Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 11, 2022, page 82 - 91

The Application of Cox Regression Model in Covid-19 Survival (Vaccination Status and Age)

Nurin Ainun Miza Mohd Dzahir, Noraslinda Mohamed Ismail*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: noraslinda@utm.my

Abstract
Survival analysis is a set of statistical processes for analysing data in which the outcome variable of interest is the amount of time before an event happens. An event is referred to be a failure in survival analysis because the majority of the event may be mortality, illness incidence, or a variety of other undesirable occurrences. This study addresses censoring, which involves the patients' incomplete period of survival. The Cox proportional hazard model, often known as a semiparametric model, is the most commonly used regression for survival study data. This model has fewer assumptions than the parametric model but more than the nonparametric model. However, no assumptions are made on the baseline hazard function, which contradicts the parametric model. The purpose of this study is to compare the rate of survivability between vaccination status and age whether it gives effect to participants' survival time. Next, we compare the estimated Cox Regression coefficient by utilizing SPSS software and R Software language. At the end of the study, the interpretation of the results revealed that each software gives equal interpretation where the vaccination status gives effect to the survival time and the age does not gives effect to the survival time of the participants.
Keywords: Survival analysis, Cox Regression, Censoring

1. Introduction

According to World Health Organization (WHO), at least 13 different vaccines have been administered that are in use such as Pfizer, AstraZeneca, Moderna, Sinovac-CoronaVac and many more. Coronavirus disease 2019 (COVID-19) is becoming increasingly dangerous due to the emergence of variants. Rapid herd immunity via vaccination is required to prevent mutation and the emergence of variants that can completely evade immune surveillance. However, there were also populations that refused to take vaccinations due to some reason whether they were on medical terms such as allergies to vaccine or not. It is so widespread that hospital beds are not available for all patients that are in needed. As a result, most hospitals accept patients with a high chance of recovery.

From this case, failure time, survival time, or event time is declared as the response variable. The responses are the time until an adverse-events, time until immunisation, and COVID-19 mortality, where the event is usually continuous. However, it can only be determined partially due to common reasons such as no event before the end of a study and failed follow-up. Censored data refers to data for incompletely observed responses [1]. [2] discovered that expressing survival functions when there is a variable influencing survival time is a common issue in survival analysis. This variable, also referred to as covariates, is a predictor variable that can be quantitative, qualitative, time-dependent, or time-independent. In this case, the time-dependent covariates are determined in order to estimate their survival effects.

The Cox regression method is a statistical method that is frequently used in medical research to predict the survival time for various patients. The Cox Regression method is used to predict the degree of effect of various features on survival, which is referred to as the hazard rate. The Cox regression method is an example of a semi-parametric model. Cox regression is used to predict each vaccine recipient's chance of survival. A threshold is used to assess the model's accuracy. As a result, recipients with a probability of survival greater than the threshold are the most likely to survive, while recipients with a probability of survival less than the threshold are the least likely to survive.

This research is aims to compare the effect of the covariates to participants' survival time and to compare the estimated Cox Regression coefficient. This research goal is using the partial likelihood model in frequentist approach and utilizing statistical software, R software and SPSS. The analysis of SPSS and R software results forecasts the effects of covariates involved to the survival time of the participants. The findings of this investigation may help give a better understanding of the Covid-19 survival rate.

2. Literature Review

2.1. Survival Analysis

2.1.1. Survival Function

A survival function is denoted mathematically by S, which is obviously a function of time. The survival function may alternatively be defined as the likelihood that an object of interest will live beyond a given period t.

The survival function can be represented as equation (2.1):

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

where:

$T =$ random lifespan drawn from the population

$S(t)$ = survival function

The Survival Function is also known as the survivor function or the dependability function.

2.1.2. Cox Proportional Hazard (PH) Model

The Cox PH model is a well-known and widely used regression technique in survival analysis that is used to investigate the impact of multiple variables at the same time when a specific event occurs. The type of regression model in this study is semi-parametric. It was introduced by [3] as a method of estimating parameters in the model without relying on the baseline hazard, where the results of the regression are not affected by improving regression coefficients, hazard ratios, and adjusted survival curve estimations.

The survival function has been defined by as [4]:

$$S(t|X_i) = h_0(t)\exp\{X_i\beta_i\} \tag{2}$$

Where

$h_0(t)$ is a baseline hazard function.

The Cox PH model has a crucial assumption which is the proportional hazard (PH) assumption. That is, the impact of the hazard function indicators does not vary over time.

This model consists of an exponential function which makes sure that the estimated hazard value is non-negative since the values of the fitted hazard must lie within the positive range [5]. The model is as below:

$$HR = exp\ [\sum_{i=1}^{p} \beta_i(X_A - X_B)] \tag{3}$$

2.1.3. Hazard Ratio

The hazard ratio is the slope of a survival curve and it helps to measure on how rapidly the subject are dying. It is calculated by using all of the data in the survival curve, rather than just one time point. When there is only one hazard ratio presented, it can only be evaluated if the population hazard ratio remains constant over time and any deviations are attributable to random sampling [6]. The hazard ratio may be used to calculate the risk of a covariate impacting survival.

2.1.4. Wald Test

It is a parametric statistical measure used to determine if a set of independent variables is "significant" for a model. It is also used to confirm whether or not each independent variable in a model is significant.

The Wald test is used to derive inferences about the unobserved true value of a parameter in a statistical model that represents the connection of data characteristics and where parameters are estimated from a sample. If the Wald test indicates that the parameters for specific explanatory variables are zero, the variables can be removed from the model. However, if the test results reveal that the parameters are not zero, we should incorporate the variables in the model [7].

2.1.5. Breslow Method

The Cox proportional hazards (PH) model is based on the assumption that the hazard function at time t for a given covariate vector is the product of an arbitrary baseline hazard function and an exponential function of the covariate linear combination. It may be written as follows:

$$\lambda(t|X) = \lambda_0(t)\exp\,(\beta^T X) \tag{4}$$

where $\lambda_0(t)$ is the baseline hazard function. Since the baseline hazard function is left unspecified, the Cox PH model is semiparametric. The Breslow estimator may be used to calculate the baseline survival function, $S_0(t)$.

3. Methodology

3.1. Research Data

The data that has been used in this chapter is obtained from Centers for Disease Control and Prevention (CDC). In this study, 52 randomly data were selected upon participants' vaccination status and their age. The data consist of censoring data where 0 is considered as alive and 1 is considered as died. Table 1.1 shows the variables involved in this study. The data consist of censoring data where 0 is considered as alive and 1 is considered as died. Table 1.1 shows the variables involved in this study.

3.2. Survival Model

This research is one of the semiparametric regression models. The type of modelling utilized is determined by the amount of information available about the form of the connection between the response variable and the explanatory factors, as well as the random error distribution.

This model has its characteristics in graph. In theory, the function of the model may be graphed as a smooth curve as t approaches infinity. The graph descends from $S(t) = 1$ at time $t = 0$ to $S(t) = 0$ at time $t = \infty$. $S(t) = S(\infty) = 0$ when time is infinite. It indicates that we must establish a time limit for the research period, or else no participants will survive as the study term lengthens.

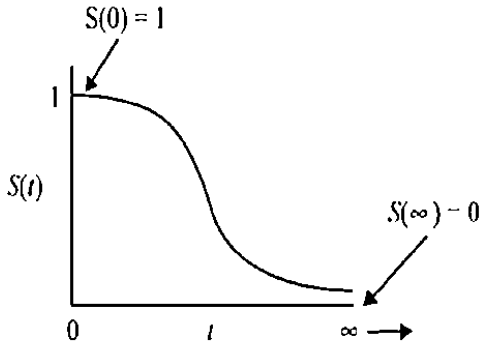

Figure 1 Theoretical survival function graph

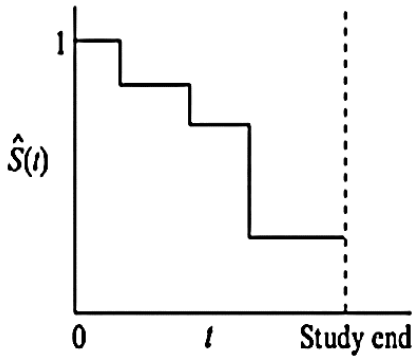

Figure 2 Practical survival function graph

Figure 1 depicts an example of theoretical survival graph. According to both figures, at the start of the research, the probability is one, indicating that all patients are alive, and as t increases to ∞, the likelihood decreases to zero, implying that no one survived at this point.

The step function graph is seen in Figure 2. Since it uses real data, the step function is not as smooth as the theoretical graph. The study time cannot be extended till ∞ since, in real life, there is always a time restriction for a study period. However, the meaning for both graphs is the same despite the difference in the graph shape.

3.3. Parameter Estimation of via Frequentist Approach

The Cox model may be expressed as:

$$h(t, X) = \exp\{\beta_1(X_1) + \beta_2(X_2)\} \quad (5)$$

where $h(t, X)$: hazard function
X_1 : vaccinated population with Covid-19 vaccine
X_2 : unvaccinated population

The test of hypotheses for the regression coefficient, as follows:

$H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ for at least one i

The *p-value* of the variables obtained determines whether the null hypothesis, H_0, is rejected or accepted. Its purpose is to determine the significance of the parameters of the factors that influence the survival time.

3.3. Censoring

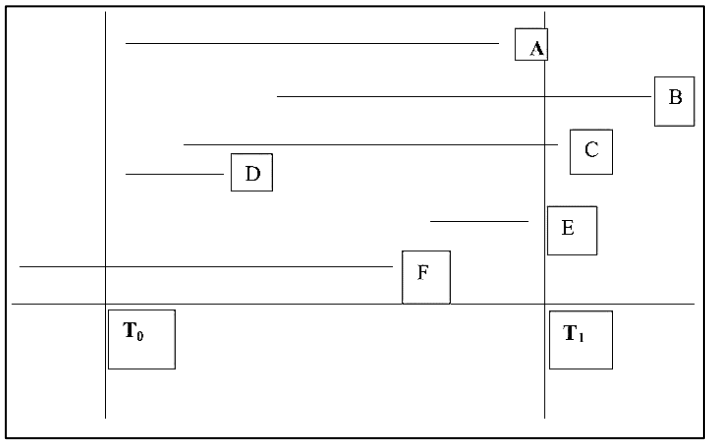

Figure 3: Illustration of censoring event

Censoring is a practically universal aspect of survival data, with right-censoring being the most common type. Before the event occurs, the period of observation expires, or an individual is withdrawn from the study. For example, some people may still be alive at the end of a clinical trial, or may drop out for reasons other than death before the experiment is completed. If the observation's beginning time at risk is uncertain, it is left-censored. Indeed, the identical observation can be censored on both the right and left sides, a phenomenon known as interval-censoring. Censoring makes survival model estimate and the likelihood function more difficult.

Censoring must be independent of the future value of the hazard for the person, conditional on the value of any covariates in a survival model and an individual's survival to a specific period. If this requirement is not met, survival distribution estimations may be substantially skewed. In this research, the censoring data depicts as 0=uncensored and 1=censored. Figure 3 above, T_0 denotes as the start of the observation period and T_1 is the end time of the observation period. Data that passed through T_1 is considered as right-censoring data, while data that did not passed T_0 denotes as left-censoring data and for data that in between T_0 and T_1 is an interval censoring. For data that stops exactly at T_1, it indicates as an uncensored data [8].

3.4. Breslow Estimator

Breslow estimator employs the profile likelihood technique. The estimates will always be positive due to the exponential component of the Breslow estimator. When the Cox PH model takes into account the covariates, the relative performance of the Breslow estimator is unclear.

Breslow provided a nonparametric maximum likelihood estimate for the cumulative baseline hazard function $\widehat{\Lambda}_{0,BR}(t)$, for patient *i* whose covariate vector is z_i=$(z_{i1}, \ldots, z_{iq})^T$, which in the absence of links between the recorded event dates may be expressed as

$$\widehat{\Lambda}_{0,BR}(t) = \sum_{j:x_j \le t} \left\{ \frac{\delta_j}{\sum_{k \in \Re} x_j \exp\,(\widehat{\beta}^T Z_k)} \right. \tag{6}$$

Breslow estimator for the baseline survival function is:

$$\hat{\mathrm{S}}_{0,BR}(t) = exp\{-\widehat{\Lambda}_{0,BR}(t)\} \tag{7}$$

As a result, the Breslow survival function estimator for a subject with covariate vector Z=z* is as follows:

$$\hat{\mathrm{S}}_{BR}(t|Z = Z^*) = [\hat{\mathrm{S}}_{0,BR}(t)]^{\exp(z)} = e^{\exp(^{Tz^*})} - \widehat{\Lambda}_{0,BR}|(t) \tag{8}$$

4. Results and discussion

4.1. Output result from SPSS

The expected graph should project as is in the shape of exponential graph. However, it is also acceptable if the output of the plot is in the shape of step function graph. From figure 4 below, the cumulative survival value at 1 indicates that all participants were considered alive at the early observation time. Besides, we can see a quick drop of survival time as the $t \to \infty$. This indicate that the increase in time of study, the lower the survival time of the participant.

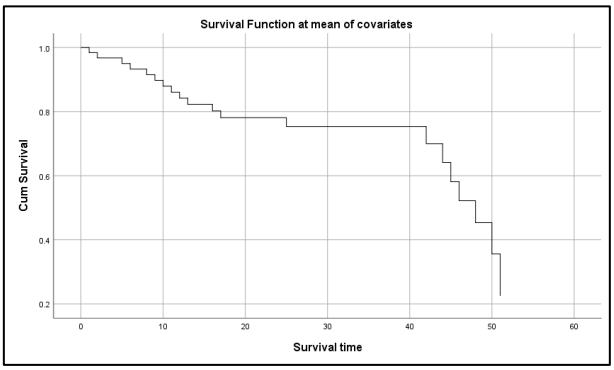

Figure 4 Survival graph

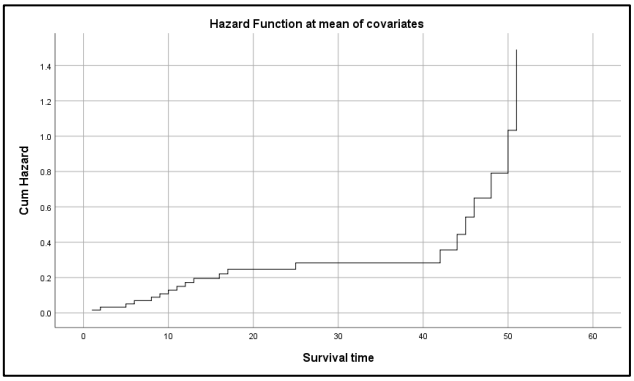

Figure 5 Hazard graph

We set up our model as follows:

$$y = \beta_1 X_1 + \beta_2 X_2 \tag{9}$$

From equation (9), y is the dependent variable of response, β_i are known as the regression coefficients, X_1 indicates the vaccination status, and X_2 indicates the age.

The following hypothesis is used to test the effect of X_1 and X_2 variables,

$H_0 : \beta_1 = \beta_2 = 0$ and $H_1 : \beta_i \neq 0$ for at least one i for i=1, 2

Table 1: Summary result from SPSS.

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
Vaccination Status	1.009	.474	4.523	1	.033	2.743	1.082	6.951
Age	.055	.236	.053	1	.817	1.056	.665	1.678

From Table 1, at the column of Sig., the *p*-value for vaccination status is 0.033. The value shows that it does have effect on the survival of the participants. This is because, the value is lower than 0.05 at 95% confidence interval. Meanwhile, for the age, the *p*-value is 0.817. The value shows that it does not have an effect on the survival of the participants. This is because the value is higher than 0.05 at 95% confidence interval.

The first column B indicates the value of coefficient of covariates for vaccination status and age that were estimated. The value of coefficient for vaccination status is 1.009 and the coefficient value for age is 0.055.

At 95% confidence interval for Exp(B), the hazard ratio value of vaccination status is 2.743 and it lies between 1.082 and 6.951. Since that hazard ratio is greater than 1, we can make an interpretation that the vaccination status gives a greater risk to the participants. The value of Exp(B) for age is 1.056 and it lies between 0.665 and 1.678. Since the value of hazard ratio is lower than 1, we can make an interpretation that the age is not in an increasing risk of dying.

4.1.1 Prognostic Factors for Survival

The Wald test was used to identify whether variables in the Cox model were significant. Table 2 shows the significance of the Wald test statistic for each variable in the Cox PH model.

Table 2: Significance of the Wald test statistic for Cox PH model.

Criteria	**|z|**	**Significance**
Vaccination Status	4.523	Greater than 1.96
Age	0.053	Lower than 1.96

The absolute values of the Wald test statistics for the vaccination status variable is 4.523 and is greater than the critical value of 1.96 at a significance level of 5%, as shown in Table 2 above. According to the results, the null hypothesis is rejected for these variables since $|z| > 1.96$ and it is significant in the model.

For age, the Wald test statistics value is 0.053 and is lower than the critical value of 1.96 at a significance level of 5%, hence the null hypothesis fails to be rejected for these variables since $|z| \leq 1.96$.

Table 3: Significance of the *p-value* for Cox PH model.

Criteria	***p-value***	**Significance**
Vaccination Status	0.033	Lower than 0.05
Age	0.817	Greater than 0.05

The *p-values* for vaccination status variable in Table 3 is 0.033 and it is lower than α = 0.05. These results suggest that these variables are not under the null hypothesis, and the parameters in the

model are significant. The *p-values* for all of the age is larger than α = 0.05. Thus, we can say that, the vaccination status is significant whereas the age range is insignificant to the model.

4.2. Output Result from R Software

The multivariate analysis of $X_1 + X_2$ on a single run of the R programming is as Figure 6 below. The coefficient of X_1 is 1.00724 and its hazard ratio is the exp(coef) = 2.73802, on the other hand, the coefficient of X_2 is 0.06405 and its hazard ratio is exp(coef) = 1.06614.
Wald test of coefficient X_1 is 2.123 while Wald test of coefficient X_2 is 0.274. The *p-value* of coefficient X_1 is |z| = 0.0337 and the *p-value* of coefficient X_2 is |z| = 0.7845.

```
> library(survival)
> args(coxph)
function (formula, data, weights, subset, na.action, init, control,
    ties = c("efron", "breslow", "exact"),
    singular.ok = TRUE, robust, model = FALSE, x = FALSE, y = TRUE,
    tt, method = ties, id, cluster, istate, statedata, ...)
NULL
> x=read.csv("data survival.csv",header=T)
> cox.mod<-coxph(Surv(survivaltime,status)~X1+X2,method="breslow",data=x)
> summary(cox.mod)
Call:
coxph(formula = Surv(survivaltime, status) ~ X1 + X2, data = x,
    method = "breslow")

  n= 52, number of events= 20

      coef exp(coef) se(coef)     z Pr(>|z|)
X1 1.00724   2.73802  0.47442 2.123   0.0337 *
X2 0.06405   1.06614  0.23416 0.274   0.7845
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   exp(coef) exp(-coef) lower .95 upper .95
X1     2.738     0.3652    1.0805     6.938
X2     1.066     0.9380    0.6737     1.687

Concordance= 0.66  (se = 0.076 )
Likelihood ratio test= 4.86  on 2 df,   p=0.09
Wald test            = 4.61  on 2 df,   p=0.1
Score (logrank) test = 4.98  on 2 df,   p=0.08
```

Figure 6 Output of multivariate analysis.

The covariate analysis of X_1 and X_2 also checked by following Figure 7 and Figure 8:

```
> library(survival)
> args(coxph)
function (formula, data, weights, subset, na.action, init, control,
    ties = c("efron", "breslow", "exact"),
    singular.ok = TRUE, robust, model = FALSE, x = FALSE, y = TRUE,
    tt, method = ties, id, cluster, istate, statedata, ...)
NULL
> x=read.csv("data survival.csv",header=T)
> cox.mod<-coxph(Surv(survivaltime,status)~X1,method="breslow",data=x)
> summary(cox.mod)
Call:
coxph(formula = Surv(survivaltime, status) ~ X1, data = x, method = "breslow")

  n= 52, number of events= 20

     coef exp(coef) se(coef)     z Pr(>|z|)
X1 1.0097    2.7448   0.4748 2.127   0.0335 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   exp(coef) exp(-coef) lower .95 upper .95
X1     2.745     0.3643     1.082     6.961

Concordance= 0.68  (se = 0.061 )
Likelihood ratio test= 4.79  on 1 df,   p=0.03
Wald test            = 4.52  on 1 df,   p=0.03
Score (logrank) test = 4.89  on 1 df,   p=0.03
```

Figure 7 Output of covariate X_1 analysis.

```
> library(survival)
> args(coxph)
function (formula, data, weights, subset, na.action, init, control,
    ties = c("efron", "breslow", "exact"),
    singular.ok = TRUE, robust, model = FALSE, x = FALSE, y = TRUE,
    tt, method = ties, id, cluster, istate, statedata, ...)
NULL
> x=read.csv("data survival.csv",header=T)
> cox.mod<-coxph(Surv(survivaltime,status)~X2,method="breslow",data=x)
> summary(cox.mod)
Call:
coxph(formula = Surv(survivaltime, status) ~ X2, data = x, method = "breslow")

  n= 52, number of events= 20

      coef exp(coef) se(coef)     z Pr(>|z|)
X2 0.06993   1.07244  0.23552 0.297    0.767

   exp(coef) exp(-coef) lower .95 upper .95
X2     1.072     0.9325    0.6759     1.702

Concordance= 0.501  (se = 0.082 )
Likelihood ratio test= 0.09  on 1 df,   p=0.8
Wald test            = 0.09  on 1 df,   p=0.8
Score (logrank) test = 0.09  on 1 df,   p=0.8
```

Figure 8 Output of covariate X_2 analysis

Figure 7 shows that the covariate X_1 is significant. This can be seen on the *p-value* that is equal to 0.0335 is lower than 0.05. Thus, reject the null hypothesis, H_0. Its hazard ratio value is HR = exp(coef) = 2.7448 with a 95% confidence interval and it lies between 1.082 and 6.961. Since the confidence interval for HR is greter 1, these results indicates that covariate X_1 associated with an increasing risk to the survivability of participant. Hence, covariate X_1 remained still to the model and continue with the next analysis for covariate X_2.

The following new hypothesis considering covariate X_2 is as below

$$H_0 : \beta_2 = 0 \qquad \text{vs} \qquad H_1 : \beta_2 \neq 0$$

Figure 8 shows that the covariate X_2 is insignificant. This can be seen on the *p-value* that is equal to 0.767 and it is greater than 0.05. Thus, the null hypothesis were failed to be rejected, H_0. Its hazard ratio HR = exp(coef) = 1.072 with a 95% confidence interval of 0.6759 and 1.072. Since the confidence interval for HR is includes 1, these results indicates that covariate X_2 makes a smaller contribution to the difference in the HR. In conclusion, only covariates X_1 give significant influence to the survival time while X_2 does not give significance influence to the survival of the participants.

Conclusion

The results from both SPSS and R software almost have a similar value on the coefficient of covariates. The coefficient for the vaccination status from SPSS IS 1.009 while from R software is 1.007. For the age, the coefficient from SPSS is 0.055 while from R software is 0.064. Both approaches show similar results in value of parameter estimates. Moreover, the model obtained from both software SPSS and R are the same.

Acknowledgement

I would like to thank my supervisor Dr. Noraslinda binti Mohamed Ismail who guided me in doing these projects. She provided me with encouragement and helped in difficult periods. Her motivation and help contributed tremendously to the successful completion of the project. A big thanks to Dr. Adina Najwa who is my examiner throughout these two semesters. Without the support and interest, this thesis would not have been the same as presented here. I would like to thank all of my supporters who helped by giving advice and providing the equipment which I needed. I would like to thank my family and friends for their support. Without that support I couldn't have succeeded in completing this project too.

References

[1] Kleinbaum, D. G. and Klein, M. (2012) Survival Analysis: a self-learning text. 3rd edn. New York: Springer.

[2] Isna Nur Aini (2011) Extended Cox Model Untuk Time-Independent Covariate Yang Tidak Memenuhi Asumsi Proportional Hazard Pada Model Cox Proportional Hazard. Bachelor Thesis, Universitas Indonesia.
[3] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.
[4] Rodriguez, G. (2007). *Lecture Notes on Generalized Linear Models*. URL: https://data.princeton.edu/wws509/notes/
[5] Kleinbaum, D. G. and Klein, M. (2012) Survival Analysis: a self-learning text. 3rd edn. New York: Springer.
[6] Bernstein L, Anderson J and MC Pike. Estimation of the proportional hazard in two-treatment-group clinical trials (http://www.jstor.org/stable/2530564). Biometrics (1981) vol. 37 (3) pp. 513-519
[7] Wald, A. (1943) 'Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large', *Transactions of the American Mathematical Society*, 54(3), p. 426. doi: 10.2307/1990256.
[8] Anthony Joe Turkson, Francis Aviah-Mensah, Vivian Nimoh, "Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review", International Journal of Mathematics and Mathematical Sciences, vol. 2021, Article ID 9307475, 16 pages, 2021. https://doi.org/10.1155/2021/9307475