Communications for Statistical Applications and Methods
2016, Vol. 23, No. 6, 555–562

http://dx.doi.org/10.5351/CSAM.2016.23.6.555
Print ISSN 2287-7843 / Online ISSN 2383-4757

Regression analysis of interval censored competing risk data using a pseudo-value approach

Sooyeon Kim[a], Yang-Jin Kim[1,a]

[a]Department of Statistics, Sookmyung Women's University, Korea

Abstract

Interval censored data often occur in an observational study where the subject is followed periodically. Instead of observing an exact failure time, two inspection times that include it are available. There are several methods to analyze interval censored failure time data (Sun, 2006). However, in the presence of competing risks, few methods have been suggested to estimate covariate effect on interval censored competing risk data. A subdistribution hazard model is a commonly used regression model because it has one-to-one correspondence with a cumulative incidence function. Alternatively, Klein and Andersen (2005) proposed a pseudo-value approach that directly uses the cumulative incidence function. In this paper, we consider an extension of the pseudo-value approach into the interval censored data to estimate regression coefficients. The pseudo-values generated from the estimated cumulative incidence function then become response variables in a generalized estimating equation. Simulation studies show that the suggested method performs well in several situations and an HIV-AIDS cohort study is analyzed as a real data example.

Keywords: competing risks, cumulative incidence function, GEE, interval censored data, pseudo-value approach

1. Introduction

A competing risk model has been applied to provide more valid statistical inferences when the subjects are at risk of failure from several causes (Geskus, 2015). For investigating the effect of covariate on the cause-related failure, two approaches have been applied: the cause-specific hazard (CSH) and the subdistribution hazard (SH). The CSH measures the instantaneous hazard of a particular cause-related failure by regarding other causes-related failures as censoring. The SH has a direct relation with a cumulative incidence function (CIF) that also make it possible to extend the result in order to interpret the covariate effect on the CIF (Fine and Gray, 1999). However, the SH utilizes a distribution of right censoring variable to reflect a possible contribution of competing failures to risk set. In a context of interval censored failure time, the distribution of such a censoring variable seems complex. For example, under a mixed case interval censored data, a right censoring is defined as the last inspection time of having no failure and should therefore be specified in the inspection process (Schick and Yu, 2000). Therefore, the direct application of Fine and Gray's method to interval censored data seems to be inappropriate.

Andersen *et al.* (2003) and Klein and Andersen (2005) presented a pseudo-value approach to directly model the effect of covariate on the cumulative incidence function. The objective is to replace

[1] Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Korea. E-mail: yjin@sookmyung.ac.kr

Published 30 November 2016 / journal homepage: http://csam.or.kr

the true but unknown survival times with pseudo-values. If we want to estimate $\theta = E(f(t))$ for a function f of failure time t, then it can be calculated with $1/n \sum_{i=1}^{n} f(t_i)$ with a known t_i. For example, $f(t) = I(t > u)$ becomes a survival function. Now, pseudo-values are defined as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i},$$

where $\hat{\theta}_{-i}$ is the estimator $\hat{\theta}$ calculated based on the sample leaving out the i^{th} observation. Our interest is to estimate covariate effect on failure time with censoring. In more detail, define $(\tilde{T}_i, \delta_i)$, $i = 1, \ldots, n$, where $\tilde{T}_i = T_i \wedge C_i$ and $\delta_i = I(T_i = \tilde{T}_i)$ with a censoring time C_i and a failure time T_i. Denote Z_i as a covariate vector of subject i and D_i as a cause indicator, respectively. Assume that the censoring time C_i is independent of $(\tilde{T}_i, \tilde{D}_i, Z_i)$, where $\tilde{D}_i = \delta_i D_i$. Define $N_{ik}(t) = I(\tilde{T}_i \leq t, \tilde{D}_i = k)$ as the count of the k^{th} cause failures over $[0, t]$ and $Y_i(t) = I(\tilde{T}_i \geq t)$ as at-risk indicator of subject i, respectively. Then, the cumulative incidence function is estimated as the simple version of the Aalen-Johansen estimator, that is, a special case of a multi-state model with a transition probability $\mathbf{P}_{lh}(s, t)$ and is defined as

$$\hat{F}_k(t) = \int_0^t \hat{S}(u-) d\hat{A}_k(u) = \hat{\mathbf{P}}_{0k}(0, t),$$

where

$$\hat{A}_k(t) = \int_0^t \frac{\sum_{i=1}^{n} dN_{ik}(u)}{\sum_{i=1}^{n} Y_i(u)}$$

is the Nelson-Aalen estimator of the cumulative cause specific hazard function and $\hat{S}(t)$ is the Kaplan-Meier estimator calculated using all causes failures. Klein and Andersen (2005) proposed the use of pseudo-values to estimate a covariate effect on the cumulative incidence function. These pseudo-values are calculated at the grids of time-points $w_1 < w_2 < \cdots < w_s$ chosen from observed failure times. Too many grids give a burden to calculate too many values; therefore, five to ten time-points equally spaced on the failure time scale have been suggested. Graw *et al.* (2009) showed the unbiasedness using a second-order von Mises expansion and Jacobsen and Martinussen (2016) utilized a U-statistic to establish the estimating equation and proposed a new type of variance.

Some studies have been proposed for an interval censored competing risk data. For a current status data, Jewell *et al.* (2003) developed a NPMLE and a pseudo MLE of a CIF and Jewell and Kalbfleisch (2004) suggested a modified pool adjacent violator algorithm (PAVA). For general interval censored data, Hudgens *et al.* (2001) suggested NPMLEs. However, only a few works have been done on a regression model. Sun and Shen (2009) proposed a two-stage estimation procedure for a cause specific event incidence probability and a hazard function. Hudgens *et al.* (2014) suggested a parametric regression model by extending Jeong and Fine (2006)'s method.

In this paper, we consider an extension of a pseudo-value approach to interval censored data. In Section 2, the estimation procedure is described and the performance of the suggested method is evaluated via several simulation studies in Section 3. Section 4 presents the result of the application to a real dataset and some discussions are commented on in Section 5.

2. Statistical model

Denote $l_i, r_i, \tilde{d}_i, z_i$ as the observable data set from a subject $i = 1, \ldots, n$. Assume that censoring times l_i and r_i satisfying $l_i < t_i < r_i$ are independent of a failure time t_i. The first step of a pseudo-value approach is to estimate CIF. By extending Turnbull's (1974) self-consistency algorithm, Hudgens *et al.*

(2001) suggested two versions of nonparametric estimators. The NPMLE of CIF has a cause specific time support that results in undefined regions when implementing with overall survival functions. As an alternative approach, a pseudolikelihood estimator (PLE) was calculated with a pooled time support derived from all causes interval censored data. In this paper, we adopt the PLE to estimate CIF for interval censored data. Denote $E = \cup_{l=1}^{m}[q_l, p_l]$ as an equivalence set which is a set of nonoverlapping intervals. $[q_l, p_l]$ is defined as the elements of left censoring times and the ones of right censoring times following immediately (Lindsey and Ryan, 1998; Peto, 1973). Now, denote α_{il}^k as an indicator variable, $\alpha_{il}^k = I([q_l, p_l] \in [l_i, r_i) \cap \tilde{d}_i = k)$ and let $\psi_l^k = F_k(p_l+) - F_k(q_l-)$. Then a pseudolikelihood is defined as $\prod_{i=1}^{n}[F_{d_i}(r_i-) - F_{d_i}(l_i-)]^{\delta_i} S(l_i)^{1-\delta_i}$. Based on Hudgens' Lemmas (Hudgens *et al.*, 2001) related with E and F_k, this likelihood is rewritten as

$$\prod_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{m}\left[\alpha_{il}^k\psi_l^k\right].$$

To estimate $\psi = \psi_l^k$, define an indicator $I_{il}^k = \{t_i \in [q_l, p_l] \cap \tilde{d}_i = k\}$. However, this quantity is unavailable owing to unknown failure times t_i. To solve this problem, the EM algorithm (Dempster *et al.*, 1977) is applied. In E-step, given the data $O_i = l_i, r_i, \tilde{d}_i$, and ψ,

$$E\left(I_{il}^k|\psi, O_i\right) = \mu_{il}^k(\psi) = \frac{\alpha_{il}^k\psi_l^k}{\sum_{k'=1}^{K}\sum_{l'=1}^{m}\alpha_{il'}^{k'}\psi_{l'}^{k'}} \tag{2.1}$$

is calculated. Then at M-step, the proportion of failure of cause k at the l^{th} interval is obtained with

$$\psi_l^k(\psi) = \sum_{i=1}^{n}\tilde{\psi}_l\frac{\mu_{il}^k}{\sum_{l'}\mu_{il'}^k}, \quad \tilde{\psi}_l = \sum_{k=1}^{K}\psi_l^k. \tag{2.2}$$

With initial values of $\{\psi_l^k = 1/(Km),\ k = 1,\ldots,K; l = 1,\ldots,m\}$, iterate (2.1) and (2.2) until a convergence criterion satisfies. Using the final values of $\{\hat{\psi}_l^k\}$, the NPMLE of CIF is defined as $\hat{F}_k(t) = \sum_{j=1}^{l}\hat{\psi}_j^k$ if $p_l < t < q_{l+1}$ and $\hat{F}_k(t) = \hat{\psi}_1^k + \cdots + \hat{\psi}_m^k$ if $t > p_m$.

The next step after estimating a CIF is to generate the Jackknife pseudo-values at the prefixed grids of time points $w_1 < \cdots < w_r$,

$$\theta_{ij} = n\hat{F}_k(w_j) - (n-1)\hat{F}_k^{(-i)}(w_j), \quad j = 1,\ldots,r. \tag{2.3}$$

In order to find grid points, the equivalence set is utilized. Define $e_l = \{(q_l + p_l)/2\}$, $l = 1,\ldots,m$. Then $w_1 < \cdots < w_r$ are determined as the points with equal distances among the sorted e_l's. By treating $\theta_i = (\theta_{i1},\ldots,\theta_{ir})'$ as the repeated response variables and applying a suitable link function g, the following generalized estimating equation (Liang and Zeger, 1986) is applied

$$U(\beta) = \sum_{i=1}^{n}U_i(\beta) = \sum_{i=1}^{n}\left(\frac{\partial g^{-1}(\beta' z_i)}{\partial\beta}\right)' V_i^{-1}\left\{\theta_i - g^{-1}(\beta' z_i)\right\} = 0, \tag{2.4}$$

where $g(\theta_i) = (g(\theta_{i1}),\ldots,g(\theta_{ir}))'$, $g(\theta_{ij}) = \alpha_j + \gamma' z_i = \beta' z_{ij}$ with $\beta = (\alpha_1,\ldots,\alpha_r,\gamma)$ and $V_i = \mathrm{Cov}(\theta_{il},\theta_{il'})$ is a working covariance matrix where AR(1) and independence covariance are most commonly used. Graw *et al.* (2009) remarked the condition of an asymptotic unbiasedness and expressed

Table 1: Estimation of a binary covariate effects using pseudo-values

p	n	Independence				AR(1)			
		Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
0.3	100	0.009	0.448	0.442	0.952	0.008	0.467	0.450	0.966
	200	0.019	0.317	0.312	0.965	0.019	0.332	0.320	0.968
	300	0.012	0.259	0.251	0.946	0.014	0.271	0.257	0.944
0.6	100	0.008	0.326	0.316	0.948	0.019	0.343	0.328	0.946
	200	0.003	0.225	0.226	0.964	0.005	0.237	0.236	0.965
	300	0.005	0.189	0.184	0.948	0.006	0.198	0.193	0.954

Table 2: Estimation of a continuous covariate effects using pseudo-values

p	n	Independence				AR(1)			
		Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
0.3	100	0.020	0.645	0.638	0.943	0.023	0.670	0.648	0.936
	200	0.012	0.442	0.449	0.950	0.020	0.465	0.452	0.947
	300	0.013	0.348	0.349	0.956	0.014	0.353	0.351	0.956
0.6	100	0.020	0.470	0.467	0.947	0.017	0.485	0.471	0.936
	200	0.009	0.320	0.310	0.950	0.020	0.334	0.324	0.943
	300	0.014	0.250	0.249	0.956	0.017	0.260	0.257	0.960

the Jackknife pseudo-values as the von Mises expansion to derive the asymptotic properties including the following sandwich variance.

$$\hat{\Sigma} = \hat{\Gamma}^{-1}\left(\hat{\beta}\right)\widehat{\mathrm{Var}}\left(U\left(\hat{\beta}\right)\right)\hat{\Gamma}^{-1}\left(\hat{\beta}\right), \tag{2.5}$$

where

$$\hat{\Gamma}^{-1}\left(\hat{\beta}\right) = \sum_{i=1}^{n}\left(\frac{dg^{-1}\left(\beta' z_i\right)}{d\beta}\right)' V_i^{-1}\left(\frac{dg^{-1}\left(\beta' z_i\right)}{d\beta}\right)\Big|_{\beta=\hat{\beta}},$$

and

$$\widehat{\mathrm{Var}}\left(U\left(\hat{\beta}\right)\right) = \sum_{i=1}^{n} U_i(\beta)U_i(\beta)'|_{\beta=\hat{\beta}}.$$

3. Simulation

In this section, the performance of our proposed method is demonstrated with some simulated data. 500 replications with sample size $n = 100$, 200 and 300 are generated. We consider two types of covariates: a binary covariate $Z = \{0, 1\}$ generated from a Bernoulli distribution with $p = 0.5$ and a continuous covariate $Z \sim N(0, 0.4)$. For simplicity, two causes are assumed ($K = 2$). Then a cause 1 failure time is generated with the following cumulative incidence function

$$F_1(t|Z) = 1 - \exp(-\Lambda_0(t)\exp(\gamma Z)),$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ with $\lambda_0(t) = pe^{-t}/(1 - p(1 - e^{-t}))$ and p is the occurrence probability of cause 1 failure, that is, $F_1(\infty|Z = 0) = p$. In this simulation, two values of p are used ($p = 0.3$ or 0.6). In order to make interval censored data, the different number of inspections for each subject is generated from a discrete uniform $h_i \sim U(10, 20)$ and then h_i's inspection gap times are generated

Table 3: Results of estimating effects of covariates for subtype E based on Independence

TGP	Covariate	logit model			cloglog		
		$\hat{\gamma}$	se($\hat{\gamma}$)	p-value	$\hat{\gamma}$	se($\hat{\gamma}$)	p-value
4	Age	−0.039	0.015	0.011	−0.036	0.014	0.014
	Gender (female = 1)	0.782	0.358	0.028	0.719	0.327	0.028
5	Age	−0.035	0.015	0.023	−0.032	0.014	0.028
	Gender (female = 1)	0.859	0.364	0.018	0.782	0.330	0.017
6	Age	−0.038	0.015	0.013	−0.035	0.014	0.016
	Gender (female = 1)	0.831	0.358	0.020	0.759	0.326	0.019

Table 4: Results of estimating effects of covariates for subtype E based on AR

TGP	Covariate	logit model			cloglog		
		$\hat{\gamma}$	se($\hat{\gamma}$)	p-value	$\hat{\gamma}$	se($\hat{\gamma}$)	p-value
4	Age	−0.037	0.015	0.014	−0.035	0.014	0.012
	Gender (female = 1)	0.789	0.360	0.028	0.719	0.327	0.028
5	Age	−0.035	0.015	0.019	−0.032	0.015	0.033
	Gender (female = 1)	0.872	0.371	0.018	0.789	0.333	0.028
6	Age	−0.035	0.015	0.016	−0.034	0.014	0.015
	Gender (female = 1)	0.837	0.363	0.021	0.762	0.328	0.020

from $g_l \sim U(0.05, 0.1)$. Then set (L_i, R_i) as an interval censored data when satisfying $L_i = a_{l-1} < T_i < R_i = a_l$ with $L_i = a_{l-1} = a_{l-2} + g_{l-1}, R_i = a_{l-1} + g_l$ and $a_0 = 0$. If $T_i > a_{h_i}$, T_i is regarded as a right censored with $(L_i, R_i) = (a_{h_i}, \infty)$. Now, pseudo-values were calculated at 6 grid time points $(w_1 < \cdots < w_6)$ selected from the equivalence sets described at the previous section.

Tables 1 and 2 show the simulation results for a binary covariate and a continuous covariate, respectively. Each table includes the absolute bias of $\hat{\gamma}$, empirical standard error (ESE), mean of sandwich standard error (SSE), and 95% coverage probabilities (CP) under two working covariance structures (independence and AR(1)). The estimates show that biases are small for all cases and standard errors calculated from the sandwich estimator and ESE are similar. Also, the empirical coverage probabilities almost satisfy nominal level. Compared results with two different p values, a larger $p(= 0.6)$ value with more cause 1 failures results in smaller standard errors than a smaller $p(= 0.3)$ one. Also, compared two covariance structures, the ESE under AR(1) is larger than one under independence working covariance.

4. Data analysis

The suggested method is applied to a HIV vaccine study designed to investigate the rates of HIV incidence and determine related risk factors (Hudgens *et al.*, 2002). This project was established to assess the feasibility of the vaccine to HIV in the injecting drug users (IDU) in Bangkok, Thailand. 1209 HIV seronegative IDU were enrolled and they were supposed to visit about every four months for counseling and assessment of HIV seroconversion. A total of 1124 people had at least one visit, with 133 diagnosed with HIV seroconversion among them. In this study, two subtypes strains such as subtype B strain and subtype E strain can occur and the occurrence of one subtype censors one of the other subtype. In detail, of the 133 converts, 27 and 99 subjects have subtype B and subtype E, respectively, and the remaining seven patients' subtypes were unknown. We investigate the effect of age and gender (female = 1) on the subtype E with the suggested method. To generate pseudo-values, three different numbers of grids (TGP = 4, 5, 6) are implemented using equivalence sets and two link functions (logit and complementary log-log) are applied. Tables 3 and 4 show estimated covariate

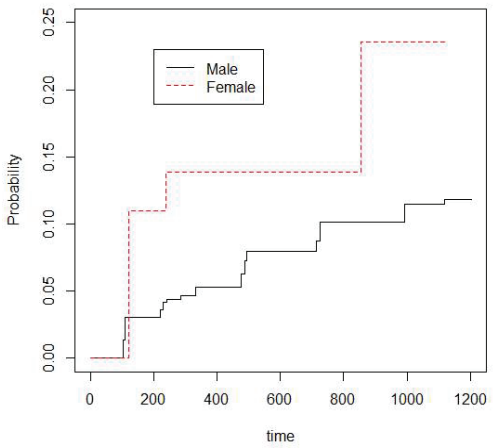

Figure 1: *Cumulative incidence functions of subtype E for male and female group.*

effects under two different covariance structures. According to results, older patients have a lower incidence rate than younger ones and female patients have a higher incidence rate than male patients. The results are similar among two covariance structures, two link functions and three different numbers of time grids. Figure 1 shows the estimated CIFs of subtype E of two genders by applying the PLE described in Section 2. From it, female patients seem to have a higher occurrence rate than male patients during the whole time period. However, this result should be dealt carefully because there were only fourteen female patients among the 133 ones.

5. Discussion

As the most widely used competing risk regression model, Fine and Gray's subdistribution hazard model is based on an inverse probability censoring weighting technique that requires a specification of the distribution of censoring times. However, for interval censored data with different censoring structure, such specification is neither amenable nor possible. Therefore, we suggest the extension of the pseudo-value regression model proposed by Klein and Andersen (2005) to interval censored data instead of applying the Fine and Gray's model. Jackknife pseudo-values are obtained from the estimated cumulative incidence function. The suggested approach can be implemented with an ordinary program since all calculations are performed using R package *pseudo* once estimating the CIFs. Simulation results show the proposed method results in consistent estimates and desirable coverage probabilities. The suggested pseudo-value approach can be applied to a multi-state model with several transition states and corresponding probabilities. In the studies on a multi-state model, Barrett *et al.* (2011) developed an interval censored semi-competing risk model that allowed a one direction transition between terminal events. Kim (2014) analyzed a bivariate current status data using a multi-state model. Commenges (2002) provided a review of the multi-state model with interval censored data. Another possible future work is about a missing cause. The IDU dataset includes seven subjects with unidentified HIV subtypes. Even though the number of the subjects with a missing cause is small in this dataset, this problem has commonly occurred in competing risk data; therefore, suitable methods should be developed for interval censored data (Goetghebeur and Ryan, 1995; Lu and Tsiatis, 2001; Moreno-Betancur and Latouche, 2013).

Acknowledgements

This work was supported by the National Research Foundation of Korea research grant NRF-2014R1 A2A2A01003567. We are grateful to Professor Michael G. Hudgens for providing with IDU cohort study dataset.

References

Andersen PK, Klein JP, and Rosthøj S (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models, *Biometrika*, **90**, 15–27.

Barrett JK, Siannis F, and Farewell VT (2011). A semi-competing risks model for data with interval-censoring and informative observation: an application to the MRC cognitive function and ageing study, *Statistics in Medicine*, **30**, 1–10.

Commenges D (2002). Inference for multi-state models from interval-censored data, *Statistical Methods and Medical Research*, **11**, 167–182.

Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B (Methodological)*, **39**, 1–38.

Fine JP and Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–509.

Geskus RB (2015). *Data Analysis with Competing Risks and Intermediate States*, CRC Press, Boca Raton, FL.

Goetghebeur E and Ryan L (1995). Analysis of competing risks survival data when some failure types are missing, *Biometrika*, **82**, 821–833.

Graw F, Gerds TA, and Schumacher M (2009). On pseudo-values for regression analysis in competing risks models, *Lifetime Data Analysis*, **15**, 241–255.

Hudgens MG, Li C, and Fine JP (2014). Parametric likelihood inference for interval censored competing risks data, *Biometrics*, **70**, 1–9.

Hudgens MG, Longini IM, Vanichseni S, Hu DJ, Kitayaporn D, Mock PA, Halloran ME, Satten GA, Choopanya K, and Mastro TD (2002). Subtype-specific transmission probabilities for human immunodeficiency virus type 1 among injecting drug users in Bangkok, Thailand, *American Journal of Epidemiology*, **155**, 159–168.

Hudgens MG, Satten GA, and Longini IM (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation, *Biometrics*, **57**, 74–80.

Jacobsen M and Martinussen T (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observation, *Scandinavian Journal of Statistics*, **43**, 845–862.

Jeong JH and Fine J (2006). Direct parametric inference for the cumulative incidence function, *Journal of Royal Statistical Society Series C (Applied Statistics)*, **55**, 187–200.

Jewell NP and Kalbfleisch JD (2004). Maximum likelihood estimation of ordered multinomial parameters, *Biostatistics*, **5**, 291–306.

Jewell NP, Van der Laan M, and Henneman T (2003). Nonparametric estimation from current status data with competing risks, *Biometrika*, **90**, 183–197.

Kim YJ (2014). Regression analysis of bivariate current status data using a multistate model, *Communications in Statistics - Computation and Simulation*, **43**, 462–475.

Klein JP and Andersen PK (2005). Regression modeling of competing risks data based on pseudo values of the cumulative incidence function, *Biometrics*, **61**, 223–229.

Liang KY and Zeger SL (1986). Longitudinal data analysis using generalized linear model, *Biometrika*, **73**, 13–22.

Lindsey JC and Ryan LM (1998). Methods for interval censored data, Statistics in Medicine, *Statistics in Medicine*, **17**, 219–238.

Lu K and Tsiatis AA (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure, *Biometrics*, **57**, 1191–1197.

Moreno-Betancur M and Latouche A (2013). Regression modeling of the cumulative incidence function with missing cause of failure using pseudo-values, *Statistics in Medicine*, **32**, 3206–3223.

Peto R (1973). Experimental survival curves for interval-censored data, *Applied Statistics*, **22**, 86–91.

Schick A and Yu Q (2000). Consistency of the GMLE with mixed case interval-censored data, *Scandinavian Journal of Statistics*, **27**, 45–55.

Sun J (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

Sun J and Shen J (2009). Efficient estimation for the proportional hazard model with competing risks and current status data, *Canadian Journal of Statistics*, **37**, 592–606.

Turnbull BW (1974). Nonparametric estimation of survivorship function with doubly censored data, *Journal of American Statistical Association*, **69**, 169–173.

Received August 11, 2016; Revised November 11, 2016; Accepted November 20, 2016