STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:1733–1747 (DOI: 10.1002/sim.1774)

Cure fraction estimation from the mixture cure models for grouped survival data

Binbing Yu[1,*,†], Ram C. Tiwari[2], Kathleen A. Cronin[2] and Eric J. Feuer[2]

[1]*Information Management Services, Inc., 12501 Prosperity Dr. Suite 200, Silver Spring, MD 20910, U.S.A.*
[2]*Statistical Applications and Research Branch, National Cancer Institute, 6116 Executive Boulevard, Bethesda, MD 20892, U.S.A.*

SUMMARY

Mixture cure models are usually used to model failure time data with long-term survivors. These models have been applied to grouped survival data. The models provide simultaneous estimates of the proportion of the patients cured from disease and the distribution of the survival times for uncured patients (latency distribution). However, a crucial issue with mixture cure models is the identifiability of the cure fraction and parameters of kernel distribution. Cure fraction estimates can be quite sensitive to the choice of latency distributions and length of follow-up time. In this paper, sensitivity of parameter estimates under semi-parametric model and several most commonly used parametric models, namely lognormal, loglogistic, Weibull and generalized Gamma distributions, is explored. The cure fraction estimates from the model with generalized Gamma distribution is found to be quite robust. A simulation study was carried out to examine the effect of follow-up time and latency distribution specification on cure fraction estimation. The cure models with generalized Gamma latency distribution are applied to the population-based survival data for several cancer sites from the Surveillance, Epidemiology and End Results (SEER) Program. Several cautions on the general use of cure model are advised. Copyright © 2004 John Wiley & Sons, Ltd.

1. INTRODUCTION

A primary focus of cancer research is the achievement of cure. The chance of being cured and the number of years of survival from diagnosis are of great interest to cancer patients and the medical community alike. As medical treatments progress, one would like to distinguish between a change in the probability of cure and an increase in the expected survival time for uncured patients. In this paper we define the probability of cure for a cohort of cancer patients as the percent of patients who are not at risk of dying from the cancer. This can also be described as the asymptotic value of the cause-specific or relative survival, $S(t)$, as time t goes to infinity (see equation (1) below).

*Correspondence to: Binbing Yu, Information Management Services, Inc., 12501 Prosperity Dr. Suite 200, Silver Spring, MD 20910, U.S.A.
†E-mail: yub@imsweb.com

Received September 2003
Accepted November 2003

Table I. Survival functions for some common parametric distributions[†].

Distribution	Parameterization			Transformation
	$G^*(t;\lambda,\rho)$	$G(t;\mu,\sigma)$	$G^s(z)$	
Lognormal	$1-\Phi(\frac{\log t-\lambda}{\rho})$	$1-\Phi(\frac{\log t-\mu}{\sigma})$	$1-\Phi(z)$	$\mu=\lambda,\sigma=\rho$
Loglogistic	$[1-(\lambda t)^{\rho}]^{-1}$	$\left[1+\exp(\frac{\log t-\mu}{\sigma})\right]^{-1}$	$[1+\exp(z)]^{-1}$	$\mu=-\log\lambda,\sigma=1/\rho$
Weibull	$\exp[-(\lambda t)^{\rho}]$	$\exp\left[-\exp(\frac{\log t-\mu}{\sigma})\right]$	$\exp[-\exp(z)]$	$\mu=-\log\lambda,\sigma=1/\rho$

[†]Where $\Phi(z)$ is the cdf of a standard normal distribution function, $G^*(t;\lambda,\rho)$ is the conventional form of survival function G; $G(t;\mu,\sigma)$ is the form in location-scale family; $G^s(z)$ is the form in standardized scale $z=\frac{\log t-\mu}{\sigma}$.

Boag [1] proposed a two-component mixture model for the analysis of survival data when a proportion of patients are cured. This cure model has since been used by many, see for example, Farewell [2], Gamel *et al.* [3], Yamaguchi [4] and Taylor [5] among others. This model postulates that a fraction of the patients, c, are long-term survivors and that the failure time for the uncured patients follows a proper survival distribution $G(t)$, referred to as latency distribution. The cause-specific survival function for the overall population survival time T is

$$S(t)=c+(1-c)G(t) \tag{1}$$

The latency distribution $G(t)$ can take the form of parametric or semi-parametric distributions. Among the parametric models, lognormal (LN), loglogistic (LL) and Weibull (WB) distributions are widely used to model the survival time. After reparametrization (Gamel *et al.* [6]), these survival functions belong to the location-scale family $G(t;\mu,\sigma)$ (See Table I). The median survival time of the uncured patients is $\exp(\mu)$. In the mixture cure models, both parameters c and μ may depend on the covariates, for example, $c=[1+\exp(-\beta_c x_c)]^{-1}$ and $\mu=\beta_\mu x_\mu$.

Let $Z=(\log T-\mu)/\sigma$. The standardized random variable (r.v.) Z follows LN, LL and WB distributions respectively and the distributions are given by $G^s(z)$. Yamaguchi [4] used an accelerated failure time (AFT) model with generalized Gamma (GG) distribution as the latency. The family of generalized Gamma distributions includes exponential, Weibull, lognormal and Gamma as its special cases and has considerable flexibility in capturing the shape of the survival distribution of T for uncured patients. The r.v. Z for the generalized Gamma models [4] has the log-Gamma distribution with density function

$$f(z;k)=\begin{cases}\dfrac{\sqrt{k}}{\Gamma(k^{-1})}\exp[k^{-1}(\sqrt{k}z-\mathrm{e}^{\sqrt{k}z}-\log k)] & \text{when } k>0,\\ \dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{z^2}{2}\right) & \text{when } k=0,\end{cases} \tag{2}$$

where k is the shape parameter. The GG distribution is lognormal when $k=0$ and is Weibull when $k=1$. The latency distribution $G(t)$ can also be semi-parametric (SP). Kuk and Chen

[7] and Sy and Taylor [8] developed a SP proportional hazards cure model with $G(t)$ modeled as

$$G(t) = \exp\left(-\exp(\mu)\int_{u=0}^{t}\lambda_0(u)\,\mathrm{d}u\right) \tag{3}$$

where $\lambda_0(t)$ be the unspecified baseline hazard function, which is only defined at the event times. In their models, the observations with censored times beyond the largest event time were classified as cured. For grouped survival data, $\lambda_0(t)$ is usually chosen to be piece-wise constant, where $\lambda_0(t_i) = \lambda_{0i}$ are defined only for the times with event [9]. The semi-parametric proportional hazards model, including piecewise exponential, does not belong to the AFT model because the covariate effect in the proportional hazards model is constant whereas the covariate effect in the AFT model is either decelerating or accelerating. In addition to the proportional hazards cure models, Li and Taylor [10] introduce a semi-parametric accelerated failure-time cure model, where the latency distribution is determined by an AFT model with unspecified baseline distribution.

However, the models mentioned above only consider continuous survival times. The mixture cure models have been applied to grouped survival data using parametric latency distribution [6] or semi-parametric proportional hazards latency distribution [11]. In this paper, the parametric cure models and semi-parametric proportional hazards cure model are applied to population-based grouped survival data to investigate the sensitivity of the parameter estimates.

Suppose that the survival times are grouped into J intervals given by $I_j = [t_{j-1}, t_j)$ for $j = 0, \ldots, J$ with $t_0 = 0$ and $t_J = \tau$, the end of follow-up. Let n_j be the number of people alive at the beginning of interval j, d_j be the number that died and l_j be the number lost to follow-up in the interval. By assuming uniform lost to follow-up, the number of people at risk during this interval is $n_j^* = n_j - \frac{1}{2}l_j$. Let P_j be the survival probability from all causes given the individual is alive at the beginning of interval j. Under the assumption that d_j is an independent binomial random variable with parameters n_j^* and P_j, the log-likelihood function of the observed survival data is

$$\log L(c, \mu, \sigma) = \sum_{j=1}^{J}\{d_j \log(1 - P_j) + (n_j^* - d_j)\log(P_j)\} \tag{4}$$

The cause-specific analysis assumes that the cause of death is known, and that P_j is the probability of surviving the cancer of interest. If the cause-of-death information is not accurate or reliable, we can use relative-survival approach [12]. Relative survival is a measure of excess mortality observed in a cohort of cancer patients as compared to a cancer-free general population. The fraction c can be interpreted as the proportion of cured patients whose survival pattern is similar to the general cancer-free population. As discussed by Weller *et al.* [13], by assuming independent forces of mortality, the probability of survival through interval j can be expressed as $P_j = R_j E_j$, where E_j is the expected survival through interval j in the population without the cancer of interest and R_j, the relative (net) survival rate, is the probability of surviving the cancer through interval j given the patient is alive at the beginning of this interval, which can be expressed as $R_j = S(t_j)/S(t_{j-1})$. For more on the binomial likelihood function (4) and the algorithm for obtaining maximum likelihood estimates (MLEs) of (c, μ, σ), see Gamel *et al.* [6].

The rest of the paper is organized as follows. Motivated by the analysis of Hodgkin's disease data, we examine the identifiability of mixture cure models for several parametric

models for $G(t)$ in Section 2. We carry out a simulation study to examine the sensitivity of cure fraction estimates from parametric or semi-parametric proportional hazards cure models in Section 3. In Section 4, the cure models are applied to the survival data for several cancer sites from SEER Program [14]. A discussion on the use of the parametric cure models is given in Section 5.

2. IDENTIFIABILITY OF PARAMETRIC MIXTURE CURE MODELS

2.1. Motivation: Hodgkin's disease data

Weller *et al.* [13] analysed the Hodgkin's disease survival data from the Connecticut Tumor Registry from 1940 to 1973, and also from the SEER Program (including Connecticut Registry) from 1974 to 1992 using parametric relative survival model (1) with both c and G depending on covariates. We applied the parametric cure model using log-normal (LN), log-logistic (LL) and Weibull (WB) survival functions as choices for G to a subset of the Hodgkin's disease survival data diagnosed after 1977 with 10-year follow-up [14].

In Figure 1(a), the Kaplan–Meier estimate of the survival curve and the estimated survival curves $\hat{S}(t)$ using LN, LL and WB for G are presented. We observe that the MLEs $(\hat{c}, \hat{\mu}, \hat{\sigma})$ of (c, μ, σ) for these models, given in Table II, are quite different. The cure fraction estimates differ significantly, from about 0 per cent for LN model to 49 per cent for WB model. This is evident from the tail behavior of $\hat{S}(t)$ plotted in Figure 1(b). The value of $\hat{S}(t)$ remains higher for WB model compared to those for LN and LL models when t is large. However, the survival curves from all three models fit the observed data equally well and they are almost not identifiable for $t \leqslant 10$ years. In order to examine the identifiability of the cure models, we propose the following definitions:

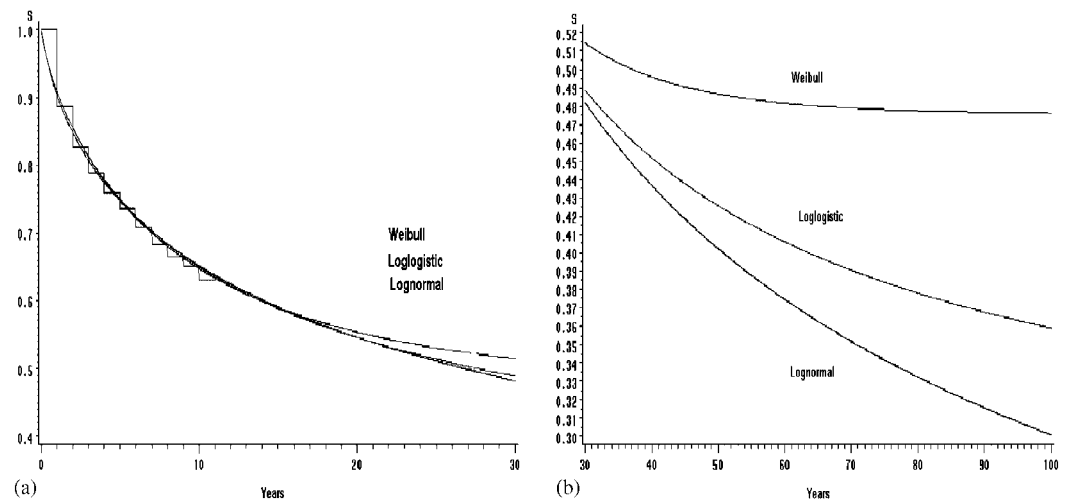

Figure 1. Estimated survival curves $\widehat{S(t)}$ for the Hodgkin's disease survival data with 10-year follow-up. (a) $\widehat{S(t)}$; $t \in (0, 30)$. (b) $\widehat{S(t)}$; $t \in (30, 100)$.

Table II. MLEs for Hodgkin's disease survival data [14] with 10-year follow-up.

Model	$\hat{c}$	$\hat{\mu}$	$\hat{\sigma}$	$\exp(\mu)$
LN	0.059	3.390	2.707	29.7
LL	0.299	2.518	1.398	12.4
WB	0.492	2.231	1.487	9.3

Definition 1
Let $S_i(t) = c_i + (1-c_i)G_i(t;\mu_i,\sigma_i), 0 \leqslant c_i \leqslant 1, i = 1,2$ be two mixture cure models. Given a small positive value ε, *the cure models are not identifiable within* ε *until time* τ if there exist $c_1 \neq c_2$ and $G_1(t;\mu_1,\sigma_1) \neq G_2(t;\mu_2,\sigma_2)$ for $t \leqslant \tau$ such that $|S_1(t) - S_2(t)| \leqslant \varepsilon$ for $t \leqslant \tau$.

Definition 2
Let $\Omega(\varepsilon,\tau) = \{S(t) : S(t)$ are not identifiable within ε until time $\tau\}$. The *maximum cure difference*, $\Delta(\varepsilon,\tau)$, is defined as $\max_{i,j}|c_i - c_j|$ for any $S_i(t), S_j(t)$ from $\Omega(\varepsilon,\tau)$.

We will examine the identifiability of the cure models within the family of three parametric distributions, i.e. LN, LL and WB, for $G_i(t)$. Let $G_1(t), G_2(t)$ and $G_3(t)$ denote the LN, LL and WB survival functions, respectively. Here, we assume that the maximum follow-up time τ is 10 or 20 years and $\varepsilon = 0.01$.

2.2. *Identifiability of* $S(t)$ *with different choices of latency distribution* $G(t)$

The mixture cure model postulates that a fraction c of the patients are cured from cancer and become long-term survivors. Thus, c can be estimated accurately if the follow-up time T is sufficiently long. If the follow-up time is not long enough, the observed survival fraction $S(t)$ will contain both cured patients with fraction c and uncured survivors with fraction $(1-c)G(t)$. The cure fraction c is related to $G(t;\mu,\sigma)$ as

$$c = 1 - \frac{1 - S(t)}{1 - G(t;\mu,\sigma)} \tag{5}$$

For an observed value of $S(t)$ at time t, as $G(t)$ increases, the value of c decreases. Even for the same μ and σ, different choice of G may yield different value of c. To examine the effect of different choice to $G(\cdot)$, we assume that the non-identifiable distributions sharing the same values μ^* and σ^* form $\Omega(\varepsilon,\tau|\mu^*,\sigma^*)$, a subset of $\Omega(\varepsilon,\tau)$. Let $S_i(t) = c_i + (1 - c_i)G_i(t;\mu^*,\sigma^*)$, $i = 1,2,3$, be the non-identifiable models in the set.

The maximum differences in cure fractions c of the non-identifiable mixture cure models from $\Omega(\varepsilon,\tau|\mu^*,\sigma^*)$ are listed in Table III. We see that maximum cure different is less than 0.06 when the follow-up time is greater than or equal to the median survival time for uncured patients $(\exp(\mu^*))$. In this case, the cure fraction estimates from different models are similar. When the median survival time is 15 years and follow-up is only 10 years, the maximum cure difference from different models is as high as 80 per cent. Hence, a follow-up time longer than median survival time is necessary for the cure difference to be small.

Table III. Maximum cure difference $\Delta(\varepsilon, \tau|\mu^*, \sigma^*)$ for the non-identifiable set $\Omega(\varepsilon, \tau|\mu^*, \sigma^*)$ within $\varepsilon = 0.01$.

		σ^*			
$\exp(\mu^*)$	τ	0.25	0.5	1	1.5
5	5	0.06	0.06	0.06	0.06
	10	0.02	0.04	0.04	0.04
	15	0.02	0.02	0.04	0.04
10	5	1.00	0.70	0.20	0.14
	10	0.06	0.06	0.06	0.06
	20	0.02	0.04	0.04	0.04
15	10	0.80	0.22	0.12	0.10
	15	0.06	0.06	0.06	0.06
	20	0.04	0.04	0.06	0.06

2.3. Identifiability of $S(t)$ with different (c, μ, σ) but the same latency distribution $G(t)$

Even for the same latency distribution G^*, different pairs of parameters $\{(c_i, \mu_i, \sigma_i), i = 1, 2\}$ may produce non-identifiable $S(t)$ for $t \leqslant \tau$. Assume that $S_i(t) = c_i + (1 - c_i)G^*(t; \mu_i, \sigma_i), i = 1, 2$, are not identifiable within distance ε until time τ with the same latency distribution G^*, where the distribution $G^*(\cdot)$ is one of the LN, LL and WB distributions. Let $\Omega(\varepsilon, \tau|G^*)$ be the non-identifiable set of $S(t)$ with the same G^* and $\Delta(\varepsilon, \tau|G^*)$ be the maximum cure difference of the non-identifiable set. Let $\underline{\Omega}(\varepsilon, \tau|G^*)$ be a subset of $\Omega(\varepsilon, \tau|G^*)$ where the $S(t)$ has the same values at $\tau/4$ and $3\tau/4$, i.e.

$$\underline{\Omega}(\varepsilon, \tau|G^*) = \left\{ S(t) \in \Omega(\varepsilon, \tau|G^*) : e^{\mu} \leqslant 15, \sigma \leqslant 1.5 \text{ and } S_i(t) = S_j(t) \text{ for } t = \frac{\tau}{4}, \frac{3\tau}{4} \right\}$$

The maximum cure difference for the set $\underline{\Omega}(\varepsilon, \tau|G^*)$, $\underline{\Delta}(\varepsilon, \tau|G^*)$, is an approximation of $\Delta(\varepsilon, \tau|G^*)$. In Table IIIb, we show the values of $\underline{\Delta}(\varepsilon, \tau|G^*)$ for different functions of G^*. The last six columns in Table IV shows the pairs of (c_i, μ_i, σ_i) that generate the non-identifiable survival curves with the maximum cure difference. The maximum cure difference usually occurs when the median survival times for uncured patients are relatively long compared with the follow-up time. For example, if G^* is Weibull with the median survival time of 15 years and the follow-up is only 10 years, the maximum cure fraction difference is as large as 99 per cent although the difference in $S(t)$ is less than 0.01.

2.4. Illustration of the Hodgkin's disease data

When the latency distribution G_i can take different forms and the values (μ_i, σ_i) can be different, the maximum cure difference for the non-identifiable set is even larger. The Hodgkin's disease data is a typical example. The difference between the three survival curves defined by (G, c, μ, σ) in Table II is less than 0.007. However, the cure takes values from 0.059 to 0.492, with maximum cure difference 0.433. The loglikelihood plot is an important tool to examine the sensitivity of the cure fraction estimate. In Figure 2, we show the plots of profile loglikelihood $\log L(c, \hat{\mu}_c, \hat{\sigma}_c)$ with respect to c for the Hodgkin's disease survival data, where

Table IV. The maximum cure difference, $\underline{\Delta}(\varepsilon, \tau | G^*)$ for the set $\underline{\Omega}(\varepsilon, \tau | G^*)$ with $\varepsilon = 0.01$ and the pair of extreme (c_i, μ_i, σ_i).

			$(c_i, \mu_i, \sigma_i), i = 1, 2$					
G^*	τ	$\underline{\Delta}(\varepsilon, \tau \mid G^*)$	$(c_1,$	$\mu_1,$	$\sigma_1)$	$(c_2,$	$\mu_2,$	$\sigma_2)$
LN	5	0.94	0.96	1.79	0.40	0.02	2.71	0.54
	10	0.34	0.02	2.71	1.50	0.36	2.04	1.21
	15	0.25	0.03	2.71	1.50	0.28	2.20	1.24
	20	0.20	0.01	2.71	1.50	0.21	2.30	1.26
LL	5	0.98	0.99	2.20	0.10	0.01	2.66	0.10
	10	0.98	0.00	2.71	0.10	0.98	2.31	0.10
	15	0.23	0.06	2.71	1.10	0.29	2.14	0.92
	20	0.17	0.01	2.71	1.00	0.18	2.34	0.87
WB	5	0.99	0.00	2.71	0.20	0.99	1.77	0.20
	10	0.99	0.00	2.71	0.10	0.99	2.24	0.10
	15	0.19	0.08	2.71	1.40	0.27	2.18	1.23
	20	0.13	0.05	2.71	1.30	0.18	2.38	1.18

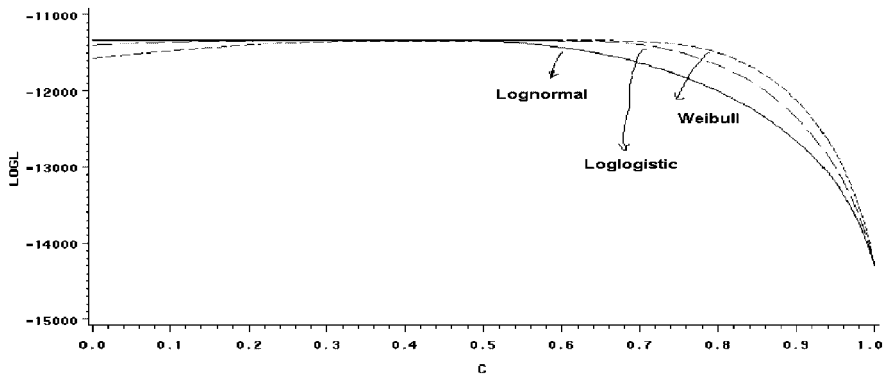

Figure 2. Plot of maximum profile loglikelihood vs. cure fraction c for Hodgkin's disease data.

$\log L(c, \hat{\mu}_c, \hat{\sigma}_c) = \max_{(\mu,\sigma)} \log L(c, \mu, \sigma)$ is the maximum loglikelihood value with fixed c. We observe that the profile loglikelihood is quite flat when $c < 0.5$ for all three latency distributions G, so the MLE's of cure fraction have large variances and are not stable. Hence, longer follow-up time is necessary to obtain an accurate cure fraction estimate.

3. SIMULATIONS: EFFECT OF FOLLOW-UP TIME AND LATENCY DISTRIBUTION SPECIFICATION ON CURE FRACTION ESTIMATION

In Section 2, we discussed the identifiability of the mixture cure models with respect to the distance ε and the maximum follow-up time τ. We observed that different combinations

of latency distribution $G(t)$ and parameters (c, μ, σ) can produce non-identifiable survival functions $S(t)$ within distance ε. A key factor for the instability in the cure fraction estimates from different models is the median survival time in relation to the length of follow-up. A longer median survival time than the follow-up time suggests that the survivors include substantial fraction of uncured patients and a significant number of deaths occur after the end of follow-up.

To improve the accuracy of cure fraction estimates, we should either have a larger population size which would increase the power of detecting smaller difference in the survival functions $S(t)$ or have a longer follow-up time which would allow the survival curve to level off. For population-based cancer registry data, e.g. SEER data, the population size is usually the current population in the registries and does not vary much. Hence we will focus on the effect of follow-up time on the cure fraction estimation and see whether increasing the length of follow-up time increases the accuracy of cure fraction estimates. For the purpose of comparison, we run simulation to examine the cure fraction estimates for different values of follow-up time, cure fraction and median survival time for uncured patients.

We also compare the estimates from mixture cure models with different specifications for the latency distribution. As seen in Section 2, besides the length of follow-up time, the choice of latency distribution function $G(t)$ could introduce the non-identifiability of the survival function $S(t)$. Also, if the analysis was carried out using a specific parametric distribution, e.g. LN, LL or WB, the cure fraction estimates may not be sufficiently robust as the fitted distribution might grossly misspecify the true latency distribution $G_0(t)$. Peng *et al.* [15] employed a generalized **F** distribution in the mixture cure model. The GG model showed an equally good fit as the generalized **F** mixture model in their application of lymphoma data. Because of the flexibility of the generalized Gamma mixture model, we use it as a choice of $G(t)$. We also use semi-parametric (SP) piecewise exponential distribution for $G(t)$ to allow the latency distribution to be less dependent on the parameters. The tail, after the last observed event, is exponential for the SP cure model. The fitted function $G(t)$ used in the simulation are LN, LL, WB, GG and SP.

The data were generated from model (1) with the true latency survival function $G_0(t)$ as either WB or LL. The total population size in the beginning, n_0, was taken to be 50,000. The data were stratified by sex and half of the population are female. The number of death in the jth interval $[t_{j-1}, t_j)$ was assumed to be binomial, i.e. $d_j \sim B(n_j, S(t_j)/S(t_{j-1}))$. The number of people at risk in the beginning of the interval $n_j = n_{j-1} - d_{j-1}, j = 1, \ldots, J$. The maximum follow-up time τ was 10 or 20 years. The values of the cure fraction were $c_0 = 0.1, 0.25, 0.5, 0.75$ and the median survival times for uncured patients were taken to be $\exp(\mu_0) = 5, 10, 15$ and $\sigma = 1$. The bias for the estimates for cure fraction and median survival was calculated as $B(\hat{c}) = \sum_{i=1}^{100}(\hat{c}_i - c_0) \times 100$ per cent/100 and $B(\mathrm{e}^{\hat{\mu}}) = \sum_{i=1}^{100}[\exp(\hat{\mu}_i) - \exp(\mu_0)]/100$ for one hundred simulated samples, where $\hat{c}_i$ and $\hat{\mu}_i$ are the estimates of c_0 and μ_0 in the ith simulation. The simulation results for different true parameters (c_0, μ_0, G_0) and fitted distributions $G(t)$ are shown in Tables V and VI.

In Table V, the follow-up time is 10 years. When the fitted distribution $G(t)$ correctly specifies the true latency distribution $G_0(t)$, i.e. both are LN or LL, the cure fraction estimates are accurate. The maximum (absolute) biases of the cure fraction estimates are 2.7 per cent for LN and 2.6 per cent for LL when $\mathrm{e}^{\mu_0} \leqslant 10$; the maximum (absolute) biases of the cure fraction estimates are 5.5 per cent for LN and 8.6 per cent for LL when $\mathrm{e}^{\mu_0} = 15$. Here we see that the longer follow-up time, more accurate the cure fraction

Table V. The biases of $(\hat{c}, e^{\hat{\mu}})$ for different models with 10-year follow-up.

			LN		LL		WB		GG		SP	
c_0	e^{μ_0}	G_0	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$
0.10	5	LN	−0.5	0.0	3.2	−0.2	−3.1	2.8	0.8	0.0	20.3	8.9
		LL	−5.4	0.7	−1.4	0.2	−8.7	5.7	9.6	0.2	29.3	−1.5
	10	LN	−0.7	0.1	14.0	−1.8	2.6	2.3	3.6	−0.6	44.6	36.9
		LL	−9.8	2.5	−2.6	0.6	−5.8	7.2	−5.6	7.0	44.3	−5.0
	15	LN	−1.6	0.2	8.0	−2.1	−4.8	2.5	−5.5	2.7	58.6	84.7
		LL	−8.6	3.8	0.4	−0.2	16.1	0.6	−4.0	8.2	53.5	−9.1
0.25	5	LN	−0.6	0.0	2.3	−0.2	−21.0	5.9	0.5	0.0	17.7	8.6
		LL	−6.0	0.9	−1.0	0.1	−8.5	6.3	6.5	0.3	24.1	−1.5
	10	LN	−2.7	0.4	14.2	−2.2	4.2	1.8	4.4	−0.5	37.2	35.6
		LL	−21.3	7.1	−2.1	0.6	−11.1	10.0	9.8	0.6	37.0	−5.0
	15	LN	−3.4	0.7	12.8	−3.2	−10.9	3.6	8.4	−1.4	44.4	99.1
		LL	−18.9	9.5	−6.1	2.1	14.8	0.2	14.5	−0.3	43.9	−8.9
0.50	5	LN	−0.4	0.0	1.6	−0.2	−15.2	6.9	0.3	0.0	11.8	8.6
		LL	−4.6	1.1	−0.7	0.1	−14.6	13.2	4.1	0.3	16.2	−1.5
	10	LN	−1.9	0.4	10.3	−2.4	15.4	−1.6	3.9	−0.6	20.2	43.3
		LL	−21.3	11.8	−0.3	−0.0	−20.4	18.5	−2.1	4.8	22.3	−4.4
	15	LN	−5.0	1.6	11.2	−4.1	−14.0	5.1	−6.2	2.3	27.8	108.6
		LL	−33.9	30.0	−8.6	5.0	−16.8	19.0	−3.2	9.6	26.9	−8.2
0.75	5	LN	−0.5	0.1	0.5	−0.2	−2.7	4.7	−0.1	0.0	2.6	10.8
		LL	−4.1	2.8	−0.7	0.2	−6.4	15.3	0.8	0.6	5.9	−1.0
	10	LN	−1.1	0.4	4.9	−2.4	7.8	−1.8	−2.4	1.8	8.5	48.6
		LL	−14.9	19.4	−2.1	1.8	−25.0	42.0	−0.7	7.0	9.8	−3.9
	15	LN	−5.5	3.4	7.6	−5.2	−7.7	5.0	4.3	−2.2	12.1	121.5
		LL	−28.7	66.3	−8.5	11.0	−17.7	34.3	−8.9	26.2	12.7	−7.6

estimates are when the fitted distribution $G(t)$ is correctly specified. When the distribution $G(t)$ misspecifies the true latency distribution $G_0(t)$, i.e. $(G, G_0) = (\mathrm{LL}, \mathrm{LN})$, (LN, LL), (WB, LN) or (WB, LL), the cure fraction estimates are very unstable, especially when $c_0 = 0.25$ or 0.50. For example, when $c_0 = 0.50, e^{\mu_0} = 5$ and $(G, G_0) = (\mathrm{LL}, \mathrm{LN})$, the absolute bias is as high as 33.9 per cent; when $c_0 = 0.25, e^{\mu_0} = 10$ and $(G, G_0) = (\mathrm{LL}, \mathrm{WB})$, the bias is 20.4 per cent.

If the fitted model is semi-parametric (SP) proportional hazards, the biases of cure fraction estimates are substantially positive. That is because the tail of the piecewise exponential is very small, hence $G(t)$ decreases very fast when t becomes large. Based on the simulation, the piecewise exponential distribution is therefore not a good choice of $G(t)$. When the fitted model $G(t)$ is GG, the biases of cure fraction estimates GG are slightly larger than those when $G(t)$ correctly specifies $G_0(t)$. However, the cure fraction estimates from GG are in general more robust when $G(t)$ and $G_0(t)$ are the same model. Except for one case where $(c_0, e^{\mu_0}, G_0) = (0.25, 15, \mathrm{LL})$, all (absolute) biases are less than 9.8 per cent. Especially when $e^{\mu_0} = 5$, the maximum (absolute) bias is only 5.5 per cent for all cases. Li and Taylor [10] found a similar phenomena that the proportional hazards latency was not likely to be ideal

Table VI. The biases of $(\hat{c}, e^{\hat{\mu}})$ for different models with 20-year follow-up.

			LN		LL		WB		GG		SP	
c_0	e^{μ_0}	G_0	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$	$B(\hat{c})$	$B(e^{\hat{\mu}})$
0.10	5	LN	−0.4	0.0	−1.4	0.1	5.6	1.7	−0.1	0.0	5.9	11.9
		LL	−0.6	0.0	−0.6	0.0	12.0	1.2	5.6	0.3	16.5	−0.6
	10	LN	−0.5	0.0	4.2	−0.7	7.0	3.2	1.0	−0.0	19.4	61.9
		LL	−8.9	2.2	−3.6	0.8	−10.0	11.5	7.0	0.4	28.1	−3.0
	15	LN	−1.8	0.3	11.7	−2.5	18.8	0.1	2.3	−0.2	34.2	159.6
		LL	−9.9	3.6	−1.9	0.5	−6.9	12.8	8.7	1.5	37.9	−6.3
0.25	5	LN	−0.4	0.0	−1.2	0.1	4.6	1.7	−0.2	0.1	4.4	12.0
		LL	−0.8	0.1	−0.9	0.1	10.0	1.2	4.6	0.3	13.4	−0.6
	10	LN	−0.4	0.0	3.5	−0.7	−15.1	10.1	0.3	0.0	17.3	61.3
		LL	−11.3	3.7	−1.2	0.2	−4.0	10.1	8.7	0.5	24.3	−3.1
	15	LN	−1.1	0.2	9.7	−2.5	17.3	−0.4	1.9	−0.2	28.3	160.4
		LL	−22.2	11.4	−1.4	0.4	−18.1	21.4	10.8	0.4	31.5	−6.2
0.50	5	LN	−0.5	0.0	−1.1	0.1	2.8	1.7	−0.3	0.1	3.5	11.8
		LL	−0.8	0.1	−0.9	0.1	6.4	1.2	2.7	0.4	9.3	−0.7
	10	LN	−0.8	0.1	2.0	−0.6	−1.8	6.4	0.3	0.1	11.6	60.2
		LL	−8.3	4.1	−1.2	0.4	−6.1	16.5	4.0	0.9	15.9	−3.0
	15	LN	−1.7	0.6	6.0	−2.3	14.5	−2.0	0.8	−0.1	18.8	154.7
		LL	−20.0	16.8	−1.2	0.5	−22.7	34.4	5.4	1.2	19.7	−5.8
0.75	5	LN	−0.6	0.1	−0.8	0.2	1.2	1.8	−0.5	0.2	−1.8	14.3
		LL	−0.9	0.3	−0.8	0.3	3.0	1.3	0.7	0.5	2.5	−0.1
	10	LN	−1.1	0.4	0.5	−0.5	2.7	2.4	−0.3	0.2	1.8	73.6
		LL	−5.9	5.2	−1.9	0.8	−0.4	11.2	1.3	1.6	4.8	−1.6
	15	LN	−1.3	0.6	2.7	−2.4	6.7	−1.6	0.2	−0.1	5.1	195.7
		LL	−13.5	24.8	−1.8	1.5	−20.8	66.0	−0.5	6.2	7.5	−4.4

for their tonsil cancer example. As pointed out by a referee, the poor performance of the semi-parametric model in Table V may therefore be due to the use of proportional hazards model for latency rather than the accelerated failure time model.

For the GG distribution, the effect of follow-up length relative to the median survival time (e^{μ_0}) is significant. When e^{μ_0} changes from 5, 10 to 15, the maximum (absolute) bias of cure changes from 5.5, 9.8 to 14.5 per cent. Generally, the bias of cure fraction and that of the median survival time are negatively correlated. The bias of the median survival time is less significant than that of the cure fraction. The length of follow-up time is also critical to the estimation of median survival time. For the GG distribution, the maximum (absolute) biases of the median survival are 0.6, 6.2 and 24.6 years when the true median survival time e^{μ_0} is 5, 10, 15 years, respectively.

Table VI shows the simulation results when the follow-up time is 20 years. The results are similar to Table V when the median survival time is 5 or 10 years. The cure estimates from SP model have substantially positive bias. The GG distributions is the most robust method and the maximum (absolute) biases for c_0 and e^{μ_0} are 8.7 per cent and 1.6 years when $e^{\mu_0} \leqslant 10$. When the median survival time is 15 years, the maximum (absolute) bias of cure fraction

decreases from 14.5 to 10.6 per cent and the maximum (absolute) bias of median survival decreases from 24.6 years to 6.2 years when the follow-up time increases from 10 to 20 years.

Based on the above simulation, we see that both latency distribution specification and length of follow-up time play important roles to improve the parameter estimates of the mixture cure models. It is necessary that the follow-up time is longer than the median survival time for the uncured individuals so that most events can be observed before the end of the follow-up. Empirically, for the cancer sites with shorter median survival time ($\leqslant 5$) for uncured patients or higher cure fraction ($c \geqslant 75$ per cent), the estimates of cure fraction from different models are similar. The generalized Gamma distribution seems to be more robust than the other distributions in the sense that GG gives smaller values of biases. However, when the true latency distribution is loglogistic, the bias of cure fraction from GG model can still be as high as 10.8 per cent. Hence, when we fit the mixture cure models to the survival data, we recommend to try different latency distributions and compare the resulted cure fraction estimates before drawing conclusions about the cure rate.

4. APPLICATION TO SEVERAL CANCER SITES USING SEER PROGRAM DATA

We fit the parametric cure model (1) using LN, LL, WB and GG as choices for G to the 10 year follow-up survival data for testicular caner, colon and rectum cancer, lung cancer and female breast cancer from SEER 9 registries. Table VII presents the MLEs of the parameters for these cancer sites with different follow-up time. Figure 3 presents the plots of estimated survival curves $\widehat{S(t)}$ for the data with 10-year follow-up. When the follow-up is only 10 years, the cure fraction estimate is stable for testicular cancer, colon and rectum cancer and lung cancer as the uncured patients for such cancers have short survival time. However, for female breast cancer, the difference in cure fraction estimates is as high as 30 per cent and uncured patients have around 10 years median survival time.

To examine the effect of longer follow-up, we analyze the female breast cancer data with 25-year follow-up (Table VII). The longer follow-up slightly reduces the difference of cure

Table VII. MLEs of cure model for survival data from serval SEER cancer sites.

Site	Follow-up		Latency distribution $G(t)$				
			LN	LL	WB	GG	Difference
Colon	10	c (per cent)	43.1	41.2	48.5	48.1	6.9
		e^{μ}	1.6	1.8	2.2	2.2	0.6
Lung	10	c (per cent)	8.8	7.6	9.9	8.8	2.3
		e^{μ}	0.6	0.6	0.8	0.6	0.2
Testicular	10	c (per cent)	90.5	90.4	91.0	90.5	0.6
		e^{μ}	1.4	1.4	1.8	1.4	0.4
Breast	10	c (per cent)	31.8	49.1	62.2	55.0	30.4
		e^{μ}	12.1	7.2	6.5	7.5	5.6
	25	c (per cent)	29.4	42.2	54.1	37.2	24.7
		e^{μ}	12.9	8.8	8.8	11.3	4.1

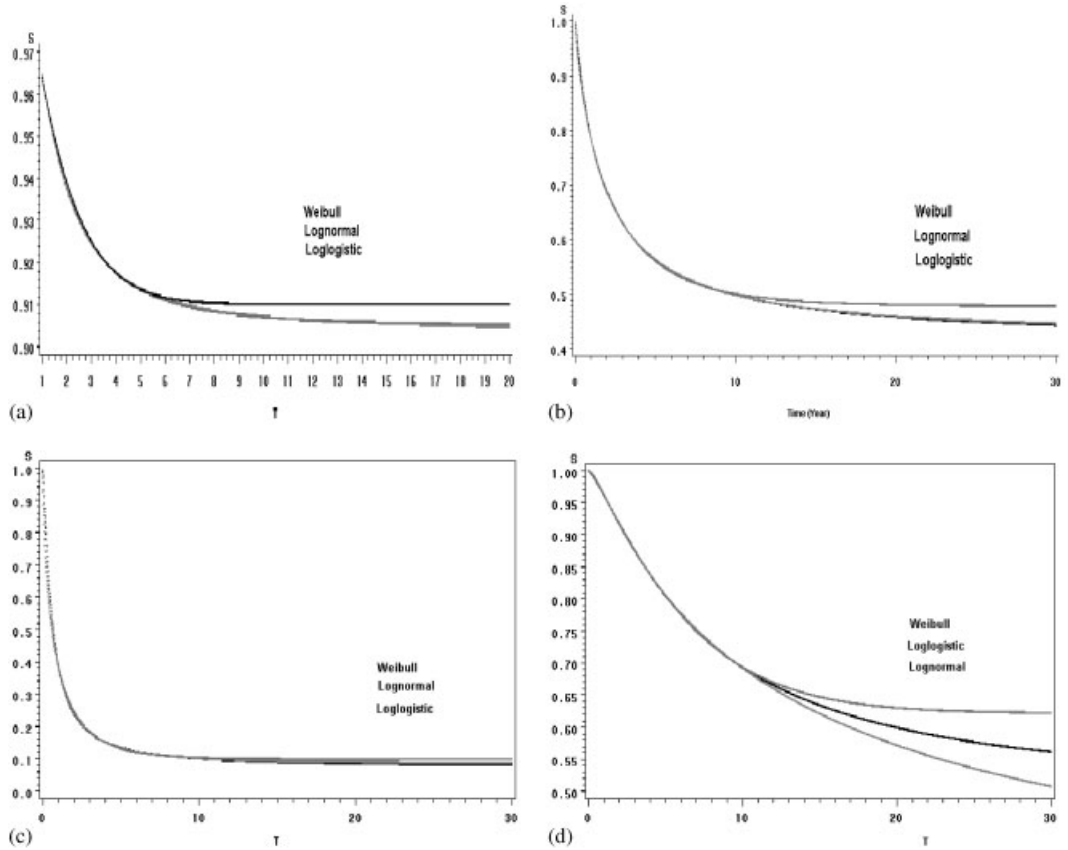

Figure 3. Estimated survival curves $\hat{S}(t)$ from different models for survival data from several cancer sites with 10 year follow-up: (a) testicular caner; (b) colon and rectum cancer; (c) lung cancer; (d) female cancer.

fraction estimates from 30.4 per cent for 10-year follow-up to 24.7 per cent for 25-year follow-up. The effect of longer follow-up time in reducing the cure fraction difference is not significant because the median survival time for uncured breast cancer patients is relatively long.

Because substantial difference in the cure fraction estimates exists for female breast cancer. One possible reason for the model influence on the results is heterogeneity in the population of cancer patients considered. To explore the issue of heterogeneity, we looked at female breast cancer data in more detail. The various stages and histological types of breast cancer and age of diagnosis exhibit differences in regard to relative frequency, site pattern within the breast and patient survival. We first stratified the female breast cancer patients by age and cancer stage and applied the parametric cure models for each stratum. The results are shown in Table VIII.

Table VIII. MLEs for female with breast cancer by age and stage after 25 year follow-up.

Age	Stage		Latency distribution $G(t)$ LN	LL	WB	GG	Difference
All	Localized	c (per cent)	63.5	70.4	75.3	63.5	11.8
		e^{μ}	14.7	11.2	11.9	14.7	3.5
	Regional	c (per cent)	30.6	34.4	42.6	30.6	12.0
		e^{μ}	6.8	6.2	7.2	6.8	1.0
	Distant	c (per cent)	3.0	0.5	6.5	5.7	6.0
		e^{μ}	1.5	1.6	2.2	2.0	0.7
<50	Localized	c (per cent)	71.2	75.8	79.7	71.2	8.5
		e^{μ}	12.4	9.2	10.0	12.4	3.2
	Regional	c (per cent)	43.7	44.3	49.8	43.7	6.1
		e^{μ}	4.4	4.3	5.5	4.4	1.2
	Distant	c (per cent)	22.3	20.7	24.4	22.6	3.7
		e^{μ}	1.2	1.2	1.9	1.2	0.7
50+	Localized	c (per cent)	57.6	68.3	73.9	57.6	16.3
		e^{μ}	18.4	12.9	13.3	18.4	5.5
	Regional	c (per cent)	22.4	22.5	33.0	22.4	10.6
		e^{μ}	3.4	3.4	4.2	3.4	0.8
	Distant	c (per cent)	4.0	2.1	6.1	4.0	4.0
		e^{μ}	0.6	0.6	1.0	0.6	0.4

If the data were stratified only by historical stage. The maximum difference in cure fraction estimates reduced from 24.7 per cent for the combined data to 12.0 per cent for regional breast cancer. For distant breast cancer, the maximum cure difference was only 6 per cent because the median survival for uncured patients is only 2 years. If the data were further stratified by age, the maximum cure difference decreased even further, which are 8.5 per cent for the younger group and 16.3 per cent for the older group. For both age groups (<50 years, 50+ years), the cure fraction estimates for the distant breast cancer patients are stable because the uncured patients have very short median survival time (<2). Because of the large median survival time (>10) for the uncured localized breast cancer patients, the difference in estimates of c for localized cancer is larger; 8.5 per cent and 16.3 per cent for the younger and the older groups, respectively. The estimates among older group have more variation, especially for localized breast cancer.

The largest difference in estimates of c occurs in the older group with localized breast cancer. For this group, we stratified the survival data by histological type. Infiltrating duct carcinoma is the largest group of female breast cancer, constituting 67.9 per cent of the total. Lobular carcinomas and Medullary carcinoma are the second and third largest groups, with 6.3 and 2.8 per cent of the total. Within each histological group, the population are more homogeneous and the difference in the estimates of c is even smaller (Table IX). This is partly because the cure fraction for these groups are very high (>75 per cent). Compared with 16 per cent difference in cure fraction estimates $\hat{c}$ in Table VIII, the corresponding difference within each histological group is less than 11 per cent. From these examples, we see that stratification of population into homogeneous groups, such as based on historic stages and histology, will reduce the sensitivity of cure fraction estimation to the model assumption of G.

Table IX. MLEs of (c, μ, σ) by histology for female localized breast cancer patients (50+) after 25 years follow-up.

Histology		LN	LL	WB	GG	Difference
Infiltrating duct carcinoma, NOS (8500)	c (per cent)	73.6	79.7	84.2	73.6	10.6
	e^{μ}	10.4	8.0	7.8	10.4	2.6
Lobular carcinoma (8520-8521)	c (per cent)	86.0	88.7	90.6	86.0	4.6
	e^{μ}	10.4	9.1	9.4	10.4	1.3
Medullary carcinoma (8510-8512)	c (per cent)	89.0	88.7	89.7	89.5	1.0
	e^{μ}	3.4	3.4	3.8	3.7	0.4

5. DISCUSSIONS

Based on the analyses presented above, the estimation of the cure fraction is subject to the non-identifiability of the cure model (1). The estimates of c may be sensitive to the model assumptions on survival function G for uncured patients. The model using generalized Gamma as the latency distribution is more robust. We also observed that relatively long follow-up time with respect to the median survival time and homogeneity of the observations are two factors influencing the accuracy of cure fraction estimates.

For cancer sites with relatively short median survival time (<5 years) or high cure rate (>75 per cent), the parametric cure model performs well and the cure fraction estimates are stable. However, if the median survival time is long and the cancer patients in the analysis are heterogeneous, then the cure fraction estimates are sensitive to the specification of G and the heterogeneity of population across different groups. To study the sensitivity of the cure fraction estimates $\hat{c}$, we suggest examining the likelihood plots and the MLEs of the parameters (c, μ, σ) corresponding to different choices of G. If the likelihood function is flat or if the MLEs of (c, μ, σ) shows much variation, then stratification or longer follow-up might be required.

An alternative approach would be not to use a mixture formulation but rather to use a survival model in which long-term survivors are accounted for via a hazard function that decreases appropriately to a value near zero as time tends to infinity, see Yakovlev *et al.* [16] and Chen *et al.* [17]. Such approaches can have computational and modelling advantages over the mixture approach. Unlike mixture models, however, they do not allow separate interpretation of the effects of the covariates on the probability of cure and on the failure time distribution of the susceptible via two sets of regression coefficients.

Another concept of cure, usually referred to as 'personal cure', defines the probability of cure as the crude survival probability with respect to other competing risks that preclude the event of interest from occurring. For example, a person is considered cured from the cancer of interest if their death is due to the other causes. The estimation of the proportion of 'personal cure' has been explored by Larson and Dinse [18] and it is of interest to extend these models for use with cancer registry data.

REFERENCES

1. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society* 1949; **11**:15–44.

2. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**:257–262.
3. Gamel JW, McLean IW, Rosenberg SH. Proportion cured and mean log survival time as functions of tumor size. *Statistics in Medicine* 1990; **9**:999–1006.
4. Yamaguchi K. Accelerate failure-time regression models with a regression model of surviving fraction: an application to the analysis of 'permanent employment' in Japan. *Journal of the American Statistical Association* 1992; **87**:284–292.
5. Taylor JMG. Semiparametric estimation in failure time mixture models. *Biometrics* 1995; **51**:899–907.
6. Gamel JW, Weller EA, Wesley MN, Feuer EJ. Parametric cure models of relative and cause-specific survival for grouped survival times. *Computer Methods and Programs in Biomedicine* 2000; **61**:99–110.
7. Kuk AYC, Chen CH. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 1992; **79**:187–200.
8. Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; **56**:227–236.
9. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 1978; **34**:57–67.
10. Li CS, Taylor JMG. A semi-parametric accelerated failure time cure model. *Statistics in Medicine* 2002; **21**: 3235–3247.
11. Yu B, Tiwari RC. Application of EM algorithm to regression models for grouped survival data with a cure fraction, *under revision.*
12. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *Monograph No. 6*, 1961; Bethesda, National Cancer Institute.
13. Weller EA, Feuer EJ, Frey CM, Wesley MN. Parametric relative survival regression using generalized linear models with application to Hodgkin's Lymphoma. *Applied Statistics* 1999; **48**(1):79–89.
14. Surveillance, Epidemiology and End Results Program. *The Portable Survival System/Mainframe Survival System* 1999; Bethesda, National Cancer Institute.
15. Peng YW, Dear KBG, Denham JW. A generalized F mixture model for cure rate estimation. *Statistics in Medicine* 1998; **17**:813–830.
16. Yakovlev AY, Asselain B, Bardou VJ, Fourquet A, Hoang T, Rochefordiere A, Tsodikov AD. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et Analyse de Donnees Spatio-Temporelles* No 12, Asselain B, Boniface M, Duby C, Lopez C, Masson JP, Tranchefort J (eds). Société Française de Biométrie: ENSA Rennes, France, 66–82.
17. Chen MH, Ibrahim JG, Sinha D. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999; **94**:909–919.
18. Larson MG, Dinse G. A mixture model for the regression analysis of competing risks data. *Applied Statistics* 1985; **34**:201–211.
19. Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982; **38**:933–942.

