American Journal of Theoretical and Applied Statistics
2021; 10(3): 167-174
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20211003.15
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Sieve Estimation for Mixture Cure Rate Model with Informatively Interval-Censored Failure Time Data

Yeqian Liu, James Plott, Yingxiao Huang

Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA

Email address:
Yeqian.liu@mtsu.edu (Yeqian Liu)

To cite this article:
Yeqian Liu, James Plott, Yingxiao Huang. Sieve Estimation for Mixture Cure Rate Model with Informatively Interval-Censored Failure Time Data. *American Journal of Theoretical and Applied Statistics*. Vol. 10, No. 3, 2021, pp. 167-174. doi: 10.11648/j.ajtas.20211003.15

Received: June 10, 2021; **Accepted**: June 23, 2021; **Published**: June 29, 2021

Abstract: In biomedical and public health studies, interval-censored data arise when the failure time of interest is not exactly observed and instead only known to lie within an interval. Furthermore, the failure time and censoring time may be dependent. There may also exist a cured subgroup, meaning that a proportion of study subjects are not susceptible to the failure event of interest. Many authors have investigated inference procedure for interval-censored data. However, most existing methods either assume no cured subgroup or apply only to limited situations such that the failure time and the observation time have to be independent. To take both cured subgroups and informative censoring into consideration for regression analysis of interval-censored data, we employ a mixture cure model and propose a sieve maximum likelihood estimation approach using Bernstein Polynomials. A novel expectation-maximization algorithm with the use of subject-specific independent log-normal latent variable is developed to obtain the numerical solutions of the model. The robustness and finite-sample performance of the proposed method in terms of estimation accuracy and predictive power is evaluated by an extensive simulation study which suggest that the proposed method works well for practical situations. In addition, we provide an illustrative example using NASA's hypobaric decompression sickness database (HDSD).

Keywords: Interval-censoring, Cure Rate Model, Informative Censoring, Sieve Maximum Likelihood Estimation, EM Algorithm, Bernstein Polynomial

1. Introduction

This paper discusses regression analysis of interval-censored data when there exists the informative censoring issue and a cured subpopulation. Interval-censored data occur naturally and frequently in randomized clinical trials, where the exact time of event occurrence is unknown but the event time is only known to lie within an interval. In regular survival analysis, it is usually assumed that every subject is susceptible to the failure event. However, there may exist a subpopulation which is cured or immune to the failure event.

Several type of cure models are proposed to deal with this scenario [1-4]. In addition, [3], [5-6] studied the cure rate model for the analysis of interval-censored failure time data.

Another challenging issue for this problem is having correlated failure time and censoring. Many authors have developed regression procedure to deal with informative censoring [2-3], [7-9]. Furthermore, [8-9] considered the cure rate models with informatively right censored data. It is also proved by [10] that ignoring of the cured subpopulation could result in an overestimation of the survival time. And the estimation could be seriously biased if the informative censoring is not considered in the model [11]. However, it does not seem to exist an established inference procedure for interval-censored data that takes both cured subgroup and informative censoring into account.

In this paper, we present a sieve estimation procedure for analyzing interval-censored data that is able to address both cured subgroups and informative censoring using the mixture cure rate model. Cox's proportional hazards model is used for modeling both failure time and censoring time. A latent variable is introduced in order to directly characterize the correlation between failure time and the dependence between failure time and censoring time. The remainder of the article is organized as follows. Section 2 introduces notation, underlying model as well as the parameter estimate procedure for informative interval censored data. A sieve maximum likelihood estimation procedure is then be

described in Section 3. An EM algorithm is developed and Bernstein polynomials is used to approximate unknown functions. Section 4 presents some results obtained from an extensive simulation study conducted to assess the performance of the proposed methodology and an illustrative example is provided in Section 5. Section 6 contains some discussion and concluding remarks.

2. Assumptions, Models and Likelihood Function

In a clinical study with a cured subpopulation, let T denote the failure time and assume the failure event of each patients is observed within a time interval [L, R]. X is the covariates of patients. Now have interval-censored survival data. Define the cure indicator variable $U = 0$ if the subject is cured and nonsusceptible and $U = 1$ otherwise, and suppose that we can write T as

$$T = U T^* + (1 - U)\infty,$$

where $T^* < \infty$ denotes the failure time of a susceptible subject. The cure indicator U is modeled by the logistic model [6]

$$P(U = 1|Z) = \frac{\exp(Z'\alpha)}{1+\exp(Z'\alpha)} \quad (1)$$

Here Z denotes the vector of covariates that may have effects on U, which may be the same as, a part of or different from X, and α denotes a vector of regression parameters as β and γ. Now assume a clinical study that has n independent subjects. For the i-th subject , let T_i denote the event time and let L_i and R_i be the left and right endpoint of the interval censored data. Also assume that the interval censored time is correlated with failure time T_i. Define $W_i = R_i - L_i$, the gap time between the two observation times. In the following, we model the hazard functions of T_i, R_i and W_i through an unobserved or latent vector $b_i = (b_{1i}, b_{2i}, b_{3i})$ by assuming that T_i, L_i and W_i are independent conditional on X_i and b_i. For subjects with $U = 1$, the cumulative hazard function of T at t is specified by

$$\lambda_i^{(T)}(t|X_i, b_i) = \lambda_{t0}(t)\exp\{\beta_t'X_i + b_{1i}\}. \quad (2)$$

Conditional on X_i and b_i, we assume that the hazard functions for L_i and W_i follow the Cox model:

$$\lambda_i^{(L)}(t|X_i, b_i) = \lambda_{l0}(t)\exp\{\beta_l'X_i + b_{2i}\} \quad (3)$$

$$\lambda_i^{(W)}(t|X_i, b_i) = \lambda_{w0}(t)\exp\{\beta_w'X_i + b_{3i}\} \quad (4)$$

respectively, where β_t, β_l and β_w are $p \times 1$ vectors of unknown regression parameters, and $\lambda_{t0}(t)$, $\lambda_{l0}(t)$ and $\lambda_{w0}(t)$ are unknown baseline hazard functions. Moreover, we assume that $b_i \sim$ i.i.d. $N(0, \Sigma)$. The baseline covariates Z and X may share some common components and Z includes 1 so that α contains the intercept term. Denote $\delta = I(R < \infty)$. Denote the observation from a single subject by $\mathcal{O}_i = (L_i, W_i, \delta_i, Z_i, X_i)$. The parameter need to be estimated $\theta = (\alpha, \beta_t, \beta_l, \beta_w, \Lambda_t(\cdot), \Lambda_l(\cdot), \Lambda_w(\cdot))$.
$\theta_0 = (\alpha_0, \beta_t, \beta_l, \beta_w, \Lambda_{t0}(\cdot), \Lambda_{l0}(\cdot), \Lambda_{w0}(\cdot))$ where $\Lambda_{t0} = \int_0^t \lambda_{t0}(s)ds$, $\Lambda_{l0} = \int_0^t \lambda_{l0}(s)ds$, $\Lambda_{w0} = \int_0^t \lambda_{w0}(s)ds$ Denote $S_p(t|X, Z, b) = p(Z) + \{1 - p(Z)\}\{1 - \exp(-\exp(\beta_t'X_i + b_1)\Lambda_{t0}(t))\}$. $F_p(t|X, Z, b) = 1 - S_p(t|X, Z, b)$. Conditional on $(L_i, W_i, X_i, Z_i, b_i)$, the likelihood of the observation from subject i has the following form:

$$L_{\delta_i|L_i,W_i,b_i} = \{S_p(L_i|X, Z, b) - S_p(R_i|X, Z, b)\}^{\delta_i}\{S_p(L_i|X, Z, b)\}^{1-\delta_i} \quad (5)$$

Conditional on (X_i, b_i), the likelihood functions related to L_i and W_i are given by

$$L_{L_i|b_i} = \{\lambda_{l0}(L_i)\exp\{\beta_l'X_i + b_{2i}\}\exp\{-\exp\{\beta_l'X_i + b_{2i}\}\Lambda_{l0}\}. \quad (6)$$

$$L_{W_i|b_i} = \{\lambda_{w0}(L_i)\exp\{\beta_w'X_i + b_{2i}\}\exp\{-\exp\{\beta_w'X_i + b_{2i}\}\Lambda_{w0}\}^{\delta_i}. \quad (7)$$

Then likelihood function of θ for a single observation $\mathcal{O} = (L, W, \delta, Z, X)$ is

$$L(\theta, \mathcal{O}) = \int L_{\delta|L,W,X,Z,b}L_{L|X,b}L_{W|X,b}f(b; \Sigma)db. \quad (8)$$

where $f(b; \Sigma)$ is the density function of b.

The full likelihood function of θ given $\mathcal{W}$ is $L(\theta) = \Pi_{i=1}^n L(\theta, \mathcal{O}_i)$.

3. Inference Procedure

We propose to use a sieve method to approximate Λ_t for alleviating the computation burden. We approximate $\Lambda_t(\cdot)$ by Bernstein polynomial functions on $I = [a, b]$, where a, b are the lower and upper bounds of the finite observation times $\{L_i, R_i\delta_i: i = 1,2, \cdots, n\}$. $\Lambda_{t0}(\cdot)$, $\Lambda_{l0}(\cdot)$ $\Lambda_{w0}(\cdot)$ are approximated by

$$\Lambda_{1n}(t) = \sum_{j=0}^m \exp(\gamma_j)B_j(t, m, a, b) = B(t)'\eta. \quad (9)$$

where $B(t) = (B_0(t, m, a, b), \cdots, B_{m+1}(t, m, a, b))'$, $\eta = (\exp(\gamma_0), \cdots, \exp(\gamma_{m+1}))'$ and $B_j(t, m, a, b) = {}_m\!_j\ (\frac{t-a}{b-a})^j(1 - \frac{t-a}{b-a})^{m-j}$, $m = o(n^v)$ for some $v > 0$.

Because the integrated form of the log-normal frailty is very complicated, the EM algorithm is used to perform the maximum likelihood estimates (MLEs) [6].

EM Algorithm

E-step:

The density function of log-normal frailty b_i is

$$f(b_i|\mathcal{O}_i, \theta) = L_i^{-1}(\theta; \mathcal{O}_i)L_{\delta_i|L_i,W_i,b_i}L_{L_i|b_i}L_{W_i|b_i}f(b_i; \Sigma) \quad (10)$$

Then write the complete data likelihood as

$$L_c(\theta;\mathcal{O},b)=\prod_{i=1}^{n} L_{\delta_i|L_i,W_i,b_i}L_{L_i|b_i}L_{W_i|b_i}f(b_i;\Sigma) \quad (11)$$

Take the logarithm of the likelihood function (11) as

$$l_c(\theta;\mathcal{O},b)=\sum_{i=1}^{n}\{\log L_{\delta_i|L_i,W_i,b_i}+\log L_{L_i|b_i}+\log L_{W_i|b_i}+\log f(b_i;\Sigma)\} \quad (12)$$

Compute the conditional expectation of the complete likelihood function, and compute the following integration

$$E\{h(b_i)|\mathcal{O}_i,\hat{\theta}^{(k)}\}=\int h(b_i)f(b_i|\mathcal{O}_i,\hat{\theta}^{(k)})db_i \quad (13)$$

where $h(b_i)$ is a function involving b_i and $\hat{\theta}^{(k)}$ is the estimate of θ after the kth iteration. We calculate this integral the Monte Carlo method. Let $\hat{E}\{h(b_i)|\mathcal{O}_i,\hat{\theta}^{(k)}\}$ denote the approximate value of $E\{h(b_i)|\mathcal{O}_i,\hat{\theta}^{(k)}\}$ using the Monte Carlo approach.

Let $Q_i(b_i;\mathcal{O}_i,\theta)=L_{\delta_i|L_i,W_i,b_i}L_{L_i|b_i}L_{W_i|b_i}$, $v_i=\Sigma^{-1/2}b_i$. It follows trivially that

$$E\{h(b_i)|\mathcal{O}_i,\theta\}=\frac{\int h(b_i(v_i))Q_i(b_i(v_i);\mathcal{O}_i,\theta)\exp(-v_{i\prime}v_i)dv_i}{\int Q_i(b_i;\mathcal{O}_i,\theta)\exp(-v_{i\prime}v_i)dv_i} \quad (14)$$

Then we approximate $E\{h(b_i)|\mathcal{O}_i,\theta\}$ by the Monte Carlo approach.

$$\hat{E}\{h(b_i)|\mathcal{O}_i,\theta\}=\frac{\sum_{l=1}^{L}h(b_i(v_{il}))Q_i(b_i(v_{il});\mathcal{O}_i,\theta)}{\sum_{l=1}^{L}Q_i(b_i(v_{il});\mathcal{O}_i,\theta)} \quad (15)$$

where $\{b_{il},l=1,\cdots,L\}$ are generated from multivariate standard normal distribution with mean zero and covariance matrix I, where is the identity matrix.

M-step:

Maximizing the conditional expectation of (13) with respect to θ at the $(k+1)$th iteration yields the updated estimator $\theta^{(k+1)}$. Denote by λ_{l0} (λ_{w0}) the vector of the discrete baseline hazard function of the observation time points Li (W_i), $i=1,\cdots,n$. Let $\psi=(\theta',\lambda_{l0}',\lambda_{w0}')'$ and $\hat{\psi}=(\hat{\theta}',\hat{\lambda}_{l0}',\hat{\lambda}_{w0}')'$ the estimate for ψ. Also let $l(\psi;\mathcal{O},b)$ be the logarithm of the complete data likelihood function. It can be checked that

$$l(\psi;\mathcal{O},b)=\sum_{i=1}^{n}[\delta_i\log\{S_p(L_i|X,Z,b)-S_p(R_i|X,Z,b)\}+(1-\delta_i)\log\{S_p(L_i|X,Z,b)\}]$$

$$+\sum_{i=1}^{n}[\log\lambda_{l0}(L_i)+\{\beta_{l\prime}X_i+b_{2i}\}-\exp\{\beta_{l\prime}X_i+b_{2i}\}\Lambda_{l0}]+\sum_{i=1}^{n}\delta_i[\log\lambda_{w0}(L_i)+\{\beta_{w\prime}X_i+b_{2i}\}-\exp\{\beta_{w\prime}X_i+b_{2i}\}\Lambda_{w0}]. \quad (16)$$

Then the components of $\partial l(\theta;\mathcal{O},b)/\partial\theta$ are

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\gamma_j}=\sum_{i=1}^{n}\frac{\partial\log L_{\delta_i|L_i,W_i,b_i}}{\partial\gamma_j},j=0,\cdots,m+1 \quad (17)$$

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\beta_t}=\sum_{i=1}^{n}\frac{\partial\log L_{\delta_i|L_i,W_i,b_i}}{\partial\beta_t} \quad (18)$$

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\beta_l}=\sum_{i=1}^{n}X_i[1-\exp\{X_{i\prime}\beta_l+b_{2i}\}\Lambda_{l0}(L_i)] \quad (19)$$

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\lambda_{l0}(L_i)}=\sum_{i=1}^{n}\frac{1}{\lambda_{l0}(L_i)}-\sum_{k=1}^{n}I(L_k\geq L_i)\exp\{X_{k\prime}\beta_l+b_{2k}\} \quad (20)$$

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\beta_w}=\sum_{i=1}^{n}\delta_iX_i[1-\exp\{X_{i\prime}\beta_w+b_{3i}\}\Lambda_{w0}(L_i)] \quad (21)$$

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\lambda_{w0}(W_i)}=\sum_{i=1}^{n}\delta_i\{\frac{1}{\lambda_{w0}(W_i)}-\sum_{k=1}^{n}I(W_k\geq W_i)\exp\{X_{k\prime}\beta_w+b_{3k}\}\} \quad (22)$$

$$\frac{\partial l(\psi;\mathcal{O},b)}{\partial\Sigma^{-1}}=n\Sigma-\sum_{i=1}^{n}b_ib_{i\prime}-\frac{1}{2}\,diag\,\{n\Sigma-\sum_{i=1}^{n}b_ib_{i\prime}\}$$

First let $S^{(T)}(\beta_t,\gamma;\mathcal{O}_i,b_i)=(\frac{\partial l(\theta;\mathcal{O}_i,b_i)}{\partial\beta_{t\prime}},\frac{\partial l(\theta;\mathcal{O}_i,b_i)}{\partial\gamma\prime})'$. Then we update the parameter estimators for $(\beta_{t\prime},\gamma')'$,

$$S^{(T)}(\beta_t,\gamma)=\sum_{i=1}^{n}\hat{E}\{S^{(T)}(\beta_t,\gamma;\mathcal{O}_i,b_i)|\mathcal{O}_i\} \quad (23)$$

Note that the estimates of β_t,γ,β_l and β_w are updated at the $(k+1)$th iteration using the one-step Newton–Raphson algorithm. Solving (20) and (22) actually yields Aalen-Breslow type estimators for λ_{l0} and λ_{w0} as follows:

$$\hat{\Lambda}_{l0}(t)=\sum_{i=1}^{n}\int_0^t\frac{dN_{li}(s)}{\sum_{k=1}^{n}I(L_k\geq s)\exp\{\beta_{l\prime}X_k\}E\{b_{2k}|\mathcal{O}_k\}} \quad (24)$$

and

$$\hat{\Lambda}_{w0}(t)=\sum_{i=1}^{n}\int_0^t\frac{dN_{wi}(s)}{\sum_{k=1}^{n}\delta_kI(W_k\geq s)\exp\{\beta_{w\prime}X_k\}E\{b_{3k}|\mathcal{O}_k\}}. \quad (25)$$

Plugging them into (19) and (21) leads to score equations

$$S^{(L)}(\beta_l) = \sum_{i=1}^{n} \int_0^{\infty} [X_i - \bar{X}^{(l)}(t;\beta_l)] dN_{li}(t) \tag{26}$$

$$S^{(W)}(\beta_w) = \sum_{i=1}^{n} \delta_i \int_0^{\infty} [X_i - \bar{X}^{(w)}(t;\beta_w)] dN_{wi}(t) \tag{27}$$

where $N_{li} = I(L_i \le t)$, $N_{wi} = I(W_i \le t)$, and

$$\bar{X}^{(l)}(t;\beta_l) = \frac{\sum_{k=1}^{n} I(L_k \ge t) X_k \exp\{\beta_{l'} X_k\} \hat{E}\{b_{2k}|\mathcal{O}_k\}}{\sum_{k=1}^{n} I(L_k \ge t) \exp\{\beta_{l'} X_k\} \hat{E}\{b_{2k}|\mathcal{O}_k\}}$$

$$\bar{X}^{(w)}(t;\beta_w) = \frac{\sum_{k=1}^{n} \delta_k I(W_k \ge t) X_k \exp\{\beta_{l'} X_k\} \hat{E}\{b_{3k}|\mathcal{O}_k\}}{\sum_{k=1}^{n} \delta_k I(W_k \ge t) \exp\{\beta_{l'} X_k\} \hat{E}\{b_{3k}|\mathcal{O}_k\}}$$

Once the new estimates of β_l and β_w are obtained, Λ_{l0} and Λ_{w0} can be estimated by the (24) and (25). Finally, the closed-form maximum likelihood estimate of Σ is

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \hat{E}\{b_i b_{i'} | \mathcal{O}_i\}. \tag{28}$$

For the variance estimation of the parameters, we employed the Louis formula [12], and calculate the observed information matrix $I(\hat{\psi})$ by

$$I(\hat{\psi}) = -E\{\frac{\partial^2 l(\psi;\mathcal{O},b)}{\partial\psi\partial\psi'} | \mathcal{O}, \hat{\psi}\} - E\{\frac{\partial l(\psi;\mathcal{O},b)}{\partial\psi} \frac{\partial l(\psi;\mathcal{O},b)}{\partial\psi'} | \mathcal{O}, \hat{\psi}\} \tag{29}$$

4. A Simulation Study

In this section, an comprehensive simulation study is conducted to evaluation the finite sample performance of the proposed estimator. The uniform distribution and normal distribution is used to generate the covariates. To generate the data, we first generated the X_i's and the b_i's and then the L_i and $R_i's$ from model (3) and (4) with $\Lambda_l(t) = 0.01\,t^2$ and $\Lambda_w(t) = 0.02\,t^2$. We first generate the cure indicator U_i's based on model (1) and generate the T_i^*'s based on model (2) with $\Lambda_T(t) = 0.05\,t^2$. The simulation results is based on 1000 replications with the sample size of 200 or 400.

The estimated regression parameters $\alpha, \beta_t, \beta_l, \beta_w$ and σ is given in tables 1-4. In the simulation setup, Z_i's follow the uniform distribution over $(0,2)$, X_i's follow $N(0,1)$, and $\alpha_0 = 1$, $(\beta_t, \beta_l, \beta_w) = (0,0,0)$, $(0,0.5,1)$, $(1,0.5,0)$, or $(0.5,0.5,0.5)$. Table 1 corresponds to the non-informative censoring situation with $\sigma_0 = 0$, and Tables 2-3 correspond to the informative censoring situation with σ_0 =0.5 and $\sigma_0 = 1$ respectively. The results in these tables include estimated bias (Bias), the sample standard deviation (SSD) of the estimates, the estimated standard errors (SEE), and the 95% empirical coverage probability (CP). Similar to [13], we took the degree of the Bernstein Polynomials $m = [n^{1/4}] = 3$ for $n = 200$ and $m = [n^{1/4}] = 4$ for $n = 400$.

The simulation results indicate that the proposed estimator performs pretty well as the bias is small and variance estimation is close to sample variance of the simulated data. In addition, the coverage probability is close to 95% which indicates that the proposed estimator is asymptotically normally distributed. Moreover, as the sample size increases, both the biased and estimated variance decreases which is to be expected.

The simulation results for the two covariate scenario are displayed in Table 4 with $\beta_t = (1,1.5)'$, $\beta_l = (0.5,1)'$, $\alpha_0 = (1,1)'$, $\beta_w = (1,1)'$ or $(1,-1)'$, and $\sigma_0 = 0.5$. The covariates for cure model Z_1 and Z_2 were generated from the standard normal distribution and the uniform distribution over $(-1,1)$, respectively, and the covariates for Cox model X_1 and X_2 were generated from the Bernoulli distribution with $p = 0.5$ and the standard normal distribution, respectively. The results indicate that our proposed method also works well for multiple covariates scenarios. Furthermore, we tried different values for the degree of Bernstein Polynomials and found they all gave similar simulation results. Thus, the results are robust for different degree of polynomial.

Table 1. *Results on estimations of the regression coefficients based on the simulated data with one covariate and σ=0.*

Z~ Unif(0,2), X~ N(0,1), α=1, Cure%=35%										
n	***J***	**Par**	**Bias**	**SSD**	**SEE**	**CP**	**Bias**	**SSD**	**SEE**	**CP**
			$\widehat{\beta_t}$=0, $\widehat{\beta_l}$=0, $\widehat{\beta_w}$=0, Cr=45%				$\widehat{\beta_t}$=0, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=1, Cr=50%			
200	3	$\hat{\alpha}$	0.023	0.145	0.131	0.935	0.063	0.148	0.169	0.955
		$\widehat{\beta_t}$	0.048	0.171	0.178	0.938	0.061	0.173	0.198	0.971
		$\widehat{\beta_l}$	-0.019	0.124	0.134	0.958	0.027	0.136	0.142	0.935
		$\widehat{\beta_w}$	0.025	0.136	0.141	0.951	0.020	0.125	0.141	0.950
		$\hat{\sigma}$	0.005	0.095	0.081	0.943	0.007	0.082	0.105	0.965
400	4	$\hat{\alpha}$	0.018	0.118	0.125	0.953	0.029	0.113	0.106	0.947
		$\widehat{\beta_t}$	0.026	0.138	0.132	0.946	0.045	0.179	0.185	0.942
		$\widehat{\beta_l}$	0.016	0.097	0.103	0.955	-0.021	0.091	0.86	0.955
		$\widehat{\beta_w}$	0.022	0.111	0.109	0.941	0.019	0.128	0.147	0.970
		$\hat{\sigma}$	-0.005	0.061	0.055	0.948	0.002	0.081	0.088	0.953
			$\widehat{\beta_t}$=1, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=0, Cr=35%				$\widehat{\beta_t}$=0.5, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=0.5, Cr=40%			

Z~ Unif(0,2), X~ N(0,1), α=1, Cure%=35%									
n	***J***	**Par**	**Bias**	**SSD**	**SEE**	**CP**	**Bias**	**SSD**	**SEE**	**CP**
200	3	$\hat{\alpha}$	0.051	0.178	0.162	0.938	0.058	0.161	0.168	0.967
		$\widehat{\beta_t}$	0.056	0.201	0.206	0.961	0.054	0.149	0.132	0.944
		$\widehat{\beta_l}$	0.031	0.141	0.144	0.951	0.047	0.138	0.142	0.938
		$\widehat{\beta_w}$	0.012	0.112	0.103	0.957	0.060	0.151	0.156	0.960
		$\hat{\sigma}$	-0.004	0.081	0.075	0.940	0.003	0.072	0.077	0.951
400	4	$\hat{\alpha}$	0.048	0.146	0.159	0.948	0.051	0.126	0.131	0.956
		$\widehat{\beta_t}$	0.044	0.128	0.126	0.962	0.062	0.151	0.160	0.958
		$\widehat{\beta_l}$	0.031	0.115	0.119	0.947	-0.027	0.132	0.145	0.944
		$\widehat{\beta_w}$	0.025	0.109	0.101	0.955	0.042	0.172	0.161	0.946
		$\hat{\sigma}$	0.008	0.041	0.052	0.951	-0.001	0.035	0.032	0.942

Table 2. *Results on estimations of the regression coefficients based on the simulated data with one covariate and σ=0.5.*

Z~ Unif(0,2), X~ N(0,1), α=1, Cure%=35%									
n	***J***	**Par**	**Bias**	**SSD**	**SEE**	**CP**	**Bias**	**SSD**	**SEE**	**CP**
			$\widehat{\beta_t}$=0, $\widehat{\beta_l}$=0, $\widehat{\beta_w}$=0, Cr=45%				$\widehat{\beta_t}$=0, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=1, Cr=50%			
200	3	$\hat{\alpha}$	0.036	0.138	0.144	0.967	0.043	0.151	0.164	0.957
		$\widehat{\beta_t}$	0.018	0.131	0.148	0.947	0.068	0.189	0.199	0.971
		$\widehat{\beta_l}$	0.020	0.129	0.131	0.953	0.032	0.166	0.142	0.946
		$\widehat{\beta_w}$	0.014	0.122	0.127	0.956	0.021	0.116	0.110	0.947
		$\hat{\sigma}$	0.009	0.072	0.077	0.944	0.004	0.057	0.063	0.959
400	4	$\hat{\alpha}$	0.031	0.108	0.115	0.953	0.027	0.113	0.106	0.947
		$\widehat{\beta_t}$	0.022	0.128	0.122	0.946	0.029	0.149	0.140	0.942
		$\widehat{\beta_l}$	0.014	0.145	0.167	0.955	0.036	0.156	0.163	0.953
		$\widehat{\beta_w}$	0.035	0.152	0.145	0.939	0.048	0.172	0.178	0.942
		$\hat{\sigma}$	-0.010	0.051	0.035	0.949	0.003	0.061	0.078	0.937
			$\widehat{\beta_t}$=1, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=0, Cr=35%				$\widehat{\beta_t}$=0.5, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=0.5, Cr=40%			
200	3	$\hat{\alpha}$	0.051	0.128	0.112	0.938	0.048	0.113	0.133	0.970
		$\widehat{\beta_t}$	0.068	0.161	0.168	0.963	0.041	0.169	0.173	0.941
		$\widehat{\beta_l}$	0.042	0.141	0.145	0.954	0.038	0.150	0.156	0.946
		$\widehat{\beta_w}$	0.038	0.132	0.135	0.943	0.046	0.138	0.141	0.953
		$\hat{\sigma}$	-0.002	0.042	0.053	0.961	0.004	0.072	0.065	0.934
400	4	$\hat{\alpha}$	0.038	0.092	0.105	0.946	0.041	0.106	0.102	0.953
		$\widehat{\beta_t}$	0.064	0.178	0.176	0.965	0.035	0.131	0.140	0.956
		$\widehat{\beta_l}$	0.042	0.125	0.139	0.945	0.027	0.112	0.105	0.944
		$\widehat{\beta_w}$	0.029	0.112	0.105	0.951	0.044	0.123	0.135	0.962
		$\hat{\sigma}$	0.001	0.038	0.047	0.953	-0.006	0.055	0.042	0.941

Table 3. *Results on estimations of regression coefficients based on the simulated data with one covariate and σ=1.*

Z~ Unif(0,2), X~ N(0,1), α=1, Cure%=35%									
n	***J***	**Par**	**Bias**	**SSD**	**SEE**	**CP**	**Bias**	**SSD**	**SEE**	**CP**
			$\widehat{\beta_t}$=0, $\widehat{\beta_l}$=0, $\widehat{\beta_w}$=0, Cr=40%				$\widehat{\beta_t}$=0, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=1, Cr=45%			
200	3	$\hat{\alpha}$	0.051	0.127	0.148	0.943	0.060	0.162	0.153	0.965
		$\widehat{\beta_t}$	0.046	0.181	0.177	0.935	0.041	0.178	0.192	0.971
		$\widehat{\beta_l}$	0.065	0.169	0.163	0.952	0.057	0.176	0.182	0.940
		$\widehat{\beta_w}$	0.058	0.203	0.199	0.959	0.071	0.210	0.203	0.953
		$\hat{\sigma}$	0.010	0.092	0.084	0.946	0.003	0.087	0.102	0.966
400	4	$\hat{\alpha}$	0.047	0.118	0.135	0.953	0.049	0.123	0.146	0.941
		$\widehat{\beta_t}$	0.052	0.138	0.142	0.949	0.038	0.159	0.168	0.942
		$\widehat{\beta_l}$	0.074	0.145	0.157	0.958	0.069	0.176	0.183	0.953
		$\widehat{\beta_w}$	0.063	0.133	0.147	0.952	0.081	0.190	0.198	0.960
		$\hat{\sigma}$	0.002	0.051	0.065	0.948	0.005	0.034	0.048	0.934
			$\widehat{\beta_t}$=1, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=0, Cr=35%				$\widehat{\beta_t}$=0.5, $\widehat{\beta_l}$=0.5, $\widehat{\beta_w}$=0.5, Cr=40%			
200	3	$\hat{\alpha}$	0.035	0.148	0.157	0.935	0.049	0.123	0.111	0.970
		$\widehat{\beta_t}$	0.077	0.185	0.180	0.944	0.054	0.154	0.161	0.957
		$\widehat{\beta_l}$	0.061	0.174	0.179	0.959	0.066	0.160	0.150	0.949
		$\widehat{\beta_w}$	0.045	0.156	0.147	0.930	0.061	0.179	0.170	0.952
		$\hat{\sigma}$	-0.006	0.072	0.063	0.937	0.012	0.082	0.090	0.948
400	4	$\hat{\alpha}$	0.038	0.116	0.109	0.942	0.041	0.119	0.128	0.959
		$\widehat{\beta_t}$	0.071	0.168	0.156	0.964	0.072	0.141	0.160	0.953
		$\widehat{\beta_l}$	0.056	0.145	0.149	0.948	0.047	0.132	0.145	0.944
		$\widehat{\beta_w}$	0.038	0.126	0.137	0.959	0.056	0.161	0.148	0.937
		$\hat{\sigma}$	0.009	0.068	0.052	0.955	-0.004	0.065	0.052	0.967

5. An Application

In this section, we illustrate our methodology by applying it to the NASA's Hypobaric decompression sickness database (HDSD). There are 238 subjects aged between 20 and 54 in the study (177 male and 61 female). The subjects are tested by a dehydrogenation process in a hypobaric environment. The response variable is the time of developing grade IV venous gas emboli (VGE). The goal of the study is to find out association between VGE and potential risk factors (NOADYN, TR360, age and gender). NOADYN is an indicator of the conditional of test subjects (NOADYN=1 for ambulatory and NOADYN=0 for lower body adynamic). TR360 represents the tissue ratio at 360 degrees.

We have interval-censored data here since the failure event (Grade IV VGE) is only observed to occur within two examination time points. We also have informative censoring scenario since the subjects who develop Grade IV VGE are more likely to have their examination earlier. Also some subjects are immune to Grade IV VGE and will never develop any related symptom. Therefore, cure rate model would fit the scenario here. It is pointed out that only covariates relate to the characteristic of the subject can affect the immunity of the failure event [14]. Thus, we only include age and gender in the logistic model for the cure rate. The estimation results are given in table 5. For comparison, we also include the estimation results given by a naïve estimation procedure that ignores the dependence between censoring time and failure time.

Table 4. *Results on estimations of regression coefficients based on the simulated data with two covariates and σ=0.5.*

Z_1 ~ N(0,1), Z_2 ~ Unif(-1,1), X_1 ~ Bernoulli(0.5), X_2 ~ N(0,1), α=(0,1), Cure%=30%										
n	***m***	**Par**	**Bias**	**SSD**	**SEE**	**CP**	**Bias**	**SSD**	**SEE**	**CP**
			$\beta_t = (1,1.5)'$, $\beta_l = (0.5,1)'$, $\beta_w = (1,1)'$, Cr=42%				$\beta_t = (1,1.5)'$, $\beta_l = (0.5,1)'$, $\beta_w = (1,-1)'$, Cr=46%			
200	3	$\widehat{\alpha_1}$	0.024	0.285	0.263	0.943	0.028	0.244	0.258	0.953
		$\widehat{\alpha_2}$	-0.019	0.261	0.245	0.931	0.015	0.245	0.249	0.961
		$\widehat{\beta_{t1}}$	0.051	0.339	0.352	0.960	0.037	0.351	0.338	0.943
		$\widehat{\beta_{t2}}$	-0.044	0.364	0.352	0.943	0.051	0.356	0.372	0.949
		$\widehat{\beta_{l1}}$	0.023	0.126	0.141	0.975	-0.012	0.154	0.161	0.935
		$\widehat{\beta_{l2}}$	0.018	0.183	0.165	0.939	0.015	0.163	0.182	0.957
		$\widehat{\beta_{w1}}$	0.038	0.158	0.167	0.945	0.042	0.191	0.211	0.951
		$\widehat{\beta_{w2}}$	0.041	0.183	0.191	0.962	0.036	0.205	0.216	0.947
		$\hat{\sigma}$	0.005	0.051	0.064	0.971	-0.006	0.068	0.051	0.933
400	4	$\widehat{\alpha_1}$	0.018	0.242	0.236	0.941	0.021	0.191	0.216	0.965
		$\widehat{\alpha_2}$	-0.017	0.184	0.191	0.956	0.008	0.162	0.178	0.951
		$\widehat{\beta_{t1}}$	0.048	0.256	0.263	0.952	0.031	0.241	0.235	0.948
		$\widehat{\beta_{t2}}$	-0.022	0.272	0.287	0.963	0.025	0.292	0.286	0.941
		$\widehat{\beta_{l1}}$	0.017	0.077	0.086	0.955	0.008	0.091	0.085	0.948
		$\widehat{\beta_{l2}}$	0.021	0.096	0.103	0.953	0.012	0.116	0.108	0.956
		$\widehat{\beta_{w1}}$	0.029	0.176	0.169	0.961	0.027	0.164	0.156	0.961
		$\widehat{\beta_{w2}}$	0.018	0.163	0.181	0.954	0.020	0.180	0.185	0.944
		$\hat{\sigma}$	-0.002	0.056	0.071	0.961	-0.005	0.046	0.057	0.937

Table 5. *Analysis results of NASA's Hypobaric Decompression Sickness Data (Assuming $v = 0.05$).*

			Ignore informative censoring			**With informative censoring**		
			EST	**SEE**	***p*-value**	**EST**	**SEE**	***p*-value**
$m = 3$	Cure model	Intercept	-2.712	0.494	<0.001	-2.943	0.573	<0.001
		Age	0.912	0.320	0.005	1.116	0.372	0.003
		Gender	1.356	0.503	0.009	1.507	0.617	0.012
	PH model	Age	-0.281	0.061	<0.001	-0.326	0.083	<0.001
		Gender	-0.142	0.161	0.327	-0.170	0.196	0.416
		TR360	-0.667	0.383	0.103	-1.128	0.517	0.032
		NOADYN	0.826	0.359	0.018	0.751	0.317	0.009
$m = 4$	Cure model	Intercept	-2.506	0.520	<0.001	-2.883	0.584	<0.001
		Age	0.893	0.323	0.005	1.018	0.361	0.006
		Gender	1.237	0.477	0.008	1.396	0.556	0.012
	PH model	Age	-0.266	0.060	<0.001	-0.342	0.082	<0.001
		Gender	-0.128	0.157	0.399	-0.178	0.210	0.472
		TR360	-0.632	0.389	0.123	-1.049	0.512	0.048
		NOADYN	0.791	0.310	0.007	0.818	0.300	0.014
$m = 5$	Cure model	Intercept	-2.501	0.521	<0.001	-2.751	0.536	<0.001
		Age	0.831	0.303	0.003	1.007	0.331	0.005
		Gender	1.121	0.480	0.015	1.274	0.511	0.011
	PH model	Age	-0.231	0.038	<0.001	-0.299	0.079	<0.001
		Gender	-0.102	0.135	0.367	-0.125	0.154	0.391
		TR360	-0.610	0.371	0.131	-0.952	0.387	0.054
		NOADYN	0.761	0.277	0.011	0.808	0.310	0.007

From table 5, we can see that the estimation is robust with respect to the degree of Bernstein polynomials. Both proposed and naïve approach give similar estimates to the coefficient of NOADYN, gender and age. Moreover, the results showed that subjects with higher TR360 had longer survival time in terms of Grade IV VGE. Nevertheless, only the proposed approach detected a significant effect of TR360. Whereas the naïve estimation procedure that ignores the informative censoring failed to detect the significance of risk factor TR360. In addition, our proposed sieve estimation procedure detected the significance of risk factor age and NOADYN. Therefore, we conclude that younger people develop Grade IV VGE more quickly and subjects who are ambulatory develop Grade IV VGE more quickly that those who are lower body adynamic. The estimation on cure rate suggest that older and male subjects are more likely to develop Grade IV VGE than younger and female subjects.

6. Discussions and Conclusions

In this article, we considered the analysis of informatively interval-censored data when there is a cured subpopulation. In order to deal with informative interval-censoring and cured subpopulation at the same time, we used a log-normal frailty variable to account for the independence between censoring time and failure time. A mixture cure rate model was developed to account for the cured subpopulation. To estimate the model parameters, we proposed a sieve maximum likelihood estimation procedure. Bernstein polynomials are used as the sieve functions to estimate the non-parametric component of the model. Furthermore, we derived an EM algorithm to obtain the numerical solutions of the model parameters. The EM algorithm has the advantage of reducing the computational burden of the problem and provide efficient estimators. We also conducted an extensive simulation study that showed our method has an advantage over the traditional method that ignores the informative censoring and cured subpopulation. In addition, the simulation results suggested our method performed well for different practical scenarios.

There are a few future research directions on this topic. In our paper, we employed the Cox model under the assumptions of proportional hazards. Obviously, in some practical situations, other survival models such as the semiparametric transformation model or proportional odds model may be more appropriate. Therefore, it would be interesting to develop different estimation procedure for these models. Moreover, we developed a mixture cure rate model for the problem. Several researchers have developed a non-mixture cure model which has the advantage of modeling event time uniformly [15-16]. Another research direction is to develop a sieve estimation procedure using the non-mixture cure model.

In general, when dealing interval-censored data with informative censoring and cured subgroup, we recommend the sieve maximum likelihood approach. However, this approach may be less reliable when the data is subject to measurement error. Note that in the application section above, the tissue ratio at 360 degrees (TR360) may subject to measurement error. It is well known that ignorance of measurement error could lead to biased estimates. In the future, it would be interesting to establish an estimation procedure to address the measurement error and informative censoring at the same time.

Acknowledgements

The authors are grateful to the editor and anonymous referee for their beneficial and accurate comments that improved this paper.

References

[1] Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with longterm survivors. Biometrics. 38, 1041–1046.

[2] Lam, K. F. & Xue, H. (2005). A semiparametric regression cure model with current status data. Biometrika. 92, 573–586.

[3] Ma, S. (2010). Mixed case interval censored data with a cured subgroup. Statistica Sinica. 20, 1165–1181.

[4] Balogun, O., Gao, X. Z., Jolayemi, E. T. & Olaleye, S. (2020). Generalized cure rate model for infectious diseases with possible co-infections. PLoS ONE. 15, 1-16.

[5] Hu, T., and L. Xiang. (2016). Partially linear transformation cure models for interval-censored data. Computational Statistics and Data Analysis. 93, 257–69.

[6] Liu, Y., Hu, T. & Sun, J. (2020). Regression analysis of intervalcensored failure time data with cured subgroup and mismeasured covariates. Communications in Statistics - Theory and Methods. 49(1): 189-202.

[7] Riester, K., Kappos, L., Selmaj, K., Lindborg, S., Lipkovich, I. & Elkins, J. (2019). Impact of informative censoring on the treatment effect estimate of disability worsening in multiple sclerosis clinical trials. Multiple Sclerosis and Related Disorders. 39, 101865.

[8] Li, Y., Tiwari, R. & Guha, S. (2007). Mixture cure survival models with dependent censoring. Journal of the Royal Statistical Society: Series B. 69, 285–306.

[9] Othus, M., Li, Y., Tiwari, R. (2007). A class of semiparametric mixture cure survival models with dependent censoring. Journal of American Statistical Association. 104, 1241–1250.

[10] Rondeau, V., Schaffner, E., Corbiere, F., Gonzalez, J. & Pelissier, S. (2011). Cure frailty models for survival data: application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. Statistical Methods in Medical Research. 22 (3): 1–18.

[11] Huang, X., Wolfe, R. A. (2002). A Frailty Model for Informative Censoring. Biometrics. 58 (3): 510–520.

[12] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B. 40, 226–233.

[13] Zhou, Q., Hu, T. & Sun, J. (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. Journal of the American Statistical Association. 112 (518): 664–72.

[14] Conkin, J. & Powell, M. (2001). Lower body adynamia as a factor to reduce the risk of hypobaric decompression sickness. Aviation. Space and Environmental Medicine. 72 (3): 202–14.

[15] Liu, H. & Shen, Y. (2009). A semiparametric regression cure model for interval censored data. Journal of the American Statistical Association. 104 (487): 1168–78.

[16] Zeng, D., Yin, G., and Ibrahim, J. G. (2006). Semiparametric Transformation Models for Survival Data with a Cure Fraction. Journal of the American Statistical Association. 101, 670–684.