BIOMETRICS 79, 1–24 DOI: 10.1111/biom.13795

November 2022

A semiparametric joint model for cluster size and subunit-specific interval-censored outcomes

Chun Yin Lee[1], Kin Yau Wong[1,*], K. F. Lam[2,3], and Dipankar Bandyopadhyay[4]

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

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

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

[4]Department of Biostatistics, Virginia Commonwealth University, Virginia, U.S.A.

[*]*email*:kin-yau.wong@polyu.edu.hk

SUMMARY: Clustered data frequently arise in biomedical studies, where observations, or subunits, measured within a cluster are associated. The cluster size is said to be informative, if the outcome variable is associated with the number of subunits in a cluster. In most existing work, the informative cluster size issue is handled by marginal approaches based on within-cluster resampling, or cluster-weighted generalized estimating equations. Although these approaches yield consistent estimation of the marginal models, they do not allow estimation of within-cluster associations, and are generally inefficient. In this paper, we propose a semiparametric joint model for clustered interval-censored event time data with informative cluster size. We use a random effect to account for the association among event times of the same cluster as well as the association between event times and the cluster size. For estimation, we propose a sieve maximum likelihood approach and devise a computationally-efficient expectation-maximization algorithm for implementation. The estimators are shown to be strongly consistent, with the Euclidean components being asymptotically normal and achieving semiparametric efficiency. Extensive simulation studies are conducted to evaluate the finite-sample performance, efficiency and robustness of the proposed method. We also illustrate our method via application to a motivating periodontal disease dataset.

KEY WORDS: dental study, EM algorithm, informative cluster size, random effect model, sieve estimation.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Interval-censored data commonly arise in epidemiological and clinical studies, where the event time of interest is not directly observed but is only known to fall within an interval. For example, in oral health studies, subjects may be examined periodically, and the onset of a dental disease is only known to occur between two examination times. A special case of interval-censored data is current-status data (also known as case-1 interval-censored data), where subjects are examined once, and the event of interest is only known to have occurred before or after the examination (Huang and Wellner, 1995; Klein et al., 2016; Zhang and Sun, 2010b).

This work is motivated by a cross-sectional periodontal disease (PD) study (Fernandes et al., 2009) conducted on Gullah-speaking African-American Type-2 Diabetics (henceforth, GAAD), where the primary goal is to investigate the association between potential risk factors and the time to onset of PD, defined by the disease status (diseased/non-diseased) of a tooth. A number of statistical challenges were observed while analyzing this dataset. First, the exact time to onset of PD of each tooth is unknown but is subject to (case-1) interval censoring. Second, the event times of interest within a subject are clustered, because the teeth share a common oral condition. Third, and most importantly, subjects with poor oral health tend to have a higher PD risk and associated risk of tooth loss (Spiekerman and Lin, 1998; Williamson et al., 2003; Cong et al., 2007; Neuhaus and McCulloch, 2011) such that they tend to have fewer teeth at observation, leading to the informative cluster size (ICS) scenario. For example, the seemingly negative association observed between the percentage of diseased teeth and the number of (available) teeth in the motivating GAAD data (see Table 1) provides preliminary empirical evidence of the ICS issue.

[Table 1 about here.]

The generalized estimating equations approach (Liang and Zeger, 1986) is commonly employed in analyzing clustered survival data (Kor et al., 2013; Niu and Peng, 2014). Therein, the unknown parameters under a marginal representation are estimated with a prespecified working correlation

matrix. An advantage of the generalized estimating equations method is that the parameter estimators are consistent when the marginal model is correctly specified, even though the dependence structure of the event times may be misspecified. However, the generalized estimating equations approach is biased in the presence of ICS. The within-cluster resampling (Hoffman et al., 2001) and cluster-weighted generalized estimating equations (Williamson et al., 2003) methods have been proposed to provide unbiased estimators in the ICS scenario. For right-censored data, Cong et al. (2007) and Williamson et al. (2008) showed empirically that cluster-weighted generalized estimating equations and within-cluster resampling are equivalent and both estimators of the regression parameters are unbiased under the Cox model. For interval-censored data, the finite-sample performance of the two methods are found to be similar and satisfactory based on the Cox model with a parametric or nonparametric cumulative baseline hazard function (Zhang and Sun, 2010a, 2013). Zhao et al. (2018) proposed to use within-cluster resampling to estimate the parameters based on a linear transformation model, while Lam et al. (2021) proposed a class of semiparametric transformation cure survival models for clustered current-status data based on the cluster-weighted generalized estimating equations approach. In the marginal approach, the association among the cluster size and the response variables within a cluster is not characterized. Also, estimating equations approaches based on the marginal model are generally statistically less efficient (Zhou et al., 2017) than maximum likelihood estimation under the full model, unless in the absence of within-cluster association and ICS.

The joint modeling approach is an appealing alternative to cluster-weighted generalized estimating equations or within-cluster resampling that overcomes the above drawbacks. In this approach, two separate regression models are considered for the response and the cluster size, with the two models linked by a common random effect quantifying the strength and direction (positive or negative) of the association between the cluster size and the response. Although the joint modeling approach has received increasing recent attention, contemporary work mainly focuses on the gen-

eralized linear mixed-effects models for continuous response variables and misspecification issues of the random effect model and/or cluster size model. Under a Bayesian framework, Dunson et al. (2003) proposed a continuation ratio probit model for the cluster size and a multivariate mixed-effects model for the response variables, where the association between cluster size and the subunit-level response variables was modeled through a multivariate normal random variable. In that vein, Chen et al. (2011) studied, both theoretically and empirically, the robustness to the distributional and functional forms of the cluster size model based on the generalized linear mixed-effects models. They reported that the estimate for the subunit-level parameter can be sensitive to the distributional assumptions of the cluster size model, but is insensitive to the functional form of the shared random effect. Furthermore, Neuhaus and McCulloch (2011) studied maximum likelihood methods and the robustness of estimation under generalized linear mixed-effects models, where they showed that the estimator can be biased if the ICS issue is ignored. Zhang et al. (2017) studied the robustness to misspecification of cluster size models (within generalized linear mixed-effects models) in terms of type I/II errors for tests of regression coefficients. However, research work for joint modeling of failure times and cluster size is scanty. Kim (2010) considered a joint modeling approach for interval-censored data, where the cluster size is modeled by the cumulative logistic regression and the failure times are modeled by the Cox model. However, theoretical properties of the methods were not studied.

In this paper, we develop a joint modeling approach with a semiparametric regression model for the event times and a binomial model for the cluster size, where the latter model is specifically motivated by oral health studies. We devise a sieve maximum likelihood (ML) estimation method based on I-splines and develop a computationally efficient expectation-maximization (EM) algorithm using a data augmentation technique based on latent Poisson variables. We utilize modern empirical process theory to rigorously establish the asymptotic properties of the proposed estimator under a general interval-censoring framework.

The rest of this paper is structured as follows. Section 2 outlines the model, likelihood, and the estimation method. Section 3 explores the theoretical properties, including consistency of the proposed estimator, and the asymptotic normality and efficiency of the estimator of the Euclidean parameters. A simulation study is presented in Section 4 to assess the finite-sample performance of the estimator. We illustrate our method via application to the GAAD data in Section 5. Finally, we present some concluding remarks in Section 6. Technical details are relegated to the Supporting Information.

2. Methods

2.1 *Model, data, and sieve estimation*

In oral health studies, a cluster refers to a subject, and the cluster size N is the number of teeth of a subject at baseline or by the time of observation. Let $\boldsymbol{Z}$ be a vector of cluster-specific time-invariant covariates such as gender, age and oral habits, and b be a cluster-specific random effect that may represent an unobservable index of dental hygiene. We assume

$$N|(\boldsymbol{Z}, b) \sim \text{Binomial}\left(m, \frac{e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}+\kappa b}}{1+e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}+\kappa b}}\right),$$

where m is the fixed maximum cluster size, and $\boldsymbol{\alpha}$ and κ are regression coefficients, respectively for $\boldsymbol{Z}$ and b. In practice, m is typically known and set to be the total number of permanent teeth (i.e., $m = 28$), excluding the 4 third-molars (which are typically not evaluated). To facilitate the formulation of the ICS mechanism, we first define the event times and covariates for all m units, although only N of them could be observed. Let $T_1, \ldots, T_m$ be the individual failure times pertaining to the cluster and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ be the corresponding vectors of cluster/subject-specific time-invariant covariates. We assume that conditional on b and the covariates, $(N, T_1, \ldots, T_m)$ are mutually independent, with the hazard function of T_j $(j = 1, \ldots, m)$ being

$$\lambda(t \mid \boldsymbol{X}_j, b) = \lambda(t) e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j+b},$$

where $\lambda(\cdot)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$ is a vector of regression coefficients. This model assumes that there is a random effect b that affects both the failure times and the number of units available to the study. We assume that $b \sim \mathrm{N}(0, \sigma^2)$.

Suppose that only the information of $T_1, \ldots, T_N$ is available. Also, the N failure times are subject to interval censoring, such that the events are only known to occur within a time interval. In particular, let M_j denote a random positive integer that represents the number of inspections for the jth unit of the cluster and $\boldsymbol{U}_j \equiv (U_{j0}, \ldots, U_{j,M_j+1})$ denotes the set of inspection time points, with $0 = U_{j0} < U_{j1} < \cdots < U_{jM_j} < U_{j,M_j+1} = \infty$ for $j = 1, \ldots, N$. Define $\boldsymbol{\Delta}_j = (\Delta_{j0}, \ldots, \Delta_{jM_j})$, where $\Delta_{jk} = I(U_{jk} < T_j \leqslant U_{j,k+1})$ for $k = 0, \ldots, M_j$, and $I(\cdot)$ is the indicator function. Clearly, $M_j = 1$ and $M_j = 2$ correspond to case-1 and case-2 interval censoring, respectively. Assume that censoring is noninformative, i.e., $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m)$ is independent of $(b, T_1, \ldots, T_m)$ given $(N, \boldsymbol{Z}, \boldsymbol{X})$, where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$. For a random sample of n clusters, the observed data consist of $(N_i, M_{ij}, \boldsymbol{Z}_i, \boldsymbol{U}_{ij}, \boldsymbol{\Delta}_{ij}, \boldsymbol{X}_{ij})_{j=1,\ldots,N_i; i=1,\ldots,n}$. Let $\boldsymbol{\zeta} \equiv (\boldsymbol{\alpha}, \kappa, \boldsymbol{\beta}, \sigma)$ be the collection of all Euclidean parameters. The observed likelihood is proportional to

$$L_n(\boldsymbol{\zeta}, \Lambda) = \prod_{i=1}^{n} \int \prod_{j=1}^{N_i} \prod_{k=0}^{M_{ij}} \left\{ e^{-\Lambda(U_{ijk}) e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij} + b}} - e^{-\Lambda(U_{ij,k+1}) e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij} + b}} \right\}^{\Delta_{ijk}} \times f_N(N_i; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{Z}_i + \kappa b) f_b(b; \sigma) \, \mathrm{d}b, \tag{1}$$

where $\Lambda(t) = \int_0^t \lambda(s) \, \mathrm{d}s$, f_N denotes the probability mass function of N, and f_b denotes the probability density function of b.

We propose a sieve ML method for the estimation of $(\boldsymbol{\zeta}, \Lambda)$. Let $[a_1, a_2]$ be an interval that contains all nonzero and finite observation times. Define a set of grid points $(\psi_1, \ldots, \psi_d)$ such that $a_1 = \psi_0 < \psi_1 < \cdots < \psi_d < \psi_{d+1} = a_2$, and let $B_1, \ldots, B_K$ be the I-spline (integral of the M-spline) basis functions (Ramsay, 1988) of order $(K - d + 1)$ on the grid. We approximate $\Lambda(\cdot)$ by $\sum_{l=1}^{K} \gamma_l B_l(\cdot)$ for some nonnegative coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^{\mathrm{T}}$, so that the approximation is nondecreasing. Let $(\widehat{\boldsymbol{\zeta}}_n, \widehat{\boldsymbol{\gamma}}_n) = \arg\max_{\boldsymbol{\zeta}, \boldsymbol{\gamma}} L_n(\boldsymbol{\zeta}, \sum_{l=1}^{K} \gamma_l B_l)$ and $\widehat{\Lambda}_n(\cdot) = \sum_{l=1}^{K} \widehat{\boldsymbol{\gamma}}_l B_l(\cdot)$. We let $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\zeta}}_n, \widehat{\Lambda}_n)$, the sieve ML estimator of $\boldsymbol{\theta} \equiv (\boldsymbol{\zeta}, \Lambda)$.

In the model, we assume that $N = 0$ is possible and observable. In practice, however, subjects with no observation on failure times may be excluded, in which case a truncated binomial model should be adopted. Nevertheless, when the binomial probability is sufficiently large, using the likelihood function (1) does not make much difference empirically but can substantially simplify the computation. In cases where $N = 0$ is more common, whether conditioning on $N > 0$ or not in the likelihood may result in a substantial difference, and we should consider methods for informative emptiness (McGee et al., 2020).

2.2 *Computation of the sieve maximum likelihood estimator*

To facilitate presentation of the computation algorithm, we introduce the following notations. For the (i, j)th observation, let $\Delta_{Lij} \equiv \Delta_{ij1}$ be the indicator of whether the observation is left-censored, $\Delta_{Rij} \equiv \Delta_{ijM_{ij}}$ be the indicator of whether the observation is right-censored, and $\Delta_{Iij} = 1 - \Delta_{Lij} - \Delta_{Rij}$. Let $[L_{ij}, R_{ij}) \equiv [U_{ijk}, U_{ij,k+1})$ be the smallest interval that brackets T_{ij}, where k is such that $\Delta_{ijk} = 1$; note that $L_{ij} = 0$ and $R_{ij} = \infty$ for left- and right-censored observations, respectively. The likelihood is then proportional to

$$
\begin{aligned}
L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int \prod_{j=1}^{N_i} \Bigg[& \left\{ 1 - e^{-\sum_{l=1}^{K} \gamma_l B_l(R_{ij}) e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij} + b}} \right\}^{\Delta_{Lij}} \left\{ e^{-\sum_{l=1}^{K} \gamma_l B_l(L_{ij}) e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij} + b}} \right. \\
& \left. - e^{-\sum_{l=1}^{K} \gamma_l B_l(R_{ij}) e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij} + b}} \right\}^{\Delta_{Iij}} e^{-\sum_{l=1}^{K} \gamma_l B_l(L_{ij}) e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij} + b} \Delta_{Rij}} \Bigg] \\
& \times \frac{e^{N_i(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{Z}_i + \kappa b)}}{(1 + e^{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{Z}_i + \kappa b})^m} \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2} b^2} \, \mathrm{d}b.
\end{aligned} \tag{2}
$$

Direct maximization of the above likelihood is difficult, especially when N_i's are large. We propose an EM algorithm involving data augmentation (Wang et al., 2016) for parameter estimation. Similar methods for correlated interval-censored data have been developed by, for example, Zeng et al. (2017), Gamage et al. (2018), and Sun et al. (2021). For $k = 1, \ldots, K$, $j = 1, \ldots, N_i$ and $i = 1, \ldots, n$, let Y_{ijk} and W_{ijk} be Poisson variables that are independent given b_i with conditional expectations $\lambda_{ijk} e^{b_i}$ and $\xi_{ijk} e^{b_i}$, respectively, where $\lambda_{ijk} = \gamma_k e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij}} \{\Delta_{Lij} B_k(R_{ij}) + \Delta_{Iij} B_k(L_{ij})\}$ and $\xi_{ijk} = \gamma_k e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_{ij}} \left[\Delta_{Iij} \{B_k(R_{ij}) - B_k(L_{ij})\} + \Delta_{Rij} B_k(L_{ij})\right]$. The likelihood under the complete

data $(Y_{ijk}, W_{ijk}, b_i, N_i)_{k=1,\ldots,K, j=1,\ldots,N_i, i=1,\ldots,n}$ is proportional to

$$\prod_{i=1}^{n}\left\{\prod_{j=1}^{N_i}\prod_{k=1}^{K} e^{-\lambda_{ijk}e^{b_i}}(\lambda_{ijk}e^{b_i})^{Y_{ijk}} e^{-\xi_{ijk}e^{b_i}}(\xi_{ijk}e^{b_i})^{W_{ijk}}\right\}\frac{e^{N_i(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i+\kappa b_i)}}{(1+e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i+\kappa b_i})^m}\frac{1}{\sigma}e^{-\frac{1}{2\sigma^2}b_i^2}.$$

Let $Y_{ij}=\sum_{k=1}^{K}Y_{ijk}$ and $W_{ij}=\sum_{k=1}^{K}W_{ijk}$. Suppose that we do not observe the complete data. Instead, for $\Delta_{Lij}=1$, we observe the event $\{Y_{ij}>0, W_{ij}=0\}$; for $\Delta_{Iij}=1$, we observe the event $\{Y_{ij}=0, W_{ij}>0\}$; and for $\Delta_{Rij}=1$, we observe the event $\{Y_{ij}=W_{ij}=0\}$. Also, the random effect b_i is not observed. It is easy to verify that the resulting observed-data likelihood is (2). Therefore, the parameter estimates can be computed using an EM algorithm with the observed data consisting of N_i and corresponding events about Y_{ij} and W_{ij}. The complete-data log-likelihood is

$$\begin{aligned}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{N_i}\sum_{k=1}^{K}\Big\{&-\lambda_{ijk}e^{b_i}+Y_{ijk}(\log\lambda_{ijk}+b_i)-\xi_{ijk}e^{b_i}+W_{ijk}(\log\xi_{ijk}+b_i)\Big\} \\ &+N_i(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i+\kappa b_i)-m\log(1+e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}_i+\kappa b_i})-\frac{1}{2}\log\sigma^2-\frac{1}{2\sigma^2}b_i^2\Bigg].\end{aligned} \tag{3}$$

In the E-step, we evaluate the conditional expectation of (3) given the observed data. We approximate expectations of functions of b_i using adaptive Gauss–Hermite quadrature (Liu and Pierce, 1994). The conditional expectations of Y_{ijk} and W_{ijk} given the observed data can be evaluated using the law of iterated expectation by first considering the conditional expectations given the observed data and b_i (the latter expectations have closed forms). In the M-step, we update γ_k's and σ^2 with closed-form solutions and update $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \kappa)$ using Newton's method. We iterate between the E- and M-steps until convergence. Details of the algorithm are presented in Section S1 of the Supporting Information.

The standard EM algorithm may converge slowly especially in large cluster size settings. To achieve faster convergence, we adopt the accelerated EM algorithm proposed by Varadhan and Roland (2008). Let $\boldsymbol{\vartheta}=(\boldsymbol{\zeta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{s}(\boldsymbol{\vartheta})$ be the updated parameter vector after one EM step, with the initial parameter set to be $\boldsymbol{\vartheta}$. Given a vector of current estimates $\boldsymbol{\vartheta}^{(k)}$, a step of the accelerated EM algorithm comprises the following four steps:

Step 1 Compute $\boldsymbol{\vartheta}_1 = \boldsymbol{s}(\boldsymbol{\vartheta}^{(k)})$.

Step 2 Compute $\boldsymbol{\vartheta}_2 = \boldsymbol{s}(\boldsymbol{\vartheta}_1)$.

Step 3 Compute $\boldsymbol{r} = \boldsymbol{\vartheta}_1 - \boldsymbol{\vartheta}^{(k)}$, $\boldsymbol{v} = \boldsymbol{\vartheta}_2 - \boldsymbol{\vartheta}_1 - \boldsymbol{r}$, and $a = -\|\boldsymbol{r}\|_2/\|\boldsymbol{v}\|_2$.

Step 4 Update the vector of parameter estimates by $\boldsymbol{\vartheta}^{(k+1)} = \boldsymbol{s}(\boldsymbol{\vartheta}^{(k)} - 2a\boldsymbol{r} + a^2\boldsymbol{v})$.

We terminate the accelerated EM algorithm when the maximum absolute difference between two consecutive estimates for $\boldsymbol{\vartheta}$ is less than a small threshold.

2.3 *Standard error estimation*

To estimate the standard error of $\widehat{\boldsymbol{\zeta}}_n$, we treat the model as fully parametric, with parameter $\boldsymbol{\vartheta}$. We then use Louis's (1982) formula to compute the observed information matrix $\boldsymbol{I}(\boldsymbol{\vartheta})$, with

$$\boldsymbol{I}(\boldsymbol{\vartheta}) = -\frac{\partial^2 Q(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\mathrm{T}}} - \sum_{i=1}^{n} E\left\{ \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}^{\otimes 2} \,\middle|\, \mathcal{O}_i \right\} + \sum_{i=1}^{n} E\left\{ \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}} \,\middle|\, \mathcal{O}_i \right\} E\left\{ \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}} \,\middle|\, \mathcal{O}_i \right\}^{\mathrm{T}}, \quad (4)$$

where $\mathcal{O}_i$ is the observed data for the ith subject, $Q(\boldsymbol{\vartheta})$ is the conditional expected complete-data log-likelihood given in Section S1, $\ell_i(\boldsymbol{\vartheta})$ is the ith term in the summation in (3), and the expectations are evaluated at the sieve ML estimator. Similar to the E-step, the expectations on the right-hand side of (4) can be evaluated using the law of iterated expectations and numerical integration. Standard errors of components of $\widehat{\boldsymbol{\zeta}}_n$ can be estimated by the corresponding diagonal elements of the inverse of the observed information matrix.

3. Asymptotic properties of the sieve maximum likelihood estimator

We present the theoretical properties of the sieve ML estimator under a general (parametric) cluster size distribution. Let $f_N(\cdot; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z} + \kappa b, \boldsymbol{\varsigma})$ be the probability mass function of N given $(\boldsymbol{Z}, b)$ and $\mathcal{S}$ be the support of N, where $\boldsymbol{\varsigma}$ is a set of unknown parameters. With an abuse of notation, we may use $\boldsymbol{\varsigma}$ to denote all unknown parameters in the conditional distribution of N and write the probability mass function as $f_N(\cdot \mid b, \boldsymbol{\varsigma})$. For a generic cluster, let $S_j(t; b, \boldsymbol{\beta}, \Lambda) \equiv \exp\{-\Lambda(t)e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_j + b}\}$ be the conditional survival probability of the jth observation. Let $\boldsymbol{\zeta}_0$ and Λ_0 be the true values

of $\boldsymbol{\zeta}$ and Λ, respectively, and let p be the dimension of $\boldsymbol{\zeta}$. In this section and the Supporting Information, we use K_n to denote the number of spline functions to emphasize its dependence on n. In the theoretical development, we assumed that the sieve space is deterministic, with prespecified grid points $\psi_0, \ldots, \psi_{d+1}$ (for each n), although the grid points are often chosen data-adaptively in practice. This is to simplify the derivations, and similar assumptions are commonly made in the theoretical development for semiparametric modeling of interval-censored data (Zhou et al., 2017; Zeng et al., 2017). We consider the following conditions; some conditions involve a generic positive constant C.

(C1) The set $\mathcal{S}$ is finite, and the support of the conditional distribution of N given any value of $\boldsymbol{Z}$ is $\mathcal{S}$. Also,

$$\sup_{m' \in \mathcal{S}} \sup_{\boldsymbol{\varsigma}} \sum_{j=1}^{3} \left\| \frac{\boldsymbol{f}_N^{(j)}(m'; b, \boldsymbol{\varsigma})}{f_N(m'; b, \boldsymbol{\varsigma})} \right\| \leqslant e^{C_1 + C_2 b}$$

for some constants C_1 and C_2, where $\boldsymbol{f}_N^{(j)}(\cdot; b, \boldsymbol{\varsigma})$ denotes the jth derivative of $f_N(\cdot; b, \boldsymbol{\varsigma})$ with respect to $(b, \boldsymbol{\varsigma}^{\mathrm{T}})^{\mathrm{T}}$.

(C2) For $j = 1, \ldots, m$, the number of monitoring times M_j is positive with $E(M_j) < C$, where $m = \max(\mathcal{S})$. Also, there exists a minimum interval spacing C^{-1} such that for any given N and M_j, $P(U_{j,k+1} - U_{jk} > C^{-1}) = 1$ for $k = 1, \ldots, M_j$ and $j = 1, \ldots, N$. In addition, for $j = 1, \ldots, m$, the union of the support of $(U_{j1}, \ldots, U_{jM_j})$ is $[a_1, a_2]$ for some constants a_1 and a_2 such that $0 < a_1 < a_2 < \infty$.

(C3) The covariates are bounded, such that $P(\|\boldsymbol{X}_j\| + \|\boldsymbol{Z}\| < C) = 1$ for $j = 1, \ldots, m$. Also, if there exist a constant c_1 and vectors $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$ such that $c_1 + \boldsymbol{c}_2^{\mathrm{T}} \boldsymbol{X}_j = 0$ and $\boldsymbol{c}_3^{\mathrm{T}} \boldsymbol{Z} = 0$ with probability 1, then $c_1 = 0$, $\boldsymbol{c}_2 = \boldsymbol{0}$, and $\boldsymbol{c}_3 = \boldsymbol{0}$.

(C4) The parameter value $\boldsymbol{\zeta}_0$ belongs to the interior of a known compact set $\mathcal{A}_\zeta \subset \mathbb{R}^p$. The function Λ_0 is strictly increasing and continuously differentiable up to the rth order on $[0, a_2]$ for some $r \geqslant 1$.

(C5) The number of grid points $K_n = O(n^\nu)$ for some $\nu \in (0,1)$, and the distance between adjacent grid points is within $(C^{-1}n^{-\nu}, Cn^{-\nu})$.

(C6) If there exists a set of parameters $(\widetilde{\boldsymbol{\zeta}}, \widetilde{\Lambda})$ such that

$$\int \prod_{j=1}^{m'} S_j(t_j; b, \widetilde{\boldsymbol{\beta}}, \widetilde{\Lambda}) f_N(m' \mid b, \widetilde{\boldsymbol{\varsigma}}) f_b(b; \widetilde{\sigma}) \, \mathrm{d}b = \int \prod_{j=1}^{m'} S_j(t_j; b, \boldsymbol{\beta}_0, \Lambda_0) f_N(m' \mid b, \boldsymbol{\varsigma}_0) f_b(b; \sigma_0) \, \mathrm{d}b$$

with probability 1 for all $t_j \in [a_1, a_2]$ $(j = 1, \ldots, m')$ and $m' \in \mathcal{M}$, then $\widetilde{\boldsymbol{\zeta}} = \boldsymbol{\zeta}_0$ and $\widetilde{\Lambda}(t) = \Lambda_0(t)$ for $t \in [a_1, a_2]$.

(C7) At any value of $(\boldsymbol{Z}, \boldsymbol{X})$ in the support of the covariates, if there exist a vector $\boldsymbol{c}_1$, constants c_2, c_3, and c_4, and functions $(h_1, \ldots, h_m)$ such that

$$\int \left\{ \prod_{j=1}^{m'} S_j(t_j; b, \boldsymbol{\beta}_0, \Lambda_0) \right\} \left\{ e^b \sum_{j=1}^{m'} h_j(t_j) + \boldsymbol{c}_1^{\mathrm{T}} \frac{\partial \log f_N(m'; \boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{Z} + \kappa_0 b, \boldsymbol{\varsigma})}{\partial \boldsymbol{\varsigma}} \bigg|_{\boldsymbol{\varsigma} = \boldsymbol{\varsigma}_0} \right.$$
$$\left. + c_2(b^2 - \sigma_0^2) + \frac{\partial \log f_N(m'; \mu, \boldsymbol{\varsigma}_0)}{\partial \mu} \bigg|_{\mu = \boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{Z} + \kappa_0 b} (c_3 + c_4 b) \right\} f_N(m' \mid b, \boldsymbol{\varsigma}_0) f_b(b; \sigma_0) \, \mathrm{d}b = 0$$

for all $t_j \in [a_1, a_2]$ $(j = 1, \ldots, m')$ and $m' \in \mathcal{M}$, then $\boldsymbol{c}_1 = \boldsymbol{0}$, $c_2 = c_3 = c_4 = 0$, and $h_j(t) = 0$ for $t \in [a_1, a_2]$ and $j = 1, \ldots, m$.

(C8) Given $(N, M_1, \ldots, M_m, \boldsymbol{X}, \boldsymbol{Z})$, the conditional density of $(U_{jk}, U_{j',k'})$ has continuous second-order partial derivatives for $(j,k) \neq (j',k')$.

REMARK 1: Condition (C1) imposes mild regularity conditions on the cluster size distribution. Conditions (C2)–(C4) are typical regularity conditions in semiparametric survival models for interval-censored data. Condition (C5) pertains to the rate at which the number of spline functions increases to infinity. Condition (C6) is for model identifiability, and condition (C7) is for the invertibility of the information operator; they are analogous to Conditions 7 and 8 of Zeng et al. (2017). Condition (C8) guarantees that the least-favorable direction of Λ is sufficiently smooth.

Let $\mathcal{B}_\Lambda = \{\Lambda \in \ L_2[0, a_2] : \Lambda(0) = 0, \Lambda(a_2) < C, \Lambda \text{ is monotonic increasing}\}$ for some large enough positive constant C and $\Theta = \mathcal{A}_\zeta \times \mathcal{B}_\Lambda$. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, we define the distance function

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \left\{ \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|^2 + \int_{a_1}^{a_2} (\Lambda_1 - \Lambda_2)(t)^2 \, \mathrm{d}t \right\}^{1/2},$$

where $\boldsymbol{\zeta}_j$ and Λ_j are the Euclidean and functional components of $\boldsymbol{\theta}_j$ $(j = 1, 2)$, respectively. Denote $\boldsymbol{\theta}_0 \equiv (\boldsymbol{\zeta}_0, \Lambda_0)$ the true value of $\boldsymbol{\theta}$. We have the following three theorems, whose proofs are given in Section S3 of the Supporting Information.

THEOREM 1 (Strong consistency): *Under conditions (C1)–(C6),* $d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \to 0$ *almost surely.*

THEOREM 2 (Rate of convergence): *Under conditions (C1)–(C7),*

$$d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p\left\{\max\left(n^{-(1-\nu)/2}, n^{-r\nu}\right)\right\}.$$

THEOREM 3 (Asymptotic normality): *Assume that conditions (C1)–(C8) hold,* $r \geqslant 2$*, and* $1/8 < \nu < 1/2$*. We have*

$$n^{1/2}(\widehat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}_0) \to_d N(0, \widetilde{\boldsymbol{I}}^{-1}),$$

where $\widetilde{\boldsymbol{I}}$ *is the efficient information matrix for* $\boldsymbol{\zeta}$ *defined in Lemma 2 in Section S2 of the Supporting Information. Also, the asymptotic variance of* $\widehat{\boldsymbol{\zeta}}_n$ *attains the semiparametric efficiency bound.*

REMARK 2: Under Theorem 2, the optimal choice of ν is $\nu = 1/(1 + 2r)$, in which case $d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-r/(1+2r)})$. This is the optimal rate of convergence in nonparametric regression under the same smoothness assumptions. Theorem 3 implies that the Euclidean parameter estimators converge to the true values at the (optimal) rate of $n^{1/2}$.

4. Simulation studies

In this section, we conduct simulation studies to evaluate the finite-sample performance, efficiency, and robustness of the proposed sieve ML estimators. We set the maximum cluster size $m = 10$, the total number of clusters $n = 50$ or 200, the true value of the cluster-specific frailty parameter $\sigma = 1$, and the cumulative baseline hazard function $\Lambda(t) = t^2/4$. Any clusters with size $N_i = 0$ were discarded. We set the inter-arrival times $(U_j - U_{j-1})$'s to follow a uniform distribution on $(0, 0.5)$ and set an administrative censoring at $t = 4$. For the estimation of the cumulative baseline hazard function, we adopted spline basis functions with degree $(K - d) = 3$, and $d = 4$ interior

knots were placed at the empirical quantiles of the finite and positive end points of the observed intervals. We used 20 nodes in the Gaussian quadrature and chose 10^{-3} as the tolerance level for the convergence criterion of the EM algorithm.

Two sets of simulation studies were performed, distinguished by whether $\boldsymbol{Z}$ and $\boldsymbol{X}$ share a common covariate or not. In Scenario I, we set $\boldsymbol{Z}_i = (1, Z_{i1})^{\mathrm{T}}$ and $X_{ij} = X_{ij1}$, where Z_{i1} and X_{ij1} are independent standard normal variables. In Scenario II, we set $\boldsymbol{Z}_i = (1, Z_{i1}, Z_{i2})^{\mathrm{T}}$, $\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2})^{\mathrm{T}}$, where Z_{i1} and X_{ij1} are independent Bernoulli variables with probability 0.5, and $Z_{i2} \equiv X_{ij2}$ is a standard normal variable for all j. The true values of $(\boldsymbol{\alpha}, \kappa, \boldsymbol{\beta})$ are given in Table 2. A positive (negative) κ refers to the case that the cluster size is positively (negatively) related to the risk of the event. Here, we set a negative value for κ to mimic the ICS nature of the GAAD data, where a subject with a better oral health condition (i.e., a small b_i) would generally have more teeth at inspection. The results (including the bias, empirical standard deviation, average standard error, and 95% coverage probability for each parameter), based on 1000 replicates for each scenario, are presented in Table 2. We also plot the averaged estimates for Λ in Figure 1.

For comparison, we implemented a marginal, cluster-weighted pseudo likelihood approach on the same simulated datasets. This approach assumes the correct marginal distribution for $T_1, \ldots, T_N$ but does not model the association among subjects within the same cluster. Similar to Zhang and Sun (2010a, 2013), Zhao et al. (2018), and Lam et al. (2021), it maximizes a cluster-weighted pseudo log-likelihood function, where the contribution of each subject to the log-likelihood is weighted by the size of its cluster. Details of the estimation and computation methods are described in Section S4 of the Supporting Information. We report the bias, empirical standard deviation, and relative efficiency of each parameter estimate in Table 2; relative efficiency is defined as the ratio of the empirical variance of the estimates from the proposed approach to that from the marginal approach.

[Figure 1 about here.]

From Table 2, we observe that the proposed estimators are virtually unbiased, and Louis' method provides accurate estimates for the standard errors of the estimators. The empirical coverage probabilities of the confidence intervals are close to the 95% nominal level. The estimator of the cumulative baseline hazard function is apparently unbiased under various scenarios. For the marginal approach, because the marginal model is correctly specified and that the cluster-weighting approach properly accounts for ICS, the parameter estimators in the survival model are unbiased. However, because the marginal approach is not based on the full likelihood, it is substantially less efficient than the proposed approach. Note that the empirical distributions of the estimates under the marginal approach tend to have heavier tails than the normal distribution, resulting in unusually large empirical standard errors.

[Table 2 about here.]

To investigate the robustness of the proposed estimators of $\boldsymbol{\beta}$ and σ in the failure time regression model when the cluster size model is misspecified, we generated the cluster size N from the proportional odds model (Kim, 2010):

$$P(N \leqslant k \mid \boldsymbol{Z}, b) = \begin{cases} 0, & \text{for } k \leqslant 4; \\ \dfrac{e^{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{Z}+\kappa b}}{1+e^{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{Z}+\kappa b}}, & \text{for } k = 5,6,7,8,9; \\ 1, & \text{for } k \geqslant 10. \end{cases}$$

We considered the same true values for the non-intercept coefficients of $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}$, κ and σ as in the first set of simulation. The intercepts of $\boldsymbol{\alpha}_k$ are set to be -3, -1, 1, 2, and 3 for $k = 5, \ldots, 9$, respectively. The estimation results of the proposed binomial-Cox model, based on 1000 replicates, are reported in Table 3. The estimates for $\boldsymbol{\beta}$ and σ are close to their true values under the misspecified model. The standard error estimates are still accurate for these parameters, which yield proper coverage probabilities. To further illustrate the fit of the proposed model under misspecification, we compute the estimated marginal survival function at $\boldsymbol{X}_{ij} = \boldsymbol{0}$, i.e., $S(t) = \int \exp\{-\Lambda(t)e^b\} f_b(b;\sigma)\,\mathrm{d}b$, where the parameters are evaluated at the sieve ML estimator. We

plot the estimated survival curves in Figure 2 under Scenarios I and II for $n = 50$ and 200, with the corresponding true survival curves overlayed. We observe that the estimated survival curves align closely with the true ones. Nevertheless, because the cluster size model is misspecified, the marginal distribution of N is not well-estimated (based on results not presented here). This suggests that estimation of the survival model is largely not sensitive to the misspecification of the cluster size model.

[Table 3 about here.]

[Figure 2 about here.]

To further investigate the robustness of the proposed method, we considered violation of assumptions on the frailty distribution, and assumptions on the association structure among the failure times and cluster size. First, we considered a Gamma frailty instead of a (log-) normal frailty. Second, we considered a two-frailty model, where an extra normal frailty u is included and shared among the failure times only, in addition to the frailty b shared among the failure times and cluster size. The models and results are presented in Section S5 of the Supporting Information. In all additional scenarios considered, the point estimates for $\boldsymbol{\beta}$ are almost identical to the true values, and the standard error estimates match closely with the empirical standard deviations. The estimator of $\boldsymbol{\alpha}$ has very satisfactory performance under the gamma-frailty model, but can be slightly biased under the two-frailty model. Overall, the results suggest that the proposed method is robust against frailty misspecification.

Finally, we considered a setting with a larger maximum cluster size of $m = 28$ and used the Akaike Information Criterion (AIC) to select d, the number of interior knots for the spline functions. This is to mimic the real data analysis presented in Section 5. The details and results are given in Section S5 of the Supporting Information. The pattern of results is very similar to those under the original simulation settings, with highly satisfactory point and interval estimates.

5. Application: GAAD data

We apply the proposed method to the GAAD dataset introduced in Section 1. In the study, each subject underwent an oral examination, and answered a detailed questionnaire on their socio-demographic, behavioral, and clinical characteristics. Diabetes control was assessed, based on the measurement of glycosylated hemoglobin (HbA1c) in their peripheral blood samples. The oral examination recorded the site-specific clinical attachment loss (in mm), an important biomarker of PD, for each tooth, excluding the third molars (Darby and Walsh, 2010). A tooth is defined as having (moderate to severe) PD if its mean clinical attachment loss is $\geqslant$ 3mm. A total of $n = 288$ subjects were recruited, among which 170 have at least one tooth suffering from PD at the time of inspection. The response variable of interest is time to onset of PD since tooth emergence. Naturally, this is a current-status response, as the PD status is observed only at the time of inspection, with $\Delta_{Iij} \equiv 0$ for $j = 1, \ldots, N_i$ and $i = 1, \ldots, n$. Also, the (true) inspection time, U_{ij1}, should be measured from the (adult) tooth emergence time, which is never recorded in any study. Hence, in this paper, we approximate it using the permanent dentition times of US adults published by the American Dental Association (2022). Furthermore, the cluster size N_i, the number of teeth at inspection, ranges from 3 to 28. The association between the proportion of diseased teeth and the number of teeth per person is illustrated in Table 1. It shows that the cluster size has non-ignorable associations with the PD outcome, as the diseased proportion decreases with the number of teeth.

[Table 4 about here.]

The primary goal of this analysis is to explore the association among covariates, failure times, and cluster size, subject to the ICS situation. The subject-specific covariates include smoking status (smoker $= 1$), gender (female $= 1$), and HbA1c (uncontrolled $= 1$). These variables along with age at inspection constitute $\boldsymbol{Z}_i$ in the binomial regression model. Furthermore, two tooth-specific variables can be identified, namely the jaw and molar indicator of the tooth, abiding by the universal

numbering system (also called the "American System") of teeth. Specifically, we set jaw $= 1$ for teeth 2–15 located on the upper jaw, and molar $= 1$ for teeth with labels 2–5, 12–15, 18–21 and 28–31, identified as premolars or molars. For the design matrix $\boldsymbol{X}_{ij}$ corresponding to the survival regression, we included both subject- and tooth-specific covariates. For the I-splines, we considered basis functions with degree $(K - d) = 3$. We selected the number of interior knots (located at equally-spaced quantiles) from $d \in \{0, \ldots, 5\}$ using the AIC (Burnham and Anderson, 2003). Our analysis yields $d = 3$, and the estimates obtained based on different candidates of d share the same direction and similar magnitude. A tolerance level of 10^{-6} was used as the stopping criterion for the EM algorithm.

The parameter estimates, standard error estimates, and 95% confidence intervals, corresponding to our proposed model fit, are presented in Table 4. Age is significantly associated with the number of teeth, while HbA1c, gender, jaw, and molar are significantly associated with the risk of PD. In particular, older subjects tend to have fewer teeth (Shimazaki et al., 2003), whereas teeth of males, of subjects with uncontrolled HbA1c, teeth located in the upper jaw, and molar teeth exhibit a significantly higher risk of PD, compared to the teeth of females, well-controlled HbA1c, teeth located in the lower jaw, and non-molar teeth, respectively. The covariates that exhibit significant effects in the failure time model, i.e., HbA1c and gender, are insignificant in the cluster size model. From the confidence intervals for α_2 and β_1, we observe that smoking status is borderline insignificant in both regressions. Smoking as a significant risk factor for PD has been established in a number of epidemiological studies (Kinane and Chestnutt, 2000; Amarasena et al., 2002; Thomson et al., 2007), with the smokers having 2.5–3.5 times greater risk of severe periodontal attachment loss, in general (Bergström, 1989). From our data analysis, the negative sign of α_2 (in the cluster size regression) resonates with these findings, i.e., an inverse association of smoking with teeth cardinality, while the positive sign of β_1 (in the failure time regression) is in tune to the aforementioned established positive association. The estimate for σ differs substantially

from zero, which demonstrates a moderate degree of within-cluster association. The estimate for the coefficient of the frailty term in the binomial regression model (i.e., κ) is negative and is significantly different from zero. This finding provides evidence that the cluster size is informative, and subjects with more teeth are less prone to PD.

As a comparison, we also analyzed the GAAD data using the standard frailty Cox model. A sieve ML estimation approach was adopted, with the same I-spline approximation for the cumulative baseline hazard function as in the proposed method. The results are presented in Table 4. Although the point estimates corresponding to the regression coefficients from the frailty Cox model and our proposed model are comparable, the standard error estimates from the frailty Cox model (which does not utilize information from the cluster size) are, on the overall, larger than those obtained under the proposed model. In particular, HbA1c becomes insignificant under the frailty Cox model. Note that HbA1c is considered a standard of care for testing and monitoring T2D. The American Diabetes Association recommends a target HbA1c level of $< 7\%$ (well-controlled), ideally within 4–6, for people with T2D (Reichard et al., 1993; Lyons and Basu, 2012). The substantial adverse effect of T2D on PD, including their bi-directional relationship, has been extensively explored in oral health studies (Mealey and Oates, 2006; Preshaw et al., 2012; Grover and Luthra, 2013; Genco and Borgnakke, 2020). The findings from the current study, i.e., uncontrolled HbA1c leads to a higher risk of PD, align with previously established conclusions.

6. Discussion

In this paper, we propose a joint model for clustered interval-censored survival data with ICS, motivated by a clinical study of PD. Under our proposed model, the strength and direction of the ICS effect are captured explicitly by the coefficient κ (corresponding to the frailty term b), which allows a formal test for the presence of ICS. This is not possible under a marginal approach. Utilizing the full likelihood, our approach leads to enhanced statistical efficiency compared to

marginal approaches. Although our motivating application involves current-status outcomes (case-1 interval censoring), our methods are established under a general interval censoring setup.

As pointed out by a reviewer, the current problem is inherently a missing data problem (with $(T_{N+1}, \ldots, T_m)$ being unobserved), where multiple imputation (MI) methods could be applicable. Nevertheless, existing MI methods are typically used to impute the exact failure times from the observed intervals (Hsu et al., 2007; Alarcón-Soto et al., 2019), thereby allowing methods for right-censored (and exact) data to be directly employed. In the current context, the data to be imputed are the unobserved event times $(T_{N+1}, \ldots, T_m)$, and relevant methods are non-trivial and have not been developed in the literature. In addition, when many of the observed cluster sizes are small compared to the maximum cluster size, a large number of missing variables have to be imputed. In this case, the updated parameter estimates at each iteration would be close to the current estimates, resulting in slow convergence. Further investigations on MI methods could be conducted in the future.

Our work can be further extended in various directions. In this current work, we propose a binomial regression model for the cluster size motivated by oral health studies where there exists a natural upper bound of $m = 28$ for the cluster size. In general settings, we can consider other cluster size models, such as the Poisson or Conway–Maxwell–Poisson model (Huang, 2017) to tackle over/under dispersed counts, or the proportional odds model. Computation of the sieve ML estimator can be performed using a similar EM-type algorithm, and its asymptotic properties can be established using arguments similar to the proofs of Theorems 1–3. Also, instead of directly modeling the number of teeth by the time of observation, we may consider a model for the time to tooth loss. Suppose that conditional on $\boldsymbol{Z}$ and b, the time to tooth loss for each tooth is independent, with a cumulative hazard function $H(t)e^{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}+\kappa b}$ where H is some positive nondecreasing function. In this case, the number of teeth at time t_0 follows the binomial distribution with probability $\exp\{-e^{\log H(t_0)+\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}+\kappa b}\}$. This alternative model is similar to the proposed model, except that the

former uses a different link function for the binomial mean and allows the observation time to have a nonlinear effect. Estimation can be performed using a similar sieve ML approach. However, a large sample size may be required for stable estimation in presence of two nonparametric components. Furthermore, we can consider joint models for the failure times and the cluster size process. In the current framework, if a unit drops out from a cluster before the failure event, the failure time is right-censored at the drop-out time. Depending on the failure event of interest, such censoring may be informative (Yu et al., 2022), and it is of interest to develop joint models for the censoring and failure times. Finally, modifications to the proposed EM algorithm are needed to attain computational scalability in light of applying our methods to data derived from large observational settings. All these are important avenues and will be considered elsewhere.

Acknowledgements

The authors would like to thank the editor, associate editor, and reviewers for their constructive comments that has led to a marked improvement in the quality of the article. The authors also thank Dr. Jyotika K. Fernandes from the Medical University of South Carolina for providing the motivating data, and the context for this work. This research was supported by a start-up research grant from the Hong Kong Polytechnic University (P0035688) and the Hong Kong Research Grants Council, University Grants Committee (Grant/Award Number: 17305819). Bandyopadhyay acknowledges partial funding support from grants R21DE031879, R01DE024984 and R01DE031134 from the United States National Institutes of Health.

Data Availability Statement

The data that support the findings of this paper are openly available in the Supporting Information of this article.

References

Alarcón-Soto, Y., Langohr, K., Fehér, C., García, F., and Gómez, G. (2019). Multiple imputation approach for interval-censored time to HIV RNA viral rebound within a mixed effects Cox model. *Biometrical Journal* **61,** 299–318.

Amarasena, N., Ekanayaka, A. N., Herath, L., and Miyazaki, H. (2002). Tobacco use and oral hygiene as risk indicators for periodontitis. *Community Dentistry and Oral Epidemiology* **30,** 115–123.

American Dental Association (2022). Eruption Charts. https://www.mouthhealthy.org/en/az-topics/e/eruption-charts. (accessed August 4, 2022).

Bergström, J. (1989). Cigarette smoking as risk factor in chronic periodontal disease. *Community Dentistry and Oral Epidemiology* **17,** 245–247.

Burnham, K. P. and Anderson, D. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Chen, Z., Zhang, B., and Albert, P. S. (2011). A joint modeling approach to data with informative cluster size: robustness to the cluster size model. *Statistics in Medicine* **30,** 1825–1836.

Cong, X. J., Yin, G., and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics* **63,** 663–672.

Darby, M. L. and Walsh, M. (2010). *Dental Hygiene: Theory and Practice*. Elsevier Health Sciences, 3rd edition.

Dunson, D. B., Chen, Z., and Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59,** 521–530.

Fernandes, J. K., Wiegand, R. E., Salinas, C. F., Grossi, S. G., Sanders, J. J., Lopes Virella, M. F., and Slate, E. H. (2009). Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina. *Journal of Periodontology* **80,** 1062–1068.

Gamage, P. W. W., McMahan, C. S., Wang, L., and Tu, W. (2018). A gamma-frailty proportional

hazards model for bivariate interval-censored data. *Computational Statistics & Data Analysis* **128,** 354–366.

Genco, R. J. and Borgnakke, W. S. (2020). Diabetes as a potential risk for periodontitis: Association studies. *Periodontology 2000* **83,** 40–45.

Grover, H. S. and Luthra, S. (2013). Molecular mechanisms involved in the bidirectional relationship between diabetes mellitus and periodontal disease. *Journal of Indian Society of Periodontology* **17,** 292–301.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88,** 1121–1134.

Hsu, C.-H., Taylor, J. M., Murray, S., and Commenges, D. (2007). Multiple imputation for interval censored data with auxiliary variables. *Statistics in Medicine* **26,** 769–781.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling* **17,** 359–380.

Huang, J. and Wellner, J. A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case 1. *Statistica Neerlandica* **49,** 153–163.

Kim, Y. J. (2010). Regression analysis of clustered interval-censored data with informative cluster size. *Statistics in Medicine* **29,** 2956–2962.

Kinane, D. and Chestnutt, I. (2000). Smoking and periodontal disease. *Critical Reviews in Oral Biology & Medicine* **11,** 356–365.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2016). *Handbook of Survival Analysis*. CRC Press.

Kor, C. T., Cheng, K. F., and Chen, Y. H. (2013). A method for analyzing clustered interval-censored data based on Cox's model. *Statistics in Medicine* **32,** 822–832.

Lam, K. F., Lee, C. Y., Wong, K. Y., and Bandyopadhyay, D. (2021). Marginal analysis of current status data with informative cluster size using a class of semiparametric transformation cure

models. *Statistics in Medicine* **40,** 2400–2412.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika* **81,** 624–629.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B* **44,** 226–233.

Lyons, T. J. and Basu, A. (2012). Biomarkers in diabetes: Hemoglobin A1c, vascular and tissue markers. *Translational Research* **159,** 303–312.

McGee, G., Weisskopf, M. G., Kioumourtzoglou, M.-A., Coull, B. A., and Haneuse, S. (2020). Informatively empty clusters with application to multigenerational studies. *Biostatistics* **21,** 775–789.

Mealey, B. L. and Oates, T. W. (2006). Diabetes mellitus and periodontal diseases. *Journal of Periodontology* **77,** 1289–1303.

Neuhaus, J. M. and McCulloch, C. E. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98,** 147–162.

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

Preshaw, P., Alba, A., Herrera, D., Jepsen, S., Konstantinidis, A., Makrilakis, K., and Taylor, R. (2012). Periodontitis and Diabetes: A two-way relationship. *Diabetologia* **55,** 21–31.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3,** 425–441.

Reichard, P., Nilsson, B.-Y., and Rosenqvist, U. (1993). The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus. *New England Journal of Medicine* **329,** 304–309.

Shimazaki, Y., Soh, I., Koga, T., Miyazaki, H., and Takehara, T. (2003). Risk factors for tooth loss in the institutionalised elderly; a six-year cohort study. *Community Dental Health* **20,**

123–127.

Spiekerman, C. F. and Lin, D. Y. (1998). Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association* **93,** 1164–1175.

Sun, L., Li, S., Wang, L., Song, X., and Sui, X. (2021). Simultaneous variable selection in regression analysis of multivariate interval-censored data. *Biometrics* [online] DOI: 10.1111/biom.13548.

Thomson, W. M., Broadbent, J. M., Welch, D., Beck, J. D., and Poulton, R. (2007). Cigarette smoking and periodontal disease among 32-year-olds: A prospective study of a representative birth cohort. *Journal of Clinical Periodontology* **34,** 828–834.

Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics* **35,** 335–353.

Wang, L., McMahan, C. S., Hudgens, M. G., and Qureshi, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* **72,** 222–231.

Williamson, J. M., Datta, S., and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59,** 36–42.

Williamson, J. M., Kim, H. Y., Manatunga, A., and Addiss, D. G. (2008). Modeling survival data with informative cluster size. *Statistics in Medicine* **27,** 543–555.

Yu, M., Feng, Y., Duan, R., and Sun, J. (2022). Regression analysis of multivariate interval-censored failure time data with informative censoring. *Statistical Methods in Medical Research* **31,** 391–403.

Zeng, D., Gao, F., and Lin, D. Y. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika* **104,** 505–525.

Zhang, B., Liu, W., Zhang, Z., Qu, Y., Chen, Z., and Albert, P. S. (2017). Modeling of correlated data with informative cluster sizes: an evaluation of joint modeling and within-cluster

resampling approaches. *Statistical Methods in Medical Research* **26,** 1881–1895.

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

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

Zhang, Z. and Sun, J. (2010b). Interval censoring. *Statistical Methods in Medical Research* **19,** 53–70.

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

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

Supporting Information

Web Appendices referenced in Sections 2–4 and the `R` code implementing the proposed methods are available with this paper at the *Biometrics* website on Wiley Online Library.

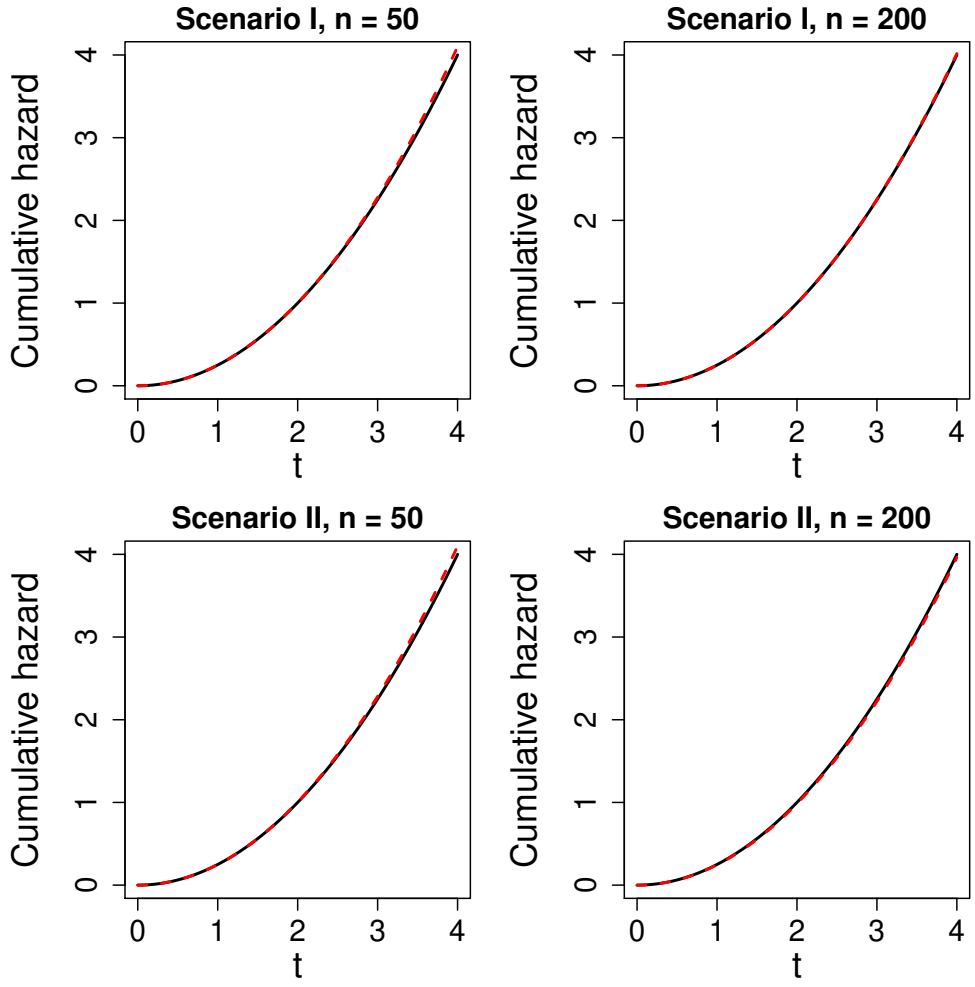

Figure 1. Simulation results corresponding to the estimation of the cumulative hazard function for the proposed approach under Scenarios I and II with sample sizes $n = 50$ and 200. The solid black lines are the true values and the dashed red lines are the averaged estimates. This figure appears in color in the electronic version of this article, and color refers to that version.

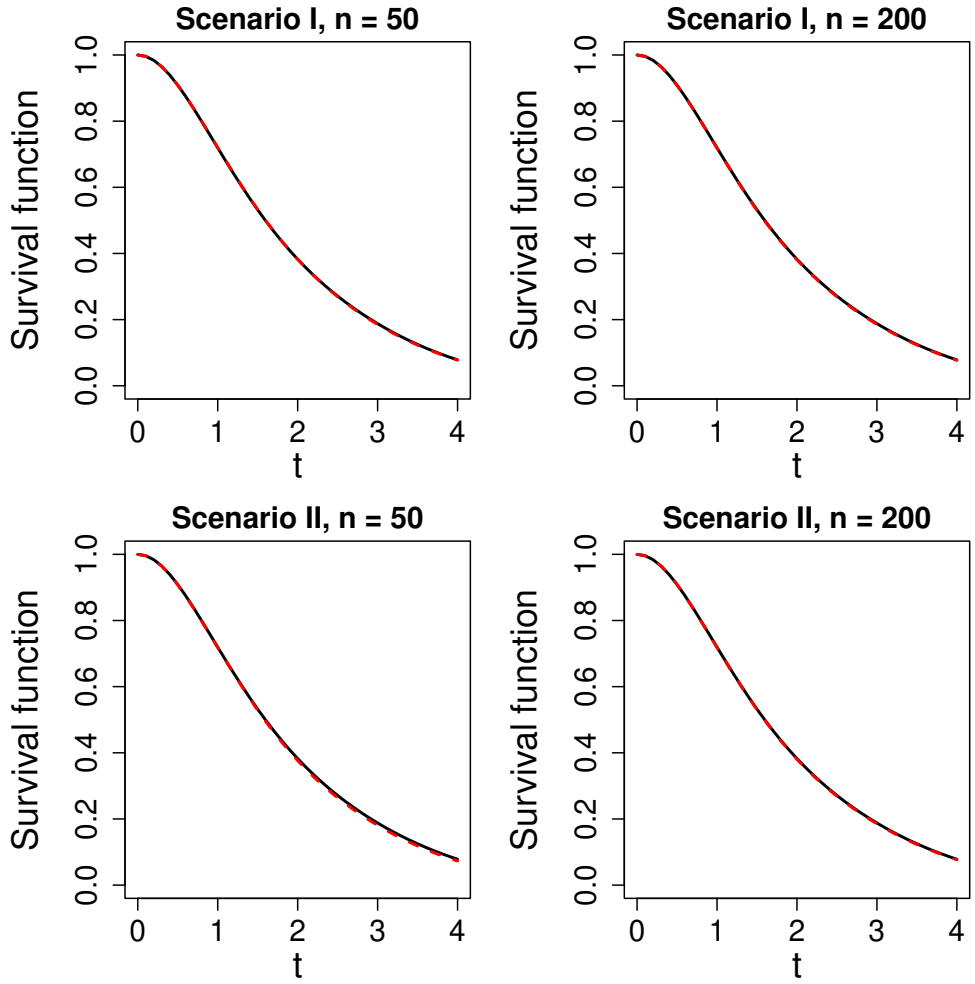

Figure 2. Simulation results corresponding to the estimation of the marginal survival function under a misspecified cluster size model with data generated under Scenarios I and II for sample sizes $n = 50$ and 200. The solid black lines are the true values and the dashed red lines are the averaged estimates under the proposed model. This figure appears in color in the electronic version of this article, and color refers to that version.

Table 1
GAAD Data: Percentage of diseased teeth and the number of teeth

Number of teeth	Number of subjects	Percentage of diseased teeth
3–10	40	34.1
11–16	43	29.1
17–19	51	14.7
20–22	57	19.4
23–25	53	13.0
26–28	44	9.0
Total	288	16.7

Table 2

Simulation summaries from fitting the proposed and marginal approaches with data generated under Scenarios I and II for sample sizes $n = 50$ and 200. Reported are True = true parameter value; Bias; ESD = empirical standard deviation; ASE = average standard error; CP = 95% coverage probability; and RE = relative efficiency

Scenario I			Proposed approach				Marginal approach		
n	Parameter	True	Bias	ESD	ASE	CP	Bias	ESD	RE
50	α_1	1	0.013	0.132	0.133	0.94		–	
	α_2	$\log(2)$	0.007	0.126	0.125	0.94		–	
	κ	-0.5	0.002	0.138	0.140	0.95		–	
	β_1	1	0.007	0.091	0.088	0.94	0.082	0.314	0.084
	σ	1	-0.014	0.143	0.139	0.94	0.075	0.593	0.058
200	α_1	1	0.004	0.067	0.066	0.94		–	
	α_2	$\log(2)$	-0.004	0.062	0.060	0.94		–	
	κ	-0.5	0.001	0.065	0.067	0.96		–	
	β_1	1	0.002	0.045	0.044	0.94	0.035	0.185	0.059
	σ	1	-0.007	0.071	0.070	0.95	0.040	0.349	0.041
Scenario II			Proposed approach				Marginal approach		
50	α_1	1	0.020	0.177	0.179	0.95		–	
	α_2	$-\log 2$	0.007	0.218	0.228	0.97		–	
	α_3	-1	0.025	0.145	0.148	0.95		–	
	κ	-0.5	0.009	0.132	0.142	0.96		–	
	β_1	1	0.000	0.150	0.149	0.96	0.027	0.393	0.145
	β_2	$\log 2$	-0.017	0.184	0.173	0.93	0.010	0.287	0.412
	σ	1	-0.026	0.145	0.144	0.94	-0.065	0.652	0.049
200	α_1	1	0.003	0.088	0.087	0.95		–	
	α_2	$-\log 2$	0.013	0.111	0.111	0.94		–	
	α_3	-1	0.036	0.068	0.072	0.93		–	
	κ	-0.5	0.016	0.066	0.067	0.93		–	
	β_1	1	0.007	0.076	0.074	0.94	0.042	0.255	0.089
	β_2	$\log 2$	-0.018	0.086	0.085	0.93	0.008	0.177	0.236
	σ	1	-0.013	0.072	0.072	0.95	0.018	0.435	0.027

Table 3

Simulation summaries under a misspecified cluster size model, with data generated from Scenarios I and II for sample sizes $n = 50$ *and* 200. *Reported are True = true parameter value; Bias; ESD = empirical standard deviation; ASE = average standard error; and CP = 95% coverage probability*

Scenario I		$n = 50$				$n = 200$			
Parameter	True	Bias	ESD	ASE	CP	Bias	ESD	ASE	CP
β_1	1	0.007	0.087	0.088	0.95	0.002	0.043	0.044	0.96
σ	1	-0.014	0.135	0.136	0.95	-0.004	0.070	0.068	0.94
Scenario II		$n = 50$				$n = 200$			
β_1	1	0.013	0.140	0.140	0.94	-0.003	0.069	0.069	0.95
β_2	$\log(2)$	0.009	0.165	0.162	0.94	-0.002	0.079	0.079	0.96
σ	1	-0.026	0.135	0.135	0.94	-0.010	0.067	0.068	0.95

Table 4

GAAD data analysis: Table shows the parameter estimates (EST), standard errors (SE), and associated 95% confidence intervals (CI), corresponding to the "Cluster size" regression and the "Failure time" regression that constitutes our proposed model, and the frailty Cox model

		Proposed model			Frailty Cox model		
Cluster size							
Covariate	Par	EST	SE	95% CI	EST	SE	95% CI
Intercept	α_1	3.639	0.286	(3.078, 4.199)		–	
Smoking	α_2	−0.173	0.120	(−0.408, 0.062)		–	
HbA1c	α_3	0.060	0.109	(−0.154, 0.273)		–	
Gender	α_4	0.156	0.131	(−0.101, 0.412)		–	
Age	α_5	−0.052	0.005	(−0.061, −0.043)		–	
	κ	−0.383	0.026	(−0.434, −0.332)		–	
Failure time							
Covariate	Par	EST	SE	95% CI	EST	SE	95% CI
Smoking	β_1	0.555	0.299	(−0.031, 1.140)	0.568	0.351	(−0.119, 1.255)
HbA1c	β_2	0.581	0.273	(0.046, 1.117)	0.609	0.323	(−0.025, 1.243)
Gender	β_3	−1.401	0.324	(−2.037, −0.766)	−1.715	0.376	(−2.452, −0.978)
Jaw	β_4	0.721	0.081	(0.563, 0.879)	0.547	0.081	(0.389, 0.706)
Molar	β_5	0.819	0.080	(0.663, 0.975)	0.729	0.081	(0.570, 0.888)
	σ	2.036	0.134	(1.774, 2.298)	2.278	0.159	(1.966, 2.590)