Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Survival analysis for diabetic retinopathy in diabetic patients using Extended Cox Model

To cite this article: S D Permai and B J Randa 2021 *J. Phys.: Conf. Ser.* **1988** 012113

View the article online for updates and enhancements.

Survival analysis for diabetic retinopathy in diabetic patients using Extended Cox Model

S D Permai[1] **and B J Randa**[2]

[1,2]Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

[1]**E-mail**: syarifah.permai@binus.ac.id

Abstract. One of the disorders of the eye that may occur to the diabetes patients is diabetic retinopathy. Diabetic retinopathy can cause vision loss and even blindness to the diabetes patients. At first, diabetic retinopathy may not have any symptoms at all. But diabetic retinopathy is the disorder which cause of blindness besides of cataracts, glaucoma and macular degeneration. The objective of this research was to determine the factors that can be used to delay diabetic retinopathy. There are several covariates that used in this research. One of the survival analysis method that can be used is Cox Proportional Hazard model. However, in this research, the result of Cox Proportional Hazard model showed that some covariates are not significant in the Cox Proportional Hazard model and the model did not fulfil the assumption, that is a constant hazard ratio over time. Therefore, an analysis was carried out using the Extended Cox Model. The results of Extended Cox model showed that all variables are concede a significant effect in delaying diabetic retinopathy in the Extended Cox model. Therefore, the Extended Cox model is more appropriate in this research than Cox Proportional Hazard model.

1. Introduction

One of the main causes of blindness in productive age is Diabetic Retinopathy, which is a retinal disorder in diabetes mellitus patients [1]. Wisconsin Epidemiology Study of Diabetic Retinopathy (WESDR) reports that 99% of type 1 diabetes mellitus patients and 60% of type 2 diabetes mellitus patients will experience diabetic retinopathy within 20 years [1]. Therefore, it is important to do early detection of diabetic retinopathy, in order to reduce the risk of severe vision loss. The way that can be done to prevent blindness is to control blood sugar and periodic eye examinations according to the degree of diabetic retinopathy [2].

The incidence of diabetic retinopathy increases with the duration of the disease and the age of the patient [1]. Beside of that, severity of hyperglycemia, other important factors include hypertension and elevated serum lipid levels also affect the increase in diabetic retinopathy [1] [2]. At first, diabetic retinopathy often shows only mild symptoms, or no symptoms at all. However, if there is no treatment, diabetic retinopathy can cause blindness [3]. There are several studies about diabetic retinopathy. Classification model using logistic regression with Least Absolute Shrinkage and Selection Operator (lasso) was performed to predict diabetic retinopathy based on patient characteristics and biochemical measures [4]. Furthermore, Kumar and Jayalalitha (2019) used Mathematical modelling to define diabetic retinopathy in eyes based on lacunarity [5]. Research of diabetic retinopathy in South Africa was conducted on type 1 and type 2 diabetes patients to examine the influence of glycaemic control and ethnic variations on the incidence and progression of diabetic retinopathy. Statistical methods that used in their research were Kaplan-Meier estimator, Log rank test and Cox Proportional Hazard regression model [6].

In this research, survival data analysis was used to determine the factors that influence the prevention of diabetic retinopathy. One of the methods that can be used in survival analysis is the Cox Proportional Hazard regression model. However, this model requires the assumption that the hazard ratio must be constant over time. If these assumptions are not fulfilled, an alternative method is needed, one of the alternative models is the Cox Extended regression model, which is the Cox model involving time dependent variables [7]. Incorrect models can occur because of the impact of violating the proportional hazard assumption in the Cox Proportional Hazard regression model [8]. Hence, the assumption is the limitation of Cox proportional Hazard regression model. If the assumptions of the Cox Proportional Hazard regression model are not fulfilled, then survival modelling with an alternative model must be carried out. Several survival modelling studies using the extended cox model include estimating survival time of breast cancer in Makassar [9], clinical improvement of leprosy patients in Lamongan [10], and research of infection that occurs in kidney patients who have catheters installed [11]. In this research, a survival analysis was performed on diabetic retinopathy data using the Cox Proportional Hazard regression model due to the assumption test on the Cox Proportional Hazard regression model is not fulfilled, an alternative modelling is carried out using the Extended Cox model.

2. Survival Analysis

Survival analysis is a statistical method for data analysis where the outcome variable is the time until an event occurs [7]. Then, there are two variables, namely survival time (T) and status (δ). Survival time refer to the time until an event occurs or that an individual has survived over some period time. Status refers to an event has occurred or not. If the event has occur then it called failure ($\delta = 1$), if not occur then it called censored ($\delta = 0$). Generally, there are three reasons censoring may occur including event do not occur until the study ends, lost to follow up and withdraws from the study.

There are two quantitative term in survival analysis, namely survivor function and hazard function. The survivor function denoted by S(t) is the probability of a person that survive longer than a specific time *t*. Theoretically, when time t = 0 then S(t) = S(0) = 1 and at time t = ~ then S(t) = S(~) = 0. The hazard function denoted by $h(t)$ is a rate of the event to occur if the person survived up to time t.

3. The Cox Proportional Hazard Model

The Cox Proportional Hazard model is a semiparametric model that usually used in survival analysis. It is popular due to its robustness. It means that the Cox Proportional Hazard model result will closely approximate the result of the parametric model [7]. The Cox Proportional Hazard model is a product of baseline hazard involving t and exponential expression involving independent variables. Let say there are p independent variables, then the general Cox Proportional Hazard model as follows.

$$h(t, \boldsymbol{X}) = h_0(t) \exp\,[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p] \tag{1}$$

where $h_0(t)$ is the baseline hazard function at time t, β_i is the parameter of Cox Proportional Hazard model, X_i is the explanatory variable or independent variable, and p is the number of independent variables, where $i = 1, 2, \dots, p$.

The parameters (β_i) in the model are estimated using the Maximum Likelihood (ML) estimation. The formula for the Cox model likelihood function is used a partial likelihood because the likelihood formula only considers failure status, do not considering the probability of censored status. The maximization process is carried out by taking partial derivatives of natural log of partial likelihood with respect to each parameter in the model equal by zero. Then use iteration to obtain the parameters (β_i). The Cox Proportional Hazard model can be interpreted using the hazard ratio (HR). HR is a measure used to determine the level of risk which can be seen from the comparison between individuals with different conditions of the independent variable (X) [10]. HR formula in terms of the regression coefficients by substituting the Cox model formula into the numerator and denominator of the hazard ratio expression. The estimated hazard ratio ($\widehat{HR}$) was computed by exponentiating the coefficient of a (0, 1) independent variable of interest as written equation.

$$\widehat{HR} = \frac{h(t,\boldsymbol{X}^*)}{h(t,\boldsymbol{X})} = \exp\,[\textstyle\sum_{i=1}^{p} \hat{\beta}_i ({X_i}^* - X_i)] \tag{2}$$

Suppose that p = 1 (that means there is one explanatory variable) and (0, 1) independent variable of interest then the hazard ratio obtained by letting ${X_1}^* = 1$ and $X_1 = 0$ in the hazard ratio formula as follows.

$$\widehat{HR} = \exp\,[\hat{\beta}_1({X_1}^* - X_1)] = e^{\widehat{\beta}_1} \tag{3}$$

An important thing in the Cox Proportional Hazard model that the baseline hazard is a function of *t*, but does not involve independent variables (X), whereas the exponential function is involved independent variables (X) but does not involve *t*. Then the Cox Proportional Hazard model assume that the hazard ratio is constant over time. The Goodness of fit (GOF) testing approach is a test that can be used to evaluate the Cox Proportional Hazard model are met.

4. The Extended Cox Model

The Extended Cox model is an alternative of the Cox Proportional Hazard model. This model can be used if the Cox Proportional Hazard model do not satisfy the assumption. The Extended Cox model is used if the independent variable is time dependent variable. A time dependent variable is defined as any variables whose value for a variable may differ over time [7]. The time dependent predictor variable must be interacted with the time function $g\ (t)$. The time function used can use *t*, *ln t* and other functions that contain t [10]. Let say there are p independent variables and q time-dependent variables, then the general Extended Cox model as equation 4.

$$h(t, \boldsymbol{X}) = h_0(t)\exp\,[\textstyle\sum_{i=1}^{p} \beta_i X_i + \sum_{j=1}^{q} \delta_j X_j g_j(t)] \tag{4}$$

The parameters in the Extended Cox model can be estimate using Maximum Likelihood estimation. The likelihood function as follows [12].

$$\ln L(\boldsymbol{\beta}) = \textstyle\sum_{k=1}^{n} d_k \left\{\boldsymbol{\beta}'\boldsymbol{x}_k(t_k) - ln \sum_{l \in R(t_{(k)})} \exp\,(\boldsymbol{\beta}'\boldsymbol{x}_l(t_k))\right\} \tag{5}$$

The maximization process is carried out by taking partial derivatives of equation 5 with respect to each parameter ($\boldsymbol{\beta}$) in the model equal by zero. Then use Newton Raphson iteration to obtain the parameters. To interpret the Extended Cox model that can be used a hazard ratio (HR). The general hazard ratio formula for the Extended Cox model as follows [7].

$$\widehat{HR}(t) = \frac{h(t,\boldsymbol{X}^*(t))}{h(t,\boldsymbol{X}(t))} = \exp\,[\textstyle\sum_{i=1}^{p} \hat{\beta}_i ({X_i}^* - X_i) + \sum_{j=1}^{q} \delta_j ({X_j}^* g_j(t) - X_j g_j(t))] \tag{6}$$

Hazard ratio formula involves differences in the values of the time dependent variables at time *t*, this hazard ratio is a function of time. Thus, in general, the Extended Cox model does not satisfy the Cox Proportional Hazard assumption if any δ_j is not equal to zero.

5. Data and Variables

The data in this research is an open data developed by Vincent [13]. The sample consisted of 197 patients. There are several variables in this research, including outcomes variable (survival time and status) and independent variables. Survival time in this research is the actual time to vision loss in months, minus the minimum possible time to event (6.5 months). Censoring was caused by death, dropout, or end of the study. The event of interest was loss of vision in the eye. The independent variables that used in this research are the variables may influence the occurrence of diabetic retinopathy, such as type of laser, the eye, age and risk score, which shown in table 1.

Table 1. Outcomes and Independent Variables.

Variable	Name of Variable	Description
X1	Laser	Type of laser used: Xenon or Argon
X2	Eye	Which eye which treated: Right or Left
X3	Age	Age at diagnosis of diabetes
X4	Type	Type of diabetes: Juvenile (diagnosis before 20) or Adult (after 20)
X5	Risk	A risk score for the eye
T	Survival Time	Time to loss of vision
δ	Status	0 = censored 1 = loss of vision in this eye

6. Results and Discussion

A log rank test was performed on categorical data before the modelling was performed to the data. There are three categorical variables namely Laser (X1), Eye (X2) and Type (X4). The result of log-rank test showed in table 2.

Table 2. The Log-rank test.

Variable	Log-rank test	p-value
Laser (X1)	4.2	0.04
Eye (X2)	0.6	0.4
Type (X4)	0.3	0.6

The hypothesis of the log-rank test as follows.
H$_0$: There is no difference in the Kaplan-Meier survival curve between different groups.
H$_1$: There is difference in the Kaplan-Meier survival curve between different groups.

Based on the table 2 result of the log-rank test showed that p-value of Eye (X2) and Type (X4) greater than $\alpha = 10\%$. It means that there is no difference in the Kaplan-Meier survival curve between right and left eye. It also means that there is no difference in the Kaplan-Meier survival curve between juvenile and adult. The result of Laser variable (X1) showed that p-value < 0.1, it means that there is difference in the Kaplan-Meier survival curve between Xenon and Argon. The next step is modelling the diabetic retinopathy data using the Cox Proportional Hazard model. Table 3 showed the estimation of parameters of the Cox Proportional Hazard model and testing the significance of the parameters. Based on the parameter estimation results in table 3, the Cox Proportional Hazard model is obtained as follows.

$$\hat{h}(t,\boldsymbol{X}) = \hat{h}_0(t)\exp\left[-0.44265\,\mathrm{X1} + 0.21588\,\mathrm{X2} - 0.00592\,\mathrm{X3} - 0.353\,\mathrm{X4} + 0.18338\,\mathrm{X5}\right] \quad (7)$$

Based on the Cox Proportional Hazard model in equation 7, all the parameters in the model must be tested for parameter significance. The hypothesis of testing the significance of the parameters are follows.
H$_0$: $\beta_i = 0$
H$_1$: $\beta_i \neq$

Table 3. The cox proportion hazard model.

Variable	Coefficient	Z test	p-value
Laser (X1) [Xenon]	-0.44265	-1.938	0.0526
Eye (X2) [Right]	0.21588	0.988	0.3232
Age (X3)	-0.00592	-0.432	0.6655
Type (X4) [Juvenile]	-0.35300	-0.813	0.4164
Risk (X5)	0.18338	2.338	0.0194

Table 3 showed that the p-values of X2, X3 and X4 are greater than $\alpha = 10\%$. It means that Eye, Age and Type variables are not significance in the Cox Proportional Hazard model. However, the p-value of Laser and Risk variables are less than 0.1. It means that Laser variable and Risk variable are significance in the Cox Proportional Hazard model. Hence, it is not the best model

Table 4. Assumption testing using goodness of fit (GOF).

Variable	Correlation (ρ)	Chi-square test	p-value
Laser (X1) [Xenon]	-0.0211	0.0467	0.8289
Eye (X2) [Right]	0.0146	0.0235	0.8781
Age (X3)	0.0679	0.4803	0.4883
Type (X4) [Juvenile]	0.0440	0.1844	0.6676
Risk (X5)	-0.1925	4.0305	0.0447

The assumption testing using Goodness of fit (GOF) showed in table 4. Based on the result of assumption testing, it showed that the p-value of Risk variable is 0.0447. This p-value is less than 0.1. It means that the assumption of proportional hazard does not met for Risk variable. However, Laser, Eye, Age and Type variables are fulfilled the proportional hazard assumption. Because the p-value of that variables are greater than $\alpha = 10\%$. Because not all variables meet the proportional hazard assumption, it is necessary to do alternative modelling using the Extended Cox model. Table 5 showed the Extended Cox model with the time function $g(t)$ is an interaction of Risk variable and survival time.

Table 5. The extended cox model.

Variable	Coefficient	Z test	p-value
Laser (X1) [Xenon]	0.175635	0.726	0.468
Eye (X2) [Right]	-0.11352	-0.469	0.639
Age (X3)	-0.005689	-0.383	0.702
Type (X4) [Juvenile]	-0.586924	-1.181	0.238
Risk (X5)	1.135427	8.208	2.25×10^{-16}
Risk $\times\ t$	-0.060995	-9.862	$< 2 \times 10^{-16}$

Based on the parameter estimation results in table 5, the Extended Cox model is obtained as follows.

$$\hat{h}(t,\boldsymbol{X}) = \hat{h}_0(t)\exp\left[0.176\text{ X1} - 0.11\text{ X2} - 0.00567\text{ X3} - 0.587\text{ X4} + 1.135\text{ X5} - 0.06\text{ X5 t}\right] \quad (8)$$

All parameters in the Extended Cox model (equation 8) need to be tested for the significance of the parameters. The hypothesis of testing the significance of the parameters as follows.

H_0 : $\beta_i = 0$
H_1 : $\beta_i \neq 0$

Based on the results in table 5 showed that Laser, Eye, Age and Type variables are not significance in the model. Because the p-values of the variables are greater than $\alpha = 10\%$. But the p-values of the Risk variables less than 0.1. It means that Risk variable and the interaction of Risk and survival time are significance in the model.

The Likelihood Ratio Test (LRT) was carried out to compare the Cox Proportional Hazard model (in equation 7) and the Extended Cox model (in equation 8). The hypothesis of testing as follows.

H_0 : Model 1 is The Cox Proportional Hazard Model
H_1 : Model 2 is The Extended Cox Model

The result of likelihood ratio test showed that test statistics is 301.62 and p-value is 2.2 x 10^{-16}. Under 0.1 significance level, the null hypothesis is rejected and conclude that we must choose the Extended Cox model in this research.

7. Conclusion

Diabetic retinopathy is a retinal disorder in diabetes mellitus patients. Diabetic retinopathy is a complication of diabetes patient that causes blockage of the blood vessels in the retina of the eye. In diabetics, high blood sugar levels will slowly clog these blood vessels, so that blood intake to the retina decreases and can cause diabetic retinopathy. Therefore, it is important to perform the diabetic retinopathy analysis in diabetic patients to prevent blindness. Based on the results and discussion in section 6, the result of log-rank test showed that there is difference in the Kaplan-Meier survival curve in the Laser variable between Xenon and Argon. The Cox Proportional Hazard model was performed to the data. But the model is not good enough because the assumption of proportional hazard is not fulfilled in the model. Then the Extended Cox model is used as an alternative model. The results of the extended cox model showed that not all independent variables are significant in the model. However, the Likelihood Ratio Test (LRT) showed that the Extended Cox model better than the Cox Proportional Hazard model. In future studies, it may be possible to add independent variables that are affect diabetic retinopathy. Because some independent variables are not significance in the model at this research, then that more independent variables that used in the model to get a better model. It also possible to use other survival model such as the parametric survival model.

Acknowledgments

This work is supported by Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's International Research Grant entitled Rainfall Modeling to Prevent Flooding in Jakarta using Machine Learning Method with contract number: No.026/VR.RTT/IV/2020 and contract date: 6 April 2020

References

[1] Elvira, Suryawijaya, E. E. 2019. Retiopati Diabetes. *Continuing Medical Education (CDK)* Vol. 46, No.3, pp. 220-224.

[2] Garg, S., Davis, R. M. 2009.Diabetic retinopathy screening update. *Clinical Dibetes.* Vol.27, No.4, pp. 140-145.

[3] Willy, T. 2019. Retinopati Diabetik. Alodokter. Kementrian Kesehatan Republik Indonesia. Accessed February 1, 2021. [https://www.alodokter.com/retinopati-diabetik].

[4] Kotsiliti, E., Al-Diri, B., Ye, X., Habib, M., Hunter, A. 2017. A classification model for predicting diabetic retinopathy based on patient characteristics and biochemical measures. *Journal for Modeling in Ophthalmology,* Vol. 4, pp. 69-85.

[5] Kumar, G. J., Jayalalitha, G. 2019. Mathematical Modelling of Diabetic Retinopathy in Eyes based on Lacunarity. *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 9, no. 2, pp. 3337-3340

[6] Thomas, R., Distiller, L., Luzio, S., Melville, V., Chowdhury, S. R., Kramer, B., Owens, D. 2015. Incidence and progression of diabetic retinopathy within a private diabetes mellitus clinic in South Africa. *Journal of Endocrinology, Metabolism and Diabetes of South Africa,* vol. 20, no. 3, pp. 127-133.

[7] Kleinbaum, D. G., Klein, M. 2005. *Survival Analysis: A Self-Learning Text.* New York: Springer.

[8] Babińska, M., Chudek, J., Chełmecka, E., Janik, M., Klimek, K., Owczarek, A. 2015. Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome. *STUDIES IN LOGIC, GRAMMAR AND RHETORIC,* vol. 43, no. 56, pp. 33-48

[9] Husain, H., Thamrin, S. A., Tahir, S., Mukhlisin, A., Mirna, M. A. 2018. The Application of Extended Cox Proportional Hazard Method for Estimating Survival Time of Breast Cancer. *IOP Conf. Series: Journal of Physics,* vol. 979, no. 012087, pp. 1-8.

[10] Nurfain, Purnami, S. W. 2017. Analisis Regresi Cox Extended pada Pasien Kusta di Kecamatan Brondong Kabupaten Lamongan. *JURNAL SAINS DAN SENI ITS,* vol. 6, no. 1, pp. D94-D100.

[11] Aini, I. N. 2011. *Extended Cox Model Untuk Time-Independent Covariate Yang Tidak Memenuhi Asumsi Proportional Hazard Pada Model Cox Proportional Hazard*. Universitas Indonesia, Depok.

[12] Collett, D. 2003. *Modelling Survival Data in Medical Research*. Boca Raton: Chapman & Hall/CRC.

[13] Arel-Bundock, V. 2020. Available Dataset. GNU General Public License. Accessed 6 June 2020. [https://vincentarelbundock.github.io/Rdatasets/csv/survival/retinopathy.csv.]