To appear in the *Journal of Nonparametric Statistics*
Vol. 00, No. 00, Month 20XX, 1–24

Mixture cure rate models with accelerated failures and nonparametric form of covariate effects

Tianlei Chen[a] and Pang Du[b]*

[a]*Celgene Corporation, 86 Morris Avenue, Summit, New Jersey 07901, USA.;* [b]*Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, U.S.A.*

(*v1.0 submitted February 2017*)

Two-component mixture cure rate model is popular in cure rate data analysis with the proportional hazards and accelerated failure time models being the major competitors for modeling the latency component. Wang, Du, and Liang (2012) first proposed a nonparametric mixture cure rate model where the latency component assumes proportional hazards with nonparametric covariate effects in the relative risk. Here we consider a mixture cure rate model where the latency component assumes accelerated failure times with nonparametric covariate effects in the acceleration factor. Besides the more direct physical interpretation than the proportional hazards, our model has an additional scalar parameter which adds more complication to the computational algorithm as well as the asymptotic theory. We develop a penalized EM algorithm for estimation together with confidence intervals derived from the Louis formula. Asymptotic convergence rates of the parameter estimates are established. Simulations and the application to a melanoma study shows the advantages of our new method.

Keywords: accelerated failure time; confidence intervals; nonparametric covariate effects; penalized EM algorithm; two-component mixture cure rate model

AMS Subject Classification: 62G05; 62G08; 62N01; 62N02

*Corresponding author. Email: pangdu@vt.edu

1. Introduction

Due to modern developments in medical treatments and procedures, many diseases considered to be deadly decades ago, such as breast cancer, non-Hodgkin's lymphoma, and melanoma, have seen substantially improved cure rates recently; see, e.g., Tai et al. (2005). This motivates the proposal of a new type of survival data, called cure rate data, where the population consists of two groups of subjects, susceptible and non-susceptible individuals. All susceptible subjects would eventually experience the failure if there is no censoring, while non-susceptible subjects are not at risk of failure anymore and can be regarded as cured. A variety of statistical methods have thus developed to analyze cure rate data. Each cure model generally has two components: *incidence* which indicates whether the event could eventually occur and *latency* which denotes when the event will occur given the subject is susceptible to the event.

Existing cure rate models can be roughly divided into two categories. One category is promotion cure models first proposed in Yakovlev and Tsodikov (1996). Some recent developments in promotion cure models can be found in Zeng, Yin, and Ibrahim (2006) and references therein. The model we consider belongs to the other category of cure rate models called two-component mixture cure model. It assumes that the population is a mixture of two sub-populations and has a survival function

$$S_{\text{pop}}(t|\mathbf{x},\mathbf{z}) = \pi(\mathbf{z})S(t|\mathbf{x}) + 1 - \pi(\mathbf{z}), \tag{1}$$

where $\pi(\mathbf{z})$ and $S(t|\mathbf{x})$ are respectively the proportion and the survival function of susceptible subjects. Here $\mathbf{z}$ and $\mathbf{x}$ are the covariates associated respectively with π and S. They may overlap or even be identical. Some earlier developments include Berkson and Gage (1952) where π was simply an unknown constant and $S(t|\mathbf{x})$ assumed a parametric model, and the extension in Farewell (1982) with $\pi(\mathbf{z})$ assuming a logistic regression model. Since then, logistic regression has become the universal approach for modeling the proportion function $\pi(\mathbf{z})$ in the incidence component. On the other hand, the model choices for the survival function $S(t|\mathbf{x})$ in the latency component may vary with two distinguished competitors, namely the proportional hazards model and the accelerated failure time model.

An early example of the proportional hazards approach is Kuk and Chen (1992), who further extended the work of Farewell (1982) by considering a semiparametric Cox proportional hazards model for the hazard function of $S(t|\mathbf{x})$. They applied a marginal likelihood approach and used an estimation method involving Monte Carlo simulation. In Peng and Dear (2000) and Sy and Taylor (2000), the model was similar in spirit to that of Kuk and Chen (1992), but the estimation was implemented through an EM algorithm Peng (2003). The same model was considered in Corbière, Commenges, Taylor, and Joly (2009). They kept the parametric form of covariate effect in the relative risk of Cox model and used splines to model the baseline hazard function. A direct optimization procedure was proposed there to estimate the parameters. However, as noted by the inventor of the proportional hazards model, Sir D. R. Cox, biological interpretation of the proportional hazards assumption can be quite tricky. He further commented that "accelerated life models are in many ways more appealing" than the proportional hazards model "because of their quite direct physical interpretation"; see, e.g., Reid (1994) and Cox (1997).

Many recent mixture cure rate models focus on the accelerated failure time modeling of the survival component. Li and Taylor (2002) developed a mixture cure rate model where the latency component assumes an accelerated failure time regression model with unspecified error distribution. An EM algorithm is used for the estimation. The work of

Zhang and Peng (2007), Xu and Zhang (2010), and Zhang, Peng, and Li (2013) focused on improving the estimation procedure in Li and Taylor (2002) with the introduction of a rank-like estimating equation at the M-step. Lu and Ying (2004) proposed a class of semiparametric transformation models incorporating cure fractions, which included both the proportional hazards model and the accelerated failure time model as special cases. Othus, Li, and Tiwari (2009) extended their model to allow for time-dependent covariates and dependent censoring. More recently, Lu (2010) proposed an accelerated failure time model with cure fraction where the unknown error density was estimated by the kernel method.

All the aforementioned papers have a major common drawback: the covariate effects in both the incidence and the latency components assume a restrictive parametric form. For example, the logistic regression model for the incidence always assumes a linear form of covariate effects. For the latency, when the proportional hazard or accelerated failure time model is used, the logarithm of the relative risk or the acceleration factor is linear in covariates. In practice, such restrictive assumption on covariate effects may not hold and the analysis tools thus derived may not be valid. This limitation motivated the nonparametric method developed in Wang et al. (2012), where the covariate effects in both incidence and latency were modeled by smoothing spline ANOVA. In a smoothing spline ANOVA decomposition, a multivariate function is decomposed into sum of orthogonal components as main effects and interactions. In Wang et al. (2012), such functional ANOVA decomposition was applied to the mean function in a nonparametric logistic regression model of $\pi(\mathbf{z})$ and the log hazard function corresponding to survival function $S(t|\mathbf{x})$. To ensure model identifiability, they had to enforce a proportional hazards structure, although the relative risk part takes a flexible nonparametric form. Thus, their model naturally extends the traditional proportional hazard modeling of the latency component to allow for nonparametric form of covariate effects. On the other hand, it leaves the nonparametric extension of accelerated failure time modeling of latency an open problem.

Our method aims to filling in this gap by modeling the logarithm of the accelerator in the latency and the regression mean in the incidence with smoothing spline ANOVA, so the covariate effects now take more flexible nonparametric forms in both components. At the same time, our method also enjoys the more direct physical interpretation than the proportional hazards model in Wang et al. (2012).

Our smoothing spline function estimates are defined as the minimizer of a penalized likelihood, which consists of the negative log likelihood representing the goodness-of-fit, a roughness penalty enforcing smooth conditions, and a smoothing parameter balancing the tradeoff. Direct optimization of the penalized likelihood is as difficult as the optimization of the likelihoods in Peng and Dear (2000), Peng (2003), and Sy and Taylor (2000), where they used the EM algorithm to obtain the MLEs. However, the classical EM algorithm can only handle parameters of finite dimensions whereas our parameters are two smooth functions residing in infinite dimensional function spaces. So we need to extend the penalized EM (PEM) algorithm in Wang et al. (2012) to estimate the two function parameters in our model simultaneously. Similar to the EM algorithm, the PEM algorithm replaces the penalized likelihood by a more optimization-friendly penalized complete log likelihood with the introduction of a latent cure status variable y. The new objective functional consists of two penalized likelihoods, one involving only the parameter in the incidence component and the other involving only the parameters in the latency component. The E-step still evaluates the conditional expectations of y_i's given the current parameter estimates. The difference from the classical EM algorithm lies in the M-step where two penalized complete likelihoods are optimized instead of the complete likelihoods. The additional roughness penalties are necessary for enforcing smoothness on the function parameters. And the extra smoothing at the M-step has been proven to accelerate the convergence of the algorithm (Silverman, Jones, Wilson, and Nychka 1990). Furthermore, our latency model contains an additional shape parameter for the baseline distribution besides the function parameter for covariate effects. This nuisance parameter adds some extra complication to the algorithm when compared with the PEM algorithm in Wang et al. (2012). Besides the point estimators, we also derive confidence intervals for the parameters through an extension of the Louis formula (Louis 1982), a classical tool for computing observed information matrix when the EM algorithm is used. And we show that our nonparametric function estimates are consistent and their convergence rates are optimal for spline estimates.

As far as we know, this is the first mixture cure rate model where the latency component assumes an accelerated failure time model with a nonparametric form of covariate effects. It provides a nice complement to the mixture cure rate model in Wang et al. (2012) where the latency component has a nonparametric form of relative risk under the proportional hazards framework. In additional to its flexibility in modeling, our method offers smooth function estimates that are appealing to practitioners especially at the exploratory stage of data analysis. Our simulations demonstrate excellent performance of the proposed method in both estimation and inference. Our new method also shows some advantage when applied to the melanoma data studied in Wang et al. (2012)

The rest of the paper is organized as follows. Section 2.1 describes the nonparametric mixture cure rate model with the accelerated failure time latency component. Section 2.2 introduces the smoothing spline ANOVA framework. Section 2.3 presents the computational procedures for parameter estimation and inference. Section 2.4 studies the asymptotic properties of our estimates. Application to a melanoma study is in Section 3, and the simulation results in Section 4. Discussions in Section 5 concludes the paper.

2. Mixture Cure Rate Model with Nonparametric Accelerator

2.1. *The Model*

Let $(t_i, \delta_i, \mathbf{z}_i, \mathbf{x}_i)$ be the observed data for the ith subject, $i = 1, \cdots, n$. Here t_i is the observed lifetime time for the ith subject, δ_i is an indicator with $\delta_i = 1$ for observed failures and $\delta_i = 0$ for censored subjects, and $\mathbf{z}_i$ and $\mathbf{x}_i$ are the covariates associated with the incidence and latency components respectively. The covariates $\mathbf{z}$ and $\mathbf{x}$, though not necessarily so, can overlap with each other or even be the same. Note that all the cured subjects are censored and have $\delta_i = 0$, but that some censored subjects may eventually experience failures after the study. Assuming independent and non-informative censoring, the observed likelihood function can be written as

$$l_{obs}(\pi(\cdot), S(\cdot)) = \prod_{i=1}^{n} [\pi(\mathbf{z}_i) f(t_i|\mathbf{x}_i)]^{\delta_i} [\pi(\mathbf{z}_i) S(t_i|\mathbf{x}_i) + 1 - \pi(\mathbf{z}_i)]^{1-\delta_i}, \tag{2}$$

where $f(t|\mathbf{x})$ and $S(t|\mathbf{x})$ are respectively the probability density function and the survival function of failure time T given the covariate $\mathbf{x}$.

For the incidence component, we propose the nonparametric logistic regression model

$$\pi(\mathbf{z}) = \exp\{\zeta(\mathbf{z})\}/[1 + \exp\{\zeta(\mathbf{z})\}],$$

where $\zeta(\cdot)$ is an unknown smooth function. For the latency component $S(t|\mathbf{x})$, we consider the accelerated failure time (AFT) model

$$\log(T) = \eta(\mathbf{x}) + \epsilon,$$

where $\eta(\cdot)$ is an unknown smooth function and the error term ϵ follows an unknown distribution with zero mean. Let $\mathbf{x}_0$ be the covariate value that $\eta(\mathbf{x}_0) = 0$ and S_0 be the survival function of failure time T given that $\mathbf{x} = \mathbf{x}_0$. Then straightforward probability derivation yields $S(t|\mathbf{x}) = S_0(te^{-\eta(\mathbf{x})})$. Therefore, the covariate effect is to change the time scale by a factor $e^{-\eta(\mathbf{x})}$, either accelerating or decelerating the time. Note that the covariate effects in both the incidence and the latency models take nonparametric forms and thus allow more flexibility in model restrictions.

There are several common choices for the error distribution in the AFT model.

Example 2.1 (Extreme value and Weibull distributions) *When ϵ follows an extreme value distribution, T follows a Weibull distribution with survival function $S(t|\mathbf{x}) = \exp[-\{te^{-\eta(\mathbf{x})}\}^\tau]$ and hazard function $h(t|\mathbf{x}) = \tau t^{-1}\{te^{-\eta(\mathbf{x})}\}^\tau$, where τ is the shape parameter of the Weibull distribution. Note that the Weibull survival function also allows for a proportional hazards model interpretation.*

Example 2.2 (Normal and log normal distributions) *When ϵ follows a normal distribution, T follows a log normal distribution with survival function $S(t|\mathbf{x}) = 1 - \Phi(\log\{te^{-\eta(\mathbf{x})}\}^\tau)$ and hazard function $h(t|\mathbf{x}) =$ $\tau t^{-1}\phi(\log\{te^{-\eta(\mathbf{x})}\}^\tau)[1 - \Phi(\log\{te^{-\eta(\mathbf{x})}\}^\tau)]^{-1}$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are respectively the cumulative distribution function and probability density function of the standard normal distribution and $1/\tau$ is the standard deviation of ϵ.*

Example 2.3 (Logistic and log logistic distributions) *When ϵ follows a logistic distribution, T follows a log logistic distribution with survival function $S(t|\mathbf{x}) = [1+\{te^{-\eta(\mathbf{x})}\}^\tau]^{-1}$ and hazard function $h(t|\mathbf{x}) = \tau t^{-1}[1 + \{te^{-\eta(\mathbf{x})}\}^{-\tau}]^{-1}$, where $1/\tau$ is the scale parameter of the logistic distribution.*

From now on, we will write $S(t|\mathbf{x}) = S(t; \eta, \tau|\mathbf{x})$ and $h(t|\mathbf{x}) = h(t; \eta, \tau|\mathbf{x})$ to emphasize the dependence of S and h on τ and η. The mixture cure model (1) now becomes

$$S_{\text{pop}}(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S(t; \eta, \tau|\mathbf{x}) + 1 - \pi(\mathbf{z}). \tag{3}$$

We first show that this model is identifiable under a mild condition on S, whose proof is in Appendix A. Note that this condition is clearly satisfied by Examples 2.1-2.3.

Proposition 2.1 *Suppose that $S(t; \eta_1, \tau_1|\mathbf{x}) = S(t; \eta_2, \tau_2|\mathbf{x})$ implies $\eta_1(\cdot) = \eta_2(\cdot)$ and $\tau_1 = \tau_2$. Then the model (3) is identifiable.*

Rewriting the observed likelihood (2) in terms of ζ and η, the smoothing spline estimate of (ζ, η, τ) is simply the minimizer of the penalized likelihood

$$-\frac{1}{n}\log l_{\text{obs}}(\zeta, \eta, \tau) + \frac{\beta}{2}J_1(\zeta) + \frac{\lambda}{2}J_2(\eta), \tag{4}$$

where the first term is negative log likelihood representing the goodness-of-fit, J_1 and J_2 are roughness penalties enforcing certain levels of smoothness on the functions ζ and η, and $\beta, \lambda > 0$ are smoothing parameters controlling the tradeoff.

2.2. *Smoothing Spline ANOVA*

The minimization of the penalized likelihood (4) is carried out in a reproducing kernel Hilbert space (RKHS) of functions. In this paper, we use cubic and tensor product cubic smoothing splines for estimation whose detailed configurations involving the RKHS are given below.

Let $\mathcal{H} = \{\eta : J(\eta) < \infty\}$ be a RKHS on the domain $\mathcal{X}$ of covariate, where J is a square seminorm in $\mathcal{H}$ with a finite dimensional null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\} \subset \mathcal{H}$. Let $R(\cdot,\cdot)$ be the reproducing kernel (RK) of $\mathcal{H}$ such that R is a non-negative definite function satisfying $\langle R(\mathbf{x},\cdot), f(\cdot)\rangle = f(\mathbf{x})$, $\forall f \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$; the RK $R(\cdot,\cdot)$ and the space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ determine each other uniquely. Typically, $\langle\cdot,\cdot\rangle = J(\cdot,\cdot) + \tilde{J}(\cdot,\cdot)$, where $J(\cdot,\cdot)$ is the semi inner product associated with $J(\cdot)$ and $\tilde{J}(\cdot,\cdot)$ is an inner product in the null space $\mathcal{N}_J$ when restricted therein. There exists a tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where the space $\mathcal{H}_J$ has $J(\eta)$ as its square norm and an RK R_J satisfying $J(R_J(\mathbf{x},\cdot), f(\cdot)) = f(\mathbf{x})$, $\forall f \in \mathcal{H}_J$. See, e.g., Section 2.1 of Gu (2013).

The following examples give the configurations for the cases of a univariate continuous $\mathbf{x}$ and a bivariate $\mathbf{x}$ with one component continuous and the other discrete. When the covariate $\mathbf{x}$ has more dimensions, one can simply incorporate them by expanding the tensor product in Example 2.5.

Example 2.4 (Cubic Spline) *Without loss of generality assume $\mathcal{X} = [0,1]$ for a univariate x. A choice of $J(\eta)$ is $\int_0^1 (\eta'')^2 dx$, which yields the popular cubic splines. If the inner product in $\mathcal{N}_J$ is $(\int_0^1 f\,dx)(\int_0^1 g\,dx) + (\int_0^1 f'\,dx)(\int_0^1 g'\,dx)$, then $\mathcal{H}_J = \{\eta : \int_0^1 \eta dx = \int_0^1 \eta' dx = 0, J(\eta) < \infty\}$ and the reproducing kernel $R_J(x_1,x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|)$, where $k_\nu(x) = B_\nu(x)/\nu!$ are scaled Bernoulli polynomials for $x \in [0,1]$. The null space $\mathcal{N}_J$ has a basis $\{1, k_1(x)\}$ of $m = 2$ functions, where $k_1(x) = x - 0.5$ for $x \in [0,1]$. See Section 2.3.3 of Gu (2013).* □

Example 2.5 (Tensor Product Spline) *Consider a bivariate variable $\mathbf{x} = (z,u)$, where $z \in \mathcal{Z} = [0,1]$ and u is a categorical variable with l levels.*

We first look at the function space corresponding to u. Note that the domain of u is $\mathcal{U} = \{1,\ldots,l\}$. Functions on $\mathcal{U}$ are essentially vectors in $\mathbb{R}^l$, so the RKHS $\mathcal{H}_{\langle u\rangle} = \mathbb{R}^l$. When u is a nominal variable, that is, its levels are not ordered, let $\bar{\eta} = \sum_{u=1}^l \eta(u)/l$. Equipped with the roughness penalty $J_{\langle u\rangle}(\eta) = \sum_{u=1}^l [\eta(u) - \bar{\eta}]^2$ and inner product $\langle f,g\rangle = \sum_{u=1}^l f(u)g(u)$, the RKHS $\mathcal{H}_{\langle u\rangle}$ decomposes as

$$\mathcal{H}_{\langle u\rangle} = \mathcal{H}_{0\langle u\rangle} \oplus \mathcal{H}_{1\langle u\rangle} = \{\eta : \eta(1) = \cdots = \eta(l)\} \oplus \Big\{\eta : \sum_{u=1}^{l} \eta(u) = 0\Big\}$$

with reproducing kernels $R_{0\langle u\rangle}(u_1,u_2) = 1/l$, $R_{1\langle u\rangle}(u_1,u_2) = I_{[u_1=u_2]} - 1/l$.

On the other hand, the construction in Example 2.4 gives a decomposition of the RKHS $\mathcal{H}_{\langle z\rangle}$ on the domain $\mathcal{Z}$

$$\begin{aligned}\mathcal{H}_{\langle z\rangle} &= \big\{\eta : \textstyle\int_0^1 (\eta'')^2 dz < \infty\big\} = \mathcal{H}_{00\langle z\rangle} \oplus \mathcal{H}_{01\langle z\rangle} \oplus \mathcal{H}_{1\langle z\rangle} \\ &= span\{1\} \oplus span\{k_1(z)\} \oplus \big\{\eta : \textstyle\int_0^1 \eta dz = \int_0^1 \eta' dz = 0,\ \int_0^1 (\eta'')^2 dz < \infty\big\},\end{aligned}$$

with reproducing kernels $R_{00\langle z\rangle}(z_1,z_2) = 1$, $R_{01\langle z\rangle}(z_1,z_2) = k_1(z_1)k_1(z_2)$, and

$R_{1\langle z\rangle}(z_1,z_2)=k_2(z_1)k_2(z_2)-k_4(|z_1-z_2|)$. The tensor product of $\mathcal{H}_{\langle z\rangle}$ and $\mathcal{H}_{\langle u\rangle}$ yields six tensor sum terms $\mathcal{H}_{\nu,\mu}=\mathcal{H}_{\nu\langle z\rangle}\otimes\mathcal{H}_{\mu\langle u\rangle}$ on $\mathcal{Z}\times\mathcal{U}$, $\nu=00,01,1$ and $\mu=0,1$, with reproducing kernels $R_{\nu,\mu}(x_1,x_2)=R_\nu(z_1,z_2)R_\mu(u_1,u_2)$, where $x_i=(z_i,u_i)$. The two subspaces with $\nu=00,01$ are of one-dimension each and can be lumped together as the null space $\mathcal{N}_J$ (thus $m=2$). The other four subspaces form $\mathcal{H}_J$ with the reproducing kernel

$$R_J=\theta_{00,1}R_{00\langle z\rangle,1\langle u\rangle}+\theta_{01,1}R_{01\langle z\rangle,1\langle u\rangle}+\theta_{1,0}R_{1\langle z\rangle,0\langle u\rangle}+\theta_{1,1}R_{1\langle z\rangle,1\langle u\rangle},$$

where $\theta_{\nu,\mu}$ are a set of extra smoothing parameters adjusting the relative weights of the roughness of different components.

For interpretation, the six subspaces readily define an ANOVA decomposition

$$\eta(z,u)=\eta_\emptyset+\eta_u(u)+\eta_z(z)+\eta_{z,u}(z,u)$$

for functions on $\mathcal{X}$, with $\eta_\emptyset\in\mathcal{H}_{00\langle z\rangle}\otimes\mathcal{H}_{00\langle u\rangle}$ being the constant term, $\eta_u\in\mathcal{H}_{00\langle z\rangle}\otimes\mathcal{H}_{1\langle u\rangle}$ the u main effect, $\eta_z\in\{\mathcal{H}_{01\langle z\rangle}\oplus\mathcal{H}_{1\langle z\rangle}\}\otimes\mathcal{H}_{0\langle u\rangle}$ the z main effect, and $\eta_{z,u}\in\{\mathcal{H}_{01\langle z\rangle}\oplus\mathcal{H}_{1\langle z\rangle}\}\otimes\mathcal{H}_{1\langle u\rangle}$ the interaction. See, e.g., Example 2.7 of Gu (2013). □

Note that the optimization of the PL is carried out on the function space $\mathcal{H}$ and may be considered infeasible since $\mathcal{H}$ is of infinite dimensions. However, by the Representer Theorem (Wahba 1990), the minimizer of the PL in some settings actually resides in a finite dimensional subspace, namely, $\mathcal{H}_0\oplus\text{span}\{R_J(\mathbf{x}_1),\dots,R_J(\mathbf{x}_n)\}$. Here $\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ are observed values for $\mathbf{x}$ in the data and sometimes called the "knots" for smoothing splines. In many other settings, the minimizer in this subspace provides a sufficient approximation to the minimizer in $\mathcal{H}$ (Gu 2013). Furthermore, Kim and Gu (2004) showed that instead of having n knots, the number of knots can be reduced to the order of $n^{2/9}$ without losing any efficiency. They suggested using $10n^{2/9}$ knots in practice, which we shall follow in our computation.

For smoothing parameter (λ and θs) selection, cross-validation scores such as the generalized cross validation (GCV) score are often used in various problems (Wahba 1990; Gu 2013). As derived in the next section, our PL (4) will be reduced to PLs similar to some existing ones. Hence, we can directly borrow the cross validation scores in those problems. Our simulation results suggest they work fine in our problems too.

2.3. *Computation: Estimation and Inference Procedures*

In this section, we introduce a penalized EM algorithm for the optimization of (4) and extend the Louis formula to compute the confidence intervals for the parameters.

Direct optimization of (4) is difficult since the parameters (ζ,η,τ) are entangled with each other. Therefore we consider a penalized version of the EM algorithm here to break up (4) into separate objective functionals. Let y_i be the unobservable susceptible indicator for the ith subject. Given $\mathbf{y}=(y_1,\dots,y_n)$, the complete log likelihood decomposes as $L_c(\zeta,\eta,\tau;\mathbf{y})=L_1(\zeta;\mathbf{y})+L_2(\eta,\tau;\mathbf{y})$, where $L_1(\zeta;\mathbf{y})=\sum_{i=1}^n\left[y_i\zeta(\mathbf{z}_i)-\log\{1+e^{\zeta(\mathbf{z}_i)}\}\right]$ and $L_2(\eta,\tau;\mathbf{y})=\sum_{i=1}^n\left\{\delta_i\log h(t_i;\eta,\tau|\mathbf{x}_i)-y_i\int_0^{t_i}h(t;\eta,\tau|\mathbf{x}_i)dt\right\}$. Note that L_1 only involves ζ and L_2 only involves (η,τ). Hence separate optimization with respect to ζ and (η,τ) become feasible now.

The E-step computes the conditional expectation of L_c with respect to the latent variable y_i's given the current estimates $\Theta^{(m)}=(\zeta^{(m)},\eta^{(m)},\tau^{(m)})$. Since L_1 and L_2 are both linear in y_i's, their expectations are readily available with

$$y_i^{(m)} = E[y_i|\Theta^{(m)}] = \delta_i + (1-\delta_i)\frac{S(t_i;\eta,\tau|\mathbf{x}_i)}{\exp\{-\zeta(\mathbf{z}_i)\} + S(t_i;\eta,\tau|\mathbf{x}_i)}\bigg|_{\Theta^{(m)}}.$$

The M-step then minimizes two penalized likelihood functionals

$$\mathrm{PL}_1(\zeta|\mathbf{y}^{(m)}) \equiv -\frac{1}{n}L_1(\zeta;\mathbf{y}^{(m)}) + \frac{\beta}{2}J_1(\zeta)$$
$$\text{and } \mathrm{PL}_2(\eta,\tau|\mathbf{y}^{(m)}) \equiv -\frac{1}{n}L_2(\eta,\tau;\mathbf{y}^{(m)}) + \frac{\lambda}{2}J_2(\eta) \quad (5)$$

in their respective RKHSs $\mathcal{H}_\zeta = \{g : J_1(g) < \infty\}$ and $\mathcal{H}_\eta = \{k : J_2(k) < \infty\}$ to obtain $\Theta^{(m+1)} = \{\zeta^{(m+1)}, \eta^{(m+1)}, \tau^{(m+1)}\}$. Note that both PL_1 and PL_2 are convex for the distributions considered here. So the optimizations in the M-step can be handled by the standard Newton-Raphson procedure. The details are given in Appendix B. The selection of the smoothing parameters β and λ in PL_1 and PL_2 are respectively through the minimization of the cross-validation scores (5.28) and (8.27) in Gu (2013).

To obtain confidence intervals for the parameters $(\zeta(\mathbf{z}), \eta(\mathbf{x}), \tau)$, we write as in Appendix B $\zeta(\mathbf{z}) = \boldsymbol{\psi}_\zeta(\mathbf{z})^T\mathbf{b}_\zeta$ and $\eta(\mathbf{x}) = \boldsymbol{\psi}_\eta(\mathbf{x})^T\mathbf{b}_\eta$, where $\boldsymbol{\psi}_\zeta$, $\boldsymbol{\psi}_\eta$ are chosen spline basis functions and $\mathbf{b}_\zeta$, $\mathbf{b}_\eta$ are coefficient vectors. Recall that $\Theta = (\zeta, \eta, \tau)$, or essentially $(\mathbf{b}_\zeta, \mathbf{b}_\eta, \tau)$. The Louis formula for computing the observed information matrix is $I_{obs}(\Theta) = E_\Theta[B(\mathbf{y};\Theta)] - E_\Theta[G(\mathbf{y};\Theta)G(\mathbf{y};\Theta)^T]$. Here G and B are respectively the gradient vector and the negative second derivative matrix of the penalized complete log likelihood $L(\mathbf{y};(\zeta,\eta,\tau)) = L_1(\zeta;\mathbf{y}) + \frac{n\beta}{2}J(\zeta) + L_2(\eta,\tau;\mathbf{y}) + \frac{n\lambda}{2}J(\eta)$.

After obtaining I_{obs} following the steps in Appendix C, we can compute the $100(1-\alpha)\%$ confidence intervals of $\zeta(\mathbf{z}_0)$ and $\eta(\mathbf{x}_0)$ at given points $\mathbf{z}_0$ and $\mathbf{x}_0$, as well as the confidence interval of τ by

$$\begin{pmatrix}\widehat{\zeta}(\mathbf{z}_0)\\ \widehat{\eta}(\mathbf{x}_0)\\ \widehat{\tau}\end{pmatrix} \pm z_{\alpha/2}\mathrm{Diag}\left\{\begin{pmatrix}\boldsymbol{\psi}_\zeta(\mathbf{z}_0)^T & 0 & 0\\ 0 & \boldsymbol{\psi}_\eta(\mathbf{x}_0)^T & 0\\ 0 & 0 & 1\end{pmatrix} I_{obs}^{-1} \begin{pmatrix}\boldsymbol{\psi}_\zeta(\mathbf{z}_0) & 0 & 0\\ 0 & \boldsymbol{\psi}_\eta(\mathbf{x}_0) & 0\\ 0 & 0 & 1\end{pmatrix}\right\}.$$

where $(\widehat{\zeta}, \widehat{\eta}, \widehat{\tau})$ are the estimates obtained at the end of the PEM algorithm.

2.4. *Asymptotic Properties*

In this section, we will present the convergence rates of our function estimates $\widehat{\zeta}$ and $\widehat{\eta}$ as well as the consistency of the parameter estimate $\widehat{\tau}$. Its technical proof is in Appendix D. Let $\pi_0(\mathbf{z})$, $\zeta_0(\mathbf{z})$, $\eta_0(\mathbf{x})$ and τ_0 be the true parameters, and r_1 and r_2 be the constants associated with $\mathcal{H}_\zeta$ and $\mathcal{H}_\eta$ that measures the smoothness levels enforced by these two function spaces. A typical value for r_1 and r_2 are $2m$ when order-m splines are used for modeling ζ and η. Then we have the following theorem.

Theorem 2.1 *Under Conditions A1-A6, we have* $\|\widehat{\zeta}-\zeta_0\|_2^2 = O_p(n^{-r/(r+1)})$, $\|\widehat{\eta}-\eta_0\|_2^2 = O_p(n^{-r/(r+1)})$, $|\widehat{\tau}-\tau_0|^2 = O_p(n^{-r/(r+1)})$, *where* $\|\cdot\|_2$ *is the* L_2*-norm and* $r = \min(r_1, r_2)$.

Note that this is the optimal convergence rate of spline estimates when splines of order $r/2$ are used. If $r_1 = r_2$, both function estimates $\widehat{\zeta}$ and $\widehat{\eta}$ obtain their optimal convergence

Table 1. Melanoma Cancer Application: Cross-validation comparison of four mixture cure rate models. PH is the model in Wang et al. (2012) and the others are the proposed models with different distribution settings.

Model	PH	AFT Weibull	AFT Log normal	AFT Log logistic
Ave. log-likelihood	-280.97	-192.66	-227.48	-282.20

rates. Otherwise, only one of them can achieve the optimal rate, and the other cannot, suffering from entangled joint estimation. Also note that the result for $\widehat{\tau}$ can be refined to have a $\sqrt{n}$-rate and asymptotic normality using the technique in Section 21.5 of Kosorok (2008). But we choose not to pursue such refinement since τ_0 is considered a nuisance parameter here.

3. Analysis of Melanoma Cancer Data

We now apply the proposed method to a data set downloaded from the Surveillance Epidemiology and End Results (SEER) (www.seer.cancer.gov) database released in 2008. The dataset selected a total of 635 white patients from the nine registered metropolitan areas who met the following criteria: (1) melanoma was their first cancer diagnosis, (2) the cancer stage was classified as local or regional, and (3) the patient only received the routine treatment. The failure time of interest was time from diagnosis of melanoma to death from melanoma. A question of interest was whether survival or cure fractions differed in this data set by gender, tumor size and age. The covariates were age at diagnosis (range: 5 to 101 years), gender (M or F) and tumor size (Big or Small). The dataset was analyzed by Wang et al. (2012) using the mixture cure rate model with a proportional hazards form of latency. They presented plots of age-stratified Kaplan-Meier curves for four age groups, each of which showing a plateau at the end of the observation interval. This suggests the possible presence of a subpopulation of cured subjects in the study and justifies that a cure rate data analysis is appropriate.

We first conducted a 5-fold cross validation to determine the best model for the data among four candidates: the mixture cure rate model with proportional hazards in Wang et al. (2012) and the proposed models with three distribution settings, namely, Weibull, log-logistic, and log-normal. The data were randomly partitioned into 5 subsets of equal size. In each of the 5 rounds, we fixed one subset as the testing data, used the remaining subsets as the training data to fit the four methods with a full model specification including all the main and interaction effects for covariates $\mathbf{z} = \mathbf{x} =$ (age, gender, size), and then computed the log likelihood on the testing subset with their corresponding function estimates. The average log likelihoods for the four methods are shown in Table 1. Clearly, the proposed AFT model with the Weibull distribution setting has the most advantage. Note that the AFT model with Weibull distribution also belongs to the category of proportional hazards as pointed out in Example 2.1. However, it is actually difficult to pinpoint the exact forms of the baseline hazard or relative risk under the proportional hazards setting in Wang et al. (2012). So this application gave a good demonstration of the proposed AFT models in showing that the AFT model with Weibull distribution, or equivalently, the proportional hazards model with Weibull distribution, might work the best for the data. Hence we analyzed the whole data set using this model specification and report the results below.

The corresponding fits together with their point-wise confidence intervals, are plotted in Figures 1 and 2. In Figure 1, we see that the CIs for female group do not cover any constant lines and the CIs for male group can barely do so. This suggests a likely association of age with the non-cure rate. For male patients, the non-cure rate increased up to age 65 and then levels off. But for female patients, the non-cure rate showed a strong and consistently increasing trend against age. For female, the non-cure rates for both tumor size groups were comparable but the increase of the non-cure rate against age for the small-size group seemed to be steeper than the one for the big-size group. For male, the big-size group clearly showed a larger non-cure rate than the small-size group. It is also reassuring to see that these results match up well with the previous findings in Wang et al. (2012).

The estimated shape parameter τ for the Weibull distribution describing the non-cured subpopulation is 1.45 with an estimated standard deviation of 0.08. The 95% confidence interval for τ is $[1.29, 1.61]$. The estimated log of the scale parameter, $\eta(\mathbf{x})$ is illustrated in Figure 2. Note that the logarithm of hazard rate has a negative linear relationship to η, so we can interpret the trends of η in terms of the log hazard rate for non-cured patients. One interesting point is that all the curves more or less showed a turning point around age 65. For the log hazard rates of each individual patient group, both males and females with small tumors showed close-to-linear trend, with males having an increasing hazard as age increases and females showing an constant hazard through all ages. The two groups with big tumors also showed similar trends. Particularly, hazard decreased with age up to age 65 and then increased afterwards. The hazard of the female group with big tumors was lower at the early age than that of their male peers.

4. Empirical Studies

We now present some simulations to evaluate the estimation performance of the proposed method and the coverage properties of the confidence intervals.

We considered the following simulation setting of true parameters

$$\pi_0(z) = c_0 + 0.7\sin\{2(z+0.6)\}, \quad \eta_0(x) = \log\frac{2.5}{\{1+0.5\sin(2\pi x)\}^{2.5}}, \quad \tau_0 = 2,$$

where z and x are continuous covariates. Two failure time distributions, namely, Weibull and log normal distributions were considered. The constant c_0 in π_0 were chosen as $c_0 = 0.1722$ for Weibull distribution and $c_0 = -0.0278$ for log normal distribution to respectively yield 20% and 40% overall cure probabilities of all subjects. Two sample sizes $n = 400$ and 800 were studied but the results for $n = 800$ are not presented here due to space concern.

The covariates x were generated as a grid of 400 equally spaced values over the range [0,1]. For covariate z, a grid of 20 equally spaced values were generated over the range $[-0.4, 0.4]$ first and then each value was repeated 20 times to generate all the 400 covariate values for z. For each observation, it was assigned as not cured according to a Bernoulli trial with probability $\pi_0(z_k)$; then failure times were randomly generated for the non-cured observations from the specific failure time distribution: either the Weibull distribution (Example 2.1) or the log normal distribution (Example 2.2) with $\tau = 2$ and $\eta(x_k)$; finally, for all the observations, censoring times were generated from Weibull distributions and the censoring status indicators were recorded. Note that all the cured samples were recorded as being censored. The Weibull distributions for the censoring times were chosen in a way such that the observed censoring rate was about 45%. One hundred data replicates were generated for each setting. The point-wise 95% confidence

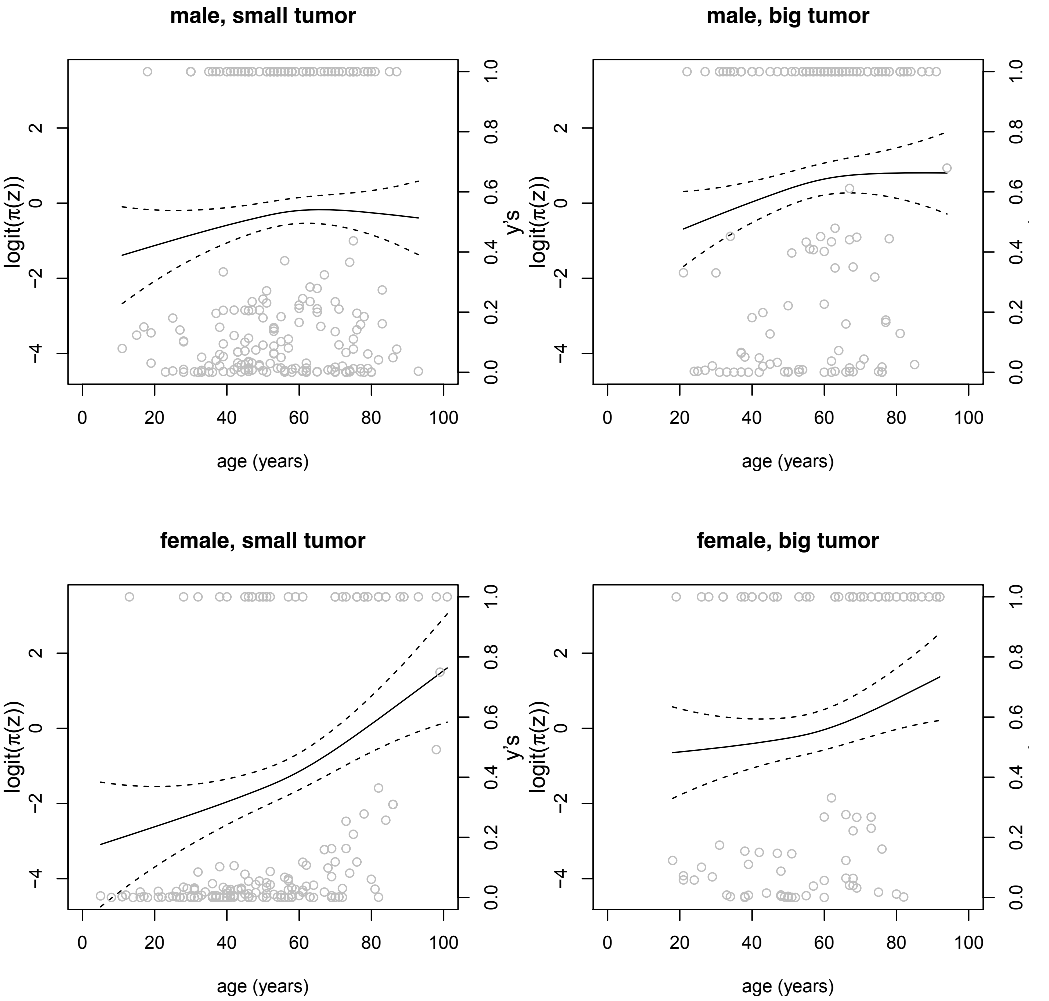

Figure 1. Melanoma Cancer Application: Estimated logit non-cure rates and their confidence intervals against age for the four patient groups determined by gender and tumor size (big or small). Superimposed are true data points with positions determined by age and converged $y's$.

intervals were calculated for $\zeta(z)$ on a z grid of size 100 equally spaced on $[-0.4, 0.4]$, for $\eta(x)$ on a x grid of size 100 equally-spaced grid points on $[0, 1]$.

Figures 3 and 4 respectively plot the simulation results for Weibull and log normal distributions. The top and bottom rows represent functions $\zeta(\cdot)$ and $\eta(\cdot)$ respectively. The left frames show point-wise coverage of the 95% interval estimates of the functions at the selected grid points, the right frames plot the true test functions (dash-dotted), the averages of point-wise function estimates (solid), the averages of point-wise 95% CIs (dashed), and the empirical 2.5% and 97.5% percentiles of point-wise function estimates (dotted). Also superimposed in the left frames are the magnitudes of the curvatures of the corresponding true curves.

For Figures 3 and 4, we first look at the plots for the η function. We can see that the mean estimates of η are close to the true functions and the mean interval estimates are close to the empirical percentiles of the 100 function estimates. The point-wise coverage

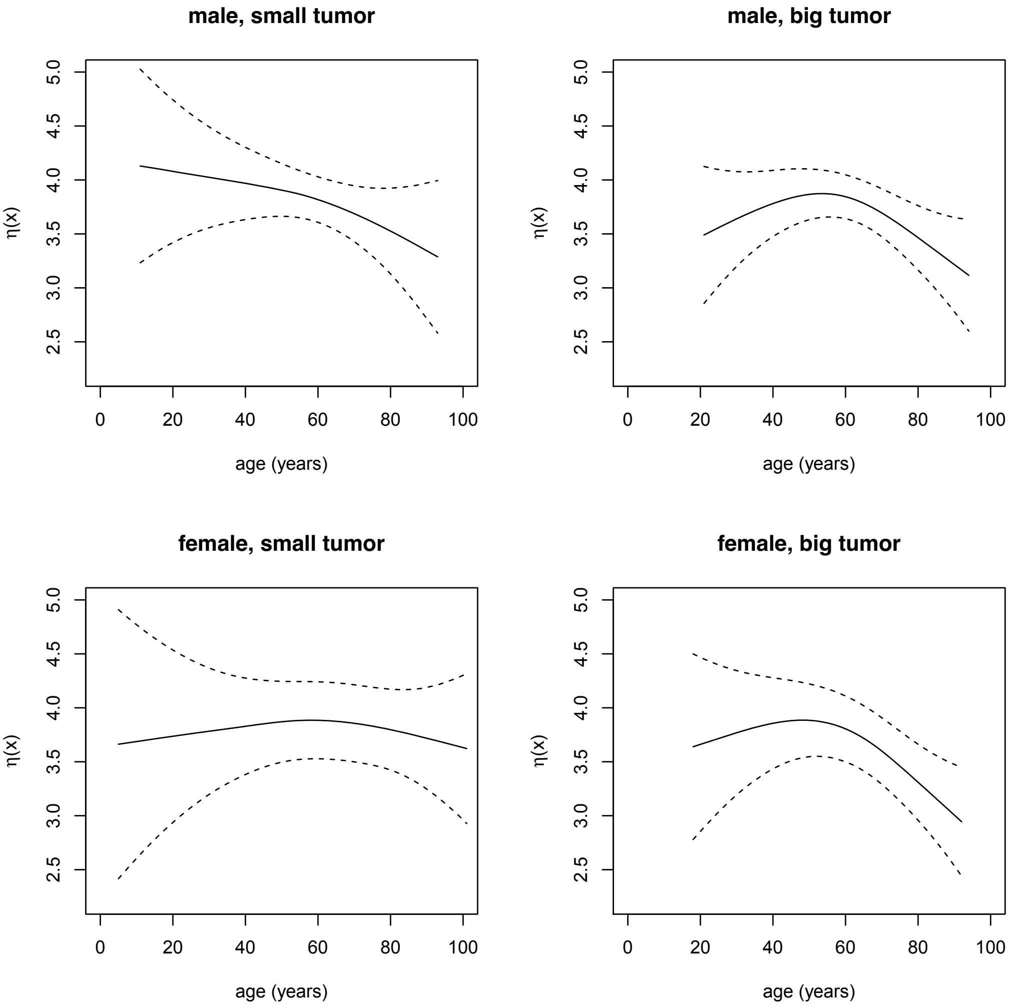

Figure 2. Melanoma Cancer Application: Estimated $\eta(\mathbf{x})$ and confidence intervals against age.

is generally close to the nominal level 0.95 with some under-coverage in areas close to the boundary or where the true curve has a high curvature. The low coverage at the boundaries is due to the dwindling information there and low coverage at high curvatures is because the rougher parts of a curve are harder to be recovered by a nonparametric smoothing method.

The plots for the ζ function in Figures 3 and 4 generally follow similar trends to those for the η function. On the other hand, we also notice that both the mean estimates and empirical coverage seem to suffer slightly in the middle of the domain.

The mean estimates of τ_0 for the Weibull and log normal distributions are respectively 2.02 and 2.04, with the average of 95% confidence intervals being respectively $[1.80, 2.25]$ and $[1.78, 2.31]$.

Though the results for $n = 800$ are not shown here, they actually show similar performance with more accurate mean estimates and narrower confidence intervals. So in conclusion, the simulation performance of the proposed method is good in both estimation and inference.

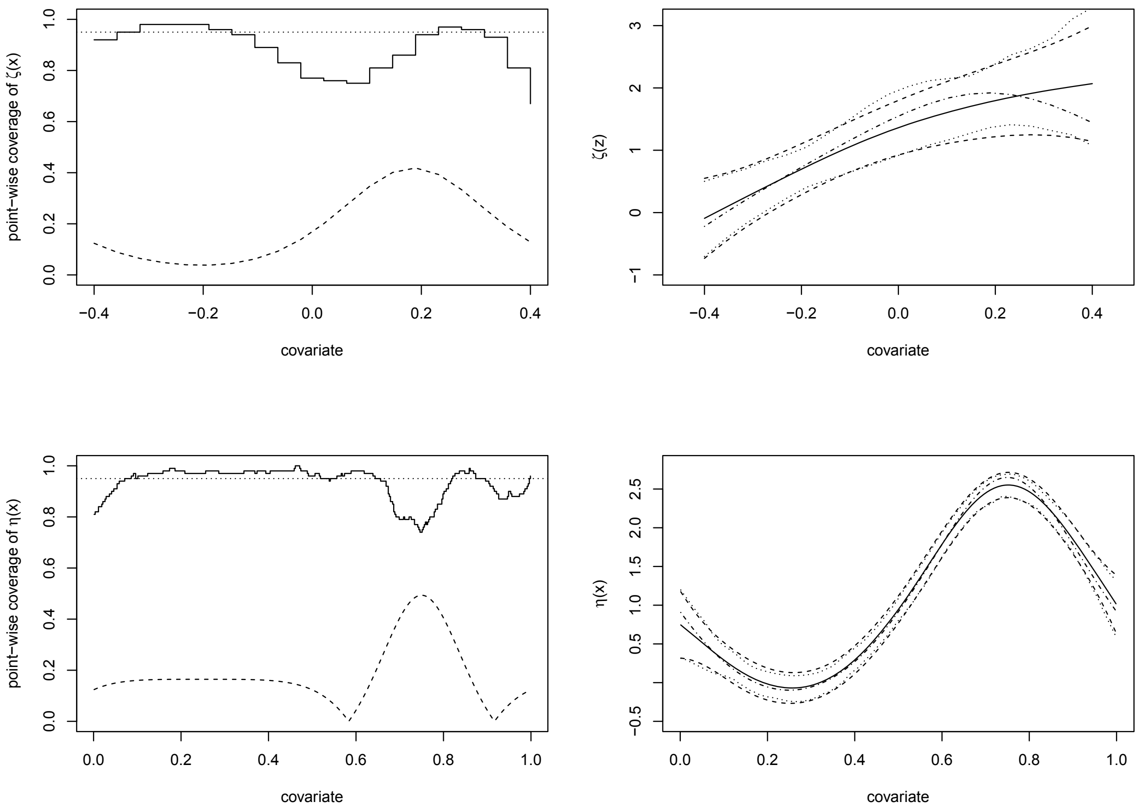

Figure 3. Simulation Results for Weibull Distribution with Test Functions $\pi_0(z)$, $\eta_0(x)$ and $n = 400$. Left column: Point-wise coverages (stepped lines). Superimposed are nominal coverage (dotted lines) and scaled $|\zeta''(z)|$ (dashed lines). Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

5. Discussion

This paper proposes a family of mixture cure rate models with nonparametric forms of covariate effects based on the framework of smoothing spline estimation. Both the probability function of being susceptible and the acceleration in the AFT hazard model of susceptible subjects feature flexible nonparametric forms of covariate effects. The parameter estimates are shown to be consistent and confidence interval estimates are derived under a penalized EM algorithm setting. This paper provides a nice complement to the previous work by Wang et al. (2012) in offering a mixture cure rate model with nonparametric covariate effects where the latency component assumes an accelerated failure time model. Since Wang et al. (2012) considered the setting of proportional hazards with a nonparametric form of relative risk, the other popular approach to modeling the latency, these two methods combined have resolved the open problem of nonparametric modeling

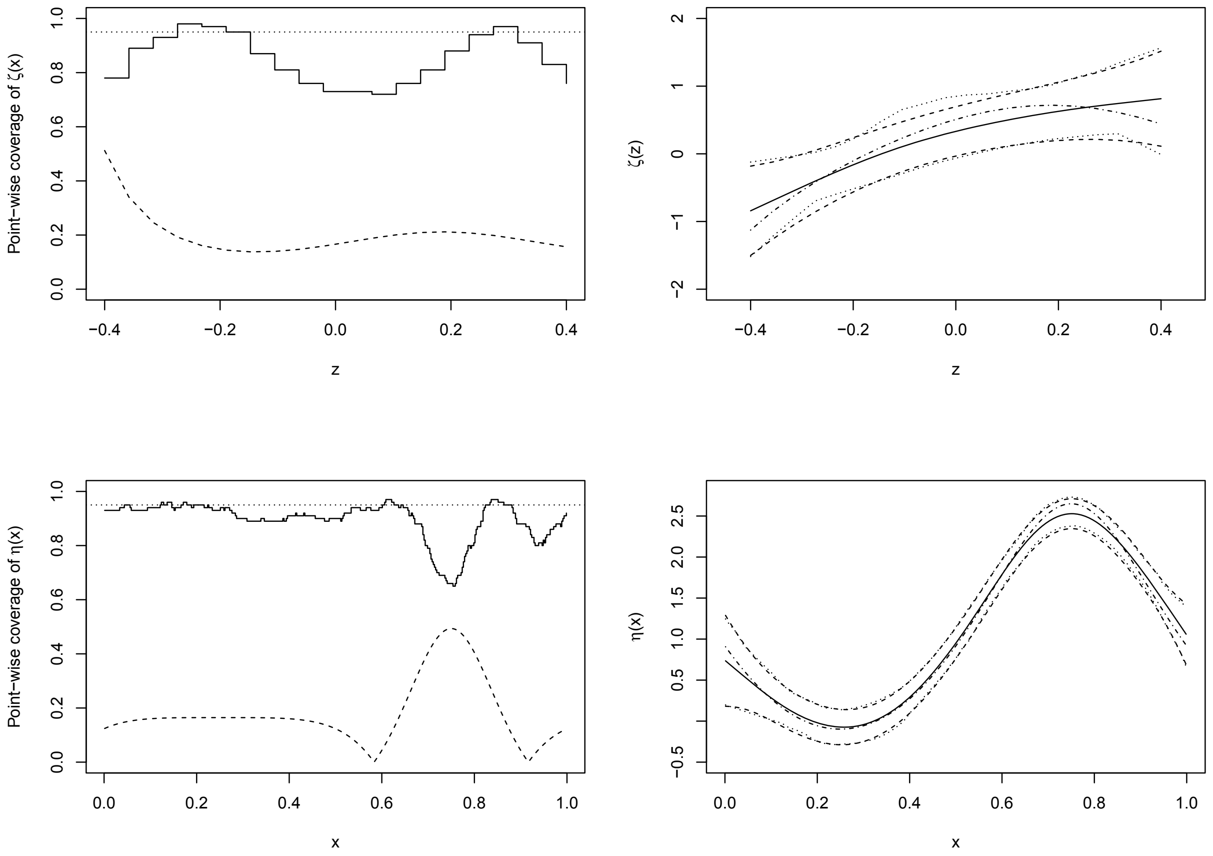

Figure 4. Simulation Results for Log Normal Distribution with Test Functions $\pi_0(z)$, $\eta_0(x)$ and $n = 400$. Left column: Point-wise coverages (stepped lines). Superimposed are nominal coverage (dotted lines) and scaled $|\zeta''(z)|$ (dashed lines). Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

of covariate effects for mixture cure rate models.

Our model is shown be theoretically identifiable. However, under parametric and semiparametric mixture cure rate models, numerical non-identifiability often occurs when censored observations occur mostly before the largest failure time. One remedy is to enforce the zero-tail constraint in Taylor (1995), which essentially assumes all the observations after the largest failure time are cured. As in Wang et al. (2012), our simulations do not show such a problem in the proposed nonparametric cure rate model. But it does not completely rule it out. In case it happens, one can incorporate the zero-tail constraint into our estimation procedure by zeroing out all the y_i's of the censored observations after the largest failure time.

Appendix A: Proof of Proposition 2.1

Suppose that for two sets of parameters (π, η, τ) and $(\tilde{\pi}, \tilde{\eta}, \tilde{\tau})$, we have $S_{\text{pop}}(t|\mathbf{x}, \mathbf{z}) = \tilde{S}_{\text{pop}}(t|\mathbf{x}, \mathbf{z})$, that is, $\pi(\mathbf{z})S(t; \eta, \tau|\mathbf{x}) + 1 - \pi(\mathbf{z}) = \tilde{\pi}(\mathbf{z})S(t; \tilde{\eta}, \tilde{\tau}|\mathbf{x}) + 1 - \tilde{\pi}(\mathbf{z})$. Then

$$\pi(\mathbf{z})\{1 - S(t; \eta, \tau|\mathbf{x})\} = \tilde{\pi}(\mathbf{z})\{1 - S(t; \tilde{\eta}, \tilde{\tau}|\mathbf{x})\}. \tag{6}$$

First, fixing $\mathbf{x}$ and $\mathbf{z}$ and letting $t \to \infty$, we can see that $\pi(\mathbf{z}) = \tilde{\pi}(\mathbf{z})$ for all $\mathbf{z}$. This means we must also have $(t/e^{\eta(\mathbf{x})})^{\tau} = (t/e^{\tilde{\eta}(\mathbf{x})})^{\tilde{\tau}}$ for all t and $\mathbf{x}$. Consequently, we must have $\tau = \tilde{\tau}$ and $\eta(\mathbf{x}) = \tilde{\eta}(\mathbf{x})$ for all $\mathbf{x}$.

Appendix B: M-step of The Penalized E-M Algorithm

We first deal with the optimization of PL$_1$ in Section 2.3. Let R_{J_1} be the reproducing kernel associated with J_1. The minimizer of PL$_1$ does not fall in a finite dimensional space. Therefore, certain approximation is necessary. Instead of $\mathcal{H}_{\zeta}$, we minimize PL$_1$ in the data-adaptive finite dimensional space $\mathcal{H}_{\zeta,n} = \mathcal{H}_{0\zeta} \oplus \text{span}\{R_{J_1}(\mathbf{z}_{i_k}, \cdot) : k = 1, \ldots, q_{\zeta,n}\}$, where $\mathcal{H}_{0\zeta} = \{g \in \mathcal{H}_{\zeta} : J_1(g) = 0\}$ is the null space of J_1, and $\{\mathbf{z}_{i_k} : k = 1, \ldots, q_{\zeta,n}\}$ is a random subset of $\{\mathbf{z}_i : i = 1, \ldots, n\}$. When $q_{\zeta,n} = n$, all the $\mathbf{z}_i$'s are selected. Kim and Gu (2004) showed that a $q_{\zeta,n}$ of order $n^{2/9}$ is sufficient for estimating a reasonably smooth multivariate function in the sense that the estimates in $\mathcal{H}_{\zeta,n}$ and $\mathcal{H}_{\zeta}$ have the same convergence rate. Without loss of generality, we express the minimizer of PL$_1$ in $\mathcal{H}_{\zeta,n}$ as

$$\zeta(\mathbf{z}) = \sum_{\nu=1}^{m_{\zeta}} d_{\nu,\zeta}\phi_{\nu,\zeta}(\mathbf{z}) + \sum_{k=1}^{q_{\zeta,n}} c_{i,\zeta} R_{J_1}(\mathbf{z}_{i_k}, \mathbf{z}) \equiv \boldsymbol{\phi}_{\zeta}^T(\mathbf{z})\mathbf{d}_{\zeta} + \boldsymbol{\xi}_{\zeta}^T(\mathbf{z})\mathbf{c}_{\zeta} \equiv \boldsymbol{\psi}_{\zeta}(\mathbf{z})^T\mathbf{b}_{\zeta}, \tag{7}$$

where $\{\phi_{\nu,\zeta} : \nu = 1, \ldots, m_{\zeta}\}$ is a set of basis functions for $\mathcal{H}_{0\zeta}$, $\boldsymbol{\psi}_{\zeta}(\mathbf{z})^T = (\boldsymbol{\phi}_{\zeta}(\mathbf{z})^T, \boldsymbol{\xi}_{\zeta}(\mathbf{z}))$ and $\mathbf{b}_{\zeta}^T = (\mathbf{d}_{\zeta}^T, \mathbf{c}_{\zeta}^T)$. Substituting (7) into PL$_1$ gives

$$-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\{\boldsymbol{\phi}_{\zeta}^T(\mathbf{z}_i)\boldsymbol{d}_{\zeta} + \boldsymbol{\xi}_{\zeta}^T(\mathbf{z}_i)\boldsymbol{c}_{\zeta}\} - \log\{1 + \exp\left(\boldsymbol{\phi}_{\zeta}^T(\mathbf{z}_i)\boldsymbol{d}_{\zeta} + \boldsymbol{\xi}_{\zeta}^T(\mathbf{z}_i)\boldsymbol{c}_{\zeta}\right)\}\right] + \frac{\beta}{2}\boldsymbol{c}_{\zeta}^T Q_{\zeta}\boldsymbol{c}_{\zeta}, \tag{8}$$

where $Q_{\zeta} = (\boldsymbol{\xi}_{\zeta}(\mathbf{z}_{i_1}), \ldots, \boldsymbol{\xi}_{\zeta}(\mathbf{z}_{i_{q_{\zeta,n}}}))$. For fixed smoothing parameters β and θ's hidden in J_1, the minimizer ζ_{β} of PL$_1$ is computed as follows. Given an initial estimate $\breve{\zeta}$, write $\breve{u}_i = -Y_i + \frac{\exp\{\breve{\zeta}(\mathbf{z}_i)\}}{1+\exp\{\breve{\zeta}(\mathbf{z}_i)\}}$, $\breve{w}_i = \frac{\exp\{\breve{\zeta}(\mathbf{z}_i)\}}{[1+\exp\{\breve{\zeta}(\mathbf{z}_i)\}]^2}$. The Newton iteration updates $\breve{\zeta}$ by the minimizer of the penalized weighted least squares functional $-\frac{1}{n}\sum_{i=1}^{n} \breve{w}_i\{\breve{Y}_i - \zeta(\mathbf{z}_i)\}^2 + \frac{\beta}{2}J_1(\zeta)$, where $\breve{Y}_i = \breve{\zeta}(\mathbf{z}_i) - \breve{u}_i/\breve{w}_i$. The selection of the smoothing parameters can be done through an outer-loop optimization of the cross-validation score (5.28) in Gu (2013).

For a fixed τ, PL$_2$ is first optimized in a data-adaptive finite dimensional space $\mathcal{H}_{\eta,n}$ with

$$\eta(\cdot) = \sum_{i=1}^{m_{\eta}} d_{\nu,\eta}\phi_{\nu,\eta}(\cdot) + \sum_{k=1}^{q_{\eta,n}} c_{k,\eta} R_{J_2}(v_{i_k}, \cdot) \equiv \boldsymbol{\phi}_{\eta}^T(\cdot)\mathbf{d}_{\eta} + \boldsymbol{\xi}_{\eta}^T(\cdot)\mathbf{c}_{\eta} \equiv \boldsymbol{\psi}_{\eta}(\cdot)^T\mathbf{b}_{\eta}, \tag{9}$$

where $\cdot$ stands for $\mathbf{x}$ and $\{v_{i_k} : k = 1, \ldots, q_{\eta,n}\}$ is a random subset of $\{\mathbf{x}_i : \delta_i = 1\}$. For a

fixed τ, define $h_1(t;\eta,\tau|\mathbf{x}) = -\partial \log h(t;\eta,\tau|\mathbf{x})/\partial\eta$ and $h_2(t;\eta,\tau|\mathbf{x}) = \partial h_1(t;\eta,\tau|\mathbf{x})/\partial\eta$. Given an initial estimate $\breve{\eta}$, write $\breve{u}_i = \partial L_2(\breve{\eta},\tau;\mathbf{y})/\partial\eta = \delta_i h_1(t_i;\eta,\tau|\mathbf{x}_i) - y_i \int_0^{t_i} h_1(t;\eta,\tau|\mathbf{x}_i)h(t;\eta,\tau|\mathbf{x}_i)dt$, $\breve{w}_i = \partial^2 L_2(\breve{\eta},\tau;\mathbf{y})/\partial\eta^2 = \delta_i h_2(t_i;\eta,\tau|\mathbf{x}_i) - y_i \int_0^{t_i} h_2(t;\eta,\tau|\mathbf{x}_i)h(t;\eta,\tau|\mathbf{x}_i)dt$
$+y_i \int_0^{t_i} h_1^2(t;\eta,\tau|\mathbf{x}_i)h(t;\eta,\tau|\mathbf{x}_i)dt$, and $\breve{Y}_i = \breve{\eta}(\mathbf{x}_i) - \breve{u}_i/\breve{w}_i$. The Newton iteration updates $\breve{\eta}$ by the minimizer of the penalized weighted least squares functional $-\frac{1}{n}\sum_{i=1}^n \breve{w}_i\{\breve{Y}_i - \eta(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}J_2(\eta)$. Similar to PL_1, the selection of the smoothing parameters can be done through an outer-loop optimization of the cross-validation score (5.28) in Gu (2013). Next, for a fixed η, we estimate τ by maximizing $L_2(\eta,\tau;\mathbf{y})$. The algorithm then alternates between the updating of η and τ.

Appendix C: Derivation of The Formula for I_{obs}

Plugging into $L(\mathbf{y};(\zeta,\eta,\tau))$ the expressions (7) and (9) for ζ and η and direct differentiation yields

$$G(\mathbf{y};\Theta) = \begin{pmatrix} \sum_{i=1}^n \left[y_i\boldsymbol{\psi}_{\zeta,i} - \left\{1+\exp(-\boldsymbol{\psi}_{\zeta,i}^T\mathbf{b}_\zeta)\right\}^{-1}\boldsymbol{\psi}_\zeta(z_i)\right] - n\beta Q_\zeta^*\mathbf{b}_\zeta \\ \sum_{i=1}^n g_i(\mathbf{b}_\eta,\tau) - n\lambda Q_\eta^*\mathbf{b}_\eta \\ \sum_{i=1}^n \left\{\delta_i h_\tau(t_i;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i) - y_i\int_0^{t_i} h_\tau(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt\right\} \end{pmatrix}, \tag{10}$$

$$B(\mathbf{y};\Theta) = \begin{pmatrix} \mathbf{B}_{\zeta\zeta} & 0 & 0 \\ 0 & \mathbf{B}_{\eta\eta} & \mathbf{B}_{\eta\tau} \\ 0 & \mathbf{B}_{\eta\tau}^T & \mathbf{B}_{\tau\tau} \end{pmatrix}, \tag{11}$$

where $\boldsymbol{\psi}_{\zeta,i} = \boldsymbol{\psi}_\zeta(z_i)$, $\boldsymbol{\psi}_{\eta,i} = \boldsymbol{\psi}_\eta(\mathbf{x}_i)$, $Q_\zeta^* = \mathrm{diag}(0,Q_\zeta)$, $Q_\eta^* = \mathrm{diag}(0,Q_\eta)$, $h_\tau(t;\eta,\tau|\mathbf{x}) = -\partial\log h(t;\eta,\tau|\mathbf{x})/\partial\tau$, $g_i(\mathbf{b}_\eta,\tau) = \delta_i\boldsymbol{\psi}_{\eta,i}h_1(t_i;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i) -$

$y_i \int_0^{t_i} \boldsymbol{\psi}_\eta(t) h_1(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) dt$, and

$$\begin{aligned}
\mathbf{B}_{\zeta\zeta} =& \sum_{i=1}^n \exp\left(\boldsymbol{\psi}_{\zeta,i}^T \mathbf{b}_\zeta\right) \left\{1 + \exp\left(\boldsymbol{\psi}_{\zeta,i}^T \mathbf{b}_\zeta\right)\right\}^{-2} \boldsymbol{\psi}_{\zeta,i} \boldsymbol{\psi}_{\zeta,i}^T + n\beta Q_\zeta^*, \\
\mathbf{B}_{\eta\eta} =& \sum_{i=1}^n \Big\{ \delta_i h_2(t_i; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) \boldsymbol{\psi}_{\eta,i} \boldsymbol{\psi}_{\eta,i}^T \\
& - y_i \int_0^{t_i} h_2(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) \boldsymbol{\psi}_\eta(t) \boldsymbol{\psi}_\eta(t)^T dt \\
& + y_i \int_0^{t_i} h_1^2(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) \boldsymbol{\psi}_\eta(t) \boldsymbol{\psi}_\eta(t)^T dt \Big\}, \\
\mathbf{B}_{\eta\tau} =& \sum_{i=1}^n \Big\{ \delta_i h_{1\tau}(t_i; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) \boldsymbol{\psi}_{\eta,i} \\
& - y_i \int_0^{t_i} h_{1\tau}(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) \boldsymbol{\psi}_\eta(t) dt \\
& + y_i \int_0^{t_i} h_1(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h_1 \tau(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) \boldsymbol{\psi}_\eta(t) \boldsymbol{\psi}_\eta(t)^T dt \Big\}, \\
\mathbf{B}_{\tau\tau} =& \sum_{i=1}^n \Big\{ \delta_i h_{\tau\tau}(t_i; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) - y_i \int_0^{t_i} h_{\tau\tau}(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) dt \\
& + y_i \int_0^{t_i} h_\tau^2(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) h(t; \boldsymbol{\psi}_\eta^T \mathbf{b}_\eta, \tau|\mathbf{x}_i) dt \Big\}
\end{aligned}$$

with $h_{1\tau} = \partial h_1(t; \eta, \tau|\mathbf{x})/\partial\tau$ and $h_{\tau\tau} = \partial h_\tau(t; \eta, \tau|\mathbf{x})/\partial\tau$. Note $E[y_i y_j] = E[y_i]E[y_j]$ for $i \neq j$ and $E[y_i^2] = E[y_i]$. Hence

$$E\left[G(\mathbf{y}; \Theta) G^T(\mathbf{y}; \Theta)\right] = E\left[G(\mathbf{y}; \Theta)\right] E\left[G(\mathbf{y}; \Theta)\right]^T + \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33}, \end{pmatrix}, \tag{12}$$

where

$$
\begin{aligned}
A_{11} =&\sum_{i=1}^{n} \left(E[y_i] - E[y_i]^2\right) \boldsymbol{\psi}_{\zeta,i}\boldsymbol{\psi}_{\zeta,i}^T, \\
A_{12} =&\sum_{i=1}^{n} \left(E[y_i] - E[y_i]^2\right) \{\boldsymbol{\psi}_{\zeta,i} \int_0^{t_i} \boldsymbol{\psi}_\eta^T(t) h_1(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt\}, \\
A_{13} =&\sum_{i=1}^{n} \left(E[y_i] - E[y_i]^2\right) \{\boldsymbol{\psi}_{\zeta,i} \int_0^{t_i} h_\tau(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt\}, \\
A_{22} =&\sum_{i=1}^{n} \left(E[y_i] - E[y_i]^2\right) \int_0^{t_i} \boldsymbol{\psi}_\eta(t) h_1(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt \\
&\times \int_0^{t_i} \boldsymbol{\psi}_\eta^T(t) h_1(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt, \\
A_{23} =&\sum_{i=1}^{n} \left(E[y_i] - E[y_i]^2\right) \int_0^{t_i} \boldsymbol{\psi}_\eta^T(t) h_1(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt \\
&\times \int_0^{t_i} h_\tau(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt, \\
A_{33} =&\sum_{i=1}^{n} \left(E[y_i] - E[y_i]^2\right) \{\int_0^{t_i} h_\tau(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)h(t;\boldsymbol{\psi}_\eta^T\mathbf{b}_\eta,\tau|\mathbf{x}_i)dt\}^2.
\end{aligned}
$$

Both G and B are linear in y_i, so their expectations are easily obtained through replacing y_i by $E[y_i]$ in (10) and (11). Further replacing the $E[y_i]$ in $E[GG^T]$ and $E[B]$ by the converged $y_i^{(\infty)}$ yields the observed information matrix I_{obs} in the Louis formula.

Appendix D: Proof of Theorem 2.1

We will need the following conditions.

A1. The domains $\mathcal{Z}$ and $\mathcal{X}$ of covariates $\mathbf{z}$ and $\mathbf{x}$ are compact sets in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively.

A2. Given the covariates $\mathbf{z}$ and $\mathbf{x}$, censoring time C is independent of susceptible indicator y and failure time T^*.

A3. Assume the observations are in a finite time interval $[0,\Omega]$. Assume that the true hazard function $h_0(t,\mathbf{x})$ for susceptible subjects is bounded away from zero and infinity.

A4. Assume the true parameter $\tau_0 \in [0,M]$ for some $M > 0$ and the true function $\eta_0 \in \mathcal{H}_\eta$. Assume that the function $S(t;\eta,\tau|\mathbf{x})$, and thus $h(t;\eta,\tau|\mathbf{x})$, are continuous with respect to η and τ. For any η in a sufficiently big convex neighborhood B_2 of η_0, there exist constants $c_3, c_4 > 0$ such that $c_3 h(t;\eta_0,\tau|\mathbf{x}) \le h(t;\eta,\tau|\mathbf{x}) \le c_4 h(t;\eta_0,\tau|\mathbf{x})$ for all $\mathbf{x}$.

A5. Assume the true function $\zeta_0 \in \mathcal{H}_\zeta$. Let $w_\zeta(\mathbf{z}) = e^{\zeta(\mathbf{z})}\{1+e^{\zeta(\mathbf{z})}\}^{-2}$. For any ζ in a sufficiently big convex neighborhood B_1 of ζ_0, there exist constants $c_1, c_2 > 0$ such that $c_1 w_{\zeta_0}(\mathbf{z}) \le w_\zeta(\mathbf{z}) \le c_2 w_{\zeta_0}(\mathbf{z})$ for all $\mathbf{z}$.

A6. The smoothing parameters $\beta \asymp n^{-r_1/(r_1+1)}$ and $\lambda \asymp n^{-r_2/(r_2+1)}$.

Condition A1 is a common boundedness assumption on covariates. Condition A2 assumes noninformative censoring. Condition A3 is the common boundedness assumption on the hazard. Besides an upper bound for τ_0, Condition A4 assumes that η_0 has proper level of smoothness and integrates to zero. The neighborhood B_2 in Condition A4 should be big enough to contain all the estimates of η_0 considered below. When the members of B_2 are all uniformly bounded, Condition A4 is automatically satisfied. Condition A5 is similar. The orders for β and λ in Condition A6 match that in standard smooth spline problems.

The estimates $\widehat{\zeta}$ and $(\widehat{\tau}, \widehat{\eta})$ are respectively the minimizers of the following two penalized likelihoods

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_i\zeta(\mathbf{z}_i) - \log\{1+e^{\zeta(\mathbf{z}_i)}\}\right] + \frac{\beta}{2}J_1(\zeta), \tag{13}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i \log h(t_i;\eta,\tau|\mathbf{x}_i) - \widehat{\mu}_i\int_0^{t_i} h(t;\eta,\tau|\mathbf{x}_i)dt\right\} + \frac{\lambda}{2}J_2(\eta), \tag{14}$$

where

$$\widehat{\mu}_i = \delta_i + (1-\delta_i)\frac{S(t_i;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)}{\exp\{-\widehat{\zeta}(\mathbf{z}_i)\} + S(t_i;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)}.$$

Let $\widetilde{\zeta}$ and $\widetilde{\eta}$ be respectively the minimizers of

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\mu_{0i}\zeta(\mathbf{z}_i) - \log\{1+e^{\zeta(\mathbf{z}_i)}\}\right] + \frac{\beta}{2}J_1(\zeta), \tag{15}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \log h(t_i;\eta,\tau_0|\mathbf{x}_i) - \mu_{0i}\int_0^{t_i} h(t;\eta,\tau_0|\mathbf{x}_i)dt\right) + \frac{\lambda}{2}J_2(\eta), \tag{16}$$

where

$$\mu_{0i} = \delta_i + (1-\delta_i)\frac{S(t;\eta_0,\tau_0|\mathbf{x}_i)}{\exp\{-\zeta_0(\mathbf{z}_i)\} + S(t;\eta_0,\tau_0|\mathbf{x}_i)}.$$

Let

$$\mu_0(t,\mathbf{x},\mathbf{z}) = \frac{\pi_0(\mathbf{z})S_0(t,\mathbf{x})}{1-\pi_0(\mathbf{z})+\pi_0(\mathbf{z})S_0(t,\mathbf{x})} = \frac{S(t;\eta_0,\tau_0|\mathbf{x})}{\exp\{-\zeta_0(\mathbf{z})\} + S(t;\eta_0,\tau_0|\mathbf{x})}.$$

Define $w_{\tau_0,\eta_0}(t,\mathbf{x},\mathbf{z}) = \mu_0(t,\mathbf{x},\mathbf{z})\int_0^t h_1^2(s;\eta_0,\tau_0|\mathbf{x})h(s;\eta_0,\tau_0|\mathbf{x})ds$. For $g \in \mathcal{H}_\zeta$ and $k \in \mathcal{H}_\eta$, define

$$V_1(g) = \int_{\mathcal{Z}} g^2(\mathbf{z})w_{\zeta_0}(\mathbf{z})f_z(\mathbf{z})d\mathbf{z}, \tag{17}$$

$$V_2(k) = \int_{\mathcal{Z}} f_z(\mathbf{z})\int_{\mathcal{X}} f_x(\mathbf{x})\int_{\mathcal{T}} k^2(\mathbf{x})w_{\tau_0,\eta_0}(t,\mathbf{x},\mathbf{z})dtd\mathbf{x}d\mathbf{z}, \tag{18}$$

where f_z and f_x are respectively the density functions for covariates $\mathbf{z}$ and $\mathbf{x}$.

We first present a lemma without proof that indicates the equivalence between $V_1(\cdot)$, $V_2(\cdot)$ and the L_2-norm $\|\cdot\|_2^2$. It is a straightforward conclusion from the boundedness conditions A1 and A3.

LEMMA A.1 *Let $g \in \mathcal{H}_\zeta$ and $k \in \mathcal{H}_\eta$. Then there exist constants $0 < c_1 \le c_2 < \infty$ and $0 < c_3 \le c_4 < \infty$ such that*

$$c_1\|g\|_2^2 \le V_1(g) \le c_2\|g\|_2^2 \text{ and } c_3\|k\|_2^2 \le V_2(k) \le c_4\|k\|_2^2.$$

Wang, Du and Liang (2012) have shown that $(V_1+\beta J_1)(\widetilde{\zeta}-\zeta_0) = O_p(\beta+n^{-1}\beta^{-1/r_1}) = O_p(n^{-r_1/(r_1+1)})$ given Conditions A5 and A6. The rest of our proof consists of two steps, dealing respectively with the convergence rates of $\widetilde{\eta}$ (step one), and $\widehat{\zeta}$ and $\widehat{\eta}$ (step three).

STEP ONE: Convergence rate of $\widetilde{\eta}$.

We will derive this through an eigenvalue analysis with two phases. In the first phase, we show the convergence rate $O_p(n^{-r_2/(r_2+1)})$ for the minimizer $\widetilde{\eta}^*$ of a quadratic approximation to (15). In the second phase, we show that the difference between $\widetilde{\eta}^*$ and $\widetilde{\eta}$ is also $O_p(n^{-r_2/(r_2+1)})$, and so is the convergence rate of $\widetilde{\eta}$.

A quadratic functional B is said to be *completely continuous* with respect to another quadratic functional A, if for any $\epsilon > 0$, there exists a finite number of linear functionals $L_1, \ldots, L_k$ such that $L_j f = 0, j = 1, \ldots, k$, implies that $B(f) \le \epsilon A(f)$; When a quadratic functional B is *completely continuous* with respect to another quadratic functional A, there exists eigenfunctions $\{\phi_\nu, \nu = 1, 2, \cdots\}$ such that $B(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $A(\phi_\nu, \phi_\mu) = \rho_\nu\delta_{\nu\mu}$, where $\delta_{\nu\mu}$ is the Kronecker delta and $0 \le \rho_\nu \uparrow \infty$. And any function satisfying $A(f) < \infty$ has a *Fourier series expansion* $f = \sum_\nu f_\nu\phi_\nu$, where $f_\nu = B(f, \phi_\nu)$ are the *Fourier coefficients*. See Gu (2013).

Two more lemmas without proof are presented next. The first one follows directly from the boundedness conditions A1 and A3, the results in Section 9.1 of Gu (2013) and Lemma A.1. The second one is exactly Lemma 9.1 in Gu (2013).

LEMMA A.2 *V_2 is completely continuous to J_2 and the eigenvalues ρ_ν of J_2 with respect to V_2 satisfy that as $\nu \to \infty$, $\rho_\nu^{-1} = O(\nu^{r_2})$.*

LEMMA A.3 *As $\lambda \to 0$, the sums $\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}$, $\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2}$, and $\sum_\nu \frac{1}{1+\lambda\rho_\nu}$ are all of order $O(\lambda^{-1/r_2})$.*

Consider the minimizer $\widetilde{\eta}^*$ of the quadratic functional

$$\frac{1}{n}\sum_{i=1}^{n}\eta(\mathbf{x}_i)\{\delta_i h_1(t_i;\eta_0,\tau_0|\mathbf{x}_i) - \mu_{0i}\int_0^{t_i} h_1(t;\eta_0,\tau_0|\mathbf{x}_i)h(t;\eta_0,\tau_0|\mathbf{x}_i)\} \\ + \frac{1}{2}V_2(\eta-\eta_0) + \frac{\lambda}{2}J_2(\eta). \quad (19)$$

Let $\eta = \sum_\nu \eta_\nu\phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ be the Fourier expansions of η and η_0. Write $\gamma_\nu = n^{-1}\sum_{i=1}^n \phi_\nu(\mathbf{x}_i)u_i(\eta_0,\nu_0)$, where $u_i(\eta,\tau) = \delta_i h_1(t_i;\eta,\tau|\mathbf{x}_i) - \mu_{0i}\int_0^{t_i} h_1(t;\eta,\tau|\mathbf{x}_i)h(t;\eta,\tau|\mathbf{x}_i)dt$. Plugging them into (19), we can show that the minimizer of (19) has Fourier coefficients $\widetilde{\eta}_\nu^* = (\gamma_\nu + \eta_{\nu,0})/(1+\lambda\rho_\nu)$. Note that $V_2(\phi_\nu) = 1$.

Straightforward calculation gives $E[\gamma_\nu]=0$ and $E[\gamma_\nu^2]=n^{-1}$. Then,

$$\begin{aligned} E[V_2(\widetilde{\eta}^*-\eta_0)] &= \frac{1}{n}\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2}+\lambda\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}\rho_\nu\eta_{\nu,0}^2, \\ E[\lambda J_2(\widetilde{\eta}^*-\eta_0)] &= \frac{1}{n}\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}+\lambda\sum_\nu \frac{(\lambda\rho_\nu)^2}{(1+\lambda\rho_\nu)^2}\rho_\nu\eta_{\nu,0}^2. \end{aligned} \tag{20}$$

Combining Lemma A.3 and (20), we obtain that $(V_2+\lambda J_2)(\widetilde{\eta}^*-\eta_0)=O_p(\lambda+n^{-1}\lambda^{-1/r_2})$, as $n\to\infty$ and $\lambda\to 0$.

For a functional $L(f)$, define the differential operator $D_{f,g}$ as $D_{f,g}(L)=\left.\frac{dL(f+\alpha g)}{d\alpha}\right|_{\alpha=0}$. Applying $D_{\widetilde{\eta},\widetilde{\eta}-\widetilde{\eta}^*}$ to (16) yields

$$\frac{1}{n}\sum_{i=1}^n u_i(\widetilde{\eta},\tau_0)\{\widetilde{\eta}(\mathbf{x}_i)-\widetilde{\eta}^*(\mathbf{x}_i)\}+\lambda J_2(\widetilde{\eta},\widetilde{\eta}-\widetilde{\eta}^*)=0. \tag{21}$$

Applying $D_{\widetilde{\eta}^*,\widetilde{\eta}-\widetilde{\eta}^*}$ to (19) yields

$$\frac{1}{n}\sum_{i=1}^n u_i(\eta_0,\tau_0)\{\widetilde{\eta}(\mathbf{x}_i)-\widetilde{\eta}^*(\mathbf{x}_i)\}+V_2(\widetilde{\eta}^*-\eta_0,\widetilde{\eta}-\widetilde{\eta}^*)+\lambda J_2(\widetilde{\eta}^*,\widetilde{\eta}-\widetilde{\eta}^*)=0. \tag{22}$$

Subtracting (22) from (21) yields

$$\begin{aligned} \lambda J_2(\widetilde{\eta}-\widetilde{\eta}^*)&-\frac{1}{n}\sum_{i=1}^n u_i(\widetilde{\eta},\tau_0)\{\widetilde{\eta}(\mathbf{x}_i)-\widetilde{\eta}^*(\mathbf{x}_i)\} \\ &= V_2(\widetilde{\eta}^*-\eta_0,\widetilde{\eta}-\widetilde{\eta}^*)+\frac{1}{n}\sum_{i=1}^n u_i(\eta_0,\tau_0)\{\widetilde{\eta}(\mathbf{x}_i)-\widetilde{\eta}^*(\mathbf{x}_i)\}. \end{aligned} \tag{23}$$

Now by the mean value theorem, Condition A4, and Lemma 9.16 in Gu (2013), (23) indicates

$$(c_3V_2+\lambda J_2)(\widetilde{\eta}-\widetilde{\eta}^*)\le\big\{(|1-c|V_2+\lambda J_2)(\widetilde{\eta}-\widetilde{\eta}^*)\big\}^{1/2}O_p\big(\{(|1-c|V_2+\lambda J_2)(\widetilde{\eta}-\eta_0)\}^{1/2}\big),$$

where $c\in[c_3,c_4]$ and c_3,c_4 are constants in Condition A(4). Hence $(V_2+\lambda J_2)(\widetilde{\eta}-\widetilde{\eta}^*)$ is $O_p(\lambda+n^{-1}\lambda^{-1/r_2})$. Consequently, $(V_2+\lambda J_2)(\widetilde{\eta}-\eta_0)=O_p(\lambda+n^{-1}\lambda^{-1/r_2})=O_p(n^{-r_2/(r_2+1)})$ given Condition A6.

Step Two: Convergence rates of $\widehat{\zeta}$ and $\widehat{\eta}$.

Let $u_\zeta(\mathbf{z})=e^{\zeta(\mathbf{z})}(1+e^{\zeta(\mathbf{z})})^{-1}$. Applying $D_{\widetilde{\zeta},\widetilde{\zeta}-\zeta_0}$ to (15) and $D_{\widehat{\zeta},\widetilde{\zeta}-\zeta_0}$ to (13), and subtracting the resulting equations give

$$\frac{1}{n}\sum_{i=1}^n(\widehat{\mu}_i-\mu_{0i})(\widetilde{\zeta}-\zeta_0)_i=\frac{1}{n}\sum_{i=1}^n(u_{\widehat{\zeta}}-u_{\widetilde{\zeta}})_i(\widetilde{\zeta}-\zeta_0)_i+\beta J_1(\widehat{\zeta}-\widetilde{\zeta},\widetilde{\zeta}-\zeta_0). \tag{24}$$

Applying $D_{\widetilde{\eta},\widetilde{\eta}-\eta_0}$ to (16) and $D_{\widehat{\eta},\widetilde{\eta}-\eta_0}$ to (14), and subtracting the resulting equations give

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_i-\mu_{0i})\{\widetilde{\eta}(\mathbf{x}_i)-\eta_0(\mathbf{x}_i)\}\int_0^{t_i}h_1(t;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)h(t;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)dt&\\ -\frac{1}{n}\sum_{i=1}^{n}\delta_i\{\widetilde{\eta}(\mathbf{x}_i)-\eta_0(\mathbf{x}_i)\}\{h_1(t_i;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)-h_1(t_i;\widetilde{\eta},\tau_0|\mathbf{x}_i)\}&\\ =-\frac{1}{n}\sum_{i=1}^{n}\mu_{0i}\{\widetilde{\eta}(\mathbf{x}_i)-\eta_0(\mathbf{x}_i)\}\Big\{\int_0^{t_i}h_1(t;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)h(t;\widehat{\eta},\widehat{\tau}|\mathbf{x}_i)dt&\\ -\int_0^{t_i}h_1(t;\widetilde{\eta},\tau_0|\mathbf{x}_i)h(t;\widetilde{\eta},\tau_0|\mathbf{x}_i)dt\Big\}+\lambda J_2(\widehat{\eta}-\widetilde{\eta},\widetilde{\eta}-\eta_0).&\quad(25)\end{aligned}$$

For $0\le\alpha_1,\alpha_2,\alpha_3\le 1$, consider

$$\begin{aligned}A_i(\alpha_1,\alpha_2,\alpha_3)&=\delta_i+(1-\delta_i)S\big(t;\eta_0+\alpha_2(\widehat{\eta}-\eta_0),\tau_0+\alpha_3(\widehat{\tau}-\tau_0)|\mathbf{x}_i\big)\\ &\times\Big(\exp[-\{\zeta_0+\alpha_1(\widehat{\zeta}-\zeta_0)\}(\mathbf{z}_i)]+S\big(t;\eta_0+\alpha_2(\widehat{\eta}-\eta_0),\tau_0+\alpha_3(\widehat{\tau}-\tau_0)|\mathbf{x}_i\big)\Big)^{-1}.\end{aligned}$$

Then $\widehat{\mu}_i=A_i(1,1,1)$ and $\mu_{0i}=A_i(0,0,0)$. By the first order Taylor expansion of A_i at $(0,0,0)$,

$$\widehat{\mu}_i\approx\mu_{0i}+(1-\delta_i)M_i\Big\{(\widehat{\zeta}-\zeta_0)(\mathbf{z}_i)+S_\eta(t_i;\eta_0,\tau_0|\mathbf{x}_i)(\widehat{\eta}-\eta_0)(\mathbf{x}_i)+S_\tau(t_i;\eta_0,\tau_0|\mathbf{x}_i)(\widehat{\tau}-\tau_0)\Big\},\quad(26)$$

where $M_i=\exp\{-\zeta_0(\mathbf{z}_i)\}S(t_i;\eta_0,\tau_0|\mathbf{x}_i)\big[\exp\{-\zeta_0(\mathbf{z}_i)\}+S(t_i;\eta_0,\tau_0|\mathbf{x}_i)\big]^{-2}$, $S_\eta=\partial S/\partial\eta$ and $S_\tau=\partial S/\partial\tau$.

Plugging (26) into (24), a combination of the mean value theorem, Conditions A4-A5, and Lemma 10.17 of Gu (2013) can show that as $\beta\to 0$ and $n\beta^{2/r_1}\to\infty$, (24) indicates

$$\begin{aligned}\{C_1\|\widehat{\zeta}-\zeta_0\|\cdot\|\widetilde{\zeta}-\zeta_0\|+C_2\|\widehat{\eta}-\eta_0\|\cdot\|\widetilde{\zeta}-\zeta_0\|+C_3|\widehat{\tau}-\tau_0|\cdot\|\widetilde{\zeta}-\zeta_0\|\}\{1+o_p(1)\}&\\ =C_4V_1(\widehat{\zeta}-\widetilde{\zeta},\widetilde{\zeta}-\zeta_0)+\beta J_1(\widehat{\zeta}-\widetilde{\zeta},\widetilde{\zeta}-\zeta_0)&\quad(27)\end{aligned}$$

for some constants $C_1,C_2,C_3,C_4>0$. Similarly, (26) and (25) indicates that as $\lambda\to 0$ and $n\lambda^{2/r_2}\to\infty$,

$$\begin{aligned}\{C_5\|\widehat{\zeta}-\zeta_0\|\cdot\|\widetilde{\eta}-\eta_0\|+C_6\|\widehat{\eta}-\eta_0\|\cdot\|\widetilde{\eta}-\eta_0\|+C_7|\widehat{\tau}-\tau_0|\cdot\|\widetilde{\eta}-\eta_0\|\}\{1+o_p(1)\}&\\ =C_8V_2(\widehat{\eta}-\widetilde{\eta},\widetilde{\eta}-\eta_0)+\lambda J_2(\widehat{\eta}-\widetilde{\eta},\widetilde{\eta}-\eta_0).&\quad(28)\end{aligned}$$

for some constants $C_4,C_5,C_6>0$. Hence, (27) and (28) combined with results in the previous step and Lemma A.1 yield $\|\widehat{\zeta}-\zeta_0\|^2=O_p(n^{-r/(r+1)})$, $\|\widehat{\eta}-\eta_0\|^2=O_p(n^{-r/(r+1)})$, and $|\widehat{\tau}-\tau_0|^2=O_p(n^{-r/(r+1)})$ for $r=\min(r_1,r_2)$.

References

Berkson, J., and Gage, R.P. (1952), ‘Survival Curve for Cancer Patients Following Treatment’, *Journal of the American Statistical Association*, 47, 501–515.

Corbière, F., Commenges, D., Taylor, J.M.G., and Joly, P. (2009), 'A penalized likelihood approach for mixture cure models', *Statistics in Medicine*, 28, 510–524.

Cox, D.R. (1997), 'Some remarks on the analysis of survival data', in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, eds. D. Lin and T.R. Fleming, New York. Springer-Verlag, pp. 1–9.

Farewell, V. (1982), 'The Use of Mixture Models for the Analysis of Survival Data With Long-Term Survivors', *Biometrics*, 38, 1041–1046.

Gu, C. (2013), *Smoothing Spline ANOVA Models (2nd Ed.)*, New York: Springer-Verlag.

Kim, Y.J., and Gu, C. (2004), 'Smoothing Spline Gaussian Regression: More Scalable Computation Via Efficient Approximation', *Journal of the Royal Statistical Society, Series B*, 66, 337–356.

Kosorok, M.R. (2008), *Introduction to empirical processes and semi parametric inference*, New York: Springer-Verlag.

Kuk, A.Y.C., and Chen, C.H. (1992), 'A Mixture Model Combining Logistic Regression with Proportional Hazards Regression', *Biometrika*, 79, 531–541.

Li, C.S., and Taylor, J.M.G. (2002), 'A semi-parametric accelerated failure time cure model', *Statistics in Medicine*, 21, 3235–3247.

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.

Lu, W. (2010), 'Efficient Estimation for an Accelerated Failure Time Model with a Cure Fraction', *Statistica Sinica*, 20, 661–674.

Lu, W., and Ying, Z. (2004), 'On Semiparametric Transformation Cure Models', *Biometrika*, 91, 331–343.

Othus, M., Li, Y., and Tiwari, R.C. (2009), 'A Class of Semiparametric Mixture Cure Survival Models With Dependent Censoring', *Journal of the American Statistical Association*, 104, 1241–1250.

Peng, Y. (2003), 'Fitting Semiparametric Cure Models', *Computational Statistics & Data Analysis*, 41, 481–490.

Peng, Y., and Dear, K.B.G. (2000), 'A Nonparametric Mixture Model for Cure Rate Estimation', *Biometrics*, 56, 237–243.

Reid, N. (1994), 'A conversation with Sir David Cox', *Statistical Science*, 9, 439–455.

Silverman, B.W., Jones, M.C., Wilson, J.D., and Nychka, D.W. (1990), 'A Smoothed EM Approach to Indirect Estimation Problems, with Particular Reference to Stereology and Emission Tomography', *Journal of the Royal Statistical Society, Series B*, 52, 271–324.

Sy, J.P., and Taylor, J.M.G. (2000), 'Estimation in a Cox Proportional Hazards Cure Model', *Biometrics*, 56, 227–236.

Tai, P., Yu, E., Cserni, G., Vlastos, G., Royce, M., Kunkler, I., and Vinh-Hung, V. (2005), 'Minimum Follow-up Time Required for the Estimation of Statistical Cure of Cancer Patients: Verification Using Data From 42 Cancer Sites in the SEER Database', *BMC Cancer*, 5, 48.

Taylor, J.M.G. (1995), 'Semi-parametric Estimation in Failure Time Mixture Models', *Biometrics*, 51, 899–907.

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia: SIAM.

Wang, L., Du, P., and Liang, H. (2012), 'Two-component mixture cure rate model with spline estimated nonparametric components', *Biometrics*, 68, 726–735.

Xu, L., and Zhang, J. (2010), 'Multiple imputation method for the semiparametric accelerated failure time mixture cure model', *Computational Statistics & Data Analysis*, 54, 1808–1816.

Yakovlev, A.Y., and Tsodikov, A.D. (1996), *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, Hackensack, NJ: World Scientific.

Zeng, D., Yin, G., and Ibrahim, J.G. (2006), 'Semiparametric Transformation Models for Survival Data with a Cure Fraction', *Journal of the American Statistical Association*, 101, 670–684.

Zhang, J., and Peng, Y. (2007), 'A new estimation method for the semiparametric accelerated failure time mixture cure model', *Statistics in Medicine*, 26, 3157–3171.

Zhang, J., Peng, Y., and Li, H. (2013), 'Multiple imputation method for the semiparametric accelerated failure time mixture cure model', *Computational Statistics & Data Analysis*, 59, 95–102.