Marginal analysis of current status data with informative cluster size using a class of semiparametric transformation cure models

Kwok Fai Lam[1,2] | Chun Yin Lee[3] | Kin Yau Wong[3] | Dipankar Bandyopadhyay[4]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

[2]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore

[3]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

[4]Department of Biostatistics, Virginia Commonwealth University, USA

Correspondence
Kwok Fai Lam, The University of Hong Kong, Pokfulam, Hong Kong.
Email: hrntlkf@hku.hk

Summary

This research is motivated by a periodontal disease dataset that possesses certain special features. The dataset consists of clustered current status time-to-event observations with large and varying cluster sizes, where the cluster size is associated with the disease outcome. Also, heavy censoring is present in the data even with long follow-up time, suggesting the presence of a cured sub-population. In this paper, we propose a computationally efficient marginal approach, namely the cluster-weighted generalized estimating equation approach, to analyze the data based on a class of semiparametric transformation cure models. The parametric and non-parametric components of the model are estimated using a Bernstein-polynomial based sieve maximum pseudo-likelihood approach. The asymptotic properties of the proposed estimators are studied. Simulation studies are conducted to evaluate the performance of the proposed estimators in scenarios with different degree of informative clustering and within-cluster dependence. The proposed method is applied to the motivating periodontal disease data for illustration.

KEYWORDS:
cure model, current status data, estimating equations, informative cluster size, survival analysis

1 | INTRODUCTION

This research is motivated by a cross-sectional periodontal disease (PD) study of Gullah-speaking African-American diabetics, or the GAAD study,[1] where the periodontal health condition of each study participant was examined once by recording the tooth site-level surrogate clinical attachment level, or CAL.[2] A CAL $\geq$ 3mm is regarded as moderate to severe incidence of PD, although the (latent) time of incidence of the PD status, both at the site-level, and tooth-level (when averaging CAL values corresponding to the sites of a tooth) remains unknown. The central objective of this study is to identify the risk factors associated with the latent time to PD incidence at the tooth-level.

The GAAD dataset has several interesting characteristics, which cannot be handled easily using traditional methods. First, the response variable of interest, i.e., the age of PD incidence T, for each tooth, cannot be observed directly, and is subject to case-1 interval censoring,[3] with either $\{T \leq Y\}$ or $\{T > Y\}$ being observed, where Y denotes the age at inspection. This type of data is commonly called *current status data*.[4] Second, the disease outcomes of the teeth within a subject are clustered in nature, and are highly correlated as they shared the same oral health condition. Third, the cluster size N, which is the number of (available) teeth of a subject at examination, varies substantially from 3 to 28. It is postulated that N is associated with the disease outcome, since PD is a leading cause of tooth loss. To motivate this further, we group the data into three strata, namely (a) $N \in [1, 10)$; (b) $N \in [10, 20)$; and (c) $N \in [20, \infty)$, and plot the nonparametric Turnbull estimates of the survival curves for current status
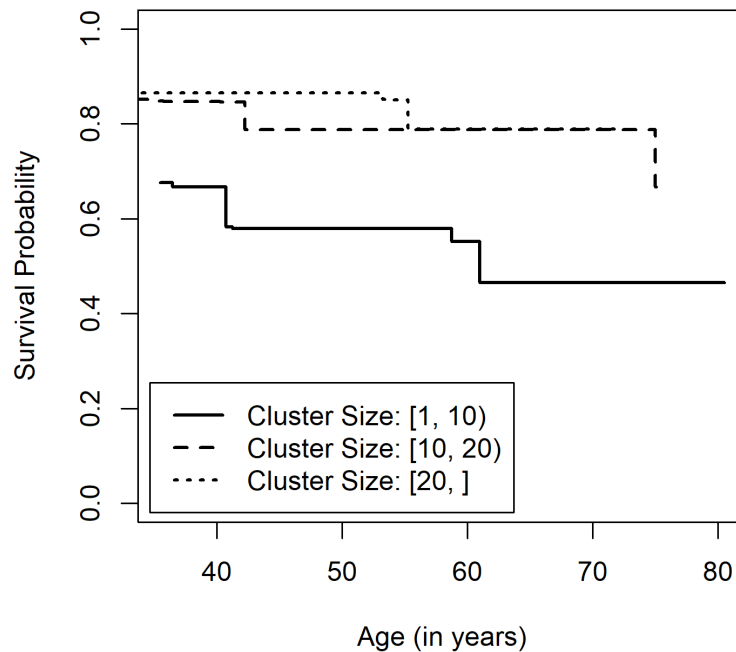
FIGURE 1 Turnbull estimates of the survival functions for the periodontal disease data, stratified by cluster size

data in Figure 1, ignoring the effect of clustering. It is obvious that the survival probabilities increase with cluster sizes, which points to the informative cluster size (ICS) paradigm.[5] Fourth, the survival curves exhibit a plateau behavior at high probability values in their right tails, suggesting the existence of teeth that are non-susceptible to PD.

In standard survival analysis, all subjects are assumed to experience the event of interest eventually, such that a non-susceptible status can never be observed in practice. However, in studies of breast cancer[6,7] and relapse-free survival of the melanoma patients,[8,9] this assumption can be violated, where not all individuals in the population are susceptible to the event of interest (e.g., death due to the specific cancer), leading to longer term follow-up and high censoring rates. Here, it may be more appropriate to postulate the existence of long term survivors, or a cure fraction in the population. The mixture cure (MC) model introduced by Berkson and Gage[10] is presumably the most popular cure-rate model. Here, a subject in the population is simply classified as either cured, or non-cured using a binary random variable U, and a logistic link function is often used to model the association between the cure probability and a set of covariates. In these models, the failure time distribution of the uncured subjects may also be modelled by another set of covariates. Based on right-censored data, various estimation methods have been proposed[6,7,11,12,13] for the MC model. For current status data, Lam and Xue[14] studied a semiparametric MC regression model. Despite its popularity, the MC setup has several limitations, including absence of the proportional hazards (PH) structure, and poor interpretation for describing the underlying biological process,[8] at least in cancer studies. Moreover, this type of mixture models cannot be extended naturally to accommodate clustered data. On the contrary, the promotion time cure (PTC) model[15,16] overcomes the above drawbacks. An appealing feature of this model is that both the cure probability and failure time distribution can be modelled by just one set of covariates, retaining the PH assumption. Later, Zeng et al.[17] proposed a class of transformation cure models by extending the PTC model.

When analyzing clustered interval-censored data, the dependence of the within-cluster observations, generally accommodated by a frailty- or copula-induced model, cannot be ignored, otherwise, loss of estimation efficiency is inevitable. There are two commonly used estimation approaches in the literature. The first one is the so-called *direct* inference, where the regression and dependence parameters are estimated simultaneously in the joint likelihood. In that vein, Zhou et al.[18] studied the semiparametric transformation models for bivariate interval-censored failure time data based on the gamma frailty. Sun and Ding[19] considered

the semiparametric transformation models for bivariate general interval-censored survival data based on the two-parameter Achimedean copula model. Lee et al.[20] studied a class of semiparametric partly linear frailty transformation models for clustered interval-censored data. All the aforementioned papers adopted the sieve maximum likelihood estimation (sieve-MLE) approach where the estimators are shown to be consistent, asymptotically normal and efficient, but their methods do not accommodate data with a cure fraction. Moreover, they only focused on the bivariate case or small cluster size settings as the direct inference procedures are computationally demanding, or even infeasible when the number of event types or the cluster sizes are large. Analysis of clustered interval-censored data with a cure fraction based on a semiparametric frailty-Cox model was considered in Lam and Wong.[21] Therein, the estimation method is able to accommodate data with moderate cluster sizes, but would be quite computationally demanding when the cluster sizes are moderately large (say ≥ 10). The other approach is the *marginal* approach, which only focuses on the estimation of the regression parameters to determine whether a covariate has a significant effect on the responses. Here, similar to generalized estimating equations (GEE), the parameter estimation is performed via the maximization of a pseudo-likelihood function under a working independence assumption, leading to a computationally efficient approach in handling clustered data, even with large cluster sizes. An appealing property of the GEE-type direction is that the estimator of the regression parameter is consistent when the marginal survival model is correctly specified (in the absence of ICS), irrespective of the underlying dependence structure within a cluster. Based on the GEE and Clayton's copula, Kor et al.[22] and Niu and Peng[23] studied Cox-type models for interval-censored data, and cure-rate right censored data, respectively. Furthermore, the cluster-weighted GEE approach[24] can be extended to deal with scenarios of ICS with the survival outcome. Under right-censoring, cluster-weighted Weibull and Cox PH models were proposed,[5,25] that are shown to be equivalent to the within-cluster re-sampling (WCR) approach.[26] For interval-censored data with ICS, comparison of the estimating equation and WCR approaches under a Cox setup[27,28] was conducted, and a linear transformation approach (including Cox PH model and a proportional odds model) was proposed.[29] However, these models do not accommodate a cure fraction. In this paper, we propose a marginal transformation PTC model[17] powered by Bernstein polynomial (BP)[30] based sieve maximum pseudo-likelihood (PL) approach to derive inference in clustered current status data with ICS and a cure fraction.

The rest of the paper is structured as follows. In Section 2, we present a class of semiparametric transformation cure models for the marginal distribution of the failure times. The estimation and inference procedures will also be discussed. In Section 3, we establish the theoretical properties for the proposed estimators. In Section 4, we study the finite sample performance of the proposed estimator under a variety of synthetic data generated scenarios. Application of the proposed model and method to the GAAD dataset is demonstrated in Section 5. Finally, some concluding remarks are provided in Section 6. Proof of theoretical results from Section 3 are relegated to Appendix A1.

2 | MODEL SPECIFICATION AND METHODS

We consider a random sample of n clusters (subjects), consisting of cluster subunits (teeth). Let N_i denote the (random) size of the ith cluster, $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, and let T_{ij} denote the event time of interest, say time to achieving moderate to severe PD (at the tooth-level). Also, let $\boldsymbol{X}_{ij} = (1, X_{ij1}, \ldots, X_{ijp})^{\mathrm{T}}$ denote the vector of covariates for the jth unit in the ith cluster. We leave the association structure of $(T_{i1}, \ldots, T_{iN_i}, N_i)$ unspecified. Conditional on $\boldsymbol{X}_{ij}$, the marginal survival function of T_{ij} has the form

$$S(t \mid \boldsymbol{X}_{ij}) = G\{F(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij}); \rho\}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of unknown regression parameters, F is an unspecified distribution function, and $G(\cdot; \rho)$ is a pre-specified transformation function indexed by a parameter ρ. Note that the cure proportion is $S(\infty \mid \boldsymbol{X}_{ij}) = G\{\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij}); \rho\}$, which is generally non-zero. In this paper, we consider the class of Box-Cox transformations for G, where

$$G(x; \rho) = \begin{cases} \exp\left[-\rho^{-1}\{(1+x)^\rho - 1\}\right], & \rho > 0; \\ (1+x)^{-1}, & \rho = 0. \end{cases}$$

This model generalizes the PTC model[8] by the introduction of a transformation function. In the PTC model, each event time of interest T_{ij} is thought of as the minimum of U independent latent event times, $\widetilde{T}_1, \ldots, \widetilde{T}_U$, where U is a Poisson random variable with mean $\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij})$, and each $\widetilde{T}_k$ ($k = 1, \ldots, U$) follows the cumulative distribution function F; T_{ij} is set to be ∞ if $U = 0$, and that observation is considered cured. With the transformation, the latent variable U is allowed to follow a large class of distributions. In particular, with $\rho = 1$, U follows the Poisson distribution with mean $\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij})$, and with $\rho = 0$, U follows

the geometric distribution with mean $\{1 + \exp(-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_{ij})\}$; these two choices of ρ result in the PH model and the proportional odds (PO) model, respectively.

Suppose that the event times are not exactly observed, but are subject to case-1 interval censoring. Let Y_{ij} be the time of inspection, and $\Delta_{ij} \equiv I(T_{ij} \leq Y_{ij})$ be the event indicator (of PD incidence), for the jth tooth in the ith cluster. The inspection times, calculated as the difference between the subject's clinic visit time and the time of (adult) tooth eruption, may *vary* with the tooth-types. We assume that the event times and the censoring/inspection times are independent, conditional on the cluster size and covariates. The observed data consist of $\mathcal{O}_i = (N_i, Y_{ij}, \Delta_{ij}, \boldsymbol{X}_{ij})_{j=1,\ldots,N_i; i=1,\ldots,n}$. We propose a marginal approach for estimating the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, F)$ in model (1). Analogous to Williamson et al.,[31] we adopt a working independence-within-cluster assumption and maximize the following PL function

$$L_n(\boldsymbol{\theta} \mid \mathcal{O}) = \prod_{i=1}^{n}\prod_{j=1}^{N_i} L_{ij}(\boldsymbol{\theta} \mid \mathcal{O}_{ij})^{w_{ij}} = \prod_{i=1}^{n}\prod_{j=1}^{N_i} \left[\left\{1 - S(Y_{ij}|\boldsymbol{X}_{ij})\right\}^{\Delta_{ij}} S(Y_{ij}|\boldsymbol{X}_{ij})^{1-\Delta_{ij}} \right]^{w_{ij}}, \qquad (2)$$

where L_{ij} denotes the likelihood contribution of the jth observation from the ith cluster, and w_{ij} is the weighting of the corresponding likelihood contribution. The specification of w_{ij}'s is given in Section 2.2.

2.1 | Sieve Maximum Pseudo Likelihood Estimation

Maximization of the PL function (2) is not straightforward, due to the presence of the nonparametric function F. Although the distribution function F has infinite dimension, one can still estimate the function using monotone step functions, or polynomial splines.[32,33] This essentially reduces the dimensionality in estimating F so that the PL function in (2) can be evaluated numerically. In this paper, we adopt a BP to approximate the distribution function F. The BP approach enjoys several merits from the perspective of implementation; it often requires only a few parameters for a decent approximation, and is free from pre-specification of the interior knots as opposed to, for example, B-splines,[34] thereby rendering significant computational efficiency. Also, we can show that the optimization based on the BP can be easily reduced to an unconstrainted nonlinear problem via reparameterization.

Following Lam et al.,[35] we impose a zero-tail constraint $F(t) = 1$ for $t \geq \tau$, where τ is called a cure threshold. In practice, one can set the cure threshold to be $\max_{i,j}(\Delta_{ij}Y_{ij})$, given that the time length of the study is sufficiently long. Let $\mathcal{A} \subset \mathbb{R}^{p+1}$ be a bounded parameter space of the regression parameters $\boldsymbol{\beta}$, and let $\mathcal{B}_F$ be the collection of all non-negative, non-decreasing functions, with an upper bound equals to 1 over the interval $[0, \tau]$. Then, the BP-based sieve for the approximation of $\mathcal{B}_F$ can be defined as

$$\mathcal{B}_{F,n} = \left\{ K_m(t; \boldsymbol{\psi}) : 0 \leq \psi_0 \leq \psi_1 \leq \cdots \leq \psi_m \text{ and } K_m(\tau; \boldsymbol{\psi}) = 1 \right\},$$

where

$$K_m(t; \boldsymbol{\psi}) = \sum_{j=0}^{m} \psi_j \binom{m}{j} \left(\frac{t}{\tau}\right)^j \left(\frac{\tau - t}{\tau}\right)^{m-j}$$

is called a BP of degree m, and $\boldsymbol{\psi}$ is an $(m+1)$-dimensional vector of coefficients of the basis polynomials. We choose m to be an integer which grows at a rate of $O(n^\nu)$ for $0 < \nu < 1$, and the function could approximate[36] arbitrarily closely any smooth true function as $n \to \infty$. Moreover, the monotonicity constraints on $\boldsymbol{\psi}$ ensure that the estimated function is nondecreasing, and $K_m(.; \boldsymbol{\psi})$ is a differentiable function on $[0, \tau]$ that starts from the origin.

Therefore, the parameter estimates for $\boldsymbol{\theta}$ can be obtained by maximizing the PL function $L_n(\boldsymbol{\theta})$ over the sieve space $\mathcal{A} \otimes \mathcal{B}_{F,n}$. Monotonicity of F and $F(\tau) = 1$ can be easily imposed by the reparameterization $\psi_q = \sum_{i=0}^{q} e^{\phi_i} / \sum_{j=0}^{m} e^{\phi_j}$ where $\phi_j \in (-\infty, \infty)$ for $j = 0, \ldots, m$ and $q = 0, \ldots, m$. Hence, the optimization can be done via typical unconstrained methods such as the Newton-Raphson or Nelder-Mead simplex algorithm. In the implementation of the proposed method, the quasi-Newton method of Broyden–Fletcher–Goldfarb–Shanno (BFGS)[37] algorithm is adopted for solving this unconstrained nonlinear optimization problem, which is readily available in standard statistical software, such as R.

2.2 | Statistical Inference for β

For a given m, the sieve maximum PL estimator $\widehat{\boldsymbol{\theta}}_n \equiv (\widehat{\boldsymbol{\beta}}_n, \widehat{F}_n)$ can be obtained by solving the weighted score function

$$\sum_{i=1}^{n} U_i(\boldsymbol{\theta} \mid \mathcal{O}_i) \equiv \sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ij} \frac{\partial \log L_{ij}(\boldsymbol{\theta}|\mathcal{O}_{ij})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

This framework reduces to a traditional unit-based analysis under $w_{ij} = 1$ ($i = 1, \ldots, n; j = 1, \ldots, N_i$) and is referred to as the GEE method in subsequent sections. To adjust for ICS, we adopt $w_{ij} = N_i^{-1}$ for $j = 1, \ldots, N_i$, and the resulting method is referred to as the cluster-weighted GEE (CWGEE) method. It is well-known that the inference of β cannot be simply performed based on the usual Fisher information matrix because the cluster-specific heterogeneity would be completely ignored. Instead, we propose to use a robust sandwich estimator[38] given by

$$\Sigma_n(\widehat{\theta}_n|\mathcal{O}) = H_n(\widehat{\theta}_n|\mathcal{O})^{-1} \left\{ \sum_{i=1}^n U_i(\widehat{\theta}_n|\mathcal{O}_i) U_i(\widehat{\theta}_n|\mathcal{O}_i)^{\mathrm{T}} \right\} H_n(\widehat{\theta}_n|\mathcal{O})^{-1}, \tag{3}$$

where

$$H_n(\theta|\mathcal{O}) = -\sum_{i=1}^n \frac{\partial U_i(\theta|\mathcal{O}_i)}{\partial \theta}.$$

The variance of $\widehat{\beta}_n$ can be approximated by the corresponding elements of Σ_n, and the performance of this variance estimator and the corresponding robust confidence interval estimation will be evaluated in Section 4.

3 | THEORETICAL RESULTS

Let $(N, T_j, Y_j, \Delta_j, X_j)_{j=1,\ldots,N}$ denote the data for a generic cluster. Let β^* and F^* be the true parameter values. The conditions required for the derivation of the asymptotic properties of the proposed estimators are listed below, with some conditions involving a generic positive constant M.

(C1) Conditional on N, the distributions of (Y_j, T_j, X_j) for $j = 1, \ldots, N$ are identical.

(C2) The true parameter value β^* belongs to the interior of the parameter space $\mathcal{A}$. The cure threshold τ satisfies $F^*(\tau) = F^*(\infty) = 1$, and the function F^* is strictly increasing, with bounded rth derivative on $(0, \tau)$ for some $r \geq 1$. The transformation function G is strictly decreasing, with bounded second derivative on $(0, \infty)$, and $G(0) = 1$.

(C3) The support of Y_1 is $[\eta, \tau]$ for some $\eta \in (0, \tau)$, and $P(Y_1 = \tau \mid X_1)$ is bounded away from 0 almost surely.

(C4) With probability 1, $\|X_1\| < M$. Also, the smallest eigenvalue of $\mathrm{E}(X_1 X_1^{\mathrm{T}} \mid Y_1)$ is bounded away from 0 almost surely. In addition, there exists $\kappa \in (0, 1)$, such that $\mathrm{Var}(a^{\mathrm{T}}\widetilde{X}_1 \mid Y_1) \geq \kappa a^{\mathrm{T}}\mathrm{E}(\widetilde{X}_1\widetilde{X}_1^{\mathrm{T}} \mid Y_1)a$ almost surely for all $a \in \mathbb{R}^p$, where $\widetilde{X}_1$ consists of the last p components of X_1.

(C5) The censoring time Y_1 is continuous on (η, τ), and its density function is twice-differentiable on (η, τ) with bounded second derivative.

Condition (C1) requires that conditional on the cluster size, the survival and censoring time distributions are identical across observations of the same cluster. Condition (C2) requires that the cure threshold τ is correctly specified, such that observations surviving beyond τ must be cured. Condition (C3) requires that the follow-up is long enough to cover the cure threshold. This condition is necessary for the identification of the cure proportion. Condition (C4) is imposed for the identifiability of the model parameters. Generally, to satisfy condition (C4), one should avoid selecting covariates that are strongly correlated with each other or with the inspection time. Condition (C5) ensures that the least-favorable direction for F exists, and can be approximated by the BP functions at a fast enough rate.

Let

$$\ell(\beta, F) = \frac{1}{N} \sum_{j=1}^N \left(\Delta_j \log\left[1 - G\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\} \right] + (1 - \Delta_j) \log G\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\} \right),$$

$$\dot{\ell}_\beta(\beta, F) = \frac{1}{N} \sum_{j=1}^N \left[-\frac{\Delta_j}{1 - G\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\}} + \frac{1 - \Delta_j}{G\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\}} \right] G'\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\} F(Y_j)e^{X_j^{\mathrm{T}}\beta} X_j,$$

$$\dot{\ell}_F(\beta, F)[h] = \frac{1}{N} \sum_{j=1}^N \left[-\frac{\Delta_j}{1 - G\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\}} + \frac{1 - \Delta_j}{G\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\}} \right] G'\{F(Y_j)e^{X_j^{\mathrm{T}}\beta}\} e^{X_j^{\mathrm{T}}\beta} h(Y_j).$$

Let $\ell^{(1)}(\boldsymbol{\beta}, F)$, $\dot{\boldsymbol{\ell}}_{\beta}^{(1)}(\boldsymbol{\beta}, F)$, and $\dot{\ell}_F^{(1)}(\boldsymbol{\beta}, F)[h]$ be the first terms of the corresponding summations above. For a d-dimensional vector of functions $\boldsymbol{h} = (h_1, \ldots, h_d)^T$, we let $\dot{\boldsymbol{\ell}}_F(\boldsymbol{\beta}, F)[\boldsymbol{h}] = (\dot{\ell}_F(\boldsymbol{\beta}, F)[h_1], \ldots, \dot{\ell}_F(\boldsymbol{\beta}, F)[h_d])^T$ and $\dot{\ell}_F^{(1)}(\boldsymbol{\beta}, F)[\boldsymbol{h}] = (\dot{\ell}_F^{(1)}(\boldsymbol{\beta}, F)[h_1], \ldots, \dot{\ell}_F^{(1)}(\boldsymbol{\beta}, F)[h_d])^T$. We can show that under conditions (C1)–(C4), the CWGEE method is consistent, such that

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\| + \left\{ \int_\eta^\tau (\widehat{F}_n - F^*)^2(t)\,\mathrm{d}t \right\}^{1/2} \to_{\text{a.s.}} 0.$$

Also, under the same conditions, the rate of convergence of the estimators is given by

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\| + \left\{ \int_\eta^\tau (\widehat{F}_n - F^*)^2(t)\,\mathrm{d}t \right\}^{1/2} = O_p\!\left[n^{-\min\{(1-\nu)/2,\,\nu r/2\}}\right],$$

where r is given in condition (C2), and ν is such that $m = O(n^\nu)$. If we further assume that $r \geq 2$, $\nu \in (1/4, 1/2)$, and condition (C5) holds, then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \to_d \mathrm{N}(\boldsymbol{0}, \widetilde{\boldsymbol{I}}),$$

where

$$\widetilde{\boldsymbol{I}} = \left\{ \mathbb{P}(\dot{\boldsymbol{\ell}}_{\beta}^{(1)} - \dot{\ell}_F^{(1)}[\widetilde{\boldsymbol{h}}])^{\otimes 2} \right\}^{-1} \left\{ \mathbb{P}(\dot{\boldsymbol{\ell}}_{\beta} - \dot{\ell}_F[\widetilde{\boldsymbol{h}}])^{\otimes 2} \right\} \left\{ \mathbb{P}(\dot{\boldsymbol{\ell}}_{\beta}^{(1)} - \dot{\ell}_F^{(1)}[\widetilde{\boldsymbol{h}}])^{\otimes 2} \right\}^{-1},$$

$\mathbb{P}$ denotes the true probability measure, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ for any vector $\boldsymbol{a}$, the score statistics are evaluated at the true parameter values, and $\widetilde{\boldsymbol{h}}$ is the least-favorable direction for F in the univariate model, such that $\mathbb{P}\dot{\ell}_F^{(1)}[\widetilde{\boldsymbol{h}}]\dot{\ell}_F^{(1)}[h] = \mathbb{P}\dot{\boldsymbol{\ell}}_{\beta}^{(1)}\dot{\ell}_F^{(1)}[h]$ for all $h \in L_2(\mathbb{P})$.

The proof of estimation consistency is outlined in Appendix A1. The proofs of the rate of convergence and the asymptotic normality of the estimators follow the arguments in Zhang et al.[27] and Zhou et al.[18] and are omitted.

4 | SIMULATION STUDIES

We conduct a simulation study to assess the finite-sample performance of the parameter estimates from our proposed method. We generate two covariates, namely X_1 from a Bernoulli distribution with probability 0.5 and X_2 from a standard normal distribution. In the simulation, we let $\boldsymbol{X}_{ij} = (1, X_{ij1}, X_{ij2})^T$. To mimic a study where the cluster size can be large and informative, we considered the Gumbel copula model in order to generate survival times from each cluster. First, a cluster-specific latent variable ξ_i is generated according to a positive stable (PS) distribution[39] with parameter γ having the Laplace transform $\Phi_\xi(s) = \exp(-s^\gamma)$ for $i = 1, 2, \ldots, n$. The ICS is generated based on the sampled value of ξ_i. Let the maximum cluster size be $\widetilde{n}$. Specifically, the cluster size N_i is randomly drawn from a binomial distribution with parameters $(\widetilde{n}, 0.75)$ if ξ_i is less than the median of its respective PS distribution, and from a binomial distribution with parameters $(\widetilde{n}, 0.25)$ otherwise. Clusters with sizes 0, 1 and $\widetilde{n}$ are discarded. Samples would be re-generated until we have n clusters. Given $N_i > 1$, the correlated failure times for the ith cluster can be generated based on the Gumbel copula[40] model with the joint survival function given by

$$S(t_1, t_2, \ldots, t_{N_i}|\boldsymbol{X}_{ij}, N_i) = \exp\left(-\left[\sum_{j=1}^{N_i}\left\{-\log S(t_j|\boldsymbol{X}_{ij})\right\}^{1/\gamma}\right]^\gamma\right),$$

where $\gamma \in (0, 1]$ characterizes the degree of association among the failure times of the uncured patients in each cluster, and $S(\cdot \mid \boldsymbol{X}_{ij})$ is given by (1). The parameter γ is set to be 0.2, 0.5, and 0.8 with respective Kendall's tau of 0.8, 0.5, and 0.2. Note that for any simulated observations with marginal survival rates $S(t_{ij}|\boldsymbol{X}_{ij}) < G\{\exp(\boldsymbol{\beta}^T\boldsymbol{X}_{ij}); \rho\}$, they would be regarded as cured. Also, a random monitoring time Y_{ij} is generated by $\min\{\mathrm{Uniform}(0, 8), 4\}$, such that the observation is either left- or right-censored. In the above setup, a small value of cluster-specific ξ_i is associated with a large cluster size, long survival times, and a high chance of being cured. Such a setting is chosen to mimic the GAAD dataset as an illustration, where the patients with more sampled teeth generally have better oral health, and are less prone to the event.

We consider $(n, \widetilde{n}) = (100, 40)$, $(200, 20)$, and $(500, 100)$ and set the parameter vector $(\beta_0, \beta_1, \beta_2)$ to be $(0.5, -1, 1)$, $(-0.5, 1, -1)$, and $(-0.5, 1, -1)$, respectively. The distribution function $F(t)$ is set to be $\{1 - \exp(-\min(t, \tau))\}/\{1 - \exp(-\tau)\}$ with the cure threshold $\tau = 4$. We set the transformation parameter $\rho = 0, 0.5, 1$. In the simulated samples, the cure probability in the population fluctuates around 50%, and the right-censoring rates fluctuates around 60%. The data are analyzed by maximizing $L_n(\boldsymbol{\theta})$ over the sieve space based on the weighted and non-weighted models, that is $w_{ij} = N_i^{-1}$ and $w_{ij} = 1$

TABLE 1 Simulation results corresponding to the estimated regression parameters for $n = 100$ and $\tilde{n} = 40$ (maximum cluster size), with varying γ and ρ. Here, ESD: empirical standard deviation; ESE: estimated standard error; EC: empirical coverage with 95% nominal level

n	γ	ρ	Regression parameter	CWGEE				GEE			
				Bias	ESD	ESE	EC	Bias	ESD	ESE	EC
100	0.2	0	β_0	0.040	0.192	0.194	0.95	−0.711	0.180	0.179	0.04
			β_1	−0.024	0.153	0.152	0.95	−0.142	0.152	0.152	0.86
			β_2	0.028	0.114	0.111	0.94	0.145	0.122	0.120	0.80
		0.5	β_0	0.029	0.137	0.144	0.95	−0.484	0.143	0.145	0.09
			β_1	−0.018	0.125	0.124	0.95	−0.257	0.137	0.137	0.56
			β_2	0.018	0.095	0.095	0.96	0.253	0.112	0.114	0.40
		1	β_0	0.035	0.106	0.113	0.96	−0.315	0.116	0.120	0.25
			β_1	−0.025	0.113	0.112	0.94	−0.309	0.128	0.128	0.32
			β_2	0.025	0.091	0.093	0.94	0.304	0.109	0.111	0.19
100	0.5	0	β_0	0.032	0.163	0.169	0.95	−0.580	0.143	0.142	0.03
			β_1	−0.016	0.140	0.138	0.95	−0.077	0.125	0.123	0.90
			β_2	0.019	0.095	0.093	0.93	0.079	0.086	0.086	0.85
		0.5	β_0	0.027	0.117	0.126	0.95	−0.400	0.110	0.114	0.06
			β_1	−0.015	0.116	0.112	0.93	−0.169	0.110	0.109	0.67
			β_2	0.018	0.085	0.080	0.94	0.170	0.087	0.081	0.47
		1	β_0	0.029	0.090	0.103	0.94	−0.271	0.091	0.093	0.18
			β_1	−0.023	0.099	0.101	0.96	−0.224	0.101	0.102	0.40
			β_2	0.023	0.080	0.079	0.94	0.220	0.082	0.081	0.23
100	0.8	0	β_0	0.037	0.123	0.135	0.95	−0.314	0.100	0.099	0.13
			β_1	−0.018	0.124	0.128	0.95	−0.018	0.108	0.108	0.94
			β_2	0.014	0.078	0.078	0.96	0.021	0.065	0.066	0.94
		0.5	β_0	0.033	0.088	0.100	0.95	−0.222	0.079	0.078	0.19
			β_1	−0.017	0.100	0.103	0.96	−0.069	0.090	0.091	0.88
			β_2	0.016	0.066	0.067	0.95	0.070	0.058	0.058	0.77
		1	β_0	0.032	0.071	0.078	0.94	−0.158	0.066	0.065	0.32
			β_1	−0.020	0.094	0.091	0.93	−0.099	0.085	0.082	0.78
			β_2	0.016	0.064	0.064	0.94	0.096	0.056	0.056	0.60

$(i = 1, \ldots, n; j = 1, \ldots, N_i)$, respectively. The degree of the BP is set to be $m = 3$ for both the weighted and non-weighted methods. For each scenario, we consider 1000 replicates.

Tables 1 and 2 summarize the results based on $(n, \tilde{n}) = (100, 40)$ and $(200, 20)$, and Table S1 in the Supplementary Materials summarizes the results based on $(n, \tilde{n}) = (500, 100)$. The bias, empirical standard deviation (ESD), and average estimated standard error (ESE) of the estimated coefficients are reported. The empirical coverage (EC) is computed for the 95% confidence intervals of the regression parameters, constructed based on the asymptotic normality of the estimators. In all settings with large or small cluster sizes, the ESE matches closely with the ESD which confirms that the robust variance-covariance matrix in (3) is valid for both the weighted or non-weighted methods. Also, in the case with ICS, the estimator based on GEE approach yields large bias in most cases, while the empirical bias based on CWGEE is small and negligible in all cases. This pattern is consistent throughout the change in the number of clusters n, and the change in parameter values over the scenarios. In particular, one can see that the intercept β_0 is severely underestimated by the GEE approach, resulting in an overestimation of the cure proportion. This phenomenon can be explained by the fact that we generate the data with positive association between cluster size and survival time, and that the GEE implicitly assigns larger weights to larger clusters. In addition, the dependence parameter γ in the copula model may actually affect the resulting performance of the GEE approach, but not the CWGEE. As expected, noticeable improvement in the estimation can be seen from the GEE when the within-cluster dependence is getting weaker (i.e., a large value of γ). Moreover, due to the bias of the point estimator, the GEE estimator always underestimates the coverage with a large

TABLE 2 Simulation results corresponding to the estimated regression parameters for $n = 200$ and $\widetilde{n} = 20$ (maximum cluster size), with varying γ and ρ. Here, ESD: empirical standard deviation; ESE: estimated standard error; EC: empirical coverage with 95% nominal level

n	γ	ρ	Regression parameter	CWGEE Bias	ESD	ESE	EC	GEE Bias	ESD	ESE	EC
200	0.2	0	β_0	0.004	0.158	0.161	0.95	−0.852	0.165	0.163	0.01
			β_1	0.016	0.149	0.141	0.94	0.132	0.142	0.133	0.83
			β_2	−0.019	0.094	0.094	0.95	−0.135	0.096	0.097	0.74
		0.5	β_0	0.003	0.128	0.132	0.95	−0.733	0.139	0.139	0.01
			β_1	0.018	0.112	0.113	0.96	0.251	0.115	0.117	0.42
			β_2	−0.022	0.079	0.080	0.95	−0.250	0.089	0.090	0.20
		1	β_0	−0.002	0.108	0.110	0.94	−0.625	0.120	0.121	0.00
			β_1	0.025	0.102	0.099	0.94	0.305	0.112	0.109	0.21
			β_2	−0.024	0.073	0.076	0.96	−0.302	0.088	0.089	0.06
200	0.5	0	β_0	0.001	0.144	0.148	0.95	−0.659	0.131	0.132	0.01
			β_1	0.021	0.135	0.134	0.94	0.078	0.115	0.116	0.89
			β_2	−0.014	0.086	0.085	0.94	−0.074	0.078	0.076	0.84
		0.5	β_0	0.007	0.116	0.121	0.95	−0.565	0.113	0.114	0.00
			β_1	0.017	0.108	0.107	0.94	0.168	0.104	0.102	0.62
			β_2	−0.016	0.072	0.071	0.94	−0.166	0.071	0.070	0.36
		1	β_0	−0.003	0.096	0.101	0.96	−0.494	0.098	0.100	0.00
			β_1	0.026	0.093	0.093	0.95	0.221	0.093	0.093	0.33
			β_2	−0.020	0.067	0.068	0.95	−0.215	0.068	0.069	0.12
200	0.8	0	β_0	0.016	0.114	0.128	0.96	−0.335	0.091	0.095	0.06
			β_1	0.014	0.127	0.129	0.95	0.019	0.106	0.107	0.95
			β_2	−0.019	0.076	0.077	0.95	−0.024	0.063	0.064	0.93
		0.5	β_0	0.018	0.094	0.101	0.96	−0.289	0.082	0.083	0.06
			β_1	0.007	0.102	0.102	0.96	0.060	0.093	0.090	0.89
			β_2	−0.011	0.063	0.064	0.96	−0.066	0.056	0.056	0.79
		1	β_0	0.006	0.084	0.084	0.94	−0.258	0.074	0.073	0.06
			β_1	0.018	0.091	0.089	0.94	0.097	0.082	0.080	0.78
			β_2	−0.019	0.059	0.061	0.93	−0.098	0.052	0.054	0.56

discrepancy, whereas the CWGEE estimator provides an empirical coverage probability of the parameter that closely resembles the nominal level of 95%.

5 | APPLICATION: GAAD DATA

In this section, we illustrate our approach via application to the GAAD dataset mentioned in Section 1. The word 'Gullah' represents unique cultural and linguistic patterns of the African-Americans living on the sea islands of South Carolina.[41] The GAAD study was primarily aimed to explore the relationship between PD and diabetes (determined by HbA1c, or 'glycosylated hemoglobin') in this population. The dataset has $n = 288$ subjects, where 170 subjects have at least one tooth identified with PD incidence. Note, although the time of clinic visit for all available teeth in a subject are the same, the *actual* inspection time of adult permanent teeth varies with tooth-types,[42] which often does not get recorded. In lieu of exact eruption times, we use the approximate permanent dentition times of US adults published by the American Dental Association, and available at https://www.mouthhealthy.org/en/az-topics/e/eruption-charts. With only 913 out of 5461 teeth recording $\Delta = 1$, we also observe heavy censoring, which can be attributed to teeth that are non-susceptible to PD. In addition, Figure 1 presented in Section 1 suggests that the cluster size is informative.

The subject-level covariates under consideration are gender (male/female), smoking status (smoker/non-smoker), glycemic level, or HbA1c (controlled/uncontrolled), and body mass index (BMI), while the only tooth-level covariate is the jaw indicator, i.e., location of the tooth in upper/lower jaw. About 26% of the subjects are smokers. The mean age is 55 years, with a range from 26–87 years. Female subjects seem to be predominant (about 73%) in our data, which is not uncommon among Gullah subjects.[41] About 74% of subjects are obese (BMI $\geq$ 30), and 64% are with uncontrolled HbA1c. In our analysis, we categorize BMI variable into three groups, namely normal (< 25 kg/m^2), overweight ($25 - 30$ kg/m^2) and obese (≥ 30 kg/m^2).

TABLE 3 GAAD data analysis: Estimates for the regression parameters, estimated standard errors (ESE), and 95% confidence intervals, based on the CWGEE and GEE methods.

		CWGEE			GEE		
(m, ρ)	Covariates	Estimates	ESE	95% CI	Estimates	ESE	95% CI
	Intercept	−0.886	0.352	(−1.576, −0.195)	−1.288	0.322	(−1.920, −0.657)
	Smoker (smoker = 1)	0.393	0.203	(−0.003, 0.792)	0.574	0.212	(0.159, 0.989)
	HbA1c (uncontrolled = 1)	0.409	0.198	(0.022, 0.797)	0.508	0.201	(0.114, 0.901)
$(1, 1)$	Gender (female = 1)	−0.760	0.198	(−1.148, −0.372)	−0.773	0.219	(−1.203, −0.343)
	Jaw (upper jaw = 1)	0.082	0.108	(−0.130, 0.294)	0.284	0.085	(0.117, 0.451)
	BMI ($25 - 30$)	0.330	0.325	(−0.307, 0.967)	0.176	0.349	(−0.509, 0.861)
	BMI (≥ 30)	−0.091	0.291	(−0.662, 0.480)	−0.051	0.324	(−0.687, 0.584)

First, our proposed survival model is fitted to the GAAD dataset, based on the CWGEE approach. For the implementation of the proposed method, one has to choose the degree of the BP, and the transformation parameter value. We perform a two-dimensional grid search over (m, ρ), and select the best model based on the Akaike information criterion (AIC).[43] The AIC is given by $-\log \hat{L}_n + 2(m + 1 + p)$, where $\hat{L}_n$ is the PL evaluated at the sieve-MLE. Although one can select ρ with the lowest AIC, a model will not have much practical value unless it has a good interpretation. We search m from 1 to 6 and ρ over $\{0, 0.1, \ldots, 1\}$. The cluster-weighted method with $(m, \rho) = (1, 1)$ achieves the smallest AIC, suggesting that the PTC model gives a better fit than the PO cure model; the estimated covariate effects are similar for different choices of m. Then, the same set of (m, ρ) is applied to the GEE approach for comparison. The estimates for the regression parameters, robust standard errors and corresponding 95% confidence intervals (CIs) are presented in Table 3. We observe that the estimated regression parameters from both approaches, although similar in terms of direction, are very different in magnitude. The estimates for the intercept β_0 (and the corresponding cure probabilities) from the two approaches are also quite different; specifically, the estimated cure probabilities for zero covariates from the GEE and CWGEE are $\exp\{-\exp(-0.886)\} = 0.662$ and $\exp\{-\exp(-1.288)\} = 0.759$, respectively. This finding echoed with the simulation results, where the intercept is severely underestimated under the GEE approach. This can be explained by the informative clustering nature of the data. The GEE approach suggests that smoking, HbA1c, gender and jaw variables are significant factors associated with the risk of developing PD. Contrarily, the CWGEE approach suggests that only HbA1c and gender variables are significant factors. Under the PTC model, the constant hazard ratio assumption is preserved. The hazard ratio for gender (female against male) is $\exp(-0.760) = 0.468$, and the hazard ratio for HbA1c (uncontrolled against controlled) is $\exp(0.409) = 1.505$; both ratios are different from 1 at 5% level of significance. This suggests that females or subjects with controlled HbA1c are less susceptible to PD and tend to have longer time to disease onset, if susceptible.

We use a cross-validation procedure, delineated in Appendix A2, to compare the predictive performance of the CWGEE and GEE approaches. In the analysis, the averaged log-likelihood values in the testing sets are −46.8 and −47.3 for CWGEE and GEE, respectively. Also, CWGEE yields larger log-likelihood in 73% of the replicates. The results suggest that CWGEE produces more accurate outcome prediction by accommodating ICS.

6 | DISCUSSION

In this paper, we consider a class of semiparametric transformation cure models as a generalization of the PTC model for marginal analysis of clustered current status data. Under ICS, the traditional GEE yields biased estimation, even if the marginal survival

model is correctly specified. As a remedy for ICS, we propose to use a CWGEE approach that weights a cluster by the inverse of the cluster size. Nonetheless, GEE and CWGEE are both valid, when the cluster size is non-informative. [31] We consider a sieve-MLE approach, and approximate the nonparametric unspecified distribution function F by a BP. Constraints on F such as monotonicity and $F(\tau) = 1$ can be imposed easily by a reparameterization of the BP coefficients. The proposed estimators are shown to be consistent, and asymptotically normal.

We illustrate our proposed method on a dataset recording current status time to event of PD incidence, where initial data exploration reveals the presence of ICS paradigm. Given that the cluster sizes in the dataset are potentially large (many subjects with > 20 teeth), the estimation methods by maximizing the joint likelihood in the frailty models can be computationally intensive. On the contrary, the marginal approach proposed here provides a computationally efficient method for parameter estimation. It is speculated that informative clustering is present, as the prevalence of the disease decreases with the number of teeth. This is plausibly why some parameter estimates based on GEE differ substantially from those based on CWGEE.

In this paper, we treat the transformation parameter ρ as prespecified and select it based on an information criterion. An alternative method for the selection/estimation of ρ is to regard it as an unknown parameter and estimate it along with other parameters using the maximum (pseudo) likelihood method. Nevertheless, as remarked by Zeng et al. [17] under a similar cure transformation model, the transformation parameter cannot be reliably estimated under sample sizes smaller than 1500. Similar situations have been encountered in our set-up, and the estimation of ρ is not numerically stable when n is moderately small as the pseudo-likelihood function of ρ would be almost flat.

There are a number of future directions to consider, stemming from our current work. First, we currently allow the cluster size and the response to be associated but do not explicitly model the association structure. Alternatively, it is also worthy to consider a joint modeling approach that regresses both the cluster size and survival times on the same or different sets of covariates. In such a model, the association between the cluster size and current-status survival outcomes can be modelled via a shared frailty term in both regression models. One advantage of a joint-modeling approach is that, under a correctly-specified model and regularity conditions, the regression parameters can be estimated with optimal statistical efficiency. Second, we may develop a statistical test to detect the presence of ICS within current-status (or more general interval-censored) scenarios with a cured proportion. Third, the present work assumes that the inspection time Y is non-informative, which may be invalid in practice. One can consider extending the work to accommodate dependent censoring for clustered time-to-event data, using a copula or a frailty approach to model the association between T and Y. [44,45]

ACKNOWLEDGMENTS

DATA AVAILABILITY STATEMENT

The synthetic data used to support our findings in Section 4 is available from the corresponding author upon reasonable request. The application data are not publicly available due to privacy, or ethical restrictions.

References

1. Fernandes JK, Wiegand RE, Salinas CF, et al. Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina. *Journal of Periodontology* 2009; 80(7): 1062–1068.

2. Darby ML, Walsh M. *Dental Hygiene: Theory and Practice*. Elsevier health sciences. 3 ed. 2010.

3. Huang J, Wellner JA. Asymptotic normality of the NPMLE of linear functionals for interval censored data, Case 1. *Statistica Neerlandica* 1995; 49(2): 153–163.

4. Klein JP, Van Houwelingen HC, Ibrahim JG, Scheike TH. *Handbook of Survival Analysis*. CRC Press . 2016.

5. Cong XJ, Yin G, Shen Y. Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics* 2007; 63(3): 663–672.

6. Peng Y, Dear KBG. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; 56(1): 237–243.

7. Lam KF, Fong DYT, Tang OY. Estimating the proportion of cured patients in a censored sample. *Statistics in Medicine* 2005; 24(12): 1865–1879.

8. Chen MH, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999; 94(447): 909–919.

9. Ibrahim JG, Chen MH, Sinha D. Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* 2001; 57(2): 383–388.

10. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952; 47(259): 501–515.

11. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982: 1041–1046.

12. Kuk AYC, Chen CH. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 1992; 79(3): 531–541.

13. Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; 56(1): 227–236.

14. Lam KF, Xue H. A semiparametric regression cure model with current status data. *Biometrika* 2005; 92(3): 573–586.

15. Tsodikov AD, Yakovlev AY, Asselain B. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. 1. World Scientific . 1996.

16. Tsodikov A. A proportional hazards model taking account of long-term survivors. *Biometrics* 1998: 1508–1516.

17. Zeng D, Yin G, Ibrahim JG. Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association* 2006; 101(474): 670–684.

18. Zhou Q, Hu T, Sun J. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association* 2017; 112(518): 664–672.

19. Sun T, Ding Y. Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics* 2019. doi: 10.1093/biostatistics/kxz032

20. Lee CY, Wong KY, Lam KF, Xu J. Analysis of clustered interval-censored data using a class of semiparametric partly linear frailty transformation models. *Biometrics* 2020. doi: 10.1111/biom.13399

21. Lam KF, Wong KY. Semiparametric analysis of clustered interval-censored survival data with a cure fraction. *Computational Statistics & Data Analysis* 2014; 79: 165–174.

22. Kor CT, Cheng KF, Chen YH. A method for analyzing clustered interval-censored data based on Cox's model. *Statistics in Medicine* 2013; 32(5): 822–832.

23. Niu Y, Peng Y. Marginal regression analysis of clustered failure time data with a cure fraction. *Journal of Multivariate Analysis* 2014; 123: 129–142.

24. Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics* 2003; 59(1): 36–42.

25. Williamson JM, Kim HY, Manatunga A, Addiss DG. Modeling survival data with informative cluster size. *Statistics in Medicine* 2008; 27(4): 543–555.

26. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika* 2001; 88(4): 1121–1134.

27. Zhang X, Sun J. Regression analysis of clustered interval-censored failure time data with informative cluster size. *Computational Statistics & Data Analysis* 2010; 54(7): 1817–1823.

28. Zhang X, Sun J. Semiparametric regression analysis of clustered interval-censored failure time data with informative cluster size. *The International Journal of Biostatistics* 2013; 9(2): 205–214.

29. Zhao H, Ma C, Li J, Sun J. Regression analysis of clustered interval-censored failure time data with linear transformation models in the presence of informative cluster size. *Journal of Nonparametric Statistics* 2018; 30(3): 703–715.

30. Farouki RT. The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design* 2012; 29(6): 379–419.

31. Williamson JM, Kim HY, Manatunga A, Addiss DG. Modeling survival data with informative cluster size. *Statistics in Medicine* 2008; 27(4): 543–555.

32. Rossini AJ, Tsiatis AA. A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* 1996; 91(434): 713–721.

33. Zhang Y, Hua L, Huang J. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics* 2010; 37(2): 338–354.

34. Eilers PHC, Marx BD. Flexible smoothing with *B*-splines and penalties. *Statistical Science* 1996; 11(2): 89–121.

35. Lam KF, Wong KY, Zhou F. A semiparametric cure model for interval-censored data. *Biometrical Journal* 2013; 55(5): 771–788.

36. Lorentz GG. *Bernstein Polynomials*. NewYork: Chelsea Publishing Co . 1986.

37. Dai YH. Convergence properties of the BFGS algoritm. *SIAM Journal on Optimization* 2002; 13(3): 693–701.

38. Royall RM. Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review/Revue Internationale de Statistique* 1986: 221–226.

39. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; 73(2): 387–396.

40. Emura T, Chen YH. *Analysis of Survival Data with Dependent Censoring: Copula-based Approaches*. Springer . 2018.

41. Johnson Spruill I, Hammond P, Davis B, McGee Z, Louden D. Health of Gullah Families in South Carolina With Type 2 Diabetes. *The Diabetes Educator* 2009; 35(1): 117–123.

42. Chaitanya P, Reddy JS, Suhasini K, Chandrika IH, Praveen D. Time and eruption sequence of permanent teeth in Hyderabad children: A descriptive cross-sectional study. *International Journal of Clinical Pediatric Dentistry* 2018; 11(4): 330–337.

43. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media . 2003.

44. Ma L, Hu T, Sun J. Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* 2015; 102(3): 731–738.

45. Liu Y, Hu T, Sun J. Regression analysis of current status data in the presence of a cured subgroup and dependent censoring. *Lifetime Data Analysis* 2017; 23(4): 626–650.

46. Wong WH, Shen X. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* 1995; 23(2): 339–362.

APPENDIX

A1 Proof of Estimation Consistency. By Theorem 1.6.2 of Lorentz,[36] there exists a Bernstein polynomial $\widetilde{F}_n$ such that $\|\widetilde{F}_n - F^*\|_\infty = O(n^{-rv/2})$. Let $\mathbb{P}_n$ denote the empirical measure. By the definition of $\widehat{\beta}_n$ and $\widehat{F}_n$,

$$\mathbb{P}_n \ell(\widehat{\beta}_n, \widehat{F}_n) \geq \mathbb{P}_n \ell(\beta^*, \widetilde{F}_n).$$

Therefore,

$$\left|(\mathbb{P}_n - \mathbb{P})\ell(\widehat{\beta}_n, \widehat{F}_n)\right| + \left|(\mathbb{P}_n - \mathbb{P})\ell(\beta^*, F^*)\right| + \left|\mathbb{P}_n\{\ell(\beta^*, \widetilde{F}_n) - \ell(\beta^*, F^*)\}\right| \geq \mathbb{P}\ell(\beta^*, F^*) - \mathbb{P}\ell(\widehat{\beta}_n, \widehat{F}_n). \tag{B1}$$

Following the proof of Lemma 2 of Zhou et al.,[18] we can show that the first two terms on the left-hand side of (B1) converges to 0 almost surely. By the mean-value theorem, the third term on the left-hand side of (B1) converges uniformly to 0. Note that for any fixed function f,

$$\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N} f(Y_j, \Delta_j, X_j)\right\} = \mathrm{E}\left[\frac{1}{N}\sum_{j=1}^{N} \mathrm{E}\{f(Y_j, \Delta_j, X_j) \mid N\}\right] = \mathrm{E}\{f(Y_1, \Delta_1, X_1)\},$$

where the second equality follows from condition (C1). Therefore, the right-hand side of (B1) is equal to $\mathbb{P}\ell^{(1)}(\beta^*, F^*) - \mathbb{P}\ell^{(1)}(\widehat{\beta}_n, \widehat{F}_n)$. This term is the Kullback–Leibler divergence for the observed data of the first subject of a cluster, so that by Wong and Shen[46] (p. 346), it is bounded below by the Hellinger distance:

$$\left\|Q(\Delta_1; \widehat{\xi}_n)^{1/2} - Q(\Delta_1; \xi^*)^{1/2}\right\|_{L_2(\mathbb{P})}^2 = \left\|\frac{(-1)^{\Delta_1} G'(\xi)}{2Q(\Delta_1; \xi)^{1/2}}\bigg|_{\xi=\widetilde{\xi}_n}\left\{\widehat{F}_n(Y_1)e^{X_1^{\mathrm{T}}\widehat{\beta}_n} - F^*(Y_1)e^{X_1^{\mathrm{T}}\beta^*}\right\}\right\|_{L_2(\mathbb{P})}^2,$$

where $\widehat{\xi}_n = \widehat{F}_n(Y_1)e^{X_1^{\mathrm{T}}\widehat{\beta}_n}$, $\xi^* = F^*(Y_1)e^{X_1^{\mathrm{T}}\beta^*}$, $\widetilde{\xi}_n$ is some value between $\widehat{\xi}_n$ and ξ^*, and

$$Q(\Delta; \xi) = \{1 - G(\xi)\}^\Delta G(\xi)^{1-\Delta}.$$

By condition (C2), $G'(\xi)$ is negative and uniformly bounded away from 0, and $Q(\Delta_1; \xi)$ is clearly uniformly bounded above by 1. Therefore, the right-hand side above is up to a scaling factor bounded below by

$$\left\|\widehat{F}_n(Y_1)e^{X_1^{\mathrm{T}}\widehat{\beta}_n} - F^*(Y_1)e^{X_1^{\mathrm{T}}\beta^*}\right\|_{L_2(\mathbb{P})}^2 \geq \mathrm{E}\left\{\left\|e^{X_1^{\mathrm{T}}\widehat{\beta}_n} - e^{X_1^{\mathrm{T}}\beta^*}\right\|^2 \bigg| Y_1 = \tau\right\}P(Y_1 = \tau).$$

By conditions (C3) and (C4) and the mean-value theorem, the right-hand side above is up to a scaling factor bounded below by $\|\widehat{\beta}_n - \beta^*\|^2$. Therefore, (B1) implies that $\|\widehat{\beta}_n - \beta^*\|^2 \leq o_p(1)$, where the right-hand side goes to 0 almost surely. Similarly, we conclude that $\|\widehat{F}_n(Y_1) - F^*(Y_1)\|_{L_2(\mathbb{P})}^2 \to_{\mathrm{a.s.}} 0$, and thus the desired result follows from condition (C3).

A2 Cross-Validation Procedure. We use the following cross-validation procedure to compare the predictive performance of CWGEE and GEE on the GAAD data:

(i) Randomly split the GAAD data ($n = 288$) into a training set and a testing set, with clusters as sampling units and a ratio of sample size of $2:1$ (i.e., 192 and 96 clusters for the training and testing sets, respectively).

(ii) Perform CWGEE and GEE on the training set with $(m, \rho) = (1, 1)$, and denote the resulting sieve-MLEs as $\widehat{\theta}_{\mathrm{CWGEE}}$ and $\widehat{\theta}_{\mathrm{GEE}}$, respectively.

(iii) Generate $B = 1000$ random samples from the testing set, where each random sample consists of a randomly selected observation from each cluster. For $b = 1, \ldots, B$, let $(Y_i^{(b)}, \Delta_i^{(b)}, X_i^{(b)})_{i=1,\ldots,n^{(b)}}$ be the bth random sample with $n^{(b)} = 96$ and

$$\ell^{(b)}(\theta) \equiv \sum_{i=1}^{n^{(b)}} \Delta_i^{(b)} \log\{1 - S(Y_i^{(b)}|\theta, X_i^{(b)})\} + (1 - \Delta_i^{(b)}) \log S(Y_i^{(b)}|\theta, X_i^{(b)})$$

be the log-likelihood value evaluated at θ.

(iv) Compute the mean log-likelihood values $B^{-1}\sum_{b=1}^{B} \ell^{(b)}(\widehat{\theta}_{\mathrm{CWGEE}})$ and $B^{-1}\sum_{b=1}^{B} \ell^{(b)}(\widehat{\theta}_{\mathrm{GEE}})$.

(v) Repeat (i)–(iv) 100 times, and report the mean log-likelihood values for both methods averaged over the 100 replicates.