BIOMETRICS 56, 237–243
March 2000

A Nonparametric Mixture Model for Cure Rate Estimation

Yingwei Peng

[1]Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland A1C 5S7, Canada
email: ypeng@math.mun.ca

and

Keith B. G. Dear

Department of Statistics, The University of Newcastle, Newcastle, New South Wales 2308, Australia

SUMMARY. Nonparametric methods have attracted less attention than their parametric counterparts for cure rate analysis. In this paper, we study a general nonparametric mixture model. The proportional hazards assumption is employed in modeling the effect of covariates on the failure time of patients who are not cured. The EM algorithm, the marginal likelihood approach, and multiple imputations are employed to estimate parameters of interest in the model. This model extends models and improves estimation methods proposed by other researchers. It also extends Cox's proportional hazards regression model by allowing a proportion of event-free patients and investigating covariate effects on that proportion. The model and its estimation method are investigated by simulations. An application to breast cancer data, including comparisons with previous analyses using a parametric model and an existing nonparametric model by other researchers, confirms the conclusions from the parametric model but not those from the existing nonparametric model.

KEY WORDS: Breast cancer; Censored data; EM algorithm; Logistic regression; Marginal likelihood; Proportional hazards assumption; Survival data.

1. Introduction

In some clinical studies, a substantial proportion of patients who respond favorably to treatment appear subsequently to be free of any signs or symptoms of the disease and may be considered cured, while the remaining patients may eventually relapse. Long-term censored survival times usually appear in data. The focus of the studies is on the estimation of the proportion of patients who are cured and of the failure time distribution of uncured patients.

Patients with long-term censored times can be found in many cancer studies. A clinical study of breast cancer patients, analyzed by Farewell (1986), presents a typical case. In this study, the time to relapse or death for three treatment arms of adjuvant therapy were observed from 139 patients, together with four other factors for each patient, including clinical stage, pathological stage, histological stage, and the number of lymph nodes having disease involvement. The Kaplan–Meier survival curves of patients from the three treatment groups are given in Figure 1. Note that all curves level off above 0.4 and there are a number of long-term censored observations at the tails of these curves, which correspond to patients who may potentially be cured in each of these groups. The problems of interest in this study include what the proportion of patients who may be cured is and what effects the treatment methods and other factors may have on the cure rate and on the failure time of uncured patients.

Mixture models have been employed to analyze clinical trials with potentially cured patients. Let T be a nonnegative random variable denoting the failure time of interest and $f(t \mid \mathbf{x}, \mathbf{z})$ and $S(t \mid \mathbf{x}, \mathbf{z})$ be the probability density function and survival function of T, respectively, where $\mathbf{x}$ and $\mathbf{z}$ are observed values of two covariate vectors on which the distribution of T may depend. In the mixture models, the distribution of T has a finite mixture, $f(t \mid \mathbf{x}, \mathbf{z}) = \pi(\mathbf{z}) f_u(t \mid \mathbf{x})$ and $S(t \mid \mathbf{x}, \mathbf{z}) = \pi(\mathbf{z}) S_u(t \mid \mathbf{x}) + 1 - \pi(\mathbf{z})$, where $\pi(\mathbf{z})$ is the proportion of uncured patients, which may depend on $\mathbf{z}$ by the logistic form $\log[\pi(\mathbf{z})/\{1 - \pi(\mathbf{z})\}] = \boldsymbol{\gamma}'\mathbf{z}$, $f_u(t \mid \mathbf{x})$ and $S_u(t \mid \mathbf{x})$ are the density and survival functions of the failure time distribution of uncured patients, respectively, which may depend on $\mathbf{x}$. The density and survival functions of cured patients are set equal to zero and one, respectively, for all finite value of t because cured patients will never experience a relapse or death due to the disease. Therefore, their failure times can be conveniently defined as infinite. The specification of $f_u(t \mid \mathbf{x})$ or $S_u(t \mid \mathbf{x})$ can be parametric or nonparametric, which will lead to parametric and nonparametric mixture models.

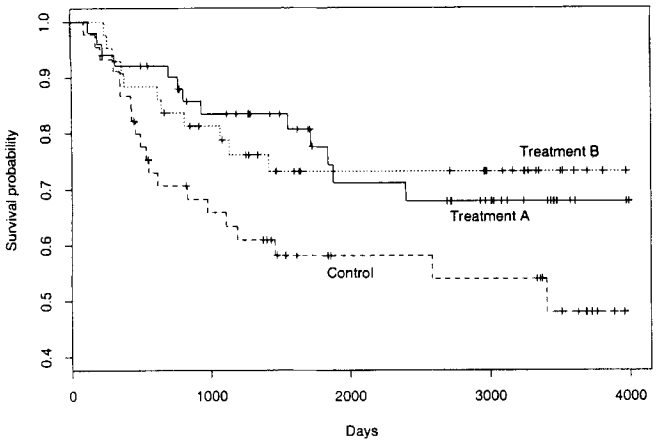

Figure 1. Relapse-free survival curves of breast cancer patients in three treatment groups.

In parametric mixture models, a particular distribution is assumed as the failure time distribution of uncured patients. The density function $f_u(t \mid \mathbf{x})$ and survival function $S_u(t \mid \mathbf{x})$ are derived from the distribution, which may also depend on one or more parameters. (Discussions of parametric mixture models can be found in Boag [1949], Jones et al. [1981], Farewell [1982, 1986], Cantor and Shuster [1992], Ghitany, Maller, and Zhou [1994], and Denham et al. [1996].) A problem with these parametric models is that it is difficult to verify the distributional assumptions used in the models. More general distributions, such as the extended generalized gamma and the generalized F, are proposed to reduce their parametric constraints (Yamaguchi, 1992; Peng, Dear, and Denham, 1998.) However, a nonparametric method may also be used to relax the parametric constraints.

Unlike parametric mixture models, nonparametric mixture models for the cure rate estimation have been under study only in recent years. Kuk and Chen (1992) applied a proportional hazards (PH) assumption to the failure time distribution of uncured patients. Their method is analogous to that used in Cox's PH model and is nonparametric. However, results from their method depend on a Monte Carlo approximation of the likelihood function involved, which is inconvenient for routine use. Taylor (1995) employed the Kaplan–Meier survivor estimator to estimate the failure time distribution of uncured patients and the EM algorithm to estimate γ, but this model does not allow covariates in the failure time distribution of uncured patients.

In this paper, we will investigate the nonparametric mixture models further and compare their performance with parametric mixture models. A new nonparametric estimation method will be proposed. In the subsequent sections, we will first outline the EM algorithm in mixture models for censored data. A PH mixture model is then proposed based on the EM algorithm. The model is similar in spirit to that of Kuk and Chen (1992), but the estimation method is different. The model and its estimation method provide a uniform framework under which other nonparametric mixture models for cure rate estimation can be derived. The proposed method is compared with existing nonparametric and parametric methods based on simulated and real data.

2. EM Algorithm for Mixture Models

Suppose we have data in the form $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, 2, \ldots, n$, where t_i denotes the observed survival time for the ith patient, δ_i is a censoring indicator with one if t_i is censored and zero otherwise and $\mathbf{x}_i$ and $\mathbf{z}_i$ are observed values of the two covariates. The likelihood function for the mixture model given in Section 1 is $\Pi_{i=1}^{n}[\pi(\mathbf{z}_i)f_u(t_i \mid \mathbf{x}_i)]^{\delta_i}[\pi(\mathbf{z}_i)S_u(t_i \mid \mathbf{x}_i) + 1 - \pi(\mathbf{z}_i)]^{1-\delta_i}$.

To use the EM algorithm to estimate unknown parameters in this mixture model, let c_i be an indicator of cure status of the ith patient, namely, c_i is one if the patient is cured and zero otherwise, $i = 1, 2, \ldots, n$. Obviously, if $\delta_i = 1$, then $c_i = 0$, but if $\delta_i = 0$, c_i is not observable and it can be one or zero. Note that $1 - \pi(\mathbf{z}_i) = P(c_i = 1 \mid \mathbf{z}_i)$. Let $\mathbf{c} = (c_1, \ldots, c_n)$. Therefore, $\mathbf{c}$ is partially missing information which will be employed in the EM algorithm.

Given c_i, i.e., the complete data are available, the complete log likelihood function is

$$L_{\mathbf{c}} = \log \prod_{i=1}^{n} \left[\{\pi(\mathbf{z}_i)f_u(t_i \mid \mathbf{x}_i)\}^{1-c_i}\right]^{\delta_i} \times \left[\{1 - \pi(\mathbf{z}_i)\}^{c_i}\{\pi(\mathbf{z}_i)S_u(t_i \mid \mathbf{x}_i)\}^{1-c_i}\right]^{1-\delta_i}. \quad (1)$$

The E-step in the EM algorithm calculates the expectation of (1) for given the current estimates of $f_u(t_i \mid \mathbf{x}_i)$, $S_u(t_i \mid \mathbf{x}_i)$ and $\pi(\mathbf{z}_i)$, which is the sum of following functions:

$$L_P = \sum_{i=1}^{n}[g_i \log \pi(\mathbf{z}_i) + (1 - g_i)\log\{1 - \pi(\mathbf{z}_i)\}] \quad (2)$$

$$L_S = \sum_{i=1}^{n}\{g_i \log S_u(t_i \mid \mathbf{x}_i) + \delta_i \log h_u(t_i \mid \mathbf{x}_i)\}, \quad (3)$$

where $h_u(\cdot) = f_u(\cdot)/S_u(\cdot)$ is the hazard function of the failure time distribution of uncured patients, g_i is the expectation of $1 - c_i$ conditional on the current estimates of γ and $S_u(t \mid \mathbf{x})$, given by

$$g_i = \delta_i + (1 - \delta_i)\pi(\mathbf{z}_i)S_u(t_i \mid \mathbf{x}_i) \div \{1 - \pi(\mathbf{z}_i) + \pi(\mathbf{z}_i)S_u(t_i \mid \mathbf{x}_i)\}, \quad (4)$$

which is the probability of the ith patient being uncured. Therefore, the E-step of the EM algorithm for this problem consists of assigning the probability g_i to each patient.

The M-step of the EM algorithm consists of maximizing (2) and (3) with respect to $f_u(\cdot)$, $S_u(\cdot)$ and γ for fixed g_i. The advantage of using EM algorithm here is that the maximum likelihood estimates of the failure time distribution of uncured patients and γ can be obtained separately because (2) only depends on γ while (3) only depends on the failure time distribution of uncured patients. Further details of the EM algorithm on general mixture models can be found in Larson and Dinse (1985). Equation (2) can be maximized by usual optimization methods such as the Newton-Raphson method. Maximizing (3) will depend on how the distribution of the failure time of uncured patients is specified and modeled by $\mathbf{x}_i$.

3. Proportional Hazards Mixture Model

The PH assumption provides a way to introduce covariates into models and to separate the effect of the covariates and the shape of a baseline hazard function. It has been successfully employed in Cox's PH regression model for survival data.

To model the effect of covariates $\mathbf{x}_i$ on the failure time distribution of uncured patients in the mixture model, we also employ the PH assumption, i.e., the effect of $\mathbf{x}_i$ on the distribution is modeled by $h_u(t_i \mid \mathbf{x}_i) = h_{u0}(t_i)\exp(\boldsymbol{\beta}'\mathbf{x}_i)$, where $h_{u0}(t)$ is the baseline hazard function, which can be any arbitrary unspecified hazard function but not a function of $\mathbf{x}_i$. This assumption implies $S_u(t_i \mid \mathbf{x}_i) = S_{u0}(t_i)^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}$, where $S_{u0}(t) = \exp\{-\int_0^t h_{u0}(w)dw\}$. Then equation (3) can be maximized after the baseline hazard $h_{u0}(t)$ is specified. A nonparametric estimation method for this model is proposed as follows.

3.1 *Estimation of $\boldsymbol{\beta}$ in the M-Step*

Let $\tau_1 < \cdots < \tau_k$ denote the distinct uncensored failure times. The set of d_j tied uncensored failures at τ_j is denoted by D_j. Let E_j be the set of individuals with censoring times in $[\tau_j, \tau_{j+1})$, $j = 0, \ldots, k$, where $\tau_0 = 0$ and $\tau_{k+1} = \infty$. R_j is the risk set at time τ_j, i.e., the set of individuals alive and uncensored just prior to τ_j. We first assume no ties among the uncensored observations and that $d_j = 1$, $j = 1, \ldots, k$. Let $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$ be covariates corresponding to $\tau_1, \ldots, \tau_k$. Following Kalbfleisch and Prentice (1973), (3) can be approximated by

$$\log \int_0^\infty \int_{\tau_1}^\infty \cdots \int_{\tau_{k-1}}^\infty \prod_{j=1}^k \left[h_{u0}(\tau_j) \exp\left\{ \boldsymbol{\beta}'\mathbf{x}_{(j)} - \left(\exp(\boldsymbol{\beta}'\mathbf{x}_{(j)}) + \sum_{i \in E_j} g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i) \right) \int_0^{\tau_j} h_{u0}(u)du \right\} \right] d\tau_k \cdots d\tau_1$$
$$= \log \prod_{j=1}^k \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{(j)})}{\sum_{i \in R_j} g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i)}.$$

If there are only a few ties, some adjustments are often very convenient to handle the ties. Following a method similar to that proposed by Breslow (1974), the adjusted likelihood function of the above is

$$\log \prod_{j=1}^k \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_j)}{\left\{ \sum_{i \in R_j} g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i) \right\}^{d_j}}, \tag{5}$$

where $\mathbf{s}_j = \Sigma_{i \in D_j} \mathbf{x}_i$ is the sum of the covariate vectors for individuals in D_j. If d_j, $j = 1, \ldots, k$, are substantially greater than one, the discrete nature of the failure time needs to be considered and the methods discussed by Cox and Oakes (1984, Section 7.6) may be used to derive a likelihood function.

Note that $g_i \equiv 1$ if all patients are uncured. In this case, (5) reduces to the usual likelihood function used in Cox's PH model. Therefore, maximizing (3) in the M-step now reduces to maximizing (5). It involves the regression parameter $\boldsymbol{\beta}$ but not the baseline hazard function, and it can be maximized by the Newton–Raphson method.

3.2 *Estimation of the Survival Function in the E-Step*

In the E-step, we update g_i by (4). This updating, however, involves the survival function $S_u(t_i \mid \mathbf{x}_i)$, which in turn involves the baseline survival function $S_{u0}(t)$ for given $\hat{\boldsymbol{\beta}}$. Therefore, the baseline distribution cannot be eliminated completely in the EM algorithm, and an estimation method of $S_{u0}(t)$ based on the current information is needed to complete the E-step.

The baseline survival function $S_{u0}(t)$ can be estimated by a method similar to that used in Cox's PH model. Following the argument of Kalbfleisch and Prentice (1980, p. 85), the contribution to the likelihood of a patient who fails at τ_j is $[S_{u0}(\tau_j)]^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)} - [S_{u0}(\tau_j + 0)]^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}$ and the contribution of a censored patient at time t is $[S_{u0}(t+0)]^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}$. The maximization is achieved by taking $S_{u0}(t+0) = S_{u0}(\tau_j + 0)$ for $t \in [\tau_j, \tau_{j+1})$, which effectively leads to a discrete baseline distribution with the probability mass only at the uncensored observations. Let $\alpha_j = S_{u0}(\tau_{j+1})/S_{u0}(\tau_j)$, $j = 0, \ldots, k$, then $\alpha_0 = 1$, $S_{u0}(t) = \Pi_{j:\tau_j < t} \alpha_j$, $S_{u0}(\tau_j) = \Pi_{l=1}^{j-1} \alpha_l$, and $S_{u0}(\tau_j + 0) = \Pi_{l=1}^{j} \alpha_l$. Then equation (3) can be rewritten as

$$\log \prod_{j=1}^k \left\{ \prod_{i \in D_j} \left(1 - \alpha_j^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right) \prod_{i \in R_j - D_j} \alpha_j^{g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i)} \right\},$$

from which the maximum likelihood estimate of α_j, given values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, can be obtained by

$$\hat{\alpha}_j \approx \exp\left(\frac{-d_j}{\sum_{i \in R_j} g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i)} \right)$$
$$\approx 1 - \frac{d_j}{\sum_{i \in R_j} g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i)}. \tag{6}$$

Under (6), the estimated baseline survival function $S_{u0}(t)$ is given by

$$\hat{S}_{u0}(t) = \exp\left(- \sum_{j:\tau_j < t} \frac{d_j}{\sum_{i \in R_j} g_i \exp(\boldsymbol{\beta}'\mathbf{x}_i)} \right).$$

Other methods using a piecewise constant parametric assumption or an approach suggested by Kalbfleisch and Prentice (1980, p. 79), as a referee pointed out, can also be employed to obtain an estimator of the survival function, which is a continuous function of time.

When $t > \tau_k$, $\hat{S}_{u0}(t)$ usually levels off if there is a censored time greater than τ_k, which is similar to the Kaplan–Meier estimator. Note that the leveling off in the Kaplan–Meier estimator is often employed as an indicator of the proportion of cured patients (Maller and Zhou, 1992), while $\hat{S}_{u0}(t)$ is the

baseline survival function of uncured patients. Therefore, it is reasonable to exclude the cured patients from $\hat{S}_{u0}(t)$ by forcing it to be zero beyond τ_k. This method can be justified for data with possible cured patients (Taylor, 1995).

It is easy to show that, when all patients are uncured and $g_i \equiv 1$, $i = 1, \ldots, n$, (6) reduces to the usual estimators for Cox's PH model. On the other hand, if no covariate affects the distribution of the failure time of uncured patients, (6) coincides with the estimates proposed in Taylor (1995), i.e., (6) is an "exact" solution of (3) when there is no covariate.

3.3 *Variances of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ in the EM Algorithm*

Variance estimates for the estimated parameters are not immediately available in the EM algorithm. As shown by Louis (1982), the variances of estimated parameters in the EM algorithm can be estimated by the difference between the conditional (conditional on the observed data) expectation of the complete-data observed information matrix and the conditional expectation of the square of the complete-data score function evaluated at the estimated values of the parameters.

For $\hat{\boldsymbol{\gamma}}$, the complete-data likelihood function corresponding to (2) is $\Sigma_{i=1}^{n} [(1 - c_i) \log \pi(\mathbf{z}_i) + c_i \log\{1 - \pi(\mathbf{z}_i)\}]$. It is not difficult to find the conditional expectation of its observed information matrix and of the square of its score function.

The computation of the variance of $\hat{\boldsymbol{\beta}}$ is not as straightforward as that of $\hat{\boldsymbol{\gamma}}$ because we maximize (5) instead of (3) with respect to $\boldsymbol{\beta}$ in the EM algorithm. If the complete-data likelihood function corresponding to (3) is used, the estimation of the variance of $\hat{\boldsymbol{\beta}}$ will be similar to that of $\hat{\boldsymbol{\gamma}}$. However, it may be appropriate to compute the variance based on the complete-data likelihood function corresponding to (5), which is not readily available, however. Heuristically, we can argue that an approximation to the complete-data likelihood function is given by

$$\log \prod_{j=1}^{k} [\exp(\boldsymbol{\beta}'\mathbf{s}_j)/\{\Sigma_{i \in R_j} (1 - c_i) \exp(\boldsymbol{\beta}'\mathbf{x}_i)\}^{d_j}].$$

Under this approximation, however, it is difficult to find out the conditional expectation of its observed information matrix and of the square of its score function. A simulation or multiple imputation approach may be employed to calculate them to obtain an approximation of the variance of $\hat{\boldsymbol{\beta}}$. This approach has been employed by other researchers.

3.4 *Kuk and Chen's Method*

Instead of applying the marginal likelihood approach to (3), Kuk and Chen (1992) applied it to the complete likelihood function (1) and obtained

$$L_{\mathbf{c}} \approx \log \prod_{i=1}^{n} \left[\pi(\mathbf{z}_i)^{1-c_i} \{1 - \pi(\mathbf{z}_i)\}^{c_i}\right] + \log \prod_{j=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{(j)})}{\left\{\sum_{i \in R_j} (1 - c_i) \exp(\boldsymbol{\beta}'\mathbf{x}_i)\right\}^{d_j}}$$

for a given $\mathbf{c}$ when there are no ties among uncensored observations. The full likelihood is then $L = \Sigma_{\mathbf{c} \in \Omega} L_{\mathbf{c}}$, where Ω is the set of all possible n-tuples of $\mathbf{c} = (c_1, \ldots, c_n)$, and the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be obtained from it.

Let n_c be the number of censored observations in the data. Because $c_i = 0$ if the ith patient is uncensored, the calculation of L involves 2^{n_c} terms, which is infeasible when n_c is large. A Monte Carlo method to approximate the likelihood was suggested and uses $2^{n_c} \Sigma_{j=1}^{r} L_{\mathbf{c}_j}/r$ to approximate the full likelihood, where $\mathbf{c}_j$, $j = 1, \ldots, r$, are r independent realizations of $\mathbf{c}$ in Ω.

Note that the baseline survival function $S_{u0}(t)$ is totally eliminated as a nuisance parameter in this method. At a price, the uniform distribution was employed to assign equal probability to each of 2^{n_c} n-tuples in Ω in the simulation, which is questionable because the probability depends on the survival function of uncured patients and the information of the survival function inherited from data is lost by using the uniform distribution.

4. Numerical Analysis

4.1 *Simulation Study*

We use simulation to verify the proposed estimation method for the parameters in the PH mixture model in Section 3 (referred to as the PH mixture model) and to compare it with Kuk and Chen's method (referred to as the K&C PH mixture model).

Following Kuk and Chen (1992), we generate a control group and a treatment group with 30 patients in each. The indicator of the treatment group is the only covariate involved. We set $\gamma_0 = 2$ and $\gamma_1 = -1$, which correspond to a cure rate of 11.9% in the control group and 26.9% in the treatment group. The standard exponential distribution is used as the baseline distribution for uncured patients in the control group. The parameter $\beta = \log(1/2) = -0.693$. The censoring times are generated according to an exponential distribution with parameter $\lambda = 0.28$ so that the expected censoring rate is 31% for the control group and 53% for the treatment group.

Under this setting, 500 samples are generated and the PH mixture model is applied to the samples. The averages, biases, and variances of $\hat{\beta}$, $\hat{\gamma}_0$, and $\hat{\gamma}_1$ are shown in Table 1. Results from the K&C PH mixture model are also included in Table 1. It can be seen that the two methods of estimation are comparable to each other in this simulated case.

We also simulated a case with a continuous covariate, and similarly there are no substantial differences in the results from the two models. Details are omitted.

4.2 *Application to Breast Cancer Data*

To make further comparisons, we apply the proposed method to the breast cancer data in Section 1. Farewell (1986) first applied the Weibull mixture model to the whole dataset to investigate the effects of the covariates on the failure time of uncured patients and the cure rate. Then Kuk and Chen (1992) considered a subset of the data with three covariates only: treatment, clinical stage I, and the number of lymph nodes having disease involvement. A graphical examination of the logarithm of cumulative hazard functions for these covariates suggests no substantial departure from the PH assumption. They applied the Weibull mixture model and the K&C PH mixture model to the subset, and results are included in Table 2.

The Weibull mixture model shows that the number of lymph nodes with disease involvement significantly increases the probability of relapse and quickens their occurrence. There is some evidence that clinical stage I reduces the probability of

Table 1
Averages, biases, and variances of estimates of regression parameters from the PH and K&C PH mixture models in the simulation study

	Method	$\hat{\beta}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
Average	PH mixture	−0.658	1.922	−1.094
	K&C PH mixture	−0.696	2.246	−0.933
Bias	PH mixture	−0.035	0.078	0.094
	K&C PH mixture	−0.003	0.246	0.067
Variance	PH mixture	0.213	0.520	1.092
	K&C PH mixture	0.212	0.565	0.916

relapse. But this effect on the delay of the occurrence of relapse is strong. Treatment B significantly reduces the probability of relapse relative to control but does not affect the time to its occurrence. Treatment A, on the other hand, does not influence the chance of relapse but does delay its occurrence.

Similar conclusions can be obtained from the K&C PH mixture model for treatment A and lymph nodes with disease involvement. However, there is no evidence that clinical stage I reduces the probability of relapse, and the effects of treatment B on the reduction of the probability of relapse and on the delay of its occurrence are opposite to those from the Weibull mixture model.

Results from the PH mixture model, given in the last column of Table 2, show strong similarity to those from the Weibull mixture model rather than to those from the K&C PH mixture model. Lymph nodes with disease involvement significantly increase the probability of relapse and quicken its occurrence. The significance for the latter is slightly weak, with a one-sided p-value of around 0.1. However, there is strong evidence for both that clinical stage I reduces the probability of relapse and that it delays the occurrence of the relapse. Relative to the control group, treatment A significantly delays the occurrence of relapse but does not affect the probability of relapse. On the other hand, treatment B reduces the probability of relapse significantly but does not influence the occurrence of relapse.

The similarity between the Weibull mixture model and the PH mixture model is also revealed by the close agreement between the estimated baseline survival functions from the two methods in Figure 2, contrary to the large disagreement between those from the Weibull mixture model and from the K&C PH mixture model.

The differences between three methods can also be examined by investigating the estimated cure rates from each of these models across all 12 groups. Table 3 provided these estimates, along with the estimates from the Kaplan–Meier estimator and the extended generalized gamma (EGG) mixture model. It can be seen that the cure rates from the PH mixture model are compatible with most of those from the Weibull mixture model, the Kaplan–Meier estimator, and the EGG mixture model, whereas the cure rates from the K&C PH mixture model are systematically low, i.e., the K&C PH

Table 2
Estimated parameters under different mixture models for the breast cancer data

	Parameters for uncured patients					
	Weibull mixture		K&C PH mixture		PH mixture	
	$\hat{\beta}$	$\hat{\beta}$/SE	$\hat{\beta}$	$\hat{\beta}$/SE	$\hat{\beta}$	$\hat{\beta}$/SE
Intercept	−9.568	−7.227	—	—	—	—
Treatment A	−1.419	−2.296	−1.113	−2.309	−0.916	−1.885
Treatment B	−0.045	−0.087	−0.981	−1.652	0.088	0.181
Clinical stage I	−1.226	−2.395	−1.284	−3.139	−0.845	−1.915
Lymph nodes	0.765	1.723	0.691	1.419	0.489	1.276
	Parameters for cure rate					
	Weibull mixture		K&C PH mixture		PH mixture	
	$\hat{\gamma}$	$\hat{\gamma}$/SE	$\hat{\gamma}$	$\hat{\gamma}$/SE	$\hat{\gamma}$	$\hat{\gamma}$/SE
Intercept	0.091	0.192	0.363	0.475	0.237	0.533
Treatment A	−0.089	−0.112	0.038	0.082	−0.634	−1.176
Treatment B	−1.019	−1.863	−0.255	−0.374	−1.141	−2.165
Clinical stage I	−0.680	−1.285	0.097	0.196	−0.919	−2.067
Lymph nodes	1.359	2.307	1.532	1.839	1.349	2.571

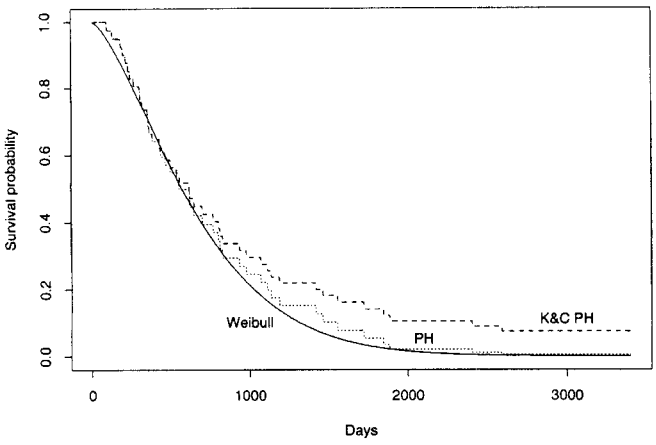

Figure 2. Estimated baseline survival functions from the Weibull, the PH, and the K&C PH mixture models for breast cancer data.

mixture model is always conservative in classifying patients as cured, and this is regarded as an advantage by Kuk and Chen (1992). Although the indistinguishability of the Weibull mixture model, which has been pointed out by Farewell (1986), may sometimes make it difficult to justify results from the Weibull mixture model, but we find little evidence from data to support this advantage.

When comparing their method with the Weibull mixture model, Kuk and Chen (1992) also questioned the goodness of fit of the Weibull mixture model. To assess the goodness of fit, we test the significance of the Weibull mixture model against the EGG mixture model (see Peng et al. [1998] for details of the test). The test gives a p-value of 0.006, which shows that the EGG mixture model significantly improves the goodness of fit of the Weibull mixture model. The estimated cure rates from the EGG mixture model, given in Table 3, however, support those from the PH mixture model instead of those from the K&C PH mixture model.

5. Conclusions and Discussion

In this paper, we set up a nonparametric mixture model for cure rate estimation problems and propose an estimation method for the model. The PH assumption is employed to include an analysis of covariate effects on the failure time of uncured patients. The estimation method is a combination of the marginal likelihood approach for Cox's PH model and the EM algorithm.

Kuk and Chen (1992) also proposed a mixture model combining logistic regression with Cox's PH model, and they employed the marginal likelihood approach and the EM algorithm as well in their estimation method. Unlike its application in Cox's PH mixture model and the K&C PH mixture model, the marginal likelihood approach in the PH mixture model cannot eliminate the baseline survival function in the EM algorithm. We estimated it and employed it to estimate the probability of a patient to be uncured in the E-step of the EM algorithm. Considering the dependence of the probability on the survival function of uncured patients, it is important that the information of the survival function be utilized in the estimation of the regression parameters in the mixture model. Ignoring this information in the estimation will lead to different estimates of regression parameters, as shown in the comparison between the K&C PH mixture model and the proposed method applied to the breast cancer dataset.

When there is no covariate considered for the failure time of uncured patients, the model and the estimation method proposed in this paper reduce to the model and the estimation method proposed by Taylor (1995). Therefore, we extend the work of Taylor (1995) to include a covariate analysis for uncured patients. When there is no cured fraction, the model and the estimation method reduce to Cox's PH model and its estimation method. Thus, we extend Cox's PH model to include a cure rate analysis for potential cured patients.

The multiple imputation method is employed to estimate the observed information matrix of regression parameters for the failure time of uncured patients. The estimation of regression parameters and their variances are verified by a simula-

Table 3
Estimated cure rates under five models for the breast cancer data

		At clinical stage I			Not at clinical stage I		
		Treatment			Treatment		
		Control	A	B	Control	A	B
<4 nodes	Weibull	0.641	0.649	0.838	0.467	0.477	0.718
	K&C PH	0.387	0.378	0.449	0.410	0.401	0.473
	PH	0.664	0.788	0.861	0.441	0.598	0.712
	Kaplan–Meier	0.661	0.805	0.923	0.508	0.676	0.644
	EGG	0.630	0.452	0.813	0.488	0.316	0.709
≥4 nodes	Weibull	0.336	0.344	0.594	0.199	0.205	0.419
	K&C PH	0.120	0.116	0.150	0.131	0.126	0.162
	PH	0.339	0.492	0.616	0.170	0.278	0.390
	Kaplan–Meier	0.200	1.000	0.625	0.200	0.190	0.667
	EGG	0.363	0.217	0.593	0.242	0.134	0.450

tion study. An estimation method for the standard error of the estimated baseline survival function for uncured patients at a given time and given values of covariates is still needed. Meanwhile, the bootstrap method for censored data may be employed.

NPCURE is a suite of C and S-PLUS functions designed to facilitate the cure rate analysis by the proposed nonparametric PH mixture model. The program is available on request from the first author.

Acknowledgements

This work was supported in part by research scholarships from the University of Newcastle and a research grant from Memorial University of Newfoundland to the first author. The authors thank Prof Vern T. Farewell at University College London for providing the breast cancer data in this research and the editor, associate editor, and two anonymous reviewers for constructive comments.

Résumé

En analyse des taux de guérison, on s'est moins intéressé aux méthodes non paramétriques qu'aux paramétriques. Nous étudions dans cet article un modèle général non paramétrique par mélange. La modélisation de l'effet des covariables sur la survie des patients non guéris repose sur l'hypothèse des risques proportionnels. L'estimation de ces effets est obtenue en maximisant la vraisemblance marginale à l'aide de l'algorithme EM et d'imputations multiples. Nous généralisons des modèles déjà proposés et nous améliorons les méthodes d'estimation. Ce modèle généralise également le modèle de Cox en soustrayant au risque une partie des patients et en étudiant l'effet des covariables sur la proportion concernée. Le modèle et la méthode d'estimation sont étudiés par simulations. Nous présentons une application en cancérologie. La comparaison de nos résultats à ceux des analyses précédentes confirme les conclusions obtenues à partir d'un modèle paramétrique mais non celles auxquelles conduisait le modèle non paramétrique antérieur.

References

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society* **11,** 15–53.

Breslow, N. E. (1974). Covariate analysis of censored survival data. *Biometrics* **30,** 89–99.

Cantor, A. B. and Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine* **11,** 931–937.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* New York: Chapman and Hall.

Denham, J. W., Denham, E. E., Dear, K. B. G., and Hudson, G. V. (1996). The follicular non-Hodgkin's lymphomas—I. The possibility of cure. *The European Journal of Cancer* **32A,** 470–479.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38,** 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *The Canadian Journal of Statistics* **14,** 257–262.

Ghitany, M. E., Maller, R. A., and Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis* **49,** 218–241.

Jones, D. R., Powles, R. L., Machin, D., and Sylvester, R. J. (1981). On estimating the proportion of cured patients in clinical studies. *Biometrie-Praximetrie* **21,** 1–11.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60,** 267–278.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Kuk, A. Y. C. and Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79,** 531–541.

Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of computing risks data. *Applied Statistics* **34,** 201–211.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* **79,** 731–739.

Peng, Y., Dear, K. B. G., and Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine* **17,** 813–830.

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

Yamaguchi, K. (1992). Accelerated 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* **87,** 284–292.

Received December 1997. *Revised June* 1999.
Accepted June 1999.