# Analysis of Partly Interval-Censored Data under Competing Risks Framework

**Yosra Yousif[1], F. A. M. Elfaki[2*] and Meftah Hrairi[1]**
[1]Department of Mechanical Engineering, Faculty of Engineering, IIUM, P.O.Box 10, 50728 Kuala Lumpur, Malaysia
[2]Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, P.O. Box 2713, Doha, Qatar

**Summary:**
Bayesian inference for partly interval-censored data is considered when there are two or more causes of failure with possibility of being masked. Cox proportional hazard model is adopted to estimate the regression coefficients. Simulation data show that the developed model is feasible and easy to implement.
***Keywords:***
*Competing Risk, Partly Interval Censored data, Bayesian Analysis.*

## 1. Introduction

Partly interval-censored (PIC) data that resulted from the studies are the subjects which are inspected periodically. This type of data consists of interval-censored (IC) observations in addition to exact observations. Suppose, there are $N$ subjects under study where $n$ of them have exact failure time, and the rest $m$are only observed for the interval that includes the true failure time. Then for the $i^{th}$ subject whose failure time is exact of the observed data is$\{(T_i, X_i)\}_{i=1}^{n}$, while for the $i^{th}$ subject whose failure time is interval of the observed data is $\{(L_i, R_i, X_i)\}_{i=1}^{m}$ where $L_i < R_i$ . Here $T_i$ and $X_i$denote the exact failure time and the vector of the covariates correspond to the $i^{th}$ subject, respectively. $L_i$ and $R_i$ are representing the lower and upper limits of the observed interval which include the true failure time. If the subject failed before the first inspection time, then it is considered as left-censored observation $(0 < T_i < R_i)$ . If the subject does not experience the event of interest until the last inspection time, then it is considered as right-censored observation $(L_i < T_i < \infty)$. Otherwise, the observation is interval-censored.

Partly interval-censored data can be found in different fields such as, medical studies and reliability studies (Odell et al., 1992; and Lu & Meeker, 1993). In the statistical literature, many researchers discussed partly interval-censored data considering different situations. For instance, Huang (1999) discussed asymptotic properties of the Nonparametric Maximum Likelihood Estimator (NPMLE) of a distribution function based on partly interval-censored data. Kim (2003) studied the Maximum Likelihood Estimator (MLE) for the proportional hazards model considering two methods to estimate the variance-covariance matrix of the MLE of the regression parameter. Zhao et al. (2008) presented a class of generalized log-rank tests for partly interval-censored data and established their asymptotic properties. Elfaki et al. (2012) proposed Cox's model with Weibull distribution using the Expectation-Maximization (EM) algorithm to compute the maximum likelihood estimator of the regression parameter and the cumulative hazard function. On the other hand, for competing risks data, various approaches have been proposed to study the regression models regarding cause-specific hazard function. For example, Goetghebeur & Ryan (1995) proposed methods to estimate the regression coefficients for competing risks with missing causes of failure by assuming that the baseline cause-specific hazards for other risks are proportional to that for the cause of interest. Lu & Tsiatis (2001) estimated proportional hazards regression parameters using parametric models to model the probability that a missing cause is the one of interest. Gao & Tsiatis (2005) developed a method to assess the effect of covariates where the relationship between them and the cause-specific hazard for the cause of interest are described using linear transformation model. Sen et al. (2010) introduced a semiparametric Bayesian approach for the regression analysis where the cause of failure was masked for some individuals.

All the literature mentioned above are about partly interval-censored data and focus on the case when there is only one event of interest. Thus, it is chosen to study the case of competing risks when there are more than one causes of failure, which can possibly be masked, by employing the Cox's proportional hazard model. The rest of this article are arranged as follows, section 2 introduces the likelihood construction under Cox proportional hazard model framework. Section 3 describes the simulation that was conducted to evaluate the developed model. Finally, a brief discussion is presented in section 4.

Manuscript received August 5, 2016
Manuscript revised August 20, 2016

## 2. Cox Proportional Hazard (PH) Model and the Likelihood Function

### 2.1 Cox Proportional Hazard Model

Cox (PH) model is a popular mathematical model which is commonly used to assess the effect of risk factors (covariates) on the failure time through the hazard function. This model has been used extensively since it was introduced by Cox (1972). Let $T$ denotes time until the unit experience failure, and let $X$ denotes the observed vector of covariates. Then under the Cox's proportional hazard model the hazard function can be expressed as,

$$\lambda(T|X) = \lambda_0(T)e^{\beta' X} \qquad (2.1)$$

where $\lambda_0$ is an unspecified nonnegative function known as baseline hazard, and $\beta$ is the vector of regression parameters. Typically, both $\beta$ and $X$ are assumed to be constant over time. A key assumption in (2.1) is that the relative risks (or hazard ratios) are constant with time. The corresponding cumulative distribution function has the form as presented in (2.2).

$$F(T|X) = 1 - e^{-\Lambda_0(T)e^{\beta' X}} \qquad (2.2)$$

Here, $\Lambda_0$ represents the cumulative baseline hazard.

### 2.2 Likelihood Function Formulation

Suppose, N subjects were inspected until the event of interest occurred. Recalling the notations from section 1, there is $n$ exact observations and $m$ interval-censored observations which may include right-censored or left-censored observations. Then the likelihood contribution of the$i^{th}$ subject from such type of data can be expressed as,

$$L = \prod_{i=1}^{n} f(T_i|X_i) \prod_{i=n+1}^{m} [S(L_i|X_i) - S(R_i|X_i)]$$

Or

$$L = \prod_{i=1}^{n} f(T_i|X_i) \prod_{i=n+1}^{m} [F(R_i|X_i) - F(L_i|X_i)] \qquad (2.3)$$

Here $f$, $S$ ,and $F$ represent the density, survival, and distribution functions, respectively.

The likelihood function in (2.3) is suitable if there is only one cause of failure and it is always observed, however, it needs to be modified in case of competing risks. Usually, in the presence of competing risks with masking, the observed data correspond to the $i^{th}$ subject consist of $(T_i, S_i, X_i)$ where $T_i$ is the failure time, $S_i$is the subset that includes the causes which might be responsible for the failure, and $X_i$is the observed vector of covariates. The likelihood contribution result from such data built on $P(T_i, S_i|X_i)$ (Kuo & Yang, 2000) can be expressed as,

$$P(T_i, S_i|X_i) = P(T_i, C_i = j|X_i)P(S_i|T_i, C_i = j, X_i) = f_j(T_i|X_i)P(S_i|T_i, C_i = j, X_i) \qquad (2.4)$$

Here, $j = (1, \dots, K)$ and $C$ denotes the true cause of failure, whereas $K$ represents the number of causes.

Suppose there are $n_1$, $n_2$, $n_3$ exact, right-censored, and interval-censored observations, respectively, of the $N$ subjects under study. Let,$Q_{ij} = P(S_i|T_i, C_i = j, X_i)$be the masking probability. Then considering the models (2.1), (2.2), and (2.3), the likelihood function for partly interval-censored data in the presence of competing risks with masked cause of failure can be written as follows.

$$= \prod_{i=1}^{n_1} \sum_{j \in S_i} Q_{ij} f_j(T_i|X_i) \prod_{i=n_1+1}^{n_2} S(L_i|X_i) \prod_{i=n_2+1}^{n_3} \sum_{j \in S_i} Q_{ij} [F_j(R_i|X_i) - F_j(L_i|X_i)]$$

Or,

$$L = \prod_{i=1}^{n_1} \sum_{j \in S_i} Q_{ij} \lambda_j(T_i|X_i) e^{-\sum_{j=1}^{K} \Lambda_{0j}(T_i) e^{\beta_j' X_i}} \prod_{i=n_1+1}^{n_2} e^{-\sum_{j=1}^{K} \int_0^{T_i} \Lambda_{0j}(T_i) e^{\beta_j' X_i}} \times \prod_{i=n_2+1}^{n_3} \sum_{j \in S_i} Q_{ij} \left[ \int_0^{R_i} \lambda_j(t|X_i) e^{-\sum_{j=1}^{K} \Lambda_{0j}(T_i) e^{\beta_j' X_i}} dt - \int_0^{L_i} \lambda_j(t|X_i) e^{-\sum_{j=1}^{K} \Lambda_{0j}(T_i) e^{\beta_j' X_i}} dt \right] \qquad (2.5)$$

Since the full likelihood function was specified (i.e. Eq 2.5), the Bayesian analysis can be developed assigning appropriate prior distributions to unknown parameters. MCMC technique is employed to implement the Bayesian approach as the full conditional posteriors distributions of the parameters are not in tractable form.

## 3. Simulation Study

In order to evaluate the proposed model performance, a simulation has been conducted to generate failure times following the cause-specific hazards-based simulation design which is proposed by Beyersmann et al. (2009). In this simulation, it was assumed that, a simple competing risks model with two causes of failure and that each failure time has independent Weibull distribution.

Different samples are generated with different sizes and masking percentages. The preliminary obtained results suggest that the proposed model is easy to implement and performs well. Tables 1, 2, and 3 show the posterior estimations of the regression coefficients, which are reasonably close to that estimated from the model with only right-censored (RC) data (i.e. Sen et al. (2010) model). Moreover, Figure1(a, b, c) compares the posterior cumulative baseline hazards of PIC model and RC model under different masking percentages, while Figure2(a, b) compares the posterior cumulative baseline hazards of three different percentages of interval-censored observations from PIC model and the posterior cumulative baseline hazards from RC model within same masking level. Obviously, there is a substantial consistency among the different estimated cumulative baseline hazards, nevertheless, this consistency can be affected by the number of the complete observations (i.e. with observed cause of failure and exact failure time).

Table1: Posterior summaries of the regression coefficients from the two approaches(sample size 50).

| No. of masked units | **Right-Censored Data** | | **Partly Interval-Censored Data** | |
|---|---|---|---|---|
| | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| 9(23%) | 0.7111 (0.7011) | 2.5390 (0.6350) | 0.0695 (0.7199) | 1.9080 (0.5828) |
| 19(48%) | 1.2060 (0.6623) | 2.2850 (0.6187) | 0.6966 (0.6168) | 1.7380 (0.5892) |
| 33(83%) | 2.2290 (0.6233) | 1.4210 (0.6106) | 1.5850 (0.6061) | 0.9654 (0.5682) |

Standard errors are given in parentheses.

Table2: Posterior summaries of the regression coefficients from the two approaches(sample size 100).

| No. of masked units | **Right-Censored Data** | | **Partly Interval-Censored Data** | |
|---|---|---|---|---|
| | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| 17(23%) | 1.0280 (0.4122) | 1.2190 (0.3509) | 0.6372 (0.4032) | 0.9197 (0.3502) |
| 36(49%) | 1.0080 (0.4457) | 1.2110 (0.3329) | 0.7304 (0.4359) | 0.8419 (0.3376) |
| 55(74%) | 0.9003 (0.3733) | 1.3790 (0.3836) | 0.5650 (0.3747) | 1.0340 (0.3852) |

Standard errors are given in parentheses.

Table3: Posterior summaries of the regression coefficients from the two approaches(sample size 150).

| No. of masked units | **Right-Censored Data** | | **Partly Interval-Censored Data** | |
|---|---|---|---|---|
| | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| 54(46%) | 0.5310 (0.1781) | 0.1149 (0.2398) | 0.4574 (0.1829) | 0.0639 (0.2316) |
| 85(73%) | 0.3265 (0.1996) | 0.3766 (0.2137) | 0.2544 (0.1970) | 0.3046 (0.2155) |
| 105(90%) | 0.2761 (0.2203) | 0.4179 (0.1908) | 0.1842 (0.2179) | 0.3548 (0.1922) |

Standard errors are given in parentheses.

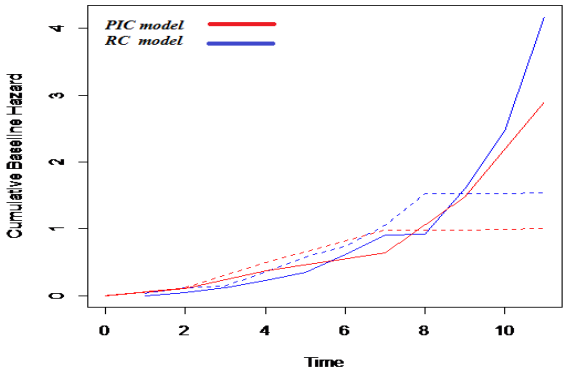


Figure1.a: Comparison of cumulative baseline hazards from the two models, PIC and RC, with 23% masked observations (simulated data with sample size 100).

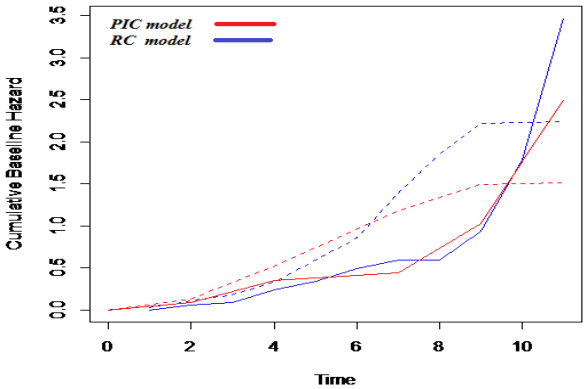


Figure1.b: Comparison of cumulative baseline hazards from the two models, PIC and RC, with 49% masked observations (simulated data with sample size 100).

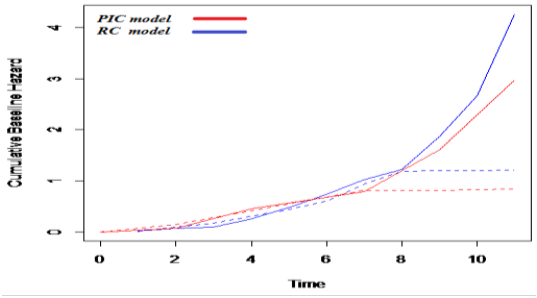


Figure1.c: Comparison of cumulative baseline hazards from the two models, PIC and RC, with 74% masked observations (simulated data with sample size 100).

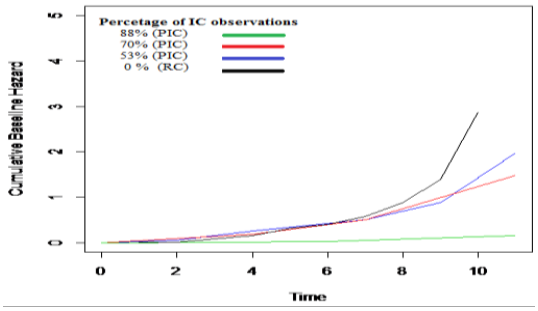


Figure2.a: Comparison of cumulative baseline hazards cause-1 of three different percentages of interval-censored (IC) observations from PIC model with the cumulative baseline hazard cause-1 from RC model.

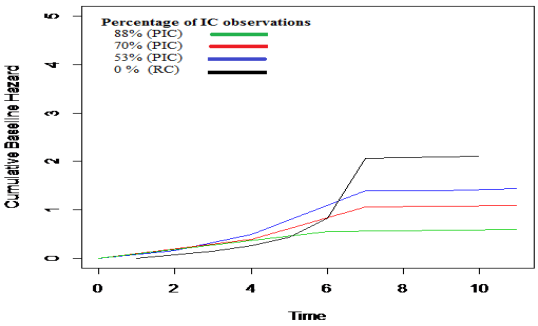


Figure2.b: Comparison of cumulative baseline hazards cause-2 of three different percentages of interval-censored (IC) observations from PIC model with the cumulative baseline hazard cause-2 from RC model.

## 4. Conclusion

A regression analysis for partly interval-censored data is studied in this paper when there are more than one causes of failure that might be masked. The Bayesian method is adopted to estimate the regression parameters and it is implemented using Markov Chain Monte Carlo techniques. The results obtained from the simulation data demonstrate that the developed model performs well and it is applicable. Further study will use real data set to analyze, based on the proposed model.

## References

[1] Beyersmann, J. et al., 2009. Simulating competing risks data in survival analysis. Statistics in Medicine, 28(6), pp.956–971. Available at: http://doi.wiley.com/10.1002/sim.3516.

[2] Cox, D.R., 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), pp.187–220. Available at: http://links.jstor.org/sici?sici=0035-9246(1972)34:2<187:RMAL>2.0.CO;2-6\nhttp://www.jstor.org.

[3] Elfaki, F.A.M., Azram, M. & Usman, M., 2012. Parametric Cox ' s Model for Partly Interval -Censored Data with Application to AIDS Studies. , 2(5), pp.352–354.

[4] Gao, G. & Tsiatis, A.A., 2005. Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. Biometrika, 92(4), pp.875–891.

[5] Goetghebeur, E. & Ryan, L., 1995. Analysis of Competing Risks Survival Data When Some Failure Types are Missing. Biomertrika, 82(4), pp.821–833.

[6] Huang, J., 1999. ASYMPTOTIC PROPERTIES OF NONPARAMETRIC ESTIMATION BASED ON PARTLY INTERVAL-CENSORED. , 9, pp.501–519.

[7] Kim, J.S., 2003. Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2), pp.489–502. Available at: http://doi.wiley.com/10.1111/1467-9868.00398.

[8] Kuo, L. & Yang, T.Y., 2000. Bayesian reliability modeling for masked system lifetime data. Statistics & Probability Letters, 47(3), pp.229–241. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167715299001601.

[9] Lu, C.J. & Meeker, W.O., 1993. Using Degradation Measures to Estimate a Time-to-Failure Distribution. Technometrics, 35(2), pp.161–174. Available at: http://www.tandfonline.com/doi/abs/10.1080/00401706.1993.10485038 [Accessed April 23, 2014].

[10] Lu, K. & Tsiatis, a a, 2001. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics, 57(4), pp.1191–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11764260.

[11] Odell, P.M., Anderson, K.M. & D'Agostino, R.B., 1992. Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull- Based Accelerated Failure Time Model. Biometrics, 48(3), pp.951–959. Available at: http://www.jstor.org/stable/2532360.

[12] Sen, A. et al., 2010. A Bayesian approach to competing risks analysis with masked cause of death. Statistics in medicine, 29(16), pp.1681–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20575048 [Accessed March 23, 2014].

[13] Zhao, X. et al., 2008. Generalized log-rank tests for partly interval-censored failure time data. Biometrical journal. Biometrische Zeitschrift, 50(3), pp.375–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18435504 [Accessed April 23, 2014].