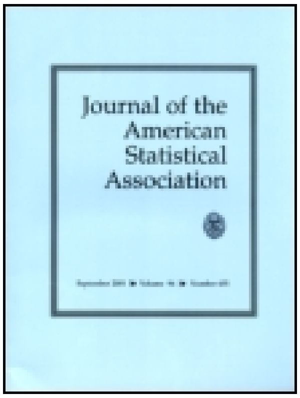

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Semiparametric Transformation Models for Survival Data With a Cure Fraction

Donglin Zeng[a], Guosheng Yin[a] & Joseph G Ibrahim[a]
[a] Donglin Zeng is Assistant Professor and Joseph G. Ibrahim is Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Guosheng Yin is Assistant Professor, Department of Biostatistics and AppliedMathematics, M. D. Anderson Cancer Center, University of Texas, Houston, TX 77030.
The authors thank the editor and the referees for helpful comments and suggestions.
Published online: 01 Jan 2012.

To cite this article: Donglin Zeng, Guosheng Yin & Joseph G Ibrahim (2006) Semiparametric Transformation Models for Survival Data With a Cure Fraction, Journal of the American Statistical Association, 101:474, 670-684, DOI: 10.1198/016214505000001122

To link to this article: http://dx.doi.org/10.1198/016214505000001122

Semiparametric Transformation Models for Survival Data With a Cure Fraction

Donglin ZENG, Guosheng YIN, and Joseph G. IBRAHIM

We propose a class of transformation models for survival data with a cure fraction. The class of transformation models is motivated by biological considerations and includes both the proportional hazards and the proportional odds cure models as two special cases. An efficient recursive algorithm is proposed to calculate the maximum likelihood estimators (MLEs). Furthermore, the MLEs for the regression coefficients are shown to be consistent and asymptotically normal, and their asymptotic variances attain the semiparametric efficiency bound. Simulation studies are conducted to examine the finite-sample properties of the proposed estimators. The method is illustrated on data from a clinical trial involving the treatment of melanoma.

KEY WORDS: Cure model; Linear transformation models; Proportional hazards model; Proportional odds model; Semiparametric efficiency.

1. INTRODUCTION

In time-to-event data arising from cancer and AIDS clinical trials, it is often observed that a proportion of subjects will never fail. For analyzing such data, cure rate models have been proposed and studied extensively. One type of commonly used cure rate model is the so-called *two-component mixture cure model* (Berkson and Gage 1952), which treats the whole population as a mixture of cured subjects and noncured subjects. This mixture model has been studied by many authors, including Gray and Tsiatis (1989), Sposto, Sather, and Baker (1992), Laska and Meisner (1992), Kuk and Chen (1992), Taylor (1995), Sy and Taylor (2000), and Lu and Ying (2004), among others. The book by Maller and Zhou (1996) provides a detailed discussion of frequentist methods of inference for the two-component mixture cure model.

Although the mixture cure model is intuitively attractive, it does have several drawbacks from both a Bayesian and frequentist perspective, as pointed out by Chen, Ibrahim, and Sinha (1999) and Ibrahim, Chen, and Sinha (2001). An alternative cure rate model with desirable properties, called the *promotion time cure model*, has been proposed and studied by Yakovlev and Tsodikov (1996), Tsodikov (1998), and Chen et al. (1999). In this model the cured subjects are assumed to have survival time equal to infinity, and the survival distribution for either cured subjects or noncured subjects can be integrated into one single formulation. For the ith individual with covariate $\mathbf{X}_i$ in the population, the survival function of subject i is given by

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

where $\theta(\cdot)$ is a known link function and $F(t)$ is a distribution function. Under the promotion time cure model (1), the cure rate is $S(\infty|\mathbf{X}_i) = \exp\{-\theta(\mathbf{X}_i)\}$ and the hazard rate at time t for subject i is equal to $\theta(\mathbf{X}_i)f(t)$, where $f(t) = dF(t)/dt$. Thus we see that model (1) has the proportional hazards structure when the covariates are modeled through $\theta(\cdot)$. Moreover, when $\theta(\mathbf{X}_i) = \exp(\boldsymbol{\beta}^T\mathbf{X}_i)$ and $\boldsymbol{\beta}$ contains an intercept term β_0, model (1) becomes the usual Cox (1972) proportional hazards model subject to the restriction of a *bounded* cumulative baseline hazard function, given by $\Lambda(t) = F(t)\exp(\beta_0)$. Thus any cure rate model has a bounded cumulative hazard, leading to an improper survival function [i.e., $S(\infty) > 0$], whereas noncure models, such as the Cox model (Cox 1972), have an unbounded cumulative hazard, thus leading to a proper survival function [i.e., $S(\infty) = 0$].

Yakovlev and Tsodikov (1996) and Chen et al. (1999) provided a biological derivation for model (1). The motivation comes from studying the time to relapse of cancer for patients with or without tumor cells. Specially, the promotion time cure model is derived as follows. For the ith subject, let N_i denote the number of tumor cells that have the potential of metastasizing, that is, the number of metastasis-competent tumor cells. The N_i's are unobservable latent variables. We assume that N_i has a Poisson distribution with Poisson rate (mean) $\theta(\mathbf{X}_i)$. We denote the promotion time for the kth tumor cell by $\tilde{T}_k$ $(k = 1, \dots, N_i)$, which is the time for the kth metastasis-competent tumor cell to produce a detectable tumor mass. The $\tilde{T}_k$'s are also unobservable quantities. Conditional on N_i, the $\tilde{T}_k$'s are independent and identically distributed (iid) as F, where F is sometimes referred to as the promotion time cumulative distribution function. Then the time to relapse of cancer, defined as $T = \min(\tilde{T}_1, \dots, \tilde{T}_{N_i})$, which is the observed event time, has the survival function

$$\begin{aligned} S(t|\mathbf{X}_i) &= P(N_i = 0) \\ &\qquad + \sum_{k\ge 1} P(\tilde{T}_1 > t, \dots, \tilde{T}_k > t | N_i = k)P(N_i = k) \\ &= \exp\{-\theta(\mathbf{X}_i)\} + \sum_{k=1}^{\infty}\{1 - F(t)\}^k \frac{\theta(\mathbf{X}_i)^k \exp\{-\theta(\mathbf{X}_i)\}}{k!} \\ &= \exp\{-\theta(\mathbf{X}_i)F(t)\}. \end{aligned}$$

In the derivation of (1), one critical assumption is that, conditional on the number of tumor cells, $N_i = k$, $(\tilde{T}_1, \dots, \tilde{T}_k)$ are mutually independent. This assumption may be unrealistic, because $(\tilde{T}_1, \dots, \tilde{T}_k)$ are unobserved random variables taken on the same subject. One possible relaxation and remedy of this assumption is to introduce a subject-specific frailty ξ_i such that conditional on both $N_i = k$ and ξ_i, $(\tilde{T}_1, \dots, \tilde{T}_k)$ are mutually

Donglin Zeng is Assistant Professor (E-mail: *dzeng@bios.unc.edu*) and Joseph G. Ibrahim is Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Guosheng Yin is Assistant Professor, Department of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center, University of Texas, Houston, TX 77030. The authors thank the editor and the referees for helpful comments and suggestions.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000001122

independent with distribution function $F(t)$. Moreover, we assume that conditional on $\mathbf{X}_i$ and ξ_i, N_i has a Poisson distribution with rate $\xi_i\theta(\mathbf{X}_i)$; thus ξ_i represents the heterogeneity of the Poisson rates in the N_i's. Following the same derivation as before, we then obtain that the survival function for the time to relapse, T, is

$$S(t|\mathbf{X}_i) = E_{\xi_i}\big[e^{-\theta(\mathbf{X}_i)F(t)\xi_i}\big],$$

where E_{ξ_i} denotes the expectation with respect to ξ_i. For example, when ξ_i has a gamma distribution with mean 1 [i.e., ξ_i has density $\{\gamma^{1/\gamma}\Gamma(1/\gamma)\}^{-1}\xi_i^{1/\gamma-1}\exp(-\xi_i/\gamma)$], after simple algebra, we obtain

$$S(t|\mathbf{X}_i) = \{1+\gamma\theta(\mathbf{X}_i)F(t)\}^{-1/\gamma}.$$

Equivalently, we can write

$$S(t|\mathbf{X}_i) = G_\gamma\{\theta(\mathbf{X}_i)F(t)\}, \tag{2}$$

where $G_\gamma(\cdot)$ is the transformation

$$G_\gamma(x) = \begin{cases} (1+\gamma x)^{-1/\gamma}, & \gamma > 0 \\ e^{-x}, & \gamma = 0. \end{cases} \tag{3}$$

Through (2) and (3), we obtain a very general class of transformation cure models and note that the proportional hazards cure rate model in (1) is a special case of this class corresponding to $\gamma = 0$. There are also other interesting special cases arising from (2) and (3). When $\gamma = 1$, we obtain a proportional odds type of cure model similar in flavor to the proportional odds models with proper survival functions considered by Pettitt (1982) and Bennett (1983). Moreover, the general form of the class in (2) not only has a strong biological motivation, but also can reduce to the usual linear transformation models studied by Cheng, Wei, and Ying (1995) under a special choice of $\theta(\cdot)$. For instance, if we choose $\theta(\mathbf{X}_i) = \exp(\beta_0 + \boldsymbol{\beta}_1^T\mathbf{Z}_i)$ with $\mathbf{X}_i = (1, \mathbf{Z}_i^T)^T$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$, and β_0 being the intercept term in the regression, then model (2) is equivalent to $S(t|\mathbf{Z}_i) = G_\gamma\{\exp(\boldsymbol{\beta}_1^T\mathbf{Z}_i)\Lambda(t)\}$, where $\Lambda(t) = F(t)\exp(\beta_0)$ is the cumulative baseline hazard. But when $\theta(\mathbf{X}_i)$ has a form other than $\theta(\mathbf{X}_i) = \exp(\boldsymbol{\beta}^T\mathbf{X}_i)$ [e.g., if $\theta(\mathbf{X}_i) = \exp(\boldsymbol{\beta}^T\mathbf{X}_i)/\{1+\exp(\boldsymbol{\beta}^T\mathbf{X}_i)\}$], then model (2) is quite different from the linear transformation model.

When γ, which specifies transformations in (3), is treated as an unknown parameter, the model parameters may not be identifiable. For example, suppose that $\theta(\mathbf{X}) = \exp(\beta_0)$. Then for any $\gamma \neq \tilde{\gamma}$, we can find a $\tilde{\beta}_0$, different from β_0, such that

$$\{1+\gamma e^{\beta_0}\}^{-1/\gamma} = \big\{1+\tilde{\gamma}e^{\tilde{\beta}_0}\big\}^{-1/\tilde{\gamma}}.$$

Thus for any distribution function $F(t)$, we define $\tilde{F}(t)$ so that

$$\{1+\gamma e^{\beta_0}F(t)\}^{-1/\gamma} = \big\{1+\tilde{\gamma}e^{\tilde{\beta}_0}\tilde{F}(t)\big\}^{-1/\tilde{\gamma}}.$$

Clearly, $\tilde{F}(t)$ is also a distribution function. Consequently, the two sets of parameters (γ, β_0, F) and $(\tilde{\gamma}, \tilde{\beta}_0, \tilde{F})$ give the same survival function, so they are not distinguishable from the observed data. More identifiability results are given in Section 4. In addition, in most practical applications, there is little information in the data from which to estimate γ with a reasonable degree of precision for small to even moderately large sample sizes. In these situations, the likelihood function of γ is flat. Our experience shows that γ can be well estimated when the sample size is very large, such as $n = 1{,}500$ or larger. Because of these limitations, we focus on the γ fixed case throughout the development of our model and asymptotic theory. However, in Section 4 we discuss estimation of γ when it is identifiable and also suggest a model selection strategy for choosing γ in the γ fixed case.

The transformation in (2) may not necessarily be from the family (3); different transformations are possible when ξ takes other distributions. For example, we may consider the following Box–Cox type transformations:

$$G_\gamma(x) = \begin{cases} \exp\Big\{-\dfrac{(1+x)^\gamma - 1}{\gamma}\Big\}, & \gamma > 0 \\ \dfrac{1}{1+x}, & \gamma = 0. \end{cases} \tag{4}$$

In this family, $\gamma = 1$ yields the proportional hazards model, whereas $\gamma = 0$ yields the proportional odds model. In this article, we study general classes of transformations $G(\cdot)$ and link functions $\theta(\cdot)$ and examine inference based on maximum likelihood estimation. However, for ease and clarity of exposition, we focus on the class in (3) or (4) and $\theta(\mathbf{X}_i) = \exp(\boldsymbol{\beta}^T\mathbf{X}_i)$ in the examples of Section 5. In addition, the promotion time cumulative distribution functions, $F(t)$, are completely unspecified and thus are estimated nonparametrically throughout.

The rest of the article is organized as follows. In Section 2 we introduce notation and propose an efficient computational algorithm for the maximum likelihood estimation procedure. In Section 3 we derive the asymptotic properties of the parameter estimates, including consistency and asymptotic normality. In Section 4 we discuss important issues of model selection, including estimation of γ when it is identifiable as well as the selection of γ when it is treated as fixed. In Section 5 we conduct simulation studies to evaluate the finite-sample properties of the estimators and also illustrate the proposed model with a real dataset. We give some concluding remarks in Section 6 and provide technical details for the proofs of the theorems in the Appendix.

2. MAXIMUM LIKELIHOOD ESTIMATION

Suppose that there are n iid right-censored observations, $\{Y_i = T_i \wedge C_i, \mathbf{X}_i, \Delta_i = I(T_i \le C_i); i = 1, \dots, n\}$, where $T_i \wedge C_i = \min(T_i, C_i)$ and $I(\cdot)$ is the indicator function. We assume that the follow-up time is infinite and that a proportion of subjects never experience failure or right-censoring, that is, $Y_i = \infty$ (so $C_i = \infty$) with probability 1 for some subjects. The right-censoring time C_i is assumed to be conditionally independent of T_i given $\mathbf{X}_i$ and to have a finite hazard rate almost everywhere. We assume that model (2) is used to link T_i with the covariate vector $\mathbf{X}_i$, where $\theta(\mathbf{X}_i) = \eta(\boldsymbol{\beta}^T\mathbf{X}_i)$, $\eta(\cdot)$ is a known and strictly positive link function and $\boldsymbol{\beta}$ includes an intercept term.

Thus the observed-data likelihood function of the parameters $(\boldsymbol{\beta}, F)$ is given by

$$\prod_{i=1}^n \Big\{\big[\big\{-G'\big(\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F(Y_i)\big)\eta(\boldsymbol{\beta}^T\mathbf{X}_i)f(Y_i)\big\}^{\Delta_i} \times \big\{G\big(\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F(Y_i)\big)\big\}^{(1-\Delta_i)}\big]^{I(Y_i<\infty)} \times \big[G(\eta(\boldsymbol{\beta}^T\mathbf{X}_i))\big]^{I(Y_i=\infty)}\Big\}, \tag{5}$$

where $G'(x)$ denotes the derivative of G with respect to x and $f(\cdot)$ is the density function corresponding to the distribution function $F(\cdot)$ with respect to Lebesgue measure. We wish to maximize the foregoing likelihood function to obtain the maximum likelihood estimators (MLEs) $\boldsymbol{\beta}$ and F; however, this maximum does not exist, because one can choose $f(Y_i) = \infty$ for some Y_i with $\Delta_i = 1$. Thus we apply a nonparametric maximum likelihood estimation approach, where F is allowed to be a right-continuous function. Instead of maximizing (5), we maximize the following modified function:

$$\prod_{i=1}^{n}\Big\{\big[\{-G'\big(\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F(Y_i)\big)\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F\{Y_i\}\}^{\Delta_i} \times \big\{G\big(\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F(Y_i)\big)\big\}^{(1-\Delta_i)}\big]^{I(Y_i<\infty)} \times \big[G(\eta(\boldsymbol{\beta}^T\mathbf{X}_i))\big]^{I(Y_i=\infty)}\Big\}, \tag{6}$$

where $F\{Y_i\}$ is the jump size of F at Y_i. The MLE for F is termed the nonparametric maximum likelihood estimator (NPMLE) for F, and it is easy to show that the estimate for F must be a distribution function only with point masses at the observed Y_i with $\Delta_i = 1$. To estimate $F(t)$ nonparametrically, we must determine a follow-up time such that all censored observations beyond that follow-up time, called the *cure threshold*, are treated as $Y_i = \infty$ (i.e., observed to be cured) and all observations lower than this threshold are treated as $Y_i < \infty$ (i.e., observed to be either a failure or right-censored). This assumption is needed so that the model is identifiable in $(\boldsymbol{\beta}, F)$, as shown in Section 3. Note that if a parametric form is assumed for F (as in Ibrahim et al. 2001), then the condition that some of the Y_i's are observed to be infinity is not needed.

To compute the MLEs, we first derive the F that maximizes (6) for fixed $\boldsymbol{\beta}$. Equivalently, we maximize the logarithm of (6), which is equal to

$$\sum_{i=1}^{n} I(Y_i<\infty)\big[\Delta_i \log p_i + \Delta_i \log\{-G'\big(\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F_i\big)\} + (1-\Delta_i)\log G\big(\eta(\boldsymbol{\beta}^T\mathbf{X}_i)F_i\big)\big],$$

subject to the constraint $\sum_j \Delta_j I(Y_j<\infty)p_j = 1$, where $p_i = F\{Y_i\}$ denotes the jump size of F at Y_i and $F_i = \sum_{Y_j\le Y_i, \Delta_j=1} p_j$. If we order the observed failure times from smallest to largest and use the indices $(1), \dots, (m)$ for the ordered times, $Y_{(1)} < \cdots < Y_{(m)}$, where $m = \sum_i \Delta_i I(Y_i < \infty)$, then, after introducing the Lagrange multiplier λ, we obtain $p_{(i)}$ by solving the equation

$$\frac{1}{p_{(i)}} + \sum_{j=1}^{n}\Big\{\Delta_j \frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)\eta(\boldsymbol{\beta}^T\mathbf{X}_j)I(Y_{(i)}\le Y_j<\infty)}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)} + (1-\Delta_j) \times \frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)\eta(\boldsymbol{\beta}^T\mathbf{X}_j)I(Y_{(i)}\le Y_j<\infty)}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)}\Big\} - \lambda = 0,$$

where $G''(x)$ denotes the second derivative of G with respect to x. Thus it follows that

$$\frac{1}{p_{(i+1)}} = \frac{1}{p_{(i)}} + \sum_{Y_{(i)}\le Y_j<Y_{(i+1)}}\Big\{\Delta_j \frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)\eta(\boldsymbol{\beta}^T\mathbf{X}_j)}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)} + (1-\Delta_j)\frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)\eta(\boldsymbol{\beta}^T\mathbf{X}_j)}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_j)}\Big\}.$$

Equivalently,

$$\frac{1}{p_{(i+1)}} = \frac{1}{p_{(i)}} + \frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})F_{(i)})\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})F_{(i)})} + \sum_{Y_{(i)}<Y_j<Y_{(i+1)}} \frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_{(i)})\eta(\boldsymbol{\beta}^T\mathbf{X}_j)}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_{(i)})}, \tag{7}$$

where $F_{(i)} = p_{(1)} + \cdots + p_{(i)}$. Using the fact that $\sum_{i=1}^{m} p_{(i)} = 1$, we can also write (7) as

$$\frac{1}{p_{(i)}} = \frac{1}{p_{(i+1)}} - \frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})(1-S_{(i+1)}))\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})(1-S_{(i+1)}))} - \sum_{Y_{(i)}<Y_j<Y_{(i+1)}} \frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)(1-S_{(i+1)}))\eta(\boldsymbol{\beta}^T\mathbf{X}_j)}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)(1-S_{(i+1)}))}, \tag{8}$$

where $S_{(i+1)} = p_{(i+1)} + p_{(i+2)} + \cdots + p_{(m)}$. From (7), we obtain a recursive formula of calculating $p_{(i+1)}$ from $p_{(i)}$ and $F_{(i)}$; whereas from (8), we obtain another recursive formula of calculating $p_{(i)}$ from $p_{(i+1)}$ and $S_{(i+1)}$. When $G'' > 0$ and $G' < 0$, we prefer to use (8), because it ensures that $0 < p_{(i)} < p_{(i+1)}$ once $p_{(i+1)} > 0$ and $S_{(i+1)} < 1$.

Hence, from (8), we can treat $\boldsymbol{\beta}$, $\alpha \equiv p_{(m)} > 0$, and λ as independent parameters and $p_{(1)}, \dots, p_{(m-1)}$ as functions of $\boldsymbol{\beta}$ and α. Then the constrained maximum likelihood equations for $\boldsymbol{\beta}$ and $p_{(1)}, \dots, p_{(m)}$ can be reduced to solving the following score equations for $\boldsymbol{\beta}$, α, and λ:

$$0 = \sum_{i=1}^{m}\frac{1}{p_{(i)}}\frac{\partial}{\partial\boldsymbol{\beta}}p_{(i)} + \sum_{i=1}^{m}\frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})F_{(i)})}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})F_{(i)})} \times \Big\{\eta'\big(\boldsymbol{\beta}^T\mathbf{X}_{(i)}\big)\mathbf{X}_{(i)}F_{(i)} + \eta\big(\boldsymbol{\beta}^T\mathbf{X}_{(i)}\big)\frac{\partial}{\partial\boldsymbol{\beta}}F_{(i)}\Big\} + \sum_{i=1}^{m}\sum_{Y_{(i)}<Y_j<Y_{(i+1)}}\frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_{(i)})}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_{(i)})} \times \Big\{\eta'(\boldsymbol{\beta}^T\mathbf{X}_j)\mathbf{X}_j F_{(i)} + \eta(\boldsymbol{\beta}^T\mathbf{X}_j)\frac{\partial}{\partial\boldsymbol{\beta}}F_{(i)}\Big\} + \sum_{j=1}^{n}\Delta_j\mathbf{X}_j \tag{9}$$

$$
\begin{aligned}
&+\sum_{j=1}^{n} I(Y_j=\infty)\frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j))}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j))}\eta'(\boldsymbol{\beta}^T\mathbf{X}_j)\mathbf{X}_j \\
&-\lambda\sum_{i=1}^{m}\frac{\partial}{\partial\boldsymbol{\beta}}p_{(i)}, \\
0=&\sum_{i=1}^{m}\frac{1}{p_{(i)}}\frac{\partial}{\partial\alpha}p_{(i)} \\
&+\sum_{i=1}^{m}\frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})F_{(i)})}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_{(i)})F_{(i)})}\eta\big(\boldsymbol{\beta}^T\mathbf{X}_{(i)}\big)\frac{\partial}{\partial\alpha}F_{(i)} \\
&+\sum_{i=1}^{m}\sum_{Y_{(i)}<Y_j<Y_{(i+1)}}\frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_{(i)})}{G(\eta(\boldsymbol{\beta}^T\mathbf{X}_j)F_{(i)})}\eta(\boldsymbol{\beta}^T\mathbf{X}_j)\frac{\partial}{\partial\alpha}F_{(i)} \\
&-\lambda\sum_{i=1}^{m}\frac{\partial}{\partial\alpha}p_{(i)}, \\
0=&\sum_{i=1}^{m}p_{(i)}-1.
\end{aligned}
$$

After eliminating λ from the first two equations, the Newton–Raphson algorithm can be used to solve the system of equations in (9). The first and second derivatives of $p_{(i)}$ with respect to $\boldsymbol{\beta}$ and α can be computed using the recursive formula (8).

We denote the MLEs for $\boldsymbol{\beta}$ and α by $\hat{\boldsymbol{\beta}}_n$ and $\hat{\alpha}_n$. We can estimate the asymptotic variance of $(\hat{\boldsymbol{\beta}}_n, \hat{\alpha}_n)$ based on the profile log-likelihood function for $(\boldsymbol{\beta}, \alpha)$, which is defined as the maximum value of the logarithm of (6) for any fixed $(\boldsymbol{\beta}, \alpha)$ and is denoted by $pl_n(\boldsymbol{\beta}, \alpha)$. The asymptotic variance of $(\hat{\boldsymbol{\beta}}_n, \hat{\alpha}_n)$ can be estimated using the negative inverse of the curvature of $pl_n(\boldsymbol{\beta}, \alpha)$ at $(\hat{\boldsymbol{\beta}}_n, \hat{\alpha}_n)$, that is,

$$
-\begin{pmatrix} \frac{\partial^2}{\partial\boldsymbol{\beta}^2}pl_n(\boldsymbol{\beta},\alpha) & \frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\alpha}pl_n(\boldsymbol{\beta},\alpha) \\ \frac{\partial^2}{\partial\alpha\,\partial\boldsymbol{\beta}}pl_n(\boldsymbol{\beta},\alpha) & \frac{\partial^2}{\partial\alpha^2}pl_n(\boldsymbol{\beta},\alpha) \end{pmatrix}^{-1}\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_n,\alpha=\hat{\alpha}_n}.
$$

Specifically, the second derivative of $pl_n(\boldsymbol{\beta}, \alpha)$ with respect to $\boldsymbol{\beta}$ and α can be calculated based on the following chain rule and the recursive formula (8):

$$
\frac{\partial}{\partial\boldsymbol{\beta}}pl_n(\boldsymbol{\beta},\alpha)=\frac{\partial}{\partial\boldsymbol{\beta}}l_n(\boldsymbol{\beta},F)+\sum_{i=1}^{m-1}\frac{\partial l_n(\boldsymbol{\beta},F)}{\partial p_{(i)}}\frac{\partial p_{(i)}}{\partial\boldsymbol{\beta}}
$$

and

$$
\frac{\partial}{\partial\alpha}pl_n(\boldsymbol{\beta},\alpha)=\frac{\partial}{\partial\alpha}l_n(\boldsymbol{\beta},F)+\sum_{i=1}^{m-1}\frac{\partial l_n(\boldsymbol{\beta},F)}{\partial p_{(i)}}\frac{\partial p_{(i)}}{\partial\alpha},
$$

where $l_n(\boldsymbol{\beta}, F)$ is the logarithm value of (6). The justification of the foregoing variance estimation method is based on the profile likelihood theory of Murphy and van der Vaart (2000), and is discussed in the Appendix.

3. ASYMPTOTIC PROPERTIES

In this section we establish theorems characterizing the asymptotic properties of $(\hat{\boldsymbol{\beta}}_n, \hat{\alpha}_n)$. To achieve consistency and asymptotic normality, we first need the following assumptions:

(C1) The covariate $\mathbf{X}$ is bounded with probability 1, and if there exists a vector $\tilde{\boldsymbol{\beta}}$ such that $\tilde{\boldsymbol{\beta}}^T\mathbf{X}=0$ with probability 1, then $\tilde{\boldsymbol{\beta}}=\mathbf{0}$.

(C2) Conditional on $\mathbf{X}$, the right-censoring time C is independent of T, and $P(C=\infty|\mathbf{X})>0$.

(C3) The true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_0$, belongs to the interior of a known compact set $\mathcal{B}_0$, and the true promotion time cumulative distribution function F_0 is differentiable with $F_0'(x)>0$ for all $x\in\mathbb{R}^+$.

(C4) The link function $\eta(\cdot)$ is strictly increasing and twice-continuously differentiable with $\eta(\cdot)>0$. Furthermore, the transformation G satisfies

$$
G(0)=1, \qquad G(x)>0, \qquad G'(x)<0,
$$
$$
G^{(3)}(x) \text{ exists and is continuous,}
$$

where $G^{(3)}(x)$ is the third derivative of $G(x)$.

Condition (C1) is the usual condition for a design matrix in regression settings. The condition $P(C=\infty|\mathbf{X})$ in (C2) ensures that at least some cured subjects are not right-censored; otherwise, if all subjects either fail or are right-censored, then, intuitively, one would be unable to identify the cure rate. In (C3), $\boldsymbol{\beta}$ is assumed to be bounded. Such an assumption is often imposed in semiparametric inference, because practical calculation is always performed within a reasonable bounded set. Many link functions $\eta(\cdot)$ and $G(\cdot)$ satisfy the conditions in (C4). Examples of $\eta(\cdot)$ include $\eta(x)=e^x$, $\eta(x)=e^x/(1+e^x)$, and $\eta(x)=\Phi(x)$, where Φ is the cumulative distribution function of the standard normal distribution. Examples of transformations satisfying (C4) include the transformations $(1+\gamma x)^{-1/\gamma}$ for $\gamma>0$ and $\exp(-x)$ for $\gamma=0$, as well as some others, such as $G(x)=\{1+\log(1+x)\}^{-\gamma}$ for $\gamma>0$ and $G(x)=\exp\{-((1+x)^\gamma-1)/\gamma\}$ for $\gamma>0$.

Before stating the main results, we first show that under conditions (C1)–(C4), the parameters $\boldsymbol{\beta}$ and F are identifiable. Suppose that two sets of parameters, $(\boldsymbol{\beta}, F)$ and $(\tilde{\boldsymbol{\beta}}, \tilde{F})$, give the same likelihood function for the observed data. We claim that $\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}$ and $F=\tilde{F}$. Because

$$
\begin{aligned}
&\Big[\big\{-G'\big(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y)\big)\eta(\boldsymbol{\beta}^T\mathbf{X})f(Y)\big\}^{\Delta} \\
&\quad\times\big\{G\big(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y)\big)\big\}^{(1-\Delta)}\Big]^{I(Y<\infty)} \\
&\quad\times[G(\eta(\boldsymbol{\beta}^T\mathbf{X}))]^{I(Y=\infty)} \\
&=\Big[\big\{-G'\big(\eta(\tilde{\boldsymbol{\beta}}^T\mathbf{X})\tilde{F}(Y)\big)\eta(\tilde{\boldsymbol{\beta}}^T\mathbf{X})\tilde{f}(Y)\big\}^{\Delta} \\
&\quad\times\big\{G\big(\eta(\tilde{\boldsymbol{\beta}}^T\mathbf{X})\tilde{F}(Y)\big)\big\}^{(1-\Delta)}\Big]^{I(Y<\infty)} \\
&\quad\times\big[G(\eta(\tilde{\boldsymbol{\beta}}^T\mathbf{X}))\big]^{I(Y=\infty)},
\end{aligned}
\tag{10}
$$

we choose $Y=\infty$. Then, from the monotonicity of both G and η, it follows that $\boldsymbol{\beta}^T\mathbf{X}=\tilde{\boldsymbol{\beta}}^T\mathbf{X}$. Thus condition (C1) gives $\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}$. Furthermore, by letting $\Delta=1$ and $Y=y$ and integrating both sides of (10) from 0 to y, we have $G(\eta(\boldsymbol{\beta}^T\mathbf{X})F(y))=G(\eta(\tilde{\boldsymbol{\beta}}^T\mathbf{X})\tilde{F}(y))$; therefore, $F(y)=\tilde{F}(y)$.

The following theorem establishes the consistency of the MLE.

Theorem 1. Under conditions (C1)–(C4), with probability 1,

$$|\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0| \to 0 \qquad \text{and} \qquad \sup_{t\in\mathbb{R}^+} |\hat{F}_n(t) - F_0(t)| \to 0;$$

that is, both $\hat{\boldsymbol\beta}_n$ and $\hat{F}_n$ are strongly consistent.

The basic idea in proving Theorem 1 is as follows. Suppose that $\hat{\boldsymbol\beta}_n$ and $\hat{F}_n$ converge to $\boldsymbol\beta^*$ and F^*. We first construct an empirical distribution function $\tilde{F}_n$ converging to F_0. Then, because $\{l_n(\hat{\boldsymbol\beta}_n, \hat{F}_n) - l_n(\boldsymbol\beta_0, \tilde{F}_n)\}/n \ge 0$, where $l_n(\boldsymbol\beta, F)$ denotes the observed log-likelihood function at $(\boldsymbol\beta, F)$, and this difference converges to the negative Kullback–Leibler divergence between $(\boldsymbol\beta^*, F^*)$ and $(\boldsymbol\beta_0, F_0)$, the identifiability result gives $\boldsymbol\beta^* = \boldsymbol\beta_0$ and $F^* = F_0$. This establishes the consistency result in Theorem 1. Constructing the empirical function $\tilde{F}_n$ and using the Kullback–Leibler divergence to prove consistency has been used by many others in semiparametric theory, including Murphy (1994), Murphy, Rossini, and van der Vaart (1997), Parner (1998), Slud and Vonta (2004), and Kosorok, Lee, and Fine (2004), among others. However, observing the fact that $\hat{F}_n$ is a distribution function, proving the convergence of the Kullback–Leibler divergence is not trivial in our case, as we show in the Appendix.

Our second result concerns the joint asymptotic distribution of $\hat{\boldsymbol\beta}_n$ and $\hat{F}_n$. To obtain the joint asymptotic distribution for $(\hat{\boldsymbol\beta}_n, \hat{F}_n)$, we first introduce the set

$$\begin{aligned}\mathcal{H} = \big\{(\mathbf{h}_1, h_2) : {} & \mathbf{h}_1 \in \mathbb{R}^d, \|\mathbf{h}_1\| < 1,\\ & h_2 \text{ is a function in } [0,\infty) \text{ with}\\ & \text{its total variation bounded by } 1\big\}.\end{aligned}$$

Here the total variation of a function h_2 is defined as the supremum of $\sum_{i=1}^m |h_2(t_{i+1}) - h_2(t_i)|$ over all finite partitions $0 = t_1 < t_2 < \cdots < t_{m+1} = \infty$. We let $\|h_2\|_V$ denote the total variation of h_2. Then $\sqrt{n}(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0, \hat{F}_n - F_0)$ can be treated as a linear functional in $l^\infty(\mathcal{H})$, the space of all bounded linear functionals on $\mathcal{H}$, defined as

$$\begin{aligned}&\sqrt{n}(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0, \hat{F}_n - F_0)[\mathbf{h}_1, h_2]\\ &\qquad = \sqrt{n}(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0)^T\mathbf{h}_1 + \sqrt{n}\int h_2(t)\, d(\hat{F}_n - F_0).\end{aligned}$$

The next theorem establishes the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0, \hat{F}_n - F_0)$ in the metric space $l^\infty(\mathcal{H})$.

Theorem 2. Under conditions (C1)–(C4), $\sqrt{n}(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0, \hat{F}_n - F_0)$ converges weakly to a mean-0 Gaussian process in $l^\infty(\mathcal{H})$. Furthermore, $\hat{\boldsymbol\beta}_n$ is efficient; equivalently, its asymptotic variance attains the semiparametric efficiency bound for $\boldsymbol\beta_0$.

The covariance matrix of the asymptotic Gaussian process is given in the Appendix. A definition of the semiparametric efficiency bound has been provided by Bickel, Klaassen, Ritov, and Wellner (1993, chap. 3). Thus Theorem 2 establishes that the MLEs are asymptotically normal and efficient. The proof of Theorem 2 is standard in most of the current semiparametric literature (including Murphy 1995; Parner 1998; and Kosorok et al. 2004). The proof relies on the linearization of the likelihood equations for $\hat{\boldsymbol\beta}_n$ and $\hat{F}_n$ and uses theorem 3.3.1 of van der Vaart and Wellner (1996). In the proof, verifying some Donsker classes and proving the invertibility of the information operator are the key steps. Both of these issues are discussed in detail in the Appendix for the proposed model.

Theorem 2 has many useful applications. By letting $h_2(\cdot) = I(\cdot \le t)$ for any $t \ge 0$, we obtain that $\sqrt{n}(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0, \hat{F}(t) - F_0(t))$ converges weakly to a mean-0 Gaussian process in $l^\infty(\mathbb{R}^d \times [0,\infty))$. As a result, for fixed t_0, $\sqrt{n}(\hat{F}_n(t_0) - F_0(t_0))$ has an asymptotic normal distribution with mean 0. If its asymptotic variance can be estimated, then one can easily construct a confidence interval for $F_0(t_0)$. Special choices of t_0 can be the quantiles of F_0. Furthermore, when interest is in testing whether the true promotion distribution function is equal to a given distribution function F_0, we can construct a test statistic $\sqrt{n}\sup_{t\ge0}|\hat{F}_n(t) - F_0(t)|$, similar to the Kolmogorov–Smirnov statistic. Then Theorem 2 implies that such a statistic has an asymptotic distribution that is the same as the supremum of a Gaussian process. We remark that in the foregoing cases, the asymptotic covariance function of the Gaussian process in Theorem 2 must be estimated. One practical way to estimate this function is through a bootstrapping approach. The justification of the bootstrapping procedure can be shown using the same techniques used by Kosorok et al. (2004). We do not pursue this issue further here, but focus only on inference for regression coefficients in the subsequent development.

4. ESTIMATION OF THE TRANSFORMATION $G(\cdot)$

In the foregoing sections, the transformation $G(\cdot)$ was assumed known. One important practical issue is how to estimate $G(\cdot)$ using the observed data. We discuss two possible methods to estimate this transformation.

The first approach is to consider $G(\cdot)$ from a parametric transformation family $\{G_\gamma : \gamma \in \Gamma\}$, where Γ is a compact set in Euclidean space. For example, G_γ arises from the family given in (3) or (4). Using the observed data, we then estimate γ along with $\boldsymbol\beta$ and F. However, as noted in Section 1, one serious problem with this approach is the possible nonidentifiability of γ. However, for some special families of transformations, the parameters $(\gamma, \boldsymbol\beta, F)$ are identifiable, as stated in the following proposition.

Proposition 1. Let $\mathbf{X} = (1, \mathbf{W}^T)^T$ and $\boldsymbol\beta_0$ as $(\beta_{01}, \boldsymbol\beta_{0w}^T)^T$. Assume that $\mathbf{W}$ has support containing a nonempty open interior and that $\boldsymbol\beta_{0w}^T\mathbf{W} \ne 0$. Then, for transformations from the family (3) and $\eta(x) = \exp(x)$, $\boldsymbol\beta_0, F_0$, and γ_0 are identifiable.

Proof. Suppose that $(\tilde{\boldsymbol\beta}, \tilde{F}, \tilde\gamma)$ gives the same observed likelihood function as $(\boldsymbol\beta_0, F_0, \gamma_0)$, that is,

$$\begin{aligned}&\Big[\big\{-G'_{\gamma_0}\big(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(Y)\big)\eta(\boldsymbol\beta_0^T\mathbf{X})f_0(Y)\big\}^{\Delta}\\ &\qquad \times \big\{G_{\gamma_0}\big(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(Y)\big)\big\}^{(1-\Delta)}\Big]^{I(Y<\infty)}\\ &\qquad \times \big[G_{\gamma_0}(\eta(\boldsymbol\beta_0^T\mathbf{X}))\big]^{I(Y=\infty)}\\ &= \Big[\big\{-G'_{\tilde\gamma}\big(\eta(\tilde{\boldsymbol\beta}^T\mathbf{X})\tilde{F}(Y)\big)\eta(\tilde{\boldsymbol\beta}^T\mathbf{X})\tilde{f}(Y)\big\}^{\Delta}\\ &\qquad \times \big\{G_{\tilde\gamma}\big(\eta(\tilde{\boldsymbol\beta}^T\mathbf{X})\tilde{F}(Y)\big)\big\}^{(1-\Delta)}\Big]^{I(Y<\infty)}\\ &\qquad \times \big[G_{\tilde\gamma}(\eta(\tilde{\boldsymbol\beta}^T\mathbf{X}))\big]^{I(Y=\infty)}, \qquad (11)\end{aligned}$$

where $G_\gamma(x) = (1+\gamma x)^{-1/\gamma}$. We choose $Y = \infty$ in (11) and obtain

$$\{1+\tilde{\gamma}\exp(\tilde{\boldsymbol{\beta}}^T\mathbf{X})\}^{1/\tilde{\gamma}} = \{1+\gamma_0\exp(\boldsymbol{\beta}_0^T\mathbf{X})\}^{1/\gamma_0}.$$

Because both sides are analytic in $\mathbf{W}$, this equality holds for any $\mathbf{W}$ in real space. If $\gamma_0 < \tilde{\gamma}$, then, from the monotonicity of $(1+\gamma x)^{1/\gamma}$, we have $\boldsymbol{\beta}_0^T\mathbf{X} > \tilde{\boldsymbol{\beta}}^T\mathbf{X}$ for any $\mathbf{X}$. Immediately, we conclude that $\tilde{\boldsymbol{\beta}}_{0w} = \boldsymbol{\beta}_{0w}$ and $\beta_{01} < \tilde{\beta}_{01}$. As a result, we have

$$\{1+\tilde{\gamma}\exp(\tilde{\beta}_{01}+\boldsymbol{\beta}_{0w}^T\mathbf{W})\}^{1/\tilde{\gamma}} = \{1+\gamma_0\exp(\beta_{01}+\boldsymbol{\beta}_{0w}^T\mathbf{W})\}^{1/\gamma_0},$$

and this holds for any real $\mathbf{W}$. Letting $\boldsymbol{\beta}_{0w}^T\mathbf{W} \to \infty$, we then obtain $\gamma_0 = \tilde{\gamma}$ and $\beta_{01} = \tilde{\beta}_{01}$. Furthermore, choosing $\Delta = 1$ and $Y = y$ and integrating from 0 to y in (11), we obtain $\tilde{F}(y) = F_0(y)$.

Proposition 1 states that if a continuous covariate has a nonzero effect, then γ can be identified. When model parameters are identifiable, with some additional regularity conditions beyond (C.1)–(C.4), the NPMLEs for $\boldsymbol{\beta}$, F, and γ are strongly consistent and asymptotically normal; the details are given in the remarks of the Appendix. This approach uses the observed data to estimate the transformation parameter, and our proposed algorithm can be easily adapted to incorporate this extra parameter estimation. However, this approach may not be useful for practical applications, for the following reasons. First, with no prior knowledge about the true covariate effects, there is always a concern about identifying all of the parameters in the model, because nonidentifiability can cause numerical instability in the computations. Second, even if the parameters are identifiable, our experience indicates that for small samples, the likelihood function is typically quite flat as a function of γ. Thus, obtaining an accurate estimate of γ requires a very large sample size, which may not be practical in many biomedical studies. Third, when the choices of transformations are from multiple families of transformations that are parameterized differently, this approach is no longer feasible.

Hence we suggest the following approach for estimating the transformation G in practice. When many transformations are under consideration, we can calculate the NPMLEs under each transformation, then choose the transformation that maximizes the Akaike information criterion (AIC). The AIC is defined as the twice log-likelihood function minus twice the number of parameters. In some applications, to obtain algebraically simple transformations, we may also penalize the complexity of the transformation. Some possible choices of a penalty can be the maximal difference between $G(x)$ and $\exp(-x)$, so that we can choose a model close to the proportional hazards model; or the choice can be the maximal difference between $G(x)$ and $1/(1+x)$, so that we can choose a model close to the proportional odds model. However, the determination of the transformation complexity remains an unsolved issue, so we defer further discussion to future work. Besides the AIC criterion, other criteria can also be used, including the Bayesian information criterion (BIC) (Schwarz 1978), the L measure (Ibrahim and Laud 1994), and likelihood-based cross-validation. As an additional note, in most practice the inference is based solely on the selected model and thus the variance estimate does not reflect the variation due to the model selection procedure. The correction of the variance estimate, sometimes called post–model selection inference, remains an open problem in semiparametric inference.

In the subsequent simulation study, we examine the performance of the NPMLEs for a fixed transformation, whereas, in the data application, we use the AIC to select the best transformation to fit the data.

5. NUMERICAL STUDIES

5.1 Simulation

We conducted simulation studies to examine the small-sample performance of our proposed methodology. In the first simulation study, the transformation cure model had a survival function of the form

$$S(t|X_1, X_2) = \left\{1+\gamma\exp(\beta_0+\beta_1X_1+\beta_2X_2)F(t)\right\}^{-1/\gamma},$$

with X_1 a uniformly distributed random variable in [0, 1], X_2 a Bernoulli random variable, $\beta_0 = .5, \beta_1 = 1, \beta_2 = -.5$, and $F(t) = 1-\exp(-t)$. We chose γ to vary from 0 to 1. Moreover, each subject had a 40% chance of being right-censored, and the censoring time was generated from an exponential distribution with mean 1. The censoring proportions varied from 17% to 22% as γ changed from 0 to 1, whereas the cure rate was as low as 8% when $\gamma = 0$ and rose to 20% when $\gamma = 1$. For each simulated dataset, the proposed method of Section 2 was implemented to calculate the MLEs of $\boldsymbol{\beta}$ and its corresponding variance estimate. In solving the score equations using the Newton–Raphson iterations, the initial values for $\boldsymbol{\beta}$ were set to 0 and the initial value for α was set to $1/n$, with n the sample size. Other initial values were also tested in the simulation study, and results were very robust to those choices. The convergence of each simulation was fast and often obtained within 10 iterations.

Table 1 summarizes the results from 1,000 replications for each combination of γ and n. The column labeled "Estimate" denotes the average values of the estimates, "SE" is the sample standard error of the estimates, "ESE" is the average of the estimated standard errors, and "CP" is the coverage proportion of 95% confidence intervals constructed based on the asymptotic normal approximation. The results in Table 1 indicate that the proposed estimation method performs well with sample sizes of 100 and 200; the biases are small, the estimated standard errors agree well with the sample standard errors, and the coverage probabilities are accurate.

In the second simulation study, we generated the failure time from the transformation cure model with survival function

$$S(t|X_1, X_2) = \exp\left[-\left\{\left(1+\gamma\exp(\beta_0+\beta_1X_1+\beta_2X_2)F(t)\right)^{\gamma}-1\right\}/\gamma\right],$$

where $F(t) = 1-\exp(-t)$ and the covariates and censoring time were generated using the same distributions as in the first simulation. In this setting we also varied γ from 0 to 1, where $\gamma = 0$ corresponds to the proportional odds cure model and $\gamma = 1$ corresponds to the proportional hazards cure model. The censoring proportion and the cure rate were 22% and 20% when $\gamma = 0$ and became 17% and 8% when $\gamma = 1$. The results, based on 1,000 repetitions for sample sizes 100 and 200, are summarized in Table 2. From Table 2, we obtain the same conclusions as in

Table 1. Simulation Results From 1,000 Replications Under the Transformation $G(x) = (1+\gamma x)^{-1/\gamma}$

Model	n	Parameter	True value	Estimate	SE	ESE	CP (%)
$\gamma = 0$	100	β_0	.5	.490	.289	.312	97.7
		β_1	1	1.033	.433	.427	94.9
		β_2	$-.5$	$-.519$	.242	.242	95.7
	200	β_0	.5	.502	.200	.218	96.1
		β_1	1	1.019	.300	.296	94.6
		β_2	$-.5$	$-.509$	.167	.168	95.9
$\gamma = .25$	100	β_0	.5	.476	.341	.350	95.6
		β_1	1	1.036	.512	.493	94.0
		β_2	$-.5$	$-.490$	.280	.281	96.0
	200	β_0	.5	.499	.236	.245	95.6
		β_1	1	1.006	.356	.344	95.1
		β_2	$-.5$	$-.507$	.194	.197	95.5
$\gamma = .5$	100	β_0	.5	.477	.380	.388	96.3
		β_1	1	1.022	.550	.554	95.4
		β_2	$-.5$	$-.518$	.320	.318	95.1
	200	β_0	.5	.488	.271	.273	95.5
		β_1	1	1.015	.400	.388	94.9
		β_2	$-.5$	$-.505$	.225	.222	95.1
$\gamma = .75$	100	β_0	.5	.487	.410	.423	95.7
		β_1	1	.995	.601	.607	95.1
		β_2	$-.5$	$-.491$	.359	.348	94.2
	200	β_0	.5	.486	.284	.298	96.5
		β_1	1	1.022	.426	.425	94.7
		β_2	$-.5$	$-.494$	.241	.244	95.4
$\gamma = 1$	100	β_0	.5	.455	.426	.458	96.7
		β_1	1	1.043	.637	.658	96.1
		β_2	$-.5$	$-.498$	.375	.378	95.4
	200	β_0	.5	.482	.310	.321	95.4
		β_1	1	1.015	.458	.460	94.8
		β_2	$-.5$	$-.502$	.258	.264	95.8

Table 2. Simulation Results From 1,000 Replications Under the Transformation $G(x) = \exp[-\{(1+x)^{\gamma} - 1\}/\gamma]$

Model	n	Parameter	True value	Estimate	SE	ESE	CP (%)
$\gamma = 0$	100	β_0	.5	.465	.442	.458	96.6
		β_1	1	1.026	.632	.658	96.4
		β_2	$-.5$	$-.510$	.387	.378	94.8
	200	β_0	.5	.498	.318	.321	95.4
		β_1	1	.995	.474	.461	93.9
		β_2	$-.5$	$-.504$	.263	.264	95.0
$\gamma = .25$	100	β_0	.5	.500	.391	.406	95.2
		β_1	1	.994	.568	.585	96.3
		β_2	$-.5$	$-.501$	.328	.335	95.7
	200	β_0	.5	.489	.283	.285	94.8
		β_1	1	1.010	.397	.409	95.9
		β_2	$-.5$	$-.502$	.237	.235	94.7
$\gamma = .5$	100	β_0	.5	.459	.356	.364	95.8
		β_1	1	1.081	.545	.523	94.7
		β_2	$-.5$	$-.500$	.297	.299	95.8
	200	β_0	.5	.502	.247	.256	96.3
		β_1	1	1.005	.360	.365	95.4
		β_2	$-.5$	$-.502$	.214	.209	93.6
$\gamma = .75$	100	β_0	.5	.471	.318	.332	96.8
		β_1	1	1.069	.479	.469	93.9
		β_2	$-.5$	$-.505$	.264	.267	95.3
	200	β_0	.5	.506	.228	.233	95.8
		β_1	1	1.000	.327	.326	94.8
		β_2	$-.5$	$-.500$	.192	.187	94.2
$\gamma = 1$	100	β_0	.5	.509	.289	.314	97.8
		β_1	1	1.008	.419	.423	95.5
		β_2	$-.5$	$-.516$	.245	.242	94.2
	200	β_0	.5	.508	.205	.219	97.1
		β_1	1	1.010	.296	.296	95.2
		β_2	$-.5$	$-.508$	.172	.168	94.1

Table 3. Simulation Results From 1,000 Replications Under Misspecified Transformation ($n = 100$)

Model	Parameter	True value	Estimate	SE	ESE	CP (%)
True transformation: $G(x) = (1 + x/2)^{-2}$						
Proportional hazards model	β_0	.5	.165	.290	.311	82.6
	β_1	1	.799	.450	.446	92.5
	β_2	−.5	−.404	.248	.255	94.2
Proportional odds model	β_0	.5	.818	.456	.466	9.8
	β_1	1	1.240	.672	.654	93.0
	β_2	−.5	−.578	.375	.373	95.1
True transformation: $G(x) = \exp[-2\{(1 + x)^{1/2} - 1\}]$						
Proportional hazards model	β_0	.5	.189	.304	.311	84.1
	β_1	1	.868	.464	.442	91.9
	β_2	−.5	−.411	.254	.252	92.7
Proportional odds model	β_0	.5	.960	.463	.472	84.4
	β_1	1	1.205	.650	.652	94.8
	β_2	−.5	−.606	.363	.373	95.0

the first simulation study; thus we conclude that maximum likelihood estimation procedure proposed here not only provides an asymptotically efficient estimator, but also yields good inferential properties for small sample sizes.

Because the proportional hazards cure model and the proportional odds cure model are commonly used in practice, we also conducted a simulation study to examine the performance of the estimates based on these two models when data were generated from a different model. Specifically, we used the same setting for generating the covariates and censoring time as in the other two simulations described earlier, while generating the survival time from either the model with a transformation $(1+x/2)^{-2}$ or $\exp\{-2((1+x)^{1/2}-1)\}$; equivalently, $\gamma = 1/2$ in both classes of (3) and (4). Both choices correspond to a model between the proportional hazards cure model and the proportional odds cure model. The results, based on 1,000 replications, are reported in Table 3. We observe that both the proportional hazards cure model and proportional odds cure models produce notable bias. Interestingly, both models estimate the direction of the coefficients correctly, and the proportional hazards cure model tends to bias towards 0, whereas the opposite is observed for the proportional odds cure model. The bias for the intercept term in both models is large, but the biases for other covariate effects are relatively small. We also observe that even with sizable bias, standard error estimates of the regression coefficients corresponding to the covariates appear to be correct.

Finally, we considered estimation of γ. We generated failure times using the cure model for the transformation class $G(x) = (1 + \gamma x)^{-1/\gamma}$. The simulation study (not shown here) indicates that the performance of the NPMLEs is poor and the convergence in calculating the NPMLEs is often problematic with a sample size of $n = 400$, due to the fact that the likelihood function tends to be flat when γ varies around the true value.

5.2 Application to Melanoma Data

As an illustration, we applied the transformation cure model in (2) to a phase III melanoma clinical trial conducted by the Eastern Cooperative Oncology Group (ECOG), labeled E1690 (Kirkwood et al. 2000). This trial consisted of two treatment arms with a total of $n = 427$ patients on the combined treatment arms, of which 241 patients experienced the event (cancer relapse). The response variable was relapse-free survival (RFS) time (in years). The covariates included in this analysis were treatment (high-dose interferon = 1, observation = 0), age (a continuous variable ranging from 19.13 to 78.05 years, with a mean of 47.93 years), sex (female = 1, male = 0), and nodal category (taking a value of 0 if there were 0 positive nodes or 1 if there were one or more positive nodes). The median follow-up time for this study was 4.33 years, which is considered a sufficient duration of follow-up for this disease. The solid and dotted curves in Figure 1 represent the Kaplan–Meier survival curves for the two treatment arms. We see that a reasonable plateau has been reached at the tails of the survival curves, and it appears that based on this follow-up period, a cure rate model is a suitable approach for the data. Cure rate models for the E1690 data were also considered by Chen, Harrington, and Ibrahim (2002) and were shown to fit better than proper survival models. Based on Figure 1, we considered subjects to be "cured" if they were censored at 5.5 years or beyond. In the dataset, 30 subjects had censored RFS times ≥5.5 years ($Y_i = \infty$). Patients with observed times <5.5 years were either failures or right-censored; and some of those right-censored subjects might indeed have been "cured" patients, but we cannot determine this because of the right-censoring.

We fit the proposed model in (2), where $G(x)$ comes from the family (3) as well as the family (4). We considered values of γ in $[0, 2]$. The MLEs for the regression coefficients of the proposed class of semiparametric transformation cure models were computed using the proposed method. Furthermore, we selected the best transformation among these two classes as the one that maximized the AIC criterion, which is equivalent to the observed log-likelihood function in this case because the number of parameters is constant. Figure 2 plots the observed log-likelihood functions obtained using the two classes of transformations. Interestingly, both classes select the same best transformation, which corresponds to the proportional hazards cure model.

Consequently, we report the results from the proportional hazards cure model in Table 4. The results show that both interferon treatment and sex did not significantly affect RFS, whereas age and nodal category did. Younger patients or those with no positive nodes had significantly better RFS and thus were more likely to be "cured," that is, to not have recurrence

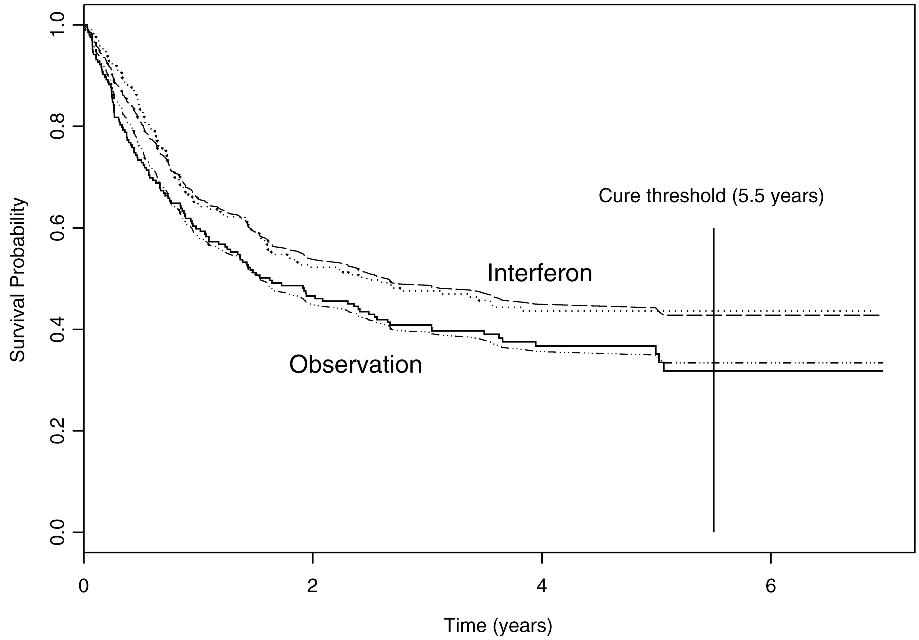

Figure 1. Kaplan–Meier Curves and Predicted Survival Curves of the Interferon and Observation Groups in the E1690 Data. The solid line and the dotted line are the Kaplan–Meier curves; the dashed line and the dot-dashed line are the predicted survival curves.

of melanoma. The results can also be used to estimate the cure rate for each group. For example, the estimated cure rates for a 50-year-old female patient with positive nodes under the interferon treatment is 41.0%. Furthermore, Figure 1 plots the fitted survival function within each treatment group, where the survival function is calculated as the empirical average of the predicted survival functions within each group. The dashed and dot-dashed lines in Figure 1 present the predicted survival functions; these agree quite well with the Kaplan–Meier curves.

As noted earlier, we treated censored subjects with RFS times 5.5 years or greater as "cured" to estimate the parameters. The choice of such a threshold value can be artificial unless it has some biological meaning. Thus we also studied the sensitivity of the estimates to the choice of this threshold value. To do

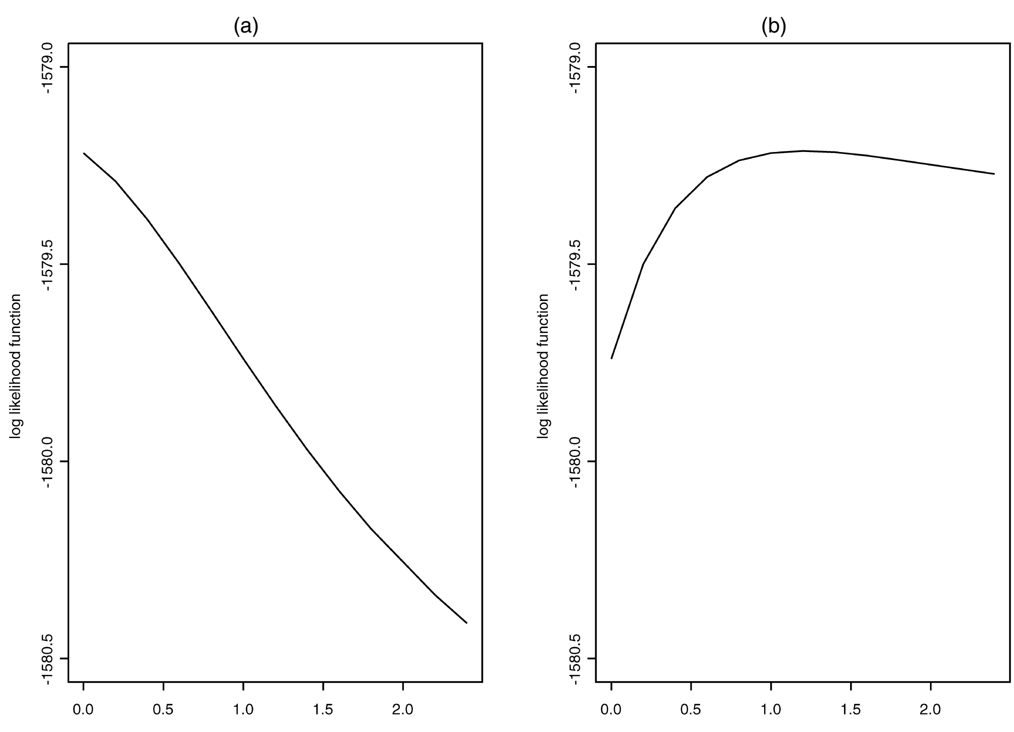

Figure 2. The Observed Log-Likelihood Functions From Different Transformations in the E1690 Data. (a) The log-likelihood functions from transformations $G(x) = (1 + \gamma x)^{-1/\gamma}$. (b) The log-likelihood functions from transformations $G(x) = \exp\{-((1 + x)^{\gamma} - 1)/\gamma\}$.

Table 4. Estimates of Regression Coefficients in the Proportional Hazards Cure Model for the E1690 Data

Cure threshold	Covariate	Estimate	SE	p value
5.1 years	Intercept	−.7977	.3147	.0113
	Treatment	−.2200	.1298	.0901
	Age	.0115	.0050	.0220
	Sex	−.2209	.1371	.1072
	Nodal category	.5519	.1599	.0006
5.5 years	Intercept	−.8027	.3156	.0110
	Treatment	−.2197	.1300	.0911
	Age	.0115	.0050	.0225
	Sex	−.2208	.1374	.1081
	Nodal category	.5520	.1603	.0006
6 years	Intercept	−.7988	.3151	.0112
	Treatment	−.2199	.1298	.0902
	Age	.0115	.0050	.0220
	Sex	−.2209	.1372	.1074
	Nodal category	.5519	.1600	.0006
6.5 years	Intercept	−.7969	.3147	.0113
	Treatment	−.2200	.1297	.0898
	Age	.0115	.0050	.0219
	Sex	−.2210	.1371	.1070
	Nodal category	.5518	.1599	.0006
7 years	Intercept	−.7972	.3148	.0113
	Treatment	−.2200	.1297	.0898
	Age	.0115	.0050	.0219
	Sex	−.2209	.1371	.1071
	Nodal category	.5518	.1599	.0006

this, we varied the threshold value larger than the last failure (5 years), using values of 5.1, 5.5, 6, 6.5, and 7 years. The estimates of the coefficients differ only in the third decimal point, as shown in Table 4.

6. DISCUSSION

We have proposed a class of semiparametric transformation cure models motivated by a specific biological process. This class is quite broad and includes the well-known proportional hazards and proportional odds structures as two special cases. We have provided an efficient algorithm for calculating the MLEs. The maximum likelihood estimation procedure yields efficient estimators of the regression parameters. As one byproduct, because model (2) reduces to a linear transformation model with a special choice of the link function $\theta(\cdot)$, the algorithm in Section 2 provides a simple way of calculating the MLEs for linear transformation models in general. Specifically, for a linear transformation model with $S(t|\mathbf{Z}_i) = G\{\exp(\boldsymbol{\beta}_1^T \mathbf{Z}_i)\Lambda(t)\}$, we can reparameterize to make it a cure rate model by defining $F(t) = \Lambda(t)/\Lambda(\tau)$ and adding an intercept term $\log \Lambda(\tau)$ into the regression. Here τ refers to the termination time of the study. Thus, treating any subjects censored at time τ as "cured," we then implement our proposed algorithm to calculate the MLEs of the parameters.

The cure threshold for the E1690 melanoma data was taken to be 5.5 years. The choice of this cutoff value depends heavily on the dataset at hand and on other practical elements, including the type of disease, the severity or stage, the corresponding treatment, and other patient prognostic factors that require an expert opinion from a physician. A simple guideline is that there should be no failures after the cure threshold. In fact, the estimates from the proposed method are very robust with respect to the choice of this threshold, as shown in Table 4.

The transformation $G(x)$ can be misspecified in practice because of limited knowledge or complex relationships between the covariates and the time-to-event variable. Kosorok et al. (2004) gave some examples in univariate survival data showing that the regression parameters can be estimated up to the correct direction even if $G(x)$ is misspecified. The same ideas can be extended to our proposed model; however, computing such estimable quantities in the presence of nonidentifiable parameters is a very challenging problem.

In deriving (2), we assumed that the promotion time survival function, $S^*(t) = 1 - F(t)$, is the same for all tumor cells. One possible generalization to this is to incorporate covariates into $S^*(t)$, for example, to allow them to be different across treatments. In this case the survival function of the tumor cell for the ith subject would be $\exp\{-\Lambda(t)e^{\zeta^T \mathbf{Z}_i}\}$, where $\mathbf{Z}_i$ is a covariate vector for treatment and other risk factors and $\mathbf{Z}_i$ may share the same components as $\mathbf{X}_i$. Thus the population survival function of interest for subject i is

$$S(t|\mathbf{X}_i, \mathbf{Z}_i) = G\{(1 - e^{-\Lambda(t)e^{\zeta^T \mathbf{Z}_i}})\theta(\mathbf{X}_i)\}.$$

Issues regarding model identifiability and maximum likelihood estimation in these general models are currently under investigation.

APPENDIX: PROOFS

A.1 Proof of Theorem 1

We introduce the following notation. Let $\mathbf{P}_n$ and $\mathbf{P}$ denote the empirical measure of n iid observations and the expectation; that is, for any measurable function $g(\Delta, Y, \mathbf{X})$ in $L_2(P)$,

$$\mathbf{P}_n[g(\Delta, Y, \mathbf{X})] = \frac{1}{n}\sum_{i=1}^n g(\Delta_i, Y_i, \mathbf{X}_i),$$

$$\mathbf{P}[g(\Delta, Y, \mathbf{X})] = E[g(\Delta, Y, \mathbf{X})].$$

From the Lagrange multiplier calculation, $\hat{F}_n$ satisfies the equation that for $Y_i < \infty$,

$$\frac{\Delta_i}{F\{Y_i\}} + \sum_{\infty > Y_j \geq Y_i} \left\{ \Delta_j \frac{G''(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)F(Y_j))\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)}{G'(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)F(Y_j))} + (1 - \Delta_j)\frac{G'(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)F(Y_j))\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)}{G(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)F(Y_j))} \right\} = n\hat{\lambda}_n.$$

We multiply both sides by $\hat{F}_n\{Y_i\}$ and sum over Y_i such that $Y_i < \infty$. We get

$$\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n \Delta_i I(Y_i < \infty) + \int_0^\infty H_n(y, \hat{\boldsymbol{\beta}}_n, \hat{F}_n)\, d\hat{F}_n(y), \qquad \text{(A.1)}$$

where

$$H_n(y, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) = \frac{1}{n}\left[\sum_{Y_j < \infty} \left\{ \Delta_j \frac{G''(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)\hat{F}_n(Y_j))\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)I(Y_j \geq y)}{G'(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)\hat{F}_n(Y_j))} + (1 - \Delta_j)\frac{G'(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)\hat{F}_n(Y_j))\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)I(Y_j \geq y)}{G(\eta(\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_j)\hat{F}_n(Y_j))} \right\}\right].$$

Hence $\hat{F}_n\{Y_i\} = \Delta_i/n(\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol\beta}_n, \hat{F}_n))$. Obviously, from (A.1), $\hat{\lambda}_n$ should be bounded by a constant with probability 1. Thus, by choosing a subsequence, still indexed by $\{n\}$, we assume that $\hat{\lambda}_n \to \lambda^*$. By choosing a further subsequence, we assume that $\hat{\boldsymbol\beta}_n \to \boldsymbol\beta^*$ and $\hat{F}_n \to F^*$ pointwise.

We consider the following class:

$$\mathcal{A}_1 = \Bigg\{ \Delta \frac{G''(\eta(\boldsymbol\beta^T\mathbf{X})F(Y))\eta(\boldsymbol\beta^T\mathbf{X})I(\infty > Y \ge y)}{G'(\eta(\boldsymbol\beta^T\mathbf{X})F(Y))} + (1-\Delta)\frac{G'(\eta(\boldsymbol\beta^T\mathbf{X})F(Y))\eta(\boldsymbol\beta^T\mathbf{X})I(\infty > Y \ge y)}{G(\eta(\boldsymbol\beta^T\mathbf{X})F(Y))} : F \text{ is a distribution function}, \boldsymbol\beta \in \mathcal{B}_0, y \in [0,\infty) \Bigg\}.$$

First, $\{\boldsymbol\beta^T\mathbf{X} : \boldsymbol\beta \in \mathcal{B}_0\}$ and $\{F(Y) : F \text{ is a distribution function}\}$ are both Donsker classes, where the latter follows from theorem 2.7.5 of van der Vaart and Wellner (1996). Because G, G', G'', and η are continuously differentiable functions, the preservation of the Donsker property based on theorem 2.10.6 of van der Vaart and Wellner (1996) implies that the classes

$$\big\{G^{(k)}\big(\eta(\boldsymbol\beta^T\mathbf{X})F(Y)\big) : \boldsymbol\beta \in \mathcal{B}_0, F \text{ is a distribution function}\big\}, \qquad k = 0, 1, 2,$$

and $\{\eta(\boldsymbol\beta^T\mathbf{X}) : \boldsymbol\beta \in \mathcal{B}_0\}$ are Donsker classes. Furthermore, we note that $G'(x)$ and $G(x)$ are both bounded away from 0 when x is in a compact set. Thus the preservation of the Donsker property under the summation, product, and quotient, as given in examples 2.10.7–2.10.9 of van der Vaart and Wellner (1996), gives that the class $\mathcal{A}_1$ is a Donsker class and so is also a Glivenko–Cantelli class. Based on the Glivenko–Cantelli theorem and the bounded convergence theorem, we conclude that uniformly in y, $H_n(y, \hat{\boldsymbol\beta}_n, \hat{F}_n) \to H^*(y)$, where

$$H^*(y) = E\Bigg[\Delta \frac{G''(\eta(\boldsymbol\beta^{*T}\mathbf{X})F^*(Y))\eta(\boldsymbol\beta^{*T}\mathbf{X})I(\infty > Y \ge y)}{G'(\eta(\boldsymbol\beta^{*T}\mathbf{X})F^*(Y))} + (1-\Delta)\frac{G'(\eta(\boldsymbol\beta^{*T}\mathbf{X})F^*(Y))\eta(\boldsymbol\beta^{*T}\mathbf{X})I(\infty > Y \ge y)}{G(\eta(\boldsymbol\beta^{*T}\mathbf{X})F^*(Y))}\Bigg].$$

Moreover, the right side of (A.1) converges to

$$\lambda^* = E\{\Delta I(Y < \infty)\} + E\bigg\{I(Y<\infty)\int_0^Y H^*(y)\,dF^*(y)\bigg\}.$$

Now we wish to show that $|\lambda^* - H^*(y)| > \delta^*$ for some positive constant δ^*. To see that, we first note that from $\sum_{i=1}^n \hat{F}_n\{Y_i\} = 1$,

$$\begin{aligned} 1 &= \sum_{i=1}^n I(Y_i < \infty)\frac{\Delta_i}{n(\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol\beta}_n, \hat{F}_n))} \\ &= \sum_{i=1}^n I(Y_i < \infty)\frac{\Delta_i}{n|\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol\beta}_n, \hat{F}_n)|} \\ &\ge \frac{1}{n}\sum_{i=1}^n I(Y_i < \infty)\frac{\Delta_i}{|\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol\beta}_n, \hat{F}_n)| + \epsilon}, \end{aligned} \qquad \text{(A.2)}$$

for any positive constant ϵ. Because $H_n(y, \hat{\boldsymbol\beta}_n, \hat{F}_n)$ converges uniformly to $H^*(y)$,

$$\frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\frac{\Delta_i}{|\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol\beta}_n, \hat{F}_n)| + \epsilon} - \frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\frac{\Delta_i}{|\lambda^* - H^*(Y_i)| + \epsilon} \to 0.$$

Then, after taking limits on both sides, we obtain $1 \ge E\{\Delta I(Y<\infty)/(|\lambda^* - H^*(Y)| + \epsilon)\}$. Letting $\epsilon \to 0$, by the monotone convergence theorem, we have

$$1 \ge \int_0^\infty \frac{c_0\,dy}{|\lambda^* - H^*(y)|}, \qquad \text{(A.3)}$$

where c_0 is a positive constant. Thus if $\inf_y|\lambda^* - H^*(y)| = 0$, then we claim that there exists a finite y_0 such that $H^*(y_0) = \lambda^*$; otherwise, $H^*(\infty) = \lambda^* = 0$. Then, for large y, $|\lambda^* - H^*(y)| < 1$, which makes (A.3) impossible. Now suppose that there exists a finite y_0 such that $\lambda^* = H^*(y_0)$. Then (A.3) becomes $1 \ge c_0\int_0^\infty dy/|H^*(y_0) - H^*(y)|$. This is impossible, because $H^*(y)$ is continuously differentiable in a neighborhood of y_0. Therefore, there exists a positive constant δ^* such that $|\lambda^* - H^*(y)| > \delta^*$. This implies that when n is large, $|\hat{\lambda}_n - H_n(y, \hat{\boldsymbol\beta}_n, \hat{F}_n)| > \delta^*$. Note that $\hat{F}_n(y) = n^{-1}\sum_{i=1}^n \Delta_i I(Y_i \le y)/|\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol\beta}_n, \hat{F}_n)|$, so $\hat{F}_n(y)$ converges uniformly to $F^*(y) = E\{\Delta I(Y \le y)/|\lambda^* - H^*(Y)|\}$.

We now show that $\boldsymbol\beta^* = \boldsymbol\beta_0$ and $F^* = F_0$. To do so, we construct another function $\tilde{F}$ that has jumps only at Y_i such that $\Delta_i = 1$ and $Y_i < \infty$. Moreover,

$$\tilde{F}_n\{Y_i\} = \frac{1}{nc_n}\frac{\Delta_i}{\tilde{\lambda}_n - H_n(Y_i, \boldsymbol\beta_0, F_0)},$$

where $\tilde{\lambda}_n$ satisfies an equation similar to (A.1) and is given by

$$\tilde{\lambda}_n = \frac{1}{n}\sum_{i=1}^n \Delta_i I(Y_i < \infty) + \int_0^\infty H_n(y, \boldsymbol\beta_0, F_0)\,dF_0(y),$$

and c_n is a constant such that $\sum_{i=1}^n \tilde{F}_n\{Y_i\} = 1$. Furthermore, using the argument of the Glivenko–Cantelli property as before, we can easily show that uniformly in y, $H_n(y, \boldsymbol\beta_0, F_0)$ converges to

$$\tilde{H}(y) = E\Bigg\{\Delta\frac{G''(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(Y))\eta(\boldsymbol\beta_0^T\mathbf{X})I(\infty > Y \ge y)}{G'(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(Y))} + (1-\Delta)\frac{G'(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(Y))\eta(\boldsymbol\beta_0^T\mathbf{X})I(\infty > Y \ge y)}{G(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(Y))}\Bigg\},$$

which, after integration by parts, is equal to $E[\eta(\boldsymbol\beta_0^T\mathbf{X})G'(\eta(\boldsymbol\beta_0^T\mathbf{X}) \times F_0(y))S_c(y|\mathbf{X})]$, where S_c is the conditional survival function of the censoring time. Consequently, direct calculation gives that $\tilde{\lambda}_n$ converges to 0. Furthermore, from

$$c_n\tilde{F}_n(y) = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i I(Y_i \le y)}{\tilde{\lambda}_n - H_n(Y_i, \boldsymbol\beta_0, F_0)},$$

we obtain that uniformly in y, $c_n\tilde{F}_n(y)$ converges to

$$E\Bigg[\frac{\Delta I(Y \le y)}{-E\{S_c(\tilde{y}|\mathbf{X})\eta(\boldsymbol\beta_0^T\mathbf{X})G'(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(\tilde{y}))\}|_{\tilde{y}=Y}}\Bigg] = F_0(y).$$

Hence $c_n \to 1$ and $\tilde{F}_n(y)$ converges to $F_0(y)$ uniformly.

Note that $\hat{F}_n$ is absolutely continuous with respect to $\tilde{F}_n(y)$ with

$$\hat{F}_n(y) = \int_0^y \frac{|\tilde{\lambda}_n - \tilde{H}_n(t, \boldsymbol\beta_0, F_0)|}{|\hat{\lambda}_n - H_n(t, \hat{\boldsymbol\beta}_n, \hat{F}_n)|}\,d\tilde{F}_n(t). \qquad \text{(A.4)}$$

From the foregoing arguments, the integrand in (A.4) is bounded and uniformly converges to $|\tilde{H}(t)|\ /|\lambda^* - H^*(t)|$. We conclude that $F^*(y) = \int_0^y |\tilde{H}(t)|\,dF_0(t)/|\lambda^* - H^*(t)|$. This implies that F^* is absolutely continuous with respect to F_0. Therefore, F^* is also differentiable, and we denote its density function by f^*.

In contrast, because the observed log-likelihood function at $(\hat{\boldsymbol\beta}_n, \hat{F}_n)$ is larger than or equal to the observed log-likelihood function at

$(\boldsymbol{\beta}_0, \tilde{F}_n)$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\Delta_i \log\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \left\{ I(Y_i=\infty)\log\frac{G(\eta(\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i))}{G(\eta(\boldsymbol{\beta}_0^T\mathbf{X}_i))}\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n I(Y_i<\infty) \\
&\quad \times \left\{ \Delta_i \log\frac{G'(\eta(\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i)\hat{F}_n(Y_i))\eta(\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i)}{G'(\eta(\boldsymbol{\beta}_0^T\mathbf{X}_i)\tilde{F}_n(Y_i))\eta(\boldsymbol{\beta}_0^T\mathbf{X}_i)} \right. \\
&\quad \left. + (1-\Delta_i)\log\frac{G(\eta(\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i)\hat{F}_n(Y_i))}{G(\eta(\boldsymbol{\beta}_0^T\mathbf{X}_i)\tilde{F}_n(Y_i))}\right\} \\
&\geq 0.
\end{aligned}$$

We take limits on both sides and note that

$$\frac{1}{n}\sum_{i=1}^n \Delta_i I(Y_i<\infty)\log\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} \to E\left\{\Delta I(Y<\infty)\log\frac{f^*(Y)}{f_0(Y)}\right\}.$$

We obtain $-K((\boldsymbol{\beta}^*, F^*), (\boldsymbol{\beta}_0, F_0)) \geq 0$, where $K(\cdot,\cdot)$ denotes the Kullback–Leibler information of $(\boldsymbol{\beta}^*, F^*)$ with respect to the true parameters. Immediately, we obtain

$$\begin{aligned}
&\left\{-G'\big(\eta(\boldsymbol{\beta}^{*T}\mathbf{X})F^*(Y)\big)\eta(\boldsymbol{\beta}^{*T}\mathbf{X})f^*(Y)\right\}^{\Delta I(Y<\infty)} \\
&\quad \times \left\{G\big(\eta(\boldsymbol{\beta}^{*T}\mathbf{X})F^*(Y)\big)\right\}^{(1-\Delta)I(Y<\infty)+I(Y=\infty)} \\
&= \left\{-G'\big(\eta(\boldsymbol{\beta}_0^T\mathbf{X})F_0(Y)\big)\eta(\boldsymbol{\beta}_0^T\mathbf{X})f_0(Y)\right\}^{\Delta I(Y<\infty)} \\
&\quad \times \left\{G\big(\eta(\boldsymbol{\beta}_0^T\mathbf{X})F_0(Y)\big)\right\}^{(1-\Delta)I(Y<\infty)+I(Y=\infty)} \qquad (A.5)
\end{aligned}$$

for almost every (Δ, X, Y) in its support. According to the second paragraph in Section 3, we obtain $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $F^* = F_0$.

We have shown that for almost every sample in the probability space, we can always choose a subsequence of $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ so that it converges to $(\boldsymbol{\beta}_0, F_0)$. Hence, with probability 1, $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ and $\hat{F}_n(y) \to F_0(y)$ for every $y \in [0,\infty)$. In particular, we obtain $\sup_y |\hat{F}_n(y) - F_0(y)| \to 0$ because of the continuity of F_0.

Remark A.1. When transformation G depends on some unknown parameter γ, where γ belongs to a compact set Γ, the proof of the consistency applies when assumptions (C1) and (C3) are replaced by the following assumptions

(C1′). Parameters $(\boldsymbol{\beta}_0, \gamma_0, F_0)$ are identifiable.

(C3′). $G_\gamma(x)$ is three times differentiable with respect to γ and x, and all of the derivatives are uniformly bounded with $G'_\gamma(x) > 0$.

In particular, (C3′) ensures that the classes of random functions in the foregoing proof are the Glivenko–Cantelli classes, whereas (C1′) ensures that the limit of $(\hat{\boldsymbol{\beta}}_n, \hat{\gamma}_n, \hat{F}_n)$ are the true parameters.

A.2 Proof of Theorem 2

To prove the asymptotic properties of $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$, we recall the definition of $\mathcal{H}$ in Section 3. Furthermore, we abbreviate $l(\boldsymbol{\beta}, F)$ as the log-likelihood function of (5), given by

$$\begin{aligned}
l(\boldsymbol{\beta}, F) &= I(Y<\infty) \\
&\quad \times \left[\Delta \log f + \Delta\log\left\{-G'\big(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y)\big)\eta(\boldsymbol{\beta}^T\mathbf{X})\right\}\right. \\
&\quad \left. + (1-\Delta)\log G\big(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y)\big)\right] \\
&\quad + I(Y=\infty)\log G(\eta(\boldsymbol{\beta}^T\mathbf{X})).
\end{aligned}$$

Let $l_{\boldsymbol{\beta}}(\boldsymbol{\beta}, F)$ denote the derivative of $l(\boldsymbol{\beta}, F)$ with respect to $\boldsymbol{\beta}$, and let $l_F(\boldsymbol{\beta}, F)[\int(h_2 - Q_F[h_2])\,dF]$ denote the derivative of $l(\boldsymbol{\beta}, F)$ along the path $(\boldsymbol{\beta}, F_\epsilon = F + \epsilon\int Q_F(h_2)\,dF)$, $\epsilon \in (-\epsilon_0, \epsilon_0)$ for a small constant ϵ_0, where $Q_F[h_2] = h_2(t) - \int_0^\infty h_2(t)\,dF(t)$. In addition, we define the derivative of $l_{\boldsymbol{\beta}}(\boldsymbol{\beta}, F)$ with respect to $\boldsymbol{\beta}$, denoted by $l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, F)$; the derivative of $l_{\boldsymbol{\beta}}(\boldsymbol{\beta}, F)$ with respect to F along the path $F + \epsilon(\hat{F}_n - F)$, denoted by $l_{\boldsymbol{\beta}F}[\hat{F}_n - F]$; the derivative of $l_F(\boldsymbol{\beta}, F)[\int Q_F(h_2)\,dF]$ with respect to $\boldsymbol{\beta}$, denoted by $l_{F\boldsymbol{\beta}}(\boldsymbol{\beta}, F)[\int Q_F(h_2)\,dF]$; and the derivative $l_F(\boldsymbol{\beta}, F)[\int Q_F(h_2)\,dF]$ with respect to F along the path $F + \epsilon(\hat{F}_n - F)$, denoted by $l_{FF}(\boldsymbol{\beta}, F)[\int Q_F(h_2)\,dF, \hat{F}_n - F]$. Furthermore, define

$$\begin{aligned}
&\Psi_1(\Delta, Y, \mathbf{X}) \\
&= I(Y<\infty)\Delta\left\{\frac{G^{(3)}(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))} - \frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))^2}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))^2}\right\} \\
&\quad + \{(1-\Delta)I(Y<\infty) + I(Y=\infty)\}\frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))}{G(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))} \\
&\quad - \{(1-\Delta)I(Y<\infty) + I(Y=\infty)\}\frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))^2}{G(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))^2}
\end{aligned}$$

and

$$\begin{aligned}
&\Psi_2(\Delta, Y, \mathbf{X}) \\
&= I(Y<\infty)\Delta\frac{G''(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))}{G'(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))} \\
&\quad + \{(1-\Delta)I(Y<\infty) + I(Y=\infty)\}\frac{G'(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))}{G(\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y))}.
\end{aligned}$$

Because $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ maximizes $\mathbf{P}_n l(\boldsymbol{\beta}, F)$, for any $(\mathbf{h}_1, h_2) \in \mathcal{H}$, it follows that

$$\mathbf{P}_n\left\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)^T\mathbf{h}_1 + l_F(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\left[\int Q_{\hat{F}_n}(h_2)\,d\hat{F}_n\right]\right\} = 0.$$

Note that $\mathbf{P}\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, F_0)^T\mathbf{h}_1 + l_F(\boldsymbol{\beta}_0, F_0)[\int Q_{F_0}(h_2)\,dF_0]\} = 0$. Thus we obtain

$$\begin{aligned}
&\sqrt{n}(\mathbf{P}_n - \mathbf{P})\left\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)^T\mathbf{h}_1 + l_F(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\left[\int Q_{\hat{F}_n}(h_2)\,d\hat{F}_n\right]\right\} \\
&= -\sqrt{n}\mathbf{P}\left\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)^T\mathbf{h}_1 + l_F(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\left[\int Q_{\hat{F}_n}(h_2)\,d\hat{F}_n\right]\right\} \\
&\quad + \sqrt{n}\mathbf{P}\left\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, F_0)^T\mathbf{h}_1 + l_F(\boldsymbol{\beta}_0, F_0)\left[\int Q_{F_0}(h_2)\,dF_0\right]\right\}. \qquad (A.6)
\end{aligned}$$

First, by the same arguments as in the consistency proof, the classes of

$$\mathcal{A}_2 = \left\{\left.\frac{G'(x)}{G(x)}, \frac{G''(x)}{G'(x)}\right|_{x=\eta(\boldsymbol{\beta}^T\mathbf{X})F(Y)} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \delta_0, \sup_y |F(y) - F_0(y)| < \delta_0\right\}$$

and

$$\mathcal{A}_3 = \left\{\eta'(\boldsymbol{\beta}^T\mathbf{X})F(Y), \eta(\boldsymbol{\beta}^T\mathbf{X})F(Y) : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \delta_0, \sup_y |F(y) - F_0(y)| < \delta_0\right\}$$

are P-Donsker. In addition, clearly both classes $\{Q_F(h_2): \|h_2\|_V \leq 1, \sup_y |F(y) - F_0(y)| < \delta_0\}$ and $\{\int_0^Y Q_F(h_2)\,dF : \|h_2\|_V \leq 1, \sup_y |F(y) - F_0(y)| < \delta_0\}$ contain the functions of Y with bounded variations, so they are also P-Donsker. Therefore, from the explicit expression of $l_{\boldsymbol{\beta}}$ and l_F, the preservation of the Donsker classes under algebraic operations implies that the class

$$\mathcal{A}_4 = \left\{ l_{\boldsymbol\beta}(\boldsymbol\beta, F)^T \mathbf{h}_1 + l_F(\boldsymbol\beta, F)\left[\int Q_F(h_2)\,dF\right] : \|\mathbf{h}_1\| \le 1, \|h_2\|_V \le 1, \|\boldsymbol\beta - \boldsymbol\beta_0\| + \sup_y |F(y) - F_0(y)| < \delta_0 \right\}$$

is P-Donsker. In contrast, it is straightforward to show that

$$l_{\boldsymbol\beta}(\hat{\boldsymbol\beta}_n, \hat F_n)^T \mathbf{h}_1 + l_F(\hat{\boldsymbol\beta}_n, \hat F_n)\left[\int Q_{\hat F_n}(h_2)\,d\hat F_n\right] \to l_{\boldsymbol\beta}(\boldsymbol\beta_0, F_0)^T \mathbf{h}_1 + l_F(\boldsymbol\beta_0, F_0)\left[\int Q_{F_0}(h_2)\,dF_0\right]$$

uniformly in $(\mathbf{h}_1, h_2) \in \mathcal{H}$. Thus the left side of (A.6) is equal to

$$\sqrt{n}(\mathbf{P}_n - \mathbf{P})\left\{ l_{\boldsymbol\beta}(\boldsymbol\beta_0, F_0)^T \mathbf{h}_1 + l_F(\boldsymbol\beta_0, F_0)\left[\left[\int Q_{F_0}(h_2)\,dF_0\right]\right]\right\} + o_p(1),$$

where $o_p(1)$ is a random variable that converges to 0 in probability in the metric space $l^\infty(\mathcal{H})$. As a result, the left side of (A.6) converges weakly to a mean-0 Gaussian process in $l^\infty(\mathcal{H})$.

Second, simple algebra shows that, uniformly in $(\mathbf{h}_1, h_2) \in \mathcal{H}$,

$$\begin{aligned}
&\bigg| l_{\boldsymbol\beta}(\hat{\boldsymbol\beta}_n, \hat F_n)^T \mathbf{h}_1 + l_F(\hat{\boldsymbol\beta}_n, \hat F_n)\left[\int Q_{\hat F_n}(h_2)\,d\hat F_n\right] \\
&\quad - l_{\boldsymbol\beta}(\boldsymbol\beta_0, F_0)^T \mathbf{h}_1 - l_F(\boldsymbol\beta_0, F_0)\left[\int Q_{F_0}(h_2)\,dF_0\right] \\
&\quad - \bigg\{ (\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0)^T l_{\boldsymbol\beta\boldsymbol\beta}(\boldsymbol\beta_0, F_0)\mathbf{h}_1 \\
&\quad + (\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0)^T l_{F\boldsymbol\beta}(\boldsymbol\beta_0, F_0)\left[\int Q_{F_0}(h_2)\,dF_0\right] \\
&\quad + \mathbf{h}_1^T l_{\boldsymbol\beta F}[\hat F_n - F_0] \\
&\quad + l_{FF}(\boldsymbol\beta_0, F_0)\left[\int Q_{F_0}(h_2)\,dF_0, \hat F_n - F_0\right]\bigg\}\bigg| \\
&\le o_p\{\|\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0\| + \|\hat F_n - F_0\|_{l^\infty}\}.
\end{aligned}$$

Thus, combining with the expressions of $l_{\boldsymbol\beta\boldsymbol\beta}, l_{\boldsymbol\beta F}, l_{F\boldsymbol\beta}$, and l_{FF}, we obtain that the right side of (A.6) equals

$$\begin{aligned}
&-\sqrt{n}\bigg\{ (\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0)^T \Omega_{\boldsymbol\beta}\big(\mathbf{h}_1, Q_{F_0}(h_2)\big) \\
&\quad + \int_0^\infty \Omega_F\big(\mathbf{h}_1, Q_{F_0}(h_2)\big)\,d(\hat F_n - F_0)(y)\bigg\} \\
&\quad + o\big\{\sqrt{n}(\|\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0\| + \|\hat F_n - F_0\|_{l^\infty})\big\},
\end{aligned}$$

where

$$\begin{aligned}
&\Omega_{\boldsymbol\beta}\big(\mathbf{h}_1, Q_{F_0}(h_2)\big) \\
&= E\left[I(Y<\infty)\Delta \frac{\eta''(\boldsymbol\beta^T\mathbf{X})\eta(\boldsymbol\beta^T\mathbf{X}) - \eta'(\boldsymbol\beta^T\mathbf{X})^2}{\eta(\boldsymbol\beta^T\mathbf{X})^2}\mathbf{X}\mathbf{X}^T\mathbf{h}_1\right] \\
&\quad + E\big[\big\{\Psi_1^0(\Delta, Y, \mathbf{X})\eta'(\boldsymbol\beta_0^T\mathbf{X})^2 F_0(Y)^2 \\
&\quad + \Psi_2^0(\Delta, Y, \mathbf{X})\eta''(\boldsymbol\beta_0^T\mathbf{X})F_0(Y)\big\}\mathbf{X}\mathbf{X}^T\mathbf{h}_1\big] \\
&\quad + E\bigg[\big\{\Psi_1^0(\Delta, Y, \mathbf{X})\eta(\boldsymbol\beta_0^T\mathbf{X})\eta'(\boldsymbol\beta_0^T\mathbf{X})F_0(Y) \\
&\quad + \Psi_2^0(\Delta, Y, \mathbf{X})\eta'(\boldsymbol\beta_0^T\mathbf{X})F_0(Y)\big\}\mathbf{X} \\
&\qquad \times \int_0^Y Q_{F_0}(h_2)\,dF_0\bigg]
\end{aligned}$$

and

$$\begin{aligned}
&\Omega_F\big(\mathbf{h}_1, Q_{F_0}(h_2)\big) \\
&= -E\big[I(Y<\infty)\Delta + \Psi_2^0(\Delta, Y, \mathbf{X})\eta(\boldsymbol\beta_0^T\mathbf{X})\{F_0(Y) - I(Y \ge y)\}\big] \\
&\quad \times Q_{F_0}[h_2] \\
&\quad + E\big[\big\{\Psi_1^0(\Delta, Y, \mathbf{X})\eta(\boldsymbol\beta_0^T\mathbf{X})\eta'(\boldsymbol\beta_0^T\mathbf{X})F_0(Y) \\
&\qquad + \Psi_2^0(\Delta, Y, \mathbf{X})\eta'(\boldsymbol\beta_0^T\mathbf{X})F_0(Y)\big\} \\
&\qquad \times \mathbf{X}^T\mathbf{h}_1 I(Y \ge y)\big] \\
&\quad + E\left[I(Y \ge y)\Psi_1^0(\Delta, Y, \mathbf{X})\eta(\boldsymbol\beta_0^T\mathbf{X})^2 \int_0^Y Q_{F_0}(h_2)\,dF_0\right],
\end{aligned}$$

where Ψ_1^0 and Ψ_2^0 have the same expressions as Ψ_1 and Ψ_2 but with $\boldsymbol\beta$ and F replaced by $\boldsymbol\beta_0$ and F_0.

Third, the linear operator $(\Omega_{\boldsymbol\beta}, \Omega_F)$ is a bounded linear operator from the linear space

$$\mathcal{S} = \mathbb{R}^d \times \left\{\tilde h_2 : \|\tilde h_2\|_V < \infty, \int_0^\infty \tilde h_2(y)\,dF_0(y) = 0\right\}$$

to itself. We wish to show that $(\Omega_{\boldsymbol\beta}, \Omega_F)$ is invertible. From the direct calculation, we have

$$\begin{aligned}
&-E\big[I(Y<\infty)\Delta + \Psi_2^0(\Delta, Y, \mathbf{X})\eta(\boldsymbol\beta_0^T\mathbf{X})\{F_0(Y) - I(Y \ge y)\}\big] \\
&\qquad = E\big[G'\big(\eta(\boldsymbol\beta_0^T\mathbf{X})F_0(y)\big)\eta(\boldsymbol\beta_0^T\mathbf{X})S_c(y|\mathbf{X})\big],
\end{aligned}$$

which is negative. Thus, $(\Omega_{\boldsymbol\beta}, \Omega_F)$ can be written as the summation of an invertible operator and a compact operator. By the approach of Rudin (1973), to prove the invertibility of $(\Omega_{\boldsymbol\beta}, \Omega_F)$, it is sufficient to show that $(\Omega_{\boldsymbol\beta}, \Omega_F)$ is one-to-one; that is, if there exists some $(\mathbf{h}_1, \tilde h_2) \in \mathcal{S}$ such that $\Omega_{\boldsymbol\beta}(\mathbf{h}_1, \tilde h_2) = 0$ and $\Omega_F(\mathbf{h}_1, \tilde h_2) = 0$, then we need to show that $\mathbf{h}_1 = 0$ and $\tilde h_2 = 0$. However, we note that, according to the derivation of the Ω's, it holds that

$$\mathbf{h}_1^T \Omega_{\boldsymbol\beta}(\mathbf{h}_1, \tilde h_2) + \int_0^\infty \Omega_{\boldsymbol\beta}(\mathbf{h}_1, \tilde h_2)\tilde h_2\,dF_0 = -E\{l_{\boldsymbol\beta}(\boldsymbol\beta_0, F_0)^T\mathbf{h}_1 + l_F(\boldsymbol\beta_0, F_0)[\tilde h_2]\}^2.$$

We thus obtain that, with probability 1,

$$l_{\boldsymbol\beta}(\boldsymbol\beta_0, F_0)^T\mathbf{h}_1 + l_F(\boldsymbol\beta_0, F_0)[\tilde h_2] = 0.$$

In particular, we choose $Y = \infty$ and obtain $\mathbf{h}_1 = 0$; then we let $Y < \infty$ and $\Delta = 1$ and obtain a homogeneous integral equation for $\tilde h_2$. Such an equation has one trivial solution, $\tilde h_2 = 0$.

Finally, using the inverse of $(\Omega_{\boldsymbol\beta}, \Omega_F)$, denoted by $(\tilde\Omega_{\boldsymbol\beta}, \tilde\Omega_F)$, (A.6) can be written as

$$\begin{aligned}
&\sqrt{n}\left\{(\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0)^T\mathbf{h}_1 + \int_0^\infty \tilde h_2\,d(\hat F_n - F_0)\right\} \\
&= -\sqrt{n}(\mathbf{P}_n - \mathbf{P}) \\
&\quad \times \{l_{\boldsymbol\beta}(\boldsymbol\beta_0, F_0)^T\tilde\Omega_{\boldsymbol\beta}(\mathbf{h}_1, \tilde h_2) + l_F(\boldsymbol\beta_0, F_0)^T\tilde\Omega_F(\mathbf{h}_1, \tilde h_2)\} \\
&\quad + o_p(1)\{\sqrt{n}(\|\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0\| + \|\hat F_n - F_0\|_{l^\infty})\},
\end{aligned}$$

where $o_p(1)$ converges to 0 in probability uniformly in $(\mathbf{h}_1, \tilde h_2) \in \mathcal{S}_0$, where $\mathcal{S}_0$ contains all $(\mathbf{h}_1, \tilde h_2) \in \mathcal{S}$ such that $\|\mathbf{h}_1\| \le 1$ and $\|\tilde h_2\|_V \le 1$. This immediately implies that

$$\sqrt{n}(\|\hat{\boldsymbol\beta}_n - \boldsymbol\beta_0\| + \|\hat F_n - F_0\|_{l^\infty}) = O_p(1).$$

Hence

$$\begin{aligned}&\sqrt{n}\left\{(\hat{\boldsymbol\beta}_n-\boldsymbol\beta_0)^T\mathbf{h}_1+\int_0^\infty \tilde h_2\,d(\hat F_n-F_0)\right\}\\&\quad=-\sqrt{n}(\mathbf{P}_n-\mathbf{P})\\&\qquad\times\left\{l_{\boldsymbol\beta}(\boldsymbol\beta_0,F_0)^T\tilde\Omega_{\boldsymbol\beta}(\mathbf{h}_1,\tilde h_2)+l_F(\boldsymbol\beta_0,F_0)^T\tilde\Omega_F(\mathbf{h}_1,\tilde h_2)\right\}\\&\qquad+o_p(1).\end{aligned}\tag{A.7}$$

Then $\sqrt{n}\{(\hat{\boldsymbol\beta}_n-\boldsymbol\beta_0)^T\mathbf{h}_1+\int_0^\infty \tilde h_2\,d(\hat F_n-F_0)\}$ converges weakly to a Gaussian process, denoted by $GP(\mathbf{h}_1,\tilde h_2)$. The covariance between $GP(\mathbf{h}_1,\tilde h_2)$ and $GP(\mathbf{h}_1^*,\tilde h_2^*)$ is given by

$$\begin{aligned}&E\Big[\left\{l_{\boldsymbol\beta}(\boldsymbol\beta_0,F_0)^T\tilde\Omega_{\boldsymbol\beta}(\mathbf{h}_1,\tilde h_2)+l_F(\boldsymbol\beta_0,F_0)[\tilde\Omega_F(\mathbf{h}_1,\tilde h_2)]\right\}\\&\qquad\times\left\{l_{\boldsymbol\beta}(\boldsymbol\beta_0,F_0)^T\tilde\Omega_{\boldsymbol\beta}(\mathbf{h}_1^*,\tilde h_2^*)+l_F(\boldsymbol\beta_0,F_0)[\tilde\Omega_F(\mathbf{h}_1^*,\tilde h_2^*)]\right\}\Big].\end{aligned}$$

Because for any h_2, $\int h_2\,d(\hat F_n-F_0)=\int Q_{F_0}(h_2)\,d(\hat F_n-F_0)$, the foregoing convergence result also implies the weak convergence result in Theorem 2.

Specifically, if we choose in (A.7) that $\tilde h_2=0$, then we conclude that $\hat{\boldsymbol\beta}_n^T\mathbf{h}_1$ is an asymptotic linear estimator for $\boldsymbol\beta_0^T\mathbf{h}_1$ with its influence function given by

$$l_{\boldsymbol\beta}(\boldsymbol\beta_0,F_0)^T\tilde\Omega_{\boldsymbol\beta}(\mathbf{h}_1,0)+l_F(\boldsymbol\beta_0,F_0)[\tilde\Omega_F(\mathbf{h}_1,0)].$$

This implies that $\hat{\boldsymbol\beta}_n$ is semiparametrically efficient, because the influence function is on the linear space spanned by the score functions for $\boldsymbol\beta_0$ and F_0.

Remark A.2. When the transformation depends on some parameter γ, the foregoing proof can be easily adapted to this case by introducing one more parameter, γ. The results hold if γ_0 is assumed to belong to the interior of Γ, (C1) and (C3) are replaced by (C1$'$) and (C3$'$), and the following assumption also holds:

(C5$'$) If with probability 1,

$$G_\gamma'(\eta(\boldsymbol\beta_0^T\mathbf{X}))\eta'(\boldsymbol\beta_0^T\mathbf{X})\mathbf{X}^T\mathbf{h}_1+\dot G_\gamma(\eta(\boldsymbol\beta_0^T\mathbf{X}))h_3=0,$$

where $\mathbf{h}_1$ and h_3 are constant vectors and $\dot G_\gamma$ denotes the derivative with respect to γ, then $\mathbf{h}_1=\mathbf{0}$ and $h_3=0$.

Note that (C5$'$) is particularly useful for proving the invertibility of the Ω's.

Remark A.3. The profile likelihood function can be used to give a consistent estimate for the asymptotic variance of $\hat{\boldsymbol\beta}_n$. Its justification follows from verifying all of the conditions of theorem 1 of Murphy and van der Vaart (2000). Especially, from the invertibility of the Ω's, we conclude that the information operator for $(\boldsymbol\beta_0,F_0)$ is invertible; therefore, there exits a vector of functions $\mathbf{h}$ with bounded variation such that $l_F^*l_F[\int Q_{F_0}(\mathbf{h})\,dF_0]=l_F^*l_{\boldsymbol\beta}$, where l_F^* is the dual operator of l_F. The function $\int Q_{F_0}(\mathbf{h})\,dF_0$ was called the "least favorable direction" by Murphy and van der Vaart (2000). We then consider the submodel $(\boldsymbol\epsilon,F_{\boldsymbol\epsilon})$, where $F_{\boldsymbol\epsilon}=F+(\boldsymbol\epsilon-\boldsymbol\beta)\int Q_F(\mathbf{h})\,dF$ and $\boldsymbol\epsilon\in\mathbb{R}^d$. It is clear that such a submodel satisfies conditions (8) and (9) in Murphy and van der Vaart (2000). Furthermore, for any $\tilde{\boldsymbol\beta}_n$, we let $\tilde F_n$ be the distribution function maximizing (6) in which $\boldsymbol\beta=\tilde{\boldsymbol\beta}_n$. From the proof of Theorem 1, the same arguments imply that $\tilde F_n$ converges uniformly to F_0 with probability 1. We thus verify condition (10) of Murphy and van der Vaart (2000). As in the proof of Theorem 2, we linearize the likelihood function for $\tilde F_n$, which is equal to

$$0=\mathbf{P}_n\left\{l_F(\tilde{\boldsymbol\beta}_n,\tilde F_n)\left[\int Q_{\tilde F_n}(h_2)\,d\tilde F_n\right]\right\}.$$

Following the same expansion and using the P-Donsker property as used in proving Theorem 2, we obtain

$$\begin{aligned}&\sqrt{n}\int\Omega_F\big(0,Q_{F_0}(h_2)\big)\,d(\tilde F_n-F_0)\\&\quad=\sqrt{n}(\mathbf{P}_n-\mathbf{P})\left\{l_F(\boldsymbol\beta_0,F_0)\left[\int Q_{F_0}[h_2]\,dF_0\right]\right\}\\&\qquad-\sqrt{n}\mathbf{P}\left[l_F(\tilde{\boldsymbol\beta}_n,F_0)\left[\int Q_{F_0}[h_2]\,dF_0\right]\right.\\&\qquad\left.-l_F(\boldsymbol\beta_0,F_0)\left[\int Q_{F_0}[h_2]\,dF_0\right]\right]+o_p(1).\end{aligned}$$

From the invertibility of $\Omega_F(0,\cdot)$, and noting that

$$\begin{aligned}&\left|\mathbf{P}\left[l_F(\tilde{\boldsymbol\beta}_n,F_0)\left[\int Q_{F_0}[h_2]\,dF_0\right]\right.\right.\\&\qquad\left.\left.-l_F(\boldsymbol\beta_0,F_0)\left[\int Q_{F_0}[h_2]\,dF_0\right]\right]\right|\le O_p(\|\tilde{\boldsymbol\beta}_n-\boldsymbol\beta_0\|),\end{aligned}$$

we obtain $\sqrt{n}\|\tilde F_n-F_0\|_{l^\infty}=O_p(\sqrt{n}+\sqrt{n}\|\tilde{\boldsymbol\beta}_n-\boldsymbol\beta_0\|)$. This immediately implies condition (11) (i.e., the no-bias condition) of Murphy and van der Vaart (2000). Furthermore, by the same arguments as used proving Theorem 1, it is straightforward to check that the class

$$\left\{\frac{\partial}{\partial\boldsymbol\epsilon}l(\boldsymbol\epsilon,F_{\boldsymbol\epsilon}):\|\boldsymbol\epsilon-\boldsymbol\beta_0\|+\|\boldsymbol\beta-\boldsymbol\beta_0\|+\|F-F_0\|<\delta_0\right\}$$

is P-Donsker and that the class

$$\left\{\frac{\partial^2}{\partial\boldsymbol\epsilon^2}l(\boldsymbol\epsilon,F_{\boldsymbol\epsilon}):\|\boldsymbol\epsilon-\boldsymbol\beta_0\|+\|\boldsymbol\beta-\boldsymbol\beta_0\|+\|F-F_0\|<\delta_0\right\}$$

is P-Glivenko–Cantelli. Thus all the conditions in theorem 1 of Murphy and van der Vaart (2000) hold, so the results of theorem 1 of Murphy and van der Vaart (2000) are true. One conclusion of this theorem shows the consistency of the variance estimator based on the profile likelihood function.

[Received July 2004. Revised August 2005.]

REFERENCES

Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273–277.

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

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Chen, M. H., Harrington, D. P., and Ibrahim, J. G. (2002), "Bayesian Cure Rate Models for Malignant Melanoma: A Case Study of Eastern Cooperative Oncology Group Trial E1690," *Applied Statistics*, 51, 135–150.

Chen, M. H., Ibrahim, J. G., and Sinha, D. (1999), "A New Bayesian Model for Survival Data With a Surviving Fraction," *Journal of the American Statistical Association*, 94, 909–919.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), "Analysis of Transformation Models With Censored Data," *Biometrika*, 82, 835–845.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

Gray, R. J., and Tsiatis, A. A. (1989), "A Linear Rank Test for Use When the Main Interest Is in Differences in Cure Rates," *Biometrics*, 45, 899–904.

Ibrahim, J. G., Chen, M., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.

Ibrahim, J. G., and Laud, P. W. (1994), "A Predictive Approach to the Analysis of Designed Experiments," *Journal of the American Statistical Association*, 89, 309–319.

Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M., and Blum, R. H. (2000), "High- and Low-Dose Interferon Alfa-2b in High-Risk Melanoma: First Analysis of Intergroup Trial E1690/S9111/C9190," *Journal of Clinical Oncology*, 18, 2444–2458.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004), "Robust Inference for Proportional Hazards Univariate Frailty Regression Models," *The Annals of Statistics*, 32, 1448–1491.

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

Laska, E. M., and Meisner, M. J. (1992), "Nonparametric Estimation and Testing in a Cure Rate Model," *Biometrics*, 48, 1223–1234.

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

Maller, R., and Zhou, X. (1996), *Survival Analysis With Long-Term Survivors*, New York: Wiley.

Murphy, S. A. (1994), "Consistency in a Proportional Hazards Model Incorporating a Random Effect," *The Annals of Statistics*, 22, 712–731.

——— (1995), "Asymptotic Theory for the Frailty Model," *The Annals of Statistics*, 23, 182–198.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997), "Maximal Likelihood Estimate in the Proportional Odds Model," *Journal of the American Statistical Association*, 92, 968–976.

Murphy, S. A., and van der Vaart, A. W. (2000), "On the Profile Likelihood," *Journal of the American Statistical Association*, 95 449–465.

Parner, E. (1998), Asymptotic Theory for the Correlated Gamma-Frailty Model," *The Annals of Statistics*, 26, 183–214.

Pettitt, A. N. (1982), "Inference for the Linear Model Using a Likelihood Based on Ranks," *Journal of the Royal Statistical Society*, Ser. B, 44, 234–243.

Rudin, W. (1973), *Functional Analysis*, New York: McGraw-Hill.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Sposto, R., Sather, H. N., and Baker, S. A. (1992), "A Comparison of Tests of the Difference in the Proportion of Patients Who Are Cured," *Biometrics*, 48, 87–99.

Slud, E., and Vonta, F. (2004), "Consistency of the NPML Estimator in the Right-Censored Transformation Model," *Scandinavian Journal of Statistics*, 31, 21–41.

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

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

Tsodikov, A. (1998), "A Proportional Hazards Model Taking Account of Long-Term Survivors," *Biometrics*, 54, 1508–1516.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

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