International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

# Survival Analysis with a Stratified Cox Regression Model in COVID-19 Patients in Makassar City, South Sulawesi Province, Indonesia

Misbahuddin[1], Anna Islamiyati[2], Georgina Maria Tinungki[3]
[1,2,3]Department of Statistics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia
Corresponding Author's: misbha.misbahuddin5@gmail.com

***Abstract***— *The Stratified Cox regression model is a modification of the Cox proportional hazard regression model, which controls predictor variables that do not meet the proportional hazards (PH) assumption by forming strata or levels. The establishment of the Stratified Cox regression model will produce a proportional hazard Cox regression model for each stratum, so that variables that do not meet the proportional hazard (PH) assumption can still be observed in the strata. The regression coefficient in the model is estimated by maximizing the partial likelihood function, then iterated using the Newton-Raphson method. The data used in this study is data on COVID-19 patients throughout Makassar City, in the form of the long healing process for 54.349 patients who are thought to be affected by 3 variables, namely gender, comorbidities, and age. The data were then analyzed using a Stratified Cox regression model without interaction. The results show that the length of recovery for COVID-19 patients is influenced by gender and comorbid factors, with male and non-comorbid patients recovering faster.*

***Keywords***— *COVID-19; Stratified Cox Regression; and Non-Proportional Hazard.*

## I. INTRODUCTION

Survival analysis is one of the procedures in statistics for analyzing data, with the variable being considered being the time until an event occurs and other variables that are thought to affect survival time [1]. In survival analysis, there are three kinds of approaches: the parametric approach, the semi-parametric approach, and the nonparametric approach. The semi-parametric approach does not require assumptions on the survival time distribution, but the results of the parameter estimation are close to those of the parametric regression method [2]. One of the semi-parametric approaches commonly used is Cox proportional hazard regression [3].

Cox proportional hazard regression can be used even though the functional form of baseline hazard is unknown, but this Cox regression model can still provide useful information in the form of a hazard ratio that does not depend on baseline hazard [4]. The response variable in the Cox proportional hazard regression model must meet the proportional hazard assumption, but under certain conditions in a study, there are independent variables that do not meet the proportional hazard assumption. If the proportional hazard assumption is not met, then the linear component of the model varies depending on time and is said to be a non-proportional hazard [5]. One technique that can be used is stratified Cox regression, which is a modification of the Cox proportional hazard model [6].

The Stratified Cox model is an extension of the Cox proportional hazard model to address independent variables that do not meet the proportional hazard assumptions. Modifications are made by stratifying the independent variables that do not meet the proportional hazard assumption [1]. The Stratified Cox regression Model provides attention to or controls variables that do not meet the proportional hazard assumption by allowing strata or levels [6]. This is done because it is suspected that variables that do not meet the proportional hazard assumption still have a contribution, and the effect is still observed by making them strata that are not included in the model [7]. Several researchers have conducted research on Cox regression on survival data, such as Bedrosian et al. [8], Breslow et al. [9], George et al. [10], Zhang et al. [11], and Porta et al. [12].

## II. STRATIFIED COX REGRESSION

The stratified Cox regression model is an extension of the Cox proportional hazard model to address independent variables that do not meet the proportional hazard assumption [13]. The proportional hazard assumption states that the ratio of the hazard functions of two individuals is constant from time to time or is equivalent to the statement that the hazard function of one individual to the hazard function of another individual is proportional [14]. Modifications are made by stratifying the independent variables that do not meet the proportional hazard assumption [15-16]. Independent variables that meet the proportional hazard assumption are included in the model, while independent variables that do not meet the assumptions are not included in the model [7].

### *A. Stratified Cox Regression Model without Interaction*

The general form of the hazard function of the stratified Cox model without interaction is as follows [16]:

$$h_s(t,X) = h_{0_s}(t)\exp[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k] \quad (1)$$

Where $s$ is the strata defined from $Z^{*}$, $s = 1,2,\dots,m^{*}$, $h_{0s}(t)$ is the basic hazard function for each stratum, and $\beta_1,\beta_2,\dots,\beta_k$ are the regression parameters.

Strata is defined as a different category from the stratification variable $Z^{*}$, and $m^{*}$ is the number of strata. In the stratified Cox model, the basic hazard function, $h_{0s}(t)$ is different for each stratum. The regression parameters $\beta_1,\beta_2,\dots,\beta_k$ for this model are the same for each stratum, so the estimated hazard ratio is the same for each stratum.

Misbahuddin, Anna Islamiyati, and Georgina Maria Tinungki, "Survival Analysis with a Stratified Cox Regression Model in COVID-19 Patients in Makassar City, Shout Sulawesi Province, Indonesia," *International Research Journal of Advanced Engineering and Science*, Volume 8, Issue 3, pp. 46-49, 2023.

### *B. Hazard Ratio*

The hazard ratio is a measure used to compare the level of risk between individual comparisons with independent variable conditions in the success category and the failure category [15]. The hazard ratio value is the hazard for category one individuals divided by the hazard for different individuals, as in the following equation:

$$\widehat{HR} = \frac{\hat{h}(t,Z^*)}{\hat{h}(t,X)} = \frac{\hat{h}_0(t)\,e^{\sum_{i=1}^{p}\hat{\beta}_y x_y^*}}{\hat{h}_0(t)\,e^{\sum_{i=1}^{p}\hat{\beta}_y x_y}}$$

$$= \exp\left[\sum_{y=1}^{p}\hat{\beta}_y\left(Z_y^* - X_y\right)\right] \qquad (2)$$

Where Z* is the stratification result variable and X is the independent variable.

In the Stratified Cox regression model, the hazard ratio value is the same in each stratification variable category [17].

## III. Parameter Estimation of the Stratified Cox Regression Model

Parameter estimation in this Stratified Cox model uses the Maximum Partial Likelihood Estimation (MPLE) method, which is called the maximum stratified partial likelihood estimation [7]. The estimation of the regression parameter with the MPLE method is the value when the partial likelihood function is at its maximum [18]. The partial likelihood function for each stratum (subscript $s$ indicating strata) is as follows:

$$L_s(\beta) = \prod_{i=1}^{54349}\frac{\exp\left[\beta^T x_{(si)}\right]}{\sum_{j\in R(t_{si})}\exp\left[\beta^T x_{(sj)}\right]} \qquad (3)$$

Estimation of the regression parameters $\beta^T = [\beta_1, \beta_2, \dots, \beta_k]$ can be obtained by multiplying together the partial likelihood functions of each stratum, where each partial likelihood function of each stratum is derived from the corresponding hazard function.

$$L_s(\beta) = \prod_{s=1}^{2} L_s(\beta)$$

$$= L_1(\beta)\times L_2(\beta) \qquad (4)$$

Then the form of the partial likelihood stratification log function is obtained as follows:

$$\ln L_p(\beta) = \sum_{s=1}^{2}\left[\begin{array}{c}\sum_{i=1}^{54349}\beta^T x \\ -\ln\left(\sum_{j\in R(t_{si})}\exp\left[\beta^T x_{(sj)}\right]\right)\end{array}\right] \qquad (5)$$

To get the estimation of the regression parameter $\beta^T = [\beta_1, \beta_2]$ by maximizing the partial likelihood function by solving the logarithmic derivative of the partial likelihood function with respect to $\beta_g$ equal to zero as in the following equation:

$$\frac{\partial}{\partial\beta_g}\ln L_p(\beta) = 0$$

$$\frac{\partial}{\partial\beta_g}\sum_{s=1}^{2}\left[\begin{array}{c}\sum_{i=1}^{54349}\beta^T x \\ -\ln\left(\sum_{j\in R(t_{si})}\exp\left[\beta^T x_{(sj)}\right]\right)\end{array}\right] = 0 \qquad (6)$$

Where, $g = 1, 2, \dots, k$

Estimation of parameters in the stratified Cox model using the maximum partial likelihood estimation (MPLE) method by finding solutions from:

$$\frac{\partial}{\partial\beta_1}\ln L_p(\beta) = \frac{\partial}{\partial\beta_1}\ln\left(\prod_{i=1}^{54349}\frac{\exp\left[\beta^T x_{(si)}\right]}{\sum_{j\in R(t_{si})}\exp\left[\beta^T x_{(sj)}\right]}\right) = 0 \qquad (7)$$

The solution to the above equation is.

$$\frac{\partial}{\partial\beta_g}\left[\sum_{i=1}^{54349}\beta^T x - \ln\left(\sum_{j\in R(t_{si})}\exp\left[\beta^T x_{(sj)}\right]\right)\right] = 0 \qquad (8)$$

The above equation can be solved numerically using the Newton-Raphson method, with the following results [19]:

$$v\hat{a}r(\hat{\beta}) = I(\hat{\beta})^{-1} \qquad (9)$$

For an approximation of the standard deviation of $\hat{\beta}$ , as follows:

$$S\hat{E}(\hat{\beta}) = \sqrt{v\hat{a}r(\hat{\beta})} \qquad (10)$$

## IV. Result and Discussion

This study used survival data for COVID-19 patients, with a population of 54.349 patients spread across all hospitals in Makassar City. Variables that are thought to influence the survival of COVID-19 patients are gender ($X_1$), comorbidity ($X_2$), and age ($X_3$).

### *A. Description of the Data on the Survival Time of COVID-19 Patients and the Factors that Influence it*

The data used in this study is survival data for COVID-19 patients, with an overview as shown in Table 1 below:

TABLE I. Description of COVID-19 patient data

| Variable | Category | Number of patients | Percentage | Status | |
|---|---|---|---|---|---|
| | | | | Event | Sensor |
| Gender | L | 25239 | 46% | 601 | 24638 |
| | P | 29110 | 54% | 496 | 28614 |
| comorbidity | Yes | 18634 | 34% | 26 | 18608 |
| | No | 35715 | 66% | 601 | 34644 |
| Age | < 45 | 38608 | 71% | 218 | 38390 |
| | ≥ 45 | 15741 | 29% | 879 | 14862 |

This section discusses the characteristics of COVID-19 patients based on survival time and factors that are thought to influence the survival of COVID-19 patients who are treated in hospitals throughout Makassar City. Characteristics of survival time can be shown by using the survival curve. In the following, the general survival curve for COVID-19 patients is presented.

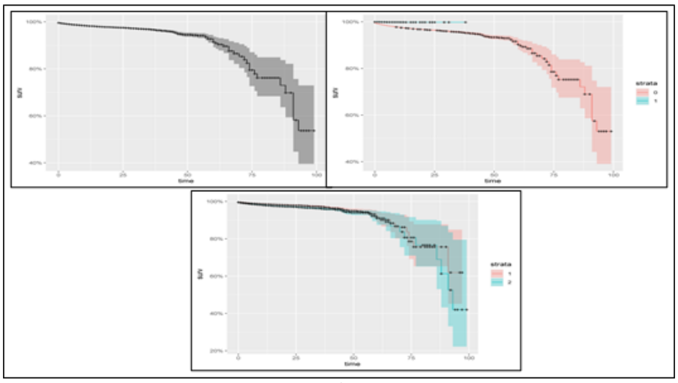

Fig. 1. Survival curve for COVID-19 patients

Figure 1 shows a graph of the survival function S(t) of the length of time to recover for COVID-19 patients in Makassar City and the factors that influence it. The median recovery time for COVID-19 patients in Makassar City is 50 days after being treated by medical personnel, and patients with a faster recovery time are COVID-19 patients who are male and non-comorbid.

## B. *Testing the Proportional Hazard Assumption*

The proportional hazard assumption test was carried out using the graphical method and the goodness of fit test, with the following results:

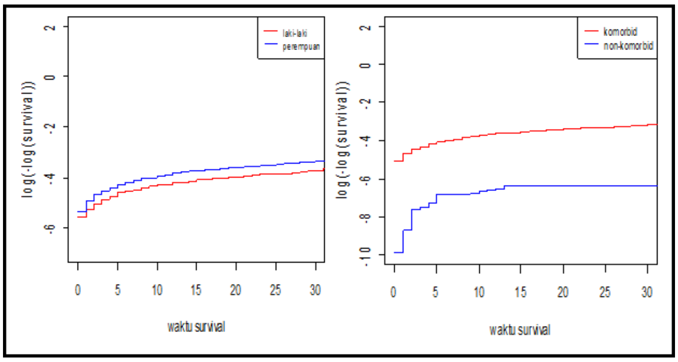

Fig. 2. Comparison plot of log (t) with log(-log(S(t))) for gender variable

Figure 2 shows that the log (t) and log(-log(S(t))) comparison plots for the sex and comorbid variables do not intersect on the two curves, so it can be said that the sex and comorbid variables fulfill the proportional hazard assumption. The age variable cannot be tested using a graph because the data is not categorized, so the test is continued with the Goodness of Fit test. The results of the calculation of the goodness of fit test for the predictor variables of gender, comorbidity, or age are presented in Table 2 below:

TABLE 2. Goodness of fit results

| Variable | P-Value | Decision |
|---|---|---|
| Gender | 0.18 | Fail Reject $H_0$ |
| Comorbidity | 0.58 | Fail Reject $H_0$ |
| Age | 0.005 | Reject $H_0$ |

In Table 2, it can be seen that the only variable that does not meet the proportional hazard assumption is the age variable because the p-value is smaller α = 0.05 and will be a stratification variable. So the Stratified Cox regression method can be used in this study.

## C. *Stratified Cox Regression Models*

The formation of the Stratified Cox regression model with stratified variables can be done with two models, namely, the model without interaction and the model with interaction. In this study, we used a Stratified Cox regression model without interaction.

### 1. *Stratified Cox regression model without interaction*

In Table 3, the results of the parameter estimation of the stratified Cox regression model are presented without interaction, with the stratification variable being age for survival time data for COVID-19 patients in Makassar City.

TABLE 3. Parameter estimation of the Stratified Cox regression model without interaction

| Variable | Estimation | P-Value |
|---|---|---|
| Gender | 0.260 | 0.011 |
| Comorbidityity | -2.906 | $2 \times 10^{-7}$ |

Based on Table 3 above, it is concluded that the Stratified Cox regression model without interaction in general can be written as follows:

The Author would like to thank the Makassar City Health Office for their willingness to provide data on COVID-19 patients.

Based on the model above, two models can be formed, namely, the hazard function for patients age < 45 years (s = 1) is:

$$h_1(t,X) = h_{01}(t)\exp[0{,}260X_1 - 2{,}906X_2]$$

The hazard function for patients aged ≥ 45 years (s = 2) is:

$$h_2(t,X) = h_{02}(t)\exp[0{,}260X_1 - 2{,}906X_2]$$

Furthermore, based on the model obtained above, parameter testing will be carried out simultaneously and partially. Simultaneously obtained a statistical value of the likelihood ratio test of 0.002. When compared with α value of 5%, α decision to reject $H_0$ is obtained. So, it can be concluded that there is at least one independent variable that has a significant effect on the model at α = 5% confidence interval. Variables that affect the survival of COVID-19 patients who receive treatment at the Makassar City Hospital in the two-year study period are gender and comorbid variables.

### 2. *Hazard Ratio*

Based on the results of parameter estimation for the Stratified Cox regression model without interaction, there are two significant variables. Hazard ratio values are used to interpret significant variables. The following is the hazard ratio value presented in Table 4:

TABLE 4. Hazard ratio for significant variables

| Variable | Hazard Ratio | Exp ($\beta$) | P-value |
|---|---|---|---|
| Gender | 1.298 | 1.298 | 0.011 |
| Comorbidity | 0.054 | 0.054 | $2 \times 10^{-7}$ |

Based on Table 4, it is known that the significant variables are sex and comorbidities because the p-value < 0.05, which means that sex and comorbidities affect the survival of

COVID-19 patients. Male patients have a survival probability 1.2 times higher than patients of female gender. Meanwhile, COVID-19 patients who do not have comorbid status have a survival probability 0.05 times higher than COVID-19 patients who have comorbid status.

## V. Conclusion

Based on the results of the analysis and discussion of estimation and Stratified Cox regression modeling in COVID-19 patients at hospitals throughout Makassar City, it can be concluded that the factors that can affect the length of recovery for COVID-19 patients are gender and comorbidities. Patients of the male gender have a survival probability 1.2 times higher than patients of the female gender. Meanwhile, COVID-19 patients who do not have comorbid status have a survival probability 0.05 times higher than COVID-19 patients who have comorbid status.

## Acknowledgment

The Author would like to thank the Makassar City Health Office for their willingness to provide data on COVID-19 patients.

## References

[1] P. K. MOULIK, R. MTONGA, and G. V. GILL, “Amputation and mortality in new-onset diabetic foot ulcers stratified by etiology,” *Diabetes Care*, vol. 26, no. 2, pp. 9–12, 2003.

[2] I. Selingerova, S. Katina, and I. Horova, “Comparison of parametric and semiparametric survival regression models with kernel estimation,” *J. Stat. Comput. Simul.*, vol. 91, no. 13, pp. 2717–2739, 2021.

[3] E. Goetghebeur and L. Ryan, “Semiparametric regression analysis of interval-censored data,” *Biometrics*, vol. 56, no. 4, pp. 1139–1144, 2000.

[4] L. D. Fisher and D. Y. Lin, “Time-dependent covariates in the Cox proportional-hazards regresion model,” *Annu. Rev. Public Health*, vol. 20, no. 6, pp. 145–157, 1999.

[5] C. V. Gonzalez, J. F. Dupuy, M. F. Lopez, P. L. Luaces, C. R. Rodriguez, G. G. Marinello, E. N. Vinagera, B. G Verdecial, B. W. Brito, L. M. Perez, M. T. de la Concepcion, C. M. Tania, “Stratified Cox regression analysis of survival under CIMAvax® EGF vaccine,” *J. Cancer Ther.*, vol. 04, no. 08, pp. 8–14, 2013.

[6] M. D. T. Chen, L. Xiang, L. Yingxue, Q. Yong, X. Eryu, Q. Yong, L. Shaoshan, X. Feng, L. Dandan, Z. Caihong, and L. Zhihong, “Prediction and risk stratification of kidney outcomes in IgA Nephropathy,” *Amerikan J. Kidney Dis.*, vol. 74, no. 3, pp. 300–309, 2019.

[7] D. V. Mehrotra, S. C. Su, and X. Li, “An efficient alternative to the stratified Cox model analysis,” *Stat. Med.*, vol. 31, no. 17, pp. 1849–1856, 2012.

[8] I. Bedrosian, C. Hu, and G. J. Chang, “Population-based study of contralateral prophylactic mastectomy and survival outcomes of breast cancer patients,” *J. Natl. Cancer Inst.*, vol. 102, no. 6, pp. 401–409, 2010.

[9] N. E. Breslow and J. A. Wellner, “Weighted likelihood for semiparametric models and two-phase stratified samples , with application to Cox regression,” *Scand. J. Stat.*, vol. 34, no. 1, pp. 86–102, 2018.

[10] B. George, S. Seals, and I. Aban, “Survival analysis and regression models,” *J. Nucl. Cardiol.*, vol. 21, no. 4, pp. 686–694, 2014.

[11] Z. Zhang, J. Reinikainen, K. A. Adeleke, M. E. Pieterse, and G. M. Catharina, “Time-varying covariates and coefficients in Cox regression models,” *Ann. Transl. Med.*, vol. 6, no. 7, p. 121, 2018.

[12] M. G. Della Porta *et al.*, “Risk stratification based on both disease status and extra-hematologic comorbidities in patients with myelodysplastic syndrome,” *Haematologica*, vol. 96, no. 3, pp. 441–449, 2011.

[13] J. In and D. K. Lee, “Survival analysis: Part II – applied clinical data analysis,” *Korean J. Anesthesiol.*, vol. 72, no. 5, pp. 441–457, 2019.

[14] S. L. Pugh, “Essence of survival analysis,” *Neuro-Oncology Pract.*, vol. 4, no. 2, pp. 77–81, 2017.

[15] M. Mohamed Ahmed Abdelaal, “Modeling survival data by using Cox regression model,” *Am. J. Theor. Appl. Stat.*, vol. 4, no. 6, p. 504, 2015.

[16] N. Ata and M. T. Sözer, “Cox regression models with nonproportional hazards applied to lung cancer survival data,” *Hacet J Math Stat*, vol. 36, no. 2, pp. 157–167, 2007.

[17] A. J. Turkson, J. A. Addor, and F. Ayiah-Mensah, “The Cox proportional hazard regression model Vis-a-Vis ITN-factor impact on mortality due to malaria,” *Open J. Stat.*, vol. 11, no. 06, pp. 931–962, 2021.

[18] M. H. Chen, J. G. Ibrahim, and Q. M. Shao, “Maximum likelihood inference for the Cox regression model with applications to missing covariates,” *J. Multivar. Anal.*, vol. 100, no. 9, pp. 2018–2030, 2009.

[19] M. Liu, W. Lu, R. E. Shore, and A. Zeleniuch-Jacquotte, “Cox regression model with time-varying coefficients in nested case-control studies,” *Biostatistics*, vol. 11, no. 4, pp. 693–706, 2010.