Survival Analysis and Critical Risk Factors in Covid-19 Patients Using Cox Regression

Jerry Dwi Trijoyo Purnomo[1*] and *Alissa* Novitasari[1]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Abstract. Coronavirus Disease 2019 (Covid-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The average incubation period of SARS-CoV-2 is 5 days (ranging from 2 to 14 days), and people experiencing symptoms occur within 12 days after infection (ranging from 8 to 16 days). Most person-to-person transmission of the virus can occur before the infected person shows symptoms (presymptomatic). A small percentage of infected people never experience symptoms but can contribute substantially to the transmission of the disease. Research continues to be carried out to determine the estimated length of recovery of Covid-19 patients. In this study, survival analysis of Covid-19 patients will be studied at Haji Hospital in Surabaya using Cox Proportional Hazard regression. Cox regression is one of the methods that can measure the relationship between Hazard rate and predictor variables without any assumptions as found in parametric models, therefore the Cox regression model is included as a semi-parametric model. This model allows the test to look at the survival time differences of two or more interest groups and can describe the effect of the predictor variables used to predict the status of the response variables to survival. The results of the Cox Proportional Hazard regression modeling showed two variables that influenced the survival time of Covid-19 patients, namely the gender variable and the symptom variable.

1 Introduction

SARS-CoV-2 was first discovered at the end of 2019 in Wuhan, China. The initial clinical symptoms of COVID-19 are mostly fever, dry cough, and fatigue. Some patients present with nasal congestion, runny nose, pharyngeal pain, myalgia, and diarrhea. Some patients may even have no symptoms. Previous studies found that the most common comorbidities were hypertension (17%), diabetes mellitus (8%), and cardiovascular disease (5%) in COVID-19 [1]. Two years since the start of the SARS-CoV-2 pandemic that caused more than 5 million deaths, the world continues to be on high COVID-19 alert. The World Health Organization (WHO), in collaboration with national authorities, public health institutions, and scientists has been closely monitoring and assessing the evolution of SARS-CoV-2 since January 2020. The emergence of certain SARS-CoV-2 variants is characterized as a Variant of Interest (VOI) and Variants of Concern (VOC), to prioritize global monitoring and research, and to inform the ongoing global response to the COVID-19 pandemic. WHO and its international sequencing

* Corresponding author : jerrypurnomo@gmail.com

network continue to monitor SARS-CoV-2 mutations and notify countries of any changes that may be necessary to respond to the variant and prevent its spread where possible [2].

The first SARS-CoV-2 virus was B.1.1.7 which is called the Alpha variant and was first detected in the UK. Variant of Concern (VOC) Alfa, or the variant phylogenetically designated as B.1.1.7 is attracting global attention, quickly displacing other variants due to its 50-70% greater transmissibility and also displayed by approximately 30% greater risk of death than others (Murayama et al., 2021). B.1.351 is the Beta variant first detected in South Africa. P.1 is the Gamma variant first detected in Brazilian travelers. The B.1.617.2 (Delta) variant was first detected in India in December 2020 and became the most frequently reported variant in the country starting in mid-April 2021. India has experienced a spike in COVID-19 cases since late March 2021, reaching more than 400,000 cases. and 4000 deaths were reported every day in early May 2021. This increase has resulted in hospital services being overwhelmed and oxygen supplies being scarce. Although only a small proportion of samples have been observed, the B.1.617 lineage of SARS-CoV-2 has dominated [2].

The next variant is Omicron which can spread much more easily from person to person. The transmission of the Omicron variant is quite worrying as the rate of spread throughout the world has increased significantly in recent days. Omicron's infection rate is twice as high in the Delta variant. However, the cell fusion caused by Omricon is not as severe as the Delta variant due to poor inter-cell transmission so it can be said that the Omicron variant has a low replication rate. Patients infected with the Omicron variant do not show severe symptoms. Thus, it is considered that this variant is mild compared to other variants. Researchers in England, Scotland, and South Africa found that the Omicron variant had a 15%-80% lower risk of hospitalization than the Delta variant. Despite the much larger number of cases, surveillance data shows that the latest wave of disease driven by Omicron had far fewer hospital admissions and deaths than previous waves [3].

The COVID-19 pandemic is still ongoing and the number of countries are even experiencing an increase in daily cases, including Indonesia. The number of cases in Indonesia continues to increase rapidly, until September 2022 there were 6,425,849 confirmed cases and 158,057 deaths. The government continues to monitor and anticipate the spread of COVID-19 subvariants, both BA.4 and BA.5, even though the pandemic situation in Indonesia is still at level 1 of WHO standards. The highest increase in cases currently still occurs in Java-Bali, which represents almost 95% of cases. The positivity rate in Indonesia experienced a significant increase from 5.12% to 10.05%. The BA.4 and BA.5 subvariants do have the ability to penetrate or avoid vaccination (vaccination evasion). This subvariant is believed to be able to penetrate vaccination two to three times more effectively than the Omicron BA.1 variant so people's chances of being infected are higher even though they have been vaccinated.

The need for COVID-19 patients to be hospitalized varies greatly from country to country as it depends on the prevalence of testing and community admission criteria. However, it is estimated that one in 5-10 adult patients with the severity of the disease and the criteria are sufficient to be hospitalized. Most patients with severe acute respiratory infection or severe acute respiratory syndrome are managed according to the WHO case definition. Criteria for intensive care also vary from country to country. Older age, chronic disease, and male gender are consistently associated with increased mortality [4]. Hospitals throughout Indonesia have been asked to prepare contingency measures related to ongoing COVID-19 cases. The still high transmission rate is shown in data as of December 19 2021 where the use of beds in COVID-19 referral hospitals nationally is 2.73%. The occupancy rate per province is no more than 30% [5].

East Java itself is the province that has the highest number of COVID-19 cases after Jakarta. More than 180 positive COVID-19 patients came from the city of Surabaya. This makes the city of Surabaya the area most infected with the COVID-19 virus considering that

Surabaya is the provincial capital which is the center for gathering residents from other regions [6]. On March 17, 2020, the East Java government reported the first case of COVID-19. The latest case developments as of September 2022, Surabaya reported 131,725 confirmed COVID-19 patients. RSU Haji Surabaya is one of the referral centers for COVID-19 patient cases

2 Methods

In this section, we briefly illustrate the literature review and the proposed method used in this paper.

2.1 Survival Analysis

Survival analysis is a collection of statistical procedures for analyzing data where the variable is the time until an event occurs. Time can refer to a person's duration when an event occurs, while an event is an event that is observed. In survival analysis, the time variable represents survival time, because it represents the length of time someone has survived. Event variables are referred to as failures because the observed event is usually a death, illness incident, or other negative individual experience.

Most survival analyses must consider a major analytical problem called censoring. Censoring occurs when there is some information about an individual's survival time, but the exact survival time is not known. Right-censored is not knowing the data on a person's survival time at the end of the observation. This occurs when the observation ends and no event occurs or when the person cannot be observed again due to other factors. Left-censored is when a person's survival time is less than or equal to the survival time observed in the study. In other words, if an observation is left-censored at time t, it can be known that there was an event between times 0 and t, but the exact time of the start of the event is not known. The next censoring method is interval-censored which occurs if the subject's survival time is within a certain known time interval. Interval-censored combines right-censored and left-censored as a special case. Left-censored data occurs whenever the value of t_1 is 0 and t_2 is a known upper limit on the true survival time. In contrast, right-censored data occurs whenever the value of t2 is infinite, and t1 is a known lower limit on the true survival time [7]. The censoring survival data illustrated in Figure 1.

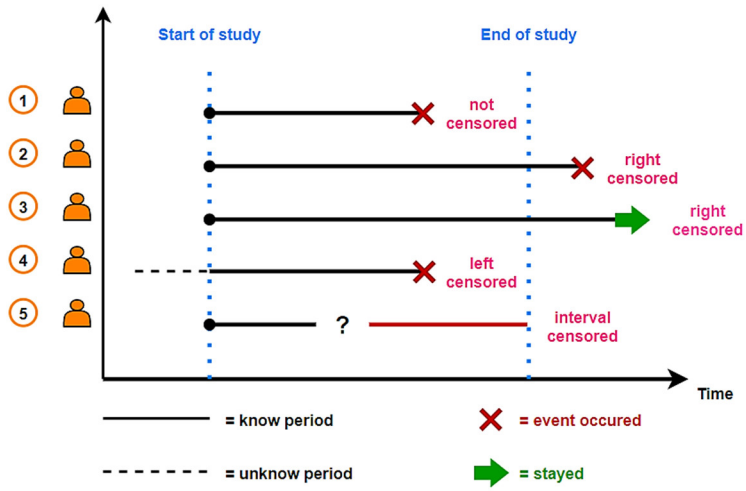

Fig. 1. Illustration of Censoring Survival Data

2.2 Survival Function

The survival function symbolized by S(t) is the probability of a random variable T surviving longer than a specified time t. the survival function is expressed in Equation 1.

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \le t) \\ &= 1 - F(t) \end{aligned} \tag{1}$$

where $F(t) = P(T \le t)$. If the probability $f(t)$ is a solid probability function that is integrated, then we obtain the survival function equation in Equation 2.

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du \tag{2}$$

The survival function is used to obtain the probability of survival time at the t value. In real cases, using actual data, a graph is obtained that is shaped like a step function. The limited research time means that not all subjects observed experience an event so the estimated survival function denoted by $S(t)$ is unlikely to fall close to zero at the end of the study. The survival function has several characteristics, namely the graph always has a decreasing pattern where the survival function will decrease when t increases.

2.3 Hazard Function

The hazard function is also called conditional failure rate, denoted by $h(t)$, the hazard function is written in Equation 3.

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t, T \ge t)}{P(T \ge t).\Delta t} \\ &= \frac{\lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t}}{1 - F(t)} \\ &= \frac{f(t)}{1 - F(t)} \end{aligned} \tag{3}$$

The hazard function $h(t)$ is the probability of an event occurring per unit time that the observed subject has survived until time t. For a given t value, the hazard function $h(t)$ has the characteristics of always being nonnegative and having no upper limit. Another characteristic is that it can be used to identify certain model shapes, such as exponential, Weibull, or lognormal curves. The $h(t)$ function is the cumulative hazard function obtained from the survival function. If $F(t) = 1 - S(t)$, then $f(t) = d\left(F(t)\right)/dt = d\left(1 - S(t)\right)/dt$.

2.4 Cox Proportional Hazard Regression

The Cox Proportional Hazard (PH) model is written using the formula in Equation 4.

$$h(t, \mathbf{X}) = h_0(t)\exp\left[\sum_{j=1}^{p}\beta_j X_j\right] \tag{4}$$

X is a matrix of predictor variables that are modeled to predict each hazard value. There are two elements in Equation 8, the baseline hazard function $h_0(t)$, a function that involves time (t) but not **X** and $\exp\left[\sum_{j=1}^{p}\beta_j X_j\right]$ is an exponential function that involves **X** instead of t (time independence). The notation **β** in equation (8) is a parameter with maximum likelihood (ML) estimation which is denoted as $\hat{\boldsymbol{\beta}}$. Maximum Likelihood estimates of the Cox model parameters are derived by maximizing the likelihood function which is denoted as L. The likelihood function is a mathematical model that describes the probability of data on a research subject as a function of the unknown parameter (**β**) in the model being observed. The likelihood function is written with the notation $L(\boldsymbol{\beta})$ where **β** denotes the set of unknown parameters. The likelihood function is a partial function that only considers the probability of subjects failing and does not consider the probability of subjects being censored. After the likelihood function is formed, the next step is to maximize the likelihood function by maximizing log L. The likelihood function for the Cox PH model is written in Equation (5).

$$L(\boldsymbol{\beta}) = \prod_{l=1}^{r}\frac{\exp\left[\boldsymbol{\beta}' x_{(t)}\right]}{\sum_{j\in R(t_{(l)})}\boldsymbol{\beta}' x_{(l)}} \tag{5}$$

$x_{(l)}$ be a variable vector of failed individuals at time l with time $t(l)$, and $R\left(t_{(l)}\right)$ be the set of all individuals who have a risk of failure at time l. After getting the partial likelihood function, the next step is to maximize the first derivative of the ln $L(\boldsymbol{\beta})$ function using the Newton-Rhapson function. Based on Equation (5) we can continue by using the ln-likelihood function and then produce the first derivative and second derivative. The first derivative of the likelihood function has been transformed into ln-likelihood. If $\mathbf{g}(\boldsymbol{\beta})$ is a vector of size p×l, it is the first derivative of the function ln $L(\boldsymbol{\beta})$ concerning parameter **β**. $\mathbf{H}(\boldsymbol{\beta})$ is a Hessian matrix measuring p×p which contains the second derivative of the likelihood function which is informed by ln-likelihood, so the parameter estimates in the (l+1) iteration can be seen in Equation 6.

$$\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} - \mathbf{H}^{-1}\left(\boldsymbol{\beta}^{(l)}\right)\mathbf{g}\left(\boldsymbol{\beta}^{(l)}\right) \tag{6}$$

We set the initial value $\boldsymbol{\beta}^{(0)}$ as the result of parameter estimation using multiple linear regression using the Least Square method. The iteration will stop if $\left\|\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{(l)}\right\| \le \varepsilon$, where ε is a very small number.

2.5 Hazard Ratio

The ratio level between individuals and predictor variables and the categories of success or failure can be seen with the Hazard Ratio [8]. The estimated value of the Hazard Ratio is obtained from the exponent of the Cox regression coefficient of each significant predictor variable with its hazard rate. The formula for the Hazard Ratio that compares two individuals can be seen in Equation 7.

$$\widehat{HR} = \exp\left[\sum_{j=1}^{p} \beta_j \left(X_j^* - X_j\right)\right] \tag{7}$$

The Hazard Ratio obtained indicates that the rate of occurrence of failure events (failure rate) in individuals with category $X = 1$ is X times greater than the rate of occurrence of the risk of failure events in individuals with category $X = 0$. The Hazard Ratio value can be obtained using Equation 8.

$$\widehat{HR} = \frac{h(t, X=1)}{h(t, X=0)} = \frac{\hat{h}_0(t)e^{\hat{\beta}}}{\hat{h}_0(t)} = e^{\hat{\beta}} \tag{8}$$

2.6 Best Model Selection

Selection of the best model with certain criteria is carried out to obtain the best model that expresses the relationship between survival time and predictor variables. Akaike Information Criterion (AlC) is a method for selecting the best model in the Cox regression model with the criteria for the smallest AlC value. The AIC value is obtained using Equation 9 [9, 10].

$$AIC = -2\ln L(\hat{\boldsymbol{\beta}}) + 2p \tag{9}$$

with p being the number of variables used.

3 Data Example

The data used in this research is medical record data of COVID-19 patients who were hospitalized at RSU Haji Surabaya in the period June 2021 to September 2021, totaling 131 observations. The dependent variable in this study is data on the survival time of patients suffering from COVID-19, with the following conditions;

a. The initial time (time origin) is the time when the initial patient was admitted to RSU Haji Surabaya for hospitalization due to COVID-19;
b. The event observed was the condition when a patient suffering from COVID-19 was declared cured and allowed to go home from the hospital;
c. The measurement scale of this research is in days;

The independent variables in this research are as follows;

a. Gender (X1)

This predictor variable is the gender of COVID-19 patients who are hospitalized at RSU Haji Surabaya and is categorized as follows.

0 = Male
1 = Female

b. Age (X2)

This predictor variable is the age of COVID-19 patients who are hospitalized at RSU Haji Surabaya and a ratio scale.

c. Patient symptoms when admitted to hospital (X3)

This predictor variable is the symptoms or conditions experienced by COVID-19 patients who are hospitalized at RSU Haji Surabaya. It is said to have mild symptoms if the patient experiences one or more symptoms in the form of body temperature ≥ 38°C, cough, runny nose, and body weakness. It is categorized as having mild symptoms if the patient experiences one of the symptoms in the form of shortness of breath and chest pain, pneumonia. These variables are categorized as follows.

0 = Mild Symptoms (body temperature ≥ 38°C, cough, runny nose, body weakness)
1 = Severe Symptoms (chest tightness and pain, pneumonia)

d. Hypertension (X4)

This predictor variable is the blood pressure of COVID-19 patients who are hospitalized at RSU Haji Surabaya and is categorized as follows.

0 = normal blood pressure (90/60 mmHg to 130/90 mmHg)

1 = high blood pressure (>130 mmHg)

e. Diabetes (X5)

This predictor variable is the blood sugar level of COVID-19 patients who are hospitalized at RSU Haji Surabaya and is categorized as follows.

0 = does not have diabetes (fasting sugar level ≤ 125mg/dl)

1 = has diabetes (fasting sugar level > 125mg/dl)

f. Cardiovascular (X6)

This predictor variable is the heart disease experienced by COVID-19 patients who are hospitalized at RSU Haji Surabaya and is categorized as follows.

0 = has no heart problems

1 = has heart problems

An explanation of the research variables is presented in Table 1.

Cox Proportional Hazard modeling was carried out using all predictor variables that are thought to influence the recovery rate of patients with confirmed COVID-19. The parameter estimation results for the Cox Proportional Hazard model are shown in Table 1.

Table 1. Parameter Estimation

Variable	Estimation	Wald	p-value
Sex (1)	-0.518	-2.308	0.021
Age	0.0003	0.041	0.967
Symptom (1)	-1.631	-6.012	0.000
Hypertension (1)	0.365	1.088	0.277
Diabetes (1)	-0.210	-0.665	0.506
Cardiovascular (1)	-0.206	0.607	0.544
Likelihood ratio	48.260		0.000

Significance testing was carried out in the form of both simultaneous tests and partial tests. Simultaneous testing of the model was carried out using the likelihood ratio. Based on Table 1, the likelihood ratio is 48.260 with a p-value of 0.000. By using α of 0.05, the decision to reject H0 is obtained, which means there is at least one variable that has a significant influence on the recovery rate of COVID-19 patients. Significant variables in the model can be identified by conducting a partial test. According to Table 1, you can see the p-value of each predictor variable. Variables that influence include age and symptoms.

3.1 Proportional Hazard Assumption

Factors that are thought to influence the recovery rate of COVID-19 patients need to be tested using the proportional hazards assumption test to find out whether these factors are independent of time. The Goodness of Fit test method is a test of the correlation between the sequenced survival times and the Schoenfeld residual. The results of testing the Goodness of Fit assumption are in Table 2.

In Table 2, it is known that all p-values for each variable are more than α (0.05) which means that all variables have met the Proportional Hazard assumption. The best model selection was carried out using the Backward elimination method to obtain the best Cox Proportional Hazard regression model. Backward elimination is carried out by eliminating one by one the least significant variables from the model. Backward elimination results are listed in Table 3.

Table 2. Goodness of Fit Test of Proportional Hazard Assumption

Variable	Estimation	*p-value*
Sex (1)	0.031	0.861
Age	0.036	0.850
Symptom (1)	0.394	0.530
Hypertension (1)	0.405	0.525
Diabetes (1)	3.353	0.067
Cardiovascular (1)	1.329	0.249

Table 3. Best Model Selection

Step	Model	Eliminated Variable	AIC
0	All variables		670.29
1	Sex, symptom, hypertension, diabetes, cardiovascular	Age	668.30
2	Sex, symptoms, hypertension, diabetes,	Cardiovascular	666.66
3	Sex, symptoms, hypertension,	Diabetes	665.07
4	Sex, symptom	Hypertension	663.88

Table 3 shows that the Backward elimination process starts by modeling all predictor variables in step 0 to produce a model in step 4. Judging from the AIC value in the last step, the best Cox Proportional Hazard model for modeling the recovery rate of confirmed COVID-19 patients is obtained. model with variable X1, namely Sex and variable X3, namely symptoms. After the best model is formed, the parameters of the best Cox Proportional Hazard model are estimated. The best Cox Proportional Hazard modeling is done using variables obtained from Backward elimination. The best Proportional Hazard model parameter estimation results are presented in Table 4.

Table 4. Best Model Parameter Estimation

Variable	Estimation	Wald	*p-value*
Sex (1)	-0.537	-2.413	0.015
Symptom (1)	-1.588	-6.443	0.000
Likelihood Ratio	46.680		0.000

Based on the parameter estimation results of the Cox Proportional Hazard model in Table 4, the following model is obtained.

$$\hat{h}(t,\mathbf{X}) = \hat{h}_0(t)\exp\left[-0.537\text{Sex}(1) - 1.588\text{Symptom}(1)\right]$$

Based on the Cox Proportional Hazard model, simultaneous and partial parameter significance tests were carried out. The simultaneous test can be seen from the likelihood ratio value in Table 4 where the likelihood ratio value is 46.680 and the p-value is 0.000. By using $\alpha = 0.05$, there is at least one variable in the model that has a significant effect on the recovery rate of COVID-19 patients. Variables that have a significant effect are obtained through partial tests. Based on the p-value of each variable in Table 4, it is known that all variables in the model have a significant effect because the p-value is less than $\alpha=0.05$. Based on the Cox Proportional Hazard regression modeling that was carried out, it was found that the gender variable and symptom variables had a significant effect on the recovery rate of COVID-19 patients at RSU Haji Surabaya. The results obtained were then interpreted by the Cox Proportional Hazard model to determine the Hazard Ratio for each variable that influenced the patient's recovery rate.

Table 5. Hazard Ratio

Variable	Estimation	Hazard Ratio
Sex (1)	-0.537	0.584
Symptom (1)	-1.587	0.204

Table 5 shows that the Hazard Ratio value for female COVID-19 patients is 0.584, which means that female COVID-19 patients have a 0.584 times lower risk of not recovering than male or female COVID-19 patients. The recovery rate for male patients is 1.712 times faster than for female patients. The Hazard Ratio value for COVID-19 patients with severe symptoms is 0.204, which means that COVID-19 patients with severe symptoms have a 0.204 times lower risk of not recovering or the recovery rate for patients with mild symptoms is 4.9 times faster than COVID-19 patients. Those with severe symptoms.

4 Conclusion

The results of Cox Proportional Hazard regression modeling showed that two variables influenced the survival time of COVID-19 patients, namely the gender variable and the symptom variable. The recovery rate for male patients is faster than for female patients. Furthermore, patients with mild symptoms have a faster recovery rate than patients with severe symptoms. For further research, other factors that influence the survival time of COVID-19 patients need to be considered. The Surabaya Haji General Hospital is expected to carry out medical treatment by continuing to monitor the patient's condition by paying attention to factors that have a significant influence on survival time, especially female patients and patients with severe symptoms so that the recovery rate is faster.

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under the project scheme of the Department of Statistics Research Project 2023 with contract number 2388/PKS/ITS/2023.

References

1. Lu, W., Yu, S., Liu, H., Suo, L., & Tang, K. (2021). Survival Analysis and Risk Factors in COVID-19. *Disaster Med Public Health Prep*, 1-6.
2. Bernal, J. L., Andrews, N., Gower, C., & Gallagher, E. (2021). Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *The New England Journal of Medicine*(385), 585-594.
3. Araf, Y., Akter, F., Tang, Y.-d., Parvez, M. S., Zheng, C., & Hossain, M. G. (2022). Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. *J Med Virol, 94*(5), 1825-1832.
4. Bintoro, S. U., Dwijayanti, N. M., Pramuda, D., & Amrita, P. N. (2021). Hematologic and coagulopathy parameter as a survival predictor among moderate to severe COVID-19 patients in the non-ICU ward: a single-center study at the main referral hospital in Surabaya, East Java, Indonesia. *PubMed Central, 10*, 791.
5. Soeroso, R. M., & Sulistiadi, W. (2022). Strategi Rumah Sakit di Indonesia dalam Mengatasi Kenaikan Kasus COVID-19 Varian Omicron: Literature Review. *Media Publikasi Promosi Kesehatan Indonesia, 5*(4), 352.
6. Albana, A. S., & Azhari, S. (2020). Prediksi Penyebaran COVID-19 Kota Surabaya dengan Simulasi Monte Carlo. *Journal of Advances in Information and Industrial Technology (JAIIT), 2*(1), 36-42.

7. Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis A Self-Learning Text* (3rd ed.). New York: Springer.
8. Hosmer, D., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis.* Wiley: Hoboken.
9. Collet, D. (2003). *Modeling Survival Data in Medical Research* (2nd ed.). London: Chapman Hall/CRS A CRC Press Company.
10. Klein, J.P., and Moeschberger, M.L. (2003). *Survival Analysis Techniques for Censored and Truncated Data* (2nd.ed.). New York: Springer.