Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Promotion time cure rate model with nonparametric form of covariate effects

Tianlei Chen and Pang Du*

Survival data with a cured portion are commonly seen in clinical trials. Motivated from a biological interpretation of cancer metastasis, promotion time cure model is a popular alternative to the mixture cure rate model for analyzing such data. The existing promotion cure models all assume a restrictive parametric form of covariate effects, which can be incorrectly specified especially at the exploratory stage. In this paper, we propose a nonparametric approach to modeling the covariate effects under the framework of promotion time cure model. The covariate effect function is estimated by smoothing splines via the optimization of a penalized profile likelihood. Point-wise interval estimates are also derived from the Bayesian interpretation of the penalized profile likelihood. Asymptotic convergence rates are established for the proposed estimates. Simulations show excellent performance of the proposed nonparametric method which is then applied to a melanoma study. **Copyright © 2016 John Wiley & Sons, Ltd.**

Keywords: Confidence intervals, Convergence rates, Nonparametric covariate effects, Promotion time cure model, Smoothing spline ANOVA.

1. Introduction

In some lifetime studies, the population under consideration consists of 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 developing such events and can be regarded as "cured". Examples include cancer studies with long-term survivors, smoking studies with permanent quitters, employment studies with long-term employees. The common focus of these studies is on the assessment of covariate effects on the probability of being cured and the failure time distribution of the susceptible individuals.

Let $\mathbf{x}$ be a vector of covariates that the distribution of failure time T may depend on. Most models for cure rate data can be loosely put into two categories. One category is the two-component mixture cure models, where the distribution of T is assumed to be a mixture of two components and have a survival function $S \quad (t|\mathbf{x}) = \pi(\mathbf{x})S(t|\mathbf{x}) + 1 - \pi(\mathbf{x})$, with $\pi(\mathbf{x})$ and $S(t|\mathbf{x})$ being respectively the proportion and the survival function of susceptible subjects. Examples can be found in some recent papers [1, 2, 3, 4, 5] and the references within.

Celgene Corporation, 86 Morris Avenue, Summit, New Jersey 07901, U. S. A.
Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, U. S. A.
**Correspondence to: Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, U. S. A.*

The focus of this paper is on the other category of models, namely the promotion time cure models. In these models, the population survival function is assumed to have the form

$$S \quad (t|\mathbf{x}) = \exp\{-\theta(\mathbf{x})F(t)\}, \tag{1}$$

where F is an unknown distribution of a nonnegative random variable and $\theta > 0$ can be modeled as $\theta(\mathbf{x})$ to incorporate the covariate $\mathbf{x}$. Note that the cure proportion in (1) is $S \quad (\infty) = e \quad ^{\mathbf{x}}$. This type of model first appeared in [6] and [7] who also noted that model (1) provided a natural way to extend the proportional hazards regression model. [8] gave a biological interpretation of the form (1) and proposed a Bayesian analysis method with $\theta(\mathbf{x}) = \exp(\mathbf{x} \ \beta)$ for some unknown parameter vector β. They also pointed out a mathematical connection between (1) and the two-component mixture cure models. [9] considered a more general form of model $S \quad (t|\mathbf{x}) = G \ (\theta(\mathbf{x})F(t))$, where G , as γ varies, offers more transformation options than the exponential one in (1). However, $\theta(\mathbf{x})$ is still assumed to have a parametric form $g(\mathbf{x} \ \beta)$ for a known and strictly positive link function g. As reviewed in [10], the promotion time cure model has a couple of advantages over the standard mixture cure model: (i) It is naturally connected to the proportional hazards model, allowing for extensions to a wide class of hazard regression models. (ii) In some settings the model can be interpreted in terms of biologically meaningful parameters.

The existing promotion time cure models have a common limitation in that they all model covariate effects in a parametric form whose validity is generally not justified in practice. The strict parametric assumption can be particularly problematic at the exploratory stage of a study. This calls for more flexible nonparametric modeling of covariate effects. Under the framework of a mixture cure model, [5] propose a method where both $\pi(\mathbf{x})$ and $S(t|\mathbf{x})$ are estimated by nonparametric smoothing spline ANOVA models. However, their nonparametric approach cannot be directly transferred to the promotion time cure model setting. For example, their penalized EM algorithm won't apply here since the cure status is assumed known here and having it in the likelihood doesn't simplify the computation. And their asymptotic theory involves only spline function estimates of two functions, $\pi(\cdot)$ and $S(\cdot|\mathbf{x})$. In our model (1), there is also a nuisance parameter $F(t)$, which is estimated by discretizing into steps and has to be analyzed separately from the spline function estimate of $\theta(\mathbf{x})$.

As far as we know, our method is the first promotion time cure rate model with a nonparametric form of covariate effects. It offers the well-needed flexibility especially necessary at the exploratory stage of data analysis. As demonstrated by our application, the method can indeed identify trends that may be missed by traditional models with a linear form of covariate effects. Our development includes a penalized profile likelihood estimation procedure via smoothing splines, a reliable inference procedure involving point-wise confidence intervals, and a rigorously proved consistency theory. The simulations and the application to a melanoma study offer numeric evidences of the excellent performance of our method.

The rest of the article is laid out as follows. Section 2 elaborates the proposed method, with model setup in §2.1, an introduction to smoothing splines in §2.2, penalized profile likelihood estimation in §2.3, point-wise confidence intervals in §2.4, and asymptotic theory in §2.5. Section 3 contains numerical studies including simulations and an application to a melanoma study. Discussion in Section 4 concludes the paper.

2. Methods

2.1. Model

Let $(t \ , \delta \ , \mathbf{x} \ )$ be the observed data for the ith subject, $i = 1, \ldots, n$. Here t is the observed lifetime time for the ith subject, δ is an indicator with $\delta = 1$ for observed failures and $\delta = 0$ for censored subjects, and $\mathbf{x}$ are the covariates. Following [9], we assume that the follow-up time for a cured subject is infinite and thus distinguishable, and that only the susceptible subjects can experience censoring. In practice, this is achieved by choosing a *cure threshold* such that

all censored observations beyond the threshold are treated as $t = \infty$ (i.e., observed to be cured) and all observations lower than the threshold are treated as $t < \infty$ (i.e., be susceptible). A common choice of this threshold is the largest observed failure time in the data; see, e.g., [9] and [11]. Our sensitivity analysis in Section 3.1 indicates that this choice of threshold is reasonably robust. Therefore, unless specified otherwise we will use it in all the data analysis in the paper. Let $\eta(\mathbf{x}) = \log\{\theta(\mathbf{x})\}$. Assuming independent and non-informative censoring for susceptible subjects, the observed likelihood function can be written as

$$l \quad (\eta(\cdot), F(\cdot)) \propto \prod \Big\{ \big([\exp\{-e^{\ \mathbf{x}_i} F(t\ )\} e^{\ \mathbf{x}_i} f(t\ )]^{\ i} [\exp\{-e^{\ \mathbf{x}_i} F(t\ )\}]^{\ i}\big)^{\ i} \\ [\exp\{-e^{\ \mathbf{x}_i}\}]^{\ i} \Big\}, \quad (2)$$

where $f(\cdot)$ is the probability density function corresponding to F. Then our estimator $(\hat{\eta}, \hat{F})$ is defined as the minimizer of

$$-\frac{1}{n} \log\{l \quad (\eta(\cdot), F(\cdot))\} + \lambda J(\eta), \quad (3)$$

where J is the roughness penalty on η and $\lambda > 0$ is the smoothing parameter. The minimization with respect to η is performed on a reproducing kernel Hilbert space $\mathcal{H}$ of functions.

2.2. *Smoothing spline estimation*

Smoothing splines will be used to estimate the covariate effect function $\eta(\mathbf{x})$ in (3). This nonparametric way of estimating η distinguishes us from the rest of the literature on promotion cure rate models where a linear form of covariate effect, $\eta(\mathbf{x}) = \mathbf{x}\ \beta$, is often used. Therefore, in this section we give a short generic review of the smoothing spline estimation procedure derived from a penalized likelihood like (3).

Given stochastic data "generated" according to an unknown "pattern" function g , the smoothing spline estimate of g is defined as the minimizer of the penalized likelihood (PL): $L(g|\text{data}) + -J(g)$. Here $L(g)$, usually the negative log likelihood, measures the goodness-of-fit of g, $J(g)$, the roughness penalty, measures the smoothness of g, and the smoothing parameter $\lambda(> 0)$ controls the trade off. The minimization of a PL is carried out in a reproducing kernel Hilbert space (RKHS) of functions. RKHS provides a theoretical basis for smoothing spline ANOVA (SSANOVA) model and unified framework for modeling various data. For a multivariate g, it can be decomposed into main effects and interactions similar to the classical ANOVA decomposition. Through proper specifications of g and $J(g)$ in a variety of problem settings, the PL yields nonparametric models for Gaussian and non-Gaussian regression, probability density estimation, hazard rate estimation, etc. See [12] for examples. In this paper, we use cubic and tensor product cubic smoothing splines for estimation whose detailed configurations involving reproducing kernel Hilbert spaces are given below. More options are available in Chapter 2 of [12].

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 $\mathcal{N}\ = \{\eta : J(\eta) = 0\} \subset \mathcal{H}$. 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 $\mathcal{N}$ when restricted therein. There exists a tensor sum decomposition $\mathcal{H} = \mathcal{N}\ \oplus \mathcal{H}$, where the space $\mathcal{H}$ has $J(\eta)$ as its square norm and an RK R satisfying $J(R\ (\mathbf{x}, \cdot), f(\cdot)) = f(\mathbf{x}), \forall f \in \mathcal{H}$. See, e.g., Section 2.1 of [12].

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 2.2. Also note that the range of a continuous variable is taken to be $[0, 1]$ here only for the simplification of notation. For a variable residing on a general compact interval $[a, b]$, a transformation

of $(\cdot - a)/(b-a)$ can bring the variable to the range of $[0,1]$.

Example 2.1 (Cubic Spline) *Without loss of generality assume $\mathcal{X} = [0,1]$ for a univariate x. A choice of $J(\eta)$ is $\int (\eta\ )\ dx$, which yields the popular cubic splines. If the inner product in $\mathcal{N}$ is $(\int f\,dx)(\int g\,dx) + (\int f\ dx)(\int g\ dx)$, then $\mathcal{H} = \mathcal{H} \ominus \mathcal{N} = \{\eta : \int \eta dx = \int \eta\,dx = 0, J(\eta) < \infty\}$ and the reproducing kernel $R(x, x) = k(x)k(x) - k(|x - x|)$, where $k(x) = B(x)/\nu!$ are scaled Bernoulli polynomials for $x \in [0,1]$. The null space $\mathcal{N}$ has a basis $\{1, k(x)\}$ of $m = 2$ functions, where $k(x) = x - 0.5$ for $x \in [0,1]$. See Section 2.3.3 of [12].* □

Example 2.2 (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, which would also be the space corresponding to $\mathcal{H}$ in Example 2.1 if U were the only covariate. 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 $\mathcal{H} = \mathbb{R}^l$. When u is a nominal variable, that is, its levels are not ordered, let $\bar{\eta} = \sum \eta(u)/l$. Equipped with the roughness penalty $J(\eta) = \sum [\eta(u) - \bar{\eta}]$ and inner product $\langle f, g\rangle = \sum f(u)g(u)$, the RKHS $\mathcal{H}$ decomposes as

$$\mathcal{H} = \mathcal{H} \oplus \mathcal{H} = \{\eta : \eta(1) = \cdots = \eta(l)\} \oplus \left\{\eta : \sum \eta(u) = 0\right\}$$

with reproducing kernels $R(u, u) = 1/l$, $R(u, u) = I_{u_1 = u_2} - 1/l$.

On the other hand, the construction in Example 2.1 gives a decomposition of the RKHS $\mathcal{H}$ on the domain $\mathcal{Z}$

$$\begin{aligned}\mathcal{H} &= \left\{\eta : \int (\eta\ )\ dz < \infty\right\} = \mathcal{H} \oplus \mathcal{H} \oplus \mathcal{H} \\ &= span\{1\} \oplus span\{k(z)\} \oplus \left\{\eta : \int \eta dz = \int \eta\,dz = 0,\ \int (\eta\ )\ dz < \infty\right\},\end{aligned}$$

with reproducing kernels $R(z, z) = 1$, $R(z, z) = k(z)k(z)$, and $R(z, z) = k(z)k(z) - k(|z - z|)$. The tensor product of $\mathcal{H}$ and $\mathcal{H}$ yields six tensor sum terms $\mathcal{H} = \mathcal{H} \otimes \mathcal{H}$ on $\mathcal{Z} \times \mathcal{U}$, $\nu = 00, 01, 1$ and $\mu = 0, 1$, with reproducing kernels $R(x, x) = R(z, z)R(u, u)$, where $x = (z, u)$. The two subspaces with $\nu = 00, 01$ are of one-dimension each and can be lumped together as the null space $\mathcal{N}$ (thus $m = 2$). The other four subspaces form $\mathcal{H}$ with the reproducing kernel

$$R = R + R + R + R,$$

where θ 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

$$\eta(z, u) = \eta + \eta(u) + \eta(z) + \eta(z, u)$$

for functions on $\mathcal{X}$, with $\eta \in \mathcal{H} \otimes \mathcal{H}$ being the constant term, $\eta \in \mathcal{H} \otimes \mathcal{H}$ the u main effect, $\eta \in \{\mathcal{H} \oplus \mathcal{H}\} \otimes \mathcal{H}$ the z main effect, and $\eta \in \{\mathcal{H} \oplus \mathcal{H}\} \otimes \mathcal{H}$ the interaction. See, e.g., Example 2.7 of [12]. □

By the Representer Theorem [13], although $\mathcal{H}$ is of infinite dimensions, the minimizer of the PL in some settings actually resides in the finite dimensional subspace $\mathcal{H} \ominus \text{span}\{R(\mathbf{x}, \mathbf{x}_1), \ldots, R(\mathbf{x}, \mathbf{x}_n)\}$, where $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 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 $\mathcal{H}$ [12]. Furthermore, [14] showed that instead of having n knots, the number of knots can be reduced to the order of n without losing any efficiency. They suggested using $10n$ knots in practice, which we shall follow in our computation.

2.3. *Penalized profile likelihood estimation*

The optimization of (3) is carried out in a profiled fashion. We first study the computation of F that minimizes (3) when η is fixed. Note that due to estimability, F can only be estimated as a step function making jumps at the observed failure times t . Order the observed failure times to obtain $t \quad < \cdots < t \quad$, where N is the number of observed failure times. Denote the corresponding jump size of F at $t \quad$ by $p \quad , j = 1, \ldots, N$, and let $F \; = \sum_{j \quad i} p \quad$. Then F, in term of its jump sizes $p \quad$, is obtained from

$$\text{minimize} \sum \log p \quad - \sum I(t \; < \infty) F \, e^{\; \mathbf{x}_i}, \quad \text{subject to} \sum p \quad = 1.$$

Simple derivation yields the recursive formula

$$\frac{1}{p} = \frac{1}{p} - e^{\; \mathbf{x}_{(i)}} - \sum_{(i) \quad j \quad (i+1)} e^{\; \mathbf{x}_j}. \tag{4}$$

Hence, from (4), we can treat $\eta, \alpha \equiv p \quad > 0$, and the Lagrange multiplier ρ introduced below as independent parameters and $p \quad , \ldots, p \qquad$ as functions of η and α. We want to minimize $-\log\{l \quad (\eta, \alpha)\} + \lambda J(\eta)$ under the constraint $\sum \quad p \quad = 1$, which is equivalent to minimize

$$-\frac{1}{n} \log\{l \quad (\eta, \alpha)\} + \lambda J(\eta) + \rho(\sum p \quad - 1) \tag{5}$$

with respect to (η, α, ρ). Furthermore, the likelihood function $l \quad$ depends on η only through its evaluation at the $\mathbf{x}$. So the Representer Theorem in [13] says $\hat{\eta}$, as the minimizer of (5), must have the expression $\eta(\mathbf{x}) = \sum \quad d \; \phi \; (\mathbf{x}) + \sum \quad c \; R \; (\mathbf{x} \; , \mathbf{x})$ for some coefficients $d \; , \ldots, d \quad$ and $c \; , \ldots, c \quad$, where $\phi \; (\cdot), \ldots, \phi \quad (\cdot)$ form a basis for the null space of J and $R \; (\cdot, \cdot)$ is the reproducing kernel induced by J. Write $\mathbf{b} = (d \; , \ldots, d \quad , c \; , \ldots, c \;)$ and $\psi = (\phi \; (\cdot), \ldots, \phi \quad (\cdot), R \; (\mathbf{x} \; , \cdot), \ldots, R \; (\mathbf{x} \; , \cdot))$. Note that now the penalty $J(\eta) = \mathbf{b} \quad Q\mathbf{b}$, where $Q = \text{diag}(\mathrm{O} \quad , Q \;)$ with $\mathrm{O} \quad$ being the $m \times m$ square matrix of zeros and $Q \quad$ being an $n \times n$ matrix with the (i, j)-th entry $R \; (\mathbf{x} \; , \mathbf{x} \;)$. Then the estimation of (η, α) reduces to solving the following score equations for $\mathbf{b}, \alpha$, and ρ.

$$\begin{aligned}
0 = & \frac{1}{n} \Big[\sum \{ \psi(\mathbf{x} \quad) F \quad + \frac{\partial}{\partial \mathbf{b}} F \quad \} \exp\{ \psi \quad (\mathbf{x} \quad) \mathbf{b} \} \\
& + \sum_{(i)} \sum_{j \quad (i+1)} \{ \psi(\mathbf{x} \;) F \quad + \frac{\partial}{\partial \mathbf{b}} F \quad \} \exp\{ \psi \quad (\mathbf{x} \;) \mathbf{b} \} - \sum \frac{1}{p} \frac{\partial}{\partial \mathbf{b}} p \quad - \sum \delta \; \psi(\mathbf{x} \;) \\
& - \sum I(t \quad = \infty) \psi(\mathbf{x} \;) \exp\{ \psi \quad (\mathbf{x} \;) \mathbf{b} \} \Big] + 2\lambda Q \mathbf{b} + \rho \sum \frac{\partial}{\partial \mathbf{b}} p \quad , \\
0 = & \sum \exp\{ \psi \quad (\mathbf{x} \quad) \mathbf{b} \} \frac{\partial}{\partial \alpha} F \quad + \sum_{(i)} \sum_{j \quad (i+1)} \exp\{ \psi \quad (\mathbf{x} \;) \mathbf{b} \} \frac{\partial}{\partial \alpha} F \\
& - \sum \frac{1}{p} \frac{\partial}{\partial \alpha} p \quad + \rho \sum \frac{\partial}{\partial \alpha} p \quad , \\
0 = & \sum p \quad - 1.
\end{aligned} \tag{6}$$

When λ is fixed, the equations can be solved by the Newton-Raphson algorithm after eliminating the Lagrange multiplier

ρ from the first two equations in (6), and the derivatives of p with respect to $\mathbf{b}$ and α can be computed using the recursive formula (4).

For the selection of λ, the classical way is to optimize a score derived from the Kullback-Leibler distance following the same principle as the well-known generalized cross validation (GCV); see, e.g., [13] and [12]. However, this approach does not seem to work well here due to the complication of having a extra nuisance parameter F. So we resort to the standard K-fold cross validation with $K = 5$ to find the optimal λ that minimizes the average log likelihood values on testing data sets.

2.4. *Inference*

We now derive an interval estimate of η based on a Bayes model interpretation of the penalized likelihood (3). The intuition is to treat (3) as the logarithm of a posterior likelihood where the roughness penalty acts like a prior distribution on the function parameter η, or essentially a prior distribution on the coefficient vector $\mathbf{c}$. Then our parameter estimates, as the minimizer of (3), become the posterior mode after the uniform priors are assigned to $\mathbf{d}$ and α. And their confidence intervals can be derived when the posterior distribution is approximated by a Gaussian distribution through a quadratic approximation of the log posterior likelihood.

Recall that $\eta(\mathbf{x}) = \sum d\ \phi\ (\mathbf{x}) + \sum c\ R\ (\mathbf{x}, \mathbf{x}) = \phi\ (\mathbf{x})\mathbf{d} + \boldsymbol{\xi}\ (\mathbf{x})\mathbf{c}$, where $\mathbf{d} = (d, \ldots, d)$, $\mathbf{c} = (c, \ldots, c)$ and $\boldsymbol{\xi} = (R\ (\mathbf{x}, \cdot), \ldots, R\ (\mathbf{x}, \cdot))$. Also the penalty $J(\eta) = \mathbf{c}\ Q\ \mathbf{c}$ with Q being an $n \times n$ matrix with the (i, j)-th entry $R\ (\mathbf{x}, \mathbf{x})$. Therefore (3) can be viewed as the posterior likelihood when one assigns the uniform prior to $\mathbf{d}$ and α, and the Gaussian prior $N(\mathbf{0}, (n\lambda)\ Q\ )$ to $\mathbf{c}$, where Q is the generalized inverse of Q. Let $(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha})$ be the coefficients corresponding to the minimizer $(\hat{\eta}, \hat{F})$ of (3) and rewrite (3) as $\mathrm{PL}(\mathbf{d}, \mathbf{c}, \alpha) = -\log\{l_{\mathrm{obs}}(\mathbf{d}, \mathbf{c}, \alpha)\} + \lambda \mathbf{c}\ Q\ \mathbf{c}$. Then $(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha})$ becomes the posterior mode and the asymptotic covariance of $(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha})$ can be calculated as

$$H = \begin{pmatrix} -\frac{\partial^2}{\partial \mathbf{d}\, \partial \mathbf{d}^T}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) & -\frac{\partial^2}{\partial \mathbf{d}\, \partial \mathbf{c}^T}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) & -\frac{\partial^2}{\partial \mathbf{d}}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) \\ -\frac{\partial^2}{\partial \mathbf{c}\, \partial \mathbf{d}^T}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) & -\frac{\partial^2}{\partial \mathbf{c}\, \partial \mathbf{c}^T}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) & -\frac{\partial^2}{\partial \mathbf{c}}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) \\ -\frac{\partial^2}{\partial \mathbf{d}^T}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) & -\frac{\partial^2}{\partial \mathbf{c}^T}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) & -\frac{\partial^2}{\partial}\mathrm{PL}(\hat{\mathbf{d}}, \hat{\mathbf{c}}, \hat{\alpha}) \end{pmatrix}$$

For any $\mathbf{x}$, we can construct the $100(1-\alpha)\%$ confidence interval of $\eta(\mathbf{x})$ as $\hat{\eta}(\mathbf{x}) \pm z\ \ s(\mathbf{x})$, where z is the $(1-\alpha/2)$-quantile of the standard normal distribution and $s(\mathbf{x})$ is the standard error function with $s\ (\mathbf{x}) = (\phi\ (\mathbf{x}), \boldsymbol{\xi}\ (\mathbf{x}), 0) H\ (\phi\ (\mathbf{x}), \boldsymbol{\xi}\ (\mathbf{x}), 0)$.

2.5. *Asymptotic Properties*

In this section, we will present the convergence rates of our function estimates $\widehat{\eta}$ as well as the consistency of the parameter estimate $\widehat{F}$. Its technical proof is in the Web Appendix A. Let $\eta\ (\mathbf{x})$ and F be the true parameters, and r be the constant associated with $\mathcal{H}$ that measures the smoothness level enforced by the function space. A typical value for r is $2m$ when order-m splines are used for modeling η. Then we have the following theorem.

Theorem 2.1 *Under Conditions A1-A6, we have* $\|\widehat{\eta} - \eta\| = O\ (n\ \ )$ *and* $\sup |\widehat{F}(t) - F\ (t)| = O\ (n\ \ )$, *where* $\|\cdot\|$ *is the* L *-norm.*

Note that this is the optimal convergence rate of spline estimates when splines of order $r/2$ are used for the estimation of η. Also note that the result for $\widehat{F}$ can be refined to have a $\sqrt{n}$-rate and asymptotic normality using the technique in Section 21.5 of [15]. But we choose not to pursue such refinement since F is considered a nuisance parameter here.

3. Numerical Studies

3.1. Simulations

3.1.1. Estimation and inference performance In this section we present some simulations to demonstrate the performance of the proposed promotion time cure model. We shall present most of the results in terms of the covariate effect function $\eta(x)$. Note that the promotion cure rate model in (1) indicates that the cure proportion and the hazard rate function are respectively $\exp\{-e \quad\}$ and $f(t)e \quad$. Clearly the cure proportion is strictly decreasing and the hazard rate function is strictly increasing with respect to η. Therefore, the trend of the η function over the covariate x is directly transferrable to the cure proportion (in the opposite direction) and the hazard rate function.

To generate failure times according to the promotion time cure model (1), we set $\eta(x) = 0.1\{10\ x\ \ (1-x)\ + 10\ x\ (1-x)\ \ \}$ and $F(t) = 1 - \exp(-t)$, the distribution function of the exponential distribution with rate 1. The covariate x took values on an equally-spaced grid of size n over the interval $[0, 1]$. We considered two total sample sizes $n = 400$ and 800. In both cases, censoring times were generated from uniform distributions with parameters tuned to yield an overall censoring rate around 50%. These choices yielded largest observed failure times around 3.0 in our simulated data. One hundred data replicates were generated for each sample size setting. For each data replicate, we computed the estimates for η and F, as well as the 95% point-wise confidence intervals of $\eta(x)$ on an equally-spaced grid of size 200 on $[0, 1]$.

Figure 1 shows the simulation results for the two total sample sizes $n = 400$ and 800. We can see that the mean function estimates are close to the true function η and the average point-wise confidences intervals match well with the empirical quantiles. Furthermore, as the sample size increases from 400 to 800, the mean function estimate becomes more accurate and the confidence intervals get narrower. The empirical coverages are generally okay, with slight under-coverage around the areas where the true η function has high curvatures. This is reasonable since higher curvature means more difficulty in accurate estimation, a phenomenon exposed in numerous literatures on various nonparametric smooth estimation problems; see, e.g., [16]. One possible remedy for the under-coverage is to place more knots in the areas of high curvatures and another more rigorous remedy would be to extend the local asymptotic inference theory in [17] for smoothing spline regression to our scenario of promotion cure rate model. But the numerical implementation of the former can be tricky and the theoretical derivation for the latter is very hard. The step function estimates of the function $F(t)$ also appear to be pretty good. In conclusion, our estimation and inference procedures really work well in the simulations.

3.1.2. Sensitivity of cure threshold The proposed method requires the specification of a cure threshold such that subjects survived beyond the threshold are claimed as cured. The threshold we use in the paper is the largest observed failure time, following the suggestions in the literature [9, 11]. In this section, we perform some sensitivity analysis on this choice of threshold. The data generation processes were the same as in the previous section with $n = 400$. Note that the largest observed failure times in our simulated data were around 3.0. Also a threshold smaller than the largest failure time is not a reasonable choice. So we created three more thresholds by respectively adding 0.5, 1 and 1.5 to our choice of threshold, the largest observed failure time. For each choice of threshold, we applied the proposed promotion time cure model to 100 data replicates. Figure 2 plotted the estimates of η and F for the four cure thresholds. As expected the estimate of η increased and the estimate of F decreased as the cure threshold increased. However, the increase and decrease were both mild indicating that the estimates are not too sensitive to the choice of cure threshold so long as it is within a reasonable range.

3.1.3. Comparison with mixture cure model Two-component mixture cure rate model and promotion time cure model are two popular choices for cure rate data analysis. Hence, in this section we present some simulation comparisons between our proposed promotion time cure model with the nonparametric mixture cure rate model in [5]. We considered two scenarios with the true model being respectively promotion time cure model and mixture cure rate model. For

the former scenario, the true functions η and F were the same as those in Section 3.1.1. For the latter scenario we adopted a simulation setting in [5]. Specifically, the data were generated from the mixture cure rate model $S \quad (t|x) = \pi(x)S(t|x) + 1 - \pi(x)$ with x from a equally-spaced grid over the range $[-0.4, 0.4]$, $\pi(x) = 0.1722 + 0.7\sin\{2(x + 0.6)\}$ and $h(t|x) = 2.5t \quad \{1 + 0.5\sin(2\pi x)\}$, where $h(t|x)$ is the hazard rate function corresponding to $S(t|x)$. We used $n = 400$ and 100 data replicates for each scenario. Note that $S \quad (t|x) = \exp\{-e \quad F(t)\}$ for promotion time cure model. Therefore to compare the estimation performances of the two methods, we computed for each data replicate the mean squared errors MSE $= n \quad \sum \quad \{\hat{S} \quad (t|x) - S \quad (t|x)\}$ for both methods. Figure 3 shows the box-plots of these MSEs under the scenarios of two different true models. Note that the MSEs in the two plots are not comparable since they were calculated from different true models, that is, the true survival functions used in the two sets of MSEs were completely different. From the two plots, one can only conclude that the right model will produce a smaller MSE. Therefore, neither of the models can be considered better than the other. In practice, it is wise to apply both models and look for common trends in covariate effects.

3.2. Application to 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 treatments including surgery and radiotherapy. Note that the data in the SEER were mostly collected from observational studies, meaning that the treatments were not randomly assigned in the studies. Hence it was not feasible to assess the treatment effect and we actually had to restrict our analysis within one treatment group in criterion 3 to remove the influence of the treatment effect here. 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). Among the 635 observations, there were 233 observed failures ranging from 1 to 129 months. The censoring times ranged from 1 to 226 months, with 118 of them larger than 129 months. In Figure 4 we also plotted the Kaplan-Meier estimates for patients younger than 55 years old, patients at least 55 years old, and all the patients, ignoring the other covariate information. These Kaplan-Meier curves all showed a clear plateau at the end of the observation interval. This suggests the possible presence of a subpopulation of cured subjects in the data as commonly seen in melanoma studies and justifies that a cure rate data analysis is appropriate here.

We applied the proposed promotion time cure model to the melanoma data with the covariate $\mathbf{x} =$ (age, gender, size). For η, we estimated it by tensor product smoothing splines with all the main effects and interactions, similar to the tensor product splines in Example 2.2 but with one more categorical variable. Following the suggestions in the literature [9, 11], we chose the largest failure time (129 months) in the data as the cure threshold such that any subjects who were censored beyond the threshold were treated as being cured. To assess the sensitivity of this choice, we also conducted the analysis with three more different thresholds: 139 months, 149 months, and 159 months. These four cure thresholds respectively yielded 118, 105, 87, and 70 cured patients. Figure 5 plotted the function estimates of η and F for these four different cure thresholds. Also plotted were the point-wise confidence intervals for the function η of the four patient groups when the cure threshold was 129 months. As demonstrated in the sensitivity simulations, an increase of the cure threshold only caused a slight increase in the estimate of η and a slight decrease in the estimate of F. Hence we draw all our following conclusions based on the analysis results with the cure threshold of 129 months.

Recall again that the cure proportion and the hazard rate function are respectively $\exp\{-e \quad\}$ and $f(t)e \quad$. The function η for three of the four patient groups showed approximately linear increasing trends against age, which meant a double-exponentially decaying trend of the cure proportion and an exponentially increasing hazard rate against age. The trends of both male groups are similar with the big size tumor male group generally having a higher value of η, or a smaller cure proportion and a higher hazard rate. The small size tumor female group showed a much steeper increasing trend of η,

or a much faster decaying trend of cure proportion and a much steeper increasing trend of hazard rate, over age than the male groups. The exception of the four groups, the big tumor female group, showed a strong nonlinear trend which slightly decreased over age at the beginning and then bounced up around age 60. This nonlinear fit also highlights a major benefit of our nonparametric modeling of η since the traditional linear modeling of η would most likely miss this interesting revelation. A version of the analysis results are plotted as cure rate function and survival function estimates in the Web Appendix B. It is also interesting to compare the analysis results with those in [5] where a mixture cure rate model with spline estimated nonparametric components was used. We can see that the general increasing/decreasing trends of cure proportion and hazard rate against age were well matched among all four patient groups between these two nonparametric methods.

4. Discussion

In this paper, we have developed a version of promotion time cure rate model where the covariate effects are modeled in a nonparametric form. As demonstrated in our application, this offers the flexibility that is often necessary at the exploratory stage of data analysis. A quick extension of the work can be the incorporation of a more complicated transformation function G than the exponential function considered here. It can surely offer more flexibility to the model. Our new method also complements the earlier work of [5] by offering a reliable nonparametric exploratory tool for analyzing cure rate data under the framework of promotion time. It inherits all the advantages of promotion time cure model against mixture cure model, namely a natural connection to the proportional hazards model and a set of more biologically meaningful parameters. However, as demonstrated in the simulations neither model can replace the other. In practice both models should be fitted to seek interpretable common trends in covariate effects.

One potential limitation we have noticed of the new method is that the point-wise confidence interval coverage drops quickly as the sample size becomes smaller than 100-200. Hence caution is needed when applying the proposed method, especially the point-wise confidence intervals, to cure rate data with a small sample size.

At last we would like to suggest a general guidance on the choices of models when cure rate data analysis is relevant. Cure rate models are by no means something appropriate for every survival data set. To be qualified for such kind of analysis, there must be strong clinical evidence of high proportion of cure for the disease and a statistical routine of reviewing the Kaplan-Meier survival curve estimates must be performed to look for early plateaus that last long after the last observed failure. Once a cure proportion is justified, it is better to first try some cure rate models with nonparametric covariate effects, such as the methods in [5] and this paper, due to their flexible model assumptions. If the estimates from these nonparametric models are approximately linear, a cure rate model with linear covariate effect can be applied, which can often result in more powerful inference when the true model is approximately linear. The choice between mixture cure rate model and promotion time cure model, however, is likely to be case by case. The theoretical comparison between them surely merits further research. A sensitivity analysis is strongly recommended when a manually selected cure threshold is required for fitting a cure rate model. This is to safeguard the scenario when a large number of censoring times are beyond the largest observed failure time but not too far away from it. In other words, the follow-up time should be long enough to claim the appropriateness of any cure rate data analysis.

Acknowledgments

The authors thank the Associate Editor and two reviewers for their constructive comments. Du's research was supported by the National Science Foundation grant DMS-1620945.

References

1. Lu W, Ying Z. On semiparametric transformation cure models. *Biometrika* 2004; **91**:331–343.
2. Fang H, Li G, Sun J. ML estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics* 2005; **32**:59–75.
3. Zhang J, Peng Y. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine* 2007; **26**:3157–3171.
4. Othus M, Li Y, Tiwari RC. A class of semiparametric mixture cure survival models with dependent censoring. *Journal of the American Statistical Association* 2009; **104**:1241–1250.
5. Wang L, Du P, Liang H. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics* 2012; **68**:726–735.
6. Yakovlev AY, Tsodikov AD. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific: Hackensack, NJ, 1996.
7. Tsodikov AD. A proportional hazards model taking account of long-term survivors. *Biometrics* 1998; **54**:1508–1516.
8. Chen MH, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999; **94**:909–919.
9. Zeng D, Yin G, Ibrahim JG. Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association* 2006; **101**:670–684.
10. Tsodikov AD, Ibrahim JG, Yakovlev AY. Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association* 2003; **98**:1063–1078.
11. Ma Y, Yin G. Cure rate model with mismeasured covariates under transformation. *Journal of the American Statistical Association* 2008; **103**(482):743–756.
12. Gu C. *Smoothing Spline ANOVA Models (2nd Ed.)*. Springer-Verlag: New York, 2013.
13. Wahba G. *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. SIAM: Philadelphia, 1990.
14. Kim YJ, Gu C. Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B* 2004; **66**(2):337–356.
15. Kosorok MR. *Introduction to empirical processes and semi parametric inference*. Springer-Verlag: New York, 2008.
16. Nychka D. Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association* 1988; **83**:1134–1143.
17. Cheng G, Shang Z. Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics* 2013; **41**(5):2608–2638.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.

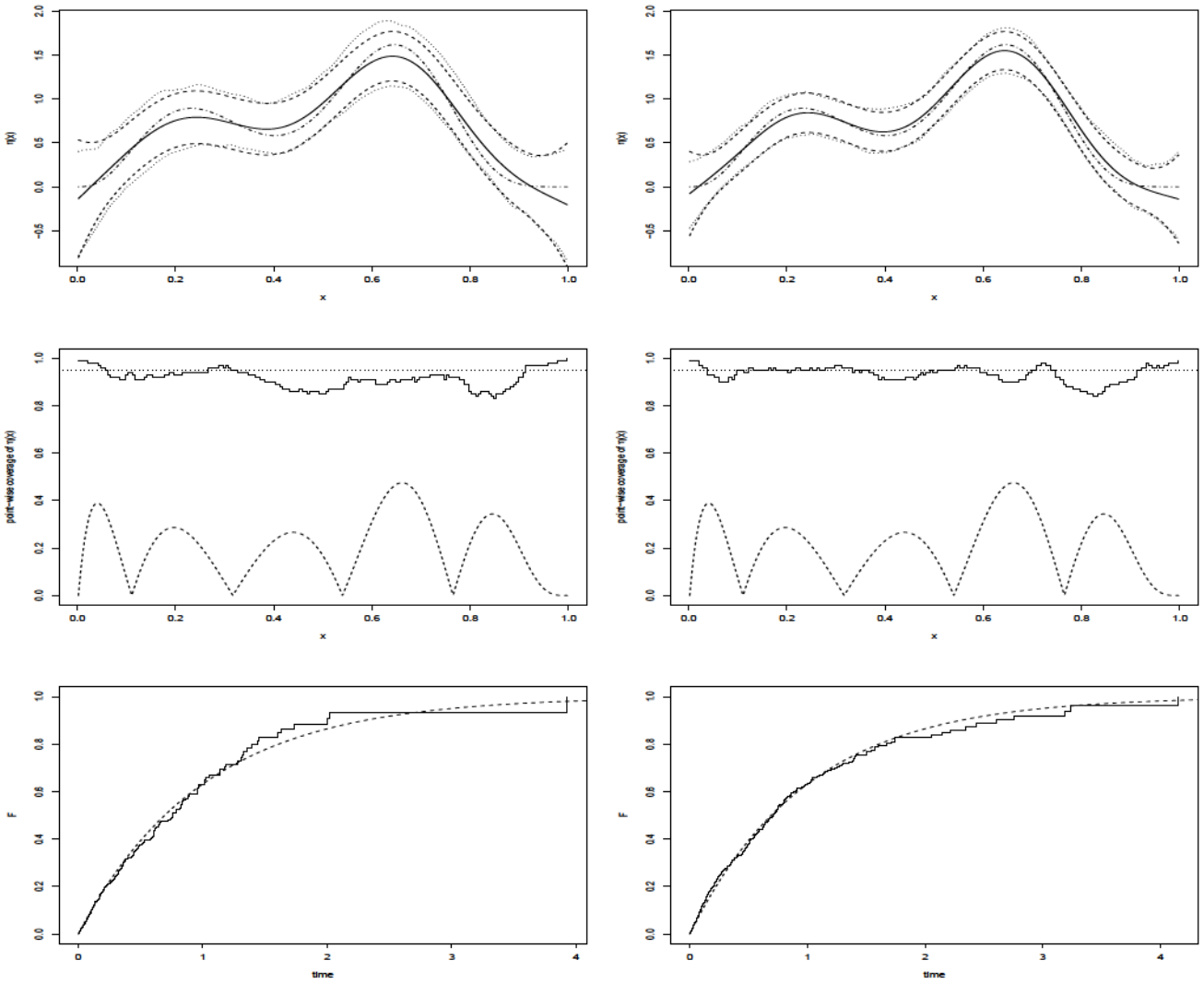

Figure 1. Simulation plots for estimation and inference: left frames for $n = 400$ and right frames for $n = 800$. Top frames: the true η function (dash-dotted lines), the mean function estimates (solid lines), the averages of 95% point-wise CIs (dashed lines), and the empirical 2.5% and 97.5% percentiles of point-wise function estimates (dotted lines). Middle frames: empirical point-wise coverages of the CIs for $\eta(x)$; curves showing the magnitude of the curvature of the true η function. Bottom frames: The step function estimates (solid lines) of the distribution function F (dashed lines) for a randomly selected data replicate.

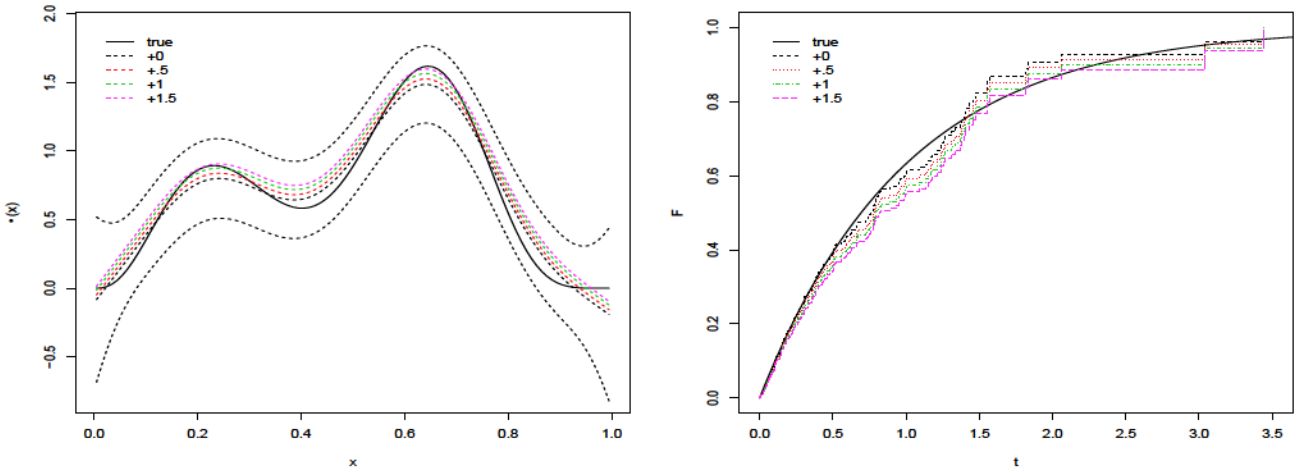

Figure 2. Simulation plots for threshold sensitivity. Left frame: the true η function (solid black), the mean estimates of η with cure threshold equal to largest failure time $+0$ (middle black dashed), $+0.5$ (red dashed), $+1$ (green dashed), and $+1.5$ (purple dashed), and the averages of 95% point-wise CIs (top and bottom black dashed lines) when cure threshold was the largest failure time. Right frame: the true F function (solid black) and the estimates of F from a random sample with cure threshold equal to largest failure time $+0$ (black dashed), $+0.5$ (red dotted), $+1$ (green dot-dashed), and $+1.5$ (purple wide-dashed).

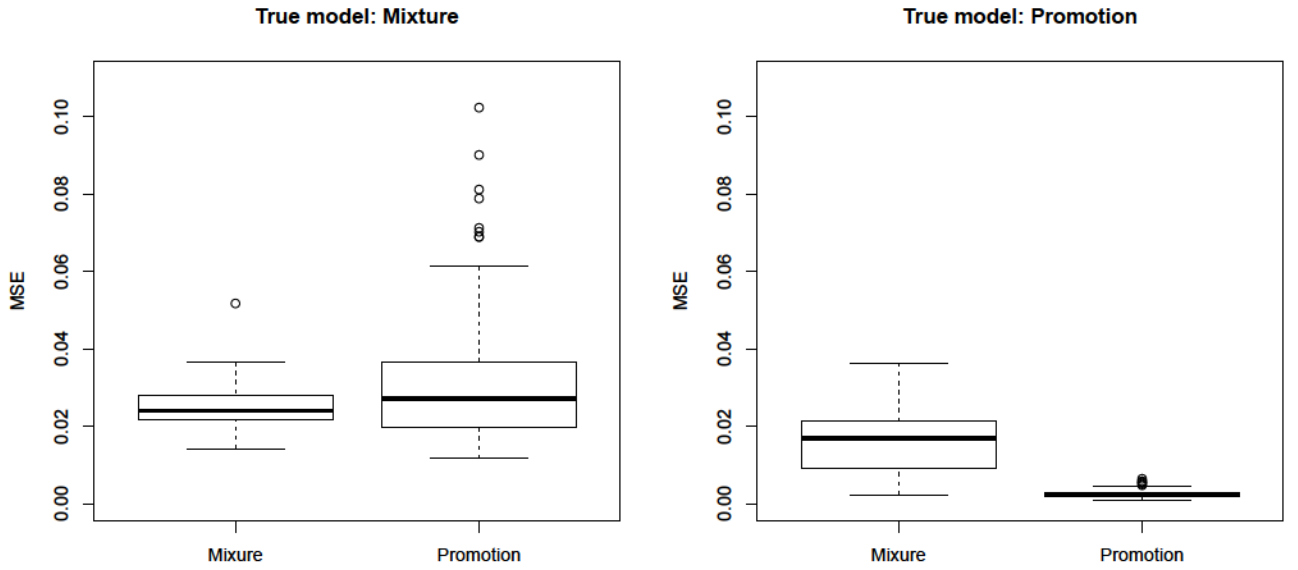

Figure 3. Simulation plots for model comparisons: Box-plots of MSEs for estimating $S_{\text{pop}}(t|x)$. "Mixture" stands for the estimates from the mixture cure rate model in [5] and "Promotion" stands for the estimates from the proposed promotion time cure model. Left frame: True model was mixture cure rate model; Right frame: True model was promotion time cure model.

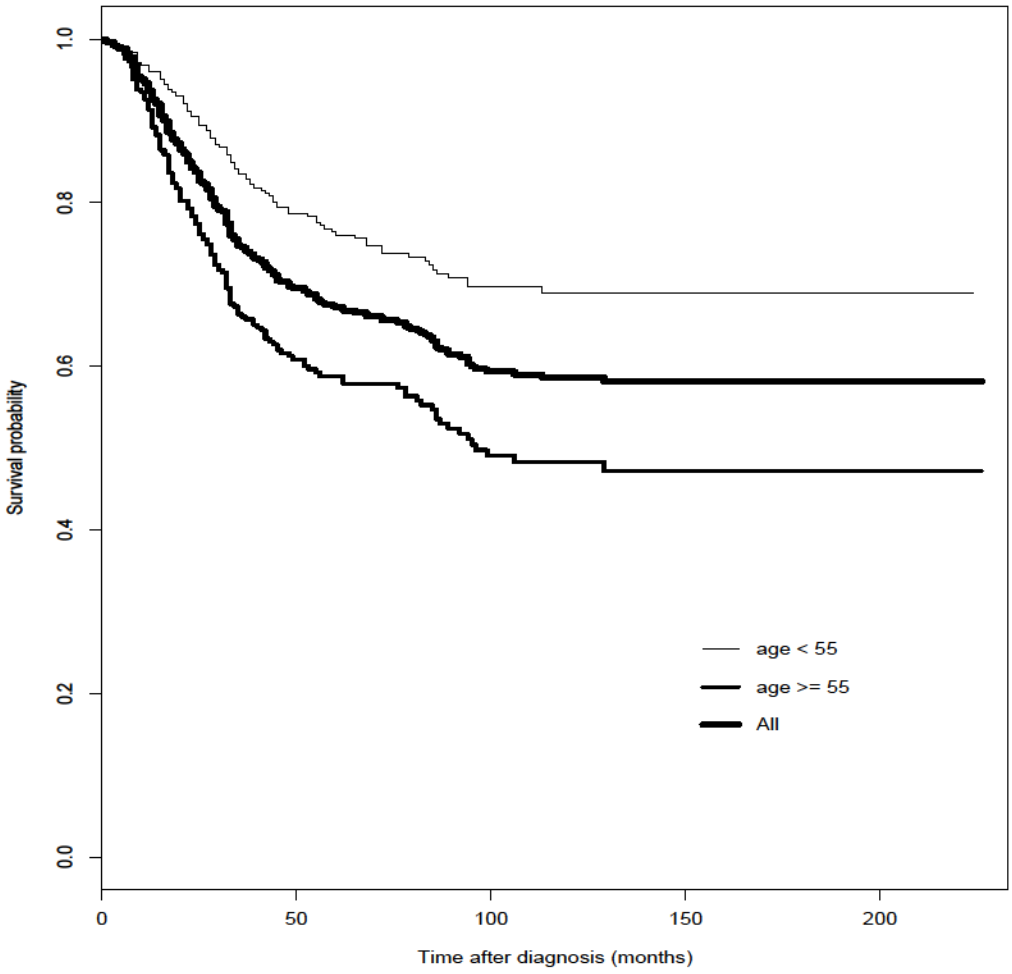

Figure 4. Melanoma Study: Kaplan-Meier plots for patients with age (thin line), patients with age (thicker line), and all the patients (thickest line).

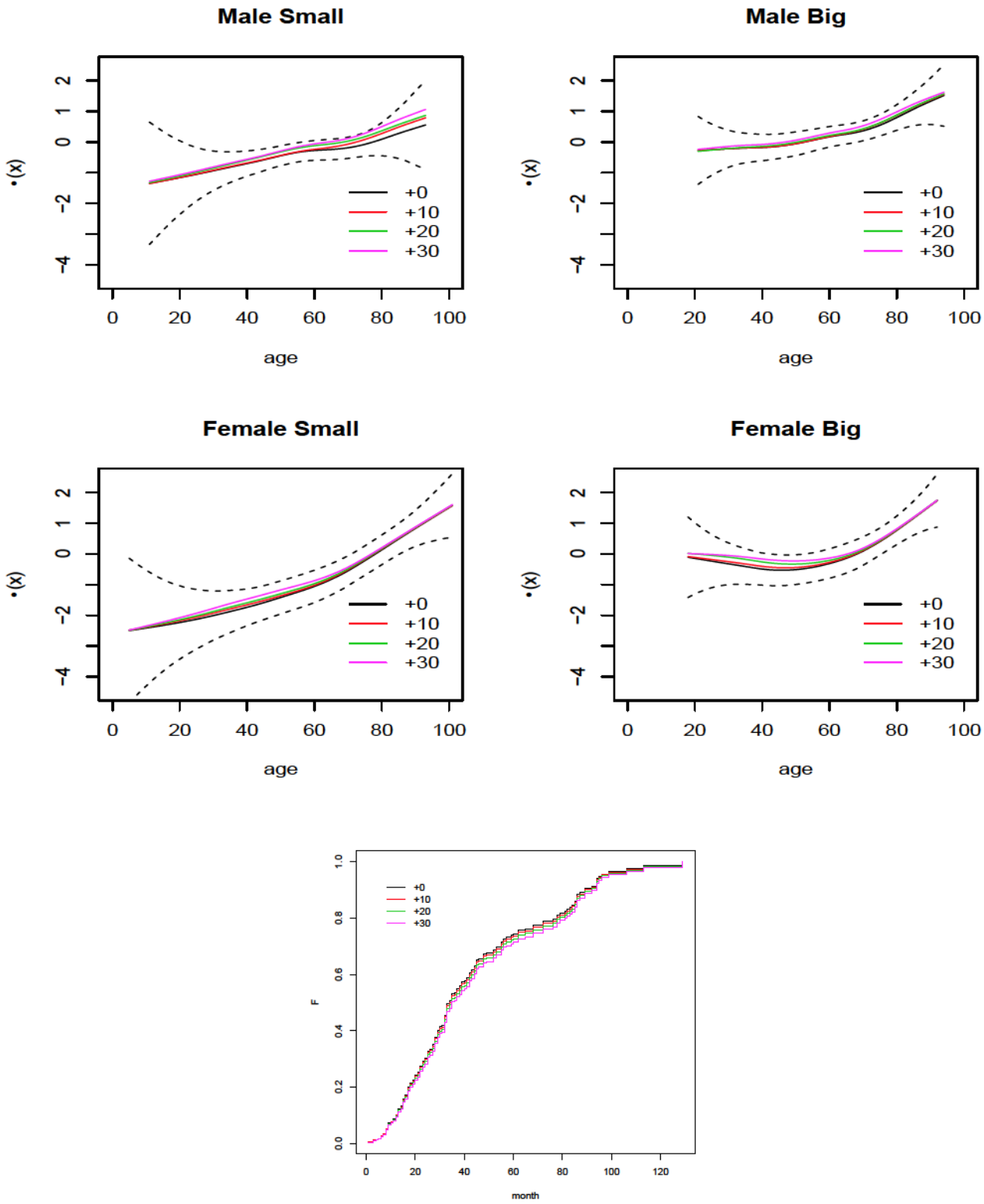

Figure 5. Top and middle frames: Plots of η estimates for the four patient groups in melanoma data; solid lines: estimates of η for cure thresholds equal to the largest failure time$+0$ (black), $+10$ months (red), $+20$ months (green), and $+30$ months (purple); dashed lines: 95% point-wise CIs of η for cure threshold equal to the largest failure time. Bottom frame: Plot of the estimates for function $F(t)$ for cure thresholds equal to the largest failure time$+0$ (black), $+10$ months (red), $+20$ months (green), and $+30$ months (purple).