BIOMETRICS 56, 227–236
March 2000

Estimation in a Cox Proportional Hazards Cure Model

Judy P. Sy

Biostatistics, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, U.S.A.

and

Jeremy M. G. Taylor

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
email: jmgt@umich.edu

SUMMARY. Some failure time data come from a population that consists of some subjects who are susceptible to and others who are nonsusceptible to the event of interest. The data typically have heavy censoring at the end of the follow-up period, and a standard survival analysis would not always be appropriate. In such situations where there is good scientific or empirical evidence of a nonsusceptible population, the mixture or cure model can be used (Farewell, 1982, *Biometrics* **38,** 1041–1046). It assumes a binary distribution to model the incidence probability and a parametric failure time distribution to model the latency. Kuk and Chen (1992, *Biometrika* **79,** 531–541) extended the model by using Cox's proportional hazards regression for the latency. We develop maximum likelihood techniques for the joint estimation of the incidence and latency regression parameters in this model using the nonparametric form of the likelihood and an EM algorithm. A zero-tail constraint is used to reduce the near nonidentifiability of the problem. The inverse of the observed information matrix is used to compute the standard errors. A simulation study shows that the methods are competitive to the parametric methods under ideal conditions and are generally better when censoring from loss to follow-up is heavy. The methods are applied to a data set of tonsil cancer patients treated with radiation therapy.

KEY WORDS: Cure model; EM algorithm; Product-limit estimate; Profile likelihood.

1. Introduction

In survival analysis, it is usually assumed that if complete follow-up were possible for all individuals, each would eventually experience the event of interest. Sometimes, however, the data come from a population where a substantial proportion of the individuals do not experience the event at the end of the observation period. In some situations, some of these survivors are actually cured in the sense that, even after an extended follow-up, no further events are observed. An example is patients with tonsil cancer treated using radiation therapy (Withers et al., 1995), in which cure occurs if the radiation kills all the cancer cells. For such data, Kaplan–Meier (K-M) estimates of the survival function for time to local recurrence level off to nonzero proportions. Furthermore, there is a time window within which most or all of the recurrences are expected to occur, and so a patient without any recurrence beyond this window can usually be considered as being cured. This type of data is typical of diseases where the biology of the disease suggests the possibility of cure. A K-M survival curve that shows a long and stable plateau with heavy censoring at the tail may be taken as empirical evidence of a cured fraction. The use of standard survival analysis for such data may be inappropriate since not all the individuals are susceptible.

In a cure model, the population is a mixture of susceptible and nonsusceptible (cured) individuals. The objective is usually to study the cure rate and survival distribution and the effect of any covariates. We are interested in whether the event can occur, which we call incidence, and when it will occur, given that it can occur, which we call latency. In the tonsil cancer application, there are two types of covariates, patient variables such as tumor stage and treatment variables such as total dose of irradiation. How these covariates influence the cure rate would be viewed as most important, but there is also interest in how they relate to when the recurrence happens.

Various parametric and nonparametric methods have been proposed for the cure model. Farewell (1982) used logistic regression for the mixture proportion and a Weibull regression model for the latency. Peng, Dear, and Denham (1998) used a generalized F for the latency distribution. Kuk and Chen (1992) proposed a semiparametric generalization of Farewell's model using a Cox proportional hazards (PH) model in the susceptible group; we call this the PH cure model. They used an estimation method involving Monte Carlo simulation. In this paper, we present a different estimation method.

The cure model should not be used indiscriminately (Farewell, 1986). There must be good empirical and biological ev-

idence of a nonsusceptible population. The model generally requires long-term follow-up and large samples, and censoring from loss to follow-up during the period when events can occur must not be excessive. Otherwise, identifiability problems between the incidence and latency parameters can occur.

2. The Proportional Hazards Cure Model

Let Y be the indicator that an individual will eventually ($Y = 1$) or never ($Y = 0$) experience the event, with $p = \Pr(Y = 1)$. Let T denote the time to occurrence of the event, defined only when $Y = 1$, with density $f(t \mid Y = 1)$ and survival function $S(t \mid Y = 1)$. For a censored individual, Y is not observed. The marginal survival function of T is $S(t) = (1-p) + pS(t \mid Y = 1)$ for $t < \infty$. Note that $S(t) \to 1 - p$ as $t \to \infty$. We assume an independent, noninformative, random censoring model and that censoring is statistically independent of Y.

Farewell (1982) used a logistic regression model for the incidence $p(x) = \text{pr}(Y = 1; x) = \exp(x'b)/(1 + \exp(x'b))$, where the covariate vector x includes the intercept, and a parametric survival model for $S(t \mid Y = 1)$. Kuk and Chen (1992) generalized this by using a Cox PH model with hazard function $\lambda(t \mid Y = 1; z) = \lambda_0(t \mid Y = 1)\exp(z'\beta)$, where z is a vector of covariates other than the intercept and $\lambda_0(t \mid Y = 1)$ is the conditional baseline hazard function. Through b and β, the model is able to separate the covariates' effects on the incidence and the latency and, in that sense, provide a flexible class of models when there is *a priori* belief in a nonsusceptible group. For the Kuk and Chen model, the conditional cumulative hazard function is $\Lambda(t \mid Y = 1; z) = \Lambda_0(t \mid Y = 1)\exp(z'\beta)$, where $\Lambda_0(t \mid Y = 1; z) = \int_0^t \lambda_0(u \mid Y = 1)du$. The conditional survival function is $S(t \mid Y = 1; z) = S_0(t \mid Y = 1)^{\exp(z'\beta)}$, where $S_0(t \mid Y = 1)$ is the conditional baseline survival function.

It is not difficult to see that a mixture of PH functions is no longer proportional and in fact, for a binary covariate, a PH cure model can have marginal survival curves that cross. However, the standard PH model is a special case of a PH cure model in which $p(x) = 1$ for all x.

The PH cure model is a special case of a multiplicative frailty model, in which the hazard for an individual, conditional on Y, can be written as $\lambda(t \mid Y; z) = Y\lambda(t \mid Y = 1; z)$. As a frailty variable, Y is not entirely unobservable since an individual becomes labeled as $Y = 1$ if an event is observed.

3. Estimation

3.1 *Maximum Likelihood Estimation*

Denote the observed data for individual i by (t_i, δ_i, z_i), $i = 1, \ldots, n$, where t_i is the observed event or censoring time, $\delta_i = 1$ if t_i is uncensored and $\delta_i = 0$ otherwise, and z_i is a vector of covariates. For convenience, we let $x_i = (1, z_i')'$, although the covariates in x_i and z_i do not have to be identical. Denote the k distinct event times by $t_{(1)} < \cdots < t_{(k)}$. It follows that, if $\delta_i = 1$, $y_i = 1$ and, if $\delta_i = 0$, y_i is unobserved, where y_i is the value taken by the random variable Y_i. The likelihood contribution of individual i is $p_i f(t_i \mid Y = 1; z_i)$ for $\delta_i = 1$ and $(1 - p_i) + p_i S(t_i \mid Y = 1; z_i)$ for $\delta_i = 0$, where $p_i = \text{pr}(Y_i = 1; x_i)$. For the PH cure model, the observed full likelihood is then

$$
\begin{aligned}
&L(b, \beta, \Lambda_0) \\
&\quad = \prod_{i=1}^{n} \left\{ p_i \lambda_0(t_i \mid Y = 1) e^{z_i'\beta} e^{-\Lambda_0(t_i \mid Y=1)\exp(z_i'\beta)} \right\}^{\delta_i} \\
&\qquad \times \left\{ (1 - p_i) + p_i e^{-\Lambda_0(t_i \mid Y=1)\exp(z_i'\beta)} \right\}^{1-\delta_i}.
\end{aligned}
$$

We want to obtain the estimates $\hat{b}$ and $\hat{\beta}$ that maximize $L(b, \beta, \Lambda_0)$. In the ordinary Cox PH model, the standard analysis is to use the partial likelihood that does not depend on $\lambda_0(t)$. Breslow (1972) used a semiparametric full likelihood construction and a profile likelihood technique in which $\Lambda_0(t)$ is replaced in the full likelihood by a nonparametric maximum likelihood estimate (MLE) given β. The estimator for $\Lambda_0(t)$ is the Aalen–Nelson estimator. Breslow showed that this partially maximized likelihood function of β is proportional to the partial likelihood. This approach, however, does not work for the PH cure model. Unlike in the ordinary PH model where little information is lost by eliminating $S_0(t)$, one cannot eliminate $S_0(t \mid Y = 1)$ in the estimation without losing information about b. We propose a maximum likelihood-based method that makes use of the full likelihood.

3.2 *The EM Algorithm*

Denote the complete data by $(t_i, \delta_i, z_i, y_i)$, $i = 1, \ldots, n$, which includes the observed data and the unobserved y_i's. The complete-data full likelihood is

$$
\begin{aligned}
&L_C(b, \beta, \Lambda_0; y) \\
&\quad = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \prod_{i=1}^{n} \left\{ \lambda_0(t_i \mid Y = 1) e^{z_i'\beta} \right\}^{\delta_i y_i} \\
&\qquad \times e^{-y_i \Lambda_0(t_i \mid Y=1)\exp(z_i'\beta)} \\
&\quad = L_1(b; y) L_2(\beta, \Lambda_0; y), \qquad (1)
\end{aligned}
$$

where y is the vector of y_i values. The likelihood factors into a logistic and a PH component. We use the notation L for likelihoods and l for log-likelihoods.

The E step takes the expectation of $l_C(b, \beta, \Lambda_0; y)$ with respect to the distribution of the unobserved y_i's, given the current parameter values and the observed data O, where $O = \{\text{observed } y_i\text{'s}, (t_i, \delta_i, z_i); i = 1, \ldots, n\}$. Note that, for censored cases, the y_i's are linear terms in the complete data log-likelihood so that we only need to compute

$$
\begin{aligned}
&\pi_i^{(m)} \\
&\quad = \text{E}(Y_i \mid \theta^{(m)}, O) \\
&\quad = \text{pr}\left(Y_i = 1 \mid T_i > t_i, \delta_i = 0, z_i; \theta^{(m)} \right) \\
&\quad = \text{pr}(Y_i = 1; b) S_0(t_i \mid Y = 1)^{\exp(z_i'\beta)} \\
&\qquad \div [1 - \text{pr}(Y_i = 1; b) \\
&\qquad\quad + \text{pr}(Y_i = 1; b) S_0(t_i \mid Y = 1)^{\exp(z_i'\beta)}] \mid_{\theta = \theta^{(m)}},
\end{aligned}
$$

for censored cases, where $\theta = (b, \beta, \Lambda_0)$, $\theta^{(m)}$ denotes the current parameter values at the mth iteration, and $S_0(t_i \mid Y = 1) = \exp\{-\Lambda_0(t_i \mid Y = 1)\}$. For uncensored i, $\text{E}(Y_i \mid \theta^{(m)}, O) = y_i = 1$. Thus, the E step replaces the y_i's in (1) with $w_i^{(m)}$, which equals one if i is uncensored and equals $\pi_i^{(m)}$ if i is censored. Denote the expected log-likelihood by $\tilde{l}_C(b, \beta, \Lambda_0; w^{(m)}) = \tilde{l}_1(b; w^{(m)}) + \tilde{l}_2(\beta, \Lambda_0; w^{(m)})$, where $w^{(m)} = \{w_i^{(m)} : i = 1, \ldots, n\}$. Note that, for i censored, the

weight $w_i^{(m)}$ represents a fractional allocation to the susceptible group.

The M step of the algorithm involves the maximization of $\tilde{l}_C$ with respect to b and β and the function Λ_0, given $w^{(m)}$. To deal with the nuisance function $\Lambda_0(t \mid Y = 1)$ or $S_0(t \mid Y = 1)$, we perform an additional maximization step in the M step using profile likelihood techniques. Two methods from the Cox PH model can be extended: the Breslow-type estimator for $\Lambda_0(t \mid Y = 1)$ and the product-limit estimator for $S_0(t \mid Y = 1)$.

Breslow-type estimator (PFL approach). This profile likelihood method is based on Breslow's (1972) likelihood for the standard PH model (see also Klein, 1992). Using $\tilde{l}_2(\beta, \Lambda_0; w^{(m)})$, it can be shown that the nonparametric MLE of $\Lambda_0(t \mid Y = 1)$ given β is a slight modification of the Aalen–Nelson estimator given by

$$\tilde{\Lambda}_0(t \mid Y = 1) = \sum_{i:t_{(i)} \le t} \left(\frac{d_i}{\sum_{l \in R_i} w_l^{(m)} e^{z_l'\beta}} \right), \tag{2}$$

where d_i is the number of events at time $t_{(i)}$ and R_i is the risk set at time $t_{(i)}^-$. Substituting $\tilde{\Lambda}_0(t \mid Y = 1)$ into $\tilde{L}_2(\beta, \Lambda_0; w^{(m)})$ leads to a partial likelihood of β,

$$\tilde{L}_3(\beta; w^{(m)}) = \prod_{i=1}^{n} \left(\frac{e^{z_i'\beta}}{\sum_{l \in R_i} w_l^{(m)} e^{z_l'\beta}} \right)^{\delta_i}. \tag{3}$$

This is similar to the Cox partial likelihood except for the inclusion of the weights $w^{(m)}$. The M step involves maximizing $\tilde{L}_3$ with respect to β, given $w^{(m)}$. Then the PFL estimator for $S_0(t \mid Y = 1)$ is $\exp\{-\tilde{\Lambda}_0(t \mid Y = 1)\}$.

Product-limit estimator (NPL approach). This method is based on a nonparametric full likelihood construction that produces the generalized MLE for $S_0(t \mid Y = 1)$. Following the argument of Kalbfleisch and Prentice (1980, p. 85), the complete-data likelihood is

$$L_2(\beta, \alpha; y) = \prod_{i=0}^{k} \left[\prod_{l \in D_i} \left\{ h\left(t_{(i)}; z_l\right) S_0\left(t_{(i)}^- \mid Y = 1\right)^{\exp(z_l'\beta)} \right\} \times \prod_{l \in C_i} S_0(t_{(i)} \mid Y = 1)^{y_l \exp(z_l'\beta)} \right],$$

where a discrete PH model is assumed and $S_0(t \mid Y = 1)$ has the product-limit form $S_0(t \mid Y = 1) = \Pi_{j:t_{(j)} \le t} \alpha_j$, with $S_0(t_{(i)}^- \mid Y = 1) = S_0(t_{(i-1)} \mid Y = 1)$. The α's are nonnegative parameters at each of the k distinct event times with $\alpha_0 = 1$, $h(t_{(i)}; z) = 1 - \alpha_i^{\exp(z'\beta)}$ is the hazard function given z, D_i is the set of individuals experiencing an event at time $t_{(i)}$, and C_i is the set of individuals censored in $[t_{(i)}, t_{(i+1)})$, $i = 0, 1, \ldots, k$. Rearranging terms and applying the E step, we obtain

$$\tilde{L}_2\left(\beta, \alpha; w^{(m)}\right) = \prod_{i=1}^{k} \left[\prod_{l \in D_i} \left\{ 1 - \alpha_i^{\exp(z_l'\beta)} \right\} \prod_{l \in (R_i - D_i)} \alpha_i^{w_l^{(m)} \exp(z_l'\beta)} \right] \tag{4}$$

Given β, we obtain independent estimating equations for each α_i,

$$\sum_{l \in D_i} \left\{ \frac{e^{z_l'\beta}}{1 - \alpha_i^{\exp(z_l'\beta)}} \right\} = \sum_{l \in R_i} w_l^{(m)} e^{z_l'\beta}, \qquad i = 1, \ldots, k. \tag{5}$$

The solution for α_i is not of closed form except when there are no ties at $t_{(i)}$, in which case the MLE of α_i given β is

$$\tilde{\alpha}_i = \left(1 - \frac{e^{z_{(i)}'\beta}}{\sum_{l \in R_i} w_l^{(m)} e^{z_l'\beta}} \right)^{\exp(-z_{(i)}'\beta)}. \tag{6}$$

We then substitute $\tilde{\alpha}_i$ into $\tilde{L}_2(\beta, \alpha; w^{(m)})$ and obtain a nonparametric profile likelihood of β and obtain its MLE. But when there are ties, the MLEs for β and α must be jointly obtained from $\tilde{L}_2(\beta, \alpha; w^{(m)})$. This requires the maximization of a potentially very high-dimensional function.

Note that equations (2)–(6) all have analogous expressions in the standard PH model except for the inclusion of the weights $w^{(m)}$. Note also that, since $w_l^{(m)}$ depends on $S_0(t_l \mid Y = 1)$, the baseline function is involved in the estimation of b and β.

3.3 *Computational Aspects*

3.3.1 *Zero-tail constraint, $S_0(t_{(k)} \mid Y = 1) = 0$.* In order to obtain a good estimate for b and β, it is important for $\hat{S}_0(t_{(k)} \mid Y = 1)$ to approach zero, where $t_{(k)}$ is the last event time. Our numerical experience suggests that the PFL approach does not work well because of the inability of $\hat{S}_0(t_{(k)} \mid Y = 1) = \exp\{-\hat{\Lambda}_0(t \mid Y = 1)\}$ to go to zero even when the data indicate a leveling off of the marginal survival curve. In contrast, the NPL approach is able to send $\hat{S}_0(t_{(k)} \mid Y = 1)$ to zero because $\hat{\alpha}_k$ can equal zero; however, it is not guaranteed that this will occur at the global MLE. Thus, despite the simpler form for the estimate of β in the PFL method, we prefer and use the NPL approach in the rest of this paper.

Taylor (1995) suggested imposing the constraint $S_0(t_{(k)} \mid Y = 1) = 0$ in the special case of the PH mixture model with $\beta = 0$. The constraint occurs automatically when the weights $w_l^{(m)}$ for censored observations after $t_{(k)}$ are set to zero in the E step, essentially classifying them as nonsusceptible. The solution with this constraint has better statistical properties and converges faster than the unconstrained MLE. For most of the data sets in simulation studies, the constrained and unconstrained MLEs were identical. In some of the cases for which they differed, the unconstrained MLE was quite unstable. A heuristic justification for this constraint is that we would only consider the model in situations where it is clear that a nonsusceptible group exists and where there is sufficient follow-up beyond the time when most of the events occur. The need for the constraint becomes clear upon examining the observed log-likelihood surface. In

the PH mixture model without the constraint, even with sufficient follow-up, the surface is sometimes not well behaved, especially for b. With the constraint, the log-likelihood surface is well behaved and approximately quadratic.

3.3.2 *Maximization algorithm in the M step.* In the M step, we use a Newton–Raphson (NR) procedure to maximize $\tilde{l}_1(b; w^{(m)})$ to find $\hat{b}$. A simultaneous NR on (β, α) using $\tilde{l}_2(\beta, \alpha; w^{(m)})$ is, however, sensitive to starting values and will easily fail to converge. The method we found to be most efficient is the two-step NR suggested by Prentice and Gloeckler (1978) in the grouped PH model wherein the updates of β and α are obtained alternately. We use the parameterization $\lambda_i = -\log \alpha_i$. Our experience was that the above algorithm did converge reliably. The observed data log-likelihood increased with each EM iteration, and different starting values gave the same mode. It is possible for the estimate for b and/or β to be infinite. This happened very rarely and only when the sample size was small and there was a very small number of either events or survivors.

4. Standard Errors and Inference

We obtain an approximation of the asymptotic variance of $(\hat{b}, \hat{\beta})$ based on the inverse of the observed full information matrix $I(b, \beta, \lambda)$. Computations are based on the observed full likelihood parameterized according to the discrete PH model

$$L(b, \beta, \lambda) = \prod_{i=0}^{k} \left[\prod_{l \in D_i} p_l \left\{ 1 - e^{-\lambda_l \exp(z_l' \beta)} \right\} \times \exp\left(- \sum_{j: t_{(j)} \le t_{(i-1)}} \lambda_j e^{z_l' \beta} \right) \right] \times \left[\prod_{l \in C_i} \left\{ (1 - p_l) + p_l \exp\left(- \sum_{j: t_{(j)} \le t_{(i)}} \lambda_j e^{z_l' \beta} \right) \right\} \right]. \quad (7)$$

Formulas for $I(b, \beta, \lambda)$ are given in the Appendix. The submatrix of $I(b, \beta, \lambda)$ corresponding to λ is not diagonal, and there is no simple way to obtain $I(b, \beta, \lambda)^{-1}$ except to directly invert the entire matrix. This approach can handle ties and also provide standard errors for $\hat{\lambda}$.

The appropriateness of the standard errors of $(\hat{b}, \hat{\beta})$ might be of concern here. The asymptotic theory developed for the standard PH model using classical (Tsiatis, 1981) or counting process (Andersen and Gill, 1982) methods is based on estimators using the Cox partial likelihood. Bailey (1984) showed that the inverse of $I(\beta, \lambda)$ for the standard PH model based on the full likelihood provides an asymptotically correct and equivalent covariance matrix for $(\hat{\beta}, \hat{\lambda})$ to that based on the partial likelihood. The estimates for β and λ using both approaches are also asymptotically equivalent. This implies that the large number of nuisance parameters does not cause serious difficulty in the joint estimation. In our numerical work, we find that the joint estimation of (b, β) with λ does not give any problems in the properties of the estimators and inferential methods as long as the zero-tail constraint is used. Other methods of estimating the variance of $(\hat{b}, \hat{\beta})$ are possible (Sy and Taylor, unpublished manuscript).

5. A Simulation Study

5.1 *Simulation Design*

In this study, we compare the performance of the parametric MLEs with the proposed semiparametric estimators. Data are generated from a logistic-exponential mixture model, where $p(z) = 1/[1 + \exp\{-(b_0 + b_1 z)\}]$, $S(t \mid Y = 1; z) = \exp(-\lambda(z)t)$, $\lambda(z) = \exp(\beta_0 + \beta z)$. The covariate z is fixed by design and is uniform between -0.5 and 0.5. Censoring times C are generated from an exponential distribution with censoring rate λ_c. Various configurations of $p(z)$, $\lambda(z)$, and λ_c are considered. For each configuration, 100 data sets are generated, each with sample size of either $n = 50$ or 100. Each observation is followed up for at most $\tau = 10$. The data for each observation are (t, δ, z), where $t = \min(T, C, 10)$ is the observed time. Maximum likelihood estimation is carried out for the Weibull and PH mixture models, with the constraint $S_0(t_{(k)} \mid Y = 1) = 0$ for the latter.

The proportion $p(z)$ is (0.35, 0.65), 0.50, 0.20, or 0.80, where the first set denotes the range of $p(z)$ on z from -0.5 to 0.5 and the rest have $b_1 = 0$. The hazard $\lambda(z)$ is (0.5, 1.5), 1.0, or (1.5, 0.5), where the first and third denote the range of $\lambda(z)$ on z from -0.5 to 0.5 while for the second $\beta = 0$. λ_c is either 0.10 or 0.40, representing mild or heavy censoring, respectively.

Fourteen configurations are considered for each sample size. The designs are grouped into three general groups: (A) intermediate $p(z)$ designs—mild censoring, (B) low/high $p(z)$ designs—mild censoring, and (C) intermediate $p(z)$ designs—heavy censoring. We consider the bias, variability, and mean square error (MSE) of $\hat{b}_0$, $\hat{b}_1$, $\hat{\beta}$, $\hat{p}(z)$, $\hat{S}_0(t \mid Y = 1)$, and $\hat{S}_0(t)$. For $\hat{b}_0$, $\hat{b}_1$, and $\hat{\beta}$, we use the median and the square of the median absolute deviation from the median (MAD) to compute the bias and variability. The MSE is replaced by bias^2 + MAD^2. The bias of the mean, variance, and MSE are used for $\hat{p}(z)$, $\hat{S}_0(t \mid Y = 1)$, and $\hat{S}_0(t)$. For $\hat{p}(z)$, we consider estimates at $z = -0.5, 0, 0.5$. For $\hat{S}_0(t \mid Y = 1)$ and $\hat{S}_0(t)$, we consider estimates at the 5th, 50th, and 95th percentiles of the true $S_0(t)$. We also compare the coverage rates of the normal approximation 95% confidence intervals for b_0, b_1, and β between the PH and Weibull mixture models. To evaluate the adequacy of the estimated variances for $\hat{b}_0$, $\hat{b}_1$, and $\hat{\beta}$, we compare the median estimated variance with the observed variance of the estimates.

5.2 *Simulation Results*

Neither the PH nor the Weibull mixture model showed any significant bias, nor did one model show consistently larger bias than the other. The bias generally contributed little to the MSE. Table 1 gives the mean of the relative MSE (MSE(Weibull)/MSE (PH)) for $\hat{b}_0$, $\hat{b}_1$, $\hat{\beta}$, $\hat{p}(z)$, $\hat{S}_0(t \mid Y = 1)$, and $\hat{S}_0(t)$ over designs according to the groups A, B, and C. For designs with mild censoring, the Weibull and PH mixture models are generally comparable for $\hat{b}_0$ and $\hat{b}_1$. For $\hat{\beta}$, the PH mixture is generally less efficient, which is to be expected since the true model is Weibull. When censoring is heavy, the PH mixture generally does better for all the parameters. For $\hat{p}(z)$, the Weibull is less efficient for designs with small $\hat{p}(z)$ or heavy censoring. For $\hat{S}_0(t \mid Y = 1)$ and $\hat{S}_0(t)$, the PH mixture model is mostly less efficient, but it improves at higher percentiles for designs with heavy censoring.

Table 1
Mean of the relative MSE (MSE(Weibull)/MSE(PH)) of $\hat{b}_0$, $\hat{b}_1$, $\hat{\beta}$, $\hat{p}(z)$, $\hat{S}_0(t \mid Y = 1)$, *and* $\hat{S}_0(t)$ *at percentiles comparing Weibull to PH over designs*

	Mild censoring	Mild censoring		Heavy censoring
	A. Intermediate $p(z)$	B. Low $p(z)$	B. High $p(z)$	C. Intermediate $p(z)$
Number of designs	10	4	4	10
$\hat{b}_0$	1.05	1.03	1.09	1.22
$\hat{b}_1$	1.00	1.09	1.02	1.24
$\hat{\beta}$	0.97	0.87	0.94	1.11
$\hat{p}(-0.5)$	1.03	1.88	1.01	1.41
$\hat{p}(0)$	1.06	1.31	1.03	1.53
$\hat{p}(0.5)$	1.03	1.16	1.05	1.21
$\hat{S}_0(t \mid Y = 1)$, 5th	0.46	0.49	0.44	0.50
$\hat{S}_0(t \mid Y = 1)$, 50th	0.73	0.80	0.72	0.93
$\hat{S}_0(t \mid Y = 1)$, 95th	0.66	1.38	0.61	0.98
$\hat{S}_0(t)$, 5th	0.48	0.60	0.46	0.59
$\hat{S}_0(t)$, 50th	0.80	0.89	0.74	0.84
$\hat{S}_0(t)$, 95th	0.96	1.02	0.90	1.08

In summary, when the true model is a Weibull mixture, under ideal conditions of sufficient follow-up and mild censoring, the PH mixture is comparable in efficiency in the estimation of the incidence parameters but is less efficient for the parameters of the conditional survival distribution. But when censoring is heavy, the PH mixture model gets an upper hand in the incidence parameters and at times even does better with the latency parameters because of the constraint.

The coverage rates (not shown) for b_0 and b_1 are reasonable and comparable between the PH and Weibull mixture. For β, there can be undercoverage for both models but more often with the Weibull, mostly for designs with heavy censoring or small $p(z)$. The undercoverage is due to underestimation in the variance of β.

Comparisons between the observed and estimated variability of $\hat{b}_0$, $\hat{b}_1$, and $\hat{\beta}$ are similar in the PH and Weibull mixture. The estimated variance of $\hat{b}_0$ and $\hat{b}_1$ generally agrees with the observed variance when censoring is mild but is much smaller when censoring is heavy. The estimated variance of $\hat{\beta}$ is generally less than the observed variance, and the discrepancy becomes greater for small $p(z)$ or heavy censoring.

We observed that, for low $p(z)$ when there are only a few events observed, the estimate $\hat{\beta}$ becomes unstable, and this gets reflected in a large estimated and observed variance. We also need a reasonable proportion of the nonsusceptibles to survive censoring to near the end of the follow-up period in order to get a reasonable estimate of the incidence proportion. We find that the Weibull mixture is more sensitive to the effects of small $p(z)$ and heavy censoring because it does not get the help the PH mixture model does from the constraint, which helps in stabilizing the tail of $\hat{S}_0(t \mid Y = 1)$ and improves the parameter estimates.

6. Radiation Therapy for Tonsil Cancer

The data consist of 672 patients from nine institutions worldwide (Withers et al., 1995). The subjects had squamous cell carcinoma of the tonsil and were treated with radiation during 1976–1985. The purpose of the study was to investigate the effects of different radiation treatment regimens on local cancer control. In this example, local recurrence is defined as the event and failure time is time from initial treatment to local recurrence. Six covariates are considered: T stage (categorical) with levels T1, T2, T3, T4; node status (binary), with level zero for having negative neck nodes (N0) and level one for having at least one positive node (N+); total dose (continuous); overall treatment duration (continuous); sex (binary); and age (continuous). All covariates are included in both parts of the model. *A priori,* we might expect the T stage, node status, and treatment variables to be more important for the incidence part of the model because incidence is directly determined by whether or not all the cancer cells are killed. If age is to be important, it may influence the latency part because the time to recurrence is determined by the growth rate of the surviving tumor cells, which is potentially determined by patient specific factors such as age. Of the 672 subjects, 206 had cancer recurring. The observed follow-up time ranged from 19 days to 14.5 years. The last three recurrences were at 4.9, 5.1, and 8.2 years from initial treatment. Of the 466 censored observations, 89 were censored after the last event and 126 between the last two events. A K-M plot for the whole data set has a level region beyond about 3 years, which together with the biology of this tumor is a clear indication of the appropriateness of a cure model. There were 170 distinct event times, 31 with ties, and the number of ties ranged from 2 to 5. The $\log\{-\log S(t)\}$ plots of the K-M survival curves against log-time for most covariates show approximately parallel curves, but not necessarily straight lines, across covariate levels, indicating that a standard PH model might provide a reasonably good fit to the observable marginal survival curves while a standard Weibull model might not.

Tests for the joint significance of each covariate on incidence and conditional latency, $(b_j, \beta_j) = 0$, in the PH mixture were performed using the likelihood ratio test (LRT) and the normal approximation Wald test. The results show that, except

Table 2
Results from the PH mixture model, Weibull mixture model, and standard Cox PH model

	PH mixture model		Weibull mixture model		Standard Cox PH model	
	Estimate	SE	Estimate	SE	Estimate	SE
Logistic Model						
Intercept	−0.181	1.03	−0.070	1.00		
T stage	—	38.9[ab]	—	43.1[ab]		
T2	0.852	0.357	0.816	0.352		
T3	1.655	0.345[a]	1.687	0.342[a]		
T4	2.198	0.430[a]	2.222	0.428[a]		
Node	0.355	0.204	0.402	0.198		
Total dose (Gray)	−0.077	0.018[a]	−0.079	0.018[a]		
Overall time (per 10 days)	0.463	0.127[a]	0.473	0.125[a]		
Sex: male	0.116	0.215	0.157	0.209		
Age (per 10 years)	0.134	0.097	0.117	0.092		
Survival Model						
Intercept	—	—	2.640	0.876	—	—
T stage	—	13.6[ab]	—	17.4[ab]	—	56.2[ab]
T2	−0.625	0.365	−0.697	0.336	0.572	0.309
T3	−0.108	0.352	−0.306	0.325	1.365	0.295[a]
T4	0.385	0.383	0.358	0.358	1.857	0.329[a]
Node	0.339	0.188	0.307	0.169	0.369	0.148
Total dose (Gray)	−0.005	0.014	−0.005	0.013	−0.049	0.011[a]
Overall time (per 10 days)	−0.007	0.078	−0.020	0.077	0.281	0.071[a]
Sex: male	0.065	0.198	−0.105	0.175	0.099	0.154
Age (per 10 years)	−0.303	0.097[a]	−0.323	0.091[a]	0.032	0.065
Shape (Weibull)	—	—	1.178	0.061	—	—

[a] p-value < 0.01.
[b] Wald χ^2 with 3 d.f.

for sex, all the covariates are significant in at least one effect. The LRT and Wald test agree quite well.

Table 2 gives the results for the PH and Weibull mixture models and the standard Cox PH model. T stage is significant on both incidence and latency. There is more probability of local recurrence with higher stage. The effect on recurrence time is, however, not monotonic. Node is marginally significant on both incidence and latency, with a higher recurrence rate and earlier recurrence times for those with positive nodes. Total dose and overall treatment duration are very significant on incidence but not on latency. A higher total dose lowers the risk of recurrence, while a longer duration results in a higher recurrence rate. Sex is neither significant on incidence nor latency. Age is not significant on incidence but is significant on latency. The positive estimate for b_j indicates a higher, though nonsignificant, recurrence rate for older patients, but the significant negative estimate for β_j indicates later recurrence times for older patients. This has a plausible biological explanation and is an example of marginal survival curves that cross. The results from the Weibull mixture model are mostly similar. The standard errors are not much different between the PH and Weibull mixture models. The results from the standard Cox PH model agree with those for the PH mixture model's global test for $(b_j, \beta_j) = 0$ except for age. The mixture model results in Table 2 use the zero-tail constraint. The parameter estimates obtained without this constraint are very similar, but not identical, to those in the table.

To compare the fits from the PH and Weibull mixture models, we construct curves for $\hat{S}(t \mid Y = 1; z)$ and $\hat{S}(t; z)$ for selected values of z. Figure 1 compares the estimated marginal survival curves and the estimated conditional survival curves for T stage and age. The fits from the PH and Weibull models are very similar for the marginal curves, while they can be different for the conditional curves, especially near the tails of the curves, although the ordering of the curves is the same. The estimated curves from PH mixture model for age at 80 years show a big jump at the tail. This is caused by the last event, which is a late event whose effect on the baseline conditional survival function is to shift it upward at the tail. Excluding this observation reduces the jump size significantly. The plots for age show how and where the curves cross and illustrate the opposite effects of age on incidence and latency. This explains why the standard PH model was not able to detect an effect of age.

7. Discussion

The cure model allows the possibility of some useful interpretations that are not available using a standard Cox PH

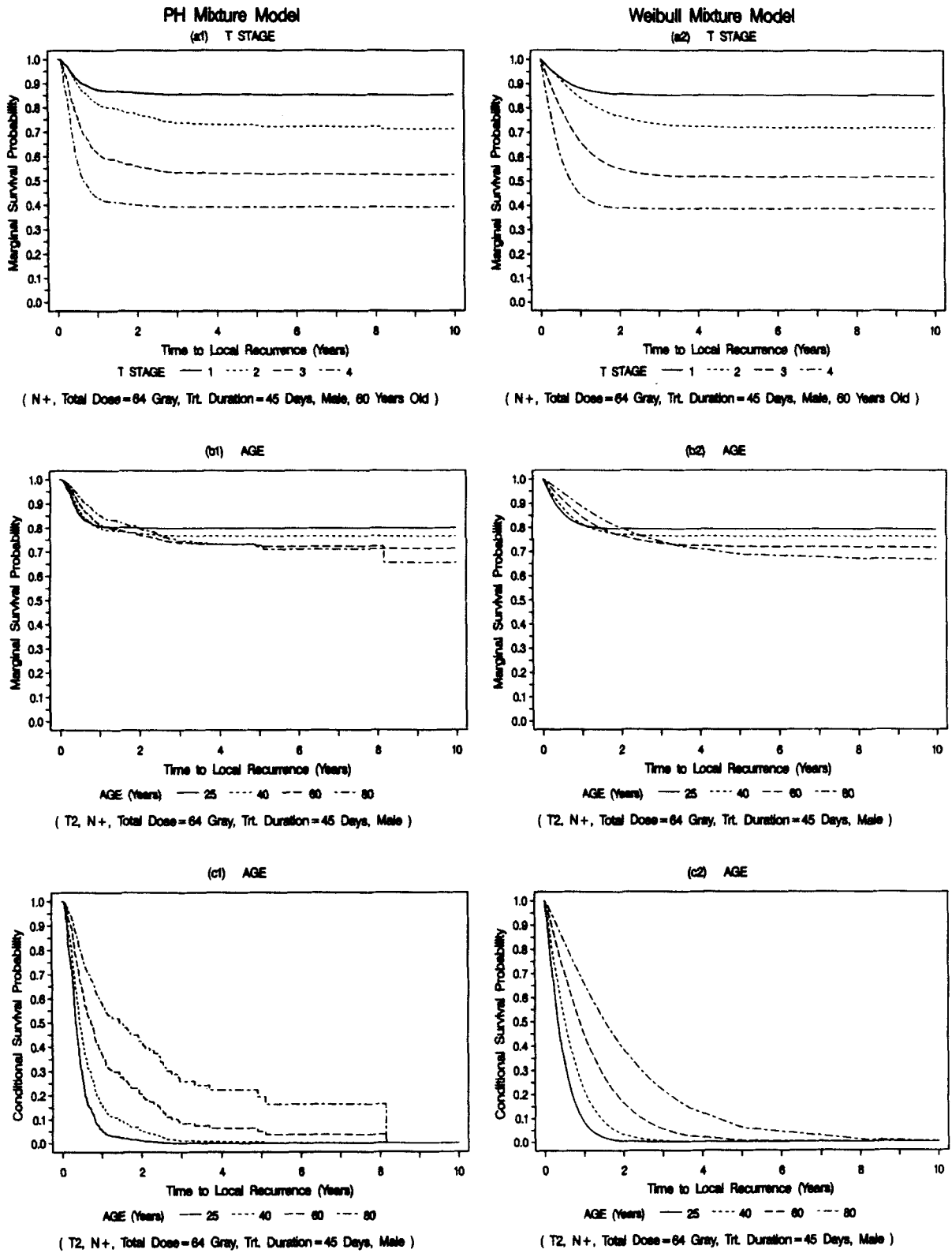

Figure 1. Estimated marginal survival curves from the PH mixture model and the Weibull mixture model for (a) T stage, (b) age, and (c) estimated conditional survival curves for age. The left panels are for the PH mixture model while the right panels are for the Weibull mixture model. For each covariate compared, the other covariates are fixed at selected values indicated in the plots.

model. For each covariate, there are two parameters, one describing how the covariate affects long-term incidence and one describing how it affects latency. Neither of these has the same interpretation as the relative hazard in a Cox model. A covariate that is important for incidence may not be important for latency and vice versa. In the tonsil cancer application, the eventual cure (i.e., incidence) is of more scientific interest than when the tumor recurred (i.e., latency). The data set also provides examples of variables that are important for incidence but not for latency (e.g., dose and treatment duration) and vice versa (e.g., age). This feature of the model that covariates can influence both the incidence and the latency allows more flexible modeling but also opens up the possibility of overparameterization. The relative importance of the two aspects of the model may differ between applications. In our experience, the incidence part of the model is usually of more scientific interest, in which case a minimal number or no covariates in the latency part may be appropriate.

An important issue in cure models is goodness-of-fit. It is possible that, even in a cure model, a marginal PH model may hold, although this did not happen for the tonsil cancer data because of the effect of age. There are a number of ways that one might address whether a cure model is appropriate, the most important of which is having a biological rationale from the underlying science. The standard PH model is a special case of the PH cure model, with an infinitely large intercept in the logistic part. For our data, the estimate and standard error for the intercept are -0.18 and 1.03, respectively, not supporting the standard PH model. For assessing the appropriateness of specific submodels, one could borrow ideas from standard approaches for these models. For example, for a major categorical variable, one could graphically examine a plot of $\log(-\log((\hat{S}(t) - (1 - \hat{p}))/\hat{p}))$ versus time for each level of the variable, where $\hat{S}(t)$ is the estimated marginal distribution and $\hat{p}$ is the estimated final level of the Kaplan–Meier curve. Approximately parallel lines would support the PH assumption for the conditional distribution.

An alternative potential approach to assessing the need for a cure model is to extend the ideas in Maller and Zhou (1996), who developed a method in the one-sample, no-covariate parametric model setting. They showed that a likelihood ratio test of the no-cure model versus cure model has, asymptotically, a mixture of chi-squared distributions. For the Kuk and Chen (1992) model, a likelihood ratio test could be constructed by fitting both the full model and a restricted model with $p(x) = 1$ for all x. Then an empirical estimate of the sampling distribution of the statistic could be obtained by simulating observations from the restricted model.

The semiparametric logistic-PH mixture model with covariates is identifiable for the parameters in the incidence probability and the conditional survival distribution. But by leaving the conditional baseline survival function arbitrary, a condition close to nonidentifiability can occur, which causes estimation problems. The constraint that sets the conditional survival function to zero beyond the last event time plays a crucial role in the procedure; it is effectively eliminating the near nonidentifiability, leading to a vastly more regular likelihood surface and good properties in the estimators. The constraint could also be viewed as an alteration of the model, e.g., to a model in which there is a finite time T_{max} after which events can never occur. Then our zero-tail constraint is implicitly estimating T_{max} by the largest event time. This is not unreasonable except in situations where there is poor follow-up beyond the period when events occur. Thus, this alteration to the model seems quite natural and appropriate to us except in situations where the cure model should not be used anyway.

The estimation method proposed in this paper is a generalization of the method of Taylor (1995) in his logistic K-M model where the conditional survival function does not depend on covariates. We simply set $\beta = 0$ and our $\hat{b}$ and $\hat{S}_0(t \mid Y = 1)$ reduce to his estimators. Statistical inference for b in Taylor (1995) was performed using likelihood ratio tests, and no standard errors were given. A special case of the information matrix given in the Appendix of the current paper can be used to find standard errors for the model considered by Taylor (1995).

Kuk and Chen (1992) in their estimation method for this model applied a marginal likelihood approach and eliminated $S_0(t \mid Y = 1)$ by simulating the Y values of the censored individuals. Their method effectively simulates $Y = 1$ or 0 with probability 1/2, ignoring the covariates and censoring times. They first maximize a Monte Carlo approximation of the marginal likelihood that is free of $S_0(t \mid Y = 1)$ to estimate (b, β). Given $(\hat{b}, \hat{\beta})$, $S_0(t \mid Y = 1)$ is then estimated using the nonparametric observed likelihood in an EM algorithm. The second step is similar to our method except that we jointly estimate (b, β) together with $S_0(t \mid Y = 1)$ within the same EM algorithm. We note that their method requires repeated application of the procedure in order to obtain the standard errors of the estimates. Furthermore, in their two-sample simulation study, it appears that their method tends to overestimate the incidence proportion for the group with more censored observations.

Acknowledgements

This research was performed when Drs Sy and Taylor were at the Department of Biostatistics, University of California–Los Angeles, and was supported by grants CA72495 and CA16042 from the National Cancer Institute.

Résumé

Certaines données de survie peuvent être issues d'une population comportant un mélange de sujets susceptibles et de sujets non susceptibles de développer un événement d'intérêt. Ces données se présentent typiquement avec un taux de censure élevé à la fin de l'étude et une analyse standard ne sera pas toujours adaptée. Dans de telles situations où l'on a de fortes présomptions scientifiques ou empiriques quant à l'existence d'une proportion de sujets non susceptibles, un modèle de mélange ou de guérison peut être utilisé (Farewell, 1982). Ce modèle assume une distribution de Bernouilli pour modéliser la probabilité d'être susceptible et un modèle paramétrique pour la distribution des temps d'événements. Kuk et Chen (1992) ont étendu ce modèle en utilisant un modèle semi-paramétrique de régression de Cox pour la distribution des délais de survenue de l'événement. Nous développons des procédures du maximum de vraisemblance, pour l'estimation jointe de l'incidence et des paramètres de régression relative à la distribution de temps d'événements, en utilisant une forme non paramétrique de la vraisemblance et un EM-algorithme. Une contrainte sur la fin de la distribution est utilisée pour

éviter un problème de non-identifiabilité. L'inverse de la matrice d'information observée est utilisée pour le calcul des écarts types. Une étude de simulation montre que cette méthode est compétitive par rapport aux méthodes paramétriques sous des conditions idéales, et est généralement plus performante quand la censure due aux perdus de vue est importante. Les méthodes sont appliquées à des données de patients atteints d'un cancer de l'amygdale et traités par radiothérapie.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10,** 1100–1120.

Bailey, K. R. (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model. *Annals of Statistics* **12,** 730–736.

Breslow, N. E. (1972). Contribution to the discussion of D. R. Cox (1972). *Journal of the Royal Statistical Society, Series B* **34,** 216–217.

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? *Canadian Journal of Statistics* **14,** 257–262.

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

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48,** 795–806.

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

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. (1996). *Survival Analysis with Long Term Survivors.* New York: Wiley.

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.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34,** 57–67.

Sy, J. P. and Taylor, J. M. G. (1999). Standard errors for the Cox proportional hazards curve model. *Mathematical and Computer Modelling,* in press.

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

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Annals of Statistics* **9,** 93–108.

Withers, H. R., Peters, L. J., Taylor, J. M. G., et al. (1995). Local control of carcinoma of the tonsil by radiation therapy: An analysis of patterns of fractionation in nine institutions. *International Journal of Radiation Oncology, Biology, Physics* **33,** 549–562.

Received March 1997. *Revised March* 1999.
Accepted April 1999.

Appendix

Observed Information Matrix

The first derivatives of the observed data log-likelihood can be obtained directly from (7). An easier, alternative way is to use the complete-data log-likelihood from the EM algorithm and derive the score function using the method of Louis (1982). Let

$$l(b,\beta,\lambda) = \log L(b,\beta,\lambda)$$

and

$$l_C(b,\beta,\lambda;y) = \log L_C(b,\beta,\lambda;y).$$

The components of the score function are

$$\frac{\partial}{\partial b} l(b,\beta,\lambda) = \frac{\partial}{\partial b} l_C(b,\beta,\lambda;y)\Big|_{y_l=w_l\forall l} = \sum_{l=1}^{n} x_l(w_l - p_l)$$

$$\frac{\partial}{\partial \beta} l(b,\beta,\lambda) = \frac{\partial}{\partial \beta} l_C(b,\beta,\lambda;y)\Big|_{y_l=w_l\forall l}$$

$$= \sum_{i=1}^{k}\left\{ \sum_{l\in D_i} \frac{z_l\lambda_i e^{z_l'\beta}}{1-e^{-\lambda_i \exp(z_l'\beta)}} - \lambda_i \sum_{l\in R_i} w_l z_l e^{z_l'\beta} \right\}$$

$$\frac{\partial}{\partial \lambda_i} l(b,\beta,\lambda) = \frac{\partial}{\partial \lambda_i} l_C(b,\beta,\lambda;y)\Big|_{y_l=w_l\forall l}$$

$$= \sum_{l\in D_i} \frac{e^{z_l'\beta}}{1-e^{-\lambda_i \exp(z_l'\beta)}} - \sum_{l\in R_i} w_l e^{z_l'\beta}, \qquad i = 1,\ldots,k,$$

where

$$w_l = \frac{p_l e^{-\Lambda_0(t_l|Y=1)\exp(z_l'\beta)}}{(1-p_l) + p_l e^{-\Lambda_0(t_l|Y=1)\exp(z_l'\beta)}}$$

for l censored and $w_l = 1$ for l uncensored and $\Lambda_0(t_l \mid Y = 1) = \Sigma_{j:t_{(j)}\le t_l}\lambda_j$. The observed information matrix $I(b,\beta,\lambda)$ has components

$$-\frac{\partial^2}{\partial b^2}\log L = \sum_{l=1}^{n} x_l x_l' p_l(1-p_l) - \sum_{l=1}^{n} x_l x_l' w_l(1-w_l)$$

$$-\frac{\partial^2}{\partial \beta^2}\log L = \sum_{i=1}^{k}\lambda_i \sum_{l\in R_i} w_l z_l z_l' e^{z_l'\beta} - \sum_{i=1}^{k}\sum_{l\in D_i} \frac{z_l z_l' h_{il}\left(1-e^{-h_{il}}-h_{il}e^{-h_{il}}\right)}{\left(1-e^{-h_{il}}\right)^2} - \sum_{l=1}^{n} w_l(1-w_l) z_l z_l' \left\{ e^{z_l'\beta}\Lambda_0(t_l \mid Y=1)\right\}^2$$

$$-\frac{\partial^2}{\partial \lambda_i^2}\log L = \sum_{l\in D_i} \frac{\left(e^{z_l'\beta}\right)^2 e^{-h_{il}}}{\left(1-e^{-h_{il}}\right)^2}$$

$$- \sum_{l \in R_i} w_l(1-w_l)\left(e^{z_l'\beta}\right)^2, \qquad i = 1, \ldots, k,$$

$$-\frac{\partial^2}{\partial \lambda_i \lambda_j} \log L = -\sum_{l=1}^{n} w_l(1-w_l)\left(e^{z_l'\beta}\right)^2 \times I\left(t_l \geq t_{(i)}\right) I\left(t_l \geq t_{(j)}\right), \qquad i \neq j,$$

$$-\frac{\partial^2}{\partial b \partial \beta} \log L = \sum_{l=1}^{n} w_l(1-w_l) x_l z_l' e^{z_l'\beta} \Lambda_0(t_l \mid Y=1)$$

$$-\frac{\partial^2}{\partial b \partial \lambda_i} \log L = \sum_{l \in R_i} w_l(1-w_l) x_l e^{z_l'\beta}, \qquad i = 1, \ldots, k,$$

$$-\frac{\partial^2}{\partial \beta \partial \lambda_i} \log L = \sum_{l \in R_i} w_l z_l e^{z_l'\beta} - \sum_{l \in D_i} \frac{z_l e^{z_l'\beta}\left(1 - e^{-h_{il}} - h_{il}e^{-h_{il}}\right)}{\left(1-e^{-h_{il}}\right)^2} - \sum_{l \in R_i} w_l(1-w_l) z_l \left(e^{z_l'\beta}\right)^2 \Lambda_0(t_l \mid Y=1), \qquad i = 1, \ldots, k,$$

where $h_{il} = \lambda_i \exp(z_l'\beta)$. Note that $-\partial^2 \log L/\partial b^2$, $-\partial^2 \log L/\partial \beta^2$, $-\partial^2 \log L/\partial \lambda^2$, and $-\partial^2 \log L/\partial \beta \partial \lambda_i$ have analogous expressions in the standard logistic regression and PH model except for negative terms for censored observations that involve $w_l(1-w_l)$, which reflects the variability in the estimated weights. When the constraint $S_0(t_{(k)} \mid Y=1) = 0$ is imposed, then $\alpha_k = 0$ and $\lambda_k = \infty$ and the dimension of λ is reduced to $k-1$.