Survival Analysis Using Cox Proportional Hazard Regression Approach in Dengue Hemorrhagic Fever (DHF) Case in Abdul Moeloek Hospital Bandar Lampung in 2019

Survival Analysis Using Cox Proportional Hazard Regression Approach in Dengue Hemorrhagic Fever (DHF) Case in Abdul Moeloek Hospital Bandar Lampung in 2019

M Irfan.[1], M Usman.[1], S Saidi.[1], Warsono.[1], D Kurniasari.[1], Widiarti.[1]

[1)]Departement of Mathematics, Faculty of Mathematics and Natural Science, University of Lampung, Indonesia

email: miftahulirfan311@gmail.com

Abstract Survival analysis is one of the statistical procedures analyzing data in the form of survival time and variables that affect survival time, namely survival time data starting from the beginning of the study (time origin/start point) until the time an event or endpoint occurs. In the field of health data is obtained from observations of patients who were observed and recorded the time event of each individual. The event in question can be in the form of death, recurrence of new diseases, or recovery. This study will discuss the application of cox proportional hazard regression to determine the cox proportional hazard model on DHF patients. Determine the factors that affect the recovery rate of DHF patients and determine the hazard ratio value of DHF patients at Abdul Moeloek Hospital in Bandar Lampung in 2019. Regression Cox proportional hazard is used because the cox proportional hazard model does not depend on the assumption of the distribution of the time of occurrence; the results are almost the same as the parametric model. They can estimate the hazard ratio without knowing the baseline hazard. Based on the selection of the best model with backward elimination and the Akaike Information Criterion (AIC), the best model is obtained with a four-variable model, namely Leukocyte, Hemoglobin, Hematocrit, and Thromobocyte. These four variables are factors that have a significant effect on the patient's length of stay. Then this study also looked at the value of the hazard ratio which thrombocyte are the variables with the largest hazard ratio value, which means the thrombocyte variable has the highest risk level.

Keywords: Survival Analysis, Cox Proportional Hazard Regression, Backward Elimination, Akaike Information Criterion, Hazard Ratio.

1.Introduction

Various studies in biology, physics, agriculture, and medicine will usually produce data related to an individual's lifetime. Survival data can be analyzed with survival analysis. Lifetime data is a nonnegative variable. Statistical analysis used to analyze lifetime data is called survival analysis [1]. Survival analysis is one of the statistical procedures for analyzing data in the form of survival time and variables that affect survival time. Survival time data starts from the beginning of the time of research (time origin/start point) until the time of occurrence of an event or endpoint [2]. In survival analysis, there are three methods, namely parametric, semi-parametric, and nonparametric. The semi-parametric method that is often used is cox regression. Cox regression is used because the cox model does not depend on the assumption distribution from the time of its occurrence. The results of the Cox regression are almost the same as the results of the parametric model. They can estimate the hazard ratio without knowing the baseline hazard function. Also, this model is a safe model to choose when in doubt to determine its parametric model, so there is no fear about the wrong choice of parametric models [3]. The Cox proportional hazard regression model has the assumption that the hazard function of different individuals is proportional or the ratio of the hazard function of two different individuals is constant [4]. Censored data is data that cannot be observed in full, due to missing individuals or for other reasons, so that data cannot be retrieved or until the end of the observation, the individual has not experienced a particular event. If it is in the opposite condition then the data is called uncensored data [4]. In survival analysis there are 3 types of censorship, they are right sensor, the left sensor, and the interval sensor. In this study, the right sensor will be used because there are patients who disappear before the expected event that occurs namely the recovery event [5]. In this study data health will be used, namely patient data for Dengue Hemorrhagic Fever in Abdul Moeloek Hospital.

Dengue Hemorrhagic Fever (DHF) is one of the diseases that every coming of the rainy season becomes a topic of discussion between the government and the community. It is a disease that can spread quickly and often cause death in a short time. One modeling estimate indicates 390 million dengue virus infections per year (95% credible interval 284–528 million), of which 96 million (67–136 million) manifest clinically (with any severity of disease) [6]. Another study on the prevalence of dengue estimates that 3.9 billion people are at risk of infection with dengue viruses. Despite the risk of infection existing in 129 countries [7], 70% of the actual burden is in Asia [6]. According to WHO, the number of dengue cases reported to WHO increased over 8 fold over the last two decades, from 505,430 cases in 2000, to over 2.4 million in 2010, and 4.2 million in 2019 [8]. Reported deaths between the years 2000 and 2015 increased from 960 to 4032. Research on DHF was conducted by [9] survival analysis in patients with Dengue Hemorrhagic Fever (DHF) at the Haji Hospital Surabaya using the Weibull regression model.

According to the Indonesian Ministry of Health, In Lampung province, in 2020 the period of January to March ranks first in the provinces with the highest DHF cases, with 3,423 cases with 11 deaths[10]. The hospital that will be used as a research sample is the Abdul Moeleok Hospital in Bandar Lampung. So based on these explanations the author is interested in conducting a study of DHF patients with the title "Survival Analysis Using Cox Proportional Hazard Regression approach in Dengue Hemorrhagic Fever (DHF) case in Abdul Moeloek Hospital Bandar Lampung in 2019"

2. Statistical Model

2.1 *Probability Density Function*

Probability density function is the probability of an individual dying or failing in the time interval t to $t + \Delta t$. Probability density function is denoted by $f(t)$ and is formulated by:

$$f(t) = \lim_{\Delta t \to 0} \left[\frac{P(t<T<(t+\Delta t))}{\Delta t} \right] = \lim_{\Delta t \to 0} \left[\frac{P(F(t+\Delta t) - F(t))}{\Delta t} \right] \quad (1)$$

2.2 *Survival Function*

The survival function $\big(S(t)\big)$ states that opportunities do not fail until the time limit t. If *T* symbolizes survival time greater than *t* [3], then the equation is as follows:

$$S(t) = P(T > t) = 1 - F(t) \tag{2}$$

2.3 *Hazard Funtion*

The hazard function or failure function of the survival time *T* denoted $h(t)$ is the probability of an individual achieving a specific event at time *t*, provided that it has survived until that time [3]. Hazard function is defined as follows:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \\ &\vdots \\ &= \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)} \end{aligned} \tag{3}$$

2.4 *Assumption of Proportional Hazard*

To test the proportional hazard assumption in a Cox proportional hazard model there are two ways to test it [2].

2.4.1 Testing proportional hazard assumption using the log-minus-log survival plot

If the log-minus-log survival plot shows a parallel or intersecting curve, then the proportional hazard assumption is not met

2.4.2 Testing the proportional hazard assumption with schoenfeld residual

In Schoenfeld residuals if $\mathrm{p} - \mathrm{value} < 0{,}05$, the covariates tested did not meet the proportional hazard assumption [11], with the test statistics as follows:

$$\chi^2_{hit} = \frac{\{\sum(g_j - \bar{g}) R_{ji}\}^2}{d\, I_j \sum(g_j - \bar{g})^2} \tag{4}$$

2.5 *Cox Proportional hazard (Cox PH) Model*

In general, the form of the model Cox proportional hazard regression is:

$$h(t) = h_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \beta_k X_k} \tag{5}$$

2.6 *Estimation of paremeters in the cox proportional hazard Model*

Cox regression uses partial likelihood for estimating parameters. Estimation of β_j can be obtained by maximizing the first derivative of the log partial likelihood function, which is as follows:

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = 0 \quad \rightarrow \quad \sum_{i=1}^{r} \left[\sum_{j=1}^{k} X_{ji} - \frac{\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{k} X_{jl} \beta_j\right) \sum_{j=1}^{k} X_{jl}}{\sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{k} X_{jl} \beta_j\right)} \right] = 0 \tag{6}$$

2.7 *Newton-Raphson Procedure*

Used to maximize the partial likelihood function. The algorithm used in the Newton Raphson's method is assumed $c = 0, 1, 2, \ldots$ and $I\big(\hat{\beta}_c\big)^{-1}$ is the inverse of $I\big(\hat{\beta}_c\big)$ Is as follows:

$$\hat{\beta}_{c+1} = \hat{\beta}_c - I\big(\hat{\beta}_c\big)^{-1}\, U\big(\hat{\beta}_c\big) \tag{7}$$

2.8 *Testing the Significance of Model parameters*

How to test the significance of the cox model parameters, namely[12]:

2.8.1 Partial Likelihood Ratio Test

This test statistic is used to test the hypothesis that one or several β_j regression parameters for the model are zero. If $p - value < 0{,}05$, indicates that there is at least one independent variable that affects survival time. With test statistics as follows:

$$G = -2[ln\, L(0) - ln\, L(\beta_j)] \tag{8}$$

2.8.2 Wald Test

This test is used to test the effect of parameters separately, which is denoted by W. If $p - value < 0{,}05$, indicates that the independent variables affect the survival time. With the test statistics as follows:

$$W = \left(\frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}\right)^2 \tag{9}$$

2.9 *Hazard ratio*

The hazard ratio value is used to determine the level of risk (tendency) that can be seen from the comparison between individuals with the condition of the free variable X in the success category with the failure category. The hazard ratio for individuals with x = 0 versus x = 1 is:

$$Hazard\ Ratio = \frac{h_0(t|x=0)}{h_0(t|x=1)} = \frac{h_0(t)}{h_0(t)e^{\widehat{\beta}}} = e^{\widehat{-\beta}} \tag{10}$$

3. Data Analysis

The data of this study are DHF patient data obtained from Abdul Moeleoek Hospital Bandar Lampung.

Table 1. Variable of Research

Variable	Explanation	Type	Category
Y	The length of time the patient was treated until healed (Survival Time)	Continuous	
X_1	Age	Continuous	
X_2	Gender	Category	1= Male 0=Female
X_3	Hemoglobin (HB)	Category	1=normal 0=abnormal
X_4	Leukocyte	Category	1=normal 0=abnormal
X_5	Hematocrit	Category	1=normal 0=abnormal
X_6	Thrombocyte	Category	1=normal 0=abnormal

3.1 *Descrptive Statistics*

The following table is a descriptive statistics that explains the average, Standart Deviation minimum ana maximum of the research variables

Table 2. Descriptive Statistics

Variable	Mean	Std Dev	Minimum	Maximum	N
Survival Time	4,4	1,45	2	7	59
Age	31,1	11,93	16	68	59
Leukocyte	5,8	2,88	1,5	17	59
Hematocrit	40,1	5,58	25	49	59
Hemoglobin	14,4	2,09	8,5	18,3	59
Thrombocyte	49.458	68.436,8	5.000	334.000	59

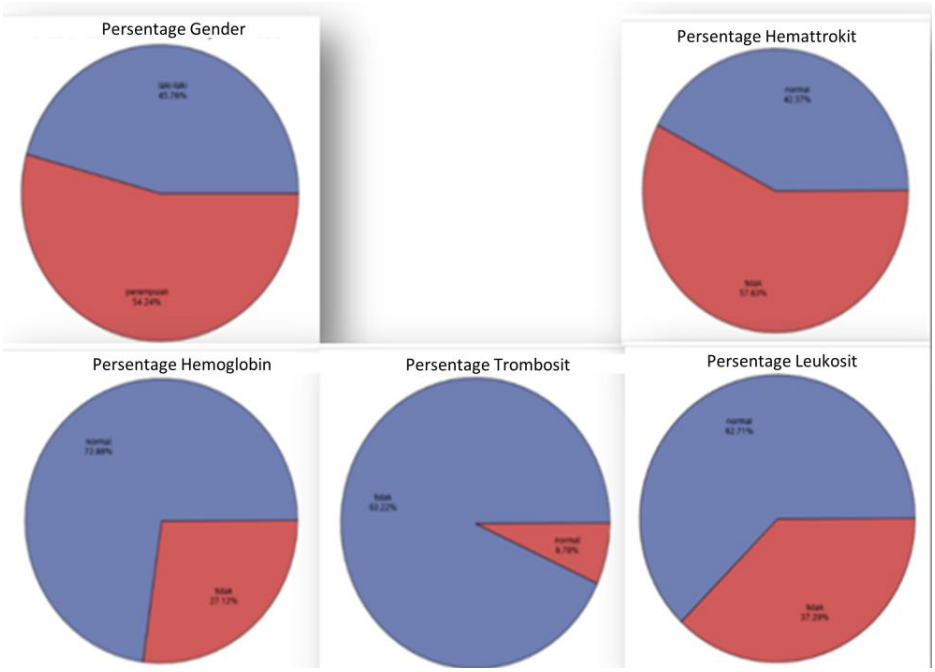

Figure 1. Percentage of gender, Thrombocyte, Leukocyte, Hemoglobin, dan Hematocrit

Based on Figure 1. shows that out of 59 DHF patients, the percentage between the number of DHF female patients is higher than male namely 45.76% and 54.24% female. The percentage of DHF patients with normal Thromobocyte was only 6.78% while those who were not normal were 93.22%, normal leukocytes were 62.71% while those who were not normal were 37.29%, normal hemoglobin was 72.88% while those who were abnormal were 27.12%, normal hematocrit was 42.37% while those who were not normal normal 57.63%. This shows that Thromobocyte and hematocytes tend to be abnormal whereas hemoglobin tends to be normal

3.2 *Checking the Proportional Hazard Assumption*

In testing the assumptions of PH, this study uses Residual Schoenfeld, as follows::

Table 3. Residual Schoenfeld

Variable	Correlation	P-Value	Decision
Age	-0,09222	0,4951	Failed to reject H_0

Gender	0,08201	0,5442	Failed to reject H_0
Hemoglobin	-0,15709	0,2432	Failed to reject H_0
Leukocyte	-0,18990	0,1571	Failed to reject H_0
Hematocrit	0,08950	0,5079	Failed to reject H_0
Thromobocyte	0,00901	0,9469	Failed to reject H_0

Based on table 3. shows that for each variable has p $-value > 0{,}05$, it can be concluded that the assumption of PH is fulfilled for all independent variables.

3.3 *Parameter Estimation and Parameter Significance Testing*

To estimate the initial parameters, enter all independent variables so that the estimated parameters obtained in the table below.

Table 4. Estimation of Parameters for All Variables

Variable	DF	Parameter Estimate
Age	1	-0.00985
Gender	1	-0.20363
Hemoglobin	1	1.07375
Leukocyte	1	0.86867
Hematocrit	1	0.97812
Thrombocyte	1	1.43470

Next, the parameter significance testing will be carried out simultaneously and partially. In testing the significance of these parameters simultaneously tested with the likelihood ratio test.

Table 5. Likelihood Ratio Test

Test	Chi-Square	Pr > ChiSq
Likelihood Ratio	28,1907	<0,0001

Seen in table 5. shows that the value of the likelihood ratio test is 28,1907 with p-value $(< 0{,}0001) < \alpha = 0{,}05$. So it can be concluded that there is at least one significant independent variable, or it can be interpreted that the model contributes to the recovery rate of DHF patients. Because there is at least one significant variable, it will be continued with separate testing.

Table 6. Wald Test

Variable	DF	Wald Chi-Square	Pr > ChiSq
Age	1	0.6960	0.4041
Gender	1	0.5126	0.4740
Hemoglobin	1	7.5065	0.0061
Leukocyte	1	7.4293	0.0064
Hematocrit	1	6.4783	0.0109
Thromobocyte	1	5.5327	0.0187

In table 6. shows that the age and gender variables have values of $\mathrm{p-value} > 0{,}05$ or more significant level. It can be concluded that the age and gender variables have no significant effect. For the variables of Hemoglobin, Leukocytes, Hematocrit and Thrombocytes have a $\mathrm{p-value} < 0{,}05$ so it can be concluded that the variables Hemoglobin, Leukocytes, Hematocrit and Thrombocytes significantly influence the length of stay variable.

3.4 *Selection of the best model*

In this study, backward elimination will be used to determine the best model, starting with removing the insignificant independent variables and the elimination process will stop when all variables are significant, then the value of AIC (Akaike Information Criterion) will be seen. Here are the results of backward elimination by removing age and gender variables

Table 7. Test the significance of the parameters eliminating the sex variables and Age

Variable	DF	Chi-Square	Pr > ChiSq
Hemoglobin	1	9.9915	0.0016
Leukocyte	1	6.7969	0.0091
Hematocrit	1	5.5416	0.0186
Thromobocyte	1	5.2491	0.0220

In Table 7. shows that the partial test results for the Hemoglobin, Leukocyte, Hematocrit and Thrombocyte variables are all significant with a $P - value < 0{,}05$ so that the backward elimination stage is stopped. Next will be seen the value of AIC (Akaike Information Criterion) of the steps that have been taken, starting from the full model to the third step by eliminating the age and gender variables. The following AIC values for each model:

Table 8. AIC Values

Model	Variable	AIC
1	Enter all the independent variables into the model (full model)	368,873
2	Without Variable X_2(Gender)	367,607
3	Without Variable X_2(Gender) and X_1(Age)	366,113

Table 8. shows that of the three models, the model that has the smallest AIC model 3, with a value of 366,113. From the stage of selecting the cox proportional hazard model 3 regression model (without Variable X_2(Gender) and X_1(Age)) is the model that meets the best model criteria because model 3 has the smallest AIC value with significant independent variables. After getting the best model, parameter estimation will be carried out on the variables that enter the model.

Table 9. Parameter estimation with the best model

Variable	DF	Parameter Estimate
Hemoglobin	1	1.20211
Leukocyte	1	0.82141
Hematocrit	1	0.86475
Thromobocyte	1	1.38290

From Table 9. shows the estimated parameters of the variable hemoglobin (X_3), leukocyte (X_4), hematocrit (X_5) dan Thromobocyte (X_6). From the parameter estimation results in table 4.11, the best cox proportional hazard model obtained after testing the parameters and selecting the model is as follows:

$$h(t) = h_0(t)\, exp(1{,}20211\ X_3 + 0{,}82141\, X_4 + 0{,}86475\, X_5 + 1{,}38290\, X_6)$$

3.5 *Hazard ratio*

To see the rate of recovery of DHF patients, hazard ratio will be used. The following are hazard ratio values for DHF patient variables:

Table 10. Hazard ratio

Variable	Parameter Estimate	Hazard Ratio
Hemoglobin	1.20211	3.327
Leukocyte	0.82141	2.274
Hematocrit	0.86475	2.374
Thromobocyte	1.38290	3.986

Based on Table 10. hazard ratio values can be seen to explain the cure rate of DHF patients.

a. The normal hemoglobin variable has a cure rate of 3.327 times compared to patients with abnormal hemoglobin.
b. The normal leukocyte variables has a cure rate of 2,274 times compared to patients with abnormal leukocytes.
c. The normal hematocrit variable has a cure rate of 2,374 times compared to patients with abnormal hematocrit.
d. The normal Thromobocyte variable has a cure rate of 3.986 times compared with Thromobocyte abnormal.

4. Conclusions

In the study of survival analysis in DHF patients in Abdul Moeloek Hospital Bandar Lampung 2019, the following conclusions are obtained:

1. The model obtained from the process of modeling the cox proportional hazard regression in DHF patient data in Abdul Moeleok Hospital Bandar Lampung 2019 is:

$$h(t) = h_0(t)\, exp(1{,}20211\ X_3 + 0{,}82141\, X_4 + 0{,}86475\, X_5 + 1{,}38290\, X_6)$$

2. From the model that has been obtained, the factors that influence the recovery of DHF patients in Abdul Moeloek Hospital are Hemoglobin, Leukocytes, Hematocrit and Thrombocyte.
3. From the value of the hazard ratio obtained, the cure rate of patients with abnormal Hemoglobin has a longer recovery rate compared to normal Hemoglobin, patients with abnormal Leukosti have a longer recovery rate compared to normal Leukocytes, patients with abnormal hematocrit have a longer recovery rate than patients with normal hematocrit, as well as patients with abnormal Thromobocyte. have a longer recovery rate than patients with normal Thromobocyte.

5. Acknowledgement

The author thanks the Abdul Moeloek Hospital in Bandar Lampung, who has provided data in this study.

References

[1] Lawless J F 1982 *Statistical Models and Methods for Lifetime Data* (United States of America, America)
[2] Collett D 2003 *Modelling Survival Data In Medical Research* Second Edition (Chapman & Hall, New York)
[3] Kleinbaum D G & Klein M 2005 *Survival analysis* : a self learning text (New York: Springer-Verlag)
[4] Lee E T & Wang J 2003 *Statistical methods for survival data analysis* (John Wiley & Sons, Canada)
[5] Jenkins S P 2005 *Survival Analysis* (Unpublished Manuscrip, New York)
[6] Bhatt S et al 2013 The global distribution and burden of dengue *Nature* 496(7446) P 504-507

[7] Brady O J et al 2012 Refining the global spatial limits of dengue virus transmission by evidence-based consensus PLOS *Neglected Tropical Diseases* 6(8) p e1760
[8] Mufidah A S & Purhadi 2016 Analisis Survival Pada Pasien Demam Berdarah Dengue (DBD) di RSU Haji Surabaya Menggunakan Model Regresi Weibull *Jurnal Sains Dan Seni ITS Vol. 5 No. 2*
[9] World Health Organization (WHO) 2020 Dengue and Severe Dengue (accessed on July 21^{st} 2020) https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue
[10] Kementerian Kesehatan RI 2020 Pasien DBD Capai 17.820, Ini 10 Provinsi dengan Jumlah Kasus Tertinggi. Kompas (accessed on July 21^{st} 2020) https://www.kompas.com/sains/read/2020/03/12/170300823/pasien-dbd-capai-17820-ini-10-provinsi-dengan-jumlah-kasus-tertinggi
[11] Therneau T M & Grambsch, P M 2000 *Modeling Survival Data Extending The Cox Modek* (Springer_Verlag, New York)
[12] Hosmer D W, Lemeshow S & May S 2008 *Applied Survival Analysis: Regression Modelling of Time to Event Data* (New Jersey: John Wiley)