Title	A semiparametric cure model for interval-censored data
Author(s)	Lam, KF; Wong, KY; Zhou, F
Citation	Biometrical Journal, 2013, v. 55 n. 5, p. 771-788
Issued Date	2013
URL	http://hdl.handle.net/10722/189464
Rights	This is the accepted version of the following article: Biometrical Journal, 2013, v. 55 n. 5, p. 771-788, which has been published in final form at: http://onlinelibrary.wiley.com/doi/10.1002/bimj.201300004/abstract

Biometrical Journal **xx** (zzzz) yy, zzz–zzz / DOI: 10.1002/bimj.200100000

A semiparametric cure model for interval-censored data

Kwok Fai Lam[*,1], **Kin Yau Wong** [2], and **Feifei Zhou** [1]

[1] Department of Statistics and Actuarial Science, the University of Hong Kong, Pokfulam Road, Hong Kong
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Received January 2013, revised zzz, accepted zzz

There is a growing interest in the analysis of survival data with a cured proportion particularly in tumor recurrences studies. Biologically, it is reasonable to assume that the recurrence time is mainly affected by the overall health condition of the patient that depends on some covariates like age, sex or treatment type received. We propose a semiparametric frailty-Cox cure model to quantify the overall health condition of the patient by a covariate-dependent frailty that has a discrete mass at zero to characterize the cured patients, and a positive continuous part to characterize the heterogeneous health conditions among the uncured patients. A multiple imputation estimation method is proposed for the right-censored case, which is further extended to accommodate interval-censored data. Simulation studies show that the performance of the proposed method is highly satisfactory. For illustration, the model is fitted to a set of right-censored melanoma incidence data and a set of interval-censored breast cosmesis data. Our analysis suggests that patients receiving treatment of radiotherapy with adjuvant chemotherapy have a significantly higher probability of breast retraction, but also a lower hazard rate of breast retraction among those patients who will eventually experience the event with similar health conditions. The interpretation is very different to those based on models without a cure component that the treatment of radiotherapy with adjuvant chemotherapy significant increases the risk of breast retraction.

Key words: Asymptotic normal data augmentation; Compound Poisson distribution; Cure model; Interval-censored data; Multiple imputation.

1 Introduction

Medical and public health researches often involve the analysis of time to a specific event where some individuals under study are highly susceptible to the event while others are at much lower risk. This type of data can be found easily in many cancer studies with the variable of interest being the time to recurrence of cancer tumors. Examples are numerous in the literature such as the study of recurrence of breast cancer tumors (Peng and Dear, 2000; Lam *et al*., 2005), and the study of melanoma incidence (Chen *et al*., 1999; Ibrahim *et al*., 2001a,b) where the melanoma patients are reported to have relapse rates of about 60% to 75% (Kirkwood *et al*., 2000). Another application of cure models is on the study of vaccine efficacy. Some vaccinated subjects will develop immunity with complete protection that they will be lifetime free of the disease, while some individuals will only have partial or even no protection from the vaccination. It is of interest to investigate whether the vaccine is effective in developing immunity and/or in delaying the onset time of the disease. For simplicity, the term *cure* is used throughout this paper even though cure models have been studied in other areas like the analysis of time to first marriage (Aalen, 1992), duration of first marriage, duration of unemployment and recidivism.

A common type of cure models is a simple mixture model that broadly classifies individuals in the population into two groups using a binary random effect U, with $U = 1$ denoting the uncured group and $U = 0$ denoting the cured group. The random effect U is more generally called frailty in standard survival analysis. Among the cured subjects, the time to recurrence T is theoretically equal to ∞ that could not

*Corresponding author: e-mail: hrntlkf@hku.hk, Phone: (852) 2857 8320 Fax: (852) 2858 9041

be observed and must be right-censored in practice. Hence, $U = 0$ is not directly observable. In the uncured group, $T < \infty$ for every individual and is assumed to follow a certain distribution $F(t \mid \boldsymbol{x})$ that may depend on a vector of covariates $\boldsymbol{x}$. The variables U and T are called the incidence and the latency variables, respectively (Sy and Taylor, 2000).

Prolific discussions on the frailty-Cox proportional hazards (PH) model are available in the literature over the past two decades. In the model, the conditional hazard function of T is given by

$$\lambda(t \mid U = u, \boldsymbol{x}) = u\lambda_0(t) \exp\left(\boldsymbol{\beta}'\boldsymbol{x}\right), \tag{1}$$

where $\lambda_0(t)$ is an arbitrary non-negative baseline hazard function and $\boldsymbol{\beta}$ is a vector of regression coefficients. The cure probability given the vector of covariates $\boldsymbol{x}$ is generally defined to be $\mathrm{P}(U = 0 \mid \boldsymbol{x}) = 1 - \Phi\left(\boldsymbol{\theta}'\boldsymbol{x}\right)$ where $\boldsymbol{\theta}$ is a vector of regression parameters associated with the incidence variable and Φ is a binary link function like the logistic or probit function. One advantage of this mixture model is that we can determine whether a certain treatment is effective in curing the disease, or is only effective in delaying the time to relapse of the disease, or both. Estimation in the frailty-Cox PH model with a nonparametric baseline hazard function is not so straightforward. Various estimation methods have been developed for right-censored data by extending the Cox partial likelihood function (Lam *et al.*, 2005; Taylor, 1995). Estimation with the nonparametric baseline hazard function is extremely complicated with interval-censored data. Ma (2009, 2010) proposed using the penalized MLE and NPMLE to handle the nonparametric baseline hazard function for interval-censored data. The procedures are computationally feasible, but the standard error of the estimate is approximated by the nonparametric bootstrap procedures. To our knowledge, no satisfactory estimation method for interval-censored data with a cure based on (1) is available in the literature.

Price and Manatunga (2001) suggested that the cure model should be able to account for heterogeneity among individuals in the uncured group when analyzing data concerning the recurrence of leukaemia among patients receiving autologous transplantation treatment. They proposed to impose another non-negative random effect V that quantifies the heterogeneity induced by some unobservable risk factors, which acts multiplicatively on the conditional hazard function such that

$$\lambda(t \mid u, v, \boldsymbol{x}) = uv\lambda_0(t) \exp\left(\boldsymbol{\beta}'\boldsymbol{x}\right). \tag{2}$$

However, it would be more natural to use a single random effect instead of two separate random effects for controlling the cure probability and the heterogeneity in the uncured group. Moreover, the frailties V_i's are generally assumed to be independently and identically distributed that they do not depend on any covariates. Nevertheless, if we interpret the frailty as a quantification of some unobservable personal risk factors, it may also be affected by some observed covariates.

Another commonly used cure model is the promotion time cure model (Chen *et al.*, 1999; Ibrahim *et al.*, 2001a,b). This model assumes that the event of interest is triggered by the onset of one of the several latent events, where the number of latent events is random and covariate-dependent. This model has sound biological interpretation particularly in cancer studies, where a latent event corresponds to the progression of a carcinogenic cell, and the relapse of cancer is caused by the production of a detectable cancer mass by any of the carcinogenic cells present. The individual is considered cured when the number of carcinogenic cell is zero. In this model, the same set of covariates affects both the time to event and the cure probability. However, it assumes that the distributions of the onset times of the latent events are independent and identical for all individuals, which may be restrictive in practice. On the other hand, Yin and Ibrahim (2005) proposed a class of cure rate models that includes the usual mixture cure model and the promotion time cure model as special cases. This class of models also allows the survival distribution of the latency variable to depend on some explanatory variables. Bayesian inference via Markov Chain Monte Carlo was suggested, but again, restricted to right censored data only.

In this paper, we consider the semiparametric frailty-Cox PH model with a single frailty to accommodate both the cured proportion and the heterogeneity in the uncured group. Two illustrating data sets, namely the right-censored melanoma incidence data (Ibrahim *et al*., 2001a) and the interval-censored breast cosmesis data (Beadle *et al*., 1984a,b), are introduced in Section 2. A class of compound Poisson mixing distributions similar to the one discussed by Aalen (1992) is extended to allow the frailties to be covariate-dependent, so that both the incidence and latency variables may depend on some observed covariates like age and type of treatment received. For mathematical convenience, the mixing distribution considered here takes the form of the non-central Chi-square distribution with zero degrees of freedom proposed by Siegel (1979). The semiparametric frailty-Cox PH model is discussed in Section 3. A simple estimation procedure by means of multiple imputation for right-censored data is proposed in Section 4.1, and the method is extended to accommodate interval-censored data in Section 4.2. Simulations are conducted to assess the performance of the proposed estimation method and the results are reported in Section 5. In Section 6, the model is fitted to the two illustrating data sets. Concluding remarks and some possible future research directions are summarized in Section 7.

2 Data Sets

Two data sets are considered in this paper with the first one being a set of possibly right censored melanoma incidence data (Ibrahim *et al*., 2001a). The data were collected in an Eastern Cooperative Oncology Group phase III clinical trial, labeled as E1690, that began in 1991 to study the effects of low-dose interferon alpha-2b (IFN) relative to OBS - a combined group of the observation and the high-dose group on survival. IFN is an adjuvant post-operative chemotherapy proposed for the high risk melanoma patients and seemed to have significant impact on relapse-free survival. There are altogether 427 participants with a median follow-up time of 1.94 years. The largest uncensored failure time in the sample is $t^* = 5.065$ years. The Kaplan-Meier plot of the survival functions of the two treatment groups for this dataset, irrespective of the covariates like age and sex, is given in Figure 1. We can see that the two estimated survival functions level off at around 0.4 and there is a large portion of data right censored between 2 to 7 years.

The second dataset is from a retrospective study to compare the effects of adjuvant chemotherapy on early breast cancer patients treated with radiotherapy to those treated with radiotherapy alone with respect to the cosmetic effects of their treatment (Beadle *et al*., 1984a,b; Finkelstein and Wolfe, 1985). The subjects of the study were treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. In the study, the early breast cancer patients not treated with mastectomy were divided into two groups, with one group being treated with primary radiation therapy and adjuvant chemotherapy and another group being treated with radiotherapy only. The variable of interest was the time until breast retraction and the objective of the study was to compare the long-term cosmetic results in patients under the two treatment regimens. The patients were followed up to 60 months, with scheduled clinic visits every 4 to 6 months and hence the time to breast retraction of each subject was interval-censored. As patients might have missed some scheduled visits, intervals that contained the exact time to breast retraction were irregular and were likely to overlap with others. Upon scrutiny, more than half of the observations were right-censored in the radiotherapy-only treatment group, with most of the censoring times being comparatively large, which probably indicated that breast retraction would not occur in some patients.

3 Model Description

We consider here a random sample of size n where T_i denotes the actual event time and $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ the observed covariates of individual i, for $i = 1, \cdots, n$. Define $\boldsymbol{x}_i^{(0)} = \left(1, x_{i1}^{(0)}, \cdots, x_{ip_0}^{(0)}\right)'$ be the set of covariates associated with the incidence part and $\boldsymbol{x}_i^{(1)} = \left(x_{i1}^{(1)}, \cdots, x_{ip_1}^{(1)}\right)'$ be the set of covariates associated with the latency part ($p_0, p_1 \leq p$) and that $x_{ij}^{(k)}$ is an element in $\boldsymbol{x}_i$. In the right censoring setup,

the data collected are of the form $(y_i, \delta_i, \boldsymbol{x}_i; i = 1, \ldots, n)$ where $Y_i = \min\{T_i, C_i\}$, C_i is the (right) censoring time independent of T_i and δ_i is the censoring indicator with $\delta_i = I(T_i \leq C_i)$ so that $\delta_i = 1$ corresponds to an uncensored observation while $\delta_i = 0$ corresponds to a censored observation. For general interval-censored data, the observed data are of the form $(l_i, r_i, \delta_i, \boldsymbol{x}_i; i = 1, \ldots, n)$, where the event time T is not directly observable for all individuals, but is only known to be within the interval $(l_i, r_i]$. The censoring indicator is defined similarly with $\delta_i = 1$ if $r_i < \infty$ and $\delta_i = 0$ otherwise.

As mentioned in Section 1, we suggest a frailty model that involves only one covariate-dependent random effect U to accommodate the cured proportion in the population as well as the heterogeneity among the uncured individuals. The conditional hazard function of the semiparametric frailty-Cox model is then specified by

$$\lambda\left(t \mid u_i, \boldsymbol{x}_i^{(1)}\right) = u_i \lambda_0(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_i^{(1)}\right),$$

where $\lambda_0(t)$ is the unknown arbitrary non-negative baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_1})'$ is a vector of regression parameters associated with the latency variable. The individual frailty U_i is assumed to follow a non-central Chi-square distribution with zero degrees of freedom, which is an extension of the standard non-central Chi-square distribution to include the case of zero degrees of freedom (Siegel, 1979). The frailty U_i can be constructed as the sum of K_i independent central Chi-square random variables $(W_{1i}, W_{2i}, \ldots, W_{K_i i})$, each with 2 degrees of freedom and a scale parameter σ. Conditioned on $K_i = k_i$, we have

$$U_i = \begin{cases} 0, & \text{if } k_i = 0; \\ W_{1i} + W_{2i} + \cdots + W_{k_i i}, & \text{if } k_i > 0. \end{cases}$$

The number of terms in the sum, K_i, is a Poisson random variable with mean $\eta_i/2$, where $\eta_i = \exp\left(\boldsymbol{\theta}' \boldsymbol{x}_i^{(0)}\right)$ with $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{p_0})'$ being a vector of regression parameters associated with the frailty. In most cancer studies, we can conceive the random variable K to be the latent number of metastasis-competent tumor cells and W_{ij} is the contribution to U_i from the j-th latent tumor cell, say the relative aggressiveness of the tumor cell, and that the contributions from different cells are non-overlapping and additive. Marginally, the frailty U_i is a compound Poisson random variable with probability density function

$$g\left(u \mid \boldsymbol{x}_i^{(0)}\right) = \begin{cases} \exp(-\eta_i/2), & \text{if } u = 0; \\ \sum_{k=1}^{\infty} \frac{e^{-\eta_i/2}(\eta_i/2)^k}{k!} \times \frac{u^{k-1} e^{-u/2}}{(2\sigma)^k \Gamma(k)}, & \text{if } u > 0. \end{cases}$$

The frailty has a point mass at 0. It is a special case of the compound Poisson distribution which was considered as the frailty distribution for modeling the proportion of cured and heterogeneity in survival analysis by Aalen (1992). Moger *et al*. (2004); Moger and Aalen (2005, 2008) further extended the compound Poisson distribution to model clustered or correlated survival data with cured proportions, but they only considered the fully parametric survival model with independent and identically distributed frailty due to the complication of the resulting model. We use a slightly different representation because we find the current formulation mathematically more convenient. Moreover, the frailty U is extended to accommodate the effects of the covariates on the incidence part through the dependence of the mean of the Poisson distribution from the covariates. This extension allows the cure probability to be covariate-dependent which is more natural in practice. The mean of the frailty can easily be shown to be $\mathrm{E}\left(U_i \mid \boldsymbol{x}_i^{(0)}\right) = \sigma \eta_i$. As the frailty is multiplied to the arbitrary baseline hazard function, we need to fix the value of σ beforehand to avoid the identifiability problem due to overparameterization. Without loss of generality, we set $\sigma = 1$ so that the logarithm of the mean of the frailty can be expressed as a simple linear function of the covariates

$\boldsymbol{x}_i^{(0)}$. The population survival function is given by

$$\begin{aligned} S\left(t \mid \eta_i, \boldsymbol{x}_i^{(1)}\right) &= e^{-\frac{\eta_i}{2}} + \int_0^\infty \sum_{k=1}^{\infty} \frac{e^{-\frac{\eta_i}{2}} \left(\frac{\eta_i}{2}\right)^k}{k!} \times \frac{u_i^{k-1} e^{-\frac{u_i}{2}}}{2^k \Gamma(k)} \exp\left\{-u_i \Lambda_0(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_i^{(1)}\right)\right\} \mathrm{d}u_i \\ &= e^{-\frac{\eta_i}{2}} + e^{-\frac{\eta_i}{2}} \sum_{k=1}^{\infty} \left(\frac{\eta_i/2}{1 + 2\Lambda_0(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_i^{(1)}\right)}\right)^k \frac{1}{k!} \\ &= \exp\left[-\frac{\eta_i}{2}\left\{1 - \frac{1}{1 + 2\Lambda_0(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_i^{(1)}\right)}\right\}\right], \end{aligned} \tag{3}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$ is the cumulative baseline hazard function. The cure probability is obtained by taking the limit $t \to \infty$ in (3) leading to $\lim_{t\to\infty} S\left(t \mid \eta_i, \boldsymbol{x}_i^{(1)}\right) = \mathrm{P}(K_i = 0) = \exp(-\eta_i/2)$. A parametric assumption makes the estimation simple and straightforward, but we will consider an arbitrary baseline hazard function to provide a more flexible model to avoid a misspecification of the parametric baseline hazard function.

In cancer studies, we can conceive the covariate-dependent frailty U as a summary index of the health status or condition of the individuals that takes into account the information of their personal characteristics like gender, age, smoking status and the type of treatment received. Intuitively, as a measurement of the overall health status of a subject, U should be a continuous variable that a smaller value of U represents better health condition leading to a much smaller risk of the event. Therefore, for some curable diseases, it is conjectured that a relatively small value of U is an indication of a cure. Individuals with U less than a threshold value, say τ, become more homogeneous. They are highly protected from developing the event and can be classified as being cured. Without loss of generality, we may set $\tau = 0$ for mathematical convenience and replace all negative U's by 0. These individuals are said to be cured or at extremely low risk of the event. As a result, U has a distribution with a point mass at 0 and is continuous in the positive region.

The proposed frailty U affects both the cure probability and the heterogeneity among the uncured group. It is more natural than using two separate random effects, as whether an individual is cured is likely to be influenced by some underlying health conditions, which would also affect his time to event. It can be assumed that low cure probability and short time to event are positively correlated induced by U. This was also noted by Kim and Jhun (2008), who used a shared random effect V in the logistic regression model which account for the cured probability, and in the Cox regression in the uncured group:

$$\mathrm{P}\left(B_i = 1 \mid \boldsymbol{x}_i^{(0)}, v_i\right) = \frac{\exp\left(\boldsymbol{\theta}' \boldsymbol{x}_i^{(0)} + v_i\right)}{1 + \exp\left(\boldsymbol{\theta}' \boldsymbol{x}_i^{(0)} + v_i\right)}$$

and

$$\lambda\left(t \mid B_i = 1, \boldsymbol{x}_i^{(1)}, v_i\right) = \lambda_0(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_i^{(1)} + v_i\right),$$

where B_i is a binary random variable with $B_i = 0$ representing a cure, and $\lambda\left(t \mid B_i = 1, \boldsymbol{x}_i^{(1)}, v_i\right)$ denotes the conditional hazard function among the uncured individuals. However, it is unrealistic that the scale of the effects of v_i on the logit of the cure probability and the hazard function are restricted to be the same. Nevertheless, their method can be extended to a multivariate setting easily, say by using a random effect $u_{ij} = v_{i0} + v_{ij}$ for the jth subject of cluster i. Our proposed model can also be extended easily to accommodate multivariate data with a similar idea, and the extension are briefly mentioned in Section 7.

4 Estimation Via Multiple Imputation

4.1 Right-censored Data

To make the presentation clear and simple, we start the discussion with right-censored data, and the estimation method is extended to the more general interval-censored case in the next subsection. The estimation is complicated by the presence of the frailty U, as the partial likelihood function cannot be evaluated. However, the estimation would be simple if the latent variables K and U are observable. Suppose K and U can also be observed, and let $\boldsymbol{D}$ be the complete data with $\boldsymbol{D} = (y_i, \delta_i, \boldsymbol{x}_i, k_i, u_i;\ i = 1, \ldots, n)$. Estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is straightforward through maximizing the complete data partial likelihood L_C given by

$$\begin{aligned} L_C(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{D}) &= \prod_{i=1}^{n} \frac{\exp\left\{-\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x}_i^{(0)})}{2}\right\}\left\{\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x}_i^{(0)})}{2}\right\}^{k_i}}{k_i!} \\ &\quad \times \prod_{k_i>0} \frac{u_i^{k_i-1}\exp(-\frac{u_i}{2})}{2^{k_i}\Gamma(k_i)} \times \prod_{i=1}^{n}\left\{\frac{u_i \exp\left(\boldsymbol{\beta}'\boldsymbol{x}_i^{(1)}\right)}{\sum_{m\in R(y_i)} u_m \exp\left(\boldsymbol{\beta}'\boldsymbol{x}_m^{(1)}\right)}\right\}^{\delta_i} \\ &\propto \prod_{i=1}^{n}\exp\left\{-\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x}_i^{(0)})}{2}\right\}\exp(k_i\boldsymbol{\theta}'\boldsymbol{x}_i^{(0)}) \times \prod_{i=1}^{n}\left\{\frac{\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_i^{(1)}\right)}{\sum_{m\in R(y_i)} u_m \exp\left(\boldsymbol{\beta}'\boldsymbol{x}_m^{(1)}\right)}\right\}^{\delta_i} \\ &= L_1(\boldsymbol{\theta}) \times L_2(\boldsymbol{\beta}), \end{aligned}$$

where $R(y_i)$ is the set of individuals at risk just prior to time y_i. Note that $L_2(\boldsymbol{\beta})$ is just the partial likelihood for $\boldsymbol{\beta}$ when the frailties U_i's are observed while $L_1(\boldsymbol{\theta})$ is just the likelihood function for $\boldsymbol{\theta}$ when K_i's are observed. The two likelihood functions are orthogonal and hence $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ can be estimated separately by maximizing $\ell_1(\boldsymbol{\theta}) = \log L_1(\boldsymbol{\theta})$ and $\ell_2(\boldsymbol{\beta}) = \log L_2(\boldsymbol{\beta})$, respectively to give the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ and the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$. Moreover the cumulative baseline hazard function $\Lambda_0(t)$ can be estimated by the usual Nelson-Aalen estimator

$$\hat{\Lambda}_0(t) = \sum_{i:t_i\le t}\left\{\frac{\delta_i}{\sum_{m\in R(y_i)} u_m \exp\left(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_m^{(1)}\right)}\right\}. \tag{4}$$

An estimation method using multiple imputation is proposed next. We adopt the data augmentation technique by Tanner and Wong (1987) with the Asymptotic Normal Data Augmentation (ANDA) introduced by Wei and Tanner (1991). It can essentially be treated as the Monte Carlo implementation of the E-step in the EM algorithm. As the subject-specific frailty U_i and its dummy variable K_i are unknown, the idea is to augment these unobserved observations according to their respective posterior distributions. In the following, we will also adopt the zero-tail constraint suggested by Taylor (1995) by letting U_j be 0 if the jth subject is censored beyond the largest uncensored failure time t^*. The algorithm to the estimation of the regression parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta}', \boldsymbol{\beta}')'$ and the variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$, namely $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$, is summarized as follows:

1. Initialize $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(0)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(0)} = b\boldsymbol{I}$ where $\boldsymbol{I}$ is a $(p_0 + p_1 + 1) \times (p_0 + p_1 + 1)$ identity matrix and b is a non-negative constant, which is usually chosen to be small, say 0.1.

2. Compute $\hat{\Lambda}_0^{(0)}(t)$ based on equation (4) by setting $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(0)}$ and $u_i = u_i^{(0)} = \delta_i$.

3. At the jth step ($j = 1, 2, \ldots$):

(a) generate $\boldsymbol{\alpha}_h$ from $N\left(\hat{\boldsymbol{\alpha}}^{(j-1)}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(j-1)}\right)$ for $h = 1, 2, \ldots, M$;

(b) by setting $\boldsymbol{\alpha} = \boldsymbol{\alpha}_h$ and $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$, generate $\boldsymbol{k}_h = (k_{h1}, \ldots, k_{hn})$ $(h = 1, \ldots, M)$ from the posterior distribution of K_i given by

$$(K_i - \delta_i) \mid (y_i, \boldsymbol{x}_i, \delta_i) \sim \text{Poisson}\left(\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x}_i^{(0)})}{2 + 4\Lambda_0(y_i)\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_i^{(1)}\right)}\right);$$

(c) by setting $\boldsymbol{k} = \boldsymbol{k}_h$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}_h$ and $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$, generate $\boldsymbol{u}_h = (u_{h1}, \ldots, u_{hn})$ $(h = 1, \ldots, M)$ from the conditional posterior distribution of U_i given by

$$U_i \mid (y_i, \delta_i, \boldsymbol{x}_i, k_i) \begin{cases} = 0, & \text{if } k_i = 0; \\ \sim \text{Gamma}\left(k_i + \delta_i, \left\{0.5 + \Lambda_0(y_i)\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_i^{(1)}\right)\right\}^{-1}\right), & \text{if } k_i > 0. \end{cases}$$

Moreover, set $u_{hi} = 0$ if $y_i > t^*$ where t^* is the largest observed time to event;

(d) for $h = 1, \ldots, M$, maximize $\ell(\boldsymbol{\alpha} \mid \boldsymbol{D}_h) = \log L_C(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{D}_h)$ with $\boldsymbol{D}_h = (y_i, \delta_i, \boldsymbol{x}_i, k_{hi}, u_{hi}; i = 1, \ldots, n)$ to obtain the estimates $\hat{\boldsymbol{\theta}}^{(j,h)}$, $\hat{\boldsymbol{\beta}}^{(j,h)}$, and use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(j,h)}$ to obtain $\hat{\Lambda}_0^{(j,h)}(t)$ based on equation (4);

(e) update the estimate by

$$\hat{\boldsymbol{\theta}}^{(j)} = \frac{1}{M}\sum_{h=1}^{M}\hat{\boldsymbol{\theta}}^{(j,h)}, \hat{\boldsymbol{\beta}}^{(j)} = \frac{1}{M}\sum_{h=1}^{M}\hat{\boldsymbol{\beta}}^{(j,h)} \text{ and } \hat{\Lambda}_0^{(j)}(t) = \frac{1}{M}\sum_{h=1}^{M}\hat{\Lambda}_0^{(j,h)}(t);$$

(f) update the estimated variance-covariance matrix by

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(j)} &= \frac{1}{M}\sum_{h=1}^{M}\left[-\frac{\partial^2}{\partial\boldsymbol{\alpha}'\partial\boldsymbol{\alpha}}\left\{\ell_1(\boldsymbol{\theta}) + \ell_2(\boldsymbol{\beta})\right\}\right]^{-1}_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}^{(j,h)}, \boldsymbol{D}=\boldsymbol{D}^{(j,h)}} \\ &\quad + \left(1 + \frac{1}{M}\right)\sum_{h=1}^{M}\frac{\left(\hat{\boldsymbol{\alpha}}^{(j,h)} - \hat{\boldsymbol{\alpha}}^{(j)}\right)\left(\hat{\boldsymbol{\alpha}}^{(j,h)} - \hat{\boldsymbol{\alpha}}^{(j)}\right)'}{M-1}. \end{aligned} \tag{5}$$

4. Repeat step 3 until convergence is achieved with final estimates $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(t)$.

The second term in (5) is to account for the "between-imputation" variance (Tanner and Wong, 1987; Rubin, 1987; Schenker and Welsh, 1988). The inflation factor $(1 + 1/M)$ in the between imputation variance is to account for the fact that only a finite number of imputations are drawn. We will further discuss the choice of M in the next section.

4.2 Interval-censored Data

For interval-censored data, the observed time y_i is only known to be within the interval $(l_i, r_i]$. Denote the complete data by $\boldsymbol{D} = (l_i, r_i, y_i, \delta_i, \boldsymbol{x}_i, k_i, u_i; i = 1, \ldots, n)$. It is natural to extend the proposed multiple imputation algorithm for right censored data by imputing the y_i's for the interval-censored data. This can be done by first setting the initial value of y_i to be

$$y_{hi} = (l_i + r_i)/2 \text{ if } \delta_i = 1, \text{ for } i = 1, \ldots, n; h = 1, \ldots, M$$

in step (1) of the algorithm, and add an additional step between 3(a) and 3(b):

3(a)*: generate $\boldsymbol{y}_h = (y_{h1}, \ldots, y_{hn})$:

1. for individuals with $\delta_i = 0$, $y_{hi} = l_i$ for $h = 1, \ldots, M$.
2. for individuals with $\delta_i = 1$, by setting $\boldsymbol{\alpha} = \boldsymbol{\alpha}_h$ and $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$ for $h = 1, \ldots, M$, generate y_{hi} from the conditional distribution

$$\mathrm{P}\left(Y > y \mid l_i, r_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i\right) = \frac{S\left(y \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i\right) - S\left(r_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i\right)}{S\left(l_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i\right) - S\left(r_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i\right)},$$

where the survival function $S\left(y \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i\right)$ is given by (3).

We will replace y_i by its realizations y_{hi} in steps 3(b) to 4 of the algorithm in Section 4.1. Since y_i varies at every iteration, we modify the zero-tail constraint by setting $U_j = 0$ if the jth subject is censored beyond r^* where $r^* = \max_j \{r_j \delta_j\}$.

It should be noted that the estimated survival function $\hat{S}\left(y \mid \boldsymbol{x}_i\right)$ is a step function due to the nature of the estimated cumulative baseline hazard function $\hat{\Lambda}_0\left(y\right)$. When we draw y_i's from the estimated survival function, it is possible to have lots of ties, which is undesirable in the maximization of the Cox partial likelihood. To remedy the problem, we propose to replace the cumulative baseline hazard function by a monotonic increasing function by joining the midpoints of adjacent steps of $\hat{\Lambda}_0\left(y\right)$ by straight lines. Let q be the number of observations with $\delta = 1$ in the data and let $\boldsymbol{z} = \left(z_{(1)}, \ldots, z_{(Mq)}\right)$ be the set of points of discontinuity on the cumulative baseline hazard function $\hat{\Lambda}_0\left(y\right)$ with $0 = z_{(0)} < z_{(1)} < \cdots < z_{(Mq)}$. There are a total of Mq points of discontinuity since, in each iteration, each of the q observations with $\delta = 1$ contributes to a jump on the estimated cumulative baseline hazard function, resulting in Mq points of discontinuity. We further let $z'_{(i)} = \left(z_{(i)} + z_{(i+1)}\right)/2$ for $i = 0, \ldots, Mq - 1$ and $z'_{(Mq)} = z_{(Mq)}$. The survival function can then be evaluated using the estimated cumulative baseline hazard function

$$\tilde{\Lambda}_0\left(y\right) = \frac{\hat{\Lambda}_0\left(z'_{(j-1)}\right)\left(z'_{(j)} - y\right) + \hat{\Lambda}_0\left(z'_{(j)}\right)\left(y - z'_{(j-1)}\right)}{z'_{(j)} - z'_{(j-1)}} \tag{6}$$

at every point y with j chosen such that $z'_{(j-1)} \leq y < z'_{(j)}$ and $\tilde{\Lambda}_0\left(y\right) = \hat{\Lambda}_0\left(y\right)$ for $y \geq z_{(Mq)}$. This modification would not induce much bias to the estimation as in the Cox partial likelihood, only the order of the times to event are used, but not the exact times.

Using this multiple imputation algorithm, the jump points on the hazard function are updated at every iteration and the values are not forced to take from a set of finite values. This is in contrast with some methods proposed in the literature where the set of jump points is a fixed subset of the boundary points of the intervals (Finkelstein, 1986; Liu and Shen, 2009; Ma, 2010). Our approach is more natural as we avoid arbitrarily setting the jump points at the initialization step which is hard to be justified. Also, it works well with different distributions of the intervals, even when the intervals are wide and irregular. We note that when the imputation size is large, the number of steps on the cumulative baseline hazard function will be large and ties in the imputed y_{hi} are improbable. The proposed method of eliminating the jumps is recommended as we can guarantee that there is no tie in the imputed values even for large sample sizes and imputation sizes to facilitate the evaluation of the partial likelihood function.

5 Simulation Study

5.1 Right-censored Data

Simulation studies were carried out to assess the performance of the proposed method. We assume a baseline hazard function $\lambda_0(t) = t$ and a censoring variable $C = \min(5, 15 \times A)$ where A is a uniform random number. We mimic a clinical trial with $p = 2$ where X_1 is the treatment indicator that takes

on the value 0 or 1, each with probability 0.5 and X_2 is a continuous explanatory variable generated independently from the standard normal distribution. Throughout the simulations, we assume $p_0 = p_1 = 2$ with $\boldsymbol{X}_i^{(0)} = (1, X_{i1}, X_{i2})'$ and $\boldsymbol{X}_i^{(1)} = (X_{i1}, X_{i2})'$. We set $\boldsymbol{\theta} = (-1, 1, 0)'$ and $\boldsymbol{\beta} = (0, 0.5)'$. With $\theta_1 = 1$ and $\beta_1 = 0$, the treated patients ($X_1 = 0$) have better health condition, leading to higher cure rate than the untreated patients ($X_1 = 1$) but the treatment has no direct effect on the hazard rate among the uncured patients. On the other hand, X_2 has no effect on the overall health condition of the patients and hence the cure probability ($\theta_2 = 0$), but patients with larger value of X_2 have higher risk among the uncured patients ($\beta_2 = 0.5$).

Two sample sizes $n = 200$ and $n = 500$ are considered. 500 data sets are generated for each case. To investigate the effect of the imputation size M, we repeat the simulations using imputation sizes $M = 10, 50, 100$. An important issue here is how to determine the convergence of the estimates as it is difficult to have some objective convergence criteria. We monitored the estimates under various setups and found that the estimates generally became very stable in less than 30 iterations. To be conservative towards early termination, we propose to take the estimate at the 100th iteration as the final estimate for each data set throughout this and the next subsections. The results are shown in Table 1.

The mean, the empirical standard deviation of the 500 estimates and the average of the estimated standard errors based on (5) for each parameter are reported. Moreover, for each data set, we compute the asymptotic 95% and 99% confidence intervals for each regression parameter, and their corresponding empirical coverages are also given in Table 1. The performance of the estimation method is highly satisfactory. Irrespective of the sample size and imputation size, the averages of the estimates are all very close to their corresponding true values. Moreover, the standard error estimates closely resemble their corresponding empirical standard deviations, and their empirical coverages are reasonably close to the nominal levels. For $M = 10$, we noticed that the empirical standard deviations and the average standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ are generally larger than those cases with a moderate M value. This indicates that under such a complicated missing mechanism of K and U, a larger value of M relative to $M = 10$ will lead to significant improvement in producing fewer extreme estimates and smaller standard errors. Nevertheless, an imputation size greater than 100 shows no further improvement in the standard error estimates and hence $M = 50$ is generally adequate. Other investigators suggested that such a large value of M may not be necessary (Glynn *et al*., 1993; Pan, 2000; Lam *et al*., 2005) but this is not true in the current situation where the amount of missing information is huge. Here, we can observe that though the imputation size does not have an effect on the accuracy of the estimates, it does affect the standard error estimates. Larger imputation size would lead to a better performance in the standard error estimates, as demonstrated by increasing the imputation size from 10 to 50, yet not much improvement is gained through further increasing it from 50 to 100. We postulate that this phenomenon is common in other estimation algorithms with augmentation when the missing information is huge and the missing mechanism is complex.

5.2 Interval-censored Data

Simulations are also conducted for the case with interval-censored data to assess the performance of the proposed estimation method. The generation of the covariates $\boldsymbol{x}$, the time to event T, and the right-censoring time C is identical to the case of right-censored data. We assume that the individuals are followed periodically through check-ups, which are not regular and are different for different individuals. For each individual, the first check-up time is at $t = 0$ and the duration between each check-up time is a random variable from a Uniform $(0.1, 0.5)$ distribution. As a result, an observation would be right-censored at time C if the time to event T is larger than C, and would be interval-censored otherwise. For an interval-censored observation, the endpoints of the interval would be the check-up times immediately before and after the event. In the current setup, among the uncured patients, the median time to event is about 0.725, and about 70% have an event time < 1.0. Hence, the widths of the intervals are not too wide nor too narrow so that the performance of the estimation method will neither be penalized nor benefited from this interval-censoring mechanism. For comparison purpose, we consider imputation sizes $M = 10, 50, 100$

and sample sizes $n = 200, 500$ as in the last subsection. 500 data sets for each setting are generated. The results are presented in Table 2.

The performance of the estimation method with interval-censored data is highly satisfactory and is better than our expectation. We observe that the estimates are all very close to their true values irrespective of the imputation size and sample size. Again, with an imputation size $M = 50$, the distribution of the estimator is obviously better approximated by its asymptotic normal distribution relative to the case with $M = 10$ in terms of the empirical coverages. Moreover, the estimates and estimated standard errors tend to be more stable, but further increasing the imputation size to $M = 100$ only provides marginal improvement. Therefore, a moderate imputation size of $M = 50$ is generally adequate in the current setting for right-censored and interval-censored data. In view of the inexpensive computation cost, we can always increase the value of M to a reasonably large number for a single data analysis attempt, say $M = 1000$.

Comparing the results with that of the right-censored case, we observe that the empirical standard deviations and standard errors of $\hat{\boldsymbol{\theta}}$ are quite close. It is reasonable as the interval-censoring mechanism should not have much effects on the estimation of $\boldsymbol{\theta}$. When comparing the empirical standard deviations and estimated standard errors of $\hat{\boldsymbol{\beta}}$, those in the interval-censored case are, however only marginally, larger than those in the right-censored case. This is expected because we have less information in the interval-censored case, but we did not expect that the difference is so small.

6 Application

6.1 Right-censored Data

We apply the model and estimation method to the melanoma incidence data discussed in Section 3. Ibrahim *et al*. (2001a) modeled the melanoma data using the semiparametric promotion time (SPT) model via a Bayesian framework. In their study, a bounded cumulative hazard model was considered and an extra parameter κ was used to control the degree of parametricity in the right tail of the survival curve. In order to compare with the proposed model and method, we consider the same set of covariates, namely treatment ($X_1 = 0$ for OBS and $X_1 = 1$ for IFN), age (X_2) and gender ($X_3 = 0$ for male and $X_3 = 1$ for female) in the following analyses. Their estimates, by setting $\kappa = 0.95$, are reproduced in Table 3.

We also fit the data to the cure model proposed by Price and Manatunga (2001) as specified in (2) with U being a binary variable that is modeled by a logistic regression model with $\text{logit}\left[\text{P}\left(U = 1 \mid \boldsymbol{x}^{(0)}\right)\right] = \boldsymbol{\theta}'\boldsymbol{x}^{(0)}$, $\lambda_0(t) = \exp(\beta_0)$ is a constant, and V follows a Gamma distribution with mean 1 and variance ω^{-1} that represents the non-covariate dependent heterogeneity among the uncured patients only. Here we simply let $\boldsymbol{X}^{(0)} = (1, X_1, X_2, X_3)$ and $\boldsymbol{X}^{(1)} = (X_1, X_2, X_3)$. The maximum likelihood estimates for $\boldsymbol{\alpha}$ and their estimated standard errors are listed in Table 3. We fit the proposed model to the data with $\boldsymbol{X}^{(0)} = (1, X_1, X_2, X_3)$ To compare the results with that based on model (2) and the SPT model, we consider model A with $\boldsymbol{X}^{(1)} = (X_1, X_2, X_3)$ and model B with $\boldsymbol{X}^{(1)} = 0$, respectively. To facilitate determination of convergence, the estimation algorithm starts with an imputation size $M = 50$ and the progress of the algorithm is monitored. As suggested by Tanner and Wong (1987) when the processes appear to be stationary, a much larger imputation size ($M = 500$) is adopted to force the fluctuations of the estimates to be within a very small range. We take the estimates at the 100-th iteration since switching to $M = 500$ as the final estimates for both models A and B. They are summarized in Table 3.

Using a 0.05 level of significance, the analyses show that the latency variable does not depend on any of the X_j's, and age and sex are not significantly associated with the incidence variables in all 4 models. The SPT model reveals that the treatment IFN significantly increases the cure probability when compare to the OBS group, but not in other models. In fact, model B and the SPT model share some common characteristics that covariates are not included in the analysis of the latency variable. We use the Nelson-Aalen type estimator to estimate the baseline cumulative hazard function in model B, taking into account of the frailty U, while the SPT-model assumed a piecewise constant hazard function for the latency variable which are very similar in nature. However, the difference mainly comes from the covariate-dependent

frailty, the model for the incidence variable. Taking the melanoma data as an example, the proposed model assumed that the frailty is the sum of the relative aggressiveness of each of the k_i latent metastasis-competent tumor cells that each cell may have different level of relative aggressiveness biologically, but the SPT-model assumes that each cell has the same level of aggressiveness with the same contribution. Moreover, different choice of κ may also lead to different results. Therefore, the proposed model may be a bit more general and flexible over the SPT-model. In addition, the proposed model allows the latency variable to depend on some explanatory variables, which is more flexible in practice.

6.2 Interval-censored Data

The proposed model is also fitted to the interval-censored breast retraction data discussed in Section 3. Forty-six early breast cancer patients received only radiotherapy ($X = 0$) and 48 patients received both radiotherapy and chemotherapy ($X = 1$). The treatment indicator X is the only available covariate. We include X in both the frailty part and the Cox regression part. Similar to the first application, we start with a moderate imputation size $M = 50$ and $M = 1000$ is adopted when the processes appear to be stationary. We take the estimates at the 100-th iteration since switching to $M = 5000$ as the final estimates. The results are shown in Table 4 under Method 1.

The treatment effect is significant in both the Poisson regression in the frailty part and the Cox regression for the uncured patients. A positive $\hat{\theta}_1$ indicates that treatment of radiotherapy with adjuvant chemotherapy is associated with a larger frailty leading to a deterioration in the health conditions of the patients, and thus a smaller cure rate relative to the radiotherapy-only group (3.39% vs 45.22%). This is in line with our observations where more observations were right-censored in the radiotherapy-only group. Based on a semiparametric model for interval-censored data without considering a cured proportion in the population, Finkelstein and Wolfe (1985) concluded that the group receiving chemotherapy in addition to radiotherapy experienced a significantly earlier cosmetic deterioration measured by the appearance of breast retraction. Finkelstein (1986) and Zhang *et al*. (2010) came up with a similar conclusion based on the standard Cox PH model for interval-censored data that the treatment of radiotherapy with adjuvant chemotherapy significantly increases the risk of the breast retraction.

In addition, our approach considers a cured proportion in the population and a Cox-type PH model among the uncured patients. The estimate $\hat{\beta}_1$ is significantly different from 0 which implies that, conditioned on $U > 0$, the treatment of radiotherapy with adjuvant chemotherapy reduces the risk of breast retraction. Our analysis result suggests that patients receiving treatment of radiotherapy with adjuvant chemotherapy have a significantly higher probability of breast retraction, but also a lower hazard rate of breast retraction among those patients who will eventually experience the events with similar health conditions.

The above results highlight the importance of the choice of an appropriate model and to identify whether the treatment is effective in curing the disease or delaying/expediting the time to onset of the disease or both. It may be more reasonable to adopt a cure model when the event time of a substantial proportion of the patients are right-censored at some reasonably large values.

The estimated survival functions obtained by using the estimated baseline hazard function (6), which look like step functions, are plotted in Figure 2. Upon scrutiny, they are not "steps" but very steep slopes as the imputed failure times, though being distinct, would cluster around several time points after several iterations. These "steps", once formed, cannot be smoothed out easily. To obtain a *smooth* estimate for the survival functions, we propose an alternative imputation method for the interval-censored failure times. To impute $Y_{i1}, \ldots, Y_{iM}$ from the conditional distribution of $Y_i \mid (l_i, r_i, \delta_i, \boldsymbol{x}_i)$, we evaluate the survival function based on a modified cumulative baseline hazard function:

$$\grave{\Lambda}_0(t) = \frac{\hat{\Lambda}_0(l_i)(r_i - t) + \hat{\Lambda}_0(r_i)(t - l_i)}{r_i - l_i}$$

for $l_i \leq t < r_i$. In other words, when imputing each Y_i, we replace the cumulative baseline hazard function by a straight line with the values of the two ends being $\hat{\Lambda}_0(l_i)$ and $\hat{\Lambda}_0(r_i)$. The imputed Y's would be more spread out that leads to smoother estimated cumulative baseline hazard function and hence the estimated survival functions.

The breast cosmesis data were analyzed using the alternative imputation method. The regression parameters estimates are reported in Table 4 under Method 2. The estimates between the two imputation methods are consistent. The *smoothed* estimated survival functions (they are indeed piecewise linear functions) are also plotted in Figure 2 with those based on the original imputation method. The two sets of estimated survival functions are also very consistent that the latter estimated survival functions are just like the smoothed versions of the original ones. Therefore, this alternative imputation method would not affect the estimates much, as we are only interested in the ranks of the observations when evaluating the partial likelihood but not the actual values on the baseline hazard function. The latter imputation method has the advantage that the computation is faster. However, when the sample size is small, its performance is inferior to the original method that the empirical standard deviations and the standard error estimates are generally larger from some simulation results not reported here. Nevertheless, the difference is negligible when the sample size is large, say $n = 500$.

7 Discussion

A frailty-Cox PH model is proposed in this paper, which accommodates both the curing status and the heterogeneity among the uncured individuals with a single frailty term. The frailty is assumed to follow a compound-Poisson distribution. The model has sound biological interpretation that the heterogeneity comes from differences in the underlying health conditions among the individuals, and is quantified by the frailty. The cured individuals are supposed to be extremely healthy and have zero frailty while a large frailty is associated with bad health condition that the subject is at a higher risk of the event. As a quantification of the underlying health status, though not being directly observable, it is reasonable that the frailty may depend on some covariates. Therefore, in the proposed model, the mean of the Poisson distribution in the compound-Poisson distributed frailty is determined by a vector of covariates. One advantage of the proposed model is that the covariates can affect the time to event in two ways, namely affecting the incidence and the latency. A particular covariate can deteriorate the health of the individuals and in turn shorten the time to event, which is usually the case for some individual characteristics such as age and gender. On the other hand, a covariate can directly decrease the hazard rate of an individual, which is usually the case for some experimentally related covariates such as treatment indicator. The proposed model shares some similarities with the promotion time model of Chen *et al*. (1999), but the proposed model is more flexible as it allows the covariates to have different effects on the incidence and latency variables.

The calculation of the maximum likelihood estimates of the regression parameters by evaluating the marginal survival function (3) is very complicated as it involves the nonparametric baseline hazard function. Fortunately, the multiple imputation method provides a conceptually straightforward alternative which is easy to implement. It takes advantage of the simple partial likelihood function when the unobserved data K and U are available in the right-censored case. Moreover, an extra step can be incorporated in the estimation algorithm to accommodate interval-censored data. From the simulations, we observe that, through careful selection of the imputation size, the performance of the ANDA estimation is very satisfactory in terms of its nearly unbiased property, accurate standard error estimation and reasonable coverages based on the asymptotic approximation. Besides the ANDA algorithm, we have also conducted simulations using the Poor Man's Data Augmentation (PMDA) algorithm (Tanner and Wong, 1987), but the results are not reported in this paper because of its poor performance. The PMDA algorithm is very similar to the ANDA algorithm but is computationally simpler in the data augmentation step. Similar to the ANDA algorithm, the PMDA algorithm also gives nearly unbiased estimates, but the standard error of

the estimates are seriously underestimated, probably because the amount of missing information is huge in this case. This underestimation cannot be remedied by increasing the imputation size. This failure to correctly estimate the true variability if the degree of missingness is severe is consistent with the findings of previous authors (Wei and Tanner, 1991; Pan, 2000). Therefore, in spite of its computational burden, the ANDA algorithm is recommended instead of the PMDA algorithm.

As commented by Li and Ma (2010), a satisfactory extension of a univariate semiparametric Cox-type cure model to a multivariate setup to explain various types of associations is not yet available for interval-censored data. The current model and approach may shed some lights to this problem. The proposed model can be extended naturally to a multivariate setup to accommodate clustered and longitudinal survival data with cured proportions. It can be done by incorporating some cluster-specific random effects for clustered survival data, and structured random effects, like dynamic random effects that progress with time (Fong *et al*., 2001) for longitudinal survival data with a cured proportion. The multiple imputation method can be extended to accommodate both right-censored and interval-censored data in general multivariate situations. Work on the multivariate case is in progress, and will be reported in another manuscript.

Acknowledgements The authors thank the editor, an Associate editor and two anonymous reviewers for their constructive comments that substantially improved the presentation and the content of the manuscript. The work was supported by Committee on Research and Conference Grants of The University of Hong Kong..

Conflict of Interest

The authors have declared no conflict of interest.

References

Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability* **2**, 951–972.

Beadle, G. F., Harris, J. R., Come, S., Henderson, C., Silver, B. and Hellman, S. (1984). The effect of adjuvant chemotherapy on the cosmetic results after radiation treatment for early stage breast cancer: A preliminary analysis. *International Journal of Radiation Oncology, Biology and Physics* **10**, 2131–2137.

Beadle, G. F., Harris, J. R., Silver, B., Botnick, L. and Hellman, S. (1984). Cosmetic results following primary radiation therapy for early breast cancer. *Cancer* **54**, 2911–2918.

Chen, M. H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**, 909–919.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845–854.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933–945.

Fong, D. Y. T., Lam, K. F., Lawless, J. F. and Lee, Y. W. (2001). Dynamic random effects models for times between repeated events. *Lifetime Data Analysis* **7**, 345–362.

Glynn, R. J., Laird, N. M. and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association* **88**, 984–993.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* 2001; **57**, 383–388.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer: New York.

Kim, Y. J. and Jhun, M. (2008). Cure rate model with interval censored data. *Statistics in Medicine* **27**, 3–14.

Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M. and Blum, R. H. (2000). The role of high- and low-does interferon Alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S9111/C9190. *Journal of Clinical Oncology* **18**, 2444–2458.

Lam, K. F., Fong, D. Y. and Tang, O. Y. (2005). Estimating the proportion of cured patients in a censored sample. *Statistics in Medicine* **24**, 1865–1879.

Li, J. and Ma, S. (2010). Interval-censored data with repeated measurements and a cured subgroup. *Journal of the Royal Statistical Society, Series C* **59**, 693–705.

Liu, H. and Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association* **104**, 1168–1178.

Ma, S. (2009). Cure model with current status data. *Statistica Sinica* **19**, 233–249.

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

Moger, T. A. and Aalen, O. O. (2005). A distribution for multivariate frailty based on the compound Poisson distribution with random scale. *Lifetime Data Analysis* **11**, 41–59.

Moger, T. A. and Aalen, O. O. (2008). Regression models for infant mortality data in Norwegian siblings, using a compound Poisson frailty distribution with random scale. *Biostatistics* **9**, 577–591.

Moger, T. A., Aalen, O. O., Heimdal, K. and Gjessing, H. K. (2004). Analysis of testicular cancer data by means of a frailty model with familial dependence. *Statistics in Medicine* **23**, 617–632.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**, 199–203.

Peng, Y. and Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56**, 237–243.

Price, D. L. and Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine* **20**, 1515–1527.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons: New York.

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16**, 1550–1566.

Siegel, A. (1979). The noncentral chi-squared distribution with zero degrees of freedom and testing for uniformity. *Biometrika* **66**, 381–386.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56**, 227–236.

Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* **51**, 899–907.

Wei, G. C. G. and Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics* **47**,1297–1309.

Yin, G. and Ibrahim, J. G. (2005). Cure rate models: A unified approach. *The Canadian Journal of Statistics* **33**, 559–570.

Zhang, Y., Hua, L. and Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**, 338–354.

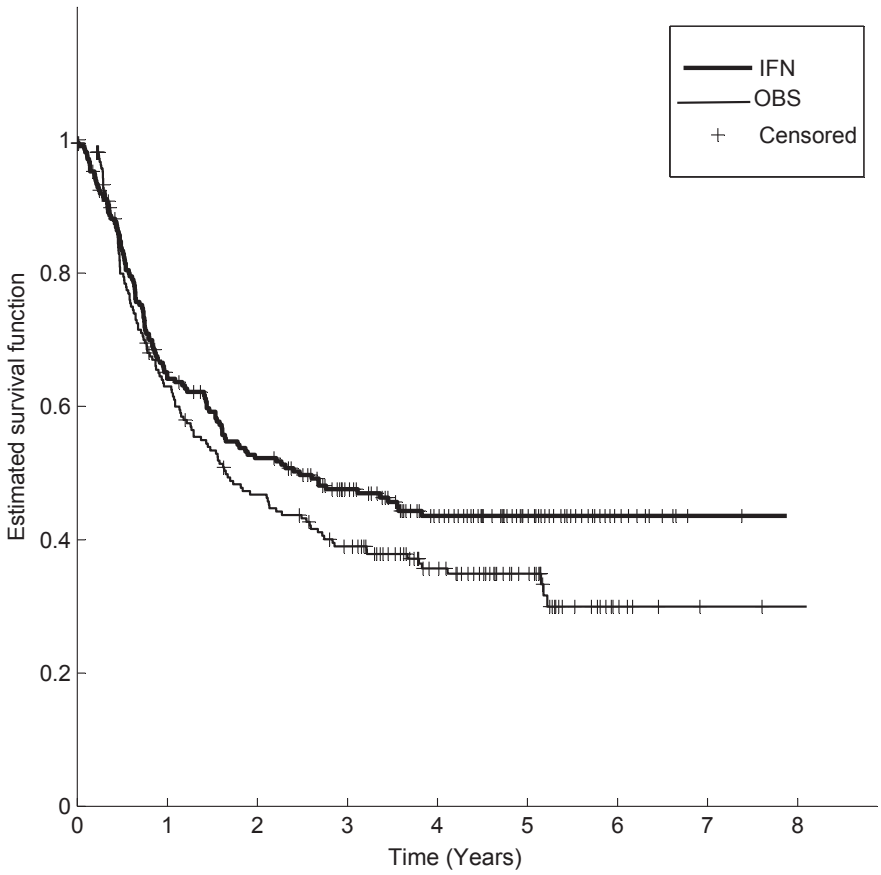

Figure 1 Kaplan Meier estimates for the melanoma data

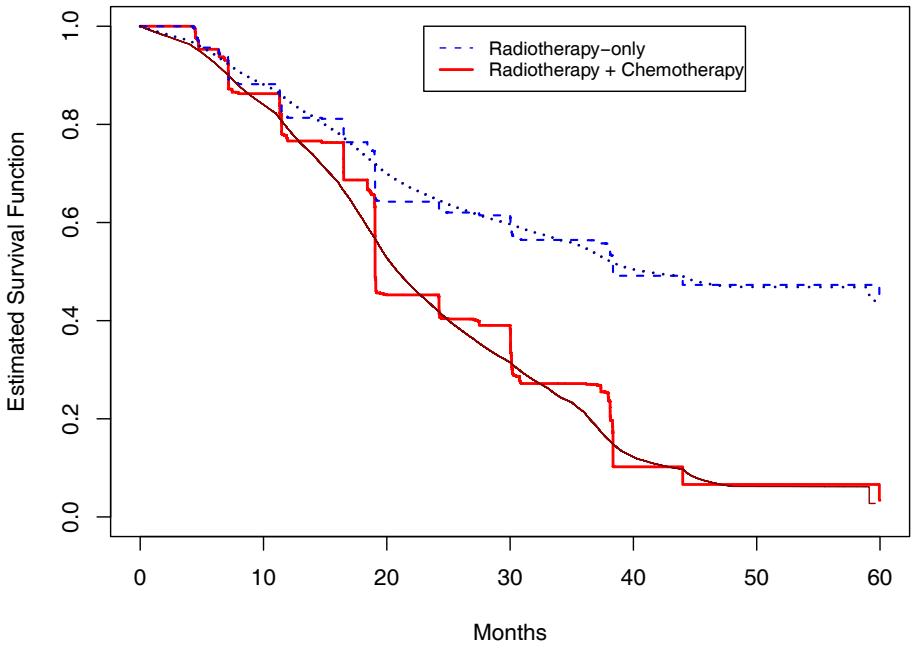

Figure 2 Estimated survival functions for the breast cosmesis data

Table 1 Simulation results for right-censored data

		θ_0	θ_1	θ_2	β_1	β_2
Sample Size $n = 200$						
	True Value	-1.0000	1.0000	0.0000	0.0000	0.5000
$M = 10$	Estimate	-1.1236	1.0065	0.0251	0.1496	0.5596
	Empirical SD	0.2660	0.3074	0.1423	0.7813	0.3626
	Average SE	0.2683	0.3260	0.1484	0.6820	0.3389
	Empirical Coverage (95%)	0.960	0.970	0.958	0.914	0.940
	Empirical Coverage (99%)	0.996	0.998	0.992	0.988	0.982
$M = 50$	Estimate	-1.1134	1.0014	0.0308	0.1381	0.5426
	Empirical SD	0.2819	0.3260	0.1364	0.6814	0.3305
	Average SE	0.2663	0.3212	0.1454	0.6505	0.3151
	Empirical Coverage (95%)	0.948	0.946	0.966	0.956	0.962
	Empirical Coverage (99%)	0.996	0.994	0.994	0.994	0.992
$M = 100$	Estimate	-1.1093	1.0070	0.0395	0.0714	0.5020
	Empirical SD	0.2500	0.2949	0.1364	0.6693	0.3087
	Average SE	0.2476	0.2988	0.1355	0.5928	0.2877
	Empirical Coverage (95%)	0.950	0.964	0.946	0.942	0.944
	Empirical Coverage (99%)	0.988	0.988	0.990	0.984	0.988
Sample Size $n = 500$						
		θ_0	θ_1	θ_2	β_1	β_2
	True Value	-1.0000	1.0000	0.0000	0.0000	0.5000
$M = 10$	Estimate	-1.0732	1.0016	0.0247	0.0570	0.4928
	Empirical SD	0.1695	0.2007	0.0929	0.3913	0.1865
	Average SE	0.1663	0.2016	0.0924	0.3816	0.1836
	Empirical Coverage (95%)	0.930	0.948	0.938	0.936	0.948
	Empirical Coverage (99%)	0.990	0.990	0.980	0.986	0.980
$M = 50$	Estimate	-1.0754	1.0078	0.0360	0.0077	0.4775
	Empirical SD	0.1683	0.1984	0.0874	0.3506	0.1678
	Average SE	0.1652	0.1999	0.0911	0.3685	0.1767
	Empirical Coverage (95%)	0.928	0.956	0.944	0.962	0.956
	Empirical Coverage (99%)	0.992	0.994	0.988	0.992	0.998
$M = 100$	Estimate	-1.0792	1.0061	0.0351	0.0253	0.4912
	Empirical SD	0.1760	0.2053	0.0846	0.3617	0.1775
	Average SE	0.1653	0.2002	0.0909	0.3681	0.1775
	Empirical Coverage (95%)	0.920	0.946	0.950	0.948	0.956
	Empirical Coverage (99%)	0.986	0.990	0.986	0.988	0.984

Table 2 Simulation results for interval-censored data

Sample Size $n = 200$						
		θ_0	θ_1	θ_2	β_1	β_2
	True Value	-1.0000	1.0000	0.0000	0.0000	0.5000
$M = 10$	Estimate	-1.0601	1.0085	0.0052	0.1078	0.5595
	Empirical SD	0.2833	0.3261	0.1418	0.7051	0.3450
	Average SE	0.2663	0.3275	0.1499	0.6648	0.3201
	Empirical Coverage (95%)	0.954	0.948	0.968	0.938	0.930
	Empirical Coverage (99%)	0.990	0.996	0.992	0.978	0.978
$M = 50$	Estimate	-1.0750	1.0279	0.0033	0.0490	0.5828
	Empirical SD	0.2700	0.3177	0.1521	0.6436	0.3343
	Average SE	0.2669	0.3239	0.1481	0.6304	0.3064
	Empirical Coverage (95%)	0.958	0.954	0.954	0.952	0.940
	Empirical Coverage (99%)	0.994	0.986	0.990	0.988	0.984
$M = 100$	Estimate	-1.0794	1.0211	0.0067	0.0081	0.5629
	Empirical SD	0.2816	0.3161	0.1479	0.6651	0.3013
	Average SE	0.2670	0.3242	0.1473	0.6251	0.3067
	Empirical Coverage (95%)	0.950	0.974	0.954	0.954	0.960
	Empirical Coverage (99%)	0.994	0.998	0.990	0.982	0.992
Sample Size $n = 500$						
		θ_0	θ_1	θ_2	β_1	β_2
	True Value	-1.0000	1.0000	0.0000	0.0000	0.5000
$M = 10$	Estimate	-1.0434	0.9967	0.0126	0.0510	0.5321
	Empirical SD	0.1883	0.2180	0.0969	0.4355	0.1980
	Average SE	0.1661	0.2050	0.0947	0.3915	0.1951
	Empirical Coverage (95%)	0.932	0.938	0.934	0.902	0.928
	Empirical Coverage (99%)	0.970	0.984	0.988	0.970	0.980
$M = 50$	Estimate	-1.0503	1.0184	0.0107	0.0368	0.5142
	Empirical SD	0.1672	0.1996	0.0958	0.3739	0.1862
	Average SE	0.1657	0.2015	0.0926	0.3790	0.1820
	Empirical Coverage (95%)	0.960	0.966	0.950	0.954	0.952
	Empirical Coverage (99%)	0.994	0.996	0.990	0.990	0.988
$M = 100$	Estimate	-1.0643	1.0309	0.0152	0.0138	0.4975
	Empirical SD	0.1812	0.2081	0.0918	0.3924	0.1793
	Average SE	0.1664	0.2023	0.0916	0.3740	0.1783
	Empirical Coverage (95%)	0.924	0.950	0.950	0.932	0.962
	Empirical Coverage (99%)	0.978	0.986	0.994	0.996	0.988

Table 3 Estimation results for melanoma data (e1690)

Model	SPT		Model (2)		Model A		Model B	
	$\hat{\boldsymbol{\alpha}}$	SE($\hat{\boldsymbol{\alpha}}$)	$\hat{\boldsymbol{\alpha}}$	SE($\hat{\boldsymbol{\alpha}}$)	$\hat{\boldsymbol{\alpha}}$	SE($\hat{\boldsymbol{\alpha}}$)	$\hat{\boldsymbol{\alpha}}$	SE($\hat{\boldsymbol{\alpha}}$)
$\boldsymbol{x}^{(0)}$				Incidence variable				
Intercept	0.016	0.097	0.2343	0.7280	0.2340	0.2723	0.3397	0.2758
Treatment	-0.244*	0.116	-0.2017	0.2639	-0.1585	0.1353	-0.2080	0.1323
Age	0.097	0.058	0.0222	0.0115	0.0096	0.0053	0.0087	0.0050
Sex	-0.115	0.120	-0.2758	0.2719	-0.1641	0.1403	-0.1375	0.1379
$\boldsymbol{x}^{(1)}$				Latency variable				
Intercept			0.3546	0.5932				
Treatment			-0.1498	0.2304	-0.0329	0.2319		
Age			-0.0069	0.0091	-0.0082	0.0092		
Sex			0.0091	0.2314	0.0221	0.2594		
ω			3.5869	2.9959				

*the regression parameter estimate is significantly different from zero at the 5% level of significance

Table 4 Estimation results for the breast cosmesis data

	θ	β
	Estimate (SE)	Estimate (SE)
	Method 1	
Intercept	0.4620 (0.2210)	/
Treatment	1.4502 (0.4228)*	-1.4153 (0.6133)*
	Method 2	
Intercept	0.4945 (0.2180)	/
Treatment	1.4782 (0.4489)*	-1.2760 (0.6365)*

*the regression parameter estimate is significantly different from zero at the 5% level of significance

