Published in final edited form as:
J Am Stat Assoc. 2016 ; 111(513): 145–156. doi:10.1080/01621459.2014.995795.

Generalizing Quantile Regression for Counting Processes with Applications to Recurrent Events

Xiaoyan Sun [Doctoral Candidate],
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322 (xsun33@emory.edu).

Limin Peng[*] **[Associate Professor]**,
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322.

Yijian Huang [Professor], and
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322 (yhuang5@emory.edu).

HuiChuan J. Lai [Professor]
Departments of Nutritional Sciences, Pediatrics, and Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706 (hlai@wisc.edu).

Abstract

In survival analysis, quantile regression has become a useful approach to account for covariate effects on the distribution of an event time of interest. In this paper, we discuss how quantile regression can be extended to model counting processes, and thus lead to a broader regression framework for survival data. We specifically investigate the proposed modeling of counting processes for recurrent events data. We show that the new recurrent events model retains the desirable features of quantile regression such as easy interpretation and good model flexibility, while accommodating various observation schemes encountered in observational studies. We develop a general theoretical and inferential framework for the new counting process model, which unifies with an existing method for censored quantile regression. As another useful contribution of this work, we propose a sample-based covariance estimation procedure, which provides a useful complement to the prevailing bootstrapping approach. We demonstrate the utility of our proposals via simulation studies and an application to a dataset from the US Cystic Fibrosis Foundation Patient Registry (CFFPR).

Keywords

accelerated failure time model; accelerated recurrence time model; censored quantile regression; counting process; recurrent events; varying covariate effects

[*] (; Email: lpeng@emory.edu).

1. INTRODUCTION

Quantile regression (Koenker and Bassett 1978) has gained increasing popularity in survival analysis for its easy interpretation and flexibility in exploring the dynamic relationship between a time-to-event outcome and covariates (Powell 1984; Ying, Jung, and Wei 1995; Yang 1999; Portnoy 2003; Neocleous, Vanden Branden, and Portnoy 2006; Peng and Huang 2008; Portnoy and Lin 2010; Huang 2010, among others). For a time-to-event response T, a quantile regression model may assume that

$$Q_T(\tau|\boldsymbol{Z}) = exp\left\{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}_0^{nr}(\tau)\right\}, \quad \tau \in (0,1), \quad (1)$$

where $Q_T(\tau|\boldsymbol{Z})$ inf$\{t: \Pr(T \le t|\boldsymbol{Z}) \ge \tau$ denotes the τ-th quantile of T given the $p \times 1$ covariate vector $\boldsymbol{Z}$, $\boldsymbol{X} = (1, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\beta}_0^{nr}(\tau)$ is a $(p+1) \times 1$ vector of regression coefficients. By formulating covariate effects on different quantiles of T, model (1) enables a comprehensive examination of covariates' impact on the distribution of a time-to-event outcome. Unless otherwise specified, this article is confined to regression settings with only time-independent covariates.

In this paper, we consider extending quantile regression to model counting processes, a more general notion for describing outcomes observed in survival studies as compared to the time-to-event formulation adopted by model (1). In the traditional setting concerning only one event time, the survival information can be characterized by a counting process with a single jump at the observed event time. In recurrent events settings, where the event of interest (e.g. infection, hospitalization) can occur repeatedly, a single event time usually fails to fully capture the event history information of interest. In contrast, the counting process of recurrent events, which allows for multiple jumps, can well depict the trajectory of event occurrence and thus capture event history in full.

Many traditional survival models have been studied for their extensions for counting processes. Examples include the Cox's regression model to counting processes by Andersen and Gill (1982), the accelerated failure time model for counting processes by Lin, Wei, and Ying (1998), and more recently, transformation models for counting processes by Zeng and Lin (2006). As quantile regression has emerged as a valuable regression tool for survival data, studying its generalization for counting processes constitutes a sensible effort that can lead to two-fold benefits. First, the new counting process model is expected to handle additional types of survival data that cannot be straightforwardly covered by quantile regression model (1), such as recurrent events data. Second, as will be explained below, counting process based modeling generally facilitates the accommodation of various incomplete follow-up scenarios, a task that can be more difficult when a time-to-event formulation is adopted, as in quantile regression model (1).

Overall, this work bears a general goal of developing new counting process models extended from quantile regression modeling of a time-to-event response. We shall expound the main ideas in a recurrent events setting that arises from our motivating study. The presented strategies for estimation and inference are readily adaptable to other survival settings where data can be meaningfully captured by counting processes.

2. THE PROPOSED COUNTING PROCESS MODEL

We begin with a review of Andersen and Gill (1982)'s counting process formulation of the Cox proportional hazards model (Cox 1972). Let T and C denote time to an event of interest and time to censoring, respectively. With T subject to right censoring by C, observables include $\tilde{T} \equiv T \wedge C$ and $\delta \equiv I(T \leq C)$, where $\wedge$ is the minimum operator. When there are only time-independent covariates, the counting process formulation of the Cox model can be given by

$$E\left\{dN^{nr}(t)\,|Y^{nr}(t), \boldsymbol{Z}\right\} = Y^{nr}(t)\,\lambda_0(t)\,e^{\boldsymbol{Z}^\top \boldsymbol{b}}dt, \quad t>0, \quad (2)$$

where $N^{nr}(t) = I(\tilde{T} \leq t, \delta = 1)$, and $Y^{nr}(t) = I(\tilde{T} \geq t)$, representing the observed counting process and the at-risk process for the event of interest, respectively. Here $\lambda_0(s)$ stands for the baseline hazard function, $\boldsymbol{Z}$ denotes a $p \times 1$ covariate vector, and $\boldsymbol{b}$ denotes the vector of regression coefficients. Model (2) formulates proportional covariate effects on $E\{dN^{nr}(t)/Y^{nr}(t), \boldsymbol{Z}\}$, which corresponds to the intensity process for the counting process $N^{nr}(t)$ given $\boldsymbol{Z}$. As shown by Andersen and Gill (1982), the counting process formulation of the Cox model not only facilitates a rigorous development of asymptotic theory, but also renders a broadened regression framework that can well accommodate recurrent events data.

Through a re-examination of the work by Peng and Huang (2008), we can derive a counting process formulation of censored quantile regression model (1). That is, when T is subject to random censoring by C, which is independent of T given $\boldsymbol{Z}$, it holds under model (1) that

$$E\left\{N^{nr}\left(e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}}(\tau)\right)|\boldsymbol{Z}\right\} = E\left\{\int_0^\tau Y^{nr}\left(e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(s)}\right)\frac{1}{1-s}ds|\boldsymbol{Z}\right\}, \quad \tau \in (0,1). \quad (3)$$

In terms of the intensity process, (3) can be rewritten as

$$E\left\{dN^{nr}\left(e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(\tau)}\right)|Y^{nr}\left(e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(\tau)}\right)\boldsymbol{Z}\right\} = Y^{nr}\left(e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(\tau)}\right)\frac{1}{1-\tau}d\tau. \quad (4)$$

By (4), the intensity for $N^{nr}(t)$ equals $\frac{Y^{nr}\left(e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(\tau)}\right)}{(1-\tau)e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(\tau)}\left\{\boldsymbol{X}^\top \dot{\boldsymbol{\beta}}_0^{nr}(\tau)\right\}}$ at time $t=e^{\boldsymbol{X}^\top \boldsymbol{\beta}_0^{nr}(\tau)}$, where $\dot{\boldsymbol{\beta}}_0^{nr}(\tau)$ denotes the derivative of $\boldsymbol{\beta}_0^{nr}(\tau)$ with regard to τ. This shows that censored quantile regression does not imply a simple relationship between event intensity process and covariates, unlike the Cox model (2). Nevertheless, there exists an analogy between (2) and (4) in the sense that both models incorporate covariate effects via specifying the link between the counting process $N^{nr}(\cdot)$ and the at-risk process $Y^{nr}(\cdot)$. Such a view implicates a general strategy for modeling a counting process, which is not limited to formulating covariate effects on its intensity process.

Following this general modeling strategy, we propose a new counting process model that takes the form,

$$E\left\{N\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(u)}\right)|\boldsymbol{Z}\right\}=E\left\{\int_0^u Y\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(s)}\right)g(s)\,ds|\boldsymbol{Z}\right\}, \quad u\in(0,U], \quad (5)$$

where $N(\cdot)$ is a counting process of interest, $Y(\cdot)$ is an at-risk process appropriately specified according to $N(\cdot)$, $g(\cdot)$ is a known positive and continuous function, $\beta_0(\cdot)$ is a $(p+1)\times 1$ vector of unknown regression coefficient functions, and U is a positive constant. Note that the at-risk process $Y(\cdot)$ is only required to be left continuous, and thus it can be well defined in many incomplete follow-up scenarios (e.g. window observations of recurrent events). This indicates the flexibility of model (5) to accommodate realistic data observation schemes in survival studies.

The proposed model (5) is motivated by the counting process model (3) implied by the quantile regression model (1). In the standard survival setting with $N(\cdot) = N^{nr}(\cdot)$, $Y(\cdot) = Y^{nr}(\cdot)$, and $g(u) = 1/(1-u)$, model (5) becomes model (3). Proposition A1 of the Supplementary Materials shows that model (3) is equivalent to a weaker version of model (1) with $\tau\in(0, 1)$ replaced by $\tau\in(0, \tau_U]$, where τ_U is a positive constant less than 1. This result reflects the connection between censored quantile regression and the proposed counting process model.

In model (5), a general function $g(u)$ is adopted in place of the function $1/(1-u)$ in (3). The specification of $g(\cdot)$ determines the scale in which covariate effects are formulated. For example, in the standard random censoring case concerning a single event time T, specifying $g(u)\equiv 1$, rather than $1/(1-u)$, would yield a model that formulates covariate effects on the inverse cumulative hazard function, namely $\inf\{t : \Lambda_T(t|\boldsymbol{Z})\geq u\}$, rather than the conditional quantile function, where $\Lambda_T(t|\boldsymbol{Z})$ denotes the cumulative hazard function of T given $\boldsymbol{Z}$. At the same time, we can show that allowing $g(\cdot)$ to take a general form in the proposed model generally does not incur extra complications in inference and computation.

3. THE APPLICATION TO RECURRENT EVENTS DATA

We shall illustrate the major methodological ideas by applying the proposed counting process model to recurrent events data.

Recurrent events data is an important type of survival data, which is frequently encountered in clinical and epidemiological studies. There has been a vast literature on recurrent events data analysis. For example, well-known methods include modeling the intensity process of recurrent events (Prentice, Williams, and Peterson 1981; Andersen and Gill 1982; Chang and Wang 1999, among others) and modeling the marginal hazard of each recurrent event (Wei, Lin, and Weissfeld 1989; Cai and Prentice 1995, 1997; Spiekerman and Lin 1998, among others) or the gap time between recurrent events (Huang and Chen 2003; Schaubel and Cai 2004; Sun, Park, and Sun 2006, among others). In addition, Hougaard (2000) provided a detailed account of how recurrent events problems can be fit into the paradigm of multi-state models, for which transition probability or transition intensity between states can be estimated to characterize event progression. Shared frailty models were also proposed to explicitly model intra-subject correlations (Liang, Self, Nanndeen-Roche, and Zeger 1995; Wei and Glidden 1997; Hougaard 2000, among others). Adopting frailty models can provide additional information on the dependency among the event time components of recurrent events data.

Another popular approach for recurrent events data is to specify covariate effects on mean or rate functions of recurrent events (Pepe and Cai 1993; Lawless and Nadeau 1995; Lin, Wei, Yang, and Ying 2000; Schaubel, Zeng, and Cai 2006; Huang and Peng 2009; Sun and Guo 2011, among others). This type of approach is attractive because mean or rate functions are more intuitive to interpret than intensity or hazard functions. Modeling these quantities can avoid assumptions about the dependency of recurrent events, which are difficult to verify but are required by intensity models, multi-state models or shared frailty models.

3.1 The proposed recurrent events model

We consider a general data situation, where the observation of recurrent events is subject to an observation window specified as a time interval $(L, R]$ (Nelson 2003). As a result, the counting process for the observed recurrent events is given by $N^{re}(t) = \sum_{j=1}^{\infty} I\left(L \leq T_j^{(i)} \leq t \wedge R\right)$, where $T^{(j)}$ denotes the jth recurrent event time ($j = 1, 2, \ldots$), and the at-risk process is given by $Y^{re}(t) = I(L < t \leq R)$. The observed recurrent events data include n i.i.d. replicates of $N^{re}(\cdot)$, $\boldsymbol{Z}$, L, and R, denoted by $\{N_i^{re}(\cdot), \boldsymbol{Z}_i, L_i, R_i\}_{i=1}^{n}$.

Take the U.S. Cystic Fibrosis (CF) Foundation Patient Registry (CFFPR) data for example. Pseudomonas aeruginosa (PA) is a common and major pathogen in cystic fibrosis (CF) lungs, often resulting in chronic infections. Since no record of PA infections is available before registry entry, the observation window for PA infections is from the age at the first registry visit (L) to the age at the last follow-up before data collection cut-off date (R).

It is worth mentioning that nonzero lower bounds of observation windows have traditionally been paid less attention in recurrent events data analyses as compared to upper bounds. However, they are commonly encountered in observational studies. In CFFPR, a large proportion of subjects did not enter the registry right after birth due to delayed CF diagnosis or other reasons, resulting in many nonzero L_i's. As implied by our numerical studies, naively treating them as zeros can considerably bias the inference regarding PA infection recurrence times.

The proposed counting process model (5) offers a natural way to model recurrent events subject to window observation. With $N(\cdot)$ and $Y(\cdot)$ specified as $N^{re}(\cdot)$ and $Y^{re}(\cdot)$ respectively, model (5) gives rise to a new recurrent events model,

$$E\left\{N^{re}\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(u)}\right)|\boldsymbol{Z}\right\} = E\left\{\int_0^u Y^{re}\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(s)}\right) g(s)\, ds|\boldsymbol{Z}\right\}, \quad u \in (0, U]. \tag{6}$$

Assume (L, R) is independent of $\tilde{N}(\cdot)$ given $\boldsymbol{Z}$. Proposition A2 of the Supplementary Materials shows that (6) is equivalent to

$$\mu_{\boldsymbol{Z}}\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(u)}\right) = G(u) \equiv \int_0^u g(s)\, ds, \quad u \in (0, U],$$

where $\mu_Z(t)$ is the so-called mean function of recurrent events, defined as $\mu_Z(t) = E\{\tilde{N}(t)|\boldsymbol{Z}$ with $\tilde{N}(t) = \sum_{j=1}^{\infty} I\left(T^{(j)} \le t\right)$. The mean function (or expected frequency) of recurrent events reflects the cumulative rate of counting process $N^{re}(t)$, in contrast to the intensity function, which characterizes the instantaneous rate of events conditional on past history (Lin et al. 2000).

Therefore, an alternative representation of model (6) is given by

$$\tau_{\boldsymbol{Z}}(G(u)) = exp\left\{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(u)\right\}, \quad u \in (0, U], \quad (7)$$

where $\tau_Z(u) = \inf\{t \ge 0 : \mu_Z(t) \ge u\}$. The quantity $\tau_Z(u)$ was termed time to expected frequency u by Huang and Peng (2009). When the event of interest can occur only once and thus frequency u is restricted to the range of [0, 1], $\tau_Z(u)$ becomes the conditional quantile of $T^{(1)}$ given $\boldsymbol{Z}$. By model representation (7), the non-intercept coefficients in $\beta_0(u)$ capture the covariate effects on time to expected frequency at $G(u)$.

When $g(u) = 1$ (and thus $G(u) = u$), model (7) becomes the accelerated recurrence time (ART) model (Huang and Peng 2009), which is known to include the accelerated failure time model for recurrent events (Lin et al. 1998) as a special case. On the other hand, equation (6) provides information on how the observed counting process and the at-risk process are related under the ART model. One may choose $g(u)$ as a positive continuous function other than $g(u) = 1$. The essential difference of the resulting model from the ART model would be the coefficient function rescaled from $\beta_0(\cdot)$ to $\beta_0(G^{-1}(\cdot))$, the result of formulating covariate effects on $\tau_Z(G(u))$ instead of $\tau_Z(u)$. Given that (7) implies (6), we shall refer to model (6) as the generalized accelerated recurrence time (GART) model.

The proposed recurrent events model has considerable practical appeal. It models time to expected frequency, $\tau_Z(u)$, which may be viewed as the inverse of the mean function of recurrent events. Consequently, like means/rates based recurrent events models, the proposed model is intuitive to interpret and does not require specification of dependency through event history. While current means/rates based models mostly target the frequency scale of the mean function, the proposed model formulates covariate effects on the time scale of the mean function, and thus permits "direct physical interpretations" (Reid 1994). Such a feature may be preferred by many practitioners.

Furthermore, as elaborated in Sections 3.2–3.4, we uncover a general theoretical and inferential framework that can unify the studies of the GART model with recurrent events data and the quantile regression model (1) with randomly censored nonrecurrent event data. Thus, it is highly feasible that current software for censored quantile regression may be extended to carry out the proposed method for recurrent events data.

As another useful contribution of this work, based on our theoretical studies, we present in Section 3.4 a sample-based procedure to estimate the asymptotic covariance of the proposed estimator. According to our simulations, the new covariance estimation procedure can save two-to-three fold computation time as compared to the prevailing bootstrapping-based inference. This feature is particularly attractive for analyzing large datasets, such as the registry data from CFFPR.

3.2 The proposed estimation procedure

In the special case where $L_i = 0$ ($i = 1, \ldots, n$) and $G(u) = u$, Huang and Peng (2009) proposed an estimator of $\beta_0(u)$ as the minimizer of the following objective function:

$$\Psi\left(\boldsymbol{\beta};\mathbf{u}\right) = n^{-1}\sum_{i=1}^{n}\left\{\sum_{j=1}^{\infty} I\left(T_i^{(j)} < R_i\right)\left(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}\wedge\log\mathbf{R_i} - \log\mathbf{T_i^{(j)}}\right)^{+} - \left(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}\wedge\log\mathbf{R_i}\right)u\right\}.$$

This estimator is a generalization of Powell (1984, 1986)'s estimator for censored quantile regression when the censoring time is always known. It is easy to see that the objective function, $\Psi(\beta; u)$, is not convex. Huang and Peng (2009) developed an algorithm which is guaranteed to find a local strict minimizer that is asymptotically equivalent to the global minimizer of the proposed objective function.

While it is rather intuitive to modify Powell (1984, 1986)'s estimator to handle double censoring to an event time (Fitzenberger 1997) based on the equivariance property of quantiles (Koenker 2005), it is quite challenging to apply the same strategy to extend Huang and Peng (2009)'s method to window observed recurrent events data with nonzero L_i's. This is because the presence of nonzero L_i's can greatly increase the amount of missing information on the underlying counting process of recurrent events,$\tilde{N}_{i(t)}$. When $L_i = 0$, $T_i^{(j)}$ is observed if $T_i^{(j)} \leq R_i$; otherwise $T_i^{(j)}$ is known to be greater than or equal to R_i. This means$\tilde{N}_{i(t)}$ is partially observable. However, with $L_i > 0$, some recurrent events may have occurred before time L_i, but when such events have occurred, and how many are generally unknown. Therefore, one always lacks definitive information on $T_i^{(1)}$, and a similar ambiguity exists for any $T^{(j)}$ with $j \geq 2$. As a result, $\tilde{N}_{i(t)}$ is never observable when $L_i > 0$. Such a significant information loss on$\tilde{N}_{i(t)}$ given $L_i > 0$ is the main factor that prevents a straightforward adaptation of Huang and Peng (2009)'s method to handle window-observed recurrent events data.

Our estimating equation for $\beta_0(u)$ is directly motivated by the counting process based model representation (6). Specifically, by (6), we can readily identify an observable stochastic process, $M\left(u\right) \equiv N\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(u)}\right) - \int_0^u Y\left(e^{\boldsymbol{X}^{\top}\boldsymbol{\beta}_0(s)}\right) g\left(s\right) ds$, such that $E\{M(u)|\boldsymbol{X}\} = 0$ for $u > 0$. Here and hereafter, $N(\cdot)$ and $Y(\cdot)$ stand for the counting process and the at-risk process adopted in our recurrent events setting, namely $N^{re}(\cdot)$ and $Y^{re}(\cdot)$ respectively. Note that $M(u)$ is not a martingale in general but possesses the same utility as a martingale to construct an estimating equation for $\beta_0(\cdot)$.

More specifically, we propose to estimate $\beta_0(u)$ based on the estimating equation:

$$n^{1/2}\boldsymbol{S}_n\left(\boldsymbol{\beta},\mathbf{u}\right) = 0, \quad (8)$$

where

$$\boldsymbol{S}_n\left(\boldsymbol{\beta},\mathbf{u}\right)=n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\left\{N_i\left(exp\left\{\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}\left(\mathbf{u}\right)\right\}\right)-\int_0^u Y_i\left(exp\left\{\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}\left(\mathbf{s}\right)\right\}\right)g\left(s\right)ds\right\}.$$

In the traditional nonrecurrent event setting with $N_i(\cdot)$ and $Y_i(\cdot)$ replaced by $N_i^{nr}\left(\cdot\right)$ and $Y_i^{nr}\left(\cdot\right)$ respectively, equation (8) boils down to the estimating equation proposed in Peng and Huang (2008) for censored quantile regression. This connection shows that equation (8) unifies the existing estimation for censored quantile regression and the proposed estimation for the new recurrent events model (6).

We can adopt a grid-based algorithm for estimating $\beta_0(\cdot)$ based on equation (8). More specifically, define a grid, $S_{L(n)} = \{0 = u_0 < u_1 < . . . < u_{L(n)} = U\}$. The size of $S_{L(n)}$ is denoted by $\|S_{L(n)}\| \equiv \max_{j=1,...,L} | u_j - u_{j-1}|$. Our proposed estimator, $\hat{\boldsymbol{\beta}}\left(\cdot\right)$, is a right-continuous piecewise-constant function that jumps only at the grid $S_{L(n)}$. Given that $\tau_Z(0) = \exp\{\boldsymbol{X}^{\mathsf{T}}\beta_0(0)\} = 0$, we set $exp\left\{\boldsymbol{X}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}\left(0\right)\right\}=0$ for all *i*. We propose to obtain $\hat{\boldsymbol{\beta}}\left(u_k\right)$, *k* - 1, 2, ..., *L*(*n*), by sequentially solving the estimating equation,

$$n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i\left\{\tilde{N}_i\left(exp\left\{\boldsymbol{X}_i^T\boldsymbol{\beta}\left(\mathbf{u_k}\right)\right\}\right)-\sum_{m=0}^{k-1}Y_i\left(exp\left\{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}\left(u_m\right)\right\}\right)\int_{u_m}^{u_{m+1}}g\left(s\right)ds\right\}=0, \quad (9)$$

for $\beta(u_k)$.

Note that equation (9) is not continuous; thus an exact solution that makes (9) strictly hold may not exist. Therefore, $\hat{\boldsymbol{\beta}}\left(u_k\right)$ is defined as a generalized solution to equation (9) (Fygenson and Ritov 1994). Because equation (9) is monotone, the set of generalized solutions is convex of diameter $O(n^{-1})$ (Fygenson and Ritov 1994). To find a generalized solution to equation (9), an equivalent alternative approach is to locate the minimizer of the L_1-type convex function,

$$\begin{aligned}l_k\left(\boldsymbol{h}\right)=&\sum_{i=1}^{n}\sum_{j=1}^{\infty}I\left(L_i\le T_i^{(j)}\le R_i\right)\left|log\,T_i^{(j)}-\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{h}\right|\\&+\left|R^*-\left\{\sum_{i=1}^{n}\sum_{j=1}^{\infty}I\left(L_i\le T_i^{(j)}\le R_i\right)\left(-\boldsymbol{X}_i\right)^{\mathsf{T}}\boldsymbol{h}\right\}\right|\\&+\left|R^*-\delta\left\{\sum_{i=1}^{n}2\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{h}\sum_{m=0}^{k-1}Y_i\left(exp\left\{\boldsymbol{X}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}\left(u_m\right)\right\}\right)\int_{u_m}^{u_{m+1}}g\left(s\right)ds\right\}\right|,\end{aligned}$$

where R^* is a very large number and $j = 1, . . . , L(n)$. Following arguments similar to those in the Appendix of Peng and Fine (2009), we can show that $\partial l_k(\beta(u_k))/\partial\beta(u_k)$ equals 2 times the estimating function in (9) when R^* is chosen large enough to bound $|\left\{\sum_{i=1}^{n}\sum_{j=1}^{\infty}I\left(L_i\le T_i^{(j)}\le R_i\right)\left(-\boldsymbol{X}_i\right)^{\mathsf{T}}\boldsymbol{h}\right\}|$ and

$|\left\{\sum_{i=1}^{n} 2\boldsymbol{X}_i^\top \boldsymbol{h} \sum_{m=0}^{k-1} Y_i \left(exp\left\{\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}(u_m)\right\}\right) \int_{u_m}^{u_{m+1}} g(s)\,ds\right\}|$. This to equation then justifies the use of the minimizer of $l_k(\boldsymbol{h})$ as a generalized solution(9).

One can solve the minimization of $l_k(h)$ by using standard statistical software–for example, the *l*1*fit*() function in S-PLUS or the *rq*() function in R package *quantreg*. As shown in our theoretical studies, a grid size of order $o(n^{-1/2})$ would be su cient for our grid-based estimator to have desirable asymptotic properties, including uniform consistency and weak convergence to a Gaussian process. In our numerical studies, our choice of $//S_{L(n)}//$ is of order $O(n^{-1})$, by which we achieve good empirical results on both estimation accuracy and computational feasibility. A grid-free algorithm may be developed by adapting the strategy of Huang (2010).

3.3 Asymptotic properties

We establish the uniform consistency and weak convergence of the proposed estimator $\hat{\boldsymbol{\beta}}(\cdot)$. Define $\boldsymbol{A}(\boldsymbol{b}) = E\{\boldsymbol{X}N(\exp(\boldsymbol{X}^\mathrm{T}\boldsymbol{b}))\}$, $\boldsymbol{B}(\boldsymbol{b}) = d\mathbf{A}(b)/d\mathbf{b}^\mathrm{T}$, $\tilde{A}\,\boldsymbol{b}) = E\{\boldsymbol{X}\,Y(\exp(\boldsymbol{X}^\mathrm{T}\boldsymbol{b}))\}$, $\boldsymbol{J}(b) = d\tilde{\mathbf{A}}(\boldsymbol{b})/d\mathbf{b}^\mathrm{T}$, $\tilde{\mu}_Z(x) = E\{N(x)|\boldsymbol{Z}\}$, $g_{\boldsymbol{Z}}(x) = d\tilde{\mu}_{\boldsymbol{Z}}(x)/dx$, and $\dot{\mu}_{\boldsymbol{Z}}(x) = d\mu_{\boldsymbol{Z}}(x)/dx$. For a vector v, define $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^\top$. We let $f_{\boldsymbol{Z}}^L(x)$ and $f_{\boldsymbol{Z}}^R(x)$ denote the conditional density functions of L and R given $\boldsymbol{Z}$ respectively. Simple algebra can show that $\boldsymbol{B}(\boldsymbol{b}) = E\left\{\boldsymbol{X}^{\otimes 2} e^{\boldsymbol{X}^\top b} g_{\boldsymbol{Z}}\left(e^{\boldsymbol{X}^\top b}\right)\right\}$ and $\boldsymbol{J}(\boldsymbol{b}) = E \quad \boldsymbol{X}^{\otimes 2} e^{\boldsymbol{X}^\top b}\left\{f_{\boldsymbol{Z}}^L\left(e^{\boldsymbol{X}^\top b}\right) - f_{\boldsymbol{Z}}^R\left(e^{\boldsymbol{X}^\top b}\right)\right\}]$.

We assume the following regularity conditions:

C1: $\boldsymbol{Z}$ and $N(R)$ are bounded.

C2: Each component of $\boldsymbol{A}(\beta_0(u))$ is a Lipschitz function of u.

C3: (a) $g_Z(\exp(\boldsymbol{X}^\mathrm{T}\boldsymbol{b})) > 0$ for any $\boldsymbol{b} \in \mathscr{B}(d_0)$ and $\boldsymbol{Z} \in \mathscr{Z}$, (b) $E\left(\boldsymbol{Z}^{\otimes 2}\right)$ is positive definite, and (c) each component of $\boldsymbol{J}(\boldsymbol{b})\boldsymbol{B}(b)^{-1}$ is uniformly bounded in $\mathbf{b} \in \mathscr{B}(d_0)$, where $\mathscr{B}(d_0)$ is a neighborhood containing $\{\beta_0(u), u \in (0, U]\}$, defined in Supplement B of the Supplementary Materials.

C4: $\inf_{u \in [v, U]}$ *eigmin*$\boldsymbol{B}(\beta_0(u)) > 0$ for any $v\,2\,(0, U]$, where *eigmin*$(\cdot)$ denotes the minimum eigenvalue of a matrix.

Condition C1 assumes bounded covariates and a bounded total number of events observed during follow-up, and condition C2 implies the smoothness of $\beta_0(u)$. Condition C3 asserts additional mild assumptions, such as positive rate function of $\tilde{N}(t)$ and positive definite $E\left(\boldsymbol{Z}^{\otimes 2}\right)$. Condition C4 is a key technical assumption which ensures the consistency of the proposed estimator. By the definition of $\boldsymbol{B}(\boldsymbol{b})$, condition C4 implies $g_Z(\exp\{\boldsymbol{X}^\mathrm{T}{}_0(u)\}) > 0$ for $0 < u < U$. This means the recurrent event should have a positive rate of being observed throughout the time interval $(0, \tau_Z(U))$ under the random window observation scheme. This would generally require that the lower bound of the support of $f_{\boldsymbol{Z}}^L(\cdot)$ equals 0 and the upper bound of the support of $f_{\boldsymbol{Z}}^R(\cdot)$ exceeds $\tau_Z(U)$.

We can establish the following theorems:

Theorem 1. Under conditions C1-C4, if $\lim_{n\to\infty} \|S_{L(n)}\| = 0$, then $sup_{u\in[v,U]} \|\hat{\beta}(u) - \beta_0(u)\| \to_p 0$, where $0 < v < U$.

Theorem 2. Under conditions C1-C4, if $\lim_{n\to\infty} n^{1/2}\|S_{L(n)}\| = 0$, then $n^{1/2}\left\{\hat{\beta}(u) - \beta_0(u)\right\}$ converges weakly to a Gaussian process for $u \in [v, U]$, where $0 < v < U$.

The proofs of Theorems 1–2 follow the same set of lines that Peng and Huang (2008) adopted for censored quantile regression. This is because of the uniformity in the estimation framework demonstrated in Section 3.2; both estimation methods are closely tied to a Volterra integral equation, which, in this work, takes the form,

$$\boldsymbol{A}(\boldsymbol{\beta}(\mathbf{u})) - \int_0^u \tilde{\boldsymbol{A}}(\boldsymbol{\beta}(\mathbf{s}))\, g(s)\, ds = 0.$$

The proposed sequential estimation procedure is essentially the first-order Euler scheme to the numerical solution to the above equation. The proofs are relegated to Supplement B of the Supplementary Materials.

3.4 Inference

To make inference on $\beta_0(u)$, bootstrapping procedures can be used. For example, one may adopt a resampling procedure along the lines of Jin, Ying, and Wei (2001) by considering a perturbed estimating equation,

$$n^{-1/2}\sum_{i=1}^{n} v_i \boldsymbol{X}_i \left\{N_i\left(e^{\boldsymbol{X}_i^T\boldsymbol{\beta}(\mathbf{u})}\right) - \int_0^u Y_i\left(e^{\boldsymbol{X}_i^T\boldsymbol{\beta}(\mathbf{s})}\right) g(s)\, ds\right\} = 0,$$

where $\{v_i\}_{i=1}^n$ are i.i.d. variates from a nonnegative distribution of unit mean and unit variance, such as an exponential distribution of unit rate. The above stochastic integral equation can be solved via a procedure similar to that proposed for equation (8). Denote the resulting solution by $\hat{\beta}^*(\cdot)$. It can be shown that the distribution of $n^{1/2}\left\{\hat{\beta}^*(u) - \hat{\beta}(u)\right\}$ conditionally on the observed data and the unconditional distribution of $n^{1/2}\left\{\hat{\beta}(u) - \beta_0(u)\right\}$ have the same limiting distribution. By repeatedly generating $\{v_i\}_{i=1}^n$, one may obtain a large number of realizations of $n^{1/2}\left\{\hat{\beta}^*(u) - \hat{\beta}(u)\right\}$, the empirical distribution of which can be used to give a covariance estimate for $\hat{\beta}(u)$ or a confidence interval for $\beta_0(u)$.

We develop a sample-based approach for covariance estimation that does not involve bootstrapping and thus can save considerable computation time. The key idea is to find consistent estimates for $\boldsymbol{B}(\beta_0(\tau))$ and $\boldsymbol{J}(\beta_0(\tau))$ and then plug them into the closed form derived for the asymptotic covariance matrix of $n^{1/2}\left\{\hat{\beta}(u) - \beta_0(u)\right\}$; see equation (B.6) in Supplement B of the Supplementary Materials. However, directly evaluating $\boldsymbol{B}(\beta_0(\tau))$ and

$\boldsymbol{J}(\beta_0(\tau))$ based on their definitions would involve density estimation, which may not be stable nor efficient with small to moderate sample sizes.

To avoid density estimation, we propose a novel adaptation of the technique of Huang (2002) and Peng and Fine (2009). Note that neither Huang (2002) nor Peng and Fine (2009) handles stochastic integral estimating equations that are involved in our estimation setting. For the proposed estimator, we need to properly design working estimating equations so that they can be theoretically justified and stably solved.

Define $\boldsymbol{L}_n(\boldsymbol{b}) = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i N_i\left(e^{\boldsymbol{X}_i^\top b}\right)$, $\tilde{\boldsymbol{L}}_n(\boldsymbol{b}) = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i Y_i\left(e^{\boldsymbol{X}_i^\top b}\right)$, $\boldsymbol{\iota}_j(u) = \boldsymbol{X}_j N_j\left(e^{\boldsymbol{X}_j^\top \hat{\beta}(u)}\right)$, and $\Omega_n(u) = n^{-1}\sum_{j=1}^{n}\{\iota_j(u)\}^{\otimes 2}$. The procedure for estimating $\boldsymbol{B}(\beta_0(u))$ and $\boldsymbol{J}(\beta_0(u))$ follows.

1. Find a symmetric and nonsingular $(p+1)\times(p+1)$ matrix $\boldsymbol{E}_n(u)$ $\{e_{n,1}(u), \ldots, e_{n,p+1}(u)\}$ such that $\Omega_n(u) = \{\boldsymbol{E}_n(u)\}^2$.

2. Solve the equation,

$$\boldsymbol{L}_n(\boldsymbol{b}) = \boldsymbol{L}_n\left(\hat{\boldsymbol{\beta}}(u)\right) + e_{n,j}(u), \quad (10)$$

   for $\boldsymbol{b}$, and denote the solution by $\boldsymbol{b}_{n,j}(u)$ $(j = 1, \ldots, p + 1)$.

3. Calculate $\boldsymbol{D}_n(u) \equiv \left(\boldsymbol{b}_{n,1}(u) - \hat{\beta}(u), \ldots, \boldsymbol{b}_{n,p+1}(u) - \hat{\beta}(u)\right)$, and $\tilde{\boldsymbol{E}}_n(u) \equiv \left(\tilde{\boldsymbol{L}}_n(\boldsymbol{b}_{n,1}(u)) - \tilde{\boldsymbol{L}}_n\left(\hat{\beta}(u)\right), \ldots, \tilde{\boldsymbol{L}}_n(\boldsymbol{b}_{n,p+1}(u)) - \tilde{\boldsymbol{L}}_n\left(\hat{\beta}(u)\right)\right)$.

4. Compute $n^{-1/2}\boldsymbol{E}_n(u)\boldsymbol{D}_n(u)^{-1}$ and $n^{-1/2}\,\hat{\mathbf{E}}_{n(u)\boldsymbol{D}n(u)}{}^{-1}$, which provide consistent estimates for $\boldsymbol{B}(\beta_0(u))$ and $\boldsymbol{J}(\beta_0(u))$, respectively.

In Step 2, we adopt a working estimating equation, $\boldsymbol{L}_n(\boldsymbol{b}) = \boldsymbol{L}_n\left(\hat{\beta}(u)\right) + \boldsymbol{e}_{n,j}(u)$, which is monotone and can be solved via L_1–minimization. The key motivation for considering this estimating equation is the asymptotic linearity associated with $\boldsymbol{L}_n(\boldsymbol{b})$ in the neighborhood of $\beta_0(u)$, which can give $\boldsymbol{L}_n(\boldsymbol{b}_{n,j}(u)) - \boldsymbol{L}_n\left(\hat{\beta}(u)\right) \approx \boldsymbol{B}(\beta_0(u))\, n^{1/2}\left\{\boldsymbol{b}_{n,j}(u) - \hat{\beta}(u)\right\}$, where $\approx$ denotes asymptotic equivalence uniformly in $u \in [\nu, U]$. Likewise, we have $\tilde{\boldsymbol{L}}_n(\boldsymbol{b}_{n,j}(u)) - \tilde{\boldsymbol{L}}_n\left(\hat{\beta}(u)\right) \approx \boldsymbol{J}(\beta_0(u))\, n^{1/2}\left\{\boldsymbol{b}_{n,j}(u) - \hat{\beta}(u)\right\}$. These results are key for justifying the proposed estimation of $\boldsymbol{B}(\beta_0(u))$ and $\boldsymbol{J}(\beta_0(u))$. More detailed justifications are given in Supplement C of the Supplementary Materials.

Remark 1. The proposed procedure for obtaining $e_{n,j}(u)$ ensures that $e_{n,j}(u)$ $(j = 1, \ldots, p + 1)$ have the desired asymptotic order. The working estimating equation (10) would remain valid when one replaces $e_{n,j}(u)$ by $\gamma_c \cdot e_{n,j}(u)$, where γ_c is a constant. This fact adds flexibility to the implementation of the proposed covariance estimation procedure.

Denote the estimators of $\boldsymbol{B}(\beta_0(u))$ and $J(\beta_0(u))$ by $\hat{\mathbf{B}}(u)$ and $\hat{\mathbf{J}}(u)$ respectively. A consistent sample-based covariance estimator may be given by

$$n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\zeta}}_i(s)\,\hat{\boldsymbol{\zeta}}_i(t),$$

where $\hat{\boldsymbol{\zeta}}_i(t) = \hat{\boldsymbol{B}}(t)^{-1}\hat{\boldsymbol{\phi}}\left(\hat{\boldsymbol{\xi}}_i\right)$, $\hat{\boldsymbol{\phi}}(\cdot)$ is the plug-in estimate for the operator $\varphi(\cdot)$ defined in Supplement B of the Supplementary Materials, and

$$\hat{\boldsymbol{\xi}}_i(u) = \boldsymbol{X}_i\left\{N_i\left(e^{\boldsymbol{X}_i^{\top}\hat{\boldsymbol{\beta}}(u)}\right) - \int_0^u Y_i\left(e^{\boldsymbol{X}_i^{\top}\hat{\boldsymbol{\beta}}(s)}\right)g(s)\,ds\right\}, \quad i=1,\ldots,n.$$

Our simulation studies suggest that the proposed sample-based covariance estimator works quite well with small to moderate sample sizes.

3.5 Second-stage inference

One practical benefit of using the GART model to analyze recurrent events data is that it allows for accommodating and exploring varying covariate effects. Second-stage inference can be employed to serve this need. Given $\hat{\boldsymbol{\beta}}(\tau)$ for a range of τ's, it is often of interest to summarize the information provided by these estimators to help understand the underlying effect mechanism, and to determine whether some covariates have constant effects, so that a simpler model may be considered.

A summary of covariate effects can be generally formulated as some functional of $\beta_0(\cdot)$, denoted by (Ψ_0). A natural estimator of (Ψ_0) is $\boldsymbol{\Psi}\left(\hat{\boldsymbol{\beta}}\right)$. This estimator may be justified by using the functional delta method provided that $\Psi(\cdot)$ is compactly differentiable at β_0 (Andersen, Borgan, Gill, and Keiding 1998). The detailed inference on $\Psi(\beta_0)$ can follow the discussions in Section 2.4 of Peng and Fine (2009).

To test the constancy of a covariate effect, it is equivalent to consider the null hypothesis that takes the form, $H_{0,j}\colon \beta_0^{(j)}(u) = \rho_0, u \in [u_L, u_U]$, where the superscript $^{(j)}$ indicates the jth component of a vector, and ρ_0 is an unspecified constant ($j = 2, \ldots,$ or $p + 1$). Appropriate test statistics for $H_{0,j}$ can be developed along similar lines of Peng and Huang (2008) and Peng and Fine (2009). Accepting $H_{0,j}$ for all $j \in \{2, \ldots, p + 1\}$ may indicate the adequacy of a constant effect model. Therefore, such a procedure may be used to test the goodness-of-fit of an accelerated failure time model for recurrent events.

4. SIMULATION STUDIES

We conducted Monte Carlo simulations to assess the finite-sample performance of the proposed method. Like in Huang and Peng (2009), a Gamma frailty on a standard homogeneous Poisson process was applied to generate recurrent event times. Two covariates, Z_1 and Z_2, were considered, following the distributions, *Bernoulli*(0.5) and *Unif*(−0.5, 0.5), respectively. The recurrent event time sequence was generated by

$$T^{(j)}=exp\left\{min\left(1,\frac{T^{*(j)}}{1.5\gamma}\right)Z_1+Z_2\right\}T^{*(j)}/\gamma,\quad j=1,2,\ldots,$$

where $\{T^{*(j)}, j = 1, 2, \ldots\}$ was a recurrent event sequence from a standard homogeneous Poisson process and the frailty γ followed a Gamma distribution. The variance of γ, denoted by σ^2, determines the level of intra-individual correlation, and was chosen to be 0 or 0.5. It can be shown that, under our simulation setup,

$$\tau_{\boldsymbol{z}}\left(u\right)=exp\left\{log\left(u\right)+min\left(1,u/1.5\right)Z_1+Z_2\right\}.$$

Covariate Z_2 has a constant effect on $\tau_Z(u)$, while the effect of Z_1 increases with expected frequency. We generated L from $\omega \cdot Unif(0, 1)$ and R from $Unif(L, 12)$, where ω was a $Bernoulli(0.8)$ variate. Under these simulation set-ups, the average number of observed recurrent events per subject was about 4.0. With each selection of σ^2, we generated 500 datasets of sample size $n = 100$. For bootstrapping-based inference, the resampling size of 100 was chosen. We set $g(u) = 1$ and adopted an equally spaced grid on $u \in (0, 3]$ with step size, 0.02.

In Figures 1 and 2, we present the simulation results from the set-up with $\sigma^2 = 0$ and the set-up with $\sigma^2 = 0.5$, respectively. In the first row, we plot the empirical bias of the proposed estimator $\hat{\boldsymbol{\beta}}\left(u\right)$ (solid lines) and the empirical bias of Huang and Peng (2009)'s estimator (dashed lines), which naively assumes $L = 0$, versus expected frequency u. It is shown that our estimates have small bias except for those corresponding to small u's. Treating all L's as 0 clearly produces very biased coefficient estimation. The plots in the second row depict the empirical mean squared errors (MSE) versus expected frequency u. The empirical MSEs indicate reasonable efficiency of the proposed estimator with $n = 100$. In addition, the empirical MSE follows a similar pattern to the empirical bias. The observation of large bias and MSE at small u's is consistent with our theoretical results, which suggest substantial variability in estimating $\beta_0(u)$ as u is close to zero.

The last row of Figures 1 and 2 presents the coverage probabilities of 95% confidence intervals (CI) obtained from the proposed sample-based covariate estimates (solid lines) and those from bootstrapping (dashed lines). It shows that the bootstrapping-based procedure and the sample-based procedure have quite comparable performance. The resulting 95% CIs are slightly under-covered and yet have coverage probabilities fairly close to the nominal value. We examined the simulation results with a larger sample size, $n = 200$ (not reported here). We observe that the empirical bias and MSE decrease as the sample size increases. The coverage probabilities from either bootstrapping or sample-based procedure are much closer to the nominal value as compared to those with $n = 100$. We also evaluated the computation time taken to construct confidence intervals in each simulation setup. The computation time ratio of the sample-based approach to the bootstrapping procedure ranges from 0.16 to 0.47 with mean=0.24 and median=0.24, suggesting a significant saving in computation time by using the proposed sample-based procedure.

We also compared the estimation efficiency between the proposed estimator and Huang and Peng (2009)'s estimator when observation windows always start from 0. In these simulations, we set $L = 0$ while keeping $(T^{(j)}, R, \boldsymbol{Z})$ generated the same way. In Figure 3, we plot the relative efficiency of $\hat{\boldsymbol{\beta}}(u)$ to Huang and Peng (2009)'s estimator. It is shown that our estimator for the ART model is always more efficient than the method of Huang and Peng (2009). The efficiency gain seems to increase with the expected frequency and can be over 100% at some large u's. Such results may be explained by the fact that Huang and Peng (2009)'s estimator only employs the accelerated recurrence time assumption, (7), at a single u. Therefore, though Huang and Peng (2009)'s estimator is more robust than the proposed estimator, it does not make a full use of the global accelerated recurrence time assumption as the proposed estimator does, and thus is less efficient. The results in Figure 3 are consistent with Koenker (2008)'s findings from comparing the efficiency between Peng and Huang (2008)'s and Powell (1984, 1986)'s methods on censored quantile regression.

We also investigated a special case in which both the Cox regression model (Andersen and Gill 1982) and the GART model hold. Specifically, we generated a recurrent events dataset based on an GART model by letting $T^{(j)} = \exp(Z_1 + Z_2)T^{*(j)}$. Here $T^{*(j)}$, (Z_1, Z_2), and (L, R) are defined in the same way as those for the other simulations. In this case, the true intensity process is given by $\exp(-Z_1 - Z_2)$, and the time to expected frequency u equals $\exp(Z_1 + Z_2)u$. Figure 4 compares the estimates for $\tau_Z(u)$ with $\boldsymbol{Z} = (0, 0)$ based on the Cox regression model with those based on the GART model with $g(u) = 1$. The 95% pointwise confidence intervals are also presented. It is shown by Figure 4 that the estimates for time to expected frequency based on both models are quite similar; both curves are close to the true line of $\tau_Z(u)$. The confidence intervals derived based on the Cox regression model are generally tighter $\Psi(\beta_0)$ than those based on the GART model. The difference in confidence interval width is evident at small u's but gradually diminishes as u increases.

The observations from Figure 4 well match our expectation given the following facts: (a) the Cox regression model imposes the constancy assumption for each coefficient function, and thus its estimation can make good use of all observations; (b) the GART model allows for varying covariate effects, and the proposed estimation of $\beta_0(u)$ essentially utilizes only observations from time 0 to $\tau_Z(u)$. The stronger modeling assumption of Cox regression may contribute to the tighter confidence intervals. Since $\tau_Z(u)$ is increasing with u, the proposed estimates for $\tau_Z(u)$ under the GART model can use fuller data information at larger u's and then become more comparable to those obtained from the Cox regression model. This may explain the observed smaller differences in estimation accuracy at larger u's.

5. AN APPLICATION TO A CYSTIC FIBROSIS DATASET

Cystic Fibrosis (CF) is a life-limiting genetic disorder with an incidence of 1:3400 in Caucasians (Boat and Acton 2007). Pseudomonas aeruginosa (PA) is the most important pathogen that shortens the survival of CF patients. According to Cystic Fibrosis Foundation (CFF) 2011 annual report, it infects more than half of CF patients and often leads to chronic conditions. Characterizing the timing of PA infections and assessing how it is influenced by potential risk factors can help improve treatment decisions and are thus of scientific interest. To address these questions, we utilized the data from 2875 children documented in

1986-2008 CFF Patient Registry (CFFPR), who were born in or after 1998, had at least one F508del mutation, and had 5 or more years of follow-up in the registry.

We applied the proposed method to this CFFPR dataset, with the recurrent event time $T^{(j)}$ being the age of a CF child when he or she had the jth PA infection. While some CF children entered the registry right after birth given the availability of early diagnosis by newborn screening, others had delayed CFFPR entries due to later diagnosis of CF or other reasons. Observation of PA infections started from the first CFFPR visit, and thus the time from birth to registry entry constitutes the L in our method framework. In this dataset, age at the first CFFPR visit ranges from 0 to 5.7 years with mean=0.7 years and median=0.4 years. The number of positive PA cultures at CFFPR visits ranges from 0 to 50; the mean and median number of PA infections are 3.9 and 2.0, respectively. We considered risk factors including sex, patient's CFTR genotype (I=F508del homozygous, II=F508del heterozygous), meconium ileus (MI)status , and pancreatic sufficiency status (defined as never on pancreatic enzymes). The covariates for a subject are coded as *Female*, 1 if the subject was female and 0 otherwise; $F508/Other$, 1 if the subject was F508del heterozygous and 0 otherwise; *MI*, 1 if the subject was diagnosed by MI and 0 otherwise; and *Pancreat*, 1 if the subject was pancreatic sufficient and 0 otherwise.

We fit the proposed model to the CFFPR dataset with the covariates described above, setting $g(u) = 1$. In Figure 5, we plot the estimated coefficients along with the 95% point-wise confidence intervals. The intercept coefficient estimates (panel A of Figure 5) represent the estimated log time to expected frequency of PA infection for the reference group, which consisted of CF boys with homozygous F508del mutations who had no MI and were pancreatic insufficient. For example, the time from birth to expected PA infection frequency of 1.0 is approximately 1.7 years. An alternative interpretation of this result is that, at age of 1.7 years old, CF patients in the reference group are expected to have acquired one PA infection on average. Figure 5 (panel A) also suggests that the time to expected PA infection frequencies of 2.0 in the reference group is about 4.3 years.

The non-intercept coefficient estimates in Figure 5 (panels B–E) depict the estimated effects of covariates, which are allowed to be frequency-varying. Negative coefficient estimates indicate more rapid progression to recurrence of PA infections. To better summarize the varying covariate effect estimates, we present in Table 1 the average covariate effects in the frequency intervals (0.4, 1.4], (1.4, 2.4], and (0.4, 2.4] respectively, along with estimated standard errors and p values. The results in Table 1 suggest a strong disadvantage to CF children with pancreatic insufficiency, who tend to experience recurrent PA infections at earlier ages. Girls with CF appear to have marginal increased risk of recurrent PA infections only in the frequency interval (1.4, 2.4], which indicates later recurrence of PA infections. The average effect estimates for *MI* demonstrate a cross-over pattern, changing from −0.37 to 0.74, though not reaching statistical significance in either frequency interval. From Table 1, we find little difference in time to expected frequency between the F508del homozygous group and the F508del heterozygous group. This is consistent with literature that reported relatively weak associations between genotype and pulmonary phenotype in CF patients. We also conducted constancy tests for each covariate effect. The results given in Table 1 indicate that *MI* had a frequency-dependent effect on the timing of PA recurrence, while other covariates displayed fairly constant effects over the frequency of PA infections.

In Figure 5, we also plot the coefficient estimates obtained from applying Huang and Peng (2009)'s method (dash dotted lines). Huang and Peng (2009)'s estimator, which assumes that the PA observation window always starts from zero, may suggest a smaller difference caused by pancreatic insufficiency in time to expected frequency less than 1.0. Such a discrepancy may be expected because, when the lower bound of PA observation window is ignored, the frequency of PA infection before registry entry is naively taken as 0, though it may be positive. Intuitively, this would result in a frequency-lag in the estimated effect as a function of frequency. We also note that the intercepts estimated by Huang and Peng (2009)'s method are significantly larger than those from the proposed method. This observation conforms to the intuition that naively treating PA infection frequency before registry entry as 0 would lead to over-optimistic estimates for time to expected frequency.

6. REMARKS

In this work, we propose a new approach to model counting processes that are naturally embedded in event history data. The new counting process model can be transformed to a quantile regression model in the traditional survival setting with random censoring or a generalized accelerated recurrence time (GART) model in the recurrent events setting with random observation windows. We are able to generalize the current methodological framework for censored quantile regression (Peng and Huang 2008) to serve the broader class of survival regression models implied by the proposed counting process model. We expect that the existing software for censored quantile regression can be extended to implement proposed regression methods for recurrent events data.

Our proposals for the GART model offer a useful and flexible alternative to current approaches for analyzing recurrent events data. The new method is easy to interpret and implement. It is also straightforward to adapt the proposed recurrent events method to accommodate the realistic on-and-off observation scheme of recurrent events, corresponding to observation windows that take the form as a union of multiple disjoint time intervals, say $\cup_{j=1}^{K}(L_j, R_j]$. In this case, the at-risk process needs to be specified as $Y(t)=\sum_{j=1}^{K} I(L_j<t \le R_j)$.

As mentioned in Section 2, the specification of $g(\cdot)$ is tied to the scale in which the covariate effects are formulated. In practice, we anticipate that the preferable forms of $g(\cdot)$ would be the ones that make model (5) easily interpretable. For example, with the standard set-up for randomly censored data, popular choices of $g(\cdot)$ may include $g(u) = 1/(1 - u)$ and $g(u) = 1$, which lead to quantile regression modeling and modeling of the inverse cumulative hazard function, respectively. In the recurrent events setting, the most compelling choice of $g(\cdot)$ for model (6) may be $g(u) = 1$. Given the common stochastic structure implied by censored quantile regression model (1) and the proposed counting process model (5), we anticipate that model diagnostics for model (5) with a specified $g(\cdot)$ can be developed along the lines of Peng and Huang (2008).

The proposed method for window observed recurrent events implies an alterative quantile regression approach for doubly censored data that include n i.i.d. replicates of $\eta \equiv I(L < T \leq R)$, $W \equiv T\eta$, L and R, denoted by $\{\eta_i, W_i, L_i, R_i\}_{i=1}^{n}$. In this data setting, both left and right censoring variables, L and R, are always observed, and T is the event time of interest assumed to follow quantile regression model (1). To estimate model (1), one may use the proposed estimating equation (8) with $N_i(t)$ and $Y_i(t)$ replaced by $I(W_i \leq t, \eta_i = 1)$ and $I(L_i < t \leq R_i)$, respectively. Inference procedures may be carried out by adapting the lines in Section 3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by National Science foundation grant DMS-1007660 and National Institutes of Health grant R01 HL113548 (to Peng), National Science foundation grant DMS-1208874 and National Institutes of Health grant P30AI050409 and P20HL113451 (to Huang) and National Institutes of Health grant R01 DK072126 (to Lai). The authors thank Dr. Preston Campbell from the US Cystic Fibrosis Foundation for providing patient registry data.

REFERENCES

Andersen P, Gill R. Cox's regression model for counting processes. Annals of Statistics. 1982

Andersen, PK.; Borgan, e.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. 2nd ed.. Springer-Verlag.; New York: 1998.

Boat, T.; Acton, J. Cystic fibrosis. In: Kliegman, R.; Stanton, B.; Geme, J.; Schor, N., editors. Nelson Textbook of Pediatrics. 18th ed.. Saunders Elsevier; Philadelphia: 2007. p. 1803-1817.

Cai J, Prentice R. Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data. Biometrika. 1995; 82:151–164.

Cai J, Prentice R. Regression Estimation Using Multivariate Failure Time Data and a Common Baseline Hazard Function Model. Lifetime Data Analysis. 1997; 3:197–213. [PubMed: 9384652]

Chang S-H, Wang M-C. Condition Regression Analysis for Recurrence Time Data. Journal of American Statistical Association. 1999; 94:1221–1230.

Cox D. Regression models and life tables (with discussion). Journal of Royal Statistical Society, Ser. B. 1972; 34

Fitzenberger, B. Handbook of statistics. Vol. 15. Robust inference; North-Holland, Amsterdam: 1997.

Fygenson M, Ritov Y. Monotone estimating equations for censored data. Annals of Statistics. 1994; 22:732–746.

Hougaard, P. Analysis of Multivariate Survival Data. Springer-Verlag; New York: 2000.

Huang Y. Calibration Regression of Censored Lifetime Medical Cost. Journal of the American Statistical Association. 2002; 98:318–327.

Huang Y. Quantile Calculus and Censored Regression. The Annals of Statistics. 2010; 38:1607–1637. [PubMed: 20592942]

Huang Y, Chen Y. Marginal regression of gaps between recurrent events. Lifetime Data Analysis. 2003; 9:293–303. [PubMed: 14649847]

Huang Y, Peng L. Accelerated Recurrence Time Models. Scandinavian Journal of Statistics. 2009; 36:636–648.

Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. Biometrika. 2001; 88:381–390.

Koenker R. Regression Quantiles. 2005

Koenker R. Censored Quantile Regression Redux. Journal of Statistical Software. 2008; 27 http://www.jstatsoft.com.

Koenker R, Bassett G. Regression Quantiles. Econometrica. 1978; 46:33–50.

Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. Technometrics. 1995; 37:158–168.

Liang K-Y, Self S, Nanndeen-Roche K, Zeger S. Some Recent Developments for Regression Ananlysis of Multivariate Failure Time Data. Lifetime Data Analysis. 1995; 1:403–415. [PubMed: 9385112]

Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000; 62:711–730.

Lin DY, Wei LJ, Ying Z. Accelerated failure time models for counting process. Biometrika. 1998; 85:605–618.

Nelson, W. Recurrent Events Data Analysis for Product Repairs, Disease Recurrences and Other Applications. ASA-SIAM; Philadelphia: 2003.

Neocleous T, Vanden Branden K, Portnoy S. Correction to Censored Regression Quantiles by Portnoy, S. (2003), 1001–1012. Journal of American Statistical Association. 2006; 101:860–861.

Peng L, Fine J. Competing risks quantile regression. Journal of the American Statistical Association. 2009; 104:1440–1453.

Peng L, Huang Y. Survival Analysis with Quantile Regression Models. Journal of American Statistical Association. 2008; 103:637–649.

Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. Journal of the American Statistical Association. 1993; 88:811–820.

Portnoy S. Censored Regression Quantiles. Journal of American Statistical Association. 2003; 98:1001–1012.

Portnoy S, Lin G. Asymptotics for Censored Regression Quantiles. Journal of Nonparametric Statistics. 2010; 22:115–130.

Powell JL. Least absolute deviations estimation for the censored regression model. Journal of Econometrics. 1984; 25:303–325.

Powell JL. Censored regression quantiles. Journal of Econometrics. 1986; 32:143–155.

Prentice R, Williams B, Peterson A. On the Regression Analysis of Multivariate Failure Time Data. Biometrika. 1981; 68:373–379.

Reid N. A conversation with Sir David Cox. Statistical Science. 1994; 9:439–455.

Schaubel D, Cai J. Regression analysis for gap time hazard functions of sequentially ordered multivariate failure time data. Biometrika. 2004; 91:291–303.

Schaubel D, Zeng D, Cai J. A Semiparametric Additive Rates Model for Reccurrent Event Data. Lifetime Data Analysis. 2006; 12:389–406. [PubMed: 17031499]

Spiekerman C, Lin D. Marginal Regression Models for Multivariate Failure Time Data. Journal of American Statistical Association. 1998; 93:1164–1175.

Sun L, Park D, Sun J. The additive hazards model for recurrent gap times. Statistica Sinica. 2006; 16:919–923.

Sun L, Z. X. Guo S. Marginal regression models with time-varying coefficients for recurrent event data. Statistics in Medicine. 2011; 30:2265–2277. [PubMed: 21590791]

Wei L, Glidden D. An Overview of Statistical Methods for Multiple Failure Time Data in Clinical Trials. Statistics in Medicine. 1997; 16:833–839. [PubMed: 9160483]

Wei L, Lin D, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association. 1989; 84:1065–1073.

Yang S. Censored Median Regression Using Weighted Empirical Survival and Hazard Functions. J. Am. Statist. Assoc. 1999; 94:137–145.

Ying Z, Jung SH, Wei LJ. Survival Analysis with Median Regression Models. J. Am. Statist. Assoc. 1995; 90:178–184.

Zeng D, Lin D. Efficient estimation of semiparametric transformation models for counting processes. Biometrika. 2006; 93:627–640.

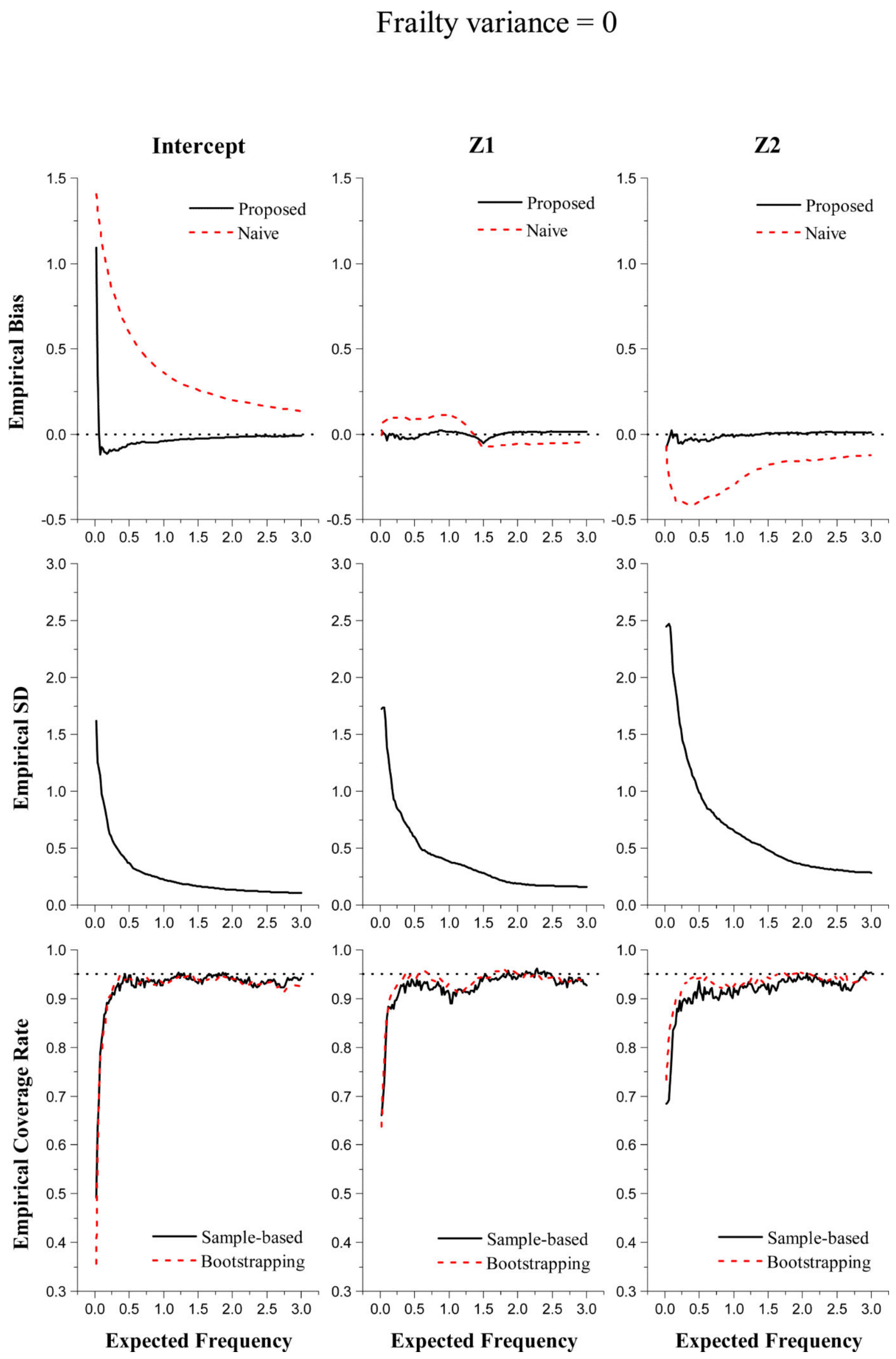

Figure 1.

Simulation results with sample size *n* = 100 for the set-up with Gamma frailty of variance 0

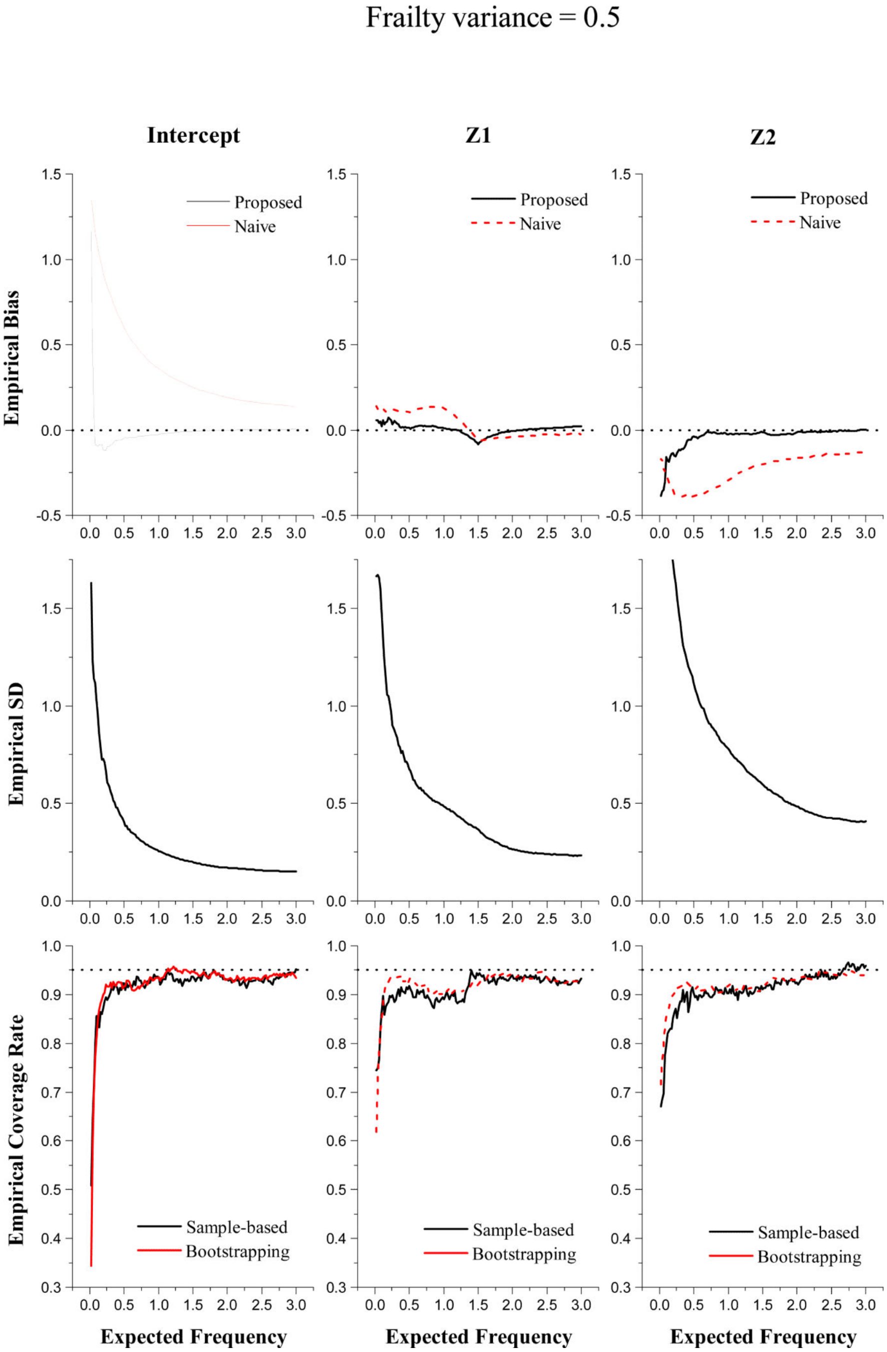

Figure 2.
Simulation results with sample size *n* = 100 for the set-up with Gamma frailty of 0.5

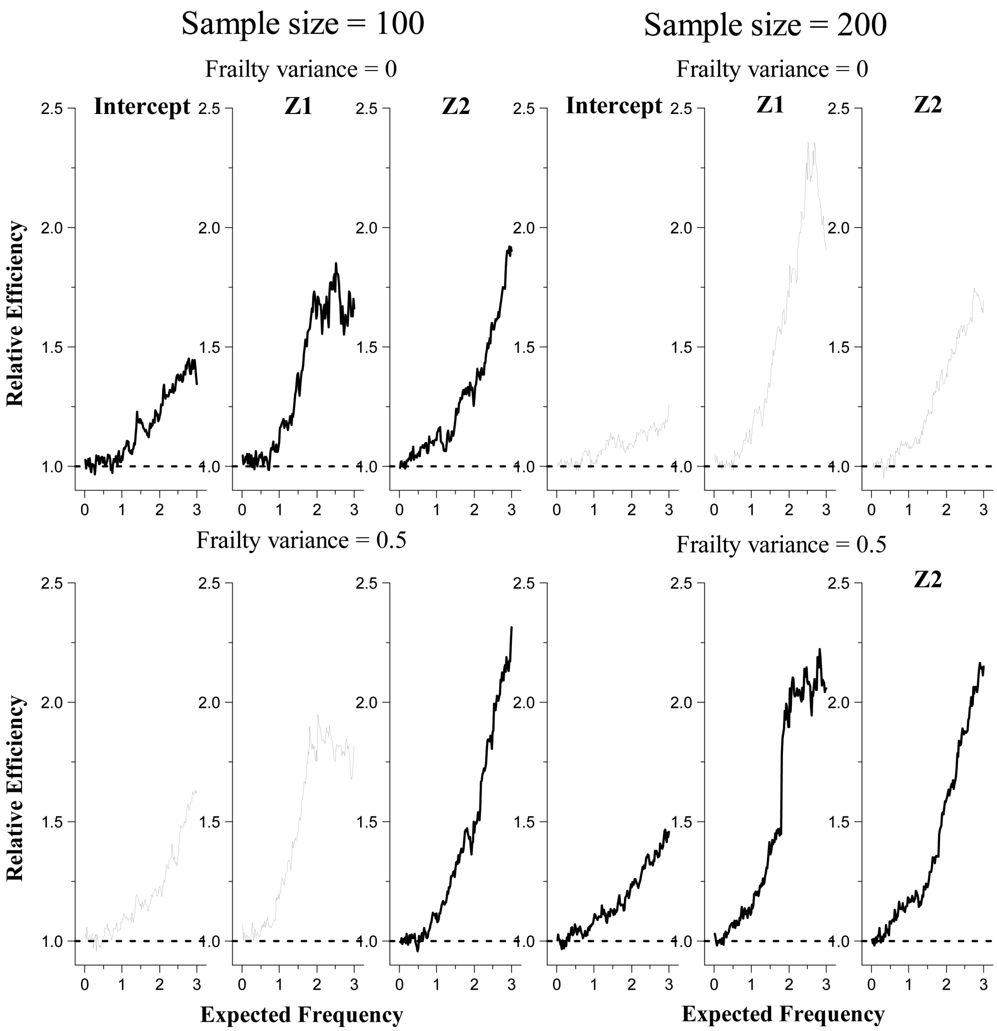

Figure 3.

Simulation results on the efficiency of the proposed estimator relative to Huang and Peng (2009)'s estimator.

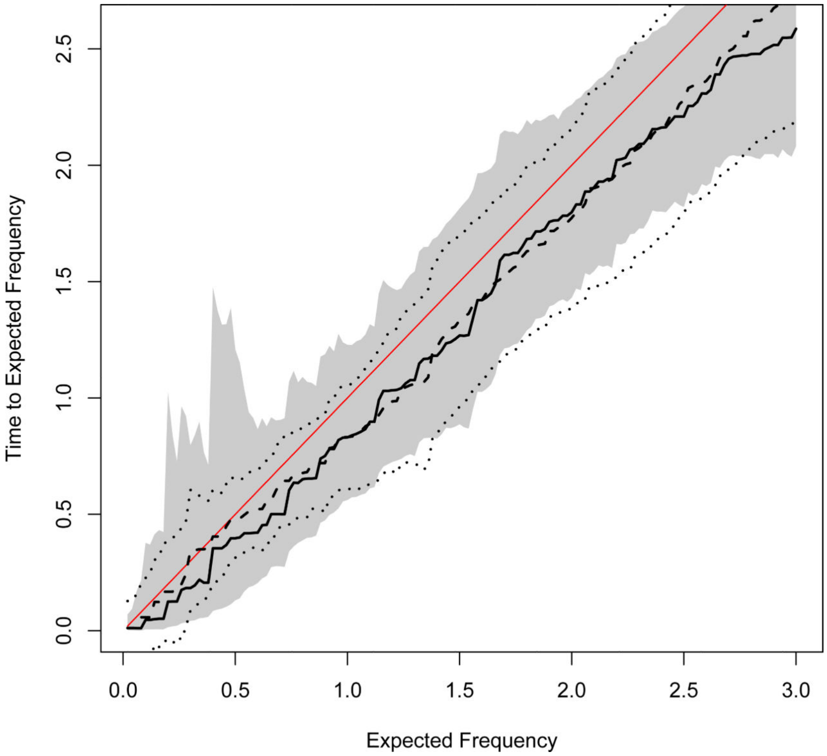

Figure 4.

Estimated $\tau_Z(u)$ curves with $\boldsymbol{Z} = (0, 0)$ based on Cox regression model (bolded dashed lines) and GART model (bolded solid lines) along with 95% pointwise confidence intervals based on Cox regression model (bolded dotted lines) and GART model (shaded area) when both Cox regression model and GART model hold. The true curve is presented in unbolded solid line.

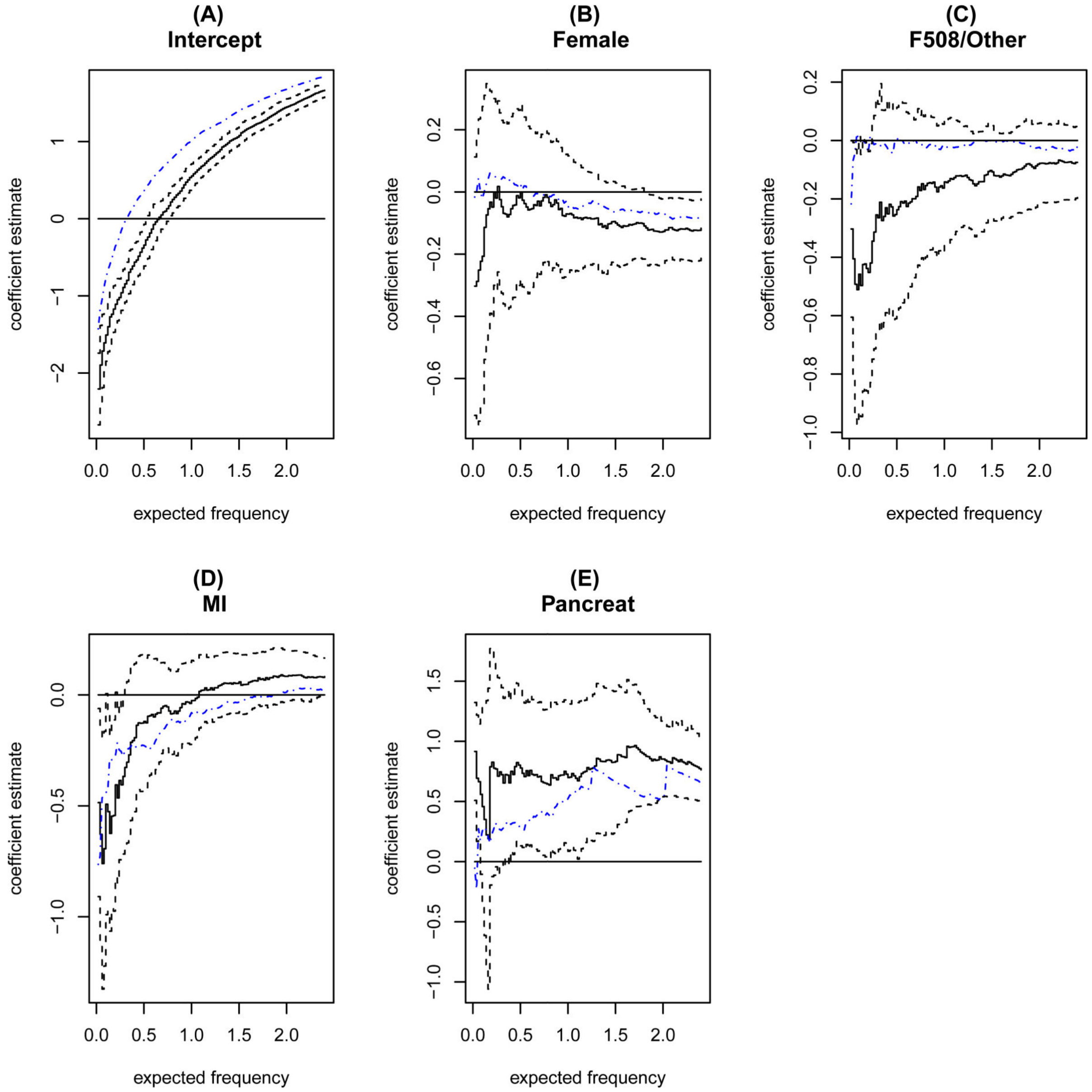

Figure 5.

CFFPR data example: coefficient estimates (solid lines) and 95% pointwise confidence intervals (dashed lines) from the proposed method and the coefficient estimates from Huang and Peng (2009)'s method (dash dotted lines).

Table 1

The CFFPR example: average covariate effect estimates along with standard errors (SE) and the corresponding p values, and the p values from constancy tests.

Average Covariate Effect Estimates					
Frequency Interval		**Sex**	***F*508/*Other***	***MI***	***Pancreat***
(0.4,1.4]	Estimate	–0.06	–0.17	–0.04	0.74
	SE	0.10	0.12	0.09	0.29
	P value	0.55	0.15	0.69	0.01
(1.4, 2.4]	Estimate	–0.11	–0.10	0.07	0.86
	SE	0.06	0.07	0.06	0.21
	P value	0.05	0.19	0.21	< 0.001
(0.4, 2.4]	Estimate	–0.09	–0.13	0.02	0.80
	SE	0.08	0.09	0.08	0.24
	P value	0.28	0.16	0.81	< 0. 001
Constancy Tests					
Constancy tests	P value	0.32	0.26	0.06	0.50