HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2018 May 31.

Published in final edited form as:

J Comput Graph Stat. 2018 ; 27(1): 48–58. doi:10.1080/10618600.2017.1349665.

Computationally Efficient Estimation for the Generalized Odds Rate Mixture Cure Model with Interval-Censored Data

Jie Zhou,
Department of Epidemiology and Biostatistics, University of South Carolina

Jiajia Zhang, and
Department of Epidemiology and Biostatistics, University of South Carolina

Wenbin Lu
Department of Statistics, North Carolina State University

Abstract

For semiparametric survival models with interval censored data and a cure fraction, it is often difficult to derive nonparametric maximum likelihood estimation due to the challenge in maximizing the complex likelihood function. In this paper, we propose a computationally efficient EM algorithm, facilitated by a gamma-poisson data augmentation, for maximum likelihood estimation in a class of generalized odds rate mixture cure (GORMC) models with interval censored data. The gamma-poisson data augmentation greatly simplifies the EM estimation and enhances the convergence speed of the EM algorithm. The empirical properties of the proposed method are examined through extensive simulation studies and compared with numerical maximum likelihood estimates. An R package "GORCure" is developed to implement the proposed method and its use is illustrated by an application to the Aerobic Center Longitudinal Study dataset.

Keywords

Cure model; Data augmentation; EM algorithm; Generalized odds rate model; Interval censoring

1 Introduction

With the development of modern technology, some diseases, such as breast cancer and prostate cancer, has an opportunity to be cured. That is, there exist a proportion of the patients being cured, who will not die from this disease in the future. To study the risk effects associated with such disease, the main interests are focused on (i) what risk factors can affect the probability of patients being cured; and (ii) what risk factors can influence the survival probabilities of uncured patients. The standard survival models, such as the proportional hazards (PH) model (Cox, 1992) and the proportional odds (PO) (Bennett, 1983) model, are inappropriate to address these questions because their unstated model assumption is that all the subjects will die from the event of interest eventually. That is, it cannot accommodate the proportion of patients being cured.

In order to capture the proportion of patients being cured, there are many discussions on the two-component mixture cure model (Boag, 1949; Berkson and Gage, 1952; Farewell, 1982) and the promotion time cure model (Tsodikov et al., 2003). Among them, the two-component mixture cure model has straightforward interpretation in practice since it models the cured and uncured patients directly by the incidence and latency part. Let T be a nonnegative random variable denoting the failure time of a patient, and $S(\cdot|\boldsymbol{x}, \boldsymbol{z})$ be the survival function of T, where $\boldsymbol{x}$ and $\boldsymbol{z}$ are two covariate vectors on which the distribution of T may depend, and may not be same in practice. The two-component mixture cure model is defined as

$$S(t|\boldsymbol{x}, z) = 1 - \pi(z) + \pi(z) \cdot S_u(t|\boldsymbol{x}), \quad (1)$$

where the incidence part, $\pi(\boldsymbol{z})$, models the uncured probability, and the latency part, $S_u(\cdot|\boldsymbol{x})$, models the survival probability of uncured patients.

The association of the covariate vector $\boldsymbol{z}$ with the probability of uncured is modeled through a link function, and the most commonly used link functions are the logit link $\text{logit}(\pi(\boldsymbol{z})) = \boldsymbol{\eta}'\boldsymbol{z}$, the probit link ($\Phi^{-1}(\pi(\boldsymbol{z})) = \boldsymbol{\eta}'\boldsymbol{z}$) and the complementary log-log link ($\log(-\log(1 - \pi(\boldsymbol{z}))) = \boldsymbol{\eta}'\boldsymbol{z}$) (Price and Manatunga, 2001). The standard survival models are applied to the survival probability of uncured patients, $S_u(t|\boldsymbol{x})$. For example, the proportional hazards mixture cure (PHMC) model, which incorporates the PH model to $S_u(\cdot|\boldsymbol{x})$, has been studied extensively in literature (Kuk and Chen, 1992; Sy and Taylor, 2000; Peng and Dear, 2000). Similarly, the proportional odds mixture cure (POMC) model assumes the PO model for $S_u(\cdot|\boldsymbol{x})$ (Gu et al., 2011). In the PHMC model, the constant hazard ratio is assumed for the uncured patients; and in the POMC model, the constant odds ratio among the survival probability of uncured patients has to be satisfied. There are cases that neither model is preferred, and a general survival function which includes both the PH and PO models is more desirable.

The generalized odds rate (GOR) class of regression models (Bickel, 1986; Dabrowska and Doksum, 1988; Scharfstein et al., 1998; Banerjee et al., 2007), which include the PH and PO models as special cases, have attracted much attention recently. The GOR model is defined as

$$g_r(S_u(t|\boldsymbol{x})) = H(t) + \boldsymbol{\beta}'\boldsymbol{x}, \quad (2)$$

where $H(\cdot)$ is a nondecreasing transformation function satisfying $H(0) = -\infty$ and $g_r(\cdot)$ is a known decreasing link function indexed by a transformation parameter r. Specifically, when $r = 0$, $g_0(s) = \log(-\log s)$ refers to the PH model, and when $r > 0$, $g_r(s) = \log(r^{-1}(s^{-r} - 1))$ corresponds to a class of generalized proportional odds (GPO) model (Mao and Wang, 2010). Note, when $r = 1$ it is the PO model with $g_1(s) = -\text{logit}(s) = -\log[s/(1 - s)]$. In this article, we incorporate the GOR model to the survival probability of uncured patients, $S_u(\cdot|\boldsymbol{x})$, and refer it as the GORMC model, which includes the PHMC ($r = 0$), POMC ($r = 1$), and GPO mixture cure (GPOMC) model ($r > 0$) as its special cases. Mao and Wang (2010) also

provide the proof of the identifiability of the covariates in the generalized odds rate mixture cure model.

In medical studies, it is common to observe interval censored data, where event times of interest are only known to occur in some intervals. For example, in the Aerobics Center Longitudinal Study (Blair et al., 1996; Lee et al., 2012), we are interested in studying the relation between diagnosis age of hypertension and potential risk factors including cardiorespiratory fitness (hereafter referred to as “fitness”) levels, gender and BMI. Here, age at diagnosis of hypertension is only known to occur between two consecutive visits, which is interval censored. The Turnbull (1976) nonparametric survival curves for different fitness levels and gender are plotted in Figure 1. From the figure, females tend to have higher survival probabilities than males at all fitness levels, and there is a leveling off of survival curves at the end of study. The leveling off can be a potential indicator of “cured”, and suggest that there exists a proportion of participants never developing hypertension.

Regression analysis with interval censored data is more challenging than right censored data in both computation and theory. Current researches for mixture cure model with interval censored data focus on the PHMC model. For the PHMC model with current status data, Ma (2009) developed a penalized maximum likelihood method with the weighted bootstrap for variance estimation. For the PHMC model with interval censored data, Kim and Jhun (2008) proposed an EM algorithm by assuming a piecewise exponential distribution for the baseline hazard with a multiple imputation approach for variance estimation; Ma (2010) assumed a piecewise constant baseline cumulative hazard function and developed a two-step maximum likelihood method; the multiple imputation technique were also considered in Lam et al. (2013) and Zhou et al. (2016). However, there are sparse work on the GORMC model with interval censored data.

Recently, a two-stage poisson data augmentation was proposed by McMahan et al. (2013) for the PH and PO model with current status data, and was extended to the PH model with interval censored data by Wang et al. (2016). In this article, we propose a new gamma-poisson data augmentation approach for the efficient estimation of the GORMC model with interval censored data. In order to account for the latent cure indicator and interval censoring, a Bernoulli random variable, a gamma frailty and poisson latent variables are introduced to facilitate the maximum likelihood estimation using an EM algorithm. Furthermore, the variance estimates can be obtained in closed forms by the Louis method. The remainder of the paper is organized as follows. Section 2 introduces the GORMC model. The proposed data augmentation assisted EM algorithm are described in Section 3. The simulation study is presented in Section 4.1 and the proposed method is applied to the Aerobic Center Longitudinal Study in Section 4.2. The discussions are given in Section 5. The closed form of Hessian Matrix for variance estimation are provided in the Appendix.

2 Model Description

Assume each subject has a sequence of examination times $0 < V_{i1} < \cdots < V_{i,ki} < \infty$, $i = 1, 2, \ldots, n$ during the study and $(L_i, R_i]$ is the observed time interval including the exact event time T_i for the ith subject. The observed data can be denoted as $\boldsymbol{O} = \{O_i = (L_i, R_i, \delta_{L,i}, \delta_{R,i},$

$\delta_{I,i}, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, 2, \ldots, n\}$, where $\delta_{L,i}$, $\delta_{R,i}$, $\delta_{I,i}$ are the censoring indicators where $\delta_{L,i} = 1$ when T_i is left censored ($L_i = 0$), $\delta_{R,i} = 1$ when T_i is right censored ($R_i = 1$), and $\delta_{I,i} = 1$ when T_i is interval censored ($0 < L_i < R_i < \infty$) with the convention that $\delta_{L,i} + \delta_{R,i} + \delta_{I,i} = 1$ for all i.

Assuming the examination times V_{ij}'s and the event time T_i are independent given the covariate $\boldsymbol{x}_i$ (Huang and Wellner, 1997). The observed likelihood function of the parameters $\boldsymbol{\theta}_o = (\boldsymbol{\eta}, \boldsymbol{\beta}, H)$ under the GORMC model with a prespecified $r > 0$ is given by

$$\mathscr{L}(\boldsymbol{\theta}_O | \boldsymbol{O}) = \prod_{i=1}^{n} \mathscr{L}_i(\boldsymbol{\theta}_o | O_i) = \prod_{i=1}^{n} \pi(\boldsymbol{z}_i)^{1-\delta_{R,i}} \{1 - S_u(R_i | \boldsymbol{x}_i)\}^{\delta_{L,i}} \{S_u(L_i | \boldsymbol{x}_i) - S_u(R_i | \boldsymbol{x}_i)\}^{\delta_{I,i}} \times \{1 - \pi(\boldsymbol{z}_i) + \pi(\boldsymbol{z}_i) S_u(L_i | \boldsymbol{x}_i)\}^{\delta_{R,i}}. \quad (3)$$

$\pi(\boldsymbol{z}_i)$ is modelled by the logistic regression, $\text{logit}(\pi(\boldsymbol{z}_i)) = \boldsymbol{\eta}' \boldsymbol{z}_i$, and the survival function of uncured patients is estimated by the GOR model (2)

$$S_u(t | \boldsymbol{x}_i) = \begin{cases} \exp\left\{ - H_e(t) e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right\}, & r = 0, \\ [1 + r H_e(t) e^{\boldsymbol{\beta}' \boldsymbol{x}_i}]^{-\frac{1}{r}}, & r > 0, \end{cases}$$

where $H_e(t) = \exp[H(t)]$. Direct maximization of (3) with respect to $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $H_e(\cdot)$ is neither easy nor reliable due to the complex structure of the likelihood function. To facilitate the maximum likelihood estimation, we propose a gamma-Poisson data augmentation method to define a proper complete data likelihood function in the next section. Then, an EM algorithm is developed based on the resulting complete data likelihood function in Section 3.

3 Data Augmentation and EM Algorithm

It is worthwhile pointing out that $\lim_{r \to 0} [1 + r H_e(t) e^{\boldsymbol{\beta}' \boldsymbol{x}_i}]^{-\frac{1}{r}} = \exp\left\{ - H_e(t) e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right\}$. Thus, all the results in the PHMC model can be derived by taking the limit of the results under the GPOMC model ($r \to 0$). Therefore, we focus discussions on the data augmentation and conditional expectations under the GPOMC model in Sections 3.1 and 3.2, and then give the results for the PHMC model in Section 3.3.

3.1 Complete Likelihood

Let $u_i \sim Ber(\pi(\boldsymbol{z}_i))$ be a latent variable with $u_i = 1$ if subject i is uncured, where $Ber(p)$ is the bernoulli distribution with the probability of success p. Conditional on $\boldsymbol{u} = (u_1, \cdots, u_n)$ and using the fact that $(1 - u_i)\delta_{R,i} = 1 - u_i$, the likelihood function of $\boldsymbol{\theta}_o$ can be written as

$$\mathscr{L}(\boldsymbol{\theta}_o|\boldsymbol{O},\boldsymbol{u}) = \prod_{i=1}^{n} \mathscr{L}_i(\boldsymbol{\theta}_o|O_i,u_i) = \prod_{i=0}^{n} \pi(z_i)^{u_i}[1-\pi(z_i)]^{1-u_i}\left\{1 - S_u(R_i|\boldsymbol{x}_i)\right\}^{\delta_{L,i}}\left\{S_u(L_i|\boldsymbol{x}_i) - S_u(R_i|\boldsymbol{x}_i)\right\}^{\delta_{I,i}} \times S_u(L_i|\boldsymbol{x}_i)^{\delta_{R,i}u_i}.$$

Furthermore, let $\boldsymbol{\phi} \sim Gamma(1/r, r)$, the survival function of uncured patients under the GPOMC model can be written as

$$S_u(t|\boldsymbol{x}) = [1 + rH_e(t)e^{\boldsymbol{\beta}'\boldsymbol{x}}]^{-\frac{1}{r}} = \int_0^{\infty} \exp\left\{-\phi H_e(t)e^{\boldsymbol{\beta}'\boldsymbol{x}}\right\} \frac{\phi^{\frac{1}{r}-1}e^{-\frac{\phi}{r}}}{\Gamma(\frac{1}{r})r^{\frac{1}{r}}} d\phi.$$

The likelihood function based on $\boldsymbol{u}$ and $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_n)$ is

$$\mathscr{L}(\boldsymbol{\theta}_o|\boldsymbol{O},\boldsymbol{u},\phi,r) = \prod_{i=1}^{n} \frac{\phi_i^{\frac{1}{r}-1}e^{-\frac{\phi_i}{r}}}{\Gamma(\frac{1}{r})r^{\frac{1}{r}}} \times \pi(z_i)^{u_i}[1-\pi(z_i)]^{1-u_i} \times \left\{1 - \exp[-\phi_i H_e(R_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}]\right\}^{\delta_{L,i}}$$

$$\times \left\{\exp[-\phi_i H_e(L_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}] - \exp[-\phi_i H_e(R_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}]\right\}^{\delta_{I,i}} \times \exp[-\delta_{R,i}u_i\phi_i H_e(L_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}],$$

where $\phi_i \overset{i.i.d}{\sim} Gamma(1/r, r)$ for $i = 1, \cdots, n$.

Poisson latent variables are introduced conditional on the gamma frailty term $\boldsymbol{\phi}$ in the following way. For subject i, we have $Y_i|\boldsymbol{\phi}_i \sim Pois(\lambda_i\boldsymbol{\phi}_i)$ and $W_i|\boldsymbol{\phi}_i \sim Pois(\omega_i\boldsymbol{\phi}_i)$, where $\lambda_i = e^{\boldsymbol{\beta}'\boldsymbol{x}_i}[\delta_{L,i}H_e(R_i) + \delta_{I,i}H_e(L_i)]$ and $\omega_i = e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\left\{\delta_{I,i}[H_e(R_i) - H_e(L_i)] + \delta_{R,i}H_e(L_i)\right\}$. Moreover, we set $Y_i \equiv 0$ if $Y_i \sim Pois(0)$. Note that Y_i and W_i are independent of each other conditional on $\boldsymbol{\phi}_i$. Under this construction, we have $Y_i > 0$ and $W_i = 0$ under left censoring, $Y_i = 0$ and $W_i > 0$ under interval censoring and $Y_i = W_i = 0$ under right censoring. Therefore, the complete likelihood function based on $\boldsymbol{\phi}$, $\boldsymbol{u}$, $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ and $\boldsymbol{W} = (W_1, \cdots, W_n)$ is

$$\mathscr{L}(\boldsymbol{\theta}_o|\boldsymbol{O},\boldsymbol{u},\phi,\boldsymbol{Y},\boldsymbol{W},r)=\prod_{i=1}^{n}\pi(z_i)^{u_i}[1-\pi(z_i)]^{1-u_i}\times\frac{\phi_i^{\frac{1}{r}-1}e^{-\frac{\phi_i}{r}}}{\Gamma(\frac{1}{r})r^{\frac{1}{r}}}\times\left[\frac{1}{Y_i!}(\lambda_i\phi_i)^{Y_i}e^{-\lambda_i\phi_i}\right]^{\delta_{L,i}}$$

$$\times\left[\frac{1}{W_i!}(\omega_i\phi_i)^{W_i}e^{-(\lambda_i+\omega_i)\phi_i}\right]^{\delta_{I,i}}\times\exp(-\delta_{R,i}^{u_i}\omega_i\phi_i),$$

(4)

under the constraint $\delta_{L,i}I(Y_i>0)+\delta_{I,i}I(W_i>0)+\delta_{R,i}=1$ for $i=1,\cdots,n$. The observed likelihood function (3) can be achieved by integrating out the latent variables $\boldsymbol{u}$, $\boldsymbol{\phi}$, $\boldsymbol{Y}$ and $\boldsymbol{W}$ in the complete likelihood function (4).

To estimate the unknown monotone transformation function H_e, the monotone splines are adopted for $H_e(\cdot)$ as

$$H_e(t)=\sum_{l=1}^{k}\gamma_l b_l(t),$$

where $b'_l s$ are nondecreasing integrated spline basis functions ranges from 0 to 1 and the coefficients $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_k)$ are non-negative to ensure the monotonicity. Assuming that $\sum_{l=1}^{k}\gamma l\to\infty$ as $k\to\infty$, the monotone spline function $H_e(t)=\sum_{l=1}^{k}\gamma_l b_l(t)\to\infty$ as $t\to\infty$. Therefore, the survival function of uncured patients is proper. The choice of knots for the spline functions can either be equally spaced or at the quartiles as suggested by Ramsay (1988). To further simplify the calculation, we let $Y_{il}|\phi_i\sim Pois(\lambda_{il}\phi_i)$ and $W_{il}|\phi_i\sim Pois(\omega_{il}\phi_i)$, where $\lambda_{il}=\gamma_l e^{\boldsymbol{\beta}'\boldsymbol{x}_i}[\delta_{L,i}^{b_l(R_i)}+\delta_{I,i}^{b_l(L_i)}]$ and $\omega_{il}=\gamma_l e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\left\{\delta_{I,i}[b_l(R_i)-b_l(L_i)]+\delta_{R,i}^{b_l(L_i)}\right\}$, $i=1,\cdots,n$; $l=1,\cdots,k$. Notice that $\sum_{l=1}^{k}\lambda_{il}=\lambda_i$ and $\sum_{l=1}^{k}\omega_{il}=\omega_i$, we have $Y_i\overset{d}{=}\sum_{l=1}^{k}Y_{il}$ and $W_i\overset{d}{=}\sum_{l=1}^{k}W_{il}$, where "$\overset{d}{=}$" means equal in distribution. The final complete likelihood function of $\boldsymbol{\theta}=(\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\gamma})$ based on $\boldsymbol{\phi}$, $\boldsymbol{u}$, $\boldsymbol{Y}_l$ and $\boldsymbol{W}_l$, $l=1,\cdots,k$, is

$$\mathscr{L}_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(z_i)^{u_i}[1-\pi(z_i)]^{1-u_i} \times \frac{\phi_i^{\frac{1}{r}-1} e^{-\frac{\phi_i}{r}}}{\Gamma(\frac{1}{r}) r^{\frac{1}{r}}} \tag{5}$$

$$\times \left\{ \prod_{l=1}^{k} \left[\frac{1}{Y_{il}!}(\lambda_{il}\phi_i)^{Y_{il}} e^{-\lambda_{il}\phi_i} \right]^{\delta_{L,i}} \times \left[\frac{1}{W_{il}!}(\omega_{il}\phi_i)^{W_{il}} e^{-(\lambda_{il}+\omega_{il})\phi_i} \right]^{\delta_{I,i}} \times \exp(-\delta_{R,i} u_i \omega_{il}\phi_i) \right\},$$

under the constraint $\delta_{L,i} I(\sum_{l=1}^{k} Y_{il} > 0) + \delta_{I,i} I(\sum_{l=1}^{k} W_{il} > 0) + \delta_{R,i} = 1$ for $i = 1, \cdots, n$. Similarly, by integrating out $\boldsymbol{u}$, $\boldsymbol{\phi}$, $\boldsymbol{Y}_l$'s and $\boldsymbol{W}_l$'s in (5), it reduces to the observed likelihood function (3).

3.2 EM Algorithm

The expectation of the logarithm of the complete likelihood function with respect to the latent variables conditional on the observed data $\boldsymbol{O}$, a prespecified transformation parameter r, and the current parameter estimate $\boldsymbol{\theta}^{(d)}$ can be expressed as the summation of two parts

$$\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(d)}) = \mathcal{Q}_1(\boldsymbol{\eta}; \boldsymbol{\eta}^{(d)}) + \mathcal{Q}_2(\boldsymbol{\beta}, \gamma; \boldsymbol{\beta}^{(d)}, \gamma^{(d)}),$$

where

$$\mathcal{Q}_1(\boldsymbol{\eta}; \boldsymbol{\eta}^{(d)}) = \sum_{i=1}^{n} E(u_i|\boldsymbol{O}, \boldsymbol{\theta}^{(d)}) \log[\pi(z_i)] + [1 - E(u_i|\boldsymbol{O}, \boldsymbol{\theta}^{(d)})] \log[1-\pi(z_i)],$$

$$\mathcal{Q}_2(\boldsymbol{\beta}, \gamma; \boldsymbol{\beta}^{(d)}, \gamma^{(d)} = \sum_{i=1}^{n} \sum_{l=1}^{k} \left\{ [\delta_{L,i} E(Y_{il}|\boldsymbol{O}, \boldsymbol{\theta}^{(d)}) + [\delta_{I,i} E(W_{il}|\boldsymbol{O}, \boldsymbol{\theta}^{(d)})](\log\gamma_l + \boldsymbol{\beta}' \boldsymbol{x}_i) - \gamma_l e^{\boldsymbol{\beta}' X_i} E(\phi_i u_i|\boldsymbol{O}, \boldsymbol{\theta}^{(d)}) \right.$$
$$\left. [(1-\delta_{R,i}) b_l(R_i) + \delta_{R,i} b_l(L_i)] \right\} + L(\boldsymbol{\theta}^{(d)}),$$

where $L(\boldsymbol{\theta}^{(d)})$ is a function of $\boldsymbol{\theta}^{(d)}$ but free of $\boldsymbol{\theta}$.

When subject i is left censored, Y_i conditionally follows the zero-truncated negative binomial distribution (ZTNB). The probability mass function (p.m.f.) can be written as

$$Pr(Y_i = y|\boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\delta_{L,i}}{c_i^{(d)}} \binom{y + \frac{1}{r} - 1}{y} [1 - (1 - c_i^{(d)})^r]^y (1 - c_i^{(d)}) I(y > 0),$$

where

$$c_i^{(d)} = 1 - \left(\frac{\delta_{L,i}}{1 + r\lambda_i^{(d)}} \right)^{\frac{1}{r}}$$

is the conditional probability for the truncation $Y_i > 0$ and $\lambda_i^{(d)}$ is the value of λ_i evaluated at $\boldsymbol{\theta}^{(d)}$. This is denoted as $Y_i \sim ZTNB(1/r, \delta_{L,i}[1 - (1 - c_i^{(d)})^r])$. Similarly, when the subject is interval censored, we have $W_i \sim ZTNB(1/r, \delta_{I,i}[1 - (1 - d_i^{(d)})^r])$, where

$$d_i^{(d)} = 1 - \left[\frac{\delta_{I,i}(1 + r\lambda_i^{(d)})}{1 + r\lambda_i^{(d)} + r\omega_i^{(d)}} \right]^{\frac{1}{r}}$$

is the conditional probability for the truncation $W_i > 0$.

The conditional expectations of Y_i and W_i are

$$E(Y_i | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\delta_{L,i}\lambda_i^{(d)}}{c_i^{(d)}}, \quad E(W_i | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\delta_{I,i}\omega_i^{(d)}}{(1 + r\lambda_i^{(d)})d_i^{(d)}}.$$

Using the fact $Y_{il} | Y_i \sim Binomial(Y_i, \lambda_{il}^{(d)}/\lambda_i^{(d)})$ and $W_{il} | W_i \sim Binomial(W_i, \omega_{il}^{(d)}/\omega_i^{(d)})$, apply the iterated rule of expectations, the conditional expectations of Y_{il} and W_{il} are

$$E(Y_{il} | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\lambda_{il}^{(d)}}{\lambda_i^{(d)}} E(Y_i | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\delta_{L,i}\lambda_{il}^{(d)}}{c_i^{(d)}}, \quad (6)$$

$$E(W_{il} | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\omega_{il}^{(d)}}{\omega_i^{(d)}} E(W_i | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = \frac{\delta_{I,i}\omega_{il}^{(d)}}{(1 + r\lambda_i^{(d)})d_i^{(d)}}. \quad (7)$$

Similarly, $\lambda_{il}^{(d)}(\omega_{il}^{(d)})$ is the value of $\lambda_{il}(\omega_{il})$ with $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^{(d)}$.

Finally, the conditional expectation of u_i and $u_i\boldsymbol{\phi}_i$ can be derived from the joint distributions of Y_i, W_i, u_i and $\boldsymbol{\phi}_i$:

$$E(u_i | \boldsymbol{O}, \boldsymbol{\theta}^{(d)}, r) = 1 - \delta_{R,i} + \frac{\delta_{R,i}\pi^{(d)}(z_i)}{e_i^{(d)}(1 + r\omega_i^{(d)})^{1/r}}, \quad e_i^{(d)} = 1 - \pi^{(d)}(z_i) + \pi^{(d)}(z_i)\left(\frac{\delta_{R,i}}{1 + r\omega_i^{(d)}} \right)^{\frac{1}{r}}, \quad (8)$$

$$E(u_i\phi_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)},r)=\frac{\delta_{L,i}[1-(1-c_i^{(d)})^{r+1}]}{c_i^{(d)}}+\frac{\delta_{I,i}[1-(1-d_i^{(d)})^{r+1}]}{(1+r\lambda_i^{(d)})d_i^{(d)}} \quad (9)$$

$$+\frac{\delta_{R,i}\pi^{(d)}(z_i)}{e_i^{(d)}(1+r\omega_i^{(d)})^{1+1/r}},$$

where $\boldsymbol{\pi}^{(d)}(\boldsymbol{z}_i)$ is the value of $\boldsymbol{\pi}(\boldsymbol{z}_i)$ with $\boldsymbol{\theta}$ being evaluated at $\boldsymbol{\theta}^{(d)}$. Once the conditional expectations of observed likelihood is obtained, the maximization can be realized as described in subsection 3.4.

3.3 Results under the PHMC Model

First notice that when $r\to 0$, $\boldsymbol{\phi}_i$ converges in probability to the value of 1. Therefore, the data augmentation for the PHMC model involves the uncured indicator $\boldsymbol{u}$ and two layers of poisson latent variables, where $\boldsymbol{u}$ is introduced the same way as before. To accommodate for the interval censored data among uncured patients, we have $Y_i\sim Pois(\lambda_i)$ and $W_i\sim Pois(\omega_i)$ in the first stage, and $Y_{il}\sim Pois(\lambda_{il})$ and $W_{il}\sim Pois(\omega_{il})$ in the second stage, where λ_i, ω_i, λ_{il} and ω_{il} have the same definition as before. The resulting complete likelihood function based on $\boldsymbol{u}$, $\boldsymbol{Y}_l$'s and $\boldsymbol{W}_l$'s is

$$\mathscr{L}_c(\boldsymbol{\theta})=\prod_{i=1}^{n}\pi(z_i)^{u_i}[1-\pi(z_i)]^{1-u_i}\times\left\{\prod_{l=1}^{k}\left[\frac{1}{Y_{il}!}\lambda_{il}^{Y_{il}}e^{-\lambda_{il}}\right]^{\delta_{L,i}}\left[\frac{1}{W_{il}!}\omega_{il}^{W_{il}}e^{-(\lambda_{il}+\omega_{il})}\right]^{\delta_{I,i}}\exp(-\delta_{R,i}u_i\omega_{il})\right\},$$

under the constraint $\delta_{L,i}I(\sum_{l=1}^{k}Y_{il}>0)+\delta_{I,i}I(\sum_{l=1}^{k}W_{il}>0)+\delta_{R,i}=1$ for $i=1,\cdots,n$.

Under the PHMC model, the conditional expectation of the logarithm of the complete likelihood function $\mathscr{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(d)})$ has a similar form as before except we have $\boldsymbol{\phi}_i\equiv 1$ in $\mathscr{Q}_2(\boldsymbol{\beta},\gamma;\boldsymbol{\beta}^{(d)},\gamma^{(d)})$.

The random variable Y_i (W_i) conditionally follows zero truncated poisson distribution, and the conditional expectations are

$$E(Y_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)})=\frac{\delta_{L,i}\lambda_i^{(d)}}{1-\exp(-\lambda_i^{(d)})}$$

$$E(W_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) = \frac{\delta_{I,i}\omega_i^{(d)}}{1-\exp(-\omega_i^{(d)})},$$

where $\lambda_i^{(d)}$ is the value of λ_i with $\boldsymbol{\theta}$ being evaluated at $\boldsymbol{\theta}^{(d)}$.

Applying the iterated rule of expectations, the conditional expectations of Y_{il} and W_{il} are

$$E(Y_{il}|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) = \frac{\lambda_{il}^{(d)}}{\lambda_i^{(d)}}E(Y_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) = \frac{\delta_{L,i}\lambda_{il}^{(d)}}{1-\exp(-\lambda_i^{(d)})} \quad (10)$$

$$E(W_{il}|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) = \frac{\omega_{il}^{(d)}}{\omega_i^{(d)}}E(W_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) = \frac{\delta_{I,i}\lambda_{il}^{(d)}}{1-\exp(-\omega_i^{(d)})}. \quad (11)$$

The conditional expectation of the uncure indicator has the form of

$$E(u_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) = 1-\delta_{R,i} + \frac{\delta_{R,i}\pi^{(d)}(z_i)\exp(-\omega_i^{(d)})}{1-\pi^{(d)}(z_i)+\pi^{(d)}(z_i)\exp(-\omega_i^{(d)})}. \quad (12)$$

Similarly, once the conditional expectations of observed likelihood is obtained, the maximization can be realized as described in subsection 3.4.

3.4 Algorithm

The Expectation Maximization (EM) Algorithm for calculating estimates for $\boldsymbol{\theta}$ is described as follows:

Step 0: give initial values $\boldsymbol{\theta}^{(0)}$ as: $\boldsymbol{\eta}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{0}$, $\boldsymbol{\gamma}^{(0)} = \mathbf{1}$ and a pre-specified value of r;

Step 1: in the $(d+1)$th iteration, calculate conditional expectations in Equations (6) to (9) based on $\boldsymbol{\theta}^{(d)}$ if $r > 0$, or Equations (10) to (12) if $r = 0$;

Step 2: update $\boldsymbol{\eta}^{(d+1)}$ by fitting a logistic regression between the outcome $E(\boldsymbol{u}|\boldsymbol{O}, \boldsymbol{\theta}^{(d)})$ and the covariate $\boldsymbol{z}$, where a quasi-binomial distribution is assumed for the outcome;

Step 3: update $\boldsymbol{\beta}^{(d+1)}$ by maximizing $\mathcal{Q}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}(\boldsymbol{\beta}); \boldsymbol{\beta}^{(d)}, \boldsymbol{\gamma}^{(d)})$, where

$$\gamma_l(\boldsymbol{\beta}) = \frac{\sum_{i=1}^n \delta_{L,i}E(Y_{il}|\boldsymbol{O},\boldsymbol{\theta}^{(d)}) + \delta_{I,i}E(W_{il}|\boldsymbol{O},\boldsymbol{\theta}^{(d)})}{\sum_{i=1}^n e^{\boldsymbol{\beta}'\boldsymbol{x}_i}E(u_i\phi_i|\boldsymbol{O},\boldsymbol{\theta}^{(d)})[(1-\delta_{R,i})b_l(R_i)+\delta_{R,i}b_l(L_i)]}, \quad l = 1,\cdots,k,$$

where ϕ_i 1 when $r = 0$;

Step 4: calculate $\gamma_l^{(d+1)} = \gamma_l\left(\beta^{(d+1)}\right)$ for $l = 1, \cdots, k$;

Step 5: repeat Step 1 - Step 4 until the absolute difference of log-likelihood values between two consecutive iterations is less than 0.001. The estimates $\hat{\theta} = (\hat{\eta}, \hat{\beta}, \hat{\gamma})$ are obtained in the last iteration.

REMARK 1—The closed form solutions for $\gamma_l(\beta)$ in Step 3 are derived by solving the first derivatives $\partial \mathcal{Q}_2(\beta, \gamma; \beta^{(d)}, \gamma^{(d)}) / \partial \gamma_l = 0$, $l = 1, \cdots, k$. The closed form expression of γ_l's is a result of the data augmentation, which simplifies the maximization task and enhance the convergence speed of the algorithm.

REMARK 2—The transformation parameter r can be chosen based on the log-likelihood values after convergence. Assume we have grids $r_1 < \cdots < r_M$, the log-likelihood values are calculated for each r_m, $m = 1, \cdots, M$, and we choose the r_m with the largest log-likelihood value.

3.5 Variance Estimation

One of the direct appeal of this approach is that we can obtain close form solutions for the variance of $\hat{\theta}$ through Louis method (Louis, 1982). That is,

$$Var(\hat{\theta}) = (-\mathcal{H})^{-1},$$

where $\mathcal{H}$ is the hessian matrix, which can be estimated by

$$\mathcal{H} = \frac{\partial^2 \mathcal{Q}(\theta; \hat{\theta})}{\partial \theta \partial \theta'} + Var\left(\frac{\partial log \mathcal{L}_c(\theta)}{\partial \theta}\right).$$

The calculation of all quantities in the above hessian matrix has closed form and is listed in the Appendix. Large sample properties are preserved and the wald-type inferences for θ can be carried out in the regular way.

REMARK 3—If the number of nodes k is large, the dimension of the parameters will be large and the direct inversion of the observed information matrix may become infeasible. An alternative way for estimating the variance of the regression parameter estimators is to invert the observed information matrix based on the profile likelihood, as commonly done in the nonparametric maximum likelihood estimation literature (Murphy and Van der Vaart, 2000).

4 Numerical Study

4.1 Simulation Study

We generate data from the GORMC model

$$S(t|\boldsymbol{x}, z) = 1 - \pi(z) + \pi(z) \times S_u(t|\boldsymbol{x}),$$

with two variables in covariate matricies $\boldsymbol{x}$ and $\boldsymbol{z}$: x_1 and z_1 follow the uniform distribution ~ $U(0, 2)$, x_2 and z_2 follow the Bernoulli distribution $Ber(.5)$. The coefficients are set to be $\boldsymbol{\beta} = (1, 1)$ and $\boldsymbol{\eta} = (0, 1, 1)$, yielding an average cure probability of 40%. The baseline transformation function has the form of $H_0(t) = \log(1 + t) + t^{3/2}$. Different models with $r = 0$, 0.5, 1 and 2 are considered.

We generate the uncure indicator $\boldsymbol{u} = (u_1, u_2, \cdots, u_n)$ through the Bernoulli distribution with probability $\pi(\boldsymbol{z})$. The interval censored survival times are generated in the following way: assume the ith subject has ordered visit times $0 = V_{i0} < V_{i1} < \cdots < V_{i\nu_i}$ during the study, where the total number of visits of each subject ν_i is generated as one plus a Poisson random variable with mean ς. The lengths between two consecutive visits are $\tau_i^{(1)}, \cdots, \tau_i^{(\nu_i)} \overset{iid}{\sim} Exponential(\kappa)$, where $\tau_i^{(j)} = V_{ij} - V_{i(j-1)}$ for $j = 1, \cdots, \nu_i$. κ and ς are adjusted to have a 10% left censoring rate and a 50% right censoring rate under different models. We have left censoring if the generated survival time $T_i < V_{i1}$ and the patient is not cured ($u_i = 1$). We have right censoring if $T_i > V_{i\nu_i}$ or the subject is cured ($u_i = 0$). Otherwise, if the patient is not cured $u_i = 1$, and $T_i \in (V_{ij}, V_{i(j+1)}]$ for $j = 1, \cdots, \nu_i - 1$, we have interval censoring.

According to Wang et al. (2016), we use 5 equally spaced knots at percentiles and order of 3 for the monotone splines. The proposed EM algorithm is compared with the numerical method that directly maximizes the observed likelihood function (3), named “Direct MLE” (Nelder and Mead, 1965). We use sample size of 500 and implement 1000 replications for each simulation setting. The bias, average estimated standard error (StErr), empirical standard deviation (StDev) and 95% empirical coverage probability (CP) are reported in Table 1. From the results, we find the biases of the proposed EM algorithm are smaller than those from the Direct MLE method in most of the settings. The estimated standard error of the EM algorithm is comparable with the empirical approach and the CP is close to the nominal level 0.95.

We further investigate the estimated baseline survival curves, and present the mean of estimated baseline survival curves along with 2.5% and 97.5% quantiles under all settings in Figure 2. The estimated baseline survival curves are close to the truth in all cases. The summaries of convergence time are listed in Table 2, which shows the proposed EM algorithm converges much faster than the Direct MLE method in all cases. This advantage is because the data augmentation simplifies the structure of the likelihood function and make the maximization steps easier.

4.2 Application to the Aerobics Center Longitudinal Study (ACLS)

Here, we apply the proposed method to the ACLS data (Blair et al., 1996; Lee et al., 2012) to investigate the association between age at diagnosis of hypertension and some potential

risk factors for young adults (baseline age less than 45 years old). The dataset contained 1,611 females and 7,703 males. The event time of interest is the diagnosis age of hypertension (HTN), which is defined as the systolic blood pressure (SBP) being greater than 140 mmHg or the diastolic blood pressure (DBP) being greater than 90 mmHg. Each subject will visit the Cooper Clinic for periodic preventive medical examination, and the blood pressure is tested in each visit. Diagnosis of HTN can be determined between two consecutive visits. The Turnbull nonparametric survival curve in Figure 1 have a leveling off at the end of study, which indicates a potential "cured" subgroup among the study population. Therefore, it is appropriate to apply the proposed model to address the question of interest.

We fit the GORMC class of regression models to this data with the fitness levels, gender and BMI in both the cure rate and survival parts. The parameter r is determined by the grid search with r ranges from 0 to 3 by 0.2. Similarly, the monotone splines have order 3 with 5 equally spaced knots at percentiles. The estimated observed log-likelihood values are plotted in Figure 3(a). The model with $r = 2$ gives the greatest log-likelihood value −8986.571. The estimates of the parameters and their standard errors for this model are reported in Table 3. The estimated survival curves for the three fitness levels are plotted for males and females in Figure 3(b), where the median values of BMI in each subgroup are used. Figure 3(b) reflects the tendency that females have lower probability to develop HTN than males.

There is no significant difference regarding whether males or females will have HTN, but males tend to have higher probability to develop HTN comparing with females at the same age after adjusting for the fitness levels and BMI. BMI plays a significant role in developing the HTN. With higher BMI, individuals are more likely to develop HTN. For individuals having potential to develop HTN, the probability to HTN increases with the increase of BMI. Fitness is not statistically significant for HTN.

4.3 Software

We develop the R package named "`GORCure`" based on the proposed method and contribute to CRAN for public use. The main function is `GORMC()`, where the formula for the survival part and the cure rate part are specified using the arguments `survfun` and `curefun`, separately. The transformation parameter `r` can be specified as any nonnegative numbers. For example, the following code fit the ACLS data with $r = 2$.

```
> fit<-GORMC(survfun = Surv(left,right) ~ BMI + Female + Medium + High,
curefun = ~ BMI + Female + Medium + High, data = ACLS, r = 2)
```

A summary table for the coefficients, including the estimates, standard errors and test results, as well as the log-likelihood value can be obtained by `summary(fit)`. The `predict` function can be used to obtain the estimated cure rate and survival curve of a new individual, and the covariates in the cure rate part and the survival part are specified by `new.z` and `new.x`, respectively.

```
> pred<-predict(fit, new.x = c(median(ACLS$BMI), 0, 0, 0), new.z = c(1,
median(ACLS$BMI), 0, 0, 0))
> pred$CureRate
> plot(pred)
```

5 Discussion

We proposed an EM algorithm for the generalized odds rate mixture cure (GORMC) model with interval censored data. The proposed algorithm is computationally efficient and the variance estimates have closed forms using the Louis method. The choice of r based on the loglikelihood values will give an indication for model selection in practice as illustrated by the real data analysis in Section 4.2.

We tried different number of knots and orders for the monotone splines in the simulation study and found little change in the coefficient estimates. Based on the experiences in the simulation, the monotone splines with a small number of k, such as 5-8, is flexible enough to accurately estimate $H_\epsilon(t)$ on the observed time window $[0, t_{max}]$, where $t_{max} = \max(L_i, R_i; i = 1,\cdots, n)$ is the last observation time. The estimated parameters including those for the uncure rate have little bias (Details in Supplementary file).

The identifiability of mixture cure models has been widely studied in the literature, for example, Li et al. (2001); Hanin and Huang (2014); Liu and Shen (2009). With the monotone spline representation considered in our paper, the identifiability of the model can be similarly shown in theory. With finite samples, the proposed estimation method may suffer the potential weak identifiability issue when there is only one binary covariate in the uncure rate $\pi(z)$, as commonly occurred in most estimation procedures for cure models. In nonparametric estimation, Taylor (1995) set the tail of $S_{u0}(t)$ to zero after the last event time, which forces the uncured survival function to be proper. Similar ideas can be applied for methods with splines. For example, Bremhorst and Lambert (2016) discussed the estimation of the promotion time cure model with the P-splines and fixed the last spline parameter to a large enough value, which forces the estimated baseline survival function of uncured subjects to be 0 at the end of the follow-up. A similar technique can also be adopted in the proposed estimation method.

The proposed method works pretty well with moderate to large sample size and moderate right censoring rate as shown in Section 4.1. When the sample size is small, say $N < 100$, or the left or right censoring is extremely large, say > 80%, the algorithm sometimes does not converge, and the estimated hessian matrix might be singular.

For mixed-case interval censored data, where some subjects have exact observation times in the study, the proposed algorithm can not be easily extended and some further research is expected in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We greatly appreciate Dr. Xuemei Sui and Dr. Steven N. Blair in the University of South Carolina for providing the ACLS study data and for their contributions in data interpretation.

Appendix: Calculation of $\mathcal{H}$

Hessian Matrix of PHMC Model

The first part of the hessian matrix $\mathcal{H}$ involves the second derivatives of $\mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$, which include the following quantities. All the expectations of latent variables are conditioned on the observed data $\boldsymbol{O}$ and the final estimate $\hat{\boldsymbol{\theta}}$. The conditions $\boldsymbol{O}$ and $\hat{\boldsymbol{\theta}}$ are omitted for notational simplicity.

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \eta_j \partial \eta_{j'}} = -\sum_{i=1}^{n} z_{ij} z_{ij'} \hat{\pi}(z_i)[1-\hat{\pi}(z_i)],$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \eta_j \partial \boldsymbol{\beta}_{j'}} = \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \eta_j \partial \gamma_{j'}} = 0,$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}} = -\sum_{i=1}^{n} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij'} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} [(1-\delta_{R,i})\widehat{H}_e(R_i) + \delta_{R,i} E(u_i) \widehat{H}_e(L_i)],$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}_j \partial \gamma_l} = -\sum_{i=1}^{n} \boldsymbol{x}_{ij} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} [(1-\delta_{R,i}) b_l(R_i) + \delta_{R,i} E(u_i) b_l(L_i)], \text{ and}$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \gamma_l \partial \gamma_{l'}} = -\sum_{i=1}^{n} \frac{1}{\hat{\gamma}_l^2} (\delta_{L,i} E(Y_{il}) + \delta_{I,i} E(W_{il})) I(l'=l).$$

The second part includes covariance between the first derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$. All of the quantities are listed below, where all the variance and covariance of latent variables are conditioned on the observed data $\boldsymbol{O}$ and the final estimate $\hat{\boldsymbol{\theta}}$.

$$Cov\left\{\frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \eta_j}, \frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \eta_{j'}}\right\} = \sum_{i=1}^{n} z_{ij} z_{ij'} Var(u_i),$$

$$Cov\left\{\frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \eta_j}, \frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{j'}}\right\} = -\sum_{i=1}^{n} \delta_{R,i} z_{ij} \boldsymbol{x}_{ij'} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \hat{H}_e(L_i) Var(u_i),$$

$$Cov\left\{\frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \eta_j}, \frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \gamma_l}\right\} = -\sum_{i=1}^{n} \delta_{R,i} z_{ij} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} b_l(L_i) Var(u_i),$$

$$Cov\left\{\frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_j}, \frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{j'}}\right\} = \sum_{i=1}^{n} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij'} [\delta_{L,i} Var(Y_i) + \delta_{I,i} Var(W_i) + \delta_{R,i} e^{2\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \hat{H}_e^2(L_i) Var(u_i)],$$

$$Cov\left\{\frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_j}, \frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \gamma_l}\right\} = \sum_{i=1}^{n} \frac{\boldsymbol{x}_{ij}}{\hat{\gamma}_l} [\delta_{L,i} Cov(Y_i, Y_{il}) + \delta_{I,i} Cov(W_i, W_{il})] + \delta_{R,i} \boldsymbol{x}_{ij} e^{2\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \hat{H}_e(L_i) b_l(L_i) Var$$

(u_i), and

$$Cov\left\{\frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \gamma_l}, \frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \gamma_{l'}}\right\} = \sum_{i=1}^{n} \frac{1}{\hat{\gamma}_l \hat{\gamma}_{l'}} [\delta_{L,i} Cov(Y_{il}, Y_{il'}) + \delta_{I,i} Cov(W_{il}, W_{il'})] + \delta_{R,i} e^{2\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} b_l(L_i) b_{l'}(L_i) Var$$

(u_i).

Next, we calculate the conditional variance and covariance of the latent variables in the above equations. First notice that we have $u_i^2 \equiv u_i$, so the conditional variance of the uncure indicator is

$$Var(u_i) = E(u_i^2) - [E(u_i)]^2 = E(u_i) - [E(u_i)]^2 = E(u_i)[1 - E(u_i)].$$

Since $Y_i(W_i)$ conditionally follows zero truncated Poisson with mean $\hat{\lambda}_i(\hat{\omega}_i)$, letting $\hat{c}_i = 1 - \exp(-\hat{\lambda}_i)(\hat{d}_i = 1 - \exp(-\hat{\omega}_i))$, we have

$$Var(Y_i) = \frac{\delta_{L,i} \hat{\lambda}_i}{\hat{c}_i^2} (\hat{c}_i - \hat{\lambda}_i + \hat{c}_i \hat{\lambda}_i), \text{ and}$$

$$Var(W_i) = \frac{\delta_{I,i} \hat{\omega}_i}{\hat{d}_i^2} (\hat{d}_i - \hat{\omega}_i + \hat{d}_i \hat{\omega}_i).$$

The conditional variance and covariance of Y_{il} (W_{il}) can be obtained by using the iterated rules for the variance/covariance and recognizing that the conditional distribution of Y_{il} (W_{il}) given Y_i (W_i) is binomial. Their explicit forms are

$$Var(Y_{il}) = \frac{\delta_{L,i}\hat{\lambda}_{il}}{\hat{c}_i^2}(\hat{c}_i - \hat{\lambda}_{il} + \hat{c}_i\hat{\lambda}_{il}),$$

$$Cov(Y_{il}, Y_{il'}) = \frac{\delta_{L,i}\hat{\lambda}_{il}\hat{\lambda}_{il'}(\hat{c}_i - 1)}{\hat{c}_i^2},$$

$$Cov(Y_{il}, Y_i) = Cov(Y_{il}, \sum_{l'=1}^{k} Y_{il'}) = Var(Y_{il}) + \sum_{l' \neq l} Cov(Y_{il}, Y_{il'}) = \frac{\delta_{L,i}\hat{\lambda}_{il}}{\hat{c}_i^2}[\hat{c}_i + \hat{\lambda}_i(\hat{c}_i - 1)],$$

$$Var(W_{il}) = \frac{\delta_{I,i}\hat{\omega}_{il}}{\hat{d}_i^2}(\hat{d}_i - \hat{\omega}_{il} + \hat{d}_i\hat{\omega}_{il}),$$

$$Cov(W_{il}, W_{il'}) = \frac{\delta_{I,i}\hat{\omega}_{il}\hat{\omega}_{il'}(\hat{d}_i - 1)}{\hat{d}_i^2}, \text{ and}$$

$$Cov(W_{il}, W_i) = Cov(W_{il}, \sum_{l'=1}^{k} W_{il'}) = Var(W_{il}) + \sum_{l' \neq l} Cov(W_{il}, W_{il'}) = \frac{\delta_{I,i}\hat{\omega}_{il}}{\hat{d}_i^2}[\hat{d}_i + \hat{\omega}_i(\hat{d}_i - 1)].$$

Hessian Matrix of GPOMC Model

The first part of the hessian matrix $\mathcal{H}$ involves the second derivatives of $\mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$, which includes the following quantities:

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \eta_j \partial \eta_{j'}} = -\sum_{i=1}^{n} z_{ij} z_{ij'} \hat{\pi}(z_i)[1 - \hat{\pi}(z_i)],$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \eta_j \partial \boldsymbol{\beta}_{j'}} = \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \eta_j \partial \gamma_{j'}} = 0,$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}} = -\sum_{i=1}^{n} x_{ij} x_{ij'} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} E(u_i \phi_i)[(1-\delta_{R,i})\widehat{H}_e(R_i) + \delta_{R,i}\widehat{H}_e(L_i)],$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \gamma_l \partial \boldsymbol{\beta}_j} = -\sum_{i=1}^{n} x_{ij} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} E(u_i \phi_i)[(1-\delta_{R,i}) b_l(R_i) + \delta_{R,i} b_l(L_i)], \text{ and}$$

$$\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})}{\partial \gamma_l \partial \gamma_{l'}} = -\sum_{i=1}^{n} \frac{1}{\hat{\gamma}_l^2} (\delta_{L,i} E(Y_{il}) + \delta_{I,i} E(W_{il})) I(l' = l).$$

The second part includes covariance between the first derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$, all of the quantities are listed below.

$$Cov\left\{\frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \eta_j}, \frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \eta_{j'}}\right\} = \sum_{i=1}^{n} z_{ij} z_{ij'} Var(u_i),$$

$$Cov\left\{\frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \eta_j}, \frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{j'}}\right\} = -\sum_{i=1}^{n} \delta_{R,i} z_{ij} \boldsymbol{x}_{ij'} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \widehat{H}_e(L_i) E(u_i \phi_i)[1 - E(u_i)],$$

$$Cov\left\{\frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \eta_j}, \frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \gamma_l}\right\} = -\sum_{i=1}^{n} \delta_{R,i} z_{ij} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} b_l(L_i) E(u_i \phi_i)[1 - E(u_i)],$$

$$Cov\left\{\frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \gamma_j}, \frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \gamma_{l'}}\right\} = \sum_{i=1}^{n} \frac{1}{\hat{\gamma}_l \hat{\gamma}_{l'}} [\delta_{L,i} Cov(Y_{il}, Y_{il'}) + \delta_{I,i} Cov(W_{il}, W_{il'})]$$
$$- e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \left\{\frac{1}{\hat{\gamma}_l} b_{l'}(R_i)[Cov(Y_{il}, \phi_i) + \delta_{I,i} Cov(W_{il}, \phi_i)] + \hat{\gamma}_{l'} b_l(R_i)[\delta_{L_i} Cov(Y_{il'}, \phi_i) + \delta_{I_i} Cov(W_{il'}, \phi_i)]\right\} + e^{2\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} Var$$
$$(u_i \phi_i)[(1-\delta_{R,i}) b_l(R_i) b_{l'}(R_i) + \delta_{R,i} b_l(L_i) b_{l'}(L_i)],$$

$$Cov\left\{\frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_j}, \frac{\partial \log \mathcal{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{j'}}\right\} = \sum_{i=1}^{n} x_{ij} x_{ij'} \left\{[\delta_{L,i} Var(Y_i) + \delta_{I,i} Var(W_i)] - 2e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \widehat{H}_e(R_i)[\delta_{L,i} Cov(Y_i, \phi_i)\right.$$
$$\left. + \delta_{I,i} Cov(W_i, \phi_i)] + e^{2\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} Var(u_i \phi_i)[(1-\delta_{R,i}) \widehat{H}_e^2(R_i) + \delta_{R,i} \widehat{H}_e^2(L_i)]\right\},$$

and

$$Cov\left\{\frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \gamma_l}, \frac{\partial \log \mathscr{L}_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_j}\right\} = \sum_{i=1}^{n} \frac{x_{ij}}{\hat{\gamma}_l}\left\{\delta_{L,i} Cov(Y_{il}, Y_i) + \delta_{I,i} Cov(W_{il}, W_i) - e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} \widehat{H}_e(R_i)[\delta_{L,i} Cov(Y_{il}, \phi_i) + \delta_{I,i} Cov(W_{il}, \phi_i)]\right\} - x_{ij} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} b_l(R_i)[\delta_{L,i} Cov(Y_i, \phi_i) + \delta_{I,i} Cov(W_i, \phi_i)] - x_{ij} e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i} Var(u_i \phi_i)[(1 - \delta_{R,i}) b_l(R_i) \widehat{H}_e(R_i) + \delta_{R,i} b_l(L_i) \widehat{H}_e(L_i)].$$

Next, we calculate the conditional variance and covariance of the latent variables in the above equations. Similar as in the PHMC model, we have

$$Var(u_i) = E(u_i)[1 - E(u_i)].$$

Let $\hat{c}_i = 1 - [\delta_{L,i}/(1 + \hat{\lambda}_i)]^{1/r}, \hat{d}_i = 1 - [\delta_{I,i}(1 + \hat{\lambda}_i)/(1 + \hat{\lambda}_i + \widehat{\omega}_i)]^{1/r}$ and $\hat{e}_i = 1 - \hat{\pi}(z_i) + \hat{\pi}(z_i)[\delta_{R,i}/(1 + \widehat{\omega}_i)]^{1/r}$, the conditional variances can be calculated directly as

$$Var(Y_i) = \frac{\delta_{L,i} \hat{\lambda}_i}{\hat{c}_i^2}[(1 + r)\hat{c}_i \hat{\lambda}_i + \hat{c}_i - \hat{\lambda}_i, \text{ and}$$

$$Var(W_i) = \frac{\delta_{I,i} \widehat{\omega}_i}{(1 + r\hat{\lambda}_i)^2 \hat{d}_i^2}[(1 + r)\widehat{\omega}_i \hat{d}_i + \hat{d}_i(1 + r\hat{\lambda}_i) - \widehat{\omega}_i].$$

The conditional variance of $u_i \boldsymbol{\phi}_i$ is

$$Var(u_i \phi_i) = \frac{\delta_{L,i}}{\hat{c}_i^2}\left\{\hat{c}_i(r + 1)[1 - (1 - \hat{c}_i)^{2r+1}] - [1 - (1 - \hat{c}_i)^{r+1}]^2\right\} + \frac{\delta_{I,i}}{\hat{d}_i^2(1 + r\hat{\lambda}_i)^2}\left\{\hat{d}_i(r + 1)[1 - (1 - \hat{d}_i)^{2r+1}] - [1 - (1 - \hat{d}_i)^{r+1}]^2\right\} + \frac{\delta_{R,i} \hat{\pi}(z_i)}{\hat{e}_i^2(1 + r\widehat{\omega}_i)^{2+1/r}}\left\{(r + 1)\hat{e}_i - \hat{\pi}(z_i)(1 + r\widehat{\omega}_i)^{-1/r}\right\}.$$

The joint pdf of $(Y_{i1}, \cdots, Y_{ik}, \boldsymbol{\phi}_i)$ is

$$f(Y_{i1}, \cdots, Y_{ik}, \phi_i) = \frac{\delta_{L,i}^{I(\sum_{l=1}^{k} Y_{il} > 0)} \phi_i^{\sum_{l=1}^{k} Y_{il} + \frac{1}{r} - 1} e^{-\phi_i(\frac{1}{r} + \lambda_i)} \prod_{l=1}^{k} \frac{1}{Y_{il}!} \lambda_{il}^{Y_{il}}}{c_i \Gamma\left(\frac{1}{r}\right) r^{\frac{1}{r}}},$$

based on which, we have the conditional covariance between Y_{il} and $\boldsymbol{\phi}_i$ is

$$Cov(Y_{il}, \phi_i) = \frac{\delta_{L,i} \hat{\lambda}_{il}}{\hat{c}_i^2}[\hat{c}_i(r + 1) + (1 - \hat{c}_i)^{r+1} - 1].$$

Furthermore, using the fact that $Y_i \overset{d}{=} \sum_{l=1}^{k} Y_{il}$, we have

$$Cov(Y_i, \phi_i) = \sum_{l=1}^{k} Cov(Y_{il}, \phi_i) = \frac{\delta_{L,i}\hat{\lambda}_i}{\hat{c}_i^2}[\hat{c}_i(r+1) + (1-\hat{c}_i)^{r+1} - 1].$$

Similarly, the joint pdf of $(W_{i1}, \cdots, W_{ik}, \phi_i)$ is

$$f(W_{i1}, \cdots, W_{ik}, \phi_i) = \frac{\delta_{I,i}(1+r\lambda_i)^{\frac{1}{r}} I(\sum_{l=1}^{k} W_{il} > 0)}{d_i \Gamma\left(\frac{1}{r}\right) r^{\frac{1}{r}}} \phi_i^{\sum_{l=1}^{k} W_{il} + \frac{1}{r} - 1} e^{-\phi_i(\frac{1}{r} + \lambda_i + \omega_i)} \prod_{l=1}^{k} \frac{1}{W_{il}!} \omega_{il}^{W_{il}},$$

and the conditional covariance between W_{il} and ϕ_i is

$$Cov(W_{il}, \phi_i) = \frac{\delta_{I,i}\hat{\omega}_{il}}{\hat{d}_i^2(1+r\hat{\lambda}_i)^2}[\hat{d}_i(r+1) + (1-\hat{d}_i)^{r+1} - 1].$$

We also have

$$Cov(W_i, \phi_i) = \sum_{l=1}^{k} Cov(W_{il}, \phi_i) = \frac{\delta_{I,i}\hat{\omega}_i}{\hat{d}_i^2(1+r\hat{\lambda}_i)^2}[\hat{d}_i(r+1) + (1-\hat{d}_i)^{r+1} - 1].$$

The conditional variance and covariance of Y_{il} (W_{il}) can be obtained by using the iterated rules for the variance/covariance and recognizing that the conditional distribution of Y_{il} (W_{il}) given Y_i (W_i) is binomial. The explicit forms are Figure 2: Estimated Baseline Survival Functions

$$Cov(Y_{il}, Y_{il'}) = \frac{\delta_{L,i}\hat{\lambda}_{il}\hat{\lambda}_{il'}}{\hat{c}_i^2}[(1+r)\hat{c}_i - 1] + \frac{\delta_{L,i}\hat{\lambda}_{il}}{\hat{c}_i} I(l' = l),$$

$$Cov(Y_{il}, Y_i) = \frac{\delta_{L,i}\hat{\lambda}_{il}}{\hat{c}_i^2}\left\{\hat{c}_i + \hat{\lambda}_i[(1+r)\hat{c}_i - 1]\right\},$$

$$Cov(W_{il}, W_{il'}) = \frac{\delta_{I,i}\hat{\omega}_{il}\hat{\omega}_{il'}}{\hat{d}_i^2(1+r\hat{\lambda}_i)^2}[(1+r)\hat{d}_i - 1] + \frac{\delta_{I,i}\hat{\omega}_{il}}{\hat{d}_i(1+r\hat{\lambda}_i)} I(l' = l), \text{ and}$$

$$Cov(W_{il}, W_i) = \frac{\delta_{I,i}\widehat{\omega}_{il}}{\hat{d}_i^2(1+r\hat{\lambda}_i)^2}\left\{\hat{d}_i(1+r\hat{\lambda}_i) + \widehat{\omega}_i[(1+r)\hat{d}_i - 1]\right.$$

References

Banerjee T, Chen MH, Dey DK, Kim S. Bayesian analysis of generalized odds-rate hazards models for survival data. Lifetime data analysis. 2007; 13(2):241–260. [PubMed: 17401683]

Bennett S. Analysis of survival data by the proportional odds model. Statistics in medicine. 1983; 2(2): 273–277. [PubMed: 6648142]

Berkson J, Gage RP. Survival curve for cancer patients following treatment. Journal of the American Statistical Association. 1952; 47(259):501–515.

Bickel, P. Efficient testing in a class of transformation models. Proceedings of the 45th Session of the International Statistical Institute; 1986. p. 63-23.

Blair SN, Kampert JB, Kohl HW, Barlow CE, Macera CA, Paffenbarger RS, Gibbons LW. Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. Jama. 1996; 276(3):205–210. [PubMed: 8667564]

Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society Series B (Methodological). 1949; 11(1):15–53.

Bremhorst V, Lambert P. Flexible estimation in cure survival models using bayesian p-splines. Computational Statistics & Data Analysis. 2016; 93:270–284.

Cox, DR. Breakthroughs in Statistics. Springer; 1992. Regression models and life-tables; p. 527-541.

Dabrowska DM, Doksum KA. Estimation and testing in a two-sample generalized odds-rate model. Journal of the American Statistical Association. 1988; 83(403):744–749.

Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics. 1982:1041–1046. [PubMed: 7168793]

Gu Y, Sinha D, Banerjee S. Analysis of cure rate survival data under proportional odds model. Lifetime data analysis. 2011; 17(1):123–134. [PubMed: 20521166]

Hanin L, Huang LS. Identifiability of cure models revisited. Journal of Multivariate Analysis. 2014; 130:261–274.

Huang, J., Wellner, JA. Interval censored survival data: a review of recent progress. Proceedings of the First Seattle Symposium in Biostatistics; Springer; 1997. p. 123-169.

Kim YJ, Jhun M. Cure rate model with interval censored data. Statistics in medicine. 2008; 27(1):3–14. [PubMed: 17516589]

Kuk AY, Chen CH. A mixture model combining logistic regression with proportional hazards regression. Biometrika. 1992; 79(3):531–541.

Lam KF, Wong KY, Zhou F. A semiparametric cure model for interval-censored data. Biometrical Journal. 2013; 55(5):771–788. [PubMed: 23720128]

Lee, D-c, Sui, X., Church, TS., Lavie, CJ., Jackson, AS., Blair, SN. Changes in fitness and fatness on the development of cardiovascular disease risk factors: hypertension, metabolic syndrome, and hypercholesterolemia. Journal of the American College of Cardiology. 2012; 59(7):665–672. [PubMed: 22322083]

Li CS, Taylor JM, Sy JP. Identifiability of cure models. Statistics & Probability Letters. 2001; 54(4): 389–395.

Liu H, Shen Y. A semiparametric regression cure model for interval-censored data. Journal of the American Statistical Association. 2009; 104(487):1168–1178. [PubMed: 20354594]

Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1982:226–233.

Ma S. Cure model with current status data. Statistica Sinica. 2009; 19(1):233.

Ma S. Mixed case interval censored data with a cured subgroup. Statistica Sinica. 2010; 20:1165–1181.

Mao M, Wang JL. Semiparametric efficient estimation for a class of generalized proportional odds cure models. Journal of the American Statistical Association. 2010; 105(489):302–311. [PubMed: 22865944]

McMahan CS, Wang L, Tebbs JM. Regression analysis for current status data using the em algorithm. Statistics in medicine. 2013; 32(25):4452–4466. [PubMed: 23761135]

Murphy SA, Van der Vaart AW. On profile likelihood. Journal of the American Statistical Association. 2000; 95(450):449–465.

Nelder JA, Mead R. A simplex method for function minimization. The computer journal. 1965; 7(4): 308–313.

Peng Y, Dear KB. A nonparametric mixture model for cure rate estimation. Biometrics. 2000; 56(1): 237–243. [PubMed: 10783801]

Price DL, Manatunga AK. Modelling survival data with a cured fraction using frailty models. Statistics in medicine. 2001; 20(9-10):1515–1527. [PubMed: 11343371]

Ramsay JO. Monotone regression splines in action. Statistical science. 1988:425–441.

Scharfstein DO, Tsiatis AA, Gilbert PB. Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. Lifetime data analysis. 1998; 4(4):355–391. [PubMed: 9880995]

Sy JP, Taylor JM. Estimation in a Cox proportional hazards cure model. Biometrics. 2000; 56(1):227–236. [PubMed: 10783800]

Taylor JM. Semi-parametric estimation in failure time mixture models. Biometrics. 1995:899–907. [PubMed: 7548707]

Tsodikov A, Ibrahim J, Yakovlev A. Estimating cure rates from survival data. Journal of the American Statistical Association. 2003; 98(464)

Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society Series B (Methodological). 1976:290–295.

Wang L, McMahan CS, Hudgens MG, Qureshi ZP. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. Biometrics. 2016; 72(1):222–231. [PubMed: 26393917]

Zhou J, Zhang J, McLain AC, Cai B. A multiple imputation approach for semiparametric cure model with interval censored data. Computational Statistics and Data Analysis. 2016; 99:105–114.

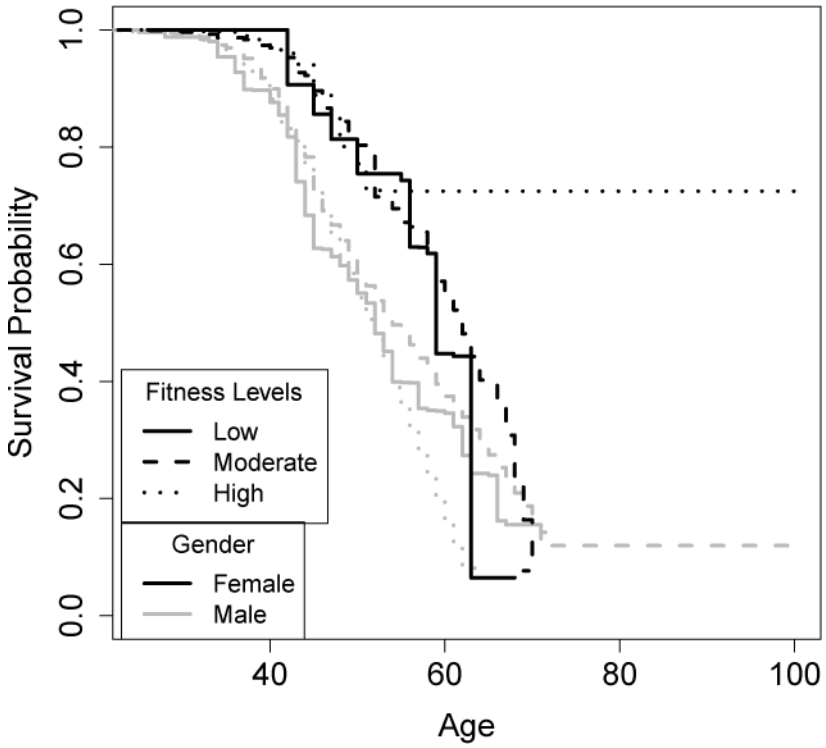


Figure 1.
Turnbull nonparametric survival curves for ACLS Dataset

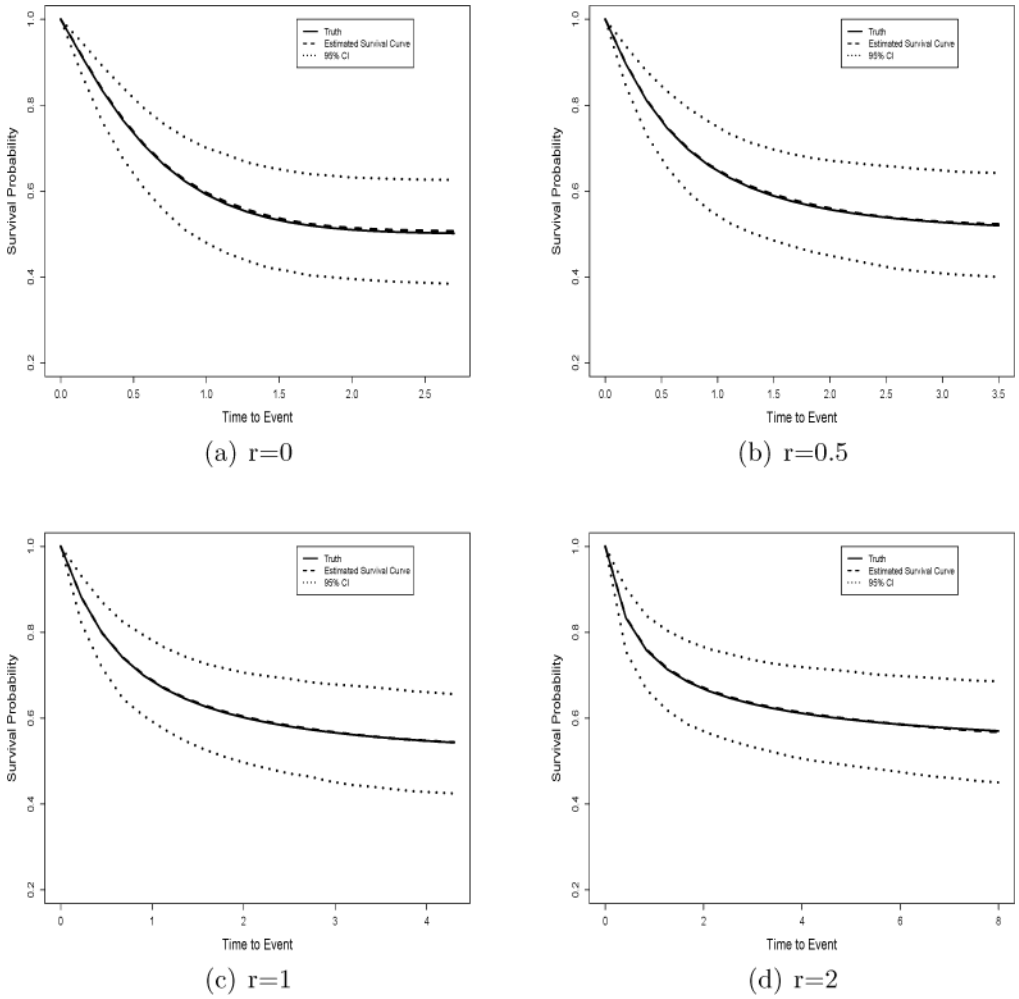
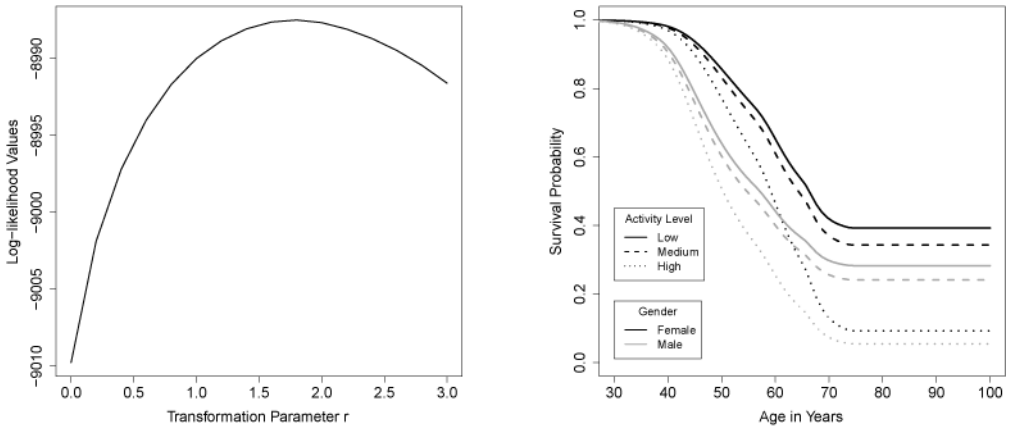


Figure 2.
Estimated Baseline Survival Functions



(a) Grid Search r (b) Survival Curve

Figure 3.
Estimated survival curves for ACLS Dataset

Table 1

Simulation Results

	Direct MLE				EM			
par.	**Bias**	**StDev**	**StErr**	**CP**	**Bias**	**StDev**	**StErr**	**CP**
$r = 0$								
η_0	0.02	0.50	0.26	0.67	0.00	0.25	0.25	0.96
η_1	−0.22	0.40	0.21	0.65	0.03	0.21	0.21	0.95
η_2	0.39	0.45	0.23	0.51	−0.02	0.24	0.24	0.96
β_1	−0.39	0.28	0.26	0.61	0.02	0.15	0.15	0.95
β_2	−0.25	0.30	0.21	0.67	−0.02	0.18	0.17	0.94
$r = 0.5$								
η_0	−0.01	0.44	0.27	0.77	0.01	0.27	0.27	0.96
η_1	−0.03	0.35	0.23	0.79	0.03	0.22	0.22	0.96
η_2	0.21	0.42	0.25	0.70	−0.03	0.26	0.25	0.94
β_1	−0.32	0.27	0.21	0.58	0.02	0.19	0.19	0.94
β_2	−0.15	0.30	0.23	0.80	−0.02	0.21	0.21	0.95
$r = 1$								
η_0	0.11	0.43	0.30	0.81	0.01	0.28	0.28	0.95
η_1	−0.03	0.37	0.26	0.82	0.01	0.25	0.24	0.94
η_2	0.14	0.47	0.28	0.73	−0.02	0.27	0.26	0.94
β_1	−0.33	0.33	0.26	0.63	0.03	0.22	0.22	0.95
β_2	0.02	0.36	0.27	0.81	−0.02	0.25	0.24	0.95
$r = 2$								
η_0	0.24	0.49	0.39	0.89	0.05	0.34	0.30	0.94
η_1	0.19	0.42	0.35	0.88	0.06	0.31	0.26	0.94
η_2	0.15	0.58	0.38	0.75	−0.06	0.33	0.29	0.94

	Direct MLE				EM			
par.	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP
β_1	−0.24	0.32	0.30	0.79	0.01	0.26	0.26	0.94
β_2	0.15	0.35	0.30	0.81	−0.01	0.29	0.29	0.95

Table 2

Convergence Time in Seconds

	Direct MLE					EM				
r	**Mean**	**Std**	**Median**	**Min**	**Max**	**Mean**	**Std**	**Median**	**Min**	**Max**
0	63.73	2.62	63.85	49.33	80.38	2.30	1.16	2.06	0.67	12.17
0.5	63.07	2.36	62.92	51.57	80.64	2.43	0.91	2.28	0.82	9.38
1	62.69	2.17	62.56	52.65	84.18	2.45	1.03	2.26	0.72	11.07
2	62.09	3.69	62.51	47.96	80.76	3.94	2.61	3.40	0.77	18.73

Table 3

Estimates for the ACLS Data

Variable	*Estimate*	*S.E.*	*P-value*
Uncure Rate			
Intercept	−3.044	1.021	0.003
BMI	0.158	0.029	< .001
Female	0.079	0.491	0.873
Medium	0.222	0.503	0.659
High	2.233	1.131	0.048
Survival Part			
BMI	0.067	0.015	< .001
female	−1.227	0.153	< .001
Medium	0.118	0.234	0.614
High	0.098	0.268	0.715