression Results

	B	SE	df	Sig.	Exp(B)
Early Age			3	0.337	
Early Age (1)	0.487	0.643	1	0.448	1.628
Early Age (2)	0.887	0.501	1	0.077	2.428
Early Age (3)	0.640	0.502	1	0.202	1.896
Clinical Stadium	-1.334	0.758	1	0.078	0.263
Tumor Size	-1.293	0.379	1	0.001	0.274
Metastasis	0.467	0.533	1	0.381	1.596

From the results of the regression cox proportional hazard above can be obtained the first model, by using Equations (8), obtained the following results:

$$h(t|X) = h_0(t)\exp(0.487\ uearly\ age\ infected\ (1) + 0.887\ early\ age\ infected(2) + 0.640\ early\ age\ infected(3) - 1.334\ clinical\ stadium - 1.293\ tumor\ size\ + 0.467\ metastasis - 0.101\ types\ of\ treatment)$$

5. Parameter Testing

Parameter testing is conducted in two stages: simultaneous testing with the likelihood ratio test and partial testing with the Wald test. Parameter testing simultaneously with likelihood ratio test using the formula in Equation (9) is obtained as follows:

$$\chi^2_{LR} = 2\log L_v - 2\log L_0 = 587.276 - 508.382 = 78.894$$

Because in the test, the likelihood ratio of the value χ^2_{LR} 78.894 greater than the $\chi^2_{5;0.05}$ 11.071 can be concluded reject H_0, meaning at least one variable significantly affects breast cancer patients' survival. After simultaneous testing is then conducted partial parameter testing with the Wald test, here is the following results on Table 5:

Table 5. First Model Wald Test Results

Variables	X^2_{wald}	Decision
Early Age (1)	0.574	Receive H_0
Early Age (2)	3.135	Reject H_0
Early Age (3)	1.625	Receive H_0
Clinical Stadium	3.097	Reject H_0
Tumor Size	0.768	Receive H_0
Metastatic	11.639	Reject H_0
Types of Treatment	0.128	Receive H_0

Based on Table 5, it can be known that variables that have a significant effect or χ^2_{wald} greater than $\chi^2_{1,0.10} = 2.706$ is a variable of the initial age of cancer that is 41-50 years, the clinic stage, and the tumor's size. In contrast, the metastatic variables and types of treatment have no significant effect.

Because there are some insignificant variables, those minor variables are excluded from the first model. Variables that significantly affected the first model have regressed Cox Proportional Hazard again and obtained results as in Table 6.

Table 6. Second Cox Proportional Hazard Regression Results

	B	SE	df	Sig.	Exp(B)
Early Age (2)	0.855	0.499	1	0.086	2.352
Clinical Stadium	-1.797	0.550	1	0.001	0.166
Tumor Size	-1.253	0.376	1	0.001	0.286

From the results of the regression cox proportional hazard above can be obtained the second model using Equation (8), got the following results:

$$h(t|X) = h_0(t)\exp(0.855\ early\ age(2) - 1.797\ clinical\ stadium - 1.253\ tumor\ size)$$

After obtaining the second model will then be conducted parameter tests simultaneously and partially test parameters again. Parameter experimenting simultaneously with likelihood ratio test using the formula in Equation (9) obtained as follows:

$$\chi^2_{LR} = 2\log L_v - 2\log L_0 = 587.276 - 527.440 = 59.836$$

Because in the test, the likelihood ratio of the value χ^2_{LR}59.836 greater than the $\chi^2_{5;0.05}$ 7.815 can be concluded reject H_0, meaning there is at least one variable that significantly affects the survival of breast cancer patients. After simultaneous testing is then conducted partial parameter testing with the Wald test. Before calculating the test Wald, so obtained as follows on Table 7.

Table 7. Second Model Wald Test Results

Variables	X^2_{wald}	Decision
Early Age (2)	2.936	Reject H_0
Clinical Stadium	10.675	Reject H_0
Tumor Size	11.105	Reject H_0

Based on Table 7 above obtained, all variables have a significant effect or χ^2_{wald} greater than $\chi^2_{1,0.10} = 2.706$, so it can be said that the second regression model is the best model with variables that affect the survival of breast cancer patients is the variable of the initial age of cancer is 41-50 years, clinical stage, and tumor size.

6. Hazard Ratio

Furthermore, hazard ratio calculation to determine how much risk the group has on each variable that affects to die. The following is the result of the hazard ratio calculation.

a. Hazard ratio variable early age of cancer is 41-50 years

$$HR = \frac{H(720)\text{UAT } 41-50 \text{ years old}}{H(720)\text{UAT } 51-60 \text{ years old}} = \frac{0.087}{0.205} = 0.424$$

From the hazard ratio, it can be said that the probability of breast cancer patients with early age infected is 51-60 years to survive within two years is 2.358 times (1/0.424) compared to breast cancer patients with an early age of 41-50 years.

b. Hazard ratio variable stage clinic

$$HR = \frac{H(720)\ advanced\ stage}{H(720) early\ stage} = \frac{0.079}{0.477} = 0.166$$

From the hazard ratio, it can be said that the chances of early-stage breast cancer patients to survive within two years is 6.024 times (1/0.166) compared to advanced breast cancer patients.

c. Hazard ratio variable tumor size

$$HR = \frac{H(2) > 5\ cm}{H(2) \leq 5\ \text{cm}} = \frac{0.197}{0.689} = 0.286$$

From the hazard ratio results, it can be said that the probability of breast cancer patients who have a tumor size of ≤5 cm survive within two years is 3.497 times (1/0.286) compared to breast cancer patients who have a tumor size of >5 cm.

D. CONCLUSION AND SUGGESTIONS

From the processing and analysis of data discussed before, it can be concluded that, based on the results of calculations, the probability of survival of breast cancer patients (with a sample of 394 patients over two years) was 0.737 or 73.7%. When reviewed from the variable of the initial age of infection then the probability of survival of breast cancer patients when the initial age of disease is 31-40 years of age by 0.656 or 65.6%, 41-50 years of 0.747 or 74.7%, 51-60 years of 0.801 or 80.1%, and >60 years of 0.778 or 77.8%.

The probability of survival of breast cancer patients based on early-stage variables is 0.944 or 94.4%, and the advanced stage is 0.503 or 50.3%. The probability of survival of breast cancer patients based on variable tumor size ≤5 cm is 0.877 or 87.7%, and that has a tumor size of 0.434 cm or 43.4%. The probability of survival of breast cancer patients with metastasis is 0.527 or 52.7%, and that there is no metastasis of 0.907 or 90.7%. The probability of survival of breast cancer patients based on variable types of radiotherapy treatment is 0.792 or 79.2%, and the type of chemotherapy treatment is 0.724 or 72.4%. Therefore, the variables that significantly affect breast cancer patients' survival are the initial age of infection, the stage of the clinic, and the size of the tumor.

In this study, patient data were analyzed in the two-year study. For further research, we recommend using data with objects observed over a more extended period, for example, for five years. The survival analysis results get a more objective picture. Also, observed variables may be more detailed types of cancer, tumor location, etc.

REFERENCES

Arshi, A., Sharifi, F. S., Ghahfarokhi, M. K., Faghih, Z., Doosti, A., Ostovari, S., … Seno, M. M. G. (2018). Expression Analysis of MALAT1 , GAS5 , SRA , and NEAT1 lncRNAs in Breast Cancer Tissues from Young Women and Women over 45 Years of Age. *Molecular Therapy - Nucleic Acids, 12*, 751–757. https://doi.org/10.1016/j.omtn.2018.07.014

Devarajan, K., & Ebrahimi, N. (2011). A semi-parametric generalization of the Cox proportional hazards regression model : Inference and applications. *Computational Statistics and Data Analysis*, *55*(1), 667–676. https://doi.org/10.1016/j.csda.2010.06.010

Dewi, G. A. T., & Hendrati, L. Y. (2015). Analisis risiko kanker payudara berdasar riwayat pemakaian kontrasepsi hormonal dan usia menarche. *Jurnal Berkala Epidemiologi*, *3*(1), 12–23.

Dwidayati, N. (2016). Asumsi Propotional Hazard (PH) Cox dalam Analisis Cure Rate Penderita Kanker Payudara. *2016: Prosiding Seminar Nasional Matematika IX 2015*, 403–417.

Ebrahimi, V., Khademian, M. H., Masoumi, S. J., Morvaridi, M. R., & Ezzatzadegan Jahromi, S. (2019). Factors influencing survival time of hemodialysis patients; Time to event analysis using parametric models: A cohort study. *BMC Nephrology*, *20*(1), 1–9. https://doi.org/10.1186/s12882-019-1382-2

Febriani, A., & Furqon, A. (2018). Metastasis Kanker Paru. *Jurnal Respirasi*, *4*(3), 94–101.

Ghodsi, Z., Salehi, A., & Hojjatoleslami, S. (2013). Knowledge of Iranian Women about Warning Signs and Risk Factors for Breast Cancer. *Procedia - Social and Behavioral Sciences*, *93*, 343–348. https://doi.org/10.1016/j.sbspro.2013.09.201

Han, S., Choi, S., Nah, S., Lee, S. U., Cho, Y. S., Kim, G. W., & Lee, Y. H. (2021). Cox regression model of prognostic factors for delayed neuropsychiatric sequelae in patients with acute carbon monoxide poisoning: A prospective observational study. *NeuroToxicology*, *82*(July 2020), 63–68. https://doi.org/10.1016/j.neuro.2020.11.006

Ihwah, A. (2015). The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product. *Agriculture and Agricultural Science Procedia*, *3*, 78–83. https://doi.org/10.1016/j.aaspro.2015.01.017

Jawitz, O. K., Fudim, M., Raman, V., Bryner, B. S., DeVore, A. D., Mentz, R. J., … Rogers, J. G. (2020). Reassessing Recipient Mortality Under the New Heart Allocation System: An Updated UNOS Registry Analysis. *JACC: Heart Failure*, *8*(7), 548–556. https://doi.org/10.1016/j.jchf.2020.03.010

Jobsen, J. J., Middelburg, J. G., Palen, J. Van Der, Riemersma, S., Siemerink, E., Struikmans, H., & Siesling, S. (2019). Breast-conserving therapy in older patients with breast cancer over three decades : progress or stagnation. *Journal of Geriatric Oncology*, *10*(2), 330–336. https://doi.org/10.1016/j.jgo.2018.08.007

Julia, K. (2012). Survival analysis. *Research & Statistics*, *33*(172). https://doi.org/10.1542/pir.33-4-172

Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology*, *22*(7), 263–270. https://doi.org/10.1016/j.mpdhp.2016.06.005

Kleibl, Z., & Kristensen, V. N. (2016). Women at high risk of breast cancer : Molecular characteristics , clinical presentation and management. *The Breast*, *28*, 136–144. https://doi.org/10.1016/j.breast.2016.05.006

Kleinbaum, D. G., & Klein, M. (2012). Kaplan-Meier Survival Curves and the Log-Rank Test. In *Statistics for Biology and Health*. https://doi.org/10.1007/978-1-4419-6646-9

Lee, J. G., Moon, S., & Salamatian, K. (2012). Modeling and predicting the popularity of online contents with Cox proportional hazard regression model. *Neurocomputing*, *76*(1), 134–145. https://doi.org/10.1016/j.neucom.2011.04.040

Nurfain, & Purnami, S. W. (2017). Analisis Regresi Cox Extended pada Pasien Kusta di Kecamatan Brondong Kabupaten Lamongan. *Jurnal Sains Dan Seni ITS*, *6*(1), 94–100.

Reyna, C., & Lee, M. C. (2014). Breast cancer in young women : special considerations in multidisciplinary care. *Journal of Multidisciplinary Healthcare*, *7*, 419–429.

RI, D. P. K. kesehatan. (2019). Penyakit Kanker di Indonesia Berada Pada Urutan 8 di Asia Tenggara dan Urutan 23 di Asia. Retrieved December 1, 2020, from http://p2p.kemkes.go.id/penyakit-kanker-di-indonesia-berada-pada-urutan-8-di-asia-tenggara-dan-urutan-23-di-asia/

RI, K. kesehatan. (2020). Jenis Kanker ini Rentan Menyerang Manusia. Retrieved December 8, 2020, from https://www.kemkes.go.id/article/view/20011400002/jenis-kanker-ini-rentan-menyerang-manusia.html

Selvaraj, S., Ilkhanoff, L., Burke, M. A., Freed, B. H., Lang, R. M., Martinez, E. E., & Shah, S. J. (2014). Association of the frontal QRS-T angle with adverse cardiac remodeling, impaired left and right ventricular function, and worse outcomes in heart failure with preserved ejection fraction. *Journal of the American Society of Echocardiography*, *27*(1), 74-82.e2. https://doi.org/10.1016/j.echo.2013.08.023

Stel, V. S., Dekker, F. W., Tripepi, G., Zoccali, C., & Jager, K. J. (2011). Survival analysis II: Cox regression. *Nephron - Clinical Practice*, *119*(3), 255–260. https://doi.org/10.1159/000328916

Suganda, A. R., Wiratmoko, W., Marhayuni, E., & Yuniastini. (2021). Sulvival life penderita kanker payudara pada wanita berdasarkan grading & kemoterapi di RSUD Dr. H. Abdul Moeloek Provinsi Lampung. *Jurnal Medika Malahayati*, *5*(2), 77–82. Retrieved from http://www.k12.wa.us/HealthFitness/Standards/HealthEducationK-12LearningStandards.pdf

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., … Zhu, H. P. (2017). Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*, *13*(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., & Schrag, D. (2015). Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. *Journal of Clinical Oncology*, *32*(22), 2380–2385. https://doi.org/10.1200/JCO.2014.55.2208

Wijaya, A. A., & Wulandari, S. P. (2015). Analisis Survival pada Pasien Penderita Sindrom Koroner Akut di RSUD Dr.Soetomo Surabaya Tahun 2013 Menggunakan Regresi Cox Proportional Hazard. *Jurnal Sains Dan Seni ITS*, *4*(2), 151–156.

Yadav, S. K., Silwal, S., Yadev, S., Kishnamoorthy, G., & Chisti, M. M. (2021). A systematic comparison of overall survival between men and women with triple negative breast cancer. *Clinical Breast Cancer*. https://doi.org/10.1016/j.clbc.2021.07.001

Yao, N., Li, W., Liu, T., Tan, S., Chen, X., Wang, W., … Qu, J. (2020). Primary tumor removal improves the prognosis in patients with stage IV breast cancer : A population-based study (cohort study). *International Journal of Surgery*, *83*, 109–114. https://doi.org/10.1016/j.ijsu.2020.08.056

Yi, Y., & Wang, X. (2011). Comparison of Wald , Score , and Likelihood Ratio Tests for Response Adaptive Designs. *Journal of Statistical Theory and Applications*, *10*(4), 553–569.

Yulianto, D., Notobroto, H. B., & Widodo. (2017). ANALISIS KETAHANAN HIDUP PASIEN PENYAKIT GINJAL KRONIS DENGAN HEMODIALISIS DI RSUD Dr. SOETOMO SURABAYA. *Jurnal Manajemen Kesehatan*, *3*(1), 99–112.

Zare, A., Mahmoodi, M., Mohammad, K., Zeraati, H., Hosseini, M., & Naieni, K. H. (2014). A Comparison between Kaplan-Meier and Weighted Kaplan-Meier Methods of Five-Year Survival Estimation of Patients with Gastric Cancer. *Acta Medica Iranica*, *52*(10), 764–767.

Zhou, B., Fine, J., & Laird, G. (2013). Goodness-of-fit test for proportional subdistribution hazards model. *Statistics in Medicine*, *32*(22), 3804–3811. https://doi.org/10.1002/sim.5815