Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia

Vol. 16, 2023, page 241 – 249

The Application of Extended Cox Regression Model on Leukemia Patients

Muhammad Riza Azruddin Mohd Aris, Noraslinda Mohamed Ismail*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: noraslinda@utm.my

Abstract
Leukemia, a complex and diverse disease, necessitates advanced statistical modeling to examine survival outcomes. This study proposes an extended Cox proportional hazard model, utilizing OpenBUGS software, to explore factors influencing leukemia patients' survival. The model incorporates time-dependent, time-independent, and random effects to account for dynamic covariates and individual-specific hazard functions. By adopting a Bayesian approach, prior knowledge is integrated, complex models are managed, and posterior distributions of parameters are obtained. OpenBUGS enables flexible modeling of time-dependent, time-independent, and random effects within the Cox framework. The model will be validated using a comprehensive dataset encompassing clinical characteristics, treatment histories, and follow-up information. The analysis aims to identify time-varying covariates and estimate random effects at the individual and group levels, uncovering novel risk factors and enhancing understanding of leukemia survival dynamics. The findings are expected to contribute to personalized treatment and management, assisting clinicians in informed decision-making and improving patient care. In conclusion, this thesis presents an extended Cox proportional hazard model that incorporates time-dependent and random effects, leading to a comprehensive understanding of leukemia prognosis and advancements in personalized medicine.

Keywords: Leukemia; Survival Analysis; Extended Cox Proportional Hazard Model;

1. Introduction

One statistical method that examines the data of time till the occurrence of an event of interest is survival analysis. Most human acute myeloid leukaemia (AML) cells have low proliferation potential, suggesting that the leukaemic clone is maintained by a small number of stem cells. Transplanting AML-initiating cells into SCID mice identified the AML-initiating cell, which, when exposed to in vivo cytokine therapy, homed to the bone marrow, proliferated significantly, and exhibited dissemination patterns and cell morphology resembling those of the original patients (Lapidot et al., 2018). Many cancers appear to rely on a tiny number of 'cancer stem cells' to continue growing and spreading (Huntly, B.J., and Gilliland, D.G., 2005). The first such cell to be characterised was the leukaemia stem cell (LSC). the Cox PH model contains an unknown baseline hazard and exponential expression. The baseline hazard is a function of t, but it excludes a time-independent covariate, while the exponential term does not include t but the covariate (Kleinbaum and Klein et al., 2012).

The problem statement aims to investigate the effect of time independent and time dependent in patients diagnosed with leukemia. Leukemia is a complex and aggressive form of cancer that requires a comprehensive understanding of various factors influencing patient survival. The analysis will consider random effects, such as the patients go to the chemotherapy treatment or take medicine or both. The data will be analysed in this study using OpenBUGS. The objective of this analysis was to estimate the effect of the time-independent and time-dependent covariate on the survival of leukemia patients and to estimate the effect of the time-independent covariate with the existence of random effects on the survival of leukemia patients.

In this study, Cox Regression was used. The 42 leukemia data from 1963 were used for the analyses in this study. Besides, Markov Chain Monte Carlo (MCMC) method was used in the extended Cox model. The exposure variable of interest is treatment status (Rx = 0 if new treatment, Rx = 1 if standard treatment). Two other variables for control are log white blood cell count (i.e., log WBC) and sex. Failure status is defined by the relapse variable (0 if censored, 1 if failure). Leukemia patients' survival rate is predicted by the study of the results using OpenBUGS. The results of this investigation may also help to better comprehend the leukemia survival rate.

2. Literature Review

Survival analysis defined as a group of statistical techniques for data analysis where the outcome variable of interest is the amount of time until a specific event happens (Emmert-Streib and Dehmer., 2019). The time variable, which can be referred to as either event time or survival time, is gathered from the beginning of the event's start follow-up until the event takes place or the study is completed, whichever comes first. While the incident is referred to as the "designated experience of interest," depending on each research case, it may involve a death, illness, recovery, or relapse (Kleinbaum and Klein, 2012). Censored data refers to the information for those imperfectly observed replies. The only strategies that can handle censored observations in this situation without throwing any data away are survival analysis methods because ordinary regression techniques are not equipped to handle the censoring data (Hazra and Gogtay, 2017).

Cox Proportional Hazard Model is one of the most widely used techniques in survival analysis. the final model using Cox regression analysis will produce a linearity equation model, where dependent variables are continuous or categorical risk factors that will affect the study endpoint, and independent variables are the incidence rate of an event (Abd Elhafeez *et al.*, 2021). According to (Kleinbaum and Klein.,2012), the extended Cox model will be applied if the independent variables fitted into the Cox regression model are not time-independent. a Cox proportional hazard regression model can be used to examine the elements that affect customers' decisions to buy a product.

In statistics, parameter estimation using Bayesian analysis and Markov chain Monte Carlo (MCMC) sampling is effective. In addition, using field test data and Bayesian theory, MCMC simulation is used to update the model of a coupled-slab system of a building structure. It is discovered that the complexity of the class of models has a significant impact on how easily the model updating problem may be identified (Lam et al.,2015). Next, in 1958, Edward L. Kaplan and Paul Meier introduced the Kaplan-Meier survival curves and estimates of survival data in a groundbreaking paper. The curves served as a valuable tool for handling incomplete observations and diverse survival times within datasets, especially in cases where not all subjects completed the study (Rich et al., 2010). When determining the survival probability of a research population, comparing the survivorship between different study groups of a covariate, and calculating the median survival on each subgroup, the Kaplan-Meier survival curve is typically utilised (N. Dudley *et al.*, 2016).

Random effect are formulated generally a level is a set of units, or equivalently a system of categories, or a classification factor in a statistical design. In statistical terminology, a level in a multilevel analysis is a design factor with random effects (Tom A.B. Snijders., 2005). When there are no theoretical or other prior guidelines about which variables should have a random effect, the researcher can be led by the substantive focus of the investigation, the empirical findings, and parsimony of modeling. Next is Cox Regression with Random Effects. This extends the standard frailty model by permitting a multivariate random effect with arbitrary design matrix in the log relative risk, in a manner similar to the modelling of random effects in linear, generalised linear, and non-linear mixed models.The random effects' distribution is commonly considered to be multivariate normal, although other (ideally symmetrical) distributions are also feasible. The reliance or connection between failure times within the same patient/cluster is a frequent aspect of such multivariate failure time data. In order to account for the patient effect, a random effect (or frailty) has been introduced into the standard hazard function (Vaida, F., & Xu, R., 2000).

3. Research Methodology

Survival Analysis

The period until a certain event of interest occurs serves as the outcome variable in the mathematical method for data analysis known as survival analysis. Where T is the non-negative continuous random variable that represents the amount of time until an event of interest occurs, the probability density function of time to event, f(t), was used to estimate the time to event with survival data. Below is the probability density function (1),

$$f(t) = \lim_{\Delta t \to 0} \left[\frac{\mathbb{P}(\ t \leq T < t+\Delta t\ )}{\Delta t}\right], \text{ for t >0} \tag{1}$$

Survival Function

The survival function, $S(t)$ is known as the probability of the event of interest not yet occurring by a specific time, *t* (a person survival longer than the specific time *t*). The survival function can also be visualized and described as a smooth curve with survival probability being the y-axis and time being the x-axis. Hence, the curve will always decrease down starting from $S(t = 0) = 1$ until the *t* goes to infinity. The formula is given as below,

$$\begin{aligned} S(t) &= \int_t^\infty f(x)\,dx, \qquad\qquad \text{for} \quad \text{t >0} \\ &= \mathbb{P}(T > t) \\ &= 1 - \mathbb{P}(T\ \leq\ t) \\ &= 1 - F(t) \end{aligned} \tag{2}$$

Cox Proportional Hazard (PH) Model

The Cox Proportional Hazard (PH) model, which considers the impact of censored observations, is known as the most popular mathematical model used to assess the association between several proportional variables and survival times of an event (hazard rate). The Cox PH model, according to Cox (1972), was expressed as a hazard function, h(t), with n variables. The formula is as follows (3),

$$\begin{aligned} ln[h(t)] &= ln[h_o(t)'] + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \\ &= ln[h_o(t)'] + \beta_0 + \sum_{k=1}^{n} \beta_k X_k \end{aligned} \tag{3}$$

Since $\exp(\beta_0)$ is a constant, the Cox regression model simplified to become (4),

$$h(t) = h_o(t) * exp(\sum_{k=1}^{n} \beta_k X_k\,) \tag{4}$$

$$h(t) = h_o(t) * exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n\,)$$

Cox Regression with Random Effects

The data presented in the leuk example actually arise via a paired design. Patients were matched according to their remission status (partial or complete) (Freireich et al.,1963). One patient from each pair received the drug 6-MP whilst the other received the placebo. We may introduce an additional vector (called pair) in the BUGS data file to indicate each of the 21 pairs of patients. We model the potential 'clustering' of failure times within pairs of patients by introducing a group-specific random effect or frailty term into the proportional hazards model. Using the counting process notation introduced in the Leuk example, this gives

$$l_i(t)dt = Y_i(t)\exp\left(\beta' z_i + b_{pair_i}\right) d\Lambda_0(t), \ \ i=1,\ldots,42; \ \ pair_i = 1,\ldots,21 \quad (5)$$

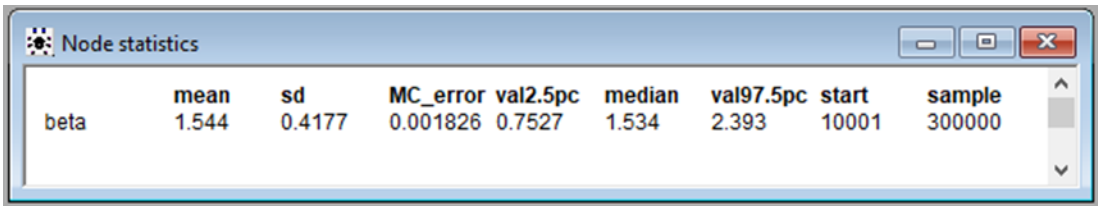
$$b_{pair_i} \sim \text{Normal}(0, t) \quad (6)$$

A non-informative Gamma prior is assumed for t, the precision of the frailty parameters. Note that the above 'additive' formualtion of the frailty model is equivalent to assuming multiplicative frailties with a log-Normal population distibution. Clayton (1991) discusses the Cox proportional hazards model with multiplicative frailties, but assumes a Gamma population distribution.

4. Results and discussion

Cox Proportional Hazard (PH) Model with Time Independent

The Proportional Hazard Assumption Test is conducted before fitting the dataset into the Cox Proportional Hazard Regression Model. In this study, the PH assumption is checked by using OpenBUGS software. The chain was run with a burn-in of 110000 iterations begin with 10001 then end with 10000000 and the thinning to every 100th draw. The parameter estimates are not sensitive to the choice of hyperparameters and initial values. Graphical representations of the posterior distribution can indicate problems with the performance of the Markov Chain Monte Carlo (MCMC) algorithm. OpenBUGS has a number of tools for reporting the posterior distribution.



```
        mean    sd      MC_error  val2.5pc  median  val97.5pc  start  sample
beta    1.544   0.4177  0.001826  0.7527    1.534   2.393      10001  300000
```

Figure 1 Results of time independent by using OpenBUGS Software

From the result the mean for beta is 1.544. To monitor whether beta is significant or not, it can be checked by looking at the 95% confidence interval for beta. If zero included, it means the result is not significant. The results shows that the interval for beta is 0.7525 ≤ beta ≤ 2.393, respectively. Hence, beta is significant.

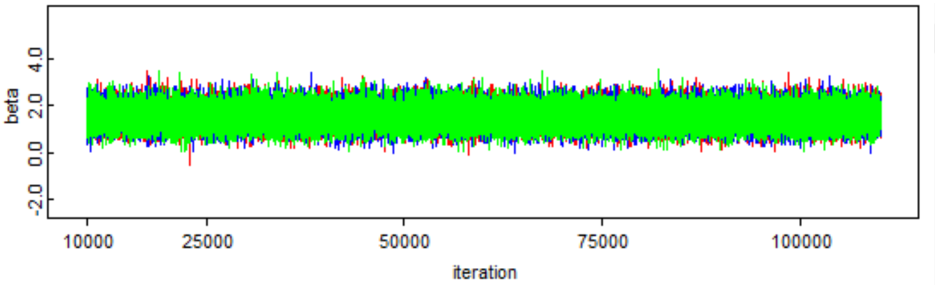


Figure 2 History diagram for beta

Figure 2 above shows a full trace plot of all stored values. It can be seen that the generated observations converge because all the values are within the zone and no strong periodicities and tendencies.

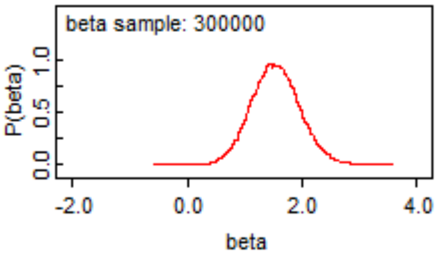


Figure 3 Density Plot for beta

Figure 3 represents the density plots. Densit plots shows the symmetrical of the posterior density. Since the graph shows bell shape, which is symmetric and fairly skewed, it can be said that the iterations converged.

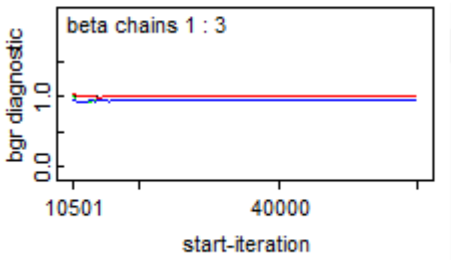


Figure 4 BGR Diagram for beta

Figure 4 shows the Brooks-Gelman-Rubin (BGR) diagnostic statistics. The graph for beta shows the iterations is converged. The BGR statistics is an ANOVA-type diagnostic. It compares within-and among-chain variance. From the figure, the BGR statistics is graphed by the red line.

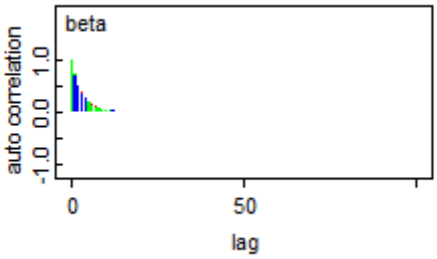
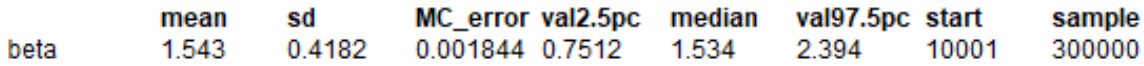


Figure 5 Auto Correlation for beta

Figure 5 shows the auto correlation for treatment and prior therapy. From the auto correlation diagram, the x-axis shows the lag and the y-axis shows the correlation coefficient. Convergence of parameter for treatment and prior therapy does not shows a decreasing curve.

Cox Proportional Hazard (PH) Model with Time Dependent
In this study, the coding that were used will be multiply with time.



```
        mean    sd      MC_error  val2.5pc  median  val97.5pc  start   sample
beta    1.543   0.4182  0.001844  0.7512    1.534   2.394      10001   300000
```

Figure 6 Results of time dependent by using OpenBUGS Software

The mean beta value based on the outcome is 1.543. The 95% confidence interval for beta can be used to determine if beta is significant or not. If a zero is present, the outcome is not significant. The data indicates that the beta range is between 0.7512 < beta < 2.394. Therefore, beta is significant.

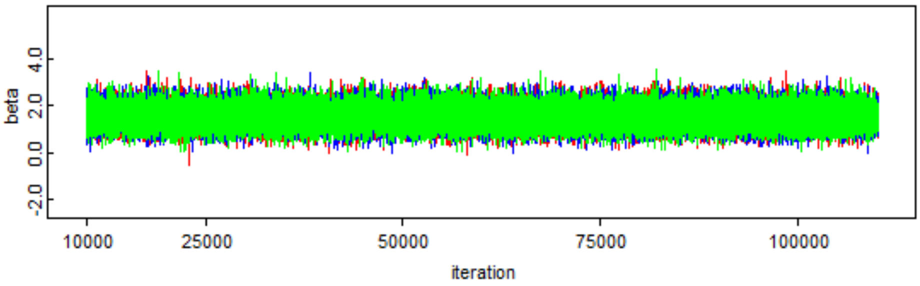


Figure 7 History Diagram for beta

A complete trace plot of all stored values is displayed in Figure 7 above. It is clear that the produced observations converge because all the values are within the zone and no strong periodicities tendencies.

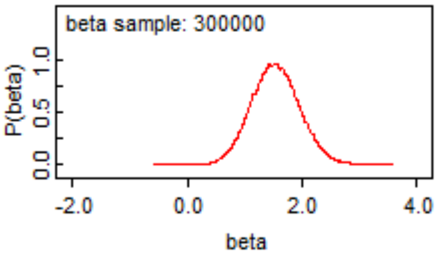


Figure 8 Density Plot for beta

The density plots are shown in the Figure 8 above. Density plots display the posterior density's symmetrical distribution. It may be claimed that the iterations converged since the graph has a bell shape that is symmetric and slightly skewed.

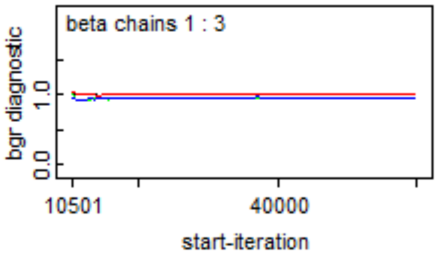


Figure 9 BGR Diagram for beta

The Brooks-Gelman-Rubin (BGR) diagnostic data are displayed in Figure 4.9. Due to the line's proximity to 1, the graph for beta suggests that iterations are becoming increasingly convergent. ANOVA-style diagnostics are used in BGR statistics. It contrasts the diversity within and between chains. The red line in the illustration graphs the BGR statistics.

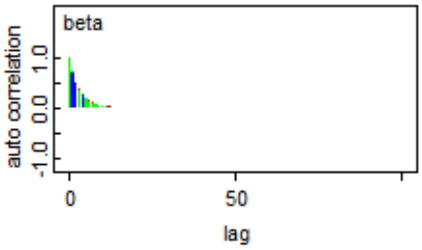
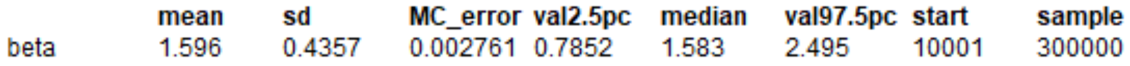


Figure 10 Auto Correlation for beta

The autocorrelation for treatment and preceding therapy is depicted in Figure 10 above. The lag is shown on the x-axis of the auto correlation diagram, and the correlation coefficient is shown on the y-axis. The parameter convergence for the treatment and preceding treatments does not depict a decelerating curve.

Cox Proportional Hazard (PH) Model with Random Effect

In this study, we will used the random effect which is the leukemia patient take the chemotherapy and medicine or not.



	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
beta	1.596	0.4357	0.002761	0.7852	1.583	2.495	10001	300000

Figure 11 Results of random effect by using OpenBUGS Software

The data shows that beta has a mean of 1.596. The 95% confidence interval for beta can be used to determine if beta is significant or not. If a zero is present, the outcome is not noteworthy. The data indicates that the beta range is between 0.7852 < beta < 2.495. Therefore, beta is significant.

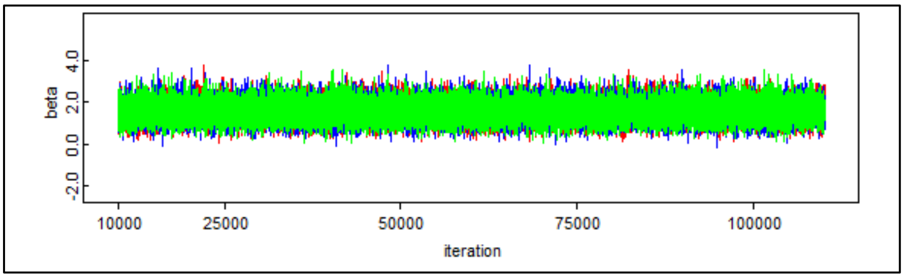


Figure 12 History Diagram for beta

A complete trace plot of all saved values is displayed in Figure 12 above. It is clear that the generated observations converge because all the values are within the zone and no strong periodicities and tendencies.

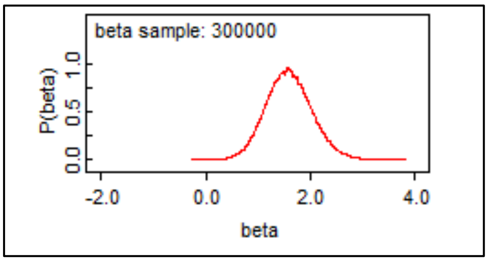


Figure 13 Density Plot for beta

The density graphs are shown in Figure 13. Density plots display the posterior density's symmetrical distribution. It may be claimed that the iterations converged since the graph has a bell shape that is symmetric and slightly skewed.

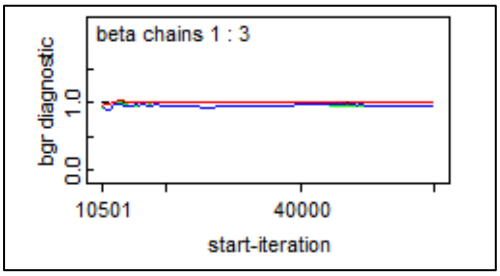


Figure 14 BGR Diagram for beta

The Brooks-Gelman-Rubin (BGR) diagnostic data are displayed in Figure 14 above. Due to the line's proximity to 1, the graph for beta suggests that iterations are becoming increasingly convergent. ANOVA-style diagnostics are used in BGR statistics. It contrasts the diversity within and between chains. The red line in the illustration graphs the BGR statistics.

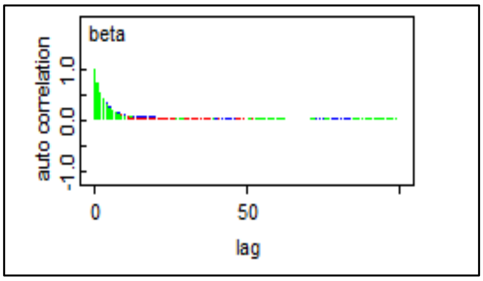


Figure 15 Auto Correlation for beta

The autocorrelation for treatment and preceding therapy is displayed in Figure 15. The lag is shown on the x-axis of the auto correlation diagram, and the correlation coefficient is shown on the y-axis. The parameter convergence for the treatment and preceding treatments does not depict a decreasing curve.

Conclusion

The significance of this study was to assist in predicting the survival of the leukemia patients. The 42 data of leukemia patients were used in this study that have censored and uncensored data. A Cox PH model was explored, and parameter estimation for the explanatory variables was performed using Markov Chain Monte Carlo (MCMC) in OpenBUGS software based on the leukemia data. Several variables such as number of expected survival time, observed survival time, censored data, uncensored data and covariates were used to fit into the Cox model. These variables were time-independent, and the model satisfied the PH assumption. In OpenBUGS software, the output of the parameter estimation consists of the mean, standard deviation, MC error, median, 95% of Confidence Interval, the starting of the iteration and number of samples.

In summary, an extended Cox model was explored to compare the time of the leukemia data. From the results obtained, the value of mean for both time-independent and time-dependent were the same, we can conclude that the leukemia data were not influenced with time. That means the leukemia data was independent. The treatment was not depending on time but it is depending on the type of treatment that given to the patients and the adaptability of the patients. In conclusion, the two objectives proposed in Chapter 1 were successfully achieved.

References

[1] Abd Elhafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G. and Roumeliotis, S. (2021). Methods to Analyze Time-to-Event Data: The Cox emm.

[2] Glen, S. (2016). Wald Test: Definition, Examples, Running the Test - Statistics How To. *StatisticsHowTo*., 1–3.

[3] Hazra, A. and Gogtay, N. (2017). Biostatistics Series Module 9: Survival Analysis. *Indian journal of dermatology*. 62(3), 251–257.

[4] Huntly, B. J., & Gilliland, D. G. (2005). Leukaemia stem cells and the evolution of cancer-stem-cell research. Nature Reviews Cancer, 5(4), 311-321.

[5] Kishore, J., Goel, M. and Khanna, P. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*. 1(4), 274.

[6] Kleinbaum, D. G., Klein, M., Kleinbaum, D. G., & Klein, M. (2012). Evaluating the proportional hazards assumption. Survival analysis: a self-learning text, 161-200.

[7] Kleinbaum, D.G. and Klein, M. (2012). Introduction to Survival Analysis. In pp.1–54.

[8] Lapidot, T., Sirard, C., Vormoor, J., Murdoch, B., Hoang, T., Caceres-Cortes, J., ... & Dick, J. E. (1994). A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature*, *367*(6464), 645-648.

[9] Snijders, T. A. (2005). Fixed and random effects. *Encyclopedia of statistics in behavioral science*, *2*(2), 664-665.

[10] Vaida, F., & Xu, R. (2000). Proportional hazards model with random effects. *Statistics in medicine*, *19*(24), 3309-3324.