Semiparametric transformation models for semicompeting survival data

Huazhen Lin[1], Ling Zhou[1], Chunhong Li[2], and Yi Li[3]

[1]Center of Statistical Research, School of Statistics,

Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

[2]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong.

[3]Department of Biostatistics, University of Michigan, USA.

email: linhz@swufe.edu.cn

SUMMARY: Semicompeting risk outcome data (e.g., time to disease progression and time to death) are commonly collected in clinical trials. However, analysis of these data is often hampered by a scarcity of available statistical tools. As such, we propose a novel semiparametric transformation model that improves the existing models in the following two ways. First, it estimates regression coefficients and association parameters simultaneously. Second, the measure of surrogacy, for example, the proportion of the treatment effect that is mediated by the surrogate and the ratio of the overall treatment effect on the true endpoint over that on the surrogate endpoint, can be directly obtained. We propose an estimation procedure for inference and show that the proposed estimator is consistent and asymptotically normal. Extensive simulations demonstrate the valid usage of our method. We apply the method to a multiple myeloma trial to study the impact of several biomarkers on patients' semicompeting outcomes—namely, time to progression and time to death.

KEY WORDS: Semicompeting risk data; Semiparametric linear transformation model; Surrogate endpoints.

1. Introduction

Terminal events such as death are often the main endpoint of clinical trials on patients with chronic life-threatening diseases (e.g., cancer). In the evolving course of the disease, landmark events (for example, disease progression) are also observed. Such nonterminal events are typically precursors of the main event and serve as important endpoints in clinical trials. It is often of substantial interest to study the association between the landmark event and death, and the marginal distribution of the time to the landmark event and the time to death given treatment and other underlying individual characteristics. The analysis often carries implications of personalized medicine. For example, in a multiple myeloma trial—the motivating study of this paper—the investigators were keen to understand the relationship between disease progression and overall survival, and their respective relationships with the treatment and certain biomarkers, including albumin and myeloma score (the expression level of myeloma cells and their normal plasma precursor cells). The resulting prognostic models shall aid patients' and physicians' decision making.

Denote time to the landmark event by S and time to death by T. Given that the occurrence of terminal events precludes the occurrence of nonterminal events—but not vice versa—S and T fall into the paradigm of semicompeting risk data (Fine, Jiang, and Chappell, 2001). A variety of methods have been proposed to model S and T. For example, Day, Bryant, and Lefkopolou (1997) considered the Clayton-Oakes model (Clayton, 1978; Oakes, 1986) and proposed a test of the independence of T and S. Fine, Jiang and Chappell (2001) provided a closed-form estimator of the association parameter in the Clayton-Oakes model using modified weighted concordance-estimating functions from Oakes (1986), along with an asymptotic variance estimator. Wang (2003) proposed an estimation procedure in this model that is more generally applicable to copula models.

In the aforementioned works, the dependence between the landmark event and death is assessed marginally, with no adjustment for covariates such as sex, age, or treatment group. In practice, the distributions of T and S in the subpopulations defined by treatment, sex or age are considered. Regression methodology offers an opportunity to investigate how patient characteristics influence the landmark event and death. The literature on regression analysis tailored to semicompeting risks is limited. Lin *et al.* (1996) introduced a semiparametric bivariate location-shift model to describe the effect of treatment on the landmark event and death in two-arm randomized studies. The model can be written as follows:

$$H(S) = X\beta + \varepsilon_1 \text{ and } H(T) = X\alpha + \varepsilon_2, \tag{1}$$

where $H(x) = \log(x)$, β and α are parameter scales, $(\varepsilon_1, \varepsilon_2)'$ are correlated error terms with unspecified distribution, and the sole covariate X is the treatment indicator. Chang (2000) extended Lin *et al.*'s method to the semicompeting risk data with a general discrete covariate. This research direction has been further extended to general regression settings in which the nonterminal event is generalized to be recurrent events (Ghosh and Lin, 2003; Lin and Ying, 2003), whereas death still serves as a terminal event. Recently, Ghosh (2009) applied Lin *et al.*'s model to assess surrogacy. It is difficult to extend Lin *et al.* and Chang's method to a high-dimensional discrete covariate or continuous covariates because the complexity of artificial censoring used by Lin *et al.* increases with the cardinality of the support of covariates. Recently, Peng and Fine (2006) proposed a rank estimator for model (1) that avoids excessive artificial censoring and thus is not limited to discrete covariates. However, because the distributions of the error terms are completely unspecified, the aforementioned methods cannot estimate or make inference on the association between S and T based on the bivariate location-shift models (1). To consider both the marginal effect of covariates on the landmark event and the association between S and T, Hsieh, *et al.* (2008) considered a method that combines the copula model and the first model of (1) with either H or the distribution of error is known; however, the authors developed their methodology for discrete covariates. Peng and Fine

(2007) proposed joint models of functional marginal regression models and a time-dependent copula model for semicompeting risks data. Their method works for continuous covariates.

When analyzing nonstandard data such as survival data, an investigator has to consider where to place assumptions and where to keep the model flexible. The methods proposed by Lin *et al.* (1996) and Chang (2000) allowed the error distributions to be unknown but required specification of the transformation functions. The method proposed by Hsieh *et al.* (2008) allows investigators to place an assumption on the transformation function or the distribution of error. However, all these methods require an extra model for the association.

In the present paper, we propose a new approach. Our model not only directly provides the marginal regression models of S and T, but also the association parameter between S and T. To illustrate our idea, we consider the case without covariates. We denote the distributions of S, T and the standard normal variable by F_1, F_2 and Φ, respectively. The probit-type transformations $\Phi^{-1}\{F_1(S)\}\hat{=}H_1(S)$ and $\Phi^{-1}\{F_2(T)\}\hat{=}H_2(T)$ follow the standard normal distribution marginally. The correlation between $H_1(S)$ and $H_2(T)$ within the traditional Gaussian framework is then imposed conventionally and leads to the normal copula model (Li and Lin, 2006). With the covariates in mind, we consider the following models:

$$H_1(S) = \mathbf{X}'\boldsymbol{\beta} + \varepsilon_1 \text{ and } H_2(T) = \mathbf{X}'\boldsymbol{\alpha} + \varepsilon_2, \tag{2}$$

where H_1 and H_2 are unknown monotonic increasing transformation functions, $(\varepsilon_1, \varepsilon_2)' \sim N(0, \Sigma_\rho)$, $\Sigma_\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Here, assume $Var(\varepsilon_2) = Var(\varepsilon_1) = 1$ and that $\mathbf{X}$ excludes the intercept term for the identification of the models. $\mathbf{X}$ can be a continuous covariate, discrete covariate or a combination of continuous and discrete covariates. Model (2) leaves the transformation functions unspecified but require the error distribution to be Gaussian. The reason for this is three-fold. First, the transformation function is more fundamental than the error distribution in estimating the regression coefficients (Lin and Zhou, 2009). Specifically, the misspecification of the transformation function leads to a seriously biased estimator of the regression coefficients, whereas the misspecification of the error distribution leads to a slightly biased or essentially unbiased estimator (Lin and Zhou, 2009). Second, the use of a Gaussian error provides an opportunity to model the association between S and T. Finally, the normal distribution is robust in some degrees (Hanley, 1988; Li and Lin, 2006). Model (2) naturally provides not only the marginal regression models of S and T, but also the association parameter of S and T. Conversely, the models proposed by Lin *et al.* (1996) and Chang (2000) cannot provide the direct association parameter, whereas those proposed by Hsieh *et al.* (2008) and Peng and Fine (2007) require an extra copula model for the association parameter.

The remainder of the paper is organized as follows. Section 2 describes an estimation procedure. Section 3 describes the derivation of the asymptotic properties. Section 4 contains the simulation results and an application to a multiple myeloma trial. Section 5 provides concluding remarks.

2. Estimation Procedure

Let $a \wedge b = \min(a, b)$ and $I(A)$ to be the indicator function for event A. Let C be the time to censoring and $\mathbf{X}$ the p-dimensional covariate vector. Assume that (S, T) and C are conditionally independent given $\mathbf{X}$. We have n observations $(U_{1i}, \delta_{1i}, U_{2i}, \delta_{2i}, \mathbf{X}_i), i = 1, \cdots, n$, a random sample from $(U_1, \delta_1, U_2, \delta_2, \mathbf{X})$, where $U_1 = S \wedge T \wedge C$, $\delta_1 = I(S \leqslant T \wedge C)$, $U_2 = T \wedge C$, and $\delta_2 = I(T \leqslant C)$. Hence, S is censored by the minimum of T and C and not just by C. The dependent censoring will complicate the analysis. For notational simplicity, denote the parameter vectors $\boldsymbol{\beta}, \boldsymbol{\alpha}$ and ρ by $\boldsymbol{\Theta}$.

2·1 *Estimation of the parameters*

The density function of the standard normal random variable is denoted ϕ. Denote $\Gamma_{1i}(\boldsymbol{\beta}, H_1) = H_1(U_{1i}) - \mathbf{X}_i'\boldsymbol{\beta}$ and $\Gamma_{2i}(\boldsymbol{\alpha}, H_2) = H_2(U_{2i}) - \mathbf{X}_i'\boldsymbol{\alpha}$. For each observation i, the likelihood will take one of the four forms defined below, depending on the values of δ_{1i} and δ_{2i}. If both events are observed, that is $\delta_{1i} = 1$, $\delta_{2i} = 1$, then $\mathcal{L}_{i1}(\mathbf{\Theta}; H_1, H_2) \propto \frac{\phi\left\{\Gamma_{1i}(\boldsymbol{\beta}, H_1)\right\} dH_1(U_{1i})}{\sqrt{1-\rho^2}}$ $\times \phi\left[\frac{\Gamma_{2i}(\boldsymbol{\alpha}, H_2) - \rho\{\Gamma_{1i}(\boldsymbol{\beta}, H_1)\}}{\sqrt{1-\rho^2}}\right] dH_2(U_{2i})$; If S_i is observed but T_i is not observed, that is $\delta_{1i} = 1$,$\delta_{2i} = 0$, then $\mathcal{L}_{i2}(\mathbf{\Theta}; H_1, H_2) \propto \phi\left\{\Gamma_{1i}(\boldsymbol{\beta}, H_1)\right\} dH_1(U_{1i}) \times \left(1 - \Phi\left[\frac{\Gamma_{2i}(\boldsymbol{\alpha}, H_2) - \rho\{\Gamma_{1i}(\boldsymbol{\beta}, H_1)\}}{\sqrt{1-\rho^2}}\right]\right)$; If S_i is not observed but T_i is observed, that is $\delta_{1i} = 0$,$\delta_{2i} = 1$, then $\mathcal{L}_{i3}(\mathbf{\Theta}; H_1, H_2) \propto \phi\left\{H_2(U_{2i}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\} dH_2(U_{2i}) \times \left(1 - \Phi\left[\frac{\Gamma_{1i}(\boldsymbol{\beta}, H_1) - \rho\{\Gamma_{2i}(\boldsymbol{\alpha}, H_2)\}}{\sqrt{1-\rho^2}}\right]\right)$; If neither event is observed, that is $\delta_{1i} = 0$, $\delta_{2i} = 0$, then $\mathcal{L}_{i4}(\mathbf{\Theta}; H_1, H_2) \propto \int_{\Gamma_{1i}(\boldsymbol{\beta}, H_1)}^{\infty} \int_{\Gamma_{2i}(\boldsymbol{\alpha}, H_2)}^{\infty} \frac{\phi(x)}{\sqrt{1-\rho^2}} \phi\left(\frac{y - \rho x}{\sqrt{1-\rho^2}}\right) dxdy$. Combining these, the likelihood resulting from observation i yields:

$$\mathcal{L}_i(\mathbf{\Theta}; H_1, H_2) \propto \mathcal{L}_{i1}(\mathbf{\Theta}; H_1, H_2)^{\delta_{1i}\delta_{2i}} \mathcal{L}_{i2}(\mathbf{\Theta}; H_1, H_2)^{\delta_{1i}(1-\delta_{2i})}$$
$$\times \mathcal{L}_{i3}(\mathbf{\Theta}; H_1, H_2)^{(1-\delta_{1i})\delta_{2i}} \mathcal{L}_{i4}(\mathbf{\Theta}; H_1, H_2)^{(1-\delta_{1i})(1-\delta_{2i})}. \quad (3)$$

The likelihood function involves parameters $\mathbf{\Theta}$ and functions H_1 and H_2. To compute the maximum likelihood estimator $\tilde{\mathbf{\Theta}}$, $\tilde{H}_1$ and $\tilde{H}_2$, we note that $\tilde{H}_1$ have and only have positive jumps at the observed uncensored landmark event time, and $\tilde{H}_2$ have and only have positive jumps at the observed uncensored terminal event time. As a result, the maximization problem reduces to a finite dimension problem. It is, however, still infeasible to maximize the likelihood (3) over a large parameter space because the dimension of the space increases with sample size. Since $\mathbf{\Theta}$ is of primary interest, to avoid complicated computation, we propose a new approach to estimate $\mathbf{\Theta}$, H_1 and H_2. Particularly, we use a series of estimating equations described in Section 2.2 to estimate H_1 and H_2, and then estimate $\mathbf{\Theta}$ by maximizing a pseudo-likelihood, which is the likelihood function $\prod_{i=1}^{n} \mathcal{L}_i(\mathbf{\Theta}; H_1, H_2)$, with H_1 and H_2 replaced by the estimated values. Andersen (2005) showed the pseudo-likelihood method is efficient for the parameters. The simulation studies also show that the proposed method, using a full likelihood for the parameters, would be quite efficient for $\mathbf{\Theta}$.

2·2 *Estimation of the transformation functions*

Model (2) is member of the family of semiparametric transformation models (Chen et al., 2002; Zhou, et al., 2009). Statistical inference procedures on the single semiparametric transformation model with independent censoring have been extensively studied. Here, we use the method proposed by Chen *et al.* (2002), which is easy to compute. Let $\boldsymbol{\alpha}_0$ and H_{20} represent the true values of $\boldsymbol{\alpha}$ and H_2, respectively, and $\Lambda(t) = -\log\{1-\Phi(t)\}$ to be the cumulative hazard function of ε_2. Suppose $N_{2i}(t) = \delta_{2i} I(U_{2i} \leqslant t),$ and $Y_{2i}(t) = I(U_{2i} \geqslant t)$. Motivated by the fact that $M_{2i}(t) = N_{2i}(t) - \int_{t_0}^{t} Y_{2i}(s) d\Lambda\{H_{20}(s) - \mathbf{X}_i'\boldsymbol{\alpha}_0\}$ is a martingale process, we estimate $H_2(t)$ by the following estimating equation:

$$\sum_{i=1}^{n} \left(dN_{2i}(t) + Y_{2i}(t) d\log\left[1-\Phi\left\{H_2(t) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}\right]\right) = 0, \tag{4}$$

where H_2 satisfies $H_2(t_0) = -\infty$. This requirement ensures that $\Lambda(a + H_2(t_0)) = 0$ for any finite a. The starting point is t_0 and is equivalent to zero if S and T are time. Here, we allow $t_0 < 0$ so that S and T can be monotonic transformations of time. It is easy to see that the estimator of H_2 is a nondecreasing step function on $[t_0, \infty)$ with $H_2(t_0) = -\infty$ and with jumps only at the observed uncensored terminal event times, denoted by $t_{d,1} < \cdots < t_{d,K}$.

Now consider the estimation of H_1. Because S and T are correlated, the direct use of Chen *et al.*'s method would yield an inconsistent estimator of H_1 because of dependent censoring. Alternatively, using the approach of Hsieh *et al.* (2008), one can estimate $H_1(t)$ based on the identity $EI(U_{1i} \geqslant t, U_{2i} \geqslant t)= S_\rho\left\{H_1(t) - \mathbf{X}_i'\boldsymbol{\beta}, H_2(t) - \mathbf{X}_i'\boldsymbol{\alpha}\right\} P(C_i > t|\mathbf{X}_i)$, where S_ρ is the survival function of $N(0, \Sigma_\rho)$. One problem with this method is that the distribution of the censoring time C must be modeled.

In this paper, we take a different approach, which does not involve the distribution of C. An important observation that leads to our estimator is $S_i \wedge (T_i \wedge C_i) = (S_i \wedge T_i) \wedge C_i$, which implies that the survival analysis in which S_i is the survival time and $T_i \wedge C_i$ is the censoring time can be regarded as the survival analysis in which $S_i \wedge T_i$ is the survival time and C_i is the censoring time. Given that $\mathbf{X}_i$, (S_i, T_i) is independent of C_i, by regarding the survival time as $W_i = S_i \wedge T_i$ and the censoring time as C_i, we obtain an independent censoring problem. Then, applying Chen *et al.*'s (2002) method to the data $\{(W_i \wedge C_i, I(W_i \leqslant C_i), \mathbf{X}_i) : i = 1, \cdots, n\}$ would yield consistent estimators of related parameters and functions. Denoting $W = S \wedge T$, under model (2), H_1 and H_2 are monotonic increasing functions, for any t, one can see that $P(W \geqslant t|\mathbf{X}) = P\{H_1(S) \geqslant H_1(t), H_2(T) \geqslant H_2(t)|\mathbf{X}\} = S_\rho\{H_1(t) - \mathbf{X}'\boldsymbol{\beta}, H_2(t) - \mathbf{X}'\boldsymbol{\alpha}\}$. Hence, the cumulative hazard function of W is given by $\tilde{\Lambda}(t) = -\log\left[S_\rho\{H_1(t) - \mathbf{X}'\boldsymbol{\beta}, H_2(t) - \mathbf{X}'\boldsymbol{\alpha}\}\right]$. Let $N_i(t) = \eta_i I(U_{1i} \leqslant t)$, $\eta_i = I(W_i \leqslant C_i)$ and $Y_i(t) = I(U_{1i} \geqslant t)$, motivated by the fact that $M_i(t) = N_i(t) + \int_{t_0}^t Y_i(s) d\log\left[S_{\rho 0}\{H_{10}(s) - \mathbf{X}_i'\boldsymbol{\beta}_0, H_{20}(s) - \mathbf{X}_i'\boldsymbol{\alpha}_0\}\right]$ is a martingale process and given $\boldsymbol{\Theta}$ and H_2, we estimate $H_1(t)$ by the following equation:

$$\sum_{i=1}^n \left(dN_i(t) + Y_i(t) d\log\left[S_\rho\{H_1(t) - \mathbf{X}_i'\boldsymbol{\beta}, H_2(t) - \mathbf{X}_i'\boldsymbol{\alpha}\}\right]\right) = 0, \tag{5}$$

where $H_1(t_0) = -\infty$. Again, following the estimating equation (5), the estimator $\widehat{H}_1(\cdot)$ of $H_1(\cdot)$ is a step function with jumps at a combination of the observed uncensored terminal and nonterminal event time, denoted by $t_1 < \cdots < t_M$. Solving the system of estimating equations of the infinite number of equations defined by (4) and (5) is equivalent to solving the system of a finite number of equations. In addition, because the estimating equation (4) is independent of H_1, the estimation of the two infinite-dimensional parameters is decomposed into two separate estimations of single infinite-dimensional parameters, which can greatly reduce computational cost (Lin, Yip and Chen, 2009).

2·3 *Algorithm to estimate* $\boldsymbol{\Theta}$, H_1 *and* H_2

Using Chen *et al.*'s (2002) approach, we provide alternative versions of (4) and (5) for easy computation. Using Taylor expansion and noting that $\sup_k |d\widehat{H}_1(t_k)| = O_p(n^{-1})$ and $\sup_k |d\widehat{H}_2(t_{d,k})| = O_p(n^{-1})$, (4) and (5) asymptotically can be rewritten as:

$$\sum_{i=1}^n \left[dN_{2i}(t) - \frac{Y_{2i}(t)\phi\{H_2(t-) - \mathbf{X}_i'\boldsymbol{\alpha}\}}{1 - \Phi\{H_2(t-) - \mathbf{X}_i'\boldsymbol{\alpha}\}} dH_2(t)\right] = 0 \text{ and} \tag{6}$$

$$\sum_{i=1}^n \Bigg(dN_i(t) + \frac{Y_i(t)}{S_\rho\{H_1(t-) - \mathbf{X}'\boldsymbol{\beta}, H_2(t-) - \mathbf{X}'\boldsymbol{\alpha}\}} \Big[S_\rho^{(10)}\{H_1(t-) - \mathbf{X}'\boldsymbol{\beta}, H_2(t-) - \mathbf{X}'\boldsymbol{\alpha}\} dH_1(t) + S_\rho^{(01)}\{H_1(t-) - \mathbf{X}'\boldsymbol{\beta}, H_2(t-) - \mathbf{X}'\boldsymbol{\alpha}\} dH_2(t)\Big]\Bigg) = 0, \tag{7}$$

with $H_1(t_0) = H_2(t_0) = -\infty$, $S_\rho^{(10)}(x_1, x_2) = dS_\rho(x_1, x_2)/dx_1$ and $S_\rho^{(01)}(x_1, x_2) = dS_\rho(x_1, x_2)/dx_2$.

It can be shown that the solution of (4) and (5) and that of (6) and (7) are asymptotically equivalent (Chen et al. 2002). Equations (6) and (7) suggest the following iterative algorithms for $\mathbf{\Theta}$, H_1 and H_2.

Step 0. Choose an initial value of $\mathbf{\Theta}$. Due to the independence of T and C given $\mathbf{X}$, we obtain a consistent estimator $(H_{2n}, \boldsymbol{\alpha}_n)$ of $(H_2, \boldsymbol{\alpha})$ by the method proposed by Chen *et al.* (2002). Then, applying Chen *et al.*'s algorithm to equation (5), we estimate $(H_1, \boldsymbol{\beta}, \rho)$ based on the following estimating equations with $H_1(t_0) = -\infty$:

$\sum_{i=1}^{n} \left(dN_i(t) + Y_i(t) d \log \left[S_\rho \left\{H_1(t) - \mathbf{X}_i'\boldsymbol{\beta}, H_{2n}(t) - \mathbf{X}_i'\boldsymbol{\alpha}_n\right\}\right]\right) = 0,$

$\int_t \sum_{i=1}^{n} \mathbf{X}_i \left(dN_i(t) + Y_i(t) d \log \left[S_\rho \left\{H_1(t) - \mathbf{X}_i'\boldsymbol{\beta}, H_{2n}(t) - \mathbf{X}_i'\boldsymbol{\alpha}_n\right\}\right]\right) = 0,$ and

$\int_t \sum_{i=1}^{n} \left(dN_i(t) + Y_i(t) d \log \left[S_\rho \left\{H_1(t) - \mathbf{X}_i'\boldsymbol{\beta}, H_{2n}(t) - \mathbf{X}_i'\boldsymbol{\alpha}_n\right\}\right]\right) = 0.$

Step 1. Obtain H_2 given $\boldsymbol{\alpha}$. First noting that $H_2(t_{d,1}-) = -\infty$ and using (4), obtain $H_2(t_{d,1})$ by solving $\sum_{i=1}^{n} \left(dN_{2i}(t_{d,1}) + Y_{2i}(t_{d,1}) \log \left[1 - \Phi\left\{H_2(t_{d,1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}\right]\right) = 0$. Then, using (6), obtain $H_2(t_{d,k}), k = 2, \cdots, K$, one-by-one by solving the equation:

$$H_2(t_{d,k}) = \frac{\sum_{i=1}^{n} dN_{2i}(t_{d,k}) + H_2(t_{d,k-1}) \sum_{i=1}^{n} \frac{Y_{2i}(t_{d,k})\phi\left\{H_2(t_{d,k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}{1-\Phi\left\{H_2(t_{d,k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}}{\sum_{i=1}^{n} \frac{Y_{2i}(t_{d,k})\phi\left\{H_2(t_{d,k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}{1-\Phi\left\{H_2(t_{d,k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}}.$$

Step 2. Obtain H_1 given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and H_2. Noting that $H_1(t_1-) = -\infty$ and using (5), obtain $H_1(t_1)$ by solving $\sum_{i=1}^{n} \left\{dN_i(t_1) + Y_i(t_1) \left(\log \left[S_\rho \left\{H_1(t_1) - \mathbf{X}'\boldsymbol{\beta}, H_2(t_1) - \mathbf{X}'\boldsymbol{\alpha}\right\}\right]\right)\right\} = 0$. Then, using (7), obtain $H_1(t_k), k = 2, \cdots, M$ one-by-one by solving the equation:

$$H_1(t_k) = H_1(t_{k-1}) - \frac{\sum_{i=1}^{n} \left[dN_i(t_k) + \frac{Y_i(t_k) S_\rho^{(01)}\left\{H_1(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\beta}, H_2(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}{S_\rho\left\{H_1(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\beta}, H_2(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}} \left\{H_2(t_k) - H_2(t_{k-1})\right\}\right]}{\sum_{i=1}^{n} \frac{Y_i(t_k) S_\rho^{(10)}\left\{H_1(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\beta}, H_2(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}{S_\rho\left\{H_1(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\beta}, H_2(t_{k-1}) - \mathbf{X}_i'\boldsymbol{\alpha}\right\}}},$$

with $H_2(t_1), \cdots, H_2(t_M)$ replaced by their estimators obtained in Step 1, noting that $H_2(t_k) = H_2(t_{k-1})$ if $t_k \notin (t_{d,1}, \cdots, t_{d,K})$.

Step 3. Obtain the estimate of $\boldsymbol{\Theta}$ by maximizing the likelihood $\mathcal{L}_i(\boldsymbol{\Theta}; H_1, H_2)$ defined in (3), with H_1 and H_2 replaced by the estimators obtained in Steps 1 and 2.

Step 4. Repeat Steps 1 to 3 until the prescribed convergence criteria are met.

3. Inference in Large Samples

In this section, we present the large sample properties of all estimators. Let $\widehat{\boldsymbol{\Theta}}$, $\widehat{H}_1(t)$ and $\widehat{H}_2(t)$ denote the estimators of $\boldsymbol{\Theta}$, $H_1(t)$ and $H_2(t)$, respectively. Let $\boldsymbol{\Theta}_0$, $H_{10}(t)$ and $H_{20}(t)$ denote the true values of $\boldsymbol{\Theta}$, $H_1(t)$ and $H_2(t)$, respectively. Regularity conditions for ensuring the central limit theorem for counting process martingales such as those assumed in Fleming and Harrington (1991) are assumed here without specific statement. Let $\tau = \inf\{t : P(S_i \wedge T_i > t) = 0\}$. We assume that τ is finite, $P(S_i \wedge T_i > \tau) > 0$ and $P(C_i = \tau) > 0$. We do this to avoid a lengthy technical discussion about the tail behavior. $\mathbf{X}_i$ is bounded, and H_{10} and H_{20} have continuous and positive derivatives.

Theorem 1. As $n \to \infty$, in probability, we have $|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0| \to 0$, $\sup_{t\in(t_0,\tau)} |\widehat{H}_1(t) - H_{10}(t)| \to 0$, and $\sup_{t\in(t_0,\tau)} |\widehat{H}_2(t) - H_{20}(t)| \to 0$.

Theorem 2. As $n \to \infty$, we have $\sqrt{n}\left(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right) \to N\left\{0, \boldsymbol{\Sigma}^{-1}\Delta\left(\boldsymbol{\Sigma}^{-1}\right)'\right\}$, where $\boldsymbol{\Sigma}$ and Δ are defined in Supplementary Material A (SM-A).

Theorem 3. As $n \to \infty$, for any $t \in (t_0, \tau)$, we have $\sqrt{n}\left\{\widehat{H}_1(t) - H_{10}(t)\right\} \to N\{0, \Sigma_1(t)\}$, and $\sqrt{n}\left\{\widehat{H}_2(t) - H_{20}(t)\right\} \to N\{0, \Sigma_2(t)\}$, where $\Sigma_1(t)$ and $\Sigma_2(t)$ are defined in SM-A.

The proofs of Theorems 1–3 are given in Supplementary Material B (SM-B). From Theorem 3, $\widehat{H}_1(t)$ and $\widehat{H}_2(t)$ converge to $H_{10}(t)$ and $H_{20}(t)$, respectively, at a rate of $n^{-1/2}$. This result shows that we estimate the nonparametric functions $H_1(\cdot)$ and $H_2(\cdot)$ with a parametric convergence rate. A similar conclusion was also reached by Horowitz (1996), Chen (2002), and Zhou, Lin and Johnson (2009).

As shown in Theorem 2, the asymptotic variance of $\widehat{\boldsymbol{\Theta}}$ takes the standard sandwich form $\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\Delta}(\boldsymbol{\Sigma}^{-1})'$. However, the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta}$ are complicated analytic forms involving complicated computations. Here, we use a resampling scheme proposed by Jin, Ying, and Wei (2001) to approximate the asymptotic variance of $\widehat{\boldsymbol{\Theta}}$. First, we generate n exponential random variables $\xi_i, i = 1, \cdots, n$ with a mean of 1 and variance of 1. We solve the following ξ_i-weighted estimation equations and denote the solutions as $\boldsymbol{\Theta}^*$, $H_1^*(t)$ and $H_2^*(t)$ for any $t \in (t_0, \tau)$: $\sum_{i=1}^n \xi_i \frac{\partial \mathcal{L}_i(\boldsymbol{\Theta}; H_1, H_2)}{\partial \boldsymbol{\Theta}} = 0$, $\sum_{i=1}^n \xi_i \left(dN_{2i}(t) + Y_{2i}(t) d \log \left[1 - \Phi\left\{H_2(t) - \mathbf{X}_i' \boldsymbol{\alpha}\right\}\right]\right) = 0$, and $\sum_{i=1}^n \xi_i \left(dN_i(t) + Y_i(t) d \log \left[S_\rho \left\{H_1(t) - \mathbf{X}_i' \boldsymbol{\beta}, H_2(t) - \mathbf{X}_i' \boldsymbol{\alpha}\right\}\right]\right) = 0$, where $H_1(t_0) = -\infty$ and $H_2(t_0) = -\infty$. The estimates $\boldsymbol{\Theta}^*$, $H_1^*(t)$ and $H_2^*(t)$ can be obtained using the same iterative algorithm in Section 2.3. We establish the validity of the proposed resampling method.

Proposition. The conditional distribution of $n^{1/2}(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}})$, given the observed data, converges almost surely to the asymptotic distribution of $n^{1/2}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)$.

The proof of the Proposition can be found in SM-B. Based on the Proposition, one can obtain a large number of realizations of $\boldsymbol{\Theta}^*$ by repeatedly generating $\xi_1, \cdots, \xi_n$ many times. The variance estimate of $\widehat{\boldsymbol{\Theta}}$ can then be approximated by the empirical variance of $\boldsymbol{\Theta}^*$.

4. Assessing the Surrogate Endpoints

An important application of semicompeting risks approaches is assessing the surrogate endpoints. Surrogate endpoints can be used in lieu of other endpoints in evaluating treatments or other interventions. They are useful because they can be measured earlier, more conveniently or more frequently than the endpoint of interest, which is refereed to as the "true" or "final" endpoint (Ellenberg and Hamilton, 1989). In the surrogacy literature, S is the surrogate endpoints and T is the true endpoint. Before a surrogate end point can replace a final end point in the evaluation of an experimental treatment, it must be formally "validated." Prentice (1989) proposed a formal definition of surrogate endpoints and outlined how potential surrogate endpoints could be validated. However, Prentice's criteria are too stringent and are not straightforward to verify. Freedmen *et al.* (1992) introduced the proportion explained, which is the proportion of the treatment effect that is mediated by the surrogate.

Suppose $\mathbf{X}$ is the covariate, S is the surrogate endpoint and T is the true endpoint. We fit the data using the proposed models (1). Then, using the multivariate normal theory, we obtain $\varepsilon_2 = \rho\varepsilon_1 + \varepsilon^*$, where $\varepsilon^* \sim N(0, 1-\rho^2)$ and is independent of ε_1. By coupling this with models (1), we get $H_2(T) = \rho H_1(S) + \mathbf{X}'(\boldsymbol{\alpha} - \rho\boldsymbol{\beta}) + \varepsilon^*$. Hence, by the definition given by Freedman, Graubard and Schatzkin (1992), if X_1 (the first element of $\mathbf{X}$) is the indicator of treatment, the proportion of treatment effect (PTE) explained by the surrogate S is $\rho\beta_1/\alpha_1$, where β_1 and α_1 are the first components of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively. This implies that one can obtain the measure of surrogacy, or the association between S and T by the models (1). In contrast, this does not happen with the proportional hazards model (Lin, Fleming and Degruttola, 1997) or the accelerated failure time model (Lin *et al.*, 1996; Chang, 2000), both of which require an extra model to estimate PTE. Buyse and Molenberghs (1998) proposed replacing the proportion explained by two new measures. The first measure, termed the *relative effect*, is the ratio of the overall treatment effect on the true endpoint over that on the surrogate endpoint. The second measure is the individual-level association between both endpoints, after accounting for the effect of treatment, referred to as the *adjusted association*. Our model also provides the relative effect $RE = \beta_1/\alpha_1$. An RE value is useful only if the variance of $H_1(T)$ and $H_2(S)$ are equivalent (Ghosh, 2009). In our model setting, the variance of $H_1(T)$ and $H_2(S)$ are equal; hence, β_1/α_1 in our models provide a useful measure of surrogacy. In contrast, $S \leqslant T$ in the bivariate location-shift model (Lin *et al.*, 1996; Chang, 2000; Ghosh, 2009), so the variance of the two random variables will generally not be the same. Denote $\beta_1/\alpha_1 = f_1(\boldsymbol{\Theta})$, $\rho\beta_1/\alpha_1 = f_2(\boldsymbol{\Theta})$, $\widehat{RE} = f_1(\widehat{\boldsymbol{\Theta}})$ and $\widehat{PTE} = f_2(\widehat{\boldsymbol{\Theta}})$, Corollary 1 below follows from Theorem 2 in a straightforward fashion.

Corollary 1. $\sqrt{n}\{\widehat{RE} - RE\} \to N(0, \dot{f}_1(\boldsymbol{\Theta})'\boldsymbol{\Sigma}^{-1}\Delta\left(\boldsymbol{\Sigma}^{-1}\right)'\dot{f}_1(\boldsymbol{\Theta}))$, and $\sqrt{n}\{\widehat{PTE} - PTE\} \to N(0, \dot{f}_2(\boldsymbol{\Theta})'\boldsymbol{\Sigma}^{-1}\Delta\left(\boldsymbol{\Sigma}^{-1}\right)'\dot{f}_2(\boldsymbol{\Theta}))$ as $n \to \infty$, where $\dot{f}(\boldsymbol{\Theta}) = df(\boldsymbol{\Theta})/d\boldsymbol{\Theta}$.

5. Simulation

In this section, we describe simulation studies conducted to assess the finite-sample performance of the proposed method by comparing it with existing methods. The existing approaches to analyze semicompeting risk data include (1) the bivariate location-shift regression model (BLSR) proposed by Lin *et al.* (1996); (2) the copula model; and (3) the combination of regression and copula model (CRC; Hsieh, Wang and Ding, 2008; Peng and Fine, 2007). The copula model is not yet ready for regression analysis, so we focus on the comparison of the proposed method with the BLSR and the CRC methods. We use the method proposed by Hsieh, Wang and Ding (2008) as a representation of the CRC methods. For each of the simulation settings, a total of 500 simulations with a sample size of 400 are conducted.

Simulation 1. We expect our method's estimates to be reliable because our method does not require specification of a parametric form for the transformation function. We also assess whether the added robustness is gained at the expense of reduced efficiency. To investigate these issues, we compare the performance of the proposed method with the correct BLSR method (termed CBLSR) and the incorrect BLSR method (termed MBLSR), in which the transformation function is correctly specified and misspecified, respectively. To make such a comparison, we generate data with the sole binary covariate X that takes the value 1 for one-half of the subjects and 0 for the other half, mimicking a binary treatment indicator. The following model generates simulation data: $H_1(S) = \beta X + \varepsilon_1,\ \ H_2(T) = \alpha X + \varepsilon_2$, where $H_1(t) = t$, $H_2(t) = \log t$, $\alpha = \beta = 1$, and $(\varepsilon_1, \varepsilon_2)'$ is a Gaussian vector with a mean of 0 and covariance matrix Σ_ρ, $\rho = 0.5$. We assume S is the time to the nonterminal event, and T is the time to the terminal event. The censoring random variable C is distributed uniformly on $(0, 20)$, so that about 15% of T is censored by C and about 15% of S is censored by $C \wedge T$.

[Table 1 about here.]

In the MBLSR, the transformations were misspecified as $H_1(t) = H_2(t) = t$. Table 1 presents the resulting estimators for β, α and ρ based on the 500 simulations using the proposed method, the CBLSR and MBLSR methods. A useful rule of thumb in evaluating bias is that biases do not have a substantial negative effect on inferences (e.g., by impairing the coverage of confidence intervals) unless the standardized bias (bias as a percentage of the SD) exceeds 40% (Olsen and Schafer, 2001). By this rule, the proposed estimator and the CBLSR method are unbiased. In contrast, the MBLSR estimator is seriously biased and inefficient, especially for α, which is the regression coefficient in the model where the transformation function is misspecified. The comparison of the CBLSR estimator with the MBLSR estimator shows that a correctly specified transformation function plays an important role in the performance of the BLSR methods. The misspecification of the transformation function can lead to the large biases and variances of the coefficient estimators. By comparing the proposed estimates with the CBLSR estimates, we see that although the estimates from the proposed method have a larger bias than the CBLSR estimators, the proposed estimators are more efficient than those of the CBLSR method. As a result, the performances of the two methods are comparable in terms of mean square errors. The CBLSR estimator is a method that leaves the error distribution unspecified, whereas our estimator leaves the transformation function unspecified. Hence, correctly putting assumptions on the transformation functions or on the error distribution may not matter to the inference about the effect of covariates. However, it does matter to the association parameter because the CBLSR cannot directly provide the estimator of the association parameter, while our method does.

Figures 1(a) and 1(b) display the average of the estimated transformation functions and their pointwise 95% confidential intervals, which suggest that the proposed method produces reasonable estimates of the transformation functions.

[Figure 1 about here.]

Simulation 2. Simulation 1 shows that the misspecification of the transformation function will lead to a seriously biased estimator for the BLSR method. Our method requires the specification of the error distribution. A natural question is whether the proposed method is sensitive to the error distribution. To investigate the issue, we generate data similar to those in Simulation 1, except that the errors $(\varepsilon_1, \varepsilon_2)'$ jointly follow a Clayton copular model as $Pr(\varepsilon_1 \geqslant x, \varepsilon_2 \geqslant y) = \phi_\gamma^{-1}\left[\phi_\gamma\{Pr(\varepsilon_1 \geqslant x)\} + \phi_\gamma\{Pr(\varepsilon_2 \geqslant y)\}\right]$ with $\phi_\gamma(\upsilon) = (\upsilon^{-\gamma} - 1)/\gamma$, $\gamma = 0.5$, and both marginal distributions of ε_1 and ε_2 follow the chi-square distribution with one degree of freedom. Therefore, the assumption on the error distribution required by our method is not satisfied, but it follows the requirement of Hsieh, Wang and Ding (2008).

For each set of simulated data, we estimate β, α and the association parameter using the proposed method, the CBLSR method, the MBLSR method, the CRC1 method and the CRC2 method. The CRC1 method is the CRC method with the transformation function correctly specified but the error distribution unspecified, and the CRC2 method is the CRC method with the error distribution correctly specified but the transformation function unspecified. The transformation functions are misspecified as $H_1(t) = H_2(t) = t$ in the MBLSR method. Table 2 presents the bias, the SD and the RMSE of β, α and the association parameter. From Table 2, one can see that our estimator is slightly biased due to the misspecification of the error distribution, while the MBLSR estimator is seriously biased and inefficient. This result implies that the estimation of the effect of covariates is driven more by the assumptions about the form of the transformation function than those about the error distribution. The conclusion is consistent with that founded by Lin and Zhou (2009).

[Table 2 about here.]

To further investigate the robustness of the proposed method, we also conduct the simulation studies with γ varying from 0.01 to 10. The resulting estimators are displayed in Table 3.

From Table 3, one can see that the proposed method has a slight bias when γ is small and essentially no bias for the other case, suggesting that our method is quite robust to the normal assumption. This occurs probably because the transformation function is nonparametric and the normal assumption is fairly robust toward some departure (Hanley, 1988).

[Table 3 about here.]

We also have conducted further simulation studies (denoted as Simulations 3 and 4) to assess the performance of the proposed method when covariates are continuous and to investigate the possible loss due to the using of the estimating equations instead of the maximum likelihood function for the transformation function. The results are reported in the Supplementary Materials C and D and point to the good performance of the proposed method and hint at the appropriateness of data analysis reported in the next section.

6. Analysis of a Multiple Myeloma Trial

The motivating example is a trial involving patients with multiple myeloma, a progressive hematological disease that represents more than 10% of all hematologic cancers. Time to disease progression and overall survival are often the two main clinical outcomes of mutiple myeloma patients, whose overall survival or time-to-disease progression ranges from a few months to more than ten years (Decaux *et al.*, 2008). In an effort to understand the efficicacy of treatment and the clinical heterogeneities among cancer patients, a total of 264 advanced multiple myeloma patients were recruited in a randomized study, wherein patients were randomly assigned to either receive proteasome inhibitor bortezomib (experimental arm) or high-dose dexamethasone (control arm). A number of clinical and laboratory features that may provide prognostic information, including age, gender, tumore proliferative index, albumin and Myeloma score (expression of myeloma markers), were also collected in the study. The purpose of the analysis is to disclose the relationship between disease progression and overall survival, and their respective relationships with the treatment and these potential prognostic factors.

In this study, overall survival was assessed from the date patients received their first dose of study drug to death or censoring, whichever comes first. Time to progression was assessed from the same starting date to disease progression, which can be censored by death. Hence, overall survival (T) and time to progression (S) are two semicompeting outcomes, as the former can censor the latter, but not vice versa. During the course of the clinical trial, patients' median follow-up time was 447 days, and 169 disease progressions and 145 deaths were observed. We consider the following model: $H_1(S) = \mathbf{X}'\boldsymbol{\beta} + \varepsilon_1$, and $H_2(T) = \mathbf{X}'\boldsymbol{\alpha} + \varepsilon_2$, where the covariate vector $\mathbf{X}$ includes treatment status (0=control, 1=experiment), gender (0=male, 1=female), Myeloma score, tumor proliferative index, age and albumin. Except for treatment status and gender, all covariates are continuous. The resulting estimates of the regression coefficients and association parameter and their standard errors are listed in Table 4. We calculated the standard errors via the resampling method described in Section 3, with 400 bootstrap samples. We chose 400 as the sample size by monitoring the stability of the standard errors.

[Table 4 about here.]

Our analysis provides several interesting results. First, Figure 2 depicting the estimated H_1, H_2 suggestes that the form of the transformation functions resembles that of a log function. Hence, the semiparametric transformation models developed for our data can roughly be interpreted as accelerated failure time models. Second, the prognostic factors act differently on the two main endpoints. Specifically, treatment status and age have significant effects on time to progression, but their effects on overall survival are not significant. This suggests that the treatment and age have only short-term effects on patients' outcome, with no long-term effect on patients' overall survival. On the other hand, the tumor proliferative index and myeloma score have significant effects for overall survival but not for time to progression, which suggests that these two prognostic factors have long-term effects on patients' outcomes, although their short-term effect is not significant. It is worth noting that albumin is has an effect on both short- and long-term outcomes, whereas gender is not significant for either outcome. Third, as measured by the correlation parameter, the two outcomes—time to progression and overall survival–are correlated even after controlling for the aforementioned prognostic factors. This hints that, for multiple myeloma, disease progression can indeed be regarded as a precursor or a surrogate for death. Such information would be helpful for designing next generation therapy for multiple myeloma patients. Finally, from the resulting estimators, the proportion of the treatment effect explained by the disease progression, is estimated as $\widehat{PTE} = \widehat{\rho}\widehat{\beta}_1/\widehat{\alpha}_1 = 1.0678$ and its 95% confidence interval is $[-1.4117, 3.5473]$. The relative effect of treatment effect is estimated as $\widehat{RE} = \widehat{\beta}_1/\widehat{\alpha}_1 = 2.7689$ and its 95% confidence interval is $[-0.4558, 5.9937]$. Both confidence intervals contain zero, implying that disease progression is not a surrogate for death if the treatment is of interest.

[Figure 2 about here.]

Finally, we propose a procedure to check the validity of the assumed semiparametric transformation models. First, we randomly divided the data into five subsets of equal size. We use four of the subsets as the training set; the remaining set is used for validation. For each subject in the validation set, we predicted the subject's event number of a landmark event and death up to time t by $\widehat{E}N_i(t) = -\int_{t_0}^{t} Y_i(t) d\log\left\{S_{\widehat{\rho}}\left(\widehat{H}_1(t) - \mathbf{X}_i'\widehat{\boldsymbol{\beta}}, \widehat{H}_2(t) - \mathbf{X}_i'\widehat{\boldsymbol{\alpha}}\right)\right\}$ and $\widehat{E}N_{2i}(t) = -\int_{t_0}^{t} Y_{2i}(t) d\log\left(1 - \Phi\left(\widehat{H}_2(t) - \mathbf{X}_i'\widehat{\boldsymbol{\alpha}}\right)\right)$, respectively. We investigated the performance of the model by examining the prediction error: $PE_{1i} = \int_{t_0}^{\tau}\left(N_i(t) - \widehat{E}N_i(t)\right) d\left\{\sum_{k=1}^{n} N_k(t)\right\}$, $PE_{2i} = \int_{t_0}^{\tau}\left(N_{2i}(t) - \widehat{E}N_{2i}(t)\right) d\left\{\sum_{k=1}^{n} N_{2k}(t)\right\}$. Figures 3 and 4 (in the Supplementary Material E) plot the prediction error against the covariates, suggesting that the prediction error is independent of the covariates—that is, the proposed model basically picks up all of the covariates' information, and so the proposed model (2) is reasonable.

7. Discussion

In the current paper, we propose semiparametric transformation models for semicompeting risk data. Our models allow the transformation function to be unknown, but the error distribution is specified to be normal. In this way, our model can provide direct estimators of the regression analysis and the association parameter. A simple algorithm is provided to estimate the transformation functions, and the proposed estimators are shown to be consistent and asymptotically normal. The simulation studies reveal that our method works well compared with existing methods.

Acknowledgements

Dr. Lin's research is supported by the Chinese Natural Science Foundation (11125104) and the Fundamental Research Funds for the Central Universities.

Supplementary Materials

The appendices referenced in Sections 3, 5 and 6, the code and the data are available with this article at the Biometrics website on Wiley Online Library.

References

Andersen, E. W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis* **11**, 333-350.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

Chang, S. H. (2000). A two-sample comparison for multiple ordered event data. *Biometrics* **56**, 183–189.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659–668.

Day R., Bryant J., and Lefkopoulou M. (1997). Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika* **84**, 45–56.

Decaux, O., Lode, L., et al. (2008). Prediction of Survival in Multiple Myeloma Based on Gene Expression Profiles Reveals Cell Cycle and Chromosomal Instability Signatures in High-Risk Patients and Hyperdiploid Signatures in Low-Risk Patients: A Study of the Intergroupe Francophone du Myelome. *J Clin Oncol* **26**, 4798–4805.

Ellenberg, S. S. and Hamilton, J.M. (1989). Surrogate endpoints in clinical trials: cancer. *Statistics in Medicine* **8**, 405–413.

Fine, J. P., Jiang, H. and Chappell, R. (2001). On semi-competing risks data. *Biometrika* **88**, 907–919.

Fleming, T. R. and Harrington, D. P. (1991). Counting Processes and Survival Analysis. New York: Wiley.

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine* **11**, 167–178.

Ghosh, D. and Lin, D. Y. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics* **59**, 877–885.

Ghosh, D. (2009). On Assessing Surrogacy in a Single Trial Setting Using a Semicompeting Risks Paradigm. *Biometrics* **65**, 521–529.

Hsieh, J., Wang, W. and Ding, A. (2008). Regression analysis based on semicompeting risks data. *J. R. Statist. Soc. B.* **70**, 3–20.

Jin, Z., Ying, Z,. and Wei, L. J. (2001). A simple Resampling Method by Perturbing the Minimand. *Biometrika* **88**, 381–390.

Li, Y. and Lin, X. (2006). Semiparametric Normal Transformation Models for Spatially Correlated Survival Data. *Journal of the American Statistical Association* **101**, 591–603.

Lin, D. Y., Fleming, T. R., and DeGruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16**, 1515– 1527.

Lin, D. Y., Robins, J. M., and Wei L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**, 381–393.

Lin, D.Y. and Ying, Z. (2003). Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics* **4**, 385–398.

Lin, H. Z., YIP, S.F. P. and Chen, F. (2009). Estimating the Population Size for a Multiple List Problem with an Open Population, *Statistica Sinica* **19**, 177-196.

Lin, H. Z. and Zhou, X. H. (2009). A semi-parametric two-part mixed-effects heteroscedastic transformation model for correlated right-skewed semi-continuous data. *Biostatistics* **10**, 640–658.

Oakes, D. (1986). Semiparametric Inference in a Model for Association in Bivariate Survival Data. *Biometrika* **73**, 353–361.

Olsen, M. K. and Schafer, J. (2001). A two-part random-effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.

Peng, L. and Fine, J. P. (2006). Nonparametric estimation with left-truncated semicompeting risks data. *Biometrika* **93**, 367–383.

Peng, L. and Fine, J. P. (2007). Regression Modeling of Semicompeting Risks Data. *Biometrics* **63**, 96–108.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* **8**, 431–440.

Wang W. (2003). Estimating the association parameter for copula models under dependent censoring. *J. Roy. Statist. Soc. B.* **65**, 257–273.

Zhou, X. H., Lin, H. Z and Johnson, E. (2009). Nonparametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 1029–1047.

Received ?? ??. Revised ?? ??. Accepted ?? ??.

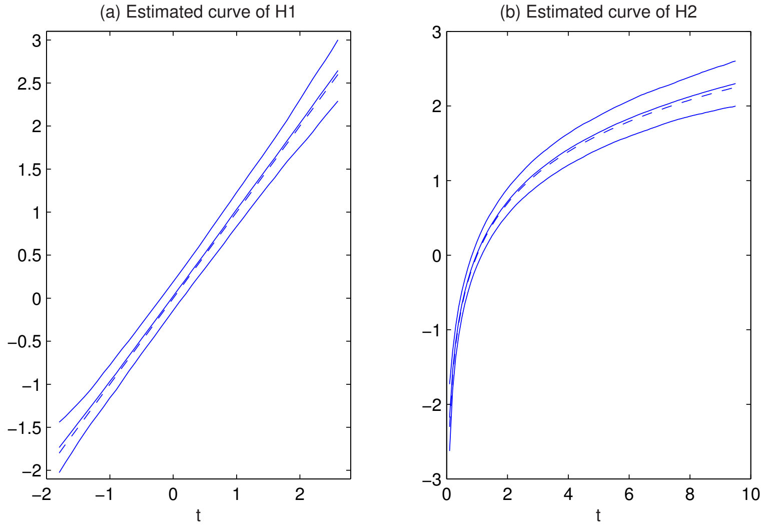

Figure 1. Results of the proposed method based on 500 replications for Simulation 1. (a) The averaged estimates of $H_1(t)$; (b) The averaged estimates of $H_2(t)$ (solid line indicates estimates and 95% confidence limit; dashed line represents the true function).

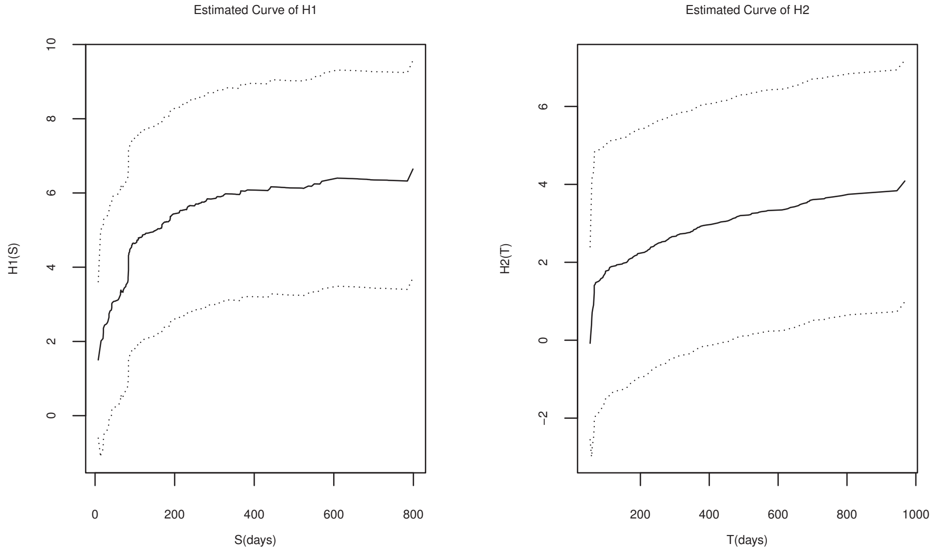

Figure 2. The proposed estimators of H_1 and H_2 for the myeloma data (solid line: estimated; dotted line: 95% confidence limit).

Table 1

The bias, empirical standard deviation (SD), root of mean square error (RMSE) and the empirical coverage probabilities (CP) of 95% *confidence intervals of estimators based on 500 replications using the proposed, the CBLSR and MBLSR methods for Simulation 1.*

	$\widehat{\alpha}$				$\widehat{\beta}$			
Method	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP
Proposed	0.0260	0.1183	0.1211	0.950	0.0277	0.1187	0.1219	0.945
CBLSR	0.0090	0.1157	0.1160		0.0063	0.1144	0.1146	
MBLSR	1.3461	0.3558	1.3923		0.1711	0.1341	0.2174	
	$\widehat{\rho}$							
Proposed	0.0086	0.0441	0.0449	0.950				

Table 2

The bias, SD and RMSE of estimators based on 500 replications using the proposed, CBLSR, MBLSR, CRC1 and CRC2 methods for Simulation 2.

	$\widehat{\beta}$			$\widehat{\alpha}$		
Method	Bias	SD	RMSE	Bias	SD	RMSE
Proposed	-0.0846	0.1064	0.1359	-0.0399	0.0994	0.1071
CBLSR	0.0047	0.0820	0.0821	0.0071	0.0845	0.0848
MBLSR	-0.0855	0.1320	0.1573	-2.0418	0.6230	2.1347
CRC1	0.0645	0.0863	0.1077			
CRC2	0.0512	0.0701	0.0868			

Table 3

The bias, SD and RMSE of estimators using the proposed method with different γ for Simulation 2.

	$\widehat{\alpha}$			$\widehat{\beta}$		
γ	Bias	SD	RMSE	Bias	SD	RMSE
0.01	-0.0279	0.0999	0.1037	-0.1197	0.1100	0.1626
0.1	-0.0277	0.1000	0.1038	-0.1058	0.1085	0.1515
0.5	-0.0243	0.1005	0.1034	-0.0677	0.1054	0.1253
1	-0.0349	0.1008	0.1067	-0.0542	0.1041	0.1174
2	-0.0530	0.1026	0.1155	-0.0465	0.1038	0.1137
4	-0.0395	0.1004	0.1079	-0.0307	0.0996	0.1042
10	-0.0344	0.1008	0.1065	-0.0264	0.1002	0.1036

Table 4
The estimates from the proposed method for the myeloma data.

	$\widehat{\alpha}$			$\widehat{\beta}$		
	Estimator	SD	P-value	Estimator	SD	P-value
Treatment	0.185	0.203	0.363	0.511	0.223	0.022
Gender	0.064	0.160	0.689	-0.012	0.194	0.949
Score	-0.018	0.006	0.007	-0.004	0.008	0.603
Index	-0.439	0.190	0.021	-0.333	0.242	0.169
Age	0.018	0.014	0.189	0.033	0.014	0.015
Albumin	0.083	0.017	0.000	0.077	0.019	0.000
	$\widehat{\rho}$					
	0.386	0.077	0.000			