Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 16(1) 2023: 67-75
http://ejournal.undip.ac.id/index.php/media_statistika

SURVIVAL ANALYSIS FOR RECURRENT EVENT DATA USING COUNTING PROCESS APPROACH: APPLICATION TO DIABETICS

Triastuti Wuryandari, Yuciana Wilandari
Department of Statistics, Diponegoro University, Indonesia

e-mail: *triastutiwuryandari1@gmail.com*

DOI: 10.14710/medstat.16.1.67-75

Article Info:
Received: 9 October 2023
Accepted: 30 November 2023
Available Online: 7 December 2023

Keywords:
Survival Analysis; Diabetics; Cox Model; Recurrent Event; AG Model

Abstract: Survival analysis is a branch of statistics for analyzing the duration of time until one or more events occur. Time to recurrence of diabetics including survival data. Diabetes can't be cured but it can be controlled. Diabetics who don't maintain their health and lifestyle will experience recurrence. Factors thought to influence the recurrence of diabetics are internal factors such as genetics and external factors such as lifestyle. The recurrence time of an object includes recurrent events because each object can experience the same recurrent event during the follow-up. One of the analysis to determine factors that are thought to influence the recurrence time of diabetics is survival analysis. Survival data can be modeled into a regression model if the survival time of an object is influenced by other factors. One of the regression models for survival data is Cox regression. One of the Cox regression models for recurrent event data is the AG model which uses a counting process approach. This study used data on the recurrence of diabetics at MH Thamrin Cileungsi Hospital. Based on data analysis, factors that influence the recurrence of diabetics are age, gender, and type of complication.

1. INTRODUCTION

Survival analysis is a branch of statistics for analyzing the time until one or more events occur (Collet, 1994). Survival analysis is a statistical procedure used to analyze data with respect to the time until an event occurs. Time starts from the beginning of observation until the occurrence of an event (Kleinbaum & Klein, 2012). One of the aims of survival analysis is to determine the relationship between survival time and variables that are thought to influence survival time. In survival analysis, events can include death, equipment failure, disease recurrence, recovery from a disease, repeated criminal acts, and so on. In survival analysis, it is known that censored data is often found, namely if the time of an object event cannot be observed completely. Several reasons for censored data occur, such as the subject not experiencing the event before the study ended or the subject disappearing during the study (Klein & Moeschberger, 2003; Lee & Wang, 2003; Sudarno, 2022).

Survival analysis aims to determine the relationship between survival time and the factors that influence survival time. One application of survival analysis is the time of recurrence in diabetics. There are three types of Diabetes Mellitus (DM), namely type 1 or Insulin Dependent Diabetes Mellitus (IDDM), type 2 or Non-Insulin Dependent Diabetes

Mellitus (NIDDM), and Gestational Diabetes Mellitus which usually occurs in women during pregnancy. The majority of diabetics are type 2. Type 2 DM sufferers are generally over 45 years old, but currently, there is an increase in the population of type 2 DM sufferers among teenagers and children (World Health Organization, 2019). The number of diabetics in Indonesia is fifth rank in the world after China, India, Pakistan, and the United States (International Diabetes Federation (IDF)). According to the IDF, the prevalence of diabetes in Indonesia is 10.8% with the number of diabetics around 19.5 million people.

Time to event could occur more than once during follow-up for an object. Such an event is called a "recurrent event". The application of the recurrent event model was discussed by Lim & Zhang (2011) used a multiplicative and additive hazard model for emergency department visit data. They concluded that multiplicative and additive hazard models have different information so that the two methods complement each other to gain a more comprehensive understanding. Ullah et al. (2014) with the Anderson Gill approach and frailty approach to sports injury. Sari & Purnami (2015) have applied the recurrent event model using the Anderson and Gill approach to cervical cancer patients at Soetomo Hospital, Surabaya. Factors that influence the recurrence of cervical cancer patients are stadium, complications, and weight. Tampubolon & Nurhayati (2018) applied an identical recurrence model to tuberculosis patients and concluded that the age factor influences the recurrence rate of tuberculous patients. Tsania et.al (2023), discussed the recurrent event model using a counting process approach in patients with asthma exacerbations at the Diponegoro National Hospital. Based on the results of the analysis, factors that influence the length of time a patient experiences an exacerbation are age, gender, and type of case. This study used data on the recurrence of diabetics treated at MH Thamrin Cileungsi Hospital, Bogor, Indonesia. The recurrence model uses the AG model with a counting process approach.

An example of a recurrent event is the recurrence of diabetics. The length of time a diabetic experiences a recurrence can be predicted using a regression model that uses time as the dependent variable. Recurrence in the majority of diabetics is caused by patient behavior that does not comply with and fulfill the recommended therapy, such as restricting diet, doing excessive physical activity, and not controlling blood sugar levels regularly, as well as not recognizing the symptoms of recurrence (Valliyot et al., 2013). Retreatment procedures for diabetes mellitus patients are carried out in the hope of correcting all the factors that cause recurrence and avoiding complications that may occur (Anandarma et al., 2021). The length of time a diabetic experiences a recurrence can be predicted using a regression Cox proportional hazard model.

There are two aims of this paper. The first aim of this research is to find out what factors influence the time of diabetics recurrence. The second aim is to determine the hazard ratio of factors that influence the time of recurrence in diabetics. The combination of the two aims is a novelty offered in this study. A case study using the medical record data of recurrence for type 2 diabetics treated from January to December 2021 at MH Thamrin Cileungsi Hospital, Bogor. The research uses risk intervals, a counting process, and a common baseline hazard. The analysis will begin by forming a Cox model prediction for recurrent event data, testing the proportional hazard assumption, calculating the hazard ratio, and interpreting it.

2. LITERATURE REVIEW

2.1. Survival Analysis

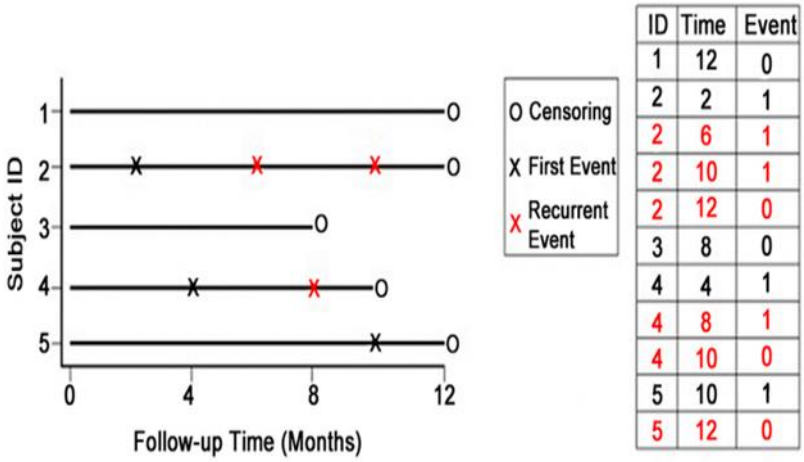
Survival analysis is used to analyze data that is related to time. Survival time starts from the origin time until an interesting event occurs. Events can include the time of death, recurrence, and another event within a certain period. If survival time is influenced by other factors then it can used regression model. One of the regression models for survival data is Cox regression. The difference between survival analysis and other statistical analysis is that there is censored data. Censored data is data that cannot be observed completely because the individual in the study is lost to follow-up, drops out, or terminated (Kleinbaum & Klein, 2012).

There are three functions in survival analysis, namely the probability density function ($f(t)$), hazard function ($h(t)$) and survival function ($S(t)$). The probability density function is the probability that an object experiences an event in the time interval from t to $t + \Delta t$. The hazard function is defined as the probability that an object fails in the time interval from t to $t + \Delta t$, if the object has survived in time t. The survival function is the probability an object survives more than time t. The relationship between the three functions is as follows $h(t) = \frac{f(t)}{S(t)}$.

2.2. Recurrent event

Recurrent event data is where the subject experiences repeated occurrences of the same type of event, for example, repeated asthma attacks, the occurrence of diabetics, etc (Kelly & Lim, 2000). There are two types of recurrent events, namely identical and non-identical recurrent events. Analysis of identical recurrent events using a counting process approach, while for non-identical recurrent event data using a stratified approach (Kleinbaum & Klein, 2012).

Recurrent event data can be modeled by the Cox proportional hazard model. There are five models for recurrent event data based on the Cox regression model: Andersen and Gill (AG); Wei, Lin, and Weissfeld (WLW); Prentice, Williams, and Peterson Total Time (PWP-TT), Prentice, Williams, and Peterson and Gap Time (PWP-GT); and Lee, Wei, and Amato (LWA) (Kelly & Lim, 2000). AG uses a counting process approach to formulate the form of increasing the number of events along a timeline. This analysis uses a common baseline hazard function for all events and estimates global parameters for the factors of interest. AG model assumes that the correlation between the times of events in a person can be explained by past events, which implies that the increase in time between events is conditionally uncorrelated, given the presence of covariates. Analysis of the PWP model based on stratification. PWP uses total time and gap time. PWP-GT evaluates the influence of covariates on events since the previous event. The PWP model may be superior to the AG model when the effects of covariates differ on subsequent events. An Illustration of recurrent event data is presented in Figure 1 and Table 1.



ID	Time	Event
1	12	0
2	2	1
2	6	1
2	10	1
2	12	0
3	8	0
4	4	1
4	8	1
4	10	0
5	10	1
5	12	0

Figure 1. Illustration from Recurrent Event Data

Table 1. Data Layout with Counting Process Approach

Object (i)	Interval (i)	Event (δ_{ij})	Start Time (t_{ij0})	Stop Time (t_{ij1})	Covariate (Z_{ijk})			
					(Z_{ij1})	(Z_{ij2})	…	(Z_{ijp})
1	1	δ_{11}	t_{110}	t_{111}	Z_{111}	Z_{112}	…	Z_{11p}
1	2	δ_{12}	t_{120}	t_{121}	Z_{121}	Z_{122}	…	Z_{12p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
1	r_1	δ_{1r_1}	t_{1r_10}	t_{1r_11}	Z_{1r_11}	Z_{1r_12}	…	Z_{1r_1p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
i	1	δ_{i1}	t_{i10}	t_{i11}	Z_{i11}	Z_{i12}	…	Z_{i1p}
i	2	δ_{i2}	t_{i20}	t_{i21}	Z_{i21}	Z_{i22}	…	Z_{i2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
i	r_i	δ_{ir_i}	t_{ir_i0}	t_{ir_11}	Z_{ir_i1}	Z_{ir_i2}	…	Z_{ir_1p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
n	1	δ_{n1}	t_{n10}	t_{n11}	Z_{n11}	Z_{n12}	…	Z_{n1p}
n	2	δ_{n2}	t_{n20}	t_{n21}	Z_{n21}	Z_{n22}	…	Z_{n2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
n	r_n	δ_{nr_n}	t_{nr_n0}	t_{nr_n1}	Z_{nr_n1}	Z_{nr_n2}	…	Z_{nr_np}

2.3. Basic Notation

There are *n* subjects and each subject can experience *K* recurrent events. Let T_{ik} is the time study of the k^{th} event for the i^{th} subject, C_{ik} is the censoring time of the k^{th} event in the i^{th} subject, and X_{ik} is the corresponding observation time, $C_{ik} = \min(T_{ik}, C_{ik})$. $X_{ik} = T_{ik}$ if the event was observed and $X_{ik} = C_{ik}$ if it is censored. Let $\delta_{ik} = I(T_{ik} \leq C_{ik})$ be the indicator variable where $I(E) = 1$ if E is true and $I(E) = 0$ otherwise. Let $h_{ik}(t)$ denote the hazard function for the k^{th} event of the i^{th} object at time t, $h_0(t)$ represent a common baseline hazard for all events and $h_{0k}(t)$ represent an event-specific baseline hazard for all event k^{th} event. Let $\vec{\boldsymbol{Z}}_{\boldsymbol{ik}} = (Z_{1ik}, \dots, Z_{pik})'$ denote the covariate vector for the i^{th} object with respect to the k^{th} event. $\overrightarrow{\boldsymbol{Z}_{\boldsymbol{i}}} = (Z_{i1}, \dots, Z_{ik})'$ is the covariate vector for the i^{th} object, where K is the maximum number of events within an object and $\vec{\boldsymbol{\beta}} = (\beta_1, \dots, \beta_p)'$ is vector unknown regression parameter.

There are four components for analysis of recurrent events based on the Cox model, namely risk interval, risk set, baseline hazard, and within-subject correlation (Kelly & Lim, 2000).

Risk Interval. Risk intervals are defined when a subject is at risk of having an event along a given time scale. There are three risk intervals: gap time, total time, and counting process formulation. The risk interval determines whether a model is marginal or conditional. Gap time and the counting process are conditional because the subject cannot take risks until the previous end event, that is, the subject is at risk due to previous events. Total time is marginal because the subjects are at risk from the start of treatment and are independent of previous events.

Risk set. The k^{th} risk set contains the individuals who are at risk for the k^{th} event. There are three possible risk sets: unrestricted, restricted, or semi-restricted. The risk set definition incorporates the choice of baseline hazard. The risk set at a given point in time depends on the individuals included in the set and when those individuals are at risk, that is the risk interval.

Baseline Hazard. There are two choices for the baseline hazard function for recurrent event models: common and event-specific. A model with a common baseline hazard has the same underlying hazard for all events. An event-specific baseline hazard is a stratified baseline hazard that allows the baseline hazard to be different for each k^{th} event. Stratifying by event is essentially fitting a separate model for each k^{th} event.

Within-subject correlation. Three approaches have been proposed for accounting for the within-subject correlation between events: conditional; marginal, and random effect. The conditional approach assumes that the current event is unaffected by earlier events that occurred to the subject. This assumption can be relaxed by introducing time-dependent covariates in the model, such as the number of prior recurrences, which may capture the dependence structure among the recurrence times. The marginal approach assumes that the events within a subject are independent and estimates robust variance using a sandwich estimator. The unadjusted variance estimates are called naive estimates. The random effect also called frailty models, introduces a random covariate into the model that induces dependence among the recurrent event times.

2.4. Multiplicative Hazard Model

One of the hazard regression models for recurrent event data is the multiplicative hazard model, known as the Cox model. The multiplicative hazard model has the same assumptions as the Cox model, namely proportional hazard. For the multiplicative hazard recurrent event model for i^{th} individual is

$$h_{ik}(t) = h_o(t)e^{\vec{\boldsymbol{\beta}}' Z_{ik}(t)} \tag{1}$$

where $h_0(t)$ is baseline hazard, $\vec{\boldsymbol{\beta}}$ is the vector regression coefficient, $Z_{ik}(t)$ is covariates.

The parameter estimates in the Cox model for recurrent events are for estimating the baseline hazard and estimating the regression coefficient. Baseline hazard estimation uses partial likelihood. The partial likelihood function is

$$L(\hat{\beta}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{e^{\beta' Z_{ik}(t)}}{\sum_{l\in R_k} Y_{lk}(t)\, e^{\beta' Z_{ik}(t)}}\right)^{\delta_{ik}} \tag{2}$$

where $Y_{lk}(t) = I(T_{ik} \geq t)$ is risk set indicator, $\vec{\boldsymbol{\beta}}$ is p vector of the regression coefficient, $Z_{ik}(t)$ is covariates.

The score function $U(\vec{\boldsymbol{\beta}})$ from model (1), are obtained by differentiating the logarithm of $L(\vec{\boldsymbol{\beta}})$ with respect to $\vec{\beta}$.

$$\log L(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{ik}\left(\overrightarrow{\boldsymbol{\beta}'} Z_{ik}(t) - log \sum_{i=1}^{n} Y_{ik}(t) e^{\overrightarrow{\boldsymbol{\beta}'} Z_{ik}(t)}\right) \quad (3)$$

The maximum partial likelihood estimate $\vec{\beta}$ is obtained by solving the corresponding score equation

$$\frac{\partial \ln L(\vec{\boldsymbol{\beta}})}{\partial \vec{\boldsymbol{\beta}}} = 0 \quad (4)$$

The results of the score equation (4) do not provide a close-form equation then a numerical method is used, namely the Newton-Raphson method. To account for the within-subject correlation, we used this robust "sandwich" method in the estimation of standard errors. The robust estimator developed by Lin & Wei (1989) was used to estimate the covariance matrix by adjusting for within-subject correlations. This can help overcome misspecification of the correlation structure assumed in the analysis of recurrent event data. The use of robust variance estimates is recommended in the analysis of recurrent event data to obtain correct and efficient statistical inferences (Kleinbaum & Klein, 2012).

3. MATERIAL AND METHOD

3.1. Sample Data

There are three types of diabetes mellitus, namely type 1 (insulin-dependent), type 2 (non-insulin-dependent), and gestational diabetes mellitus which usually occurs in women during pregnancy. Type 2 diabetes mellitus is a type of diabetes that is more common than type 1 diabetes and gestational diabetes. Type 2 diabetics are generally aged over 45 years, but currently, there is an increasing population of type 2 diabetics among teenagers and children (World Health Organization, 2019).

The data used is medical record data of inpatients with type 2 diabetes at MH Thamrin Hospital, Bogor, Indonesia. The dependent variable is the time of recurrence of diabetics. The independent variables used are: Age (X_1), Gender (X_2), Type of complication (X_3) and Hypertension (X_4). The "age" variable consists of two categories, namely adults $(X_{11} < 60)$ and elderly $(X_{12} \geq 60)$. The "gender" variable consists of two categories, namely Male (X_{21}) and Female (X_{22}). The "complication type" variable consists of two categories, namely macrovascular (X_{31}) and microvascular (X_{32}). Macrovascular complications are complications in patients with peripheral vascular disease and others; nutritional/metablock disorders; foot procedures; as well as kidney and urinary tract diseases. patients with microvascular complications, namely patients with cranial and peripheral nerve disorders (Yuhema, et al., 2015). The "hypertension" variable consists of two categories, namely patients with a history of hypertension and patients without a history of hypertension

3.2. Statistic Descriptive

In this study, data was taken from medical record data with type 2 diabetes mellitus patients treated at MH Thamrin Hospital Cileungsi, Bogor in January–December 2021. Data characteristic is presented in Table 2.

Table 2. Data Characteristic

No	Variable	Categoric
1	Time of recurrence	Start
		Stop
2	Status	0: censored
		1: observed
3	Age	0: Adult
		1: Elderly
4	Gender	0: Man
		1: Woman
5	Type of complication	0: Macrovasculer
		1: Microvasculer
6	History of hypertension	0: No
		1: Yes

Table 3. Number of Recurrence Diabetic

Number of events	1	2	3	Total
Number of subjects (%)	29 (78.4%)	6 (16.2%)	2 (5.4%)	37(100%)

Data analysis consists of the proportional hazard assumption test, the significance parameter tests, and the model interpretation

4. RESULTS AND DISCUSSION

The first step in data analysis is the proportional hazard test. Data processing using R software (R Core Team, 2021), which output for the proportional hazard assumption test is described in Table 4.

Table 4. Test for Proportional Hazard Assumption

Variable	chisq	df	p-value
Age	1.336	1	0.248
Gender	1.007	1	0.316
Type complication	0.297	1	0.586
History hypertension	2.826	1	0.093
Global	5.268	4	0.261

Based on Table 4, it is concluded that all variables have the proportional hazard assumption because the p-value from all variables has a p-value > 0.05. The output of the significant parameter test is described in Table 5.

Table 5. Multiplicative Hazard Model for Recurrent Event with A Varying Baseline and Common Coefficient Effect

Model	Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	Robust SE	Z	p-value
Multiplicative	Age (X_1)	0.5845	0.3039	0.1876	3.12	0.00184**
	Gender (X_2)	0.5369	0.3341	0.2408	2.23	0.02577*
	Complication (X_3)	0.4871	0.3239	0.2023	2.41	0.01606*
	Hypertension (X_4)	0.2200	0.3319	0.2354	0.94	0.35002

Likelihood ratio test=3.87 on 1 df, p=0.04919

Based on Table 5, the multiplicative recurrent event hazard model is as follows

$$h_{ik}(t) = h_0(t)\exp(0.5845X_1 + 0.5369X_2 + 0.4871X_3 + 0.2200X_4)$$

The model is suitable to use because the $LR > \chi^2_{(0,05;4)}$ or p-value < 0.05

Based on Table 5, variables that influence the recurrence of diabetes are age, gender, and type of complication. Interpretation of the model is as follows, the Hazard Ratio for the variable "age" is $e^{0.5845} = 1.794$, this means, elderly patients have an increased risk 1.794 times compared with adult patients. The Hazard Ratio for the variable "Gender" is $e^{0.5369} = 1.711$ this means, that male patients have an increased risk 1.711 times compared with woman patients. The Hazard Ratio for the variable "Type of Complication" is $e^{0.4871} = 1.628$, this means, patients with microvascular complications have an increased risk 1.794 times compared with adult patients.

In this research, each object has a certain number of time intervals doesn't have to be the same. Start and stop times may be different for different objects. This study used the start and stop times for each line of data for a particular object according to the data layout for the counting process approach while the previous research used gap time (Kelly & Lim, 2000; Lim & Zhang 2011).

5. CONCLUSION

Data recurrence of diabetics can be modeled using the Cox model because all variables fulfill the proportional hazard assumption. Based on data analysis, it was concluded that the variable that influences the duration of diabetes recurrence is the age, gender, and type of complication. The Cox hazard recurrent model is as follows $h_{ik}(t) = h_0(t)\exp(0.5845X_1 + 0.5369X_2 + 0.4871X_3 + 0.2200X_4)$. The Cox model measures this excess risk in relative terms and the relative risk may be constant over time. By knowing the Hazard Ratio, you can know the ratio of an object experiencing an event from each influencing variable. Assuming that the data on the recurrence of diabetics is an identical recurrent event, then the analysis uses the counting process approach.

REFERENCES

Anandarma, S.O., Asmaningrum, N., & Nur, K.R.M. (2021). Hubungan Efikasi Diri Pasien Diabetes Mellitus Tipe 2 dengan Risiko Rawat Ulang di Rumah Sakit Umum Daerah Dr. Harjono Kabupaten Ponorogo. *Jurnal Keperawatan Sriwijaya, 8*(2), 39–49.

Andersen, P.K., Borgan, O., Gill, R.D., & Keiding, N. (2012). *Statistical Models Based on Counting Processes*. New York: Springer Science & Business Media.

Collett, D. (2015). *Modelling Survival Data in Medical Research* 3rd ed. New York: Chapman and Hall.

Kelly, P.J. & Lim, L. (2000). Survival Analysis For Recurrent Event Data: Application to Childhood Infectious Diseases. *Statistics in Medicine*, *19*, 13–33.

Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis Techniques Truncated Data,* 2nd ed. New York: Springer Science.

Kleinbaum, D.G. & Klein, M. (2012). *Survival Analysis A Self Learning Text*, 2nd ed. New York: Springer Science.

Lee, E. T. & Wang, J. (2003). *Statistical Methods for Survival Data Analysis* (Vol. 476). USA: John Wiley & Sons.

Lim, H.J. & Zhang, X. (2011). Additive and Multiplicative Hazard Modeling for Recurrent Event Data Analysis. *BMC Medical Research Methodology*, *11*, 1–12.

R Core Team. (2021). R: A Language And Environment For Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Sari, N.W. & Purnami, S.W. (2015). Survival Analysis for Recurrent Event Data with Anderson Gill Approach. *Proceedings of the ICONSSE FSM SWC*, 51–54

Scaubel, D.E. & Cai, J. (2004). Regression Methods for Gap Time Hazard Functions of Sequentially Ordered Multivariate Failure Time Data. *Biometrika*, *91*. 291–303

Sudarno, S. & Setiani, E. (2019). Hazard Proportional Regression Study to Determine Stroke Risk Factors Using Breslow Method. *Media Statistika*, *12*(2), 200–213

Tampubolon, R & Noeryanti. (2018). Model Regresi Cox pada Data Kejadian Berulang Identik untuk Analisis Penyakit Tuberkulosis Terhadap Pasien Laki-laki. *Jurnal Statistika Industri dan Komputasi*, *3*(2), 33–41.

Tsaniya, U., Wuryandari, T., & Ispriyanti, D. 2023. Analisis Survival pada Data Kejadian Berulang Menggunakan Pendekatan Counting Process. *Jurnal Gaussian*, 11(3), 377–385.

Ullah, S., Gabbet, S., & Finch, C.F. (2014). Statistical Modelling for Recurrent Events: an Application to Sports Injuries. *BMJ Sport*. *Br, J. 48*(17), 1287–1293.

Valliyot, B., Sreedharan, J., Muttappallymyalil, J., & Valliyot, S.B. (2013). Risk Factors of Type 2 Diabetes Mellitus in The Rural Population of North Kerala, India: A Case Control Study. *Diabetologia Croatica*, *42*(1).

Yuhelma, Hasneli, Nauli. (2015). Identifikasi dan Analisis Komplikasi Makrovaskuler dan Mikrovaskuler pada Pasien Diabetes Mellitus. *Jurnal Online Mahasiswa Program Studi Ilmu Keperawatan Universitas Riau*, 2(1), 569–579.