Statistica Sinica **27** (2017), 1079-1100
doi:https://doi.org/10.5705/ss.202014.0153

SEMIPARAMETRIC REGRESSION ANALYSIS OF REPEATED CURRENT STATUS DATA

Baosheng Liang[1], Xingwei Tong[1], Donglin Zeng[2] and Yuanjia Wang[3]

[1]*Beijing Normal University,* [2]*University of North Carolina at Chapel Hill and* [3]*Columbia University*

Abstract: In many clinical studies, patients may be asked to report their medication adherence, presence of side effects, substance use, and hospitalization information during the study period. However, the exact occurrence time of these recurrent events may not be available due to privacy protection, recall difficulty, or incomplete medical records. Instead, the only available information is whether the events of interest have occurred during the past period. In this paper, we call these incomplete recurrent events as repeated current status data. Currently, there are no valid standard methods for this kind of data. We propose to use the Andersen-Gill proportional intensity assumption to analyze such data. Specifically, we propose a maximum sieve likelihood approach for inference and we show that the proposed estimators for regression coefficients are consistent, asymptotically normal and attain semiparametric efficiency bounds. Simulation studies show that the proposed approach performs well with small sample sizes. Finally, our method is applied to study medication adherence in a clinical trial on non-psychotic major depressive disorder.

Key words and phrases: Andersen-Gill model, current status data, recurrent events, semiparametric efficiency, sieve estimation.

1. Introduction

During many clinical studies, patients may be asked to report their medication adherence, presence of side effects, substance use, and hospitalization information, which occur repeatedly over time. However, due to privacy issues, recall difficulty or incomplete survey questionnaires or medical records, the exact times and frequencies of these events may not be observed. Instead, the only available information is whether such an event has occurred or not in the period since the most recent visit. One motivating example of our work comes from studying medication adherence in a Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. The primary goal of this study was to evaluate feasible treatment strategies to improve clinical outcomes for patients

with treatment-resistant depression. In the study, those patients who failed to achieve remission or response for current treatments after a specified number of weeks were encouraged to enter the next treatment level (Gaynes et al. (2008)), because evidence showed that patients who achieved remission function better were less prone to relapse than those who only achieved partial improvement in symptoms (Sinyor, Schaffer and Levitt (2010)). Since some anti-depressant drugs tend to have strong side effects (Trivedi et al. (2006), Howland (2008)), many patients chose to leave the study or were likely to miss taking medication from time to time. Therefore, understanding factors of the medication adherence was important for evaluating the effectiveness of treatments and trial management. However, the exact frequency and the exact event times for medication non-adherence were not observed due to recall difficulty. Instead, patients were only asked to report non-adherence since their last medication visit, resulting in incomplete information which only indicated whether or not they missed taking prescribed medicine since the most recent visits.

We term these repeated events data with incomplete information as repeated current status data. They share similarity with traditional current status data in univariate survival analysis, but the events are repeatedly observed over periods within the same subject. One unique feature is that the exact frequency of events is not observed, but only the presence or absence of the event in each time period is known. Such data pose significant statistical challenges when the goal is to assess the association between risk factors and the true recurrent events. There are no valid standard methods to analyze the repeated current status data. One intuitive method is to treat the observed current status outcomes, i.e., whether at least one event occurred or not, as a sequence of binary outcomes collected longitudinally and to apply a generalized linear model or generalized linear mixed effect model (Liang and Zeger (1986), Liang, Zeger and Albert (1988)). However, this approach cannot provide an accurate estimate of the risk effects on the recurrent events of interest and is shown in our numerical studies to lead to substantial bias and incorrect inference. Alternative approaches for analyzing incomplete recurrent events data mainly focus on analyzing panel count data (e.g., Kalbfleisch and Lawlessa (1985), Sun and Kalbfleisch (1995), Sun and Wei (2000)), where the observed information consists of the total count of events up to a sequence of given time points. Based on a working model of a non-homogeneous Poisson process, Wellner and Zhang (2000, 2007) studied the panel count data without and with covariates based on some pseudo-likelihood functions; Zhu et al. (2014) analyzed a mixed recurrent-event and panel-count data using maximum

likelihood estimation procedure with a non-homogeneous Poisson process. These approaches are not applicable for analyzing the repeated current status where the total count of recurrent events is indeed not available.

In this paper, we propose an efficient method to analyze repeated current status data. We assume that recurrent events follow the proportional intensity model (Andersen and Gill (1982)). For inference, we propose a spline-based sieve maximum likelihood estimation approach. One significant difference of our method, as compared to the estimation approach for traditional current status data or interval censored data (e.g., Huang and Rossini (1997), Cai and Betensky (2003)), is that we need to account for the dependence among repeated current status outcomes from the same subject. Proving the invertibility of the information operator is necessary for establishing asymptotic distributions and semiparametric efficiency of the proposed estimators, should account for the distribution of the counting process for the visit time points.

The remainder of this paper is organized as follow. In Section 2, we approximate the baseline intensity function using a sequence of B-splines, then maximize the observed likelihood function for inference. The consistency and asymptotic efficiency of the parameter estimators are established in Section 3. Section 4 shows numerical results from extensive simulation studies and an application of the proposed method to analyze data from the STAR*D study.

2. Method and Inference Procedure

Let $N(t)$ denote the total number of recurrent events before time t, and X be the time-independent covariate that relates to $N(t)$. To study the effect of X on the recurrent events, given the covariate X, $N(t)$ is assumed to be a non-homogeneous Poisson process with intensity function

$$\lambda(t|X) = \lambda(t)\exp(X'\beta), \tag{2.1}$$

where $\beta \in R^p$ is a p-dimensional regression parameter of interest, and $\lambda(t)$ is an unspecified baseline hazard function.

Let $0 = T_{i0} < T_{i1} < \cdots < T_{iK_i}$ be the visit time points for subject i, where K_i is the total number of observation (or visit) time points for subject i and T_{iK_i} is the last time point (censoring time) when subject i is lost to follow-up. As these visit time points might depend on the covariates, we assume that they are independent of recurrent events given the covariates. For repeated current status data, we only observe whether any recurrent events have occurred in the interval $(T_{i,j-1}, T_{ij}]$ for $j = 1, \cdots, K_i$, not the number of events. We observe a

sequence of indicators $\Delta_{ij} = I(\delta N_{ij} > 0)$ with $\delta N_{ij} = \int_{T_{i,j-1}}^{T_{ij}} dN(s) = N(T_{ij}) - N(T_{i,j-1})$, $j = 1, \cdots, K_i$. Thus, the data from n i.i.d subjects consists of $O_i = \{T_{i1}, \cdots, T_{iK_i}, \Delta_{i1}, \cdots, \Delta_{iK_i}, X_i\}$, $i = 1, \cdots, n$. Under a Poisson assumption and intensity (2.1), the observed likelihood function is

$$L_n(\beta, \Lambda) = \prod_{i=1}^{n} \prod_{j=1}^{K_i} \left[\exp\left\{ -(\Lambda(T_{ij}) - \Lambda(T_{i,j-1})) \exp(X_i'\beta) \right\} \right]^{1-\Delta_{ij}} \times \left[1 - \exp\left\{ -(\Lambda(T_{ij}) - \Lambda(T_{i,j-1})) \exp(X_i'\beta) \right\} \right]^{\Delta_{ij}}, \tag{2.2}$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is the cumulative baseline hazard function. Our goal is to estimate (β, Λ) by maximizing $L_n(\beta, \Lambda)$ over the parameter space $\Theta = \mathcal{B} \times \mathcal{A}$, where $\mathcal{B}$ is a bounded open subset of R^p, and $\mathcal{A} = \{\Lambda : \Lambda$ is a non-decreasing function over $[0, \tau]\}$, where τ denotes the study duration.

We develop a spline-based sieve maximum likelihood estimation motivated by the penalized spline method of Cai and Betensky (2003). Specifically, we use B-spline functions to model the baseline hazard function $\lambda(t)$: let the set of spline knots be $0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < t_{m_n+l} < t_{m_n+l+1} = \cdots = t_{m_n+2l} = 1$, where m_n is knots number depending on n, and l is the order of B-spline. Let $\{B_i(\cdot),\ i = 1, \cdots, k_n = m_n + l\}$ be the B-spline basis functions corresponding to these knots (Schumaker (2007)). Then the log-transformed hazard function, $\log \lambda(\cdot)$, can be approximated by a linear combination of these B-spline functions, denoted as $B_n(t)'\alpha$, where $B_n(t) = (B_1(t), \cdots, B_{k_n}(t))'$ and $\alpha = (\alpha_1, \cdots, \alpha_{k_n})' \in R^{k_n}$. Equivalently, $\Lambda(t)$ can be approximated by $\int_0^t \exp(B_n(s)'\alpha)\, ds$, which guarantees that the approximation of $\Lambda(t)$ is both positive and non-decreasing for $t > 0$. Under this sieve approximation, the observed log-likelihood function can be written as

$$l_n(\beta, \alpha) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} (1 - \Delta_{ij}) \Big[- \int_{T_{i,j-1}}^{T_{ij}} \exp\left(B_n(s)'\alpha\right) ds \cdot e^{X_i'\beta} \Big] + \Delta_{ij} \log\Big(1 - \exp\Big\{ - \int_{T_{i,j-1}}^{T_{ij}} \exp\left(B_n(s)'\alpha\right) ds \cdot e^{X_i'\beta} \Big\}\Big). \tag{2.3}$$

To prevent unrealistically large values of α's, we impose some additional bound on the α's but allow the bound to increase as the sample size increases. It is then equivalent to restrict Λ to belong to the space

$$\mathcal{S}_n = \Big\{ \tilde{\Lambda}(\cdot) : \tilde{\Lambda}(t) = \int_0^t \exp(B_n(s)'\alpha) ds,\ t \in [0, \tau], \text{ where } \alpha \in R^{k_n} \text{ satisfies } \max_{j=1,\cdots,k_n} |\alpha_j| \le D_n \Big\},$$

where D_n is some pre-specified number increasing with n. Thus, we maximize $\log L_n(\beta, \Lambda)$ over the space $(\beta, \Lambda) \in \mathcal{B} \times \mathcal{S}_n$.

Many optimization algorithms can be implemented for the maximization. In numerical studies, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to solve this optimization problem. In our numerical studies, optimization is usually quick and robust to the choice of initial values.

We denote the maximum likelihood estimate by $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$, where $\hat{\Lambda}_n = \int_0^t \exp(B_n(s)'\hat{\alpha}_n)ds$ and $(\hat{\beta}_n, \hat{\alpha}_n)$ is the maximizer of (2.3). In the next section, we show that $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is asymptotically normal with covariance matrix Σ^{-1}, where β_0 is the true parameter. Thus, to make inference for β_0, we need to estimate Σ^{-1} consistently. Following Huang and Rossini (1997) and Murphy and van der Vaart (2000), we apply a sieve profile likelihood approach: if $pl_n(\beta)$ is the profile log-likelihood function for β, $pl_n(\beta) = \max_{\Lambda \in \mathcal{S}_n}\{l_n(\beta, \Lambda)\}$, then the (r, s)th element of Σ can be estimated by

$$-\frac{pl_n(\hat{\beta}_n + h_n e_r - h_n e_s) - pl_n(\hat{\beta}_n + h_n e_r) - pl_n(\hat{\beta}_n - h_n e_s) + pl_n(\hat{\beta}_n)}{nh_n^2},$$

where e_r and e_s are the rth and sth canonical vectors respectively, and h_n is some perturbation constant usually set to be $cn^{-1/2}$ for some positive constant c. Using the estimate for Σ, we can construct asymptotic confidence intervals for β_0 and perform Wald's test for the hypothesis that β_0 is zero.

3. Asymptotic Results

We establish asymptotic properties of the sieve maximum likelihood estimators $(\hat{\beta}_n, \hat{\Lambda}_n)$. Since the results depend on the distribution of the observation time points, we need some notations. Let K be the number of observation points and $\Theta = \mathcal{B} \times \mathcal{A}$. For any Borel sets B_1, B_2 in $[0, \tau]$ and B_3 in R^p, we define measures P, ν and μ as

$$\begin{aligned} P(B_1 \times B_2 \times B_3) = \int_{B_3} \sum_{k=1}^{\infty} & P(K = k|X = x) \\ & \times \sum_{j=1}^{k} P(T_{j-1} \in B_1, T_j \in B_2 | K = k, X = x) dF(x), \end{aligned}$$

$$\nu(B_1 \times B_3) = \int_{B_3} \sum_{k=1}^{\infty} P(K = k|X = x) \sum_{j=1}^{k} P(T_j \in B_1 | K = k, X = x) dF(x),$$

and $\mu(B_1) = \nu(B_1 \times R^p)$, where $F(\cdot)$ is the distribution of X on R^p. Hence, the

measure P corresponds to the joint distribution of a randomly selected observed intervals and X, ν is the joint distribution of a randomly selected observed time point and X, and μ is the marginal distribution for a randomly selected observation time point.

For the asymptotic properties of the proposed estimators, we require some regularity conditions.

(C.1) The true value β_0 in the interior of compact set $\mathcal{B}$. The true baseline function $\lambda_0(t)$ is strictly positive in $[0,\tau]$ and is rth continuously differentiable in $[0,\tau]$, for some $r > 2$.

(C.2) X is uniformly bounded by a positive constant M; $P(K < k_0|X) = 1$ almost surely for some constant k_0, and $P(K \geq 1|X) > 0$.

(C.3) The observation time points satisfy $P(\min_{j=1,\cdots,K}(T_j - T_{j-1}) \geq t_0|X) = 1$ for some $t_0 > 0$ and $P(T_K = \tau|X) > 0$. The measure μ is dominated by Lebesgue measure in $[0,\tau]$ with a positive Radon-Nikodym derivative.

(C.4) If $g(t) + X'\beta = 0$ almost surely for some determination function g and vector β, then $g \equiv 0$ and $\beta = 0$.

(C.5) The number of knots m_n satisfies $m_n^{3/2}/\sqrt{n} \to 0$, $\sqrt{n}/m_n^{2r} \to 0$.

(C.6) D_n satisfies $D_n \to \infty$ and $D_n/\log n \to 0$.

Remark. (C.1) is a boundedness and smoothness condition for the true value. (C.2) and (C.3) require that some subjects have at least two observation time points and that the observed time intervals have at least the length t_0. (C.3) ensures that the union of the supports of T_j's contains $[0,\tau]$ so that Λ is estimable on this interval. These conditions usually hold in practice when the follow-up visits are scheduled in the studies with certain variability. (C.4) is an identifiability condition concerning the linear independence of covariate X. (C.5) and (C.6) set the constraints on the size of the sieve space in terms of the order of knot number and the bound of the sieve functions. Particularly, we can choose $m_n = n^a$ for $a \in (1/(4r), 1/3)$ and $D_n = \log\log n$.

Let $\theta_0 = (\beta_0, \Lambda_0)$ be the true parameter, and $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$ be the proposed estimator.

Theorem 1. *Under conditions (C.1) – (C.6),*

$$\|\hat{\beta}_n - \beta_0\| \to 0 \quad a.s., \text{ and } \sup_{t\in[0,\tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \to 0.$$

Additionally, $d^2(\hat{\theta}_n, \theta_0) = \|\hat{\beta}_n - \beta_0\|^2 + \|\hat{\Lambda}_n - \Lambda_0\|^2_{L_2(\mu)} \leq o_P(n^{-1/2}) + O_P(m_n^{-2r})$, *and* $n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N_d\big(0, \Sigma^{-1}\big)$, *where* Σ^{-1} *is the semiparametric efficiency bound for* β_0 *(Bickel et al. (1993)).*

We conclude therefore that $\hat{\beta}_n$ is an efficient estimator for β_0. The proof mainly follows standard arguments for sieve estimation with the challenge of handling dependent multivariate observation time points. Apart from what is need for the traditional current status data (see Huang (1996)), our proof of efficiency theory relies on the derivation of the efficient score function in the presence of multivariate observation time points. The details of the proof are given in the Appendix.

4. Simulation Studies

To assess the behavior of the proposed method with finite sample sizes, we conducted two simulation studies.

Simulation 1. Two covariates $X_1 \sim \text{Bernoulli}(0.5, 0, 1)$ (discrete) and $X_2 \sim N(0, 1)$ (continuous) were used for the intensity function $\lambda(t|X) = \lambda_0(t) \exp(X_1\beta_1 + X_2\beta_2)$, where $\lambda_0(t) = 0.5$ for *Case I* and $\lambda_0(t) = t$ for *Case II*. Combinations of different β_1 and β_2 were introduced to test the proposed estimate method and algorithm. For the time points with $t_0 = 0.6$, for each i, we generated T_{i1} randomly from $Unif(t_0, 2)$, then took $T_{i,j+1} = T_{ij} + t_0$, $j = 0, 1, \cdots, K_i - 1$ until T_{iK_i} and K_i satisfied $\tau - T_{iK_i} < t_0$, where $\tau = 4$. To calculate the sieve maximum likelihood estimates, we chose even quantiles of the observation time points as the knots and the number of the knots, m_n, was chosen to be three, since the differences in the simulation results were small when we varied m_n from 2 to 5. We used a sieve space with $D_n = C \log\{\log(n)\}$, and set $C = 10$ in both simulations and data analyzes. In the simulation, we took $n = 100$ and 300 and repeated each simulation setting 1,000 times. To estimate the asymptotic variance using profile likelihood approach, we considered $h_n = cn^{-1/2}$ with $c = 0.5, 1$ and 3.

For most simulated data, the BFGS algorithm converged within 50 iterations and showed relative robustness for different initial values. Table 1 presents the simulation results under two cases of cumulative baseline hazard functions and different settings of β_1 and β_2 with the proposed method. Column 'Bias' denotes the average of the bias of the estimates for β_1 and β_2, 'SD' is the empirical standard deviation of estimates, 'SE' stands for the mean of the estimated standard errors by the profile likelihood method with $c = 0.5, 1, 3$, and 'CP' presents the coverage proportion of 95% confidence intervals corresponding to each 'SE'. Further, we compared the proposed method with a naive approach based on generalized estimating equations that treated Δ_{ij}'s as a sequence of binary outcomes

Table 1. Results of Simulation 1 under different scenarios.

					SE			CP %			GEE Model			
Case	n	True	Bias	SD	$c=0.5$	$c=1$	$c=3$	$c=0.5$	$c=1$	$c=3$	Bias	SD	SE	CP
I	100	$\beta_1=$ 1	0.012	0.160	0.159	0.160	0.166	95.2	95.2	96.0	0.365	0.218	0.223	65.0
		$\beta_2=$ 1	0.025	0.099	0.101	0.103	0.113	95.4	95.6	97.2	0.383	0.141	0.141	21.4
		$\beta_1=$ 0	0.001	0.167	0.164	0.164	0.165	95.2	95.4	95.6	−0.018	0.223	0.212	94.0
		$\beta_2=$ 1	0.018	0.102	0.103	0.105	0.114	95.6	95.6	97.4	0.270	0.142	0.137	52.2
		$\beta_1=-1$	−0.015	0.217	0.199	0.198	0.195	94.2	93.6	92.2	−0.193	0.254	0.242	87.4
		$\beta_2=$ 1	0.031	0.118	0.116	0.118	0.125	95.4	95.8	97.2	0.209	0.150	0.144	67.2
	300	$\beta_1=$ 1	0.005	0.087	0.090	0.090	0.092	96.6	96.6	97.0	0.360	0.127	0.128	18.8
		$\beta_2=$ 1	0.008	0.058	0.056	0.057	0.060	93.2	93.8	95.2	0.376	0.082	0.081	0.0
		$\beta_1=$ 0	0.005	0.096	0.093	0.093	0.094	93.6	93.6	93.8	−0.001	0.125	0.122	94.4
		$\beta_2=$ 1	0.005	0.058	0.057	0.058	0.061	94.6	94.8	96.2	0.273	0.080	0.079	5.0
		$\beta_1=-1$	0.004	0.110	0.112	0.112	0.111	95.4	95.2	95.0	−0.202	0.135	0.139	70.2
		$\beta_2=$ 1	0.000	0.064	0.063	0.064	0.066	94.4	94.4	94.8	0.220	0.082	0.084	26.8
II	100	$\beta_1=$ 1	0.033	0.177	0.170	0.171	0.177	94.8	95.0	96.0	0.575	0.276	0.273	45.0
		$\beta_2=$ 1	0.035	0.111	0.119	0.123	0.137	96.4	96.8	98.4	0.557	0.169	0.168	5.8
		$\beta_1=$ 0	0.001	0.144	0.141	0.141	0.142	94.4	94.4	94.4	−0.010	0.241	0.224	94.4
		$\beta_2=$ 1	0.024	0.100	0.104	0.107	0.118	96.8	97.2	98.6	0.484	0.149	0.149	6.2
		$\beta_1=-1$	−0.021	0.154	0.154	0.154	0.153	94.8	94.4	94.0	−0.387	0.227	0.223	61.0
		$\beta_2=$ 1	0.030	0.103	0.103	0.105	0.113	95.4	96.0	97.4	0.390	0.147	0.142	20.4
	300	$\beta_1=$ 1	0.016	0.098	0.095	0.095	0.097	92.8	93.2	94.0	0.551	0.167	0.157	5.4
		$\beta_2=$ 1	0.010	0.064	0.065	0.066	0.071	95.6	96.8	97.8	0.545	0.098	0.097	0.0
		$\beta_1=$ 0	−0.001	0.080	0.080	0.080	0.080	94.6	94.6	94.6	−0.003	0.131	0.129	94.8
		$\beta_2=$ 1	0.005	0.059	0.058	0.059	0.065	95.0	95.2	97.4	0.460	0.090	0.085	0.0
		$\beta_1=-1$	−0.005	0.085	0.088	0.087	0.087	95.6	95.6	95.6	−0.373	0.125	0.128	16.4
		$\beta_2=$ 1	0.007	0.057	0.057	0.058	0.060	95.2	96.0	97.0	0.376	0.079	0.082	0.0

measured longitudinally, since currently there are no other options for analyzing such kind of longitudinal data with repeated current status. The results by the GEE model with independent correlation structure are also reported in Table 1.

From Table 1, we observe that the proposed estimates perform satisfactorily under both Case I and Case II scenarios in terms of fairly small bias and accurate agreement between the estimated and the empirical standard errors. The proposed variance estimation method was not sensitive to the choice of tuning parameter c, although the choice of $c = 1$ might yield a better coverage. Comparing the results of proposed method to that by GEE, we see that treating repeated current status as longitudinal outcomes can result in severe bias and incorrect inference in assessing the true association between the covariates and the recurrent events. Figure 1 displays the estimated curves for $\Lambda(t)$ under Case I and Case II baseline hazard functions, respectively. It shows that the estimated curves by the proposed B-spline method were close to the true curves on average.

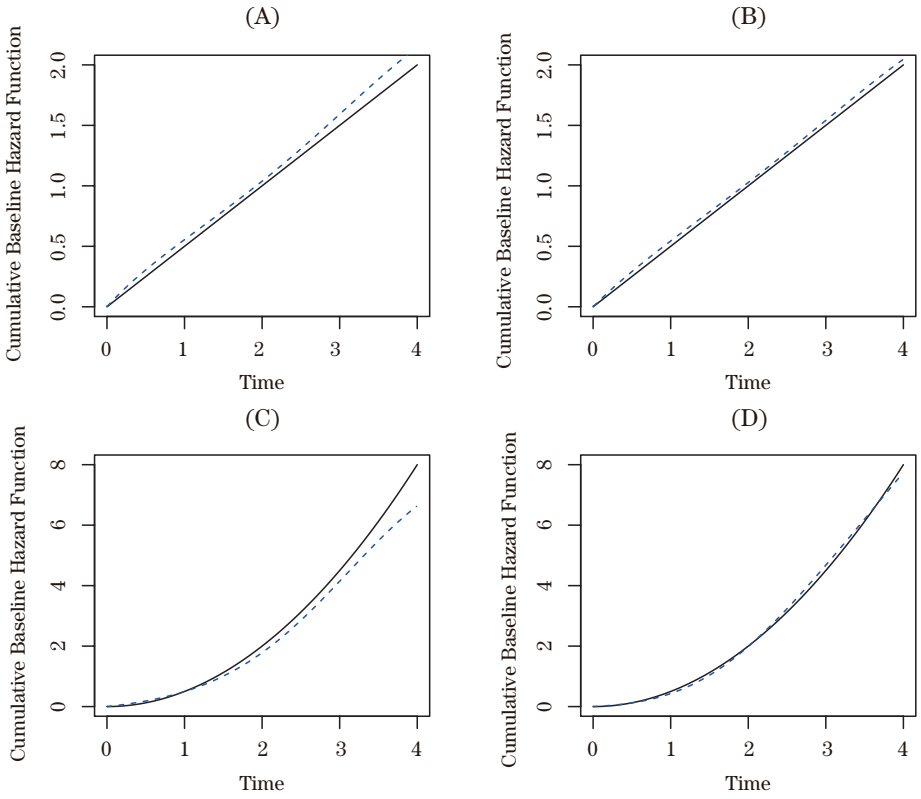

Figure 1. Estimated cumulative baseline hazard functions: the solid lines are the true ones, dash lines are the estimated ones; (A),(B) correspond to Case I and (C), (D) correspond to Case II with sample sizes 100 (left) and 300 (right).

Simulation 2. In the second simulation, we used a polynomial function as the baseline,

$$\lambda(t) = 1.225 + 2.627t - 4.708t^2 + 2.901t^3 - 0.716t^4 + 0.06t^5,\ 0 \le t \le \tau,$$

which corresponds to the cumulative hazard function

$$\Lambda(t) = 1.225t + 2.627\frac{t^2}{2} - 4.708\frac{t^3}{3} + 2.901\frac{t^4}{4} - 0.716\frac{t^5}{5} + 0.01t^6.$$

We generated covariates from two non-symmetric distributions in this simulation, $X_1 \sim$ Bernoulli (0.6, 0, 1) and $X_2 \sim \exp(1)$. What is more, we generated the visit time points T_{ij}s using a Poisson process with intensity 0.5.

Results are presented in Table 2 and Figure 2. Table 2 shows that the proposed method still performs well, and consistent with the results in Simulation 1. Figure 2 shows that as the sample size increases, the accuracy of estimated baseline hazard function is significantly improved.

Table 2. Results of Simulation 2 under different scenarios.

				SE			CP %			GEE Model			
n	True	Bias	SD	$c=0.5$	$c=1$	$c=3$	$c=0.5$	$c=1$	$c=3$	Bias	SD	SE	CP
100	$\beta_1=$ 1	0.123	0.347	0.323	0.325	0.337	93.9	94.3	95.3	0.337	0.186	0.298	85.0
	$\beta_2=$ 1	0.169	0.337	0.307	0.310	0.331	94.5	94.8	95.8	0.417	0.273	0.270	72.6
	$\beta_1=$ 0	−0.012	0.259	0.247	0.247	0.247	94.7	95.0	94.9	0.468	0.266	0.238	49.5
	$\beta_2=$ 1	0.100	0.240	0.240	0.243	0.257	95.4	95.6	96.6	0.316	0.168	0.232	78.5
	$\beta_1=-1$	−0.057	0.254	0.238	0.239	0.234	94.4	94.2	93.9	0.494	0.280	0.218	37.6
	$\beta_2=$ 1	0.081	0.197	0.194	0.198	0.206	95.1	95.6	96.3	0.256	0.124	0.200	79.9
300	$\beta_1=$ 1	0.029	0.159	0.162	0.162	0.164	95.2	95.4	96.3	0.337	0.140	0.169	50.5
	$\beta_2=$ 1	0.033	0.142	0.146	0.147	0.151	96.1	96.2	96.4	0.373	0.166	0.151	28.9
	$\beta_1=$ 0	0.004	0.129	0.133	0.133	0.133	95.5	95.6	95.6	0.465	0.232	0.135	4.9
	$\beta_2=$ 1	0.025	0.124	0.122	0.122	0.125	94.4	94.5	95.4	0.287	0.102	0.131	42.9
	$\beta_1=-1$	−0.024	0.131	0.130	0.129	0.129	94.8	94.8	94.6	0.504	0.266	0.124	1.7
	$\beta_2=$ 1	0.027	0.103	0.101	0.102	0.104	94.6	94.5	95.0	0.228	0.067	0.113	51.0

5. Application to the STAR*D Study

We applied our method to study medication adherence in the Sequenced Treatment Alternatives to Relieve Depression the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial. This trial was a phase-IV multi-site, multi-stage randomized clinical trial to compare various treatment strategies for patients with non-psychotic major depressive disorder (Rush et al. (2004)). The aim of the STAR*D study was to find the best subsequent treatment for subjects who failed to achieve adequate response to an initial antidepressant treatment (citalopram, Level 1 treatment). All patients received the Level 1 treatment, and for those who did not benefit from Level 1 treatment, Level 2 treatments were designed to help determine an appropriate next treatment step. Medications used in Level 2 treatment included sertraline (Zoloft), bupropion-SR (Wellbutrin), or venlafaxine-XR (Effexor). These medications were chosen for comparison because they represent three different classes of medications. Sertraline is an selective serotonin reuptake inhibitor (SSRI), the same class as the citalopram used in Level 1. Bupropion belongs to another class of antidepressant medications that work on different neurotransmitters than SSRIs. Venlafaxine is a dual-action medication that works on two neurotransmitters simultaneously. The STAR*D trial enrolled 4,041 outpatients with non-psychotic depression at 23 psychiatric and 18 primary care sites and obtained 80,820 observations in total. After excluding the incomplete observations, the final data set for our analysis consisted of 1,958 patients with 9,150 observations and the maximum follow-up time up to 168 days.

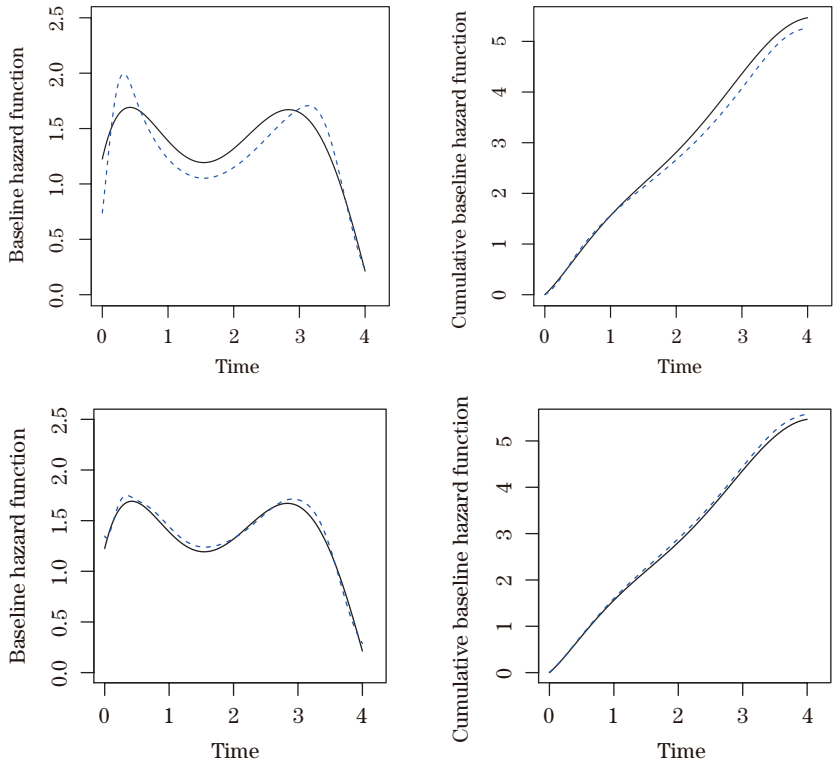

Figure 2. Estimated baseline hazard functions (left) and cumulative baseline hazard functions (right) in Simulation 2: the solid lines are the true ones, dash lines are the estimated ones. The figures in first and second rows correspond to $n = 100$ and 300.

Our interest is to assess the association between various baseline factors and the patient's medication adherence in Level 2 treatment. In the study, medication adherence is assessed by a categorical variable collected at baseline approximately, week 2, 4, 6, 9, 12, and 14 after entering the Level 2 treatment. The variable collecting information on "How often missed medication since the last visit" takes values in the categories "never", "rarely", "sometimes" and so on. Therefore, the exact incidence of non-adherence is not available but we do know whether there was any non-adherence during specific follow-up time intervals. Here, Δ is the indicator of whether the the event of "missing medication" at least some times has occurred or not during each observational interval. The proportion of missing at least some medication is about 28%. For our analysis, the baseline covariates are demographic variables including sex, race (white versus others), and age (range 18 to 75, median 40), impact of your family and friends, parent history of depression, and a baseline clinician-rated Quick Inventory of

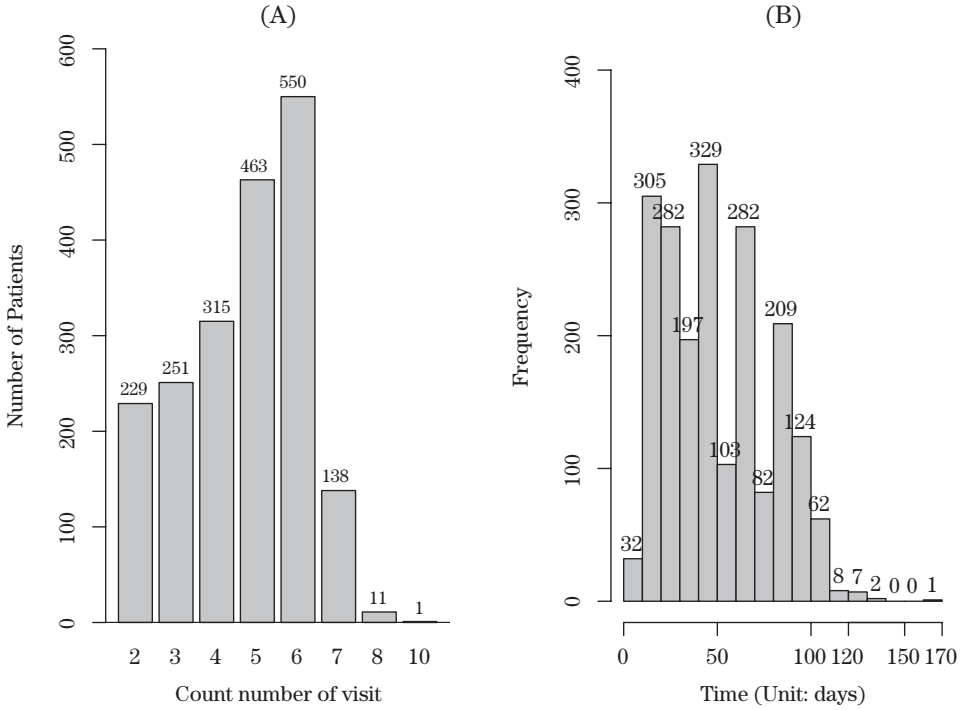

Figure 3. (A) Bar plot for the visit count number K_i; (B) The histogram of the visit times with events occurred $\Delta = 1$ in the corresponding time intervals.

Depressive Symptomatology (QIDS) score ranging from 0 to 26 in the sample (median 16), a measure of the severity of a patient's depressive symptoms.

Figure 3 (A) and (B) describe the distributions of the frequencies of follow-up intervals and the actual follow-up time periods when some non-adherence has occurred prior to the time, respectively. There is a wide variability in the follow-up assessment patterns between the patients. For example, 229 patients have a minimum number of 2 assessments, and most patients ($n = 550$) have 6 follow ups. Figure 3 (B) suggests that the intensity of the non-adherence events may be adequately estimated in the time interval $[0, 100]$ since Level 2 baseline, due to dense observations of occurrences. In addition, the histogram in Figure 3 (B) suggests some sinusoidal pattern of the event intensity that may be well captured by the B-spline basis expansion of the baseline hazard. The Pearson correlation coefficients of paired $(\Delta_{iL}, \Delta_{iR})$'s in neighboring observation intervals (like [0,20] and [20,40], [20,40] and [40,60]) are all very small and the p-values of independence tests corresponding to these intervals are all significantly smaller than 0.05.

Table 3. Association between baseline characteristics and medication non-adherence in the STAR*D study.

	Proposed Method			GEE Model		
Covarates	Est.	SE	P-value	Est.	SE	P-value
Family and Friends Impact	0.059	0.012	0.000	0.068	0.020	0.001
Female	−0.096	0.048	0.045	−0.082	0.082	0.318
White	−0.379	0.059	0.000	−0.518	0.104	0.000
Age	−0.022	0.002	0.000	−0.026	0.003	0.000
Parent History of Depression	−0.016	0.053	0.765	−0.037	0.087	0.676
Baseline QIDS	−1.132	0.643	0.078	0.382	0.204	0.062

To determine the number of knots and the degree of the B-splines, we considered m_n from 2 to 9 and l between 2 and 3. Using the BIC criterion, we selected $m_n = 2$ and $l = 3$. With this choice, the results are reported in Table 3, where the standard errors are estimated using the profile likelihood approach with $h_n = n^{-1/2}$. From this table, we conclude that family and friends impact, sex, white race, and age are significant predictors for medication non-adherence, while parent history of depression is not significant, and baseline QIDS is marginally significant. Particularly, the results show that non-whites are more likely to be non-adherent and males are more likely to miss medication. Interestingly, using GEE to analyze the sequence of binary outcomes yields insignificant gender effect and shows that the effect of baseline QIDS to be the opposite sign as the proposed method. Higher QIDS score at the baseline indicates more severe depressive symptoms, and a sicker patient is expected to be more adherent. Therefore, the coefficient of this variable on medication non-adherence is expected to be negative, consistent with that of the proposed method.

Figure 4 shows the estimated baseline hazard function and the estimated cumulative baseline hazard function. The sine wave shape of the estimated hazard function up to day 120 is consistent with the observed event intensity described in Figure 3(B). The estimated cumulative baseline hazard function demonstrates a significant increase of the cumulative hazard rate before day 120. The flattened portion of $\Lambda_0(t)$ after day 120 may be due to insufficient events toward the end of the study. On average, one expects about 20 episodes of non-adherent events in the 4 month study follow-up period, or about 5 episodes per month.

6. Discussion

In this work, we study the Andersen-Gill model for analyzing repeated current status data. This type of data occurs frequently in medical studies due

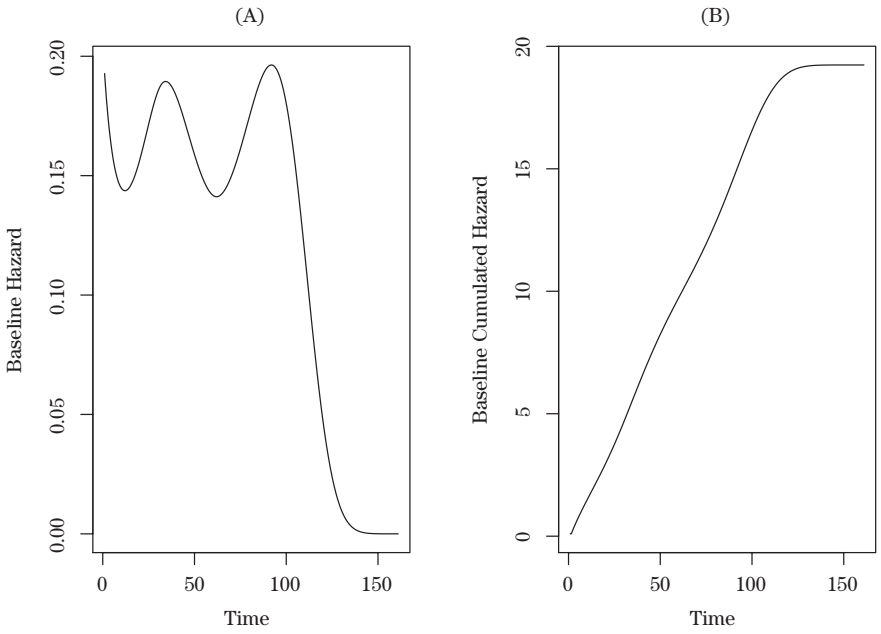

Figure 4. (A) The estimated baseline hazard function, and (B) the estimated cumulative baseline hazard function.

to practical constraints or incomplete information. However, no valid analysis method has been yet available and clinical researchers usually resort to transforming the problem into one of analyzing sequences of binary longitudinal data and using standard GEE methods. This can lead to severe bias and incorrect inference.

The proposed sieve estimators are shown to be consistent, asymptotic normal and, more importantly, seimparametrically efficient. The estimation of the regression coefficients and their asymptotic variances is computationally easy. Numerical studies show that the proposed estimators performs well under moderate sample sizes. Although we have focused on a nonhomogeneous Poisson process with proportional intensity model, the approach can be generalized to allow latent frailty and non-proportional intensities. It can also be adapted to allow time-dependent covariates. The Poisson assumption may be restrictive in practice, and a possible generalization is to include some latent frailty in our current model. This would be a challenge, both computationally and theoretically. We will pursue this in future work.

The main reasons for using the sieve likelihood approach over the nonparametric maximum likelihood approach used in Zeng, Cai and Shen (2006) are two

fold. The smooth sieve approach yields a smooth estimate of the baseline function and so can be easily visualized for patterns, as in Figure 2. Then too, our experience is that the computation for the nonparametric maximum likelihood estimate may not be stable due to sparse information of observed time points, while the sieve estimation uses fewer parameters and in borrowing strength from neighboring intervals due to smoothing, is computationally more stable.

Acknowledgements

This work is supported by BCMIS, NSFC (No. 11371062), China Zhongdian Project (No. 11131002), and China Scholarship Council. Wang's work is partly supported by grants NS073671 and NS082062.

Appendix: Proof of Theorem 1

For convenience, we assume $\tau = 1$ and let $G_n = \sqrt{n}(P_n - P)$, where P_n is the empirical measure and P is the true probability measure.

Consistency. By Schumaker (2007) and (C.1), there exists a function $\tilde{\Lambda}_0 \in \mathcal{S}_n$ such that $\|\tilde{\Lambda}_0 - \Lambda_0\|_\infty = O(m_n^{-r})$, where $\|\cdot\|_\infty$ is the supreme norm. We use Hellinger distance (see van de Geer (1993), Zeng, Cai and Shen (2006)) to prove consistency. Because

$$\sup_{t\in[0,1]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \le \sup_{t\in[0,1]} |\hat{\Lambda}_n(t) - \tilde{\Lambda}_0(t)| + \sup_{t\in[0,1]} |\tilde{\Lambda}_0(t) - \Lambda_0(t)|,$$

we only need to show $\|\hat{\beta}_n - \beta_0\| \to 0$, and $\sup_{t\in[0,1]} |\hat{\Lambda}_n(t) - \tilde{\Lambda}_0(t)| \to 0$, a.s.

Let $G(t) = \exp\{-\tilde{\Lambda}(t)\}$, $\tilde{G}_0(t) = \exp\{-\tilde{\Lambda}_0(t)\}$ and $G_0(t) = \exp\{-\Lambda_0(t)\}$. We consider a class of the likelihood functions denoted by

$$\mathcal{F} = \Big\{ f_{\beta,G} = \prod_{j=1}^{K} f_{\beta,G,j} : f_{\beta,G,j} = \Big[1 - \Big(\frac{G(T_j)}{G(T_{j-1})}\Big)^{e^{X'\beta}}\Big]^{\Delta_j} \Big(\frac{G(T_j)}{G(T_{j-1})}\Big)^{e^{X'\beta}(1-\Delta_j)} \text{ with } (\beta, G) \in \Theta^*,\ T_j \in [0,1] \text{ and } T_0 = 0 \Big\},$$

where $\Theta^* = \mathcal{B} \times \{e^{-\tilde{\Lambda}} : \tilde{\Lambda} \in \mathcal{S}_n\}$. First, we calculate the bracket covering number for $\mathcal{F}$. For any $\varepsilon > 0$ and (β_1, G_1), $(\beta_2, G_2) \in \Theta^*$, such that $\|\beta_1 - \beta_2\| < \varepsilon$, $\sup_{t\in[0,1]} |G_1(t) - G_2(t)| < \varepsilon$, we wish to set boundaries for the difference between $\sqrt{f_{\beta_1,G_1}}$ and $\sqrt{f_{\beta_2,G_2}}$. There are two scenarios:

Case I. If $f_{\beta_1,G_1} > \eta$, where η is a constant, satisfying $\eta > \varepsilon K_0 e^{BM}(MO(D_n) + 1) > 0$, we have $f_{\beta_2,G_2} > \eta - (f_{\beta_1,G_1} - f_{\beta_2,G_2})$. By the definition, we know

$$|f_{\beta_1,G_1} - f_{\beta_2,G_2}| \le |f_{\beta_1,G_1,1} - f_{\beta_2,G_2,1}| + \cdots + |f_{\beta_1,G_1,K} - f_{\beta_2,G_2,K}|.$$

Note that $|f_{\beta_1,G_1,j} - f_{\beta_2,G_2,j}| \le \int_{T_{j-1}}^{T_j} d\Lambda_1(t) \cdot |e^{X'\beta_1} - e^{X'\beta_2}| + |\int_{T_{j-1}}^{T_j} d(\Lambda_1(t) - \Lambda_2(t))|e^{X'\beta_2}$. Furthermore, since only l B-spline basis functions are non-zero at each s,

$$\tilde{\Lambda}(t) \le \int_0^t \exp\big\{\sum_{j=1}^{k_n} |\alpha_j| \, \|B_{nj}(s)\|_\infty\big\} ds \le O(1)\exp(lD_n)\tau$$

holds for any $\tilde{\Lambda} \in \mathcal{S}_n$, hence $|f_{\beta_1,G_1,j} - f_{\beta_2,G_2,j}| \le \varepsilon e^{BM}(MO(D_n)+1)$ and $f_{\beta_2,G_2} \ge \eta - \varepsilon K_0 e^{BM}(MO(D_n)+1) > 0$. Therefore,

$$\begin{aligned} |\sqrt{f_{\beta_1,G_1}} - \sqrt{f_{\beta_2,G_2}}| &= \Big|\frac{f_{\beta_1,G_1} - f_{\beta_2,G_2}}{\sqrt{f_{\beta_1,G_1}} + \sqrt{f_{\beta_2,G_2}}}\Big| \\ &\le \frac{\varepsilon K_0 e^{BM}(MO(D_n)+1)}{2\sqrt{\eta - \varepsilon K_0 e^{BM}(MO(D_n)+1)}}. \end{aligned}$$

Case II. If $f_{\beta_1,G_1} \le \eta$, then we have $f_{\beta_2,G_2} \le \eta + \varepsilon K_0 e^{BM}(MO(D_n)+1)$. Therefore, $|\sqrt{f_{\beta_1,G_1}} - \sqrt{f_{\beta_2,G_2}}| \le \sqrt{\eta} + \sqrt{\eta + \varepsilon K_0 e^{BM}(MO(D_n)+1)}$. If we choose $\eta = 2\varepsilon K_0 e^{BM}(MO(D_n)+1)$, then in either case we have

$$\begin{aligned} |\sqrt{f_{\beta_1,G_1}} - \sqrt{f_{\beta_2,G_2}}| &\le 4\sqrt{\varepsilon K_0 e^{BM}(MO(D_n)+1)} \\ &= C_1\sqrt{\varepsilon K_0 e^{BM}(MO(D_n)+1)}. \end{aligned}$$

Consequently, if we define $c_n = \varepsilon K_0 e^{BM}(MO(D_n)+1)$, it holds

$$\log N_{[]}\big(O(1)\sqrt{c_n\varepsilon}, \sqrt{\mathcal{F}}, \|\cdot\|_\infty\big) \le \log N_{[]}(\varepsilon, \Theta^*, \|\cdot\|_\infty) \le O(\varepsilon^{-1}),$$

where $N_{[]}(\cdot)$ denotes the bracket covering number, or

$$\log N_{[]}\big(O(1)\sqrt{\varepsilon}, \sqrt{\mathcal{F}}, \|\cdot\|_\infty\big) \le O(c_n\varepsilon^{-1}).$$

According to (C.6) and the results of Theorem 2.4 and Lemma 1.1 in van de Geer (1993), plus the fact that $f_{\hat{\beta}_n,\hat{G}_n} \in \mathcal{F}$, we obtain that the Hellinger distance between $f_{\hat{\beta}_n,\hat{G}_n}$ and $f_{\beta_0,\tilde{G}_0}$ converges to zero as $n \to \infty$, so $E(\sqrt{f_{\hat{\beta}_n,\hat{G}_n}} - \sqrt{f_{\beta_0,\tilde{G}_0}})^2 \to 0, a.s.$ Since $\hat{\beta}_n$ is in a compact set $\mathcal{B}$ and $\hat{G}_n$ is a bounded non-increasing function, by Helly's Selection Theorem, for any subsequence we can find a sub-subsequence, still subscripted by n, such that $\hat{\beta}_n \to \beta^*$, $\hat{G}_n(\cdot) \to \tilde{G}^*(\cdot)$ point-wise with probability 1. Because $f_{\hat{\beta}_n,\hat{G}_n}$ is bounded by 1, and $f_{\hat{\beta}_n,\hat{G}_n}$ converges to $f_{\beta_0,\tilde{G}_0}$ in Hellinger distance, we conclude that $f_{\beta^*,\tilde{G}^*} = f_{\beta_0,\tilde{G}_0}$ hold for any $t \in [0,1]$, $\tilde{G}_0(t) > 0$. For $K = 1$, we have $\ln(G^*(T_1)) \cdot e^{X'\beta^*} = \ln(\tilde{G}_0(T_1)) \cdot e^{X'\beta_0}$, $a.s.$ Then by (C.4), we obtain that $\beta^* = \beta_0$, $G^*(t) = \tilde{G}_0(t)$, $t \in [0,1]$. Hence,

$$\|\hat{\beta}_n - \beta_0\| \to 0, \ \sup_{t\in[0,1]} |\hat{G}_n(t) - \tilde{G}_0(t)| \to 0, \ a.s..$$

Since the transform function $G(\cdot)$ is one-to-one and continuous, we can easily obtain $\sup_{t\in[0,1]}|\hat{\Lambda}_n(t)-\tilde{\Lambda}_0(t)|\to 0,\ a.s.$, and consistency holds.

Convergence rate. Denote the log-likelihood function for one subject as

$$l(\theta;O)=\sum_{j=1}^{K}\Big\{\Delta_j\log\big[1-\exp(-\delta\Lambda_j e^{X'\beta})\big]-(1-\Delta_j)\delta\Lambda_j e^{X'\beta}\Big\},$$

where $\delta\Lambda_j=\Lambda(T_j)-\Lambda(T_{j-1})$. The first derivatives of log-likelihood $l(\theta;O)$ respect to β and Λ along the submodels $\beta+\epsilon b$ and $\Lambda+\epsilon h$, respectively, are

$$\dot{l}_\beta(\theta;O)=\sum_{j=1}^{K}\Big\{\frac{\Delta_j}{1-\exp(-\delta\Lambda_j e^{X'\beta})}-1\Big\}\delta\Lambda_j e^{X'\beta}X,$$

$$\dot{l}_\Lambda(\theta;O)[h]=\sum_{j=1}^{K}\Big\{\frac{\Delta_j}{1-\exp(-\delta\Lambda_j e^{X'\beta})}-1\Big\}e^{X'\beta}\delta h_j,$$

and the second derivatives of log-likelihood $l(\theta;O)$ respect to β and Λ along the submodels $\beta+\epsilon b$ and $\Lambda+\epsilon h$, respectively, are

$$\dot{l}_{\beta\beta}(\theta;O)=\sum_{j=1}^{K}\Psi_j(\theta,O)\delta\Lambda_j e^{X'\beta}XX',$$

$$\dot{l}_{\Lambda\beta}(\theta;O)[h]=\sum_{j=1}^{K}\Psi_j(\theta,O)Xe^{X'\beta}\delta h_j,$$

$$\Psi_j(\theta,O)=\frac{\Delta_j}{1-\exp(-\delta\Lambda_j e^{X'\beta})}-\frac{\Delta_j\exp(-\delta\Lambda_j e^{X'\beta})\delta\Lambda_j e^{X'\beta}}{[1-\exp(-\delta\Lambda_j e^{X'\beta})]^2}-1,$$

$$\dot{l}_{\Lambda\Lambda}(\theta;O)[h_1,h_2]=\sum_{j=1}^{K}\frac{-\Delta_j e^{2X'\beta}\exp(-\delta\Lambda_j e^{X'\beta})\delta h_{1j}\delta h_{2j}}{[1-\exp(-\delta\Lambda_j e^{X'\beta})]^2},$$

where $\delta h_j=\int_{T_{j-1}}^{T_j}dh(s)$ for $h\in\mathcal{H}(\Lambda)$, $\mathcal{H}(\Lambda)=\big\{h:(\partial\Lambda_\epsilon)/(\partial\epsilon)\big|_{\epsilon=0}=h,\ h\in L_2([0,1]),\ h(0)=0$, and Λ_ϵ is a parametric path in $\mathcal{A},\ \Lambda_\epsilon|_{\epsilon=0}=\Lambda\big\}$ is the tangent space.

If $\big(\dot{l}_\beta,\dot{l}_\Lambda\big)^*$ is the dual operator of $\big(\dot{l}_\beta,\dot{l}_\Lambda\big)$, then the information operator $\mathcal{I}(\beta,\Lambda)=\big(\dot{l}_\beta,\dot{l}_\Lambda\big)^*\big(\dot{l}_\beta,\dot{l}_\Lambda\big)$ satisfies

$$-\big\langle\mathcal{I}(\beta,\Lambda)[b,h],[b,h]\big\rangle_{L_2(P)}=P\big\{b'\dot{l}_{\beta\beta}b+2(\dot{l}_{\Lambda\beta}[h])'b+\dot{l}_{\Lambda\Lambda}[h,h]\big\},$$

where $b\in R^p$, $h\in\mathcal{H}(\Lambda)$. By Lemma 1 (see the end of the proof) and (C.2), we know $\mathcal{I}(\beta_0,\Lambda_0)$ is an invertible operator, which implies

$$\|\mathcal{I}(\beta_0,\Lambda_0)[b,h]\|^2_{L_2(P)}\geq O(1)\big(\|b\|^2+\|h\|^2_{L_2(\mu)}\big).\tag{A.1}$$

For every θ in a neighborhood of θ_0, we expand $P\{l(\theta;O)-l(\theta_0;O)\}$ at the true

values θ_0. The first derivative in the expansion vanishes and

$$P\{l(\theta_0;O)-l(\theta;O)\}=\big\langle \mathcal{I}(\beta^*,\Lambda^*)[\beta-\beta_0,\Lambda-\Lambda_0],[\beta-\beta_0,\Lambda-\Lambda_0]\big\rangle_{L_2(P)}$$

where (β^*,Λ^*) is between (β,Λ) and (β_0,Λ_0). Hence, for the neighborhood of (β_0,Λ_0) sufficiently small, by (A.1) we have $P\{l(\theta_0;O)-l(\theta;O)\}\geq O(1)d^2(\theta,\theta_0)$.

Let $\mathcal{M}_\zeta(\theta_0)=\{l(\theta;O)-l(\theta_0;O):\ (\|\beta-\beta_0\|^2+\|\Lambda-\Lambda_0\|^2_{L_2(\mu)})^{1/2}\leq\zeta,\ \beta\in\mathcal{B},\ \Lambda\in\mathcal{S}_n\}$ with some $\zeta>0$. For any $0<\epsilon<\zeta$, similar to the calculation of Shen and Wong (1994), we obtain $\log N_{[]}(\epsilon,\mathcal{M}_\zeta(\theta_0),L_2(P))\leq O(1)m_n\log(\zeta/\epsilon)$. What is more, with consistency and (C.1), (C.2), and (C.3), we know that there exists a $M_0>0$ such that $\|l(\theta;O)-l(\theta_0;O)\|_\infty\leq M_0$. Hence, by Lemma 3.4.2 of van der Vaart and Wellner (1996), it holds

$$E^*\|G_n\|_{\mathcal{M}_\zeta(\theta_0)}\leq O(1)\tilde{J}_{[]}(\zeta,\mathcal{M}_\zeta(\theta_0),L_2(P))\big[1+\frac{\tilde{J}_{[]}(\zeta,\mathcal{M}_\zeta(\theta_0),L_2(P))}{\zeta^2\sqrt{n}}M_0\big],$$

where $\tilde{J}_{[]}(\zeta,\mathcal{M}_\zeta(\theta_0),L_2(P))=\int_0^\zeta\big\{1+\log N_{[]}(\epsilon,\mathcal{M}_\zeta(\theta_0),L_2(P))\big\}^{1/2}d\epsilon$, hence

$$\tilde{J}_{[]}(\zeta,\mathcal{M}_\zeta(\theta_0),L_2(P))\leq O(1)m_n^{1/2}\zeta. \tag{A.2}$$

If $\phi_n(\zeta)=\tilde{J}_{[]}(\zeta,\mathcal{M}_\zeta(\theta_0),L_2(P))\big[1+(\tilde{J}_{[]}(\zeta,\mathcal{M}_\zeta(\theta_0),L_2(P)))/(\zeta^2\sqrt{n})M_0\big]$, and r_n satisfies $r_n^2\phi(1/r_n)\leq n^{1/2}$, we obtain $r_n^2\tilde{J}_{[]}(1/r_n,\mathcal{M}_{1/r_n}(\theta_0),L_2(P))\leq n^{1/2}$. Hence, by (A.2) and Theorem 3.2.5 of van der Vaart and Wellner (1996), we obtain $d(\hat{\theta}_n,\theta_0)\leq O_P((m_n/n)^{1/2})+O_P(m_n^{-r})$. Furthermore, by (C.5) it holds $d^2(\hat{\theta}_n,\theta_0)\leq o_P(n^{-1/2})+O_P(m_n^{-2r})$.

Asymptotic normality. The proof is divided into four steps.

Step 1. By Lemma 1, we know there exists a least favorable direction for β_0, which is defined as a tangent function h^* for Λ that satisfies

$$P\big\{\dot{l}_{\beta\Lambda}(\theta_0;O)[h]-\dot{l}_{\Lambda\Lambda}(\theta_0;O)[h^*,h]\big\}=0,$$

for any $h\in\mathcal{H}(\Lambda_0)$. Therefore, we have the efficient score function for (β,Λ),

$$\begin{aligned}l^*(\theta_0;O)&=\dot{l}_\beta(\theta_0;O)-\dot{l}_\Lambda(\theta_0;O)[h^*]\\&=\sum_{j=1}^K\Big\{\frac{\Delta_j\exp(-\delta\Lambda_{0j}e^{X'\beta_0})}{1-\exp(-\delta\Lambda_{0j}e^{X'\beta_0})}-1+\Delta_j\Big\}\big(\delta\Lambda_{0j}e^{X'\beta_0}X-e^{X'\beta_0}\delta h_j^*\big).\end{aligned}$$

By Schumaker (2007), there exists a function $h_n^*\in\mathcal{S}_n$ such that $\|h_n^*-h^*\|_\infty=O(m_n^{-r})$.

Step 2. According to (C.5) and the fact that

$$\|\dot{l}_\Lambda(\theta_0;O)[h_n^*]-\dot{l}_\Lambda(\theta_0;O)[h^*]\|_\infty\leq C\,O(m_n^{-r}),$$

we obtain $P\big\{\dot{l}_\beta(\theta_0;O)-\dot{l}_\Lambda(\theta_0;O)[h_n^*]\big\}=P\dot{l}_\Lambda(\theta_0;O)[h_n^*]\leq o_P(n^{-1/2})$. By the

definition of h^*, (C.3), (C.5), and the fact that

$$\begin{aligned}&\|\dot{l}_{\Lambda\Lambda}(\theta_0;O)[h_n^*,\hat{\Lambda}_n-\Lambda_0]-\dot{l}_{\beta\Lambda}(\theta_0;O)[\hat{\Lambda}_n-\Lambda_0]\\&\quad-\dot{l}_{\Lambda\Lambda}(\theta_0;O)[h^*,\hat{\Lambda}_n-\Lambda_0]+\dot{l}_{\beta\Lambda}(\theta_0;O)[\hat{\Lambda}_n-\Lambda_0]\|_\infty\\=&\|\dot{l}_{\Lambda\Lambda}(\theta_0;O)[h_n^*-h^*,\hat{\Lambda}_n-\Lambda_0]\|_\infty\le C\,O(m_n^{-r}),\end{aligned}$$

we conclude that $P\{\dot{l}_{\Lambda\Lambda}(\theta_0;O)[h_n^*,\hat{\Lambda}_n-\Lambda_0]-\dot{l}_{\beta\Lambda}(\theta_0;O)[\hat{\Lambda}_n-\Lambda_0]\}\le o_P(n^{-1/2})$.

Step 3. We show $-P\{\dot{l}_{\beta\beta}(\theta_0;O)-\dot{l}_{\Lambda\beta}(\theta_0;O)[h^*]\}$ is positive. As

$$-P\{\dot{l}_{\beta\beta}(\theta_0;O)+\dot{l}_{\Lambda\beta}(\theta_0;O)[h^*]\}=P\{\dot{l}_\beta(\theta_0;O)+\dot{l}_\Lambda(\theta_0;O)[h^*]\}^{\otimes2}\ge0,$$

we need only show that $P\{\dot{l}_\beta(\theta_0;O)+\dot{l}_\Lambda(\theta_0;O)[h^*]\}^{\otimes2}>0$. Otherwise, if $\dot{l}_\beta(\theta_0;O)+\dot{l}_\Lambda(\theta_0;O)[h^*]\equiv0$, we have $\sum_{j=1}^K\{\Delta_j/(1-\exp\{-\delta\Lambda_{0j}e^{X'\beta_0}\})-1\}\delta\Lambda_{0j}(\alpha_{Kj}^*+X)\equiv0$, which indicates $-(\alpha_{Kj}^*+X)\equiv0$, a contradiction.

Step 4. Since $P_n\{\dot{l}_\beta(\hat{\theta}_n;O)-\dot{l}_\Lambda(\hat{\theta}_n;O)[h_n^*]\}=0$, it holds that

$$(P_n-P)\{\dot{l}_\beta(\hat{\theta}_n;O)-\dot{l}_\Lambda(\hat{\theta}_n;O)[h_n^*]\}=-P\{\dot{l}_\beta(\hat{\theta}_n;O)-\dot{l}_\Lambda(\hat{\theta}_n;O)[h_n^*]\}.$$

An expansion on (β_0,Λ_0) of the right-hand side of the above equation yields

$$\begin{aligned}&G_n\{\dot{l}_\beta(\hat{\theta}_n;O)-\dot{l}_\Lambda(\hat{\theta}_n;O)[h_n^*]\}\\&=-\sqrt{n}P\{\dot{l}_\beta(\theta_0;O)-\dot{l}_\Lambda(\theta_0;O)[h_n^*]+[\dot{l}_{\beta\beta}(\theta_0;O)-\dot{l}_{\Lambda\beta}(\theta_0;O)[h_n^*]](\hat{\beta}_n-\beta_0)\\&\quad+\dot{l}_{\Lambda\Lambda}(\theta_0;O)[h_n^*,\hat{\Lambda}_n-\Lambda_0]-\dot{l}_{\beta\Lambda}(\theta_0;O)[\hat{\Lambda}_n-\Lambda_0]\}\\&\quad+\sqrt{n}O_P(\|\hat{\beta}_n-\beta_0\|^2+\|\hat{\Lambda}_n-\Lambda_0\|_{L_2}^2).\end{aligned}$$

For the left-hand side, it is easy to check that $\dot{l}_\beta(\hat{\theta}_n;O)-\dot{l}_\Lambda(\hat{\theta}_n;O)[h_n^*]$ belongs to the P-Donsker class $\{\dot{l}_\beta(\theta;O)-\dot{l}_\Lambda(\theta;O)[h]:\ \theta\in\mathcal{B}\times\mathcal{S}_n,\ h\in L_2([0,1])\}$, and by the result in *Step 2*,

$$\begin{aligned}&P_n\{\dot{l}_\beta(\theta_0;O)-\dot{l}_\Lambda(\theta_0;O)[h^*]\}+o_P(n^{-1/2})\\&=-\big(P\{\dot{l}_{\beta\beta}(\theta_0;O)-\dot{l}_{\Lambda\beta}(\theta_0;O)[h^*]\}+o_P(n^{-1/2})\big)(\hat{\beta}_n-\beta_0)\\&\quad+O_P\{\|\hat{\beta}_n-\beta_0\|^2+\|\hat{\Lambda}_n-\Lambda_0\|_{L_2}^2\}.\end{aligned}$$

By the result in *Step 3*, we have

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n-\beta_0)=&-P^{-1}\{\dot{l}_{\beta\beta}(\beta_0,\Lambda_0;O)-\dot{l}_{\Lambda\beta}(\beta_0,\Lambda_0;O)[h^*]\}\\&\times\sqrt{n}P_n\{\dot{l}_\beta(\beta_0,\Lambda_0;O)-\dot{l}_\Lambda(\beta_0,\Lambda_0;O)[h^*]\}\\&+o_P(1)+\sqrt{n}O_P\{\|\hat{\beta}_n-\beta_0\|^2+\|\hat{\Lambda}_n-\Lambda_0\|_{L_2}^2\}.\end{aligned}$$

By (C.5) and the convergence rate obtained previously, the last term here is also $o_P(1)$. Hence, the asymptotic normality of $\sqrt{n}(\hat{\beta}_n-\beta_0)$ holds.

Lemma 1. *Under (C.1)–(C.4), there exists a unique $h^*\in\mathcal{H}(\Lambda_0)$ such that $P\{\dot{l}_{\beta\Lambda}(\beta_0,\Lambda_0)[h]-\dot{l}_{\Lambda\Lambda}(\beta_0,\Lambda_0)[h^*,h]\}=0$, for any $h\in\mathcal{H}(\Lambda_0)$.*

Proof. By $P\{\dot{l}_{\beta\Lambda}(\beta_0,\Lambda_0)[h]-\dot{l}_{\Lambda\Lambda}(\beta_0,\Lambda_0)[h^*,h]\}=0$ and simple algebra, we need only prove that there exists a unique $h^*\in\mathcal{H}(\Lambda_0)$ such that

$$P\Big\{\sum_{j=1}^{K}(\delta h_j^*-\delta\Lambda_{0j}X)\frac{e^{2X'\beta_0}\exp(-\delta\Lambda_{0j}e^{X'\beta_0})}{1-\exp(-\delta\Lambda_{0j}e^{X'\beta_0})}\delta h_j\Big\}=0,$$

for any $h\in\mathcal{H}(\Lambda_0)$. Taking $\delta h^*(s,t)=h^*(t)-h^*(s)$ and $\delta\Lambda_0(s,t)=\Lambda_0(t)-\Lambda_0(s)$, we can rewrite the above equation as

$$\begin{aligned}0&=\int_0^\tau\int_0^t\big\{\delta h^*(s,t)\Gamma_1(s,t)-\delta\Lambda_0(s,t)\Gamma_2(s,t)\big\}(h(t)-h(s))ds\,dt\\&=\int_0^\tau h(t)\Big\{\int_0^t\delta h^*(s,t)\Gamma_1(s,t)-\delta\Lambda_0(s,t)\Gamma_2(s,t)ds\\&\quad-\int_t^\tau\delta h^*(t,s)\Gamma_1(t,s)-\delta\Lambda_0(t,s)\Gamma_2(t,s)ds\Big\}dt,\end{aligned}$$

where $\Gamma_i(s,t)=\sum_{k=1}^{\infty}P(K=k)\sum_{j=1}^{k}I(T_{j-1}=s,T_j=t)\,\alpha_{ij}(s,t)\,P_{T_{j-1},T_j|K}(s,t)$, $i=1,2$, $\alpha_{1j}(s,t)=E\big\{(e^{2X'\beta_0}\exp(-\delta\Lambda_0(T_{j-1},T_j)e^{X'\beta_0}))/(1-\exp(-\delta\Lambda_0(T_{j-1},T_j)e^{X'\beta_0}))\big|T_{j-1}=s,T_j=t\big\}$, and $\alpha_{2j}(s,t)=E\big\{(Xe^{2X'\beta_0}\exp(-\delta\Lambda_0(T_{j-1},T_j)e^{X'\beta_0}))/(1-\exp(-\delta\Lambda_0(T_{j-1},T_j)e^{X'\beta_0}))\big|T_{j-1}=s,T_j=t\big\}$. Since this holds for any $h\in\mathcal{H}(\Lambda_0)$, we conclude that

$$\begin{aligned}&\int_0^t\delta h^*(s,t)\Gamma_1(s,t)ds-\int_t^\tau\delta h^*(t,s)\Gamma_1(t,s)ds\\&\equiv\int_0^t\delta\Lambda_0(s,t)\Gamma_2(s,t)ds-\int_t^\tau\delta\Lambda_0(t,s)\Gamma_2(t,s)ds.\end{aligned}\tag{A.3}$$

Since $\delta h^*(s,t)=h^*(t)-h^*(s)$, we can rewrite (A.3) as

$$\begin{aligned}&h^*(t)\int_0^\tau\Gamma_1(s,t)I_{[0,t]}(s)+\Gamma_1(t,s)I_{[t,\tau]}(s)ds\\&\quad-\int_0^\tau h^*(s)\big[\Gamma_1(s,t)I_{[0,t]}(s)+\Gamma_1(t,s)I_{[t,\tau]}(s)\big]ds\\&\equiv\int_0^t\delta\Lambda_0(s,t)\big[\Gamma_2(s,t)I_{[0,t]}(s)+\Gamma_2(t,s)I_{[t,\tau]}(s)\big]ds.\end{aligned}\tag{A.4}$$

If we can show the left-side of (A.4) is a Fredholm operator with respect to h^* and the operator is one-to-one, then the conclusion follows from Rudin (1973). This gives the invertibility of the information operator $\mathcal{I}(\beta_0,\Lambda_0)$, given the non-singularity of the information matrix for β_0 as proved earlier.

It is easy to see that

$$\int_0^\tau\Gamma_1(s,t)I_{[0,t]}(s)+\Gamma_1(t,s)I_{[t,\tau]}(s)ds>0,$$

hence the first term of the left-side is an invertible operator with respect to h^*. Then, under (C.1), (C.2), and (C.3), we know $\Gamma_1(s,t)\big|_{0<s<t<\tau}$ is bounded. Hence, we conclude that the second term of the left-side is a compact operator with respect to h^*. Thus the left-side of (A.4) is a Fredholm operator. Finally, we only need to show it is an one-to-one operator, that is, it is 0 if and only if $h^* \equiv 0$. But it is

$$\int_0^\tau \int_0^t \delta h^*(s,t)\Gamma_1(s,t)(h(t)-h(s))ds\,dt = 0$$

for any $h \in \mathcal{H}(\Lambda_0)$. As $\Gamma_1(s,t) > 0$, we get $h^* \equiv 0$.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes a large sample study. *Ann. Statist.* **10**, 1100–1120.

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

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570–579.

Gaynes, B. N., Rush, A. J., Trivedi, M., Wisniewski, S., Spencer, D. and Fava, M. (2008). The STAR* D study: treating depression in the real world. *Cleve. Clin. J. Med.* **75**, 57–66.

Howland, R. H. (2008). Sequenced treatment alternatives to relieve depression (STAR*D) – Part 2: study outcomes. *J. Psychosoc. Nurs. Ment. Health Serv.* **46**, 21–24.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540–568.

Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Am. Statist. Assoc.* **92**, 960–967.

Kalbfleischa, J. D. and Lawlessa, J. F. (1985). The analysis of panel data under a Markov assumption. *J. Am. Statist. Assoc.* **80**, 863–871.

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

Liang, K. Y., Zeger, S. L. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–465.

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

Rush, A., Fava, M., Wisniewski, S., Lavori, P., Trivedi, M., Sackeim, H., Thase, M., Nierenberg, A., Quitkin, F., Kashner, T., Kupfer, D., Rosenbaum, J., Alpert, J., Stewart, J., Mcgrath, P., Biggs, M., Shoreswilson, K., Lebowitz, B., Ritz, L. and Niederehe, G. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials* **25**, 119–142.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, Cambridge.

Shen, X. and Wong, W. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580–615.

Sinyor, M., Schaffer, A. and Levitt, A. (2010). The sequenced treatment alternatives to relive depression (STAR*D) trial: a review. *The Canadian Journal of Psychiatry* **55**, 126–135.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica* **5**, 279–290.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Statist. Soc.* B **62**, 293–302.

Trivedi, M., Rush, A., Wisniewski, S., Nierenberg, A., Warden, D., Ritz, L., Norquist, G., Howland, R., Lebowitz, B., Mcgrath, P., Shoreswilson, K., Biggs, M., Balasubramani, G., Fava, M. and Team, S. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am. J. Psychiatry* **163**, 28–40.

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

van de Geer, S. A. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14–44.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779–814.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35**, 2106–2142.

Zeng, D., Cai, J. and Shen, Y. (2006). Semiparametric additive risks model for interval-censored data. *Statistica Sinica* **16**, 287–302.

Zhu, L., Tong, X., Sun, J., Chen, M., Srivastava, D. K., Leisenring, W. and Robison, L. L. (2014). Regression analysis of mixed recurrent-event and panel-count data. *Biostatistics* **15**, 555–568.

School of Mathematical Sciences, Beijing Normal University, Room 1232, Rear Main Building, No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China.

E-mail: liangbs@mail.bnu.edu.cn

School of Statistics, Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China.

E-mail: xweitong@bnu.edu.cn

Department of Biostatistics, Gillings School of Global Public Health, CB 7420, The University of North Carolina at Chapel Hill, 3103B McGavran-Greenberg Hall, Chapel Hill, NC 27599-7420, USA.

E-mail: dzeng@bios.unc.edu

Department of Biostatistics, Mailman School of Pubic Health, Columbia Univeresity, 722 West 168th Street, Room 211, New York, NY 10032, USA.

E-mail: yw2016@columbia.edu

(Received December 2014; accepted May 2016)