World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:7, No:5, 2013

# Survival Model for Partly Interval-Censored Data with Application to Anti D in Rhesus D Negative Studies

F. A. M. Elfaki, Amar Abobakar, M. Azram, M. Usman



***Abstract*—**This paper discusses regression analysis of partly interval-censored failure time data, which is occur in many fields including demographical, epidemiological, financial, medical and sociological studies. For the problem, we focus on the situation where the survival time of interest can be described by the additive hazards model in the present of partly interval-censored. A major advantage of the approach is its simplicity and it can be easily implemented by using R software. Simulation studies are conducted which indicate that the approach performs well for practical situations and comparable to the existing methods. The methodology is applied to a set of partly interval-censored failure time data arising from anti D in Rhesus D negative studies.

***Keywords*—**Anti D in Rhesus D negative, Cox's model, EM algorithm.

## I. Introduction

BY partly interval-censored failure time data we means, for some subjects, the exact failure times are observed, but for the remaining subjects, the survival time of interest is observed only to belong to an interval instead of being exactly [14], [18], [12].

General partly interval censored data arise often in follow-up studies. An example of such data is provided by the Framingham Heart Disease Study; see [16] for a description. In this study, times of the first occurrence of anti D infection through to contaminated blood factor disease patients are of interest. For some patients, time of the first occurrence of infection is recorded exactly. But for others, time is recorded only between two clinical examinations. Another example of such data is provided by the study on incidence of protein urea in insulin-dependent diabetic patients in Denmark; see [8] for a detailed description.

Suppose time to event random variable, or failure time $T_1, T_2, ..., T_n$ are independent and identically distributed as $F_0$. If all the random variables are observable, then it is well known that the semiparametric maximum likelihood estimator of $F_0$ is the empirical distribution function and it is asymptotically efficient.

However many reliability and medical studies, observations are subject to censoring. The goal of this paper is to discuss a semi-parametric Cox's proportional hazards regression model with the subdistribution of $F_0$ based on incomplete partly interval-censored data in which some of the failure times are observed, but some of the failure times are subject to interval censoring [13], [18].

There are many cases of partly censored data; here we consider the case that is; for the some subjects, the exact failure times $T_1, T_2, ..., T_n$ are observed. But for the remaining subjects, only the information pertaining to their current status is available. That is for subject in this group, we only know whether or not failure has occurred at the examination time $U_i$, so the observed data is;

$$(\delta_i, U_i) \quad i = n_1 + 1, ..., n$$

where $\delta_i = 1$ if the unknown failure time $T_i \le U_i$ and $\delta_i = 0$ otherwise. Note that this censored model is different from doubly-censored data studied by [4], [3] and [11].

In the competing risks model, a unit is exposed to several risks simultaneously, but it is assumed that the eventual failure of the unit is due to only one of these risks, which is called "cause of failure" [1]. The standard analysis for competing risks data involves modeling the cause specific hazard function of the different failure types under Cox's model assumption [15], [17]. The cause specific hazard function is known as subdistribution function, also historically was known as the cumulative incidence function, the marginal probability function, the crude incidence or the absolute cause-specific risk [2]. In this paper, we propose the semi-parametric proportional hazard model of the subdistribution function for partly interval-censored of a competing risks survival data based on EM algorithm to estimate the parameters.

## II. Competing Risks Model Formulation

Reference [9] developed a class of estimation procedures for semi-parametric proportional hazards regression model for the subdistribution of a competing risks model using the partial likelihood principle and weighting techniques.

F. A. M. Elfaki and M. Azram are with the Department of Science in Engineering, Faculty of Engineering, International Islamic University Malaysia, 50728, Kuala Lumpur, Malaysia (Phone: +60361964480, e-mail: faizelfaki@yahoo.com).

Amar Abobakar is with the Department of Hematology, Faculty of Medical Laboratory Science, University of Medical Science and Technology, Khartoum, Sudan.

M. Usman is with the Department of Mathematics, Faculty of Science, Universitas Lampung, Bandar Lampung, Indonesia.

Specifically, let $T$, $C$ and $Z$ denote the failure time, the censoring time and $p\times 1$ bounded time-independent covariate vector. Let $\varepsilon \in (1,2,...,K)$ because of failure, for which the $K$ causes are assumed to be observable. We only observe $X_i = \min(T_i, C_i)$, $\Delta_i = \delta_i I(T_i \le C_i)$, where $I(.)$ is the indicator function of the event. Here $\delta_i = 1$ if $T_i$ is observed and $0$ otherwise, and $C_i$ is the independent censoring variable. The main interest is the modeling of the cumulative incidence function for failure from say cause 1 conditional on the covariates, i.e. $F_1(t;Z) = \Pr(T \le t, \varepsilon = 1/Z)$, and the hazard of the subdistribution as originally described by [10]. Gray constructed K-sample tests for differences in the cumulative incidence function based on integrated difference of nonparametric estimates of the within-group subdistribution hazard functions. The subdistribution hazard as defined by Gray is,

$$\lambda_i(t;Z) = \lim_{\Delta t\to 0} \Pr(t \le T < t+\Delta t, \varepsilon = 1/T \ge t, Z = z)/\Delta t,$$
$$= \frac{1}{1-F_1(t;Z)}.\frac{d}{dt}\big(F_i(t;Z)\big) = \frac{-d}{dt}\big[\log\{1-F_1(t;Z)\}\big]$$

The cumulative incidence function and subdistribution hazard functions are estimable from the competing risks data [17]. We use Cox proportional hazards models to specify each $\lambda(t;Z)$ and assume that censoring is conditionally independent of the latent failure times for given $Z$. Then, under Cox model;

$$\lambda(t;Z) = \lambda_0(t)\exp(Z^T(t)\beta) \quad (1)$$

where $\lambda_0(t)$ is a completely unspecified, nonnegative function in $t$, $\beta$ is regression coefficients and $Z(t)$ is the original time-dependent covariates (time-varying covariates). For simplicity, we restrict our attention to a time-independent covariates. Thus the regression coefficients and baseline hazard form the Cox model for $F$ have straightforward interpretation that does not depend on the probabilistic structure of the subdistribution hazard and given as;

$$F(t;Z) = 1-\exp\Big[-\int_0^t \lambda_0(s)\exp\{Z^T(s)\beta\}ds\Big] \quad (2)$$

The familiar form of this proportional hazards model is intended to be a convenient empirical representation for the cumulative risk of a competing risk and should be evaluated on the extent to which it permits the analyst to assess the effect of covariates on the cumulative incidence function [9]. Previous work in survival data as mentioned above used techniques when there is only a single cause of failure. In this paper, we use the model for the cumulative incidence function in the competing risks setting for survival data analysis.

## III. Semi-Parametric Maximum Likelihood Estimation

### A. Complete Data

In this section we present the partial likelihood for the subdistribution function for complete data. By complete data we mean that $t$ and $\varepsilon$ are observed for all individuals. The type of data considered is complete in all failure times (not censored) and incomplete only in terms of failure modes. It may be mentioned that [6] considered different cases when the data are censored.

As mentioned above, an individual who has not failed from the cause of interest by time $t$ is at risk. This includes two distinct groups: those who have not failed from any cause and those who have previously failed from another cause. The partial likelihood for the improper distribution, $F(t:Z)$ as proposed by [9] is;

$$L(\beta) = \prod_{i=1}^{n}\left[\frac{\lambda_0(t_i)\exp\{Z_i^T(t_i)\beta\}\Delta t_i}{\sum_{j\in R_{(t_i)}}\exp\{Z_i^T(t_i)\beta\}}\right]^{I(\varepsilon_i=1)} \quad (3)$$

where $R_{(t_i)}$ is the risk set at time of failure for the $ith$ individual.

The log partial likelihood is;

$$\log\{L(\beta)\} = \sum_{i=1}^{n} I(\varepsilon_i = 1)\times\Big(Z_i^T(t_i)\beta - \log\Big[\sum_{j\in R_{(t_i)}}\exp\{Z_i^T(t_i)\beta\}\Big]\Big) \quad (4)$$

where $\log\{L(\beta)\}$ indicates that the a function depends on the unknown parameters $\beta$, the values of $Z$ being known.

The asymptotic theory of maximum likelihood estimation requires that the likelihood function satisfy some "regularity conditions" which are met in most applications. The regression coefficients $\beta$ are estimated by the values $\hat{\beta}$, which maximize the logarithm of the partial likelihood. The values $\hat{\beta} = (\hat{\beta}_1,...,\hat{\beta}_p)$ are obtained by equating to zero the $p$ first derivatives of $\log\{L(\beta)\}$ with respect to $\beta_i (i = 1,..., p)$. An iterative process such as the EM algorithm or Newton-Raphson is adopted to solve this system of equations for $\beta$. The score vector is obtained by taking the 1st derivative of (4) with respect to $\beta_i (i = 1,..., p)$ and is given by;

$$\frac{\partial \log\{L(\beta)\}}{\partial \beta_i} = \sum_{i=1}^{n} I(\varepsilon_i = 1) \times \left[ Z_i(t_i) - \frac{\sum_{j \in R_i} Z_j(t_i) \exp\{Z_j^T(t_i)\beta\}}{\sum_{j \in R_i} \exp\{Z_i^T(t_i)\beta\}} \right] \quad (5)$$

This derivative is the difference between the value of the $pth$ covariate on the subject who fails at $t$ and the weighted average of the covariate over the risk set $R_{(t_i)}$, with exponential weights $\exp\{Z_i^T(t_i)\beta\}$. Reference [9] adapt (5) in terms of counting processes, by letting $N_i(t) = I(T_i \le t, \varepsilon_i = 1)$ and $Y_i(t) = 1 - N_i(t-)$, so (5) become;

$$U(\beta) = \sum_{i=1}^{n} \left[ Z_i(s) - \frac{\sum_j Y_j(s) Z_j(s) \exp\{Z_j^T(s)\beta\}}{\sum_j Y_j(s) \exp\{Z_i^T(s)\beta\}} \right] \times dN_i(s) \quad (6)$$

The EM algorithm is used to estimate $\hat{\beta}$ by setting $U_\beta(\hat{\beta}) = 0$.

## IV. Weighted Score Function Method

To modify the second model of [9], let $\{T_i, C_i, Z_i, i = 1,...,n\}$ to be $n$ independent copies of $\{T, C, Z\}$. However, one can only observe $X_i = \min(T_i, C_i)$ and $\Delta_i = I(C_i \le T_i)$ for $i = 1,...,n$, where $I(.)$ defined as before in section 2. In the case when the survival distribution $G(.)$ of the censoring variable $C$ does not depend on $Z$, the weight at time $t$ proposed by [9] is, $w_i(t) = r_i(t)\hat{G}(t)/\hat{G}(X_i \wedge t)$ can make a simple modification of the weight at time $t$ [7] as follows;

$$w_i(t) = \frac{r_i(t) G(t)}{G_{Z_i}(t) G_{Z_j}(t)}$$

where $G_Z(.)$ is the Kaplan-Meier estimator for the survival function and $r_i(t) = I(C_i \ge T_i \wedge t)$ is the vital status on individual $i$ at time $t$. Censored individuals are observed until time $C_i$; thereafter, vital status is uncertain. If $r_i(t) = 0$, then $Y_i(t)$ and $N_i(t)$ are not observable. If $r_i(t) = 1$, then $Y_i(t)$ and $N_i(t)$ are observed data up to time $t$. Moreover, consider the standard extreme value distribution $F$ for Cox model $F(t) = 1 - \exp(-\exp(t))$. The covariate vector that has a finite number of possible values, form the basis of the weighted score function, which when applied to (6) become;

$$U_\beta^*(\beta) = \sum_{i=1}^{n} \int_0^\infty \left[ Z_i(u) - \bar{Z}_*(\beta, u) \right] w(Z_i^T \beta) dN_i(u) \quad (7)$$

where

$$\bar{Z}_*(\beta, u) = \frac{\sum_{i=1}^n w(Z_i^T\beta) Z_i(u) Y_i(u) \exp(Z_i^T(t)\beta)}{\sum_{i=1}^n w(Z_i^T\beta) Y_i(u) \exp(Z_i^T(t)\beta)} \quad (8)$$

and $w(.)$ is a positive weight function. The weight function becomes identically 1 when $F$ is the standard extreme value distribution, [5]. Reference [5] shows that the estimation procedure with $w = 1$ works well for the proportional hazards odds model. The solution of $U_\beta^*(\beta) = 0$, using the same arguments given in Appendices A of [9].

## V. Illustrative Example

The proposed method is illustrated Anti D in Rhesus D negative pregnant Sudanese women who were treated in two hospitals in Sudan, that is, Khartoum hospital and Aldayat hospital. They were 100 patients of the study were at risk for anti D infection through the contaminated blood factor. At the end of the study, there were 50 patients found to be prophylactic standard dose of the anti-D immunoglobulin administration in pregnant women who are Rhesus D-negative, but the infection times were interval-censored. Among them 14% positive for anti body and 86% negative in women who receive anti D appropriate postnatal prophylaxis. The patients were classified into either the positive treated group or negative treated group according to the amount of blood received (when treated for anti D in Rhesus). The goal here is to investigate the possible association between the treatment and the anti D in Rhesus time. We code the covariate $z_i = 0$ or $z_i = 1$ if the patient was positively or negatively treated. To see the effect of covariates on development of complications, we fitted our proposed model that is competing risks model based on EM algorithm. Appling the procedures described early, we obtained the result as shown in Table I. The first cause of failure show better result compare to second cause of failure based on standard deviation and smallest variance. We conclude that the covariates do not have a significant different. However, it is confirmed that the negative treated group had a significantly higher risk of the onset of anti D rhesus after infection.

TABLE I
Estimate Obtained under the Competing Risks Model Based on Subdistribution Using EM Algorithm for Anti D in Rhesus D Negative Data

| **First Causes** | | | | |
|---|---|---|---|---|
| Eq | $\beta_1$ | $\beta_2$ | var($\hat{\beta}$) | $E$(var) |
| W | 0.724(0.0223) | 0.712(.034) | 0.015 | 0.015 |
| CC | 0.742 (0.0223) | 0.723(.034) | 0.015 | 0.015 |
| **Second Causes** | | | | |
| W | 0.602 (0.134) | 0.653(1.27) | 0.0672 | 0.0672 |
| CC | 0.632(0.134) | 0.653(1.27) | 0.0672 | 0.0672 |

## VI. Conclusion

We have proposed a simple modification of estimating functions for partly-interval censored data using the semi-parametric Cox's proportional hazards regression models of the subdistribution of a two competing risks models namely, the censoring complete model and a weighting technique model. Simulations studies (which is not addressed here) indicate that under the assumed modification models of [9], the weighted estimating equation with censored data can be as efficient as the censoring complete score function. However, both proposed models give similar results from the two simulation studies [7]. Similar results are also obtained when using anti D infection. EM algorithm was used to estimate the parameters of the model. The simulation studies strongly support the generalized missing information principle in a semi-parametric context and use of the generalized profile information for non-identically distributed samples. From the real data set we find that the covariates do not have a significant difference. The first cause of failure show better result compare to second cause of failure based on standard deviation and smallest variance. Fixing the age at diagnosis, a very young patients have a lower hazard rate than relatively young patients. Even with many exact observations (100), the additional interval-censored observations (23) help to give a more accurate estimate of the regression parameter. However, it is confirmed that the negative treated group had a significantly higher risk of the onset of anti D Rhesus after infection.

## References

[1] Aly, A. A. E, Kochar, S. C., and Mckeague, I. W., Some Tests For Comparing Cumulative Incidence Functions and Cause-Specific Hazard Rates. *American. S. A. J*. 89 (1994): 994- 999.

[2] Benichou, J., and Gail, M. H., Estimates of Absolute Cause Specific Risk in Cohort Studies. *Biometrics*, 46(1992): 813-826.

[3] Chang, M. N. Weak Convergence of a Self-Consistent Estimator of the Survival Function with Doubly Censored Data. *Ann. Statist*. 18(1990): 391-404.

[4] Chang, M. N. and Yang, G. L., Strong Consistency of a Nonparametric Estimator of the Survival Function with Doubly Censored Data. *Ann. Statist*. 15(1987): 1536-1547.

[5] Cheng, S. C., Wei, L. J. and Ying, Z., Analysis of Transformation Models with censored Data. *Biometrika*, 83(1995): 835-846.

[6] David, H. A. and Mosechberger, M. L., The Theory of Competing Risks. (1978) London: Griffin.

[7] Elfaki, F. A. M., Parametric and semi parametric competing risks models for statistical process control with reliability analysis. (2004). *Ph.D., Thesis*. University Putra Malaysia.

[8] Enevoldsen, A. K., Borch-Johnson, K., Kreiner, S., Nerup, J, and Deckert, T., De-clining Incidence of Presistent Proteinuria in Type I (insulin-dependent) diabetic Patient in Denmark, Diabetes. 36 (1986): 205-209.

[9] Fine. J. P. and Gray. R. J., A Proportional Hazards Model For the Subdistribution of a competing Risk. *American. S. A. J*, 94 (1999): 496-509.

[10] Gray, R. J., A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of Statistics*, 80 (1988): 557-572.

[11] Gu, M. G. and Zhang, C. H., Asymptotic Properties of Self-Consistent Estimation Based on Doubly Censored Data. *Ann. Statist*. 21 (1993): 611-624.

[12] Huang, J., Asymptotic Properties of Nonparametric Estimation Based on Partly Interval- censored Data. *Statistica Sinica*. 9 (1999): 501-519.

[13] Kalbfleisch, J. D and Prentice, R. L., the Statistical Analysis of Failure Time Data. (1980).New York: Wiley.

[14] Kim. J. S. Maximim Likelihood Estimation for the Proportional Hazards Model with Perty Interval-Censored Data. *J R. Statist. Soc*., Series B 65 (2003): 489-502.

[15] Larson, M. G., Covariate Analysis of Competing Risks Models with Log-Linear Models. *Biometrics*, 40 (1984): 459-469.

[16] Odell, P. M., Anderson, K. M. and D'Agostino, R. B., Maximum Likelihood Estimation for Interval-Censored Data using a Weibull-based Accelerated Failure Time Model. *Biometrics* 48 (1992): 951-959.

[17] Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Berslow, N. E., The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics*, 34 (1978): 541-554.

[18] Zhao. X., Zhao. Q. Sun. J., and Kim S. J. Generralized Log-Rank Test for Partly Interval-Censored Failure Time Data. *Biometrical Journal*, 3 (2008): 375-385.