Published in final edited form as:

Comput Stat Data Anal. 2013 March ; 59: 95–102. doi:10.1016/j.csda.2012.09.017.

A New Semiparametric Estimation Method for Accelerated Hazards Mixture Cure Model

Jiajia Zhang^a, Yingwei Peng^b, and Haifen Li^{c,a}

^aDepartment of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208, USA

^bDepartment of Community Health and Epidemiology, Queen's University, Kingston, ON K7L 3N6, Canada

^cSchool of Finance and Statistics, East China Normal University, Shanghai, 200241, P. R. China

Abstract

The semiparametric accelerated hazards mixture cure model provides a useful alternative to analyze survival data with a cure fraction if covariates of interest have a gradual effect on the hazard of uncured patients. However, the application of the model may be hindered by the computational intractability of its estimation method due to non-smooth estimating equations involved. We propose a new semiparametric estimation method based on a smooth estimating equation for the model and demonstrate that the new method makes the parameter estimation more tractable without loss of efficiency. The proposed method is used to fit the model to a SEER breast cancer data set.

Keywords

EM algorithm; Kernel-smoothed approximation; Non-smooth estimating equation; Profile likelihood

1. Introduction

Survival data with a cure fraction has received a great deal of attention in recent years. In such data, a fraction of subjects will not experience the event of interest, however long the follow-up is, and they will be considered as cured. The remaining subjects will develop the event given sufficient follow-up and no drop-out and will be considered as uncured. Due to the presence of censoring in survival data, a sufficient follow-up is required in a study so that the cure rate can be reliably estimated (Maller and Zhou, 1994), and is assumed throughout this paper.

There are several statistical models proposed recently for survival data with a cure fraction. Each of them is based on certain model structures and assumptions that may be suitable for some problems, but not others. For example, Peng and Dear (2000), Sy and Taylor (2000) and Lu and Ying (2004) considered a mixture cure model where the distribution for the

failure time of uncured subjects is modeled by the proportional hazards (PH) model (referred to as PHMC model). They proposed semiparametric estimation methods to estimate the parameters in the models. Wang et al. (2011) propose a flexible mixture cure rate model with spline nonparametric forms for both the cure probability and the hazard rate function. The accelerated failure time (AFT) model is also considered to model the failure time of uncured subjects in the mixture cure model by Li and Taylor (2002) and Zhang and Peng (2007) (referred to as AFTMC model).

Both the PH model (Cox, 1972) and the AFT model (Cox and Oakes, 1984) in the cure models are popular in survival analysis because of simple estimation methods and easy interpretation. However, the assumptions in the models are not always applicable. One alternative to the models when the assumptions are not appropriate is the accelerated hazards (AH) model. This model was first considered by Chen and Wang (2000) and Chen (2001) for survival data without a fraction of cured subjects. Different from the PH model and the AFT model, the AH model is particularly useful to model the situation when the effect of treatments or other covariates is gradually released on the failure time distribution. That is, there is a lag period before the treatment is fully effective (Zucker and Lakatos, 1990). For example, Chen and Wang (2000) considered the effects of biodegradable carmustine (BCNU) polymers for treating patients with recurrent brain malignant gliomas in a randomized, placebo controlled clinical trial conducted by Brem et al. (1995). In the trial, the participants were randomized to the BCNU polymer treatment group and the placebo polymer control group. After the recurrent tumor was removed, a BCNU or placebo polymer was placed into the cavity. To reach a higher local drug concentration, the BCNU polymer was supposed to gradually release BCNU in the first 3-week period following placement. Due to the gradual effect, the AH model can be considered for the data while the PH model or the AFT model are not appropriate because the latter models implicitly expect the treatment to take an effect immediately following its application.

The AH assumption was considered in the mixture cure model by Zhang and Peng (2009a). They assumed the AH model for covariate effects on the failure time distribution of uncured subjects (referred to as AHMC model), and proposed a semiparametric rank-based method to estimate the parameters in the cure model. Zhang and Peng (2009b) investigated the conditions under which the hazard functions in the AH model corresponding to different covariate values do not cross. However, the application of the AHMC model is still very limited in the literature, partly due to the non-smooth estimating equations, which makes the semiparametric method computationally intractable. Therefore, finding a computationally tractable semiparametric estimation method is crucial to widen the application of the AHMC model and is the motivation of this work.

Zeng and Lin (2007) proposed an efficient semiparametric estimation method for the AFT model. The resulting estimating equation is smooth and thus computationally tractable. They also showed that the proposed method is efficient. This novel method demonstrated the possibility of converting non-smooth estimating equations into smooth estimating equations without loss of efficiency. The potential of this method was examined recently by Zhang et al. (2011) for the AH model and they demonstrated that the method can be applied to the AH model and it achieved a higher efficiency than the existing method. Lu (2010) considered this method for the AFTMC model and showed an improvement in estimation for the cure model. In this paper, we consider this method for the AHMC model and show that it also makes the estimation tractable for the AHMC model without loss of efficiency.

The paper is organized as follows. Section 2 describes the AHMC model and the proposed estimation method for the parameters in the model. A simulation study to investigate the performance of the proposed estimation method is given in Section 3, along with an

application of the model and estimation method to analyze a SEER breast cancer data set. The paper is concluded with some discussions in Section 4.

2. Accelerated Hazards Mixture Cure Model

2.1. The model

Let Y be the indicator variable for an uncured subject with $Y = 1$ if the subject is uncured and 0 if cured, T be the failure time of a subject. Define $\pi = P(Y = 1)$, $S(t) = P(T > t)$ and $S_u(t) = P(T > t | Y = 1)$. That is, π is the probability of being uncured, and $S(t)$ and $S_u(t)$ are the survival functions of the failure time of a subject and the survival function of the failure time of an uncured subject respectively. The mixture cure model is given by

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

where x and z are two sets of covariates that have effects on $\pi(\cdot)$ and $S_u(t)$. The use of the mixture cure model dates back to Berkson and Gage (1952). The advantage of the mixture cure model is that the proportion of cured subjects and the survival distribution of uncured subjects are modeled separately and the interpretation of the effects of x and z in the model is straightforward.

The common method to specify the effects of z on π is via a logit link function:

$$\pi(z) = \frac{\exp(\gamma^T z)}{1 + \exp(\gamma^T z)} \quad (2)$$

where γ is a vector of unknown parameters. Other link functions may be considered, such as the complementary log-log and the probit link functions in the generalized linear models for binary data. In this paper, we will use the logit link function only because of its simplicity and popularity.

The effects of x on $S_u(t)$ can be specified in various ways. Let $S_{u0}(t)$ be the baseline survival function of uncured subjects when $x = 0$. If $S_u(t|x) = S_{u0}(t)^{\exp(\beta^T x)}$, the cure model is the PHMC model. If $S_u(t|x) = S_{u0}(te^{\beta^T x})$, it is the AFTMC model. In this paper, we focus on the AHMC model, which is specified by equation (1), (2), and the AH model for $S_u(t|x)$:

$$S_u(t|x) = S_{u0}(te^{\beta^T x})^{\exp(-\beta^T x)} \quad (3)$$

If $h_{u0}(t)$ and $h_u(t)$ are the corresponding hazard functions of $S_{u0}(t)$ and $S_u(t)$ in (3), then $h_u(t|x) = h_{u0}(te^{\beta^T x})$, which clearly indicates a gradual effect of covariate x on the hazard function of uncured subjects in the AHMC model. This is in contrast to the PHMC model and the AFTMC model where the effect of x on the hazard rate of uncured subjects is immediate. Therefore, when a gradual effect of covariates on the hazard of uncured subjects is justifiable for data with a cured fraction, the AHMC model in (1), (2) and (3) provides a useful alternative to the existing cure models.

2.2. Estimation Procedure

Zhang and Peng (2009a) proposed a semiparametric method to estimate the parameters in the AHMC model. The method is based on the EM algorithm as follows. Let $(t_i, \delta_i, z_i, x_i)$ denote the observed data for the i th individual $i = 1, \dots, n$, where t_i is the observed survival time of T for the i th subject (may be censored), δ_i is a censoring indicator with $\delta_i = 1$ for uncensored t_i and $\delta_i = 0$ for censored t_i , and z_i, x_i are observed values of z and x for the i th

subject. The value of Y for the i th subject is denoted as y_i . Clearly for a censored subject, Y is a latent variable and its value is not observable. Denote $y = (y_1, \dots, y_n)$. The complete log likelihood in the EM algorithm when assuming all values of y are available is given by $l_c(\beta, h_{u0}(t), \gamma; y) = l_{c1}(\gamma; y) + l_{c2}(\beta, h_{u0}(t); y)$, where

$$l_{c1}(\gamma; y) = \sum_{i=1}^n y_i \log[\pi(z_i)] + (1 - y_i) \log[1 - \pi(z_i)],$$

$$l_{c2}(\beta, h_{u0}(\cdot); y) = \sum_{i=1}^n y_i \delta_i \log [h_{u0}(e^{\beta^T x_i} t_i)] + y_i e^{-\beta^T x_i} \log [S_{u0}(e^{\beta^T x_i} t_i)].$$

The E-step computes $E[l_c(\gamma, \beta, h_{u0}(t); y) | \Theta^{(m)}]$, the conditional expectation of the complete log-likelihood with respect to y , given the current estimates $\Theta^{(m)} = \{\gamma^{(m)}, \beta^{(m)}, h_{u0}^{(m)}(t)\}$. It is not difficult to see that

$$w_i^{(m)} = E(y_i | \Theta^{(m)}) = \delta_i + (1 - \delta_i) \frac{\pi(z_i) S_{u0}(e^{\beta^{(m)T} x_i} t_i) e^{-\beta^{(m)T} x_i}}{1 - \pi(z_i) + \pi(z_i) S_{u0}(e^{\beta^{(m)T} x_i} t_i)} \Big|_{\Theta = \Theta^{(m)}}. \quad (4)$$

Let $w^{(m)} = (w_1^{(m)}, \dots, w_n^{(m)})$. Then

$$E[l_c(\gamma, \beta, h_{u0}(t); y) | \Theta^{(m)}] = l_c[\gamma, \beta, h_{u0}(t); w^{(m)}] = l_{c1}[\gamma; w^{(m)}] + l_{c2}[\beta, h_{u0}(t); w^{(m)}]. \quad (5)$$

The M-step maximizes $l_{c1}(\gamma; w^{(m)})$ and $l_{c2}(\beta, h_{u0}(t); w^{(m)})$ with respect to the unknown parameters γ , β and $h_{u0}(t)$. Maximizing $l_{c1}(\gamma; w^{(m)})$ with respect to γ can be easily carried out using the Newton-Raphson algorithm, and the updated value of γ is denoted as $\hat{\gamma}$. Maximizing $l_{c2}(\beta, h_{u0}(t); w^{(m)})$ with respect to β and $h_{u0}(t)$ is a challenge task. Since $\delta_i w_i^{(m)} \equiv \delta_i, \delta_i \log w_i^{(m)} = 0, l_{c2}(\beta, h_{u0}(\cdot); w^{(m)})$ can be written as

$$\sum_{i=1}^n \delta_i \log [w_i^{(m)} h_{u0}(t_i e^{\beta^T x_i})] + w_i^{(m)} e^{-\beta^T x_i} \log [S_{u0}(t_i e^{\beta^T x_i})]$$

which can be treated as the log-likelihood function of the AH model considered by Chen

(2001) with the hazard function $w_i^{(m)} h_{u0}(t_i e^{\beta^T x_i})$. Based on this new form of the log-likelihood function, Zhang and Peng (2009a) proposed to update β from the following rank-type estimating equation:

$$\Psi(\beta; k(\cdot)) = \sum_{i=1}^n \delta_i k(t_i e^{\beta^T x_i}) \left(x_i - \frac{\sum_{j=1}^n x_j w_j^{(m)} e^{-\beta^T x_j} I(t_j e^{\beta^T x_j} \geq t_i e^{\beta^T x_i})}{\sum_{j=1}^n w_j^{(m)} e^{-\beta^T x_j} I(t_j e^{\beta^T x_j} \geq t_i e^{\beta^T x_i})} \right),$$

where $k(\cdot)$ is a general (predictable) weight function. If $k(u) = \frac{1}{n} \sum_{j=1}^n w_j^{(m)} e^{-\beta^T x_j} I(t_j e^{\beta^T x_j} \geq u)$, the corresponding estimating equation can be written as

$$\Psi(\beta) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \delta_i (x_i - x_j) w_j^{(m)} e^{-\beta^T x_j} I(t_j e^{\beta^T x_j} \geq t_i e^{\beta^T x_i}) \quad (6)$$

Like other rank-type estimating equations (Jin et al., 2003), this estimating equation is discontinuous and non-smooth, which makes the semiparametric estimation computationally intractable. Therefore, finding a computationally tractable semiparametric estimation method to update β and $h_{u0}(t)$ from $I_{c2}(\beta, h_{u0}(t); w^{(m)})$ is crucial to widen the application of the AHMC model.

Note $I_{c2}(\beta, h_{u0}(\cdot); w^{(m)})$ can also be written as

$$\sum_{i=1}^n \delta_i \log [h_{u0}(t_i e^{\beta^T x_i})] + w_i^{(m)} e^{-\beta^T x_i} \log [S_{u0}(t_i e^{\beta^T x_i})]$$

since $\delta_i w_i^{(m)} \equiv \delta_i$. Following Zeng and Lin (2007) and Zhang et al. (2011), we first assume a piecewise constant hazard function for $h_{u0}(t)$ by partition an interval containing all $t_i e^{\beta^T x_i}$ into J_n equally spaced intervals, $0 = t_0 < t_1 < \dots < t_{J_n} = M$, where M denotes an upper bound for the $t_i e^{\beta^T x_i}$ over all possible β 's in a bounded set. That is

$$h_{u0}(t) = \sum_{k=1}^{J_n} C_k I(t_{k-1} \leq t < t_k),$$

where $I(\cdot)$ is the indicator function. The corresponding $H_{u0}(t)$ is

$$H_{u0}(t) = \sum_{k=1}^{J_n} C_k (t - t_{k-1}) I(t_{k-1} \leq t < t_k) + \frac{M}{J_n} \sum_{k=1}^{J_n} C_k I(t \geq t_k).$$

Then the likelihood function (6) can be rewritten as

$$\begin{aligned} & \frac{1}{n} \sum_{k=1}^{J_n} \log \left\{ C_k \sum_{i=1}^n \delta_i I(t_{k-1} \leq t_i e^{\beta^T x_i} < t_k) \right\} \\ & - \frac{1}{n} \sum_{k=1}^{J_n} C_k \left\{ \sum_{i=1}^n w_i^{(m)} e^{-\beta^T x_i} (t_i e^{\beta^T x_i} - t_{k-1}) I(t_{k-1} \leq t_i e^{\beta^T x_i} < t_k) + \frac{M}{J_n} \sum_{i=1}^n w_i^{(m)} e^{-\beta^T x_i} I(t_i e^{\beta^T x_i} \geq t_k) \right\} \end{aligned}$$

By differentiating with respect to C_k and setting the score equation equal to 0, we obtain

$$\begin{aligned} C_k &= \left[\sum_{j=1}^n \delta_j I(t_{k-1} \leq t_j e^{\beta^T x_j} < t_k) \right] \\ & \times \left[\sum_{j=1}^n w_j^{(m)} e^{-\beta^T x_j} (t_j e^{\beta^T x_j} - t_{k-1}) I(t_{k-1} \leq t_j e^{\beta^T x_j} < t_k) + \frac{M}{J_n} \sum_{j=1}^n w_j^{(m)} e^{-\beta^T x_j} I(t_j e^{\beta^T x_j} \geq t_k) \right]^{-1}. \end{aligned}$$

After plugging the solutions of C_k into the log-likelihood function (6), and discarding those components irrelevant to β , we obtain a profile log-likelihood function for β :

$$l_{c2p}(\beta) = \frac{1}{n} \sum_{k=1}^{J_n} \left\{ \left[\sum_{i=1}^n \delta_i I(t_{k-1} \leq t_i e^{\beta^T x_i} < t_k) \right] \times \log \left[\frac{J_n}{Mn} \sum_{j=1}^n \delta_j I(t_{k-1} \leq t_j e^{\beta^T x_j} < t_k) \right] \right. \\ \left. - \frac{1}{n} \sum_{k=1}^{J_n} \left[\sum_{i=1}^n \delta_i I(t_{k-1} \leq t_i e^{\beta^T x_i} < t_k) \right] \right. \\ \left. \times \log \left[\frac{J_n}{Mn} \sum_{j=1}^n w_j^{(m)} e^{-\beta^T x_j} (t_j e^{\beta^T x_j} - t_{k-1}) I(t_{k-1} \leq t_j e^{\beta^T x_j} < t_k) \right. \right. \\ \left. \left. + \frac{1}{n} \sum_{j=1}^n w_j^{(m)} e^{-\beta^T x_j} I(t_j e^{\beta^T x_j} \geq t_k) \right] \right\}. \quad (7)$$

Following similar discussions in Zeng and Lin (2007) and Lu (2010), we can show that $l_{c2p}(\beta)$ converges uniformly, on a compact set of β , to

$$l(\beta) = \int_0^\infty \log \frac{dP(\delta=1, T e^{\beta^T x} \leq s)/ds}{E w e^{-\beta^T x}} dP(\delta=1, T e^{\beta^T x} \leq s), \quad (8)$$

as $n \rightarrow \infty$, $J_n \rightarrow \infty$ and $n/J_n \rightarrow \infty$. Then we can approximate (8) using kernel functions and the empirical observations. Let $K(\cdot)$ be a kernel function with bandwidth a_n . A kernel-smoothed approximation of (8) is

$$l_{c2p}^s(\beta) = -\frac{1}{n} \sum_{i=1}^n \delta_i e_i(\beta) + \frac{1}{n} \sum_{i=1}^n \delta_i \log \left[\frac{1}{na_n} \sum_{j=1}^n \delta_j K \left(\frac{e_j(\beta) - e_i(\beta)}{a_n} \right) \right] - \frac{1}{n} \sum_{i=1}^n \delta_i \log \left[\frac{1}{n} \sum_{j=1}^n \int_{-\infty}^{[e_j(\beta) - e_i(\beta)]/a_n} w_j^{(m)} e^{-\beta^T x_j} K(s) ds \right], \quad (9)$$

where $e_i(\beta) = \log t_i + \beta^T x_i$. As a result, instead of maximizing $l_{c2p}(\beta)$ in (7) with respect to β , we will maximize $l_{c2p}^s(\beta)$ in (9) and J_n is not involved in the maximization. If we choose a smooth kernel function for $K(\cdot)$, $l_{c2p}^s(\beta)$ will be a smooth function of β and it can be maximized by the Newton-Raphson method. For the sake of convenience, we choose the standard normal density function as the kernel function $K(\cdot)$. There are several choices of bandwidths as mentioned in Zeng and Lin (2007), such as $\sigma n^{-1/3}$, $\sigma n^{-1/5}$ and $\sigma n^{-1/7}$, where σ is the standard deviation of $\log T$ which can be obtained from observed data.

Denote the estimate of β from (9) as $\hat{\beta}$. Given $\hat{\beta}$, the baseline hazard function and the survival function can be estimated by

$$\hat{h}_{u0}(t) = \left[\frac{1}{(na_n t)} \sum_{i=1}^n \delta_i K \left(\frac{e_i(\hat{\beta}) - \log t}{a_n} \right) \right] \times \left[\frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{[e_i(\hat{\beta}) - \log t]/a_n} w_i^{(m)} e^{-\beta x_i} K(u) du \right]^{-1},$$

and

$$\hat{S}_{u0}(t) = \exp \left[- \int_{-\infty}^t \hat{h}_{u0}(s) ds \right]. \quad (10)$$

With the updated $\hat{\gamma}$, $\hat{\beta}$ and $\hat{S}_{u0}(t)$, $w_i^{(m)}$ can be updated in (4) in the E-step and the EM algorithm can be proceeded until convergence.

2.3. Variance Estimation

Let $\hat{\gamma}$, $\hat{\beta}$ and $\hat{h}_{u0}(t)$ denote the estimates from the EM algorithm, w_i be the value of $w_i^{(m)}$ in the last iteration, $l_{c1}[\gamma; w_i^{(m)}] = \sum_{i=1}^n l_{c1,i}(\gamma; w_i)$, and $l_{c2}[\beta, h_{u0}(t); w_i^{(m)}] = \sum_{i=1}^n l_{c2,i}(\beta, h_{u0}(t_i); w_i)$, where $l_{c1,i}(\gamma, w_i) = w_i \log[\pi(z_i)] + (1-w_i) \log[1-\pi(z_i)]$ and $l_{c2,i}(\beta, h_{u0}(t_i); w_i) = \delta_i \log[h_{u0}(e^{\beta^T x_i t_i})] + w_i e^{-\beta^T x_i} \log[S_{u0}(e^{\beta^T x_i t_i})]$. Let

$$S_i = \left(\frac{\partial l_{c1,i}(\gamma; w_i)}{\partial \gamma} \Big|_{\hat{\gamma}}, \frac{\partial l_{c2,i}(\beta, h_{u0}(t_i); w_i)}{\partial \beta} \Big|_{(\hat{\beta}, \hat{h}_{u0}(t_i))} \right)'$$

We propose to estimate the variances of the parameters by inverting the empirical Fisher

information matrix $\hat{I} = \sum_{i=1}^n S_i S_i'$. The first component in S_i can be obtained easily due to its closed form. To obtain the second component in S_i , we use a numerical differentiation method. For example, following the suggestion of Jamshidian and Jennrich (2000), a first-order Richardson extrapolation of the central difference to approximate $l_{c2,i}(\beta, h_{u0}(t_i); w_i)/\beta_j$ is

$$\frac{1}{12d} \left[l_{c2,i}(\beta - 2dv^{(j)}, h_{u0}(t_i); w_i) - 8l_{c2,i}(\beta - dv^{(j)}, h_{u0}(t_i); w_i) + 8l_{c2,i}(\beta + dv^{(j)}, h_{u0}(t_i); w_i) - l_{c2,i}(\beta + 2dv^{(j)}, h_{u0}(t_i); w_i) \right] \Big|_{(\hat{\beta}, \hat{h}_{u0}(t_i))} \quad (11)$$

where d is a small positive value, $v^{(j)}$ is a vector with its j th component equal to 1 and all others equal to 0.

3. Numerical Studies

3.1. Simulations

To evaluate the performance of the proposed method and compare it with existing estimation methods, we conduct a simulation study. In the simulation study, the data sets are generated from the AHMC model (1), (2) and (3). We consider a single binary covariate indicating a standard treatment and a new treatment, and the binary covariate has effects on both $\pi(\cdot)$ and $S_{u0}(t)$ with corresponding $(\gamma_0, \gamma_1) = (2, -1)$ and $\beta = 1$. These coefficient values indicate that the cure rates are about 12% for the standard treatment and 27% for the new treatment. For uncured subjects, the hazard of a subject at time t in the new treatment is equal to the hazard of a subject in the standard treatment at time $2t$. The baseline hazard function $h_{u0}(t)$ is assumed to be lognormal distribution with mean 1.65 and variance 4.67. The censoring time is generated from a uniform distribution between 0 and a and the value of a is chosen so that the corresponding censoring rate is about 30%. We generate 1000 simulation data sets with sample size of 250, 500 and 800.

We fit the AHMC model to each data set using two methods: the proposed method with the baseline survival estimator (10) (referred to as Profile), and the semiparametric method of Zhang and Peng (2009a) (referred to as Zhang & Peng). We compare the bias and variance of the parameter estimates from the two methods under various sample sizes. We also compare their computational time as a measure of computational tractability of these methods.

Table 1 shows biases (Bias), estimated empirical variances (Var) of γ_0 , γ_1 and β from the two methods under different sample sizes. The average of the asymptotic variance Var^* , obtained using the perturbation method of Section 2.3, and the corresponding coverage

probabilities of 95% confidence intervals (CP%) for the proposed method are also reported in the table. In estimating the parameters and their variances in the proposed method, we chose the bandwidth as $\sigma n^{-1/3}$ and let $d=1$ in (11) to estimate the variances. The results show that the proposed method performs well in estimating the parameters in the AHMC model. The biases of the estimates tend to be smaller than those from the Zhang and Peng's method. The estimated variances are close to the empirical variances and the corresponding CPs are close to the nominal 95%. Moreover, the proposed method is computationally more tractable than Zhang and Peng's method. We record the computational times of the two methods for data sets with different sample sizes and plot them in Figure 1. While the computational time required for Zhang and Peng's method rapidly increases as the sample size increases, the time required for the proposed method increases rather mildly. It is also easier to program the proposed method than Zhang and Peng's method because of the smooth estimating functions involved.

As we mentioned before, the proposed method can be extended to other link functions, such as the probit link and the complementary log-log (cloglog) link functions. A simulation study is conducted to investigate the performances of the model under the different link functions and different estimating methods. Under the three link functions, that is, the logit, probit, and cloglog link functions, we consider data sets with 250 sample size, 30% censoring rate and lognormal baseline distribution. Cure rate is set to 12% in the control group and 27% in the treatment group for all link functions, which means $\gamma_0 = 2$, $\gamma_1 = -1$ for the logit link, $\gamma_0 = 1.175$, $\gamma_1 = -0.562$ for the probit link, and $\gamma_0 = 0.75$, $\gamma_1 = -0.48$ for the cloglog link. The results from the true link function and the misspecified link functions are reported respectively in Table 2. From Table 2, we can see that even though the estimates of γ_0 and γ_1 are no longer valid estimates of the true parameter values when the link function is misspecified, the estimated values of β are still close to the true value 1 under different link functions and estimated cure rates (CR) from the estimated values of γ_0 and γ_1 are still close to the true cure rates.

3.2. SEER breast cancer data

As an application of the proposed estimation method, we consider a breast cancer data of Polk, Iowa from the SEER program. The data was analyzed by Zhang and Peng (2009a) using their estimation method. The purpose of this section is to estimate the parameters in the AHMC model using the proposed method and compare them with those obtained in Zhang and Peng (2009a).

The data set includes 1584 breast cancer patients diagnosed between 1995 and 2004. The maximum follow-up is near 10 years. There has four categories for the stage of the cancer: local, regional, distant and unstaged. Unstaged means information is not sufficient enough to assign a stage for the cancer. Thus, we exclude the unstaged cases when we extracted data from the SEER cancer incidence public-use database. Observations with missing values on stage are also excluded. To investigate the impact of stage of breast cancer on cancer survival, we consider the AHMC model because breast cancer is one of a few cancer diseases that is deemed to be curable for some patients and cure models have been used in analyzing survival data from breast cancer patients. The presence of cured patients in this data set is demonstrated by a large number of long term survivors. The AHMC model allows us to assess the effect of the stage on the cure rate and on the survival probability of uncured breast cancer patients.

We fit the AHMC model with stage as the single covariate in z and x . The parameters in the AHMC model are estimated using the proposed semiparametric estimation method and the results are compared with those reported in Zhang and Peng (2009a). Details of the estimates and their standard errors are presented in Table 3. In the table, γ_1 and β_1 represent

the effects of distant stage relative to local stage, and γ_2 and β_2 represent the effects of regional stage relative to local stage.

The results demonstrate that the standard error estimates from the proposed method tend to be smaller than those from Zhang and Peng (2009a), which turns γ_2 from insignificant in Zhang and Peng (2009a) into significant in the proposed method. It implies that both distant stage and regional stage have significantly smaller cure rates than the local stage. The significance of the differences in survival distribution of uncured patients between distant stage and local stage and between regional stage and local stage does not change very much under the two methods. But the difference between distant stage and local stage becomes stronger and the difference between regional stage and local stage becomes weaker in the proposed method than in Zhang and Peng (2009a).

For illustration, we plot the estimated marginal survival functions of the three stage groups based on the proposed method and Zhang and Peng's method in Figure 2 along with the Kaplan-Meier survival curve. The survival curves from both estimation methods fit reasonably well to the Kaplan-Meier survival curves. The smooth survival function curves from the proposed method appear to fit the Kaplan-Meier curves slightly better than the curves from Zhang and Peng's method, especially for the distant stage group.

4. Conclusions and Discussions

The accelerated hazards assumption is a useful alternative to the proportional hazards and accelerated failure time assumptions when a covariate's effect on failure time is gradual, not immediate. It has been employed in models for survival data with and without a cure fraction to assess the gradual effect of covariates. However, the existing estimation methods for the models are not computationally tractable because of non-smooth estimating functions involved. The standard errors of estimates often have to be estimated using the computationally intensive bootstrap method. In this paper, we demonstrate that the non-smooth estimating equation for the semiparametric AHMC model can be replaced with a smooth estimating equation. It greatly simplifies the estimation of the semiparametric AHMC model without loss of efficiency and the standard errors of the estimates can be readily obtained from the new method without resort to the bootstrap method.

The program for the proposed estimation method is available upon request from the first author.

Acknowledgments

The project is supported by an award to Drs. Jiajia Zhang and Yingwei Peng from the National Cancer Institute (NCI, Award Number R03CA137790). The content is solely the responsibility of the authors and does not necessarily represent the official views of NCI. We thank the Editor, the Associate Editor, and the two reviewers for their insightful comments, which have led to great improvements in the paper.

References

- Berkson J, Gage R. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*. 1952; 47:501–515.
- Brem H, Piantadosi S, Burger P, Walker M, Selker R, Vick N, Black K, Sisti M, Brem S, Mohr G, et al. Placebo-controlled trial of safety and efficacy of intraoperative controlled delivery by biodegradable polymers of chemotherapy for recurrent gliomas. *The Lancet*. 1995; 345:1008–1012.
- Chen Y. Accelerated hazards regression model and its adequacy for censored survival data. *Biometrics*. 2001; 57:853–860. [PubMed: 11550937]
- Chen Y, Wang M. Analysis of accelerated hazards models. *Journal of the American Statistical Association*. 2000; 95:608–618.

- Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 1972; 34:187–220.
- Cox, D.; Oakes, D. *Analysis of survival data*. New York: Chapman & Hall; 1984.
- Jamshidian M, Jennrich R. Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 2000; 62:257–270.
- Jin Z, Lin D, Wei L, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003; 90:341–353.
- Li C, Taylor J. A semi-parametric accelerated failure time cure model. *Statistics in Medicine*. 2002; 21:3235–3247. [PubMed: 12375301]
- Lu W. Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*. 2010; 20:661–674. [PubMed: 20414460]
- Lu W, Ying Z. On semiparametric transformation cure models. *Biometrika*. 2004; 91:331–343.
- Maller RA, Zhou S. Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association*. 1994; 89:1499–1511.
- Peng Y, Dear K. A nonparametric mixture model for cure rate estimation. *Biometrics*. 2000; 56:237–243. [PubMed: 10783801]
- Price D, Manatunga A. Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*. 2001; 20:1515–1527. [PubMed: 11343371]
- Sy J, Taylor J. Estimation in a Cox proportional hazards cure model. *Biometrics*. 2000; 56:227–236. [PubMed: 10783800]
- Wang L, Du P, Liang H. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*. 2011 (to appear).
- Zeng D, Lin D. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*. 2007; 102:1387–1396.
- Zhang J, Peng Y. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistical in Medicine*. 2007; 26:3157–3171.
- Zhang J, Peng Y. Accelerated hazards mixture cure model. *Lifetime Data Analysis*. 2009a; 15:455–467. [PubMed: 19697127]
- Zhang J, Peng Y. Crossing hazard functions in common survival models. *Statistics & Probability Letters*. 2009b; 79:2124–2130. [PubMed: 20613974]
- Zhang J, Peng Y, Zhao O. A new semiparametric estimation method for accelerated hazard model. *Biometrics*. 2011; 67:1352–1360. [PubMed: 21457194]
- Zucker D, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika*. 1990; 77:853–864.

Figure 1. Computational times required by the proposed method (profile) and the method of Zhang and Peng's method.

Figure 2.
Estimated survival curves for three stages based on the proposed method (profile), Zhang and Peng's method, and the Kaplan-Meier survival estimator

Table 1

Simulation Results for 1000 replications

<i>n</i>		Profile				Zhang & Peng			
		Bias	Var	Var*	CP%	Bias	Var	Var*	CP%
250	$\hat{\gamma}_0$	-0.040	0.152	0.127	92.1	-0.080	0.146		
	$\hat{\gamma}_1$	0.031	0.182	0.179	96.0	0.047	0.179		
	$\hat{\beta}$	0.098	0.224	0.120	93.2	-0.039	0.143		
500	$\hat{\gamma}_0$	-0.033	0.072	0.060	91.2	-0.060	0.065		
	$\hat{\gamma}_1$	0.010	0.090	0.086	94.2	0.022	0.085		
	$\hat{\beta}$	0.014	0.048	0.057	95.2	-0.024	0.106		
800	$\hat{\gamma}_0$	-0.034	0.044	0.040	91.1	-0.050	0.041		
	$\hat{\gamma}_1$	0.009	0.053	0.056	95.8	0.016	0.052		
	$\hat{\beta}$	-0.006	0.025	0.038	97.3	0.000	0.024		

Table 2

Simulation Results for 1000 replications from different link functions

True link	Parameter	Link function in the estimation					
		Logit		Probit		Cloglog	
		Bias	Var	Bias	Var	Bias	Var
Logit	$\hat{\gamma}_0$	-0.022	0.163	-0.845	0.042	-1.274	0.025
	$\hat{\gamma}_1$	0.020	0.198	0.455	0.056	0.538	0.039
	$\hat{\beta}$	0.084	0.094	0.096	0.105	0.092	0.093
	CR($z=0$)	12.76%		12.89%		12.99%	
	CR($z=1$)	27.19%		27.31%		27.28%	
Probit	$\hat{\gamma}_0$	0.806	0.168	-0.018	0.043	-0.446	0.025
	$\hat{\gamma}_1$	-0.430	0.205	0.009	0.058	0.090	0.039
	$\hat{\beta}$	0.042	0.079	0.058	0.080	0.085	0.092
	CR($z=0$)	12.75%		12.85%		12.94%	
	CR($z=1$)	27.38%		27.48%		27.54%	
Cloglog	$\hat{\gamma}_0$	1.222	0.136	0.403	0.036	-0.022	0.021
	$\hat{\gamma}_1$	-0.501	0.173	-0.067	0.050	0.011	0.035
	$\hat{\beta}$	0.035	0.140	0.047	0.111	0.065	0.116
	CR($z=0$)	12.75%		12.85%		12.91%	
	CR($z=1$)	27.34%		27.43%		27.48%	

Table 3

Parameter estimates and standard errors in the AHMC model for breast cancer data set

	Profile		Zhang & Peng	
	Estimate	Std.err	Estimate	Std.err
$\hat{\gamma}_0$	-0.938	0.106	-0.831	0.167
$\hat{\gamma}_1$	1.656	0.324	1.671	0.679
$\hat{\gamma}_2$	0.583	0.195	0.431	0.293
$\hat{\beta}_1$	1.681	0.232	1.212	0.239
$\hat{\beta}_2$	0.178	0.210	0.412	0.322