Survival Analysis in Breast Cancer Patients: Cox Proportional Hazard Regression Model to Evaluate Risk Factors

Gede Nayaka Baswara[1], Bramastya Z H[1], Nadiya Mujahidatul Farhani[1], Krisna Bayu C.D.[1], Bunga Citra S.L[1], Ratih Ardiati Ningrum[1]

[1] Data Science Technology, Airlangga University, Surabaya, Indonesia
ratih.an@ftmm.unair.ac.id

Abstract. This study aims to analyze the factors that influence the survival of breast cancer patients using the Cox Proportional Hazard model. The data used comes from the NKI breast cancer dataset which includes information about patients, treatment and survival. Analysis was carried out using the Cox Stratified and Cox Extended models, by testing the Proportional Hazard assumption and selecting the best model based on the Akaike Information Criterion (AIC). The results of the study showed that several variables such as age, type of treatment, histology type, tumor diameter, and number of positive nodes had a significant influence on the risk level in breast cancer sufferers. These findings provide valuable insight into identifying key factors influencing survival, thereby helping to improve treatment strategies and management of breast cancer patients.

Keywords: breast cancer, survival analysis, treatment strategy.

1 Background

Cancer is a disease where cells in the body grow abnormally and uncontrollably and suppress normal cells. Cancer cells appear when genetic mutations caused by DNA damage in normal cells have occurred. In the case of breast cancer, this abnormal cell growth occurs in the cells in the breast. This uncontrolled cell growth can cause a lump to appear on the body which is then called a tumor. Tumors are divided into two, namely benign tumors and malignant tumors or cancer. Breast cancer is a non-communicable disease that can occur in both women and men, but the possibility of experiencing breast cancer in women is up to 100 times greater than in men. [1].

Breast cancer is one of the most common malignant diseases experienced by women. This disease is the second most common cause of cancer deaths in women after cervical cancer. WHO stated that in 2020, as many as 2.3 million women in the world were diagnosed with breast cancer and 685,000 women died from the disease. Based on Riskesdas data in 2019, breast cancer cases in Indonesia reached a prevalence of 42.1 per 100,000 population with an average death rate of 17 per 100,000 population. Breast cancer is the biggest cancer that causes death in women with statistics showing the death rate reaches 14% per year [2].

T. Amrillah et al. (eds.), *Proceedings of the International Conference on Advanced Technology and Multidiscipline (ICATAM 2024)*, Advances in Engineering Research 245,
https://doi.org/10.2991/978-94-6463-566-9_24

The high incidence of breast cancer is caused by sufferers who are late in realizing the symptoms of the disease. If cancer is detected at a higher stage, treatment will be more expensive but the results will not be optimal and can even hasten death. Success in breast cancer treatment is also influenced by several factors such as the patient's age, chemotherapy, hormonal therapy, amputation, type of cancer history, tumor diameter, number of positive lymph nodes, level of cancer differentiation, angiolymphatic invasion, and lymphocyte infiltration. The main indicator for assessing the success of breast cancer treatment is the patient survival rate. One method that can be used to explain this is to use the Cox Proportional Hazard model [3].

The Cox Proportional Hazard model has the ability to display unadjusted and adjusted hazard ratio (HR) values with appropriate confidence intervals (CI). In addition, unlike other regressions, Cox Proportional Hazard is not bound by assumptions regarding the nature and shape of the distribution. Therefore, the Cox Proportional Hazard model was used to analyze the survival of breast cancer patients by evaluating risk factors. This study uses the Cox Stratified and the Cox Extended model, then implement the Akaike Information Criterion to evaluate and the best model. It is hoped that the results of this research will help the probability of survival of breast cancer sufferers and the risk factors that influence their survival, so that they can evaluate the effectiveness of the treatment given.

2 Literature Review

2.1 Survival Analysis

Survival Analysis is a statistical method used to analyze data which the variable that is considered is the time until an event occurs. Time can be expressed in: months, weeks, or days from when observations are made on an individual until an event occurs to that individual [4]. Survival analysis or survival analysis is a statistical method related to the time of an object starting from the time origin or start point until the occurrence of a certain predetermined event (failure event or end point). In this case, the event in question is death, contracting a disease, recurrence of an illness, recovery and other events that can happen to a person. In general, the purpose of survival analysis is to estimate and interpret the survival function and/or hazard function from survival data, compare the survival function and/or hazard function, determine the influence of predictor variables on survival time [5].

2.2 Proportional Hazard Assumption

The Proportional Hazard assumption is said to be fulfilled if a line on the survival curve (between groups) does not intersect each other. The Proportional Hazard assumption is very important in survival analysis. Survival analysis that meets the Proportional Hazard assumption will be analyzed using independent time analysis, while survival that does not meet the Proportional Hazard assumption will be analyzed using full model analysis or reduced model analysis. One approach that is often used to check the Proportional Hazard assumption is the graphical approach. The graphic approach is a

method used to check the Proportional Hazard assumption by looking at the survival line on the Kaplan–Meier curve. The assumption is met if the survival lines do not intersect each other [4].

2.3 Cox Proportional Hazard Model

One of the objectives of survival analysis is to determine the relationship between time to failure and covariates measured at the time of the research. The unknown distribution of the data means that parametric methods cannot be used to analyze the data, so the semiparametric survival analysis method is the appropriate method. The semiparametric regression model that is often used is the Cox Proportional Hazard (PH) regression model. Suppose there are p predictor variables (covariates) $X = (x_1, x_2, \dots, x_p)$ and the covariate values are expressed in vector form where $= (x_1, x_2, \dots, x_p)$ then the general Cox PH regression model can be seen in (1).

$$h(t,x) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \quad (1)$$

Below are the explanation of each notations:

$h_0(t)$: Failure function
$\beta_1, \beta_2, \dots, \beta_p$: Regression parameters
$x_1, x_2, \dots, x_p$: Value of the independent variable

The Cox Proportional Hazard model can provide useful information in the form of the Hazard Ratio (HR), which does not depend on HR, which is the ratio of the hazard level of one individual to the hazard level of another individual [5].

2.4 Hazard Ratio

a. Hazard ratio continuous data

The hazard ratio for continuous data can be seen by the formula (2).

$$\widehat{HR}(c) = e^{c\beta} \quad (2)$$

The interpretation of the HR estimator value in the equation is that the risk of a late death variable will increase by each additional unit of change.

b. Hazard ratio categorical data.

The hazard ratio for categorical data is as follows.

$$\widehat{HR} = \frac{h(t|x=1|)}{h(t|x=0|)} = \frac{h(t)e^{\widehat{\beta}}}{h(t)} = e^{\widehat{\beta}} \quad (3)$$

The measure used to determine the risk of dying later can be determined by the hazard ratio value. The hazard ratio is a comparison between individuals with the condition of the independent variable x in the success category and the failure category. Then the hazard ratios for individuals with $x = 1$ are compared $x = 0$ [4].

2.5 Stratified Cox Model

The Stratified Cox model is an extension of the Cox PH model to deal with independent variables that do not meet the Proportional Hazard assumption. Modifications are carried out by stratifying independent variables that do not meet the Proportional Hazard assumption. According to Kleinbaum & Klein (2012), the general form of the hazard function of the Stratified Cox model is as follows [4]

$$h_s(t, X) = h_{0s}(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k] \quad (4)$$

Where:

s : Strata defined from $Z^*, s = 1,2,..,m$.

$h_{0s}(t)$: Basic failure function for each stratum.

$\beta_1, \beta_2, \ldots, \beta_k$: Regression parameters.

This stratified Cox model without interaction is a general form of the stratified Cox model which shows that there is no interaction between the independent variables.

2.6 Extended Cox Model

In the Cox model there is a variable involving time t. This variable is called a time dependent variable. A time-dependent variable is defined as a variable that has a changing value over time (t). If there are time-dependent variables in the model, the extended Cox model can be used. In this model, the Cox model is expanded with a model containing time-dependent covariates. If $x_1, x_2, \ldots, x_p$ is a time-independent covariate that meets the proportional hazard assumption $x_{p1+1}, x_{p2+}, \ldots, x_p$ is a time−independent covariate that does not meet the proportional hazard assumption and $x_1(t), x_2(t), \ldots, x_p(t)$ is a time-dependent covariate then the Extended Cox model is defined as follows.

$$h\big(t, X(t)\big) = h_0(t) \exp\left[\sum_{a=1}^{p_1} \beta_a X_a + \sum_{b=1}^{p_2} \delta_b X_b g_b(t)\right] \quad (5)$$

Where β and δ are vector coefficients of covariates, p_1 is the number of covariates that meet the PH assumption and p_2 is the number of covariates that do not meet the PH assumption. Some of the time functions used include: $g_1(t) = t$, $g_2(t) = log$, and $g_3(t)$ is the Heaviside function [8].

2.7 Akaike Information Criterion (AIC)

Akaike Information Criterion (AIC) is a value that measures the relative quality of a model for selecting the best model from several models. AIC combines the model's ability to fit data with the complexity of the model, with the aim of selecting a model that is good at making predictions on data that has never been seen before. The AIC formula is defined as follows.

$$AIC = -2ln\big(\hat{L}\big) + 2k \quad (6)$$

Where $\hat{L}$ is the maximum likelihood value of model and k is the number of parameters to be estimated. Akaike explains that choosing a model with a low level of information loss is asymptotically the same as choosing the model that has the smallest AIC value. This means that the model that has the smallest AIC value is the better model [7].

2.8 Breast Cancer

Breast cancer (carcinoma mammae) is a malignancy that originates from breast tissue, both from the ductal epithelium and the lobules. Mammary carcinoma occurs due to the condition of cells that have lost their normal control and mechanisms, so they experience abnormal, fast and uncontrolled growth. Breast cancer is the most diagnosed cancer in women, accounting for more than 1 in 10 new cancer diagnoses each year. It is the second most common cause of death from cancer among women worldwide. Breast

cancer develops silently, and most of the disease is discovered during routine examinations [6].

3 Methodology

The data used in this research comes from the NKI Breast Cancer dataset. This data presents information regarding breast cancer cases in various states in the United States which includes information regarding patients, treatment, and survival. This data has 272 breast cancer patients (as rows) with 1570 columns. This research focuses on 12 specific variables. The following is an explanation of the variables used for research.

Table 1. Variables

Variable	Information
eventdeath	Indicates whether an event occurred or not, coded as 0 for no death and 1 for death.
timerecurrence	Shows the time period before the cancer returns after initial treatment.
age	Cancer patient's age.
chemo	Chemotherapy use was coded as 0 for no and 1 for otherwise.
hormonal	Use of hormonal therapy, coded as 0 for no and 1 for otherwise.
amputation	Indicates whether amputation was performed or not, coded as 0 for no and 1 otherwise.
histtype	Type of cancer history.
diam	Tumor diameter.
posnodes	The number of positive lymph nodes indicates the presence of cancer cells.
grade	Cancer differentiation rate
angionv	Angiolymphatic invasion was coded as 0 for no and 1 for the presence of cancer that had invaded the blood vessels or lymphatics.
lymphinfil	Lymphocyte infiltration reflects the presence of lymphocytes in tumor tissue.

By focusing on these variables, this study aims to analyze the factors that influence survival in breast cancer cases. This analysis was carried out using the Python programming language. The analysis steps in this research include:

a. Data preprocessing.
b. Descriptive statistics.
c. Exploratory data analysis.
d. Proportional Hazard Assumption.
e. Stratified Cox Model.
f. Extended Cox Model
g. Looking for the best model using the Akaike Information Criterion (AIC).

4 Results and Discussion

4.1 Data Prepocessing

In this study, the data has no duplicate data and no missing values. Certain variables are also of the categorical data type so they do not need to be encoded.

4.2 Descriptive Statistics

Table (2) is descriptive statistics for each numerical variable in this research.

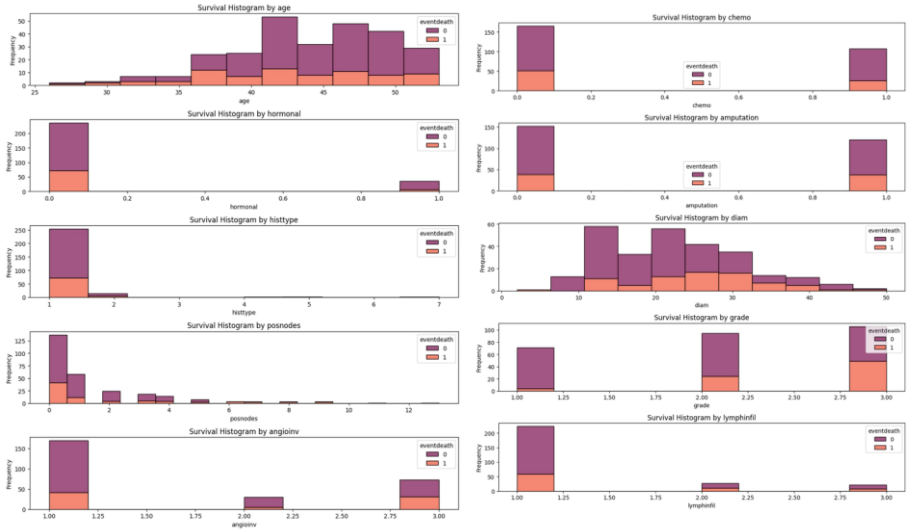
Table 2. Descriptive Statistics

	age	diam	posnodes	time
Count	272	272	272	272
Mean	44.04	22.53	1.34	7.25
Std	5.46	8.7	2.11	4.17
Min	26	2	0	0.27
25%	40.75	15	0	4.39
50%	45	20	0	6.95
75%	49	29.25	2	9.98
Max	53	50	13	18.34

In the variable “age” which indicates the age of cancer sufferers, it can be seen that the average age of sufferers is 44.04 years old, with a distribution between 26 and 53 years old. The relatively low standard deviation, namely 5.46, indicates that the age data tends to cluster around the mean value. The “diam” variable is the tumor diameter. It can be seen that the average tumor diameter is 22.53 with a standard deviation of 8.7. Tumor diameter varies between 2 to 50, and most of the data is in the first to third quartile, namely 15 to 29.25. High variability in tumor diameter is indicated by a fairly large standard deviation. In the “posnodes” variable most patients had 0 positive lymph nodes, but the maximum value reached 13, indicating significant variation in extreme cases. There is significant variation in the “time” period before the cancer returns, with most cases in the range of 0.27 to 18.34 years.

A. Exploratory Data Analysis (EDA*)*

A.1. Survival Histogram Based on Variables



(a) (b)

Fig. 1. Histogram: (a) and (b) Histogram of each variable.

In the histogram above, it can be seen that the diagram representing the no death status (0), consistently outperforms the diagram for the death status (1) in all variables. This illustrates a clear difference in the distribution of variables between patients who did not survive and those who did. The histogram shows a higher concentration of survivors within a certain range, indicating a potential relationship between the levels of these variables and an increased probability of survival. The consistent elevation of disalt for cases without death (0) across variables implies that these factors may have predictive power in determining the probability of survival. This is in line with the notion that certain characteristics or treatments may contribute to a higher chance of survival.

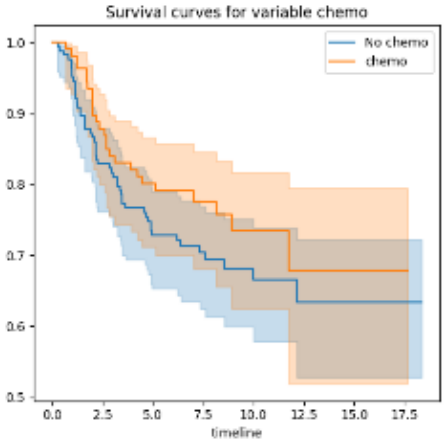
A.2. Heat Map

The heat map between variables shows that the majority of independent variables have a very low or insignificant correlation with each other, except for the relationship between the variables “posnodes” (number of positive lymph nodes indicating the presence of cancer cells) and “chemo” (chemotherapy treatment) which shows a moderate positive correlation of 0.53. Apart from that, there is also a moderate positive correlation of 0.43 between variant “grade” (level of cancer differentiation) and “lymphinfil” (lymphocyte infiltration which reflects the presence of lymphocytes in tumor tissue).

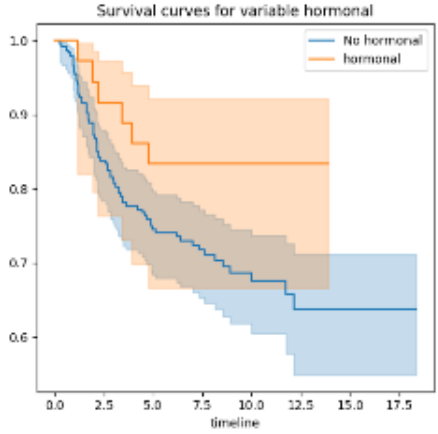
Fig. 2. Heat map of each research variable

A.3. Kaplan-Meier Survival Curves for Chemotherapy, Hormonal, & Amputation Treatment

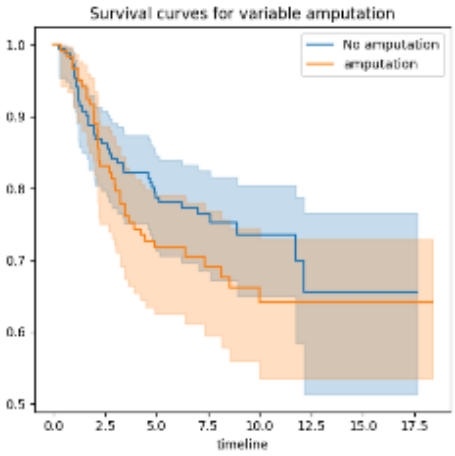
Based on the Kaplan-Meier curve for chemotherapy treatment factors (a) and hormonal therapy (b), it can be concluded that there is no overlap between the two category lines. Therefore, it can be interpreted that there is a significant difference in the survival curves between the two categories. Shows that the group of patients who received chemotherapy & hormonal therapy tended to have a higher survival rate. Meanwhile, for amputation treatment (c), it can be observed that there is an overlap between the two category lines, indicating that there is variation in the survival curve between the two categories.



(a)



(b)



(c)

Fig. 3. Kaplan-Meier: (a) chemo variables, (b) harmonic variables, and (c) amputation variables

B. Proportional Hazard Assumption

Table (3) is the result of assumption testing using the lifeline library. Based on table (3), it can be seen that most of the p-values for the variables tested are greater than alpha > 0.05, except for the “grade” variable which has a p-value of 0.04. Therefore, it can be concluded that it failed to reject H_0 for all variables except grade, which means there is not sufficient evidence to state that there is a significant correlation between these variables and survival time. In other words, variables such as "age", "amputation", "angioinv", etc., do not show sufficient evidence to violate the proportional hazard assumption.

Table 3. Proportional Hazard Assumption Test Results

Feature	Test Statistic (with Chi Sqared)	p-value	-log2(p)
age	3.08	0.08	3.66

Feature	Test Statistic (with Chi Sqared)	p-value	-log2(p)
amputation	0.48	0.49	1.04
angioinv	1.95	0.16	2.62
chemo	0.12	0.73	0.45
diam	0.11	0.74	0.44
grade	4.28	0.04	4.70
histtype	2.72	0.10	3.33
hormonal	0.20	0.65	0.61
lymphinfil	0.01	0.93	0.11
posnodes	0.59	0.44	1.18

Therefore, it can be concluded that except for the "grade" variable, all other variables meet the proportional hazard assumption. For all variables except the “grade” variable, the Cox proportional hazards regression model can be continued without modification based on the assumptions that have been tested.

C. Cox Model

C.1 Cox Stratified Model

Based on table (4), the partial p-values for the variables "age", "chemo", and "histtype" are all smaller than alpha 0.05. This resulted in the decision to reject H_0 which partially means that the variables have a significant influence on the patient's survival time.

Table 4. Parameter Significance Test Results Using the Stratified Cox Model

covariate	coef	exp(coef)	p
age	-0.053646	0.947768	0.013472
chemo	-0.551889	0.575861	0.070890
hormonal	-0.277172	0.757924	0.538575
amputation	-0.008961	0.991079	0.971430
histtype	0.536118	1.709359	0.010147
diam	0.021437	1.021669	0.550379
posnodes	0.081206	1.084595	0.147941
angioinv	0.212153	1.236336	0.120024
lymphinfil	-0.289397	0.748715	0.152684

From table (4) a Stratified Cox regression model can be created as in the calculation (4).

$$h(t) = h0(t)\exp\left[-0.053646\ \text{age} - 0.551889\ \text{chemo} - 0.277172\ \text{hormonal} - 0.008961\ \text{amputation} + 0.536118\ \text{histtype} + 0.021437\ \text{diam} + 0.081206\ \text{posnodes} + 0.212153\ \text{angioinv} + -0.289397\text{lymphinfil}\right.$$

C.2 Cox Extended Model

Based on table (5), the partial p-values for the variables "angioinv" and "grade_timerecurrence" are 7.17753×10^{-22}, and 8.674173×10^{-19} respectively, which are all smaller than alpha 0.05. This resulted in the decision to reject H_0 which partially means that the variables "angioinv" and "grade_timerecurrence" have a significant influence on the patient's survival time. The variable "lymphinfil" with a p-value of 0.075456 shows a tendency to influence survival time, but is not significant enough at the alpha level of 0.05.

Table 5. Parameter Significance Test Results Using the Extended Cox Model

covariate	coef	exp(coef)	p
age	-0.008661	0.991377	0.7037922
chemo	-0.177296	0.837532	0.6148958
hormonal	-0.530672	0.588210	0.2561964
amputation	-0.025125	0.975188	0.9227557
histtype	0.063926	1.066013	0.7606734
diam	-0.008436	0.991599	0.5503794
posnodes	0.026843	1.027206	0.6816867
grade	4.676774	107.422951	0.4089278
angioinv	0.114671	1.121504	7.177553×10^{-22}
lymphinfil	-0.361046	0.696947	0.075456
grade_timerecurrence	-0.778585	0.459055	8.674173×10^{-19}

From table (5) a Cox Extended Hazard regression model can be created as in the calculation (5).

$$h(t) = h0(t)\exp[-0.008661\ \text{age} - 0.177296\ \text{chemo} - 0.530672\ \text{hormonal} - 0.025125\ \text{amputation} + 0.063926\ \text{histtype} - 0.008436\ \text{diam} + 0.026843\ \text{posnodes} + 4.676774\ \text{grade} + 0.114671\ angioinv - 0.361046\ lymphinfil - 0.778585\ grade\_timerecurrence$$

D. Interpretation of Hazard Ratio Values

Based on table (5), the following hazard values for Cox Extended model are obtained.

a. **Age:** The coefficient -0.008661 shows that with each year of increasing age, the hazard or risk level decreases by 0.991377 times.
b. **Receipt of chemotherapy (chemo):** The coefficient -0.177296 indicates that receipt of chemotherapy reduces the hazard by 0.837532 times.
c. **Hormonal therapy (hormonal):** The coefficient -0.530672 indicates that hormonal therapy reduces the hazard by 0.588210 times.
d. **Amputation:** The coefficient -0.025125 shows that amputation reduces the hazard by 0.975188 times.

e. **Histology type (histtype):** The coefficient 0.063926 indicates that with a certain histology type, the hazard increases by 1.066013 times.
f. **Tumor diameter (diam):** The coefficient -0.008436 indicates that with each increase in tumor diameter, the hazard decreases by 0.991599 times.
g. **Number of positive nodes (posnodes):** The coefficient 0.026843 indicates that with each increase in the number of positive nodes, the hazard increases 1.027206 times.
h. **Grade:** The coefficient 4.676774 shows that with increasing grade, the hazard increases by 107.422951 times.
i. **Angiolymphatic invasion (angioinv):** The coefficient of 0.114671 indicates that with angiolymphatic invasion, the hazard increases by 1.121504 times.
j. **Lymphocyte infiltration (lymphinfil):** The coefficient -0.361046 indicates that with the presence of lymphocyte infiltration, the hazard decreases by 0.696947 times.
k. **Grade at the time of occurrence or recurrence (grade_g(t))**: The coefficient -0.778585 indicates that at the time of occurrence or recurrence, the hazard decreases by 0.459055 times.

E. Akaike Information Criterion (AIC)

From table (6), it can be seen that the results from AIC which have the lowest value are Cox Extended. In this context, a lower AIC value is considered better.

Table 6. Akaike Information Criterion Results

Model	Nilai AIC
Stratified Cox	652.2163
Extended Cox	453.0128

So from table (6), it can be concluded that the Cox Extended model has better quality or is more efficient than the Cox Stratified model. This means that the Extended Cox model provides a better balance between fitting ability to the data and model complexity compared to the Stratified Cox model.

5 Conclusion

This study investigates factors that influence the survival of breast cancer patients using the Cox Proportional Hazard model. The results of the analysis show that several variables such as age, receipt of chemotherapy, hormonal therapy, amputation, histological type, tumor diameter, number of positive nodes, grade, angiolymphatic invasion, lymphocyte infiltration, and grade at the time of occurrence or recurrence have a significant influence on the level of risk or hazard. in breast cancer sufferers. For example, receipt of chemotherapy and hormonal therapy carries a reduced risk, whereas increasing the grade and number of positive nodes increases the risk significantly. These findings provide important insights for evaluating the effectiveness of treatment and management of breast cancer patients in improving survival rates.

Disclosure of Interests. The authors declares that this paper has no competing interests.

References

1. E. Widyawati, A. Fadli, and M. S. Aliim.: Purwarupa Sistem Pakar Dengan Metode Forward Chaining Dan Certainty Factor Untuk Mendeteksi Penyakit Kanker Payudara. *Jurnal Pendidikan dan Teknologi Indonesia* **1**(6), 247–259 (2021)
2. A. N. Azmi, B. Kurniawan, A. Siswandi, and A. U. Detty, *Hubungan Faktor Keturunan Dengan Kanker Payudara DI RSUD Abdoel Moeloek*. in Statistics for Biology and Health. New York, NY: Springer New York, 2020. doi: 10.35816/jiskh.v10i2.373.
3. Y. Farida, E. A. Maulida, L. N. Desinaini, W. D. Utami, and D. Yuliati, "Breast Cancer Survival Analysis Using Cox Proportional Hazard Regression and Kaplan Meier Method," vol. 5, no. 2, pp. 340–358, 2021, doi: 10.31764/jtam.v5i2.4653.
4. D. G. Kleinbaum and M. Klein, *Survival Analysis*. New York, NY: Springer New York, 2012. doi: 10.1007/978-1-4419-6646-9.
5. S. W. Purnami and I. N. Pertiwi, "Regresi Cox Proportional Hazard Untuk Analisis Survival Pasien Kanker Otak di C-Tech Labs Edwar Technology Tangerang," *Inferensi*, vol. 3, no. 2, p. 65, Sep. 2020, doi: 10.12962/j27213862.v3i2.7727.
6. A. Rizka, M. Khalilul Akbar, and N. A. Putri, "CARCINOMA MAMMAE SINISTRA T4bN2M1 METASTASIS PLEURA," 2022.
7. I. A. Trivanni, "PERBANDINGAN MODEL REGRESI LOGISTIK ORDINAL," 2018.
8. M. Puji Paramitha, N. Satyahadewi, and N. Nessyana Debataraja INTISARI, "MODEL EXTENDED COX UNTUK MENGATASI NONPROPORTIONAL HAZARD," 2022.