Comparison of Different Underlying Distributions in The Accelerated Failure Time (AFT) Model on Mortality of Covid-19 Patients

Asti Meiza[1]*, Asep Solih Awaluddin[1], Nurapni Oktapia Hidayah[1], and Adnan Taha[2]

[1]Department of Mathematics, UIN Sunan Gunung Djati Bandung, Indonesia
[2]Oxford Business College, Oxford, United Kingdom
Email: asti.meiza@uinsgd.ac.id

Abstract

In 2022, the COVID-19 virus is still making headlines in various mass media because it is a virus that is very dangerous to health. The world health organization, WHO, explained that the virus caused a global pandemic that infected the whole world. The condition of a pandemic has not yet turned into an endemic. Based on the total confirmed COVID-19 positive cases, Indonesia ranks 18th in the world out of 222 infected countries. To determine the influence factors on COVID-19 cases, survival analysis is one of the techniques that could be applied. One of the most commonly used models in survival analysis is Accelerated Failure Time (AFT) model. In the AFT model, it is required to check assumptions regarding the feasibility of the distribution form. In this study, the distributions used are Weibull, Exponential, Log-normal, and Log-logistics distributions. We compare each distribution to get the best model to analyze death cases due to COVID-19. Comparisons are made by comparing the AIC values of each distribution. The best model is selected based on the smallest AIC value. The AFT model with a log-normal distribution is selected as the best model with an AIC value of 142.763. The AIC value for this log-normal distribution is the smallest compared to the AIC value for other distributions.
Keywords: accelerated failure time model; COVID-19; mortality analysis; survival analysis.

Abstrak

Tahun 2022, virus COVID-19 masih menjadi berita utama di berbagai media massa karena merupakan salah satu virus yang sangat berbahaya bagi kesehatan. Badan organisasi kesehatan dunia, WHO menjelaskan bahwa virus menyebabkan terjadinya pandemi global yang menginfeksi seluruh dunia. Kondisi pandemi masih belum berubah menjadi endemi. Dari total yang terkonfirmasi positif COVID-19 Indonesia menduduki posisi ke-18 di dunia dari 222 negara yang terinfeksi. Untuk mengetahui faktor-faktor yang berpengaruh terhadap COVID-19 dan untuk menentukan model dari COVID-19 ini salah satunya dapat dilakukan dengan analisis survival. Salah satu model survival yang digunakan yaitu model Accelerated Failure Time (AFT). *Dalam model AFT ini diharuskan melakukan pengecekan asumsi-asumsi mengenai kelayakan bentuk distribusi. Pada penelitian ini distribusi yang digunakan yaitu distribusi Weibull, Eksponensial, Log-normal, dan Log-logistik. Dilakukan perbandingan antar tiap distribusi untuk mendapatkan model terbaik yang dapat digunakan dalam menganalisis kasus kematian akibat COVID-19. Perbandingan dilakukan dengan membandingkan nilai AIC dari setiap distribusi. Hasil penelitian memilih model AFT dengan distribusi log-normal sebagai model terbaik dengan nilai AIC sebesar 142,763. Nilai AIC untuk distribusi log-normal ini paling kecil dibandingkan dengan nilai AIC untuk distribusi lainnya.*
Kata Kunci: *analisis mortalitas; analisis survival; COVID-19; model* accelerated failure time.

---

1. INTRODUCTION

On December 31, 2019, the World Health Organization (WHO) or the World Health Organization China Country Office reported pneumonia in Wuhan City, Hubei Province, China, whose cause was still unknown. This disease is Coronavirus Disease 2019 (COVID-19). COVID-19 comes from the beta-coronavirus genus with the same genus as the causative agents of Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) [1]. The initial symptoms of COVID-19 are non-specific, starting with the appearance of fever and cough, which can then resolve spontaneously or develop into shortness of breath, dyspnea, and pneumonia which can cause Acute Respiratory Distress Syndrome (ARDS). In addition, it can cause kidney damage, coagulation or blood clotting dysfunction, and death [2] [3].

China stated that COVID-19 is a new type of pneumonia case of coronavirus on January 7, 2020 [4]. In Indonesia, confirmed cases of COVID-19 are shown in Figure 1 [5]. Based on this figure, it can be seen that in June, the confirmed positive cases of COVID-19 increased after the January 2021 period. On June 5, 2021, confirmed positive COVID-19 reached 1,843,612 people with 51,296 deaths. Of the total confirmed positive for COVID-19, Indonesia ranks 18th in the world out of 222 infected countries. Several studies have been developed to determine the factors that influence the deaths of COVID-19 patients, including [6] [7]. In this article, several significant factors that affect the mortality of COVID-19 patients will be determined using the survival analysis with the Accelerated Failure Time (AFT) model.
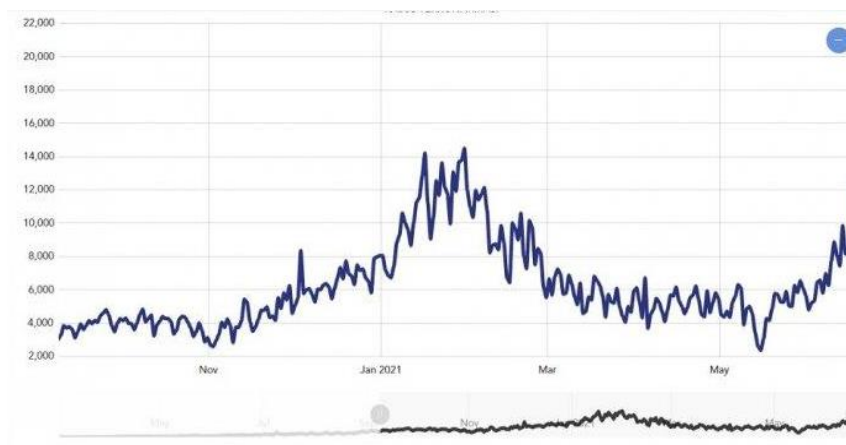
Figure 1. Daily graph of confirmed COVID-19 in 2021 [5].

The AFT model is one of the parametric models based on the data distribution that can predict the time of an event in the observation data. The distributions used are Exponential, Weibull, Log-normal, and Log-logistics distributions [8]. The AFT model can describe the relationship between the probability of survival and the set of covariates, where covariates can affect the survival time of an acceleration factor called the accelerated factor.

Recently, the AFT model has been widely applied to several problems. Jal-Usman et al. [9] identified covariates associated with the recovery time of Covid-19 patients using the AFT model. Khan and Howlader [10] implement the AFT to evaluate the stability in the selection process of some models with data sets containing many covariates. Granville and Fan [11] study the Buckley-James

estimator of accelerated failure time models with auxiliary covariates. They use a local polynomial approximation method to accommodate the covariates into the Buckley-James estimating equations. Shi-ang Qi et al. [12] describe the predictions of survival models to help professionals make decisions about patient conditions and resource allocation. They used a gradient-boosting Cox machine.

The AFT model is also widely used in medical research. Rachmaniyah et. al. [13] determine the chance of a faster death in people with diabetes mellitus. Crowther et al. [14] use the AFT model to accommodate time-dependent acceleration factors in breast cancer. Sulantari and Hariadi [15] model the time recovery of COVID-19 patients using the Kaplan-Meier method and the Log-Rank test on COVID-19 patients. The application of the AFT model in the insurance sector [16]. In this paper, we determine the factors that significantly affect the mortality of COVID-19 patients, determine a suitable distribution for COVID-19 cases, and determine the AFT model of the best distribution in cases of COVID-19 [17].

2. METHODS

Survival analysis is a statistical method used to analyze data about the time until an event occurs [18]. Survival time is the primary variable in survival analysis. The survival time can be interpreted as a variable that can measure the observation time from the initial point of observation to the end of the observation.

Definition 1. Suppose T is a non-negative random variable of the individual in survival time which states the time until a failure occurs. Survival function $S(t)$, define as the probability that the random variable T beyond time $t, (t > 0)$ (experiencing events after time t), which can be written as [19]

$$S(t) = P(T > t). \tag{1}$$

Definition 2. Hazard function $h(t)$ is commonly referred to as the force of mortality or the conditional failure rate. This hazard function is the probability of failure at a time interval [20], which can be expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2}$$

2.1. Accelerated Failure Time (AFT) Model

The AFTmodel is a parametric model based on the data distribution that can predict the time of an event in the observation data. The Accelerated Failure Time model is an alternative to the Cox Proportional Hazards (CPH) model [7]. The AFT method can describe the relationship between the probability of survival and the set of covariates, which can affect the survival time of an accelerated factor.

Hazard function of accelerated failure time with covariate $x_1, x_2, \ldots, x_p$ can be written as

$$h_i(t) = \left(\frac{1}{\eta(x)}\right) h_0 \left(\frac{t}{\eta(x)}\right), \tag{3}$$

where $h_0(t)$ is the baseline hazard function, $\eta(x)$ is the accelerated factor, and x is an independent variable. Research using the AFT model has been carried out by several researchers, including Khanal

et al. 2012 who examined the use of AFT to see the resilience of Acute Liver Failure Patients in India [21].

2.2. Distributions for the AFT Model

As stated in the previous section, the AFT is a parametric model. In this section, we will briefly discuss several distributions expressed in the AFT model, namely the Weibull distribution, Exponential Distribution, Log-normal distribution, and Log-logistic distribution. One distribution that best fits the data will be selected from these four distributions to be expressed in the AFT model.

The following is the hazard function for several distributions to be compared.

a) Weibull distribution

Weibull distribution hazard function can be written as

$$h_0(t) = \lambda\gamma(t)^{\gamma-1}, \text{ for } \lambda > 0 \text{ and } \gamma > 0. \tag{4}$$

b) Exponential distribution

The exponential distribution hazard function can be written as

$$h_0(t) = \lambda, \tag{5}$$

with parameters $t \geq 0$ and $\lambda \geq 0$.

c) Log-normal distribution

The log-normal distribution hazard function is as follows

$$h_0(t) = \frac{\frac{1}{\sqrt{2\pi}\sigma t}exp\left[\frac{-1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right]}{1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)}. \tag{6}$$

d) Log-logistic distribution

The log-logistics distribution hazard function is as follows

$$h_0(t) = \frac{e^\theta k t^{k-1}}{1+e^\theta t^k}. \tag{7}$$

The function $h_0(t)$ is a baseline hazard function that describes the risk for individuals with $X_i = 0$ [22]. This function is left unspecified [23]. Unknown means it is non-parametric [15], although the AFT is a parametric model.

2.3. The Parameter Significance Test

a) The Simultaneous Test

The Likelihood Ratio Test or G test is a test to see whether the independent variables as a whole have an effect or not on the model. The statistical formula for the G test used is [20]

$$G = -2(\log L_0 - \log L_p), \tag{8}$$

where L_0 is the likelihood function of the model without independent variables, and L_p is the likelihood function of the model consisting of p independent variables. Test statistics G follows a chi-square distribution with degrees of freedom p. If $G > \chi^2_{(\alpha,p)}$, then H_0 must be rejected, meaning that at least one independent variable can affect the model and vice versa.

b) The Partial Test

The partial test used is the Wald test statistic. This Wald test statistic follows a chi-square distribution with p degrees of freedom [21]. The hypothesis used is as follows:

$H_0: \beta_j = 0$ (independent variable is not significant),

$H_1: \beta_j \neq 0$ (independent variable is significant).

The statistical formula for Wald's test is

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

with $\hat{\beta}_j$ are coefficients that accompany the variables, and $SE(\hat{\beta}_j)$ is the error standard. If the significance value, $W > \chi^2_{(\frac{\alpha}{2})}$, H_0 is rejected, the independent variable is significant to the model, and vice versa.

2.4. Model Fit Test

Akaike's Information Criterion (AIC) method is one method that can be used to determine the best model found by Akaike. According to Klein and Moeschberger [18], the value of the AIC can be written as follows

$$AIC = -2l + 2p, \tag{10}$$

where l is log-likelihood and p is the total degrees of freedom used in the model. In determining the best model with this method, we use the smallest value of the AIC.

3. RESULTS AND DISCUSSION

In this research, we use primary data from the Bandung City Regional General Hospital in the period of January to June 2021. The sample was 100 patients, where 22 patients were uncensored, and 78 were censored. The data analysis method is divided into parameter significance and model feasibility tests. The parameter significance test is further divided into two jobs: simultaneous test and partial test.

The results of this study are divided into several parts, starting with descriptive results, then comparing the AFT model for several distributions. Finally, the model feasibility test results obtained the best distribution for the AFT model.

3.1. Descriptive Analysis

There are four variables in this study, i.e., age, gender, comorbid, and symptom. The descriptive statistics are presented in Table 1.

Table 1. Descriptive analysis of all variables

		Number of Patients	Mortalitas %
Age (X_1)	Age < 25.45	19	5.26%
	$25.45 \leq$ Age < 67.27	70	24.28%
	$67.27 \leq$ Age	11	36.36%
Gender (X_2)	Female	49	16.33%
	Male	51	27.45%
Comorbid (X_2)	None	77	6.49%
	Comorbid	23	73.91%
Symptom (X_4)	Mild	35	0%
	Moderate	37	8.11%
	Severe	28	67.86%

3.2. Estimation Parameter of The AFT Model

Before determining the AFT model, it is necessary to check the distribution assumptions first. Weibull, Exponential, Log-normal, and Log-logistics distributions are the distributions used in checking this assumption. Below is a provisional model of each distribution as follows:

a) Weibull Distribution

Table 2 shows the parameter estimation of the AFT Weibull model.

Table 2. Parameter estimation of the AFT Weibull Model

	Value	Std.Error	Z	P
Age	-0.220	0.215	-1.03	0.30491
Gender	-0.405	0.319	-1.27	0.20377
Comorbid	-1.354	0.363	-3.74	0.00019
Symptom	-0.796	0.309	-2.58	0.00992

Based on Table 2, the provisional model of the AFT Weibull model is obtained as follows:

$$h(t|X) = \frac{\left(\frac{1}{\exp(-0.220x_1-0.405x_2-1.354x_3-0.796x_4)}\right)\exp\left(-\frac{\mu}{\sigma}\right)}{\sigma}\left(\frac{t}{\exp(-0.220x_1-0.405x_2-1.354x_3-0.796x_4)}\right)^{\frac{1}{\sigma}-1}.(11)$$

b) Exponential Distribution

Table 3 shows the parameter estimation of the AFT Exponential model.

Table 3. Parameter estimation of the AFT Exponential model

	Value	Std.Error	Z	P
Age	-0.482	0.397	-1.21	0.22463
Gender	-0.835	0.564	-1.48	0.13896
Comorbid	-2.115	0.614	-3.45	0.00057
Symptom	-1.412	0.567	-2.49	0.01277

Based on Table 3, the provisional model of the AFT Exponential model is obtained as follows:

$$h(t|X) = \exp\left(\left(-\frac{\mu}{\sigma}\right) - (-0.482x_1 - 0.835x_2 - 2.115x_3 - 1.412x_4)\right). \tag{12}$$

Remember, in an Exponential distribution, the value of the basic hazard function is the same as the rate.

c) Log-normal Distribution

Table 4 shows the parameter estimation of the AFT Log-normal model.

Table 4. Parameter estimation of the AFT Log-normal model

	Value	Std.Error	Z	P
Age	-0.411	0.274	-2.11	0.1331
Gender	-0.695	0.330	-4.52	0.0348
Comorbid	-1.421	0.314	-2.88	6.0e-06
Symptom	-0.828	0.287	-2.06	0.0039

Based on Table 4, the temporary model of the AFT Log-normal model is obtained as follows:

$$h(t|X) = \left(\frac{1}{\exp(-0.411x_1 - 0.695x_2 - 1.421x_3 - 0.828x_4)}\right) \times$$

$$\left[\frac{\frac{1}{\sqrt{2\pi}\sigma t}exp\left[\frac{-1}{2}\left(\frac{\log\left(\frac{t}{\exp(-0.411x_1 - 0.695x_2 - 1.421x_3 - 0.828x_4)}\right) - \mu}{\sigma}\right)^2\right]}{1 - \Phi\left(\frac{\log\left(\frac{t}{\exp(-0.411x_1 - 0.695x_2 - 1.421x_3 - 0.828x_4)}\right) - \mu}{\sigma}\right)}\right]. \tag{13}$$

d) Log-logistic Distribution

Table 5 shows the parameter estimation of the AFT Log-logistics model. Based on this table, the provisional model of the AFT Log-Logistics model is obtained as follows:

$$h_i(t) = \frac{1}{\sigma t}\left\{1 + exp\left[-\left(\frac{\log t - \mu + 0.435x_1 - 0.605x_2 - 1.486x_2 - 0.737x_4}{\sigma}\right)\right]\right\}^{-1}. \tag{14}$$

Table 5. Parameter estimation of the AFT Log-Logistics model

	Value	Std.Error	Z	P
Age	-0.435	0.263	-1.65	0.098
Gender	-0.605	0.310	-1.95	0.051
Comorbid	-1.486	0.314	-4.73	2.3e-06
Symptom	-0.737	0.273	-2.70	0.007

3.3. Model Fit Test

Table 6 shows that the AFT Log-normal model has the smallest AIC value compared to the other three models, which is 142.763. Hence, the best model for the length of stay for the death of COVID-19 patients at the Bandung City Regional General Hospital is the AFT Log-normal model.

Table 6. The AIC values

Model	AIC
Weibull AFT	143.9453
Exponential AFT	151.7752
Log-normal AFT	142.7630
Log-logistic AFT	142.8104

Table 7 shows the parameter significance of the AFT Log-normal model. This table shows that the p-value for the AFT with the log-normal distribution is 5.6e-15, which is less than the 0.05 significance level. This value indicates that the log-normal distribution significantly affects the AFT model. The significant variables are gender, comorbidity, and symptoms. In contrast, the age variable is insignificant.

Table 7. Log-normal distribution

	p - value	Wald	Conclusion
Age	0.1331	2.25	Not significant
Gender	0.0348	4.435	Significant
Comorbid	6.0e-06	20.479	Significant
Symptom	0.0039	8.323	Significant
p - value			5.6e-15

From the parameter estimation, the final model of the AFT Log-normal model is

$$h(t|X) = \left(\frac{1}{\exp(-0.695x_2 - 1.421x_3 - 0.828x_4)}\right) \left[\frac{\frac{1}{\sqrt{2\pi}\sigma t}exp\left[\frac{-1}{2}\left(\frac{\log\left(\frac{t}{\exp(-0.695x_2 - 1.421x_3 - 0.828x_4)}\right) - \mu}{\sigma}\right)^2\right]}{1 - \Phi\left(\frac{\log\left(\frac{t}{\exp(-0.695x_2 - 1.421x_3 - 0.828x_4)}\right) - \mu}{\sigma}\right)}\right], \quad (15)$$

where x_2 is gender, x_3 is comorbid, x_4 is a symptom, t is age, μ is mean, and σ is the standard deviation.

4. CONCLUSIONS

This paper compares the AFT model using the Weibull, Exponential, log-normal, and log-logistic distributions. Based on the AIC value, the AFT model with a log-normal distribution is better than other distributions. There are three significant variables, i.e., gender, comorbidity, and symptoms. In

contrast, the age variable is insignificant. These significant variables mean that gender, the presence of comorbidities, and the onset of severe symptoms can affect death in COVID-19 patients.

REFERENCES

[1] J. Harlan, Analisis Survival, Depok: Gunadarma, 2017.

[2] R. Saikia and M. P. Barman, "A Review on Accelerated Failure Time Models," *International Journal of Statistics and Systems,* vol. 12, no. 2, pp. 311-322, 2017.

[3] A. A. Sayampanathan, C. S. Heng, P. H. Pin, J. Pang, T. Y. Leong and V. J. Lee, "Infectivity of asymptomatic versus symptomatic COVID-19," *Lancet,* vol. 397, pp. 93-94, 2021.

[4] B. K. Mishra, A. K. Keshri, Y. S. M. B. K. Rao, B. Mahato, S. Ayesha, B. P. Rukhaiyyar, D. K. Saini and A. K. Singh, "COVID-19 created chaos across the globe: Three novel quarantine epidemic models," *Chaos, Solitons & Fractals,* vol. 138, p. 109928, 2020.

[5] S. T. P. COVID-19, "Peta Sebaran," 1 August 2021. [Online]. Available: https://covid19.go.id/. [Accessed 1 August 2021].

[6] H. Nisa, A. S. F. Utami and A. D. Oktaviani, "Risk Factors for Mortality Among COVID-19 Patients in Asia: A Literature Review," *Jurnal Ilmu dan Teknologi Kesehatan,* vol. 9, no. 2, pp. 139-152, 2022.

[7] D. Raut, P. Shrivastav and S. Parwe, "Causes of death in COVID-19 patients: A literature review," *International Journal of Research in Pharmaceutical Sciences,* vol. 11, no. 1, pp. 1918-1924, 2021.

[8] W. Sanusi, A. Alimuddin and S. Sukmawati, "Model Regresi Cox dan Aplikasinya dalam Menganalisis Ketahanan Hidup Pasien Penderita Diabetes Melitus di Rumah Sakit Bhayangkara Makassar," *Journal of Mathematics, Computations, and, Statistics,* vol. 1, no. 1, pp. 62-77, 2018.

[9] E. A. Jal-Usman, A. A. Nads and D. G. Langamin, "Accelerated Failure Time (Aft) Model: Determining The Factors Associated in the Recovery of Patients Diagnosed with Covid-19," *Advances and Applications in Statistics,* vol. 79, pp. 1-9, 2022. DOI: 10.17654/0972361722056.

[10] M. H. R. Khan and T. Howlader, "Stability Selection for Lasso, Ridge and Elastic Net Implemented with AFT Models, Running Title: Stability Selection for AFT Models," *arXiv:1604.07311v1 [stat.ME] 25 Apr 2016,* 2016.

[11] K. Granville and Z. Fan, "Buckley-James Estimator of AFT Models with Auxiliary Covariates," *PLoS ONE,* vol. 9, no. 8, p. e104817, 2014. DOI: 10.1371/journal.pone.0104817.

[12] M. Pang, R. W. Platt, T. Schuster and M. Abrahamowicz, "Flexible Extension of The Accelerated Failure Time Model to Account for Nonliniear and Time-Dependent Effects of Covariates on The Hazard," *Statistical Methods in Medical Research,* vol. 30, no. 11: 096228022110417, 2021. DOI: 10.1177/09622802211041759.

[13] Rachmaniyah, Erna and Saleh, "Analisis Survival dengan Model Accelerated Failure Time Berdistribusi Log-normal," *Universitas Hasanudin,* 2014. Available at https://core.ac.uk/download/pdf/77620917.pdf.

[14] M. J. Crowther, P. Royston and M. Clements, "A Flexible Parametric Accelerated Failure Time Model and The Extension to Time-Dependent Acceleration Factors," *Biostatistics,* vol. kxac009, 2022.

[15] Sulantari and W. Hariadi, "Analisis Survival waktu sembuh pasien COVID-19 di Kabupaten Banyuwangi," *Transformasi: Jurnal Pendidikan Matematika dan Matematika,* vol. 4, no. 2, pp. 375-386, 2020.

[16] A. F. Majeed, "Accelerated Failure Time Models: An Application in Insurance Attrition," *The Journal of Risk Management and Insurance,* vol. 24, no. 2, pp. 12-35, 2020.

[17] N. Kumar, S.-A. Qi, L.-H. Kuan, W. Sun, J. Zhang and R. Greiner, "Learning Accurate Personalized Survival Models for Predicting Hospital Discharge and Mortality of COVID-19 Patients," Scientific Reports, 12, 4472, 2022.

[18] J. P. Klein and M. L. Moeschberger, Survival Analysis Techniques for Censored and Truncated Data, Second Edition, New York: Springer, 2003.

[19] N. I. Fitriani, "Tinjauan Pustaka COVID-19:Virologi, Patogenesis, dan Manifestasi Klinis," *Medika Malahayati,* vol. 4, no. 3, pp. 194-201, 2020.

[20] D. F. Moore, Applied Survival Analysis Using R., Switzerland: Springer, 2016.

[21] S. P. Khanal, V. Sreenivas and S. K. Acharya, "Accelerated Failure Time Models: An Application in the Survival of Acute Liver Failure Patients in India," *International Journal of Science and Research (IJSR),* vol. 3, no. 6, pp. 161-166, 2012.

[22] G. Rodriguez, Generalized Linear Models, lecture notes, Princeton.edu, 2022.

[23] A. Faruk, "The Comparison of Proportional Hazards and Accelerated Failure Time Models in Analyzing the First Birth Interval Survival Data," *Journal of Physics: Conference Series,* vol. 974 012008, 2018.