Published in final edited form as:

J Am Stat Assoc. 2018 ; 113(521): 357–368. doi:10.1080/01621459.2016.1255637.

Conditional modeling of longitudinal data with terminal event

Shengchun Kong [Senior Biostatistician],

Gilead Sciences, Inc., Foster City, CA 94404

Bin Nan [Professor],

Departments of Biostatistics, University of Michigan, Ann Arbor, MI 48109

John D. Kalbfleisch [Professor],

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

Rajiv Saran [Professor], and

Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109

Richard Hirth [Professor]

Department of Health Management and Policy, University of Michigan, Ann Arbor, MI 48109

Abstract

We consider a random effects model for longitudinal data with the occurrence of an informative terminal event that is subject to right censoring. Existing methods for analyzing such data include the joint modeling approach using latent frailty and the marginal estimating equation approach using inverse probability weighting; in both cases the effect of the terminal event on the response variable is not explicit and thus not easily interpreted. In contrast, we treat the terminal event time as a covariate in a conditional model for the longitudinal data, which provides a straight-forward interpretation while keeping the usual relationship of interest between the longitudinally measured response variable and covariates for times that are far from the terminal event. A two-stage semiparametric likelihood-based approach is proposed for estimating the regression parameters; first, the conditional distribution of the right-censored terminal event time given other covariates is estimated and then the likelihood function for the longitudinal event given the terminal event and other regression parameters is maximized. The method is illustrated by numerical simulations and by analyzing medical cost data for patients with end-stage renal disease. Desirable asymptotic properties are provided.

Keywords

Cox regression; Empirical process; Mixed effects model; Pseudo-maximum likelihood estimation

Correspondence to: Richard Hirth.

SUPPLEMENTARY MATERIAL

The online supplement contains general theorems about M-estimators, technical lemmas, and additional simulation. It also contains R code for implementing the methods developed here.

1 Introduction

In longitudinal studies, the collection of information can be stopped at the end of the study, at the time of dropout of a study participant, or at the time of a terminal event. Death, the most common terminal event, often occurs in cohort studies of older populations and in fatal disease follow-up studies, e.g., organ failure or cancer studies. Other types of terminal events also exist, for example, the final menstrual period is a terminal event for menstrual cycle data.

The current literature has primarily focused on modeling the longitudinally measured response variable and covariates given that the terminal event has not yet happened; see e.g. Tsiatis and Davidian (2004), Hsieh et al. (2006), Ding and Wang (2008), Albert and Shih (2010). If the terminal event is ignorable (Little and Rubin, 2002), then a likelihood-based estimation of regression parameters is straightforward. Oftentimes, however, the terminal event time is non-ignorable. Two types of approaches are widely used for longitudinal data analysis with non-ignorable terminal events: the joint modeling approach using latent frailty and the marginal estimating equation approach using inverse probability weighting. In the former, the relationship between the terminal event and the longitudinal data is indirectly modeled through the shared random effect. The latter approach is appropriate when the terminal event is simply censoring the observations of the longitudinal process, which is in fact continuing but unobserved; its use when the terminal event stops the longitudinal process is more controversial. Similar approaches have also been used in the context of recurrent events correlated with a terminal event; for example, see Ghosh and Lin (2002), Huang and Wang (2004), Zeng and Lin (2009), Albert and Shih (2010), Kalbfleisch et al. (2013), among many others.

These modeling strategies, however, are not as useful as one might wish for many longitudinal studies where the explicit effect of the terminal event time on the longitudinal measures is of interest. For example, medical payments in dialysis patients (Liu et al., 2007) and cancer patients (Chan and Wang, 2010) tend to increase when patients approach death; functional limitations in an aging population (Sowers et al., 2007) become more severe when people are closer to the end of life; and menstrual cycles become longer and more variable when women approach menopause (Harlow et al., 2008). In these cases, a question of interest is how does the impending terminal event affect the longitudinal measures, and for this question, a model for the longitudinal event conditional on the terminal event seems particularly useful and appropriate.

In this article, we propose a random effects model for repeated measures which includes the event time as an additional (fixed effect) covariate, and thus provides a more intuitive and meaningful interpretation of the effect of the terminal event time. The proposed conditional modeling strategy keeps the usual relationship of interest between the longitudinally measured response variable and covariates when the data collection time is far from the occurrence of the terminal event, but the response variable becomes increasingly dependent on the terminal event time when the data collection time is close to the terminal event. Since the terminal event time is subject to right censoring, the regression model with the terminal event time as a covariate falls into a general framework of regression with censored

covariate. For this situation, the complete case analysis by dropping observations with censored event times will be shown to be a valid estimating approach under the usual noninformative conditional independent censoring assumption for the censoring time.

We propose a semiparametric, likelihood-based approach for parameter estimation in a linear regression model with a nonlinear component for the censored covariate, that utilizes both the complete and censored data. The proposed method is shown to be consistent and asymptotically normal under a set of mild regularity conditions, and is more efficient than the complete case analysis. The proofs of the asymptotic properties rely heavily on empirical process theory. A referee drew our attention to Li et al. (2013), which has a similar aim of recovering information from censored data. We comment further on this work later in the Discussion Section.

The rest of the article is organized as follows. We describe the proposed model in Section 2 and the two-stage estimating method in Section 3. The asymptotic properties are outlined in Section 4 with proofs given in the Appendix. Section 5 contains numerical results followed by a brief discussion. Detailed technical preparations are provided in the online Supplementary Material.

2 A Nonlinear Regression Model with Mixed Effects and Censored Covariate

2.1 Complete data model with observed terminal event time

For a subject i , denote the terminal event time by S_i , the baseline covariates by a vector X_i where the first element is 1, the longitudinal response by Y_{ij} and the prespecified visit time by t_{ij} where $i = 1, \dots, n$ and $j = 1, \dots, n_i$. For given S_i , we model Y_{ij} with the following mixed effect model for longitudinal data:

$$Y_{ij} = X_i' \beta + g(S_i - t_{ij}, \xi) + Z_i' b_i + U_i(t_{ij}) + \varepsilon_{ij}, \quad (1)$$

where β is a vector of regression coefficients with length p_1 , b_i is an independent random effects vector of length q_1 associated with covariates Z_i , $U_i(t)$ is an independent stochastic processes, ε_{ij} , $j = 1, \dots, n_i$ are independent measurement errors, g is a known function that satisfies Condition 1 in Appendix, and ξ is a vector with length p_2 . The function $g(t, \xi) \rightarrow 0$ when $t \rightarrow \infty$ so that model (1) reduces to a simpler relationship of interest between the longitudinally measured response variable Y_{ij} and covariates X_i when t_{ij} is distant from S_i and should become increasingly related to the terminal event when t_{ij} is close to the terminal event S_i . Motivated by figure 5 in Chan and Wang (2010), we can choose $g(S - t, \xi)$ to be a normal kernel where $g(S - t, \xi) = \xi_1 e^{-(S-t-\xi_2)^2 \xi_3}$, $\xi = (\xi_1, \xi_2)$. Other examples include an exponential kernel where $g(t, \xi) = \xi_1 e^{-(t-\xi_2)}$.

We make the following additional assumptions: (i) b_i follows a normal distribution $N(0, D(\varphi))$, where D is a positive definite matrix depending on a parameter vector φ with length q_2 ; (ii) $U_i(t)$ is a mean zero Gaussian process with a given covariance function $\text{cov}(U_i(t_1), U_i(t_2)) = \kappa(\nu, \rho; t_1, t_2)$ that depends on a parameter vector ν with length q_3 and a scalar ρ ;

for example, $U_\lambda(t)$ can be the nonhomogeneous Ornstein-Uhlenbeck (NOU) process satisfying $\text{var}(U_\lambda(t)) = \nu(t)$ with $\log(\nu(t)) = \nu_0 + \nu_1 t$ and $\text{corr}(U_\lambda(t_1), U_\lambda(t_2)) = \rho^{|t_1 - t_2|}$; (iii) ε_{ij} follows a normal distribution $N(0, \sigma^2)$; and (iv) b_i , $U_\lambda(t)$, and ε_{ij} are mutually independent.

For a vector $t = (t_1, \dots, t_m)$, denote $\mathbf{g}(t, \xi) = (g(t_1, \xi), \dots, g(t_m, \xi))'$. Let $Y_i = (Y_{i1}, \dots, Y_{im})'$, $t_i = (t_{i1}, \dots, t_{im})'$, $\mathbf{X}_i = (X_i, \dots, X_i)_{p_1 \times n_i}'$ and $\mathbf{Z}_i = (Z_i, \dots, Z_i)_{q_1 \times n_i}'$. When S_i is observed, from

(1) we have

$$f_{\theta, \phi}(Y_i | S_i, X_i) = (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp \left\{ - (Y_i - \mathbf{X}_i \beta - \mathbf{g}(S_i \mathbf{1}_i - t_i, \xi))' \Sigma_i^{-1} (Y_i - \mathbf{X}_i \beta - \mathbf{g}(S_i \mathbf{1}_i - t_i, \xi)) / 2 \right\},$$

(2)

where $\mathbf{1}_i = (1, \dots, 1)'$ with length n_i , $\theta = (\beta, \xi)'$ with length $p = p_1 + p_2$, $\phi = (\varphi, \nu, \rho, \sigma^2)'$ with length $q = q_2 + q_3 + 2$, and $\Sigma_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \Gamma_i + \sigma^2 \mathbf{I}_i$, where $\mathbf{I}_i$ is the $n_i \times n_i$ identity matrix and Γ_i is the covariance matrix of $(U(t_{i1}), \dots, U(t_{im}))'$.

A semiparametric mixed effects model could also be considered, where g is an unknown function that can be estimated by smoothing splines. We focus on the parametric model (1) to more simply illustrate the proposed methodology.

2.2 Observed data model with potentially censored terminal event time

Let C_i be the censoring time for the i th subject. If $S_i < C_i$ then S_i is observed; otherwise S_i is right-censored by C_i . We denote the observed time by $V_i = \min(S_i, C_i)$ and the censoring indicator by $\Delta_i = 1(S_i < C_i)$. Note that $t_{ij} < V_i$ for all $i = 1, \dots, n$, $j = 1, \dots, n_i$. Here, we assume that C_i and (S_i, Y_i) are conditionally independent given X_i .

For notational simplicity, assume that the random effect Z is a sub-vector of X . For a single subject, we observe (V, Δ, Y, X) . The likelihood function for the observed data (V, Δ, Y, X) can be factored into

$$f_1(V, \Delta, Y, X) = f_2(V, \Delta | Y, X) f_3(Y | X) f_4(X),$$

where f_1 denotes the joint density of (V, Δ, Y, X) , f_2 denotes the conditional density of (V, Δ) given (Y, X) , f_3 denotes the conditional density of Y given X , and f_4 denotes the marginal density of X . Since the conditional independence of C and (S, Y) given X implies that C and S are conditionally independent given (Y, X) , we have

$$f_2(V, \Delta | Y, X) = \{f_S(S | Y, X) \bar{G}_C(S | Y, X)\}^\Delta \{\bar{F}_S(C | Y, X) g_C(C | Y, X)\}^{1-\Delta}, \quad (3)$$

where f_S denotes the conditional density of S given (Y, X) , g_C denotes the conditional density of C given (Y, X) , with $\bar{F}_S$ and $\bar{G}_C$ being the corresponding conditional survival functions. Further assuming noninformative censoring, we can drop $g_C(C|Y, X)$ and $\bar{G}_C(C|Y, X)$. Going through conditional arguments using the Bayes' rule and dropping $f_4(X)$, we obtain the likelihood function

$$L(V, \Delta, Y, X) = \{f_{\theta, \phi}(Y|S, X)f_5(S|X)\}^\Delta \left\{ \int_C^\infty f_{\theta, \phi}(Y|s, X)dF_5(s|X) \right\}^{1-\Delta}, \quad (4)$$

where $f_5(S|X)$ is the conditional density of S given X , and $F_5(S|X)$ is the corresponding cumulative distribution function. In (4), only $f_{\theta, \phi}$ contains the parameter of interest θ and nuisance parameter ϕ , whereas F_5 (or f_5) is an additional nuisance parameter.

In (4), $\{f_{\theta, \phi}(Y|S, X)f_5(S|X)\}$ is for a subject with observed terminal event time, which yields the fully observed data likelihood, and $\{\int_C^\infty f_{\theta, \phi}(Y|s, X)dF_5(s|X)\}^{1-\Delta}$ is for a subject with censored terminal event time. In section 4, we show that the complete case analysis by dropping the second part in (4) yields a consistent and asymptotically normally distributed estimator, but is inefficient compared to an approach that also utilizes the censored data. From the second part in (4), we see that the amount of efficiency gain depends on how well we can estimate the right tail of the conditional distribution $F_5(s|X)$ beyond C . We consider a semiparametric approach that allows reliable extrapolation beyond C and is robust against any parametric assumption.

Since C_i and (S_i, Y_i) are conditionally independent given X_i and C_i is random, all the commonly used semiparametric models for right-censored data allow extrapolation beyond C_i . Here, we propose the most widely used Cox regression model (Cox, 1972). Other viable models include the accelerated failure time model, the additive hazard model, and the transformation model (Kalbfleisch and Prentice, 2002). Suppose the hazard function of S given X has the following form:

$$\lambda(s|X) = \lambda(s) \exp(\alpha'X), \quad (5)$$

where α is the regression parameter with an unknown true value α_0 , and $\lambda(\cdot)$ is the baseline hazard function. The conditional cumulative distribution function is then given by

$$\eta(s; X) \equiv F_5(s|X) = 1 - \exp\{-\Lambda(s) \exp(\alpha'X)\},$$

where $\Lambda(s) = \int_0^s \lambda(u)du$ is the cumulative baseline hazard function with an unknown true value Λ_0 . Note that X appears in both models (2) and (5), but these two instances may refer to different regressions. For example, X_1 might be a covariate in (2) whereas X_1^2 is a covariate in (5). The same X is used to denote all fully observed covariates for notational simplicity. The log-likelihood function then becomes

$$\log L = \Delta \log f_{\theta, \phi}(Y|S, X) + \Delta \log \dot{\eta}(S; X) + (1 - \Delta) \log \int_C^\infty f_{\theta, \phi}(Y|u, X) d\eta(u; X). \quad (6)$$

A similar idea has been used by Lu et al. (2010), but for a different problem. Lu et al. (2010) considered longitudinal data analysis with an event time, which does not terminate the observed data.

3 The Pseudo-likelihood Method

The log likelihood function (6) involves an unknown distribution function η and the corresponding density function $\dot{\eta}$. Hence a maximum likelihood estimate, if it exists, can be complicated. We propose a tractable two-stage pseudo-likelihood approach in which the nuisance parameters (ϕ, η) are estimated in stage 1, and the parameter of interest θ is then estimated by maximizing (6) in stage 2 with nuisance parameters replaced by their estimators obtained in stage 1 (Kong and Nan, 2016). Details are given below:

Stage 1. Nuisance parameter estimation. The dispersion parameter ϕ is estimated by the complete case analysis of the nonlinear regression model (2); the Cox model regression coefficient α is estimated by maximizing the partial likelihood, and the cumulative baseline hazard Λ is estimated with the Breslow estimator (Breslow, 1972). Denote the estimators by $\tilde{\phi}_n$, $\tilde{\alpha}_n$ and $\tilde{\Lambda}_n$ respectively. The c.d.f $\eta(s; X)$ is estimated by $\tilde{\eta}_n(s; X) = 1 - \exp\{-\tilde{\Lambda}_n(s) \exp(\tilde{\alpha}'_n X)\}$, which is asymptotically equivalent to the product integral expression. It can be shown that all the estimates obtained in Stage 1 have desirable statistical properties. In particular, $\tilde{\eta}_n$ is $n^{1/2}$ -consistent in a finite interval, see Lemma A.3 in Supplementary Material; $\tilde{\phi}_n$ obtained from the complete case analysis is $n^{1/2}$ -consistent, see Theorem 4.1 in section 4.

Stage 2. Pseudo-likelihood estimation of θ . Replacing (ϕ, η) by their Stage 1 estimates $(\tilde{\phi}_n, \tilde{\eta}_n)$ in the log likelihood function yields the following log pseudo-likelihood function for a random sample of n subjects:

$$pl_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left\{ \Delta_i \log f_{\theta, \tilde{\phi}_n}(Y_i|S_i, X_i) + (1 - \Delta_i) \log \int_{C_i}^\infty f_{\theta, \tilde{\phi}_n}(Y_i|u, X_i) d\tilde{\eta}_n(u; X_i) \right\}$$

(7)

Note that the term $\log \dot{\eta}$ in (6) is dropped because it does not involve θ . However, if one wants to maximize the log-likelihood directly without using the two-stage approach, then this term cannot be omitted.

Let $\hat{\theta}_n$ denote the pseudo-likelihood estimator. Since it is obtained by maximizing the objective function (7), its asymptotic properties can be obtained from M-estimation theory, see van der Vaart (2002), Wellner and Zhang (2007) and Li and Nan (2011).

The estimates $(\tilde{\eta}_n, \tilde{\Lambda}_n)$ are obtained using a standard package for the Cox regression model. The estimates $(\tilde{\theta}_n, \tilde{\phi}_n)$ from complete case analysis are obtained by maximizing $\frac{1}{n} \sum_{i=1}^n \{\Delta_i \log f_{\theta, \phi}(Y_i | S_i, X_i)\}$ using a Newton-Raphson algorithm, where multiple initial values are tried. The two-stage estimator $\hat{\theta}_n$ is also obtained from a Newton-Raphson algorithm with the complete case analysis estimator $\tilde{\theta}_n$ as the initial value.

4 Asymptotic Properties

Let $l_0(\theta, \phi; Y, X, \Delta, V) = \log f_{\theta, \phi}(Y | S, X)$. This is the first part in the log-likelihood for the observed data. Then

$$\begin{aligned} l(\theta, \phi, \eta; Y, X, \Delta, V) &\equiv l(\theta, \phi, \eta(\alpha, \Lambda); Y, X, \Delta, V) = l_0(\theta, \phi; Y, X, \Delta, V) + (1 - \Delta) \log \int_C f_{\theta, \phi}(Y | u, X) d\eta(u; X) \\ &= l_0(\theta, \phi; Y, X, \Delta, V) + (1 - \Delta) \log \int_C f_{\theta, \phi}(Y | u, X) d[1 - \exp\{-\Lambda(u) \exp(\alpha' X)\}], \end{aligned}$$

which is (6) with $\log \dot{\eta}$ dropped.

A set of regularity conditions is introduced in the Appendix. Some conditions are commonly assumed for the Cox regression model; other conditions are for the mixed effects model, which are easily verified for a smooth function g and the NOU process. We will use standard empirical process notation from now on. In particular, $\mathbb{P}_n$ is the empirical measure and $Pf = \int f dP$ for a probability measure P and a function f .

Under the conditional independent censoring assumption, the estimators from the complete case analysis and the two-stage procedure, respectively, are consistent and asymptotically normal. These results are given in the following two theorems.

Theorem 4.1. (Complete case)

Assume that C and (S, Y) are independent given X . Under Conditions 1, 2(a), and 3–5, the complete case analysis estimator $(\tilde{\theta}_n, \tilde{\phi}_n)$ that maximizes $\mathbb{P}_n l_0(\theta, \phi; Y, X, \Delta, V)$ converges in outer probability to (θ_0, ϕ_0) ; and $\sqrt{n}((\tilde{\theta}_n, \tilde{\phi}_n) - (\theta_0, \phi_0))$ converges in distribution to a mean zero normal random variable with variance $J_1^{-1} Q_1 J_1^{-1}$, where J_1 and Q_1 are provided in the Appendix.

Theorem 4.2. (Two-stage)

Assume that C and (S, Y) are independent given X . Under Conditions 1–8, the two-stage pseudo-likelihood estimator $\hat{\theta}_n$ that maximizes (7) converges in outer probability to θ_0 ; and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a mean zero normal random variable with variance $J_2^{-1} Q_2 J_2^{-1}$, where J_2 and Q_2 are defined in the Appendix

The proof of consistency is similar to Li and Nan (2011) and van der Vaart (2002). The proof of asymptotic normality is given in the appendix, and is based on the general M-estimation theory similar to Li and Nan (2011) and Wellner and Zhang (2007). The detailed proof relies heavily on empirical process theory and is given in the Appendix.

Because the asymptotic variance of $\hat{\theta}_n$ has a very complicated expression that does not yield a simply computed estimate from the observed data, we use the bootstrap variance estimator.

5 Numerical Results

5.1 Simulations

We conduct simulations to investigate the finite sample performance of the proposed method. Simulation data sets are generated from the nonlinear model with mixed effects,

$$Y_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma \exp\{-(S_i - T_{ij} - \mu)^2 \xi\} + b_i + U_i(T_{ij}) + \varepsilon_{ij},$$

where $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = -3$, $\mu = 1$, and $\gamma = 4$. The random effect $b_i \sim \mathcal{N}(0, \exp(-0.5))$, the error term $\varepsilon_{ij} \sim \mathcal{N}(0, \exp(-0.1))$, and $U_i(t)$ is an NOU process with $v_0 = 1$, $v_1 = -1$ and $\rho = \exp(-1)/(1 + \exp(-1))$. The two fully observed covariates are X_{1i} and X_{2i} , where $X_{1i} \sim \text{Bernoulli}(0.5)$ and $X_{2i} \sim \mathcal{N}(0, 1)$ truncated at ± 3 . The terminal event time is $S_i = 4 + S_{0i}$, where S_{0i} follows an exponential distribution with conditional hazard function $\exp(-1 - 6X_{1i} + 4X_{2i})$. To generate the censoring time C_i , we first generate $C_{0i} = \kappa C_{0i}^*$, where C_{0i}^* follows an exponential distribution with conditional hazard function $\exp(-3 - X_{1i} + X_{2i})$, then set $C_i = t_{ij}$, where j satisfies $t_{ij} - C_{0i}$ and $t_{ij+1} > C_{0i}$ assuming $t_{in+1} = \infty$. The constant κ is chosen to yield 40% censoring. For each subject i , there are 10 scheduled visit times, and the first visit time t_{i1} is 0. There are two different settings to generate the subsequent visit times: (1) equally spaced time intervals with $t_{ij} = j - 1$, $j = 2, \dots, 10$; (2) non-equally spaced time intervals with the subsequent visit times generated recursively from $t_{ij} = t_{ij-1} + \min(4, W_j)$ for $j = 2, \dots, 10$, where W_j follows an exponential distribution with conditional hazard function $\exp(-3 - X_{1i} + X_{2i})$. In each setting, ξ takes two different values, 1.2 and 0.2, corresponding to a flat and a sharp nonlinear predictor in the regression model, respectively.

We simulate 500 replications for each scenario with sample size 300. The biases and variances of the proposed method are compared with those of full data and complete case analyses. The full data analysis represents the case that all data are available; in other words, there is no censoring, which has more visits and serves as a benchmark. The complete case analysis simply eliminates subjects with censored terminal event time. For the proposed two-stage method, we report the 90% and 95% coverage proportions for which the variances estimators are obtained from 100 bootstrap samples. The results are presented in Tables 1–4.

The results suggest that the biases for the proposed two-stage method are minimal and comparable to both the full data analysis and the complete case analysis. From the tables, it can be seen that the proposed method is much more efficient than the complete case analysis, and the bootstrap method performs well in estimating the variance, yielding reasonable coverage rates for all the scenarios.

We run additional simulations to further investigate the impact of survival model misspecification in (5). The results are provided in the Supplementary Material, where it is shown that misspecification of the Cox regression model can yield biased results, and that the bias increases as the severity of misspecification of (5) grows; this indicates the importance of model-checking before implementing the proposed two-stage method.

5.2 End-stage renal disease

We consider data on inpatient hospital costs of patients with end-stage renal disease (ESRD) as reported in an analysis file provided by the United States Renal Data System (USRDS); this provides an illustrative example of longitudinal data with a terminal event. These costs are of substantial interest, since Medicare paid about \$10.5 billion in 2012 for inpatient costs (USRDS 2014 annual report). We focus on the monthly inpatient costs paid by Medicare; these costs are terminated by the occurrence of death, and Chan and Wang (2010) and Liu et al. (2007) suggested that the medical payment pattern changes when patients approach death. We explore this issue taking account of patient level covariates.

For illustrative purposes, we selected a 2% random sample of the white and black patients whose service started in the calendar year 2007, and who were 65 years or older at baseline. The average age at baseline was 76.1 and follow-up ended on December 31st, 2010. Of the 840 patients selected for analysis, 65.5% died during the follow-up period. Others were censored through loss to follow up or at the end of the study. The average follow-up time for medical payment was 23.4 months. For convenience, we assume that the inpatient cost rate is constant within each hospitalization. For example, if a hospitalization starts from April 21st and ends on May 10th with the amount \$3,000, then the Medical payment is \$2,000 for April and \$1,000 for May. Usually the month of death is shorter than other months. For example, a death on April 15th only has a half month to accrue spending. We consider “spending rate” for the month in which death occurs. For example, for a death on April 15th with the April Medical payment amount \$3,000, we scale up the payment to \$6,000 for that month in the analysis. Age, log transformed body mass index (BMI), heart disease and lung disease are used as predictors for the death hazard. All of them are significant with p-values < 0.0001 , 0.0011 , < 0.0001 and 0.0127 , respectively. The goodness of fit for the Cox regression model is checked in Figure 1. Dotted lines in the first row of Figure 1 are the plots of 20 realizations from the distributions of the score processes. The observed score processes are presented with solid lines which randomly fluctuate around zero. From Figure 1, we see that the proportional hazards model for age and log transformed BMI fits the data reasonably well, with respective goodness-of-fit empirical p-values of 0.485, and 0.284, respectively, based on 1000 simulated martingale residual score processes (Lin et al., 1993). A plot of $\log \hat{\Lambda}_0(t)$ versus $\log t$ is displayed in the lower panel of Figure 1. The approximate parallelism of the curves suggests that the proportional hazards model for lung disease and heart disease provides a reasonably good approximation, all except for early times with lung disease.

Since the distribution of monthly Medicare payment, Y , is highly skewed, we consider a log transformation $\log(Y/1000 + 1)$. Figure 2 shows the final six-month trajectories of monthly inpatient costs (log transformed) for 30 randomly selected patients who died during follow-

up (dotted lines). Many show an increasing and then decreasing pattern before death. We consider a normal kernel in the nonlinear mixed model.

Exploration of the data showed a similar pattern after entry as described in Liu et al. (2007). Inpatient costs tended to increase over the first two months after entry, and then showed an approximately linear decreasing pattern through to the eighth month. Hence, we create three variables to capture this effect, where $Start1 = 1(Month = 1)$, $Start2 = Month \times 1(2 \leq Month \leq 7)$ and $Start3 = 1(Month = 8)$. Diabetes, heart disease and race are also the covariates of interest, whereas age, BMI, sex and lung disease are not significantly associated. The final models are

$$\log(Y_i(t_{ij})/1000 + 1) = \beta_0 + \beta_1 Start1 + \beta_2 Start2 + \beta_3 Start3 + \beta_4 Diabetes + \beta_5 Heart + \beta_6 Race + \gamma \exp\{-\xi(S_i - t_{ij} - \mu)^2\} + b_i + U_i(t_{ij}) + \varepsilon_{ij}$$

$$\lambda_i(s) = \lambda_0(s) \exp(\alpha_1 Age + \alpha_2 \log(BMI) + \alpha_3 Lung + \alpha_4 Heart).$$

Table 5 shows the regression coefficient estimates, where we see that the proposed two-stage method yields similar point estimates with smaller estimated variances compared to the complete case analysis, indicating the efficiency gain of the proposed method.

The estimated averages of log transformed Medicare payments are presented with a solid line in Figure 2. A Q-Q plot is used to check the normal error assumption, where only one residual is randomly selected for each patient to avoid the correlation within subjects, see Figure 3. The linear pattern is consistent with the nonlinear mixed model with normal kernel. We checked many random selections, and the plot is typical.

6 Discussion

We consider the identity link and a Gaussian error in this article. The proposed two-stage method could be generalized to non-Gaussian error, logistic or Poisson regression provided the model is identifiable and regularity conditions are suitably modified.

We allow only time-independent covariates in this article for simplicity. Time dependent covariates are often of interest in longitudinal data analysis and survival analysis. Implementation of the two-stage method for time-dependent covariates involves extrapolating $\eta(u, \bar{X}(C))$ beyond C , where $\bar{X}(C)$ is the history of the time-dependent covariates X up to time C . It involves predicting the censored covariate process, which will be explored elsewhere. An alternative is an estimating equation approach using inverse probability weighting, which would only use the subjects with observed terminal event time.

The function $g(S - t, \xi)$ in (1) that we consider in this article is a known nonlinear function up to the parameter ξ . In practice, smoothing techniques can be used to determine an appropriate parametric functional form of g or to examine the fit of the data to a hypothesized g . For example, we fitted the model (1) but approximated g with cubic B-

splines with 20 knots over the entire observation window of 48 months. This yielded an estimate of g that was very similar to the proposed Gaussian form.

We only considered the intercept parameter as a function of $S - t$ in this article to illustrate the basic concept of the proposed methodology. This modeling strategy extends naturally to regression models with a time-varying coefficient for each regressor. Such an extension is under investigation.

The major difference between our work and that of Li et al. (2013) is that the function g in model (1), together with β_0 , can be viewed as the intercept parameter which depends on $S - t$, but all other variables in the model are with reference to time t in the same way as in the usual regression models for longitudinal data. On the other hand, none of the regression parameters in Li et al. (2013) varies with $S - t$, but all the variables in their model (including the error terms) are with reference to the reverse time scale, $S - t$. See their equation (2). This leads to different model interpretations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The data used in this paper were made available by the U.S. Renal Data System. This study was supported in part by the U.S. Renal Data System under Contract No. NO1-DK-9-2344 (National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland). The data analysis was completed while Shengchun Kong was Assistant Professor of Statistics at Purdue University.

The research is supported in part by NIH grant R01-AG036802 and NSF grants DMS-1007590 and DMS-1407142, and with Federal funds from the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN276201400001C. The data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government.

A Appendix

A.1 Regularity conditions

Denote the true value of θ by θ_0 , the true value of ϕ by ϕ_0 , the sample space of response variable Y by $\mathcal{Y}$, the sample space of covariate X by $\mathcal{X}$, the sample space of random effect Z by $\mathcal{Z} \subset \mathcal{X}$, the parameter space of θ by Θ , the parameter space of ϕ by Φ , and the parameter space of η by $\mathcal{A}$. In addition to the assumptions of bounded support for X , bounded parameter spaces Θ and Φ , conditional independence between C and (S, Y) given X , and non-informative censoring, we provide a set of regularity conditions in the following:

Condition 1

The third derivatives $|^3 g(t, \xi) / (\xi_i \xi_j \xi_k)|$ and $|^3 g(t, \xi) / (t \xi_j \xi_k)|$ are bounded uniformly for all $\xi \in \Xi$ and bounded t .

Condition 2

- a. $P_{l0}(\theta, \phi; Y, X, \cdot, V)$ has a unique maximizer (θ_0, ϕ_0) ;
- b. $P_{ll}(\theta, \phi_0, \eta_0; Y, X, \cdot, V)$ has a unique maximizer θ_0 .

Condition 3

The eigenvalues for $\Sigma(\phi)$ are bounded between $[\lambda_1, \lambda_2]$, where $0 < \lambda_1 < \lambda_2 < \infty$ for any $\phi \in \Phi$ and $Z \in \mathcal{Z}$.

Condition 4

The absolute values of all the elements in ${}^3\Sigma(\phi)/(\phi_i \phi_j \phi_k)$ are bounded uniformly for all $\phi \in \Phi$ and $Z \in \mathcal{Z}$.

Condition 5

The study stops at a finite time $\tau > 0$ such that $\inf_{x \in \mathcal{X}} P(C = \tau | X = x) = \omega_1 > 0$ and $\inf_{x \in \mathcal{X}} P(S = \tau | X = x) = \omega_2 > 0$ for constants ω_1 and ω_2 .

Condition 6

The conditional distribution of S given X possesses a continuous Lebesgue density.

Condition 7

The information matrix of the partial likelihood for the Cox regression model at the true parameter values is positive definite.

Condition 8

There exist constants $\delta_1 > 0$ and $\delta_2 > 0$, such that $\int_C f_{\theta, \phi}(Y|s, X) d\eta(s; X) \geq \delta_1$ with probability 1 for any $\theta \in \Theta$ and $|\phi - \phi_0| + \|\eta - \eta_0\| < \delta_2$.

REMARK

Condition 1 holds for many smooth function g , e.g. $g(t, \xi) = \xi_1 \exp\{(t - \xi_2)^2 \xi_3\}$ or $g(t, \xi) = \xi_1 \exp\{-(t - \xi_2)\}$. Bounded third derivatives implies bounded second derivatives, which is adequate for the proof of consistency. We implemented the numerical studies with g being the normal kernel. When $g(t, \xi) = \xi_1 \exp\{(t - \xi_2)^2 \xi_3\}$, Condition 2(a) implies $\xi_1 \xi_3 > 0$; by Theorem 2.1 of Lehmann (1998), this condition holds provided model (1) is identifiable. Condition 2(b) is for the consistency of the proposed two-stage estimator $\hat{\theta}_n$, which may be unnecessarily strong as can be seen from the following. In the proof of Theorem 4.2, we can show $P_{l1}^{\ddot{}}(\theta_0, \phi_0, \eta_0; Y, X, \cdot, V) = P\{\{ {}^2l(\theta, \phi_0, \eta_0; Y, X, \cdot, V) / (\theta - \theta') \}_{\theta=\theta_0}\}$ is negative definite by Condition 2(a). Thus $P_{l1}^{\ddot{}}(\theta, \phi_0, \eta_0; Y, X, \cdot, V)$, a continuous matrix of θ , is also negative definite in a neighborhood of θ_0 , which guarantees that θ_0 is a unique maximizer of $P_{ll}(\theta, \phi_0, \eta_0; Y, X, \cdot, V)$ in a neighborhood of θ_0 . The initial value we use in the algorithm for maximizing (7) is obtained from the complete case analysis, which is shown to be $n^{1/2}$ -consistent; thus, the solution of the proposed two-stage method is likely to be in the same neighborhood, and therefore also consistent without the uniqueness requirement in Condition 2(b).

Conditions 3–4 automatically hold for model (1) with the NOU process if $|\rho| = 1 - \delta$, and $t_{i,k+1} - t_{i,k} \leq \varepsilon$, $i = 1, \dots, n$, $k = 1, \dots, n_i - 1$, where $\delta > 0$ and $\varepsilon > 0$; they are parallel to the conditions of bounded derivatives of the log likelihood in Theorem 1.1 and Theorem 2.3 of Lehmann (1998).

Conditions 5–7 are usual assumptions for Cox regression models (Andersen and Gill, 1982; Nan and Wellner, 2013). From Condition 5, we have

$$l(\theta, \phi, \eta; Y, X, \Delta, V) = \Delta \log f_{\theta, \phi}(Y|S, X) + (1 - \Delta) \log \int_C f_{\theta, \phi}(Y|u, X) d[1 - \exp\{-\Lambda(u)\} \exp(\alpha'X)].$$

(8)

Condition 8 is mainly for technical convenience. One way to obtain Condition 8 might be to truncate the response variable Y such that $|Y| \leq M < \infty$ for a large constant M . In our simulations, however, we do not implement such truncation but still obtain satisfactory results.

A.2 Proofs of Theorem 4.1 and 4.2

All the Lemmas A.1 – A.5 used in the following proofs are provided in the online Supplementary Material.

A.2.1 Proof of consistency in Theorem 4.1 for complete case analysis estimator

Proof—From Corollary 3.2.3 in van der Vaart and Wellner (1996), we need to show that (i) $P l_0(\theta_0, \phi_0; Y, X, \Delta, V) > \sup_{(\theta, \phi) \in G} P l_0(\theta, \phi; Y, X, \Delta, V)$ for any open set G that contains (θ_0, ϕ_0) ; (ii) $\sup_{(\theta, \phi)} \|(\mathbb{P}_n - P) l_0(\theta, \phi; Y, X, \Delta, V)\| \rightarrow 0$. Condition (i) is satisfied from Condition 2(a) and non-informative censoring assumption. Condition (ii) is satisfied because the class of functions $\{-(Y - \mathbf{X}\beta - \mathbf{g}(S1 - t, \xi))' \Sigma(\phi)^{-1} (Y - \mathbf{X}\beta - \mathbf{g}(S1 - t, \xi))/2 - \log |\Sigma(\phi)|/2 : \theta \in \Theta, \phi \in \Phi\}$ belongs to Glivenko-Cantelli from Lemma A.4.

A.2.2 Proof of asymptotic normality in Theorem 4.1 for complete case analysis estimator

Denote the element-wise product of two matrices A and B by $A * B$. Let

$$A_{jj}(\phi) = \partial \sum (\phi) l \partial \phi_j, \quad A_{jk}(\phi) = \partial^2 \sum (\phi) l (\partial \phi_j \partial \phi_k), \quad j = 1, \dots, q, \quad k = 1, \dots, q;$$

$$r(\theta; V, Y, X) = Y - \mathbf{X}\beta - \mathbf{g}(S1 - t, \xi).$$

Proof—The proof follows Lemma A.1 with $\psi = (\theta, \phi)$. Here

$$m(\theta, \phi; Y, X, \Delta, V) = l_0(\theta, \phi; Y, X, \Delta, V).$$

The first order derivative of $l_0(\theta, \phi; Y, X, \Delta, V)$ equals

$$\dot{l}_0(\theta, \phi; Y, X, \Delta, V) = \begin{pmatrix} \dot{l}_{01}(\theta, \phi; Y, X, \Delta, V) \\ \dot{l}_{02}(\theta, \phi; Y, X, \Delta, V) \end{pmatrix},$$

where

$$\dot{l}_{01}(\theta, \phi; Y, X, \Delta, V) = \partial l_0(\theta, \phi; Y, X, \Delta, V) / \partial \theta = \Delta D_2(\theta; V, X)' \sum (\phi)^{-1} r(\theta; V, Y, X)$$

with

$$D_2(\theta; V, X) = (X, \partial \mathbf{g}(V1 - t, \xi) / \partial \xi), \quad (9)$$

and

$$\begin{aligned} \dot{l}_{02}(\theta, \phi; Y, X, \Delta, V) &= \partial l_0(\theta, \phi; Y, X, \Delta, V) / \partial \phi = C(\theta, \phi; Y, X, \Delta, V) \quad (10) \\ &= (C_1(\theta, \phi; Y, X, \Delta, V), \dots, C_q(\theta, \phi; Y, X, \Delta, V))' \end{aligned}$$

with

$$\begin{aligned} C_j(\theta_0, \phi_0; Y, X, \Delta, V) &= -\Delta \text{tr} \left[\sum (\phi_0)^{-1} A_j(\phi_0) \right] / 2 \quad (11) \\ &+ \Delta r(\theta_0; V, Y, X)' \sum (\phi_0)^{-1} A_j(\phi_0) \sum (\phi_0)^{-1} r(\theta_0; V, Y, X) / 2. \end{aligned}$$

The second order derivative of $l_0(\theta, \phi; Y, X, \Delta, V)$ equals

$$\ddot{l}_0(\theta, \phi; Y, X, \Delta, V) = \begin{pmatrix} \ddot{l}_{011}(\theta, \phi; Y, X, \Delta, V) & \ddot{l}_{021}(\theta, \phi; Y, X, \Delta, V) \\ \ddot{l}_{021}(\theta, \phi; Y, X, \Delta, V) & \ddot{l}_{022}(\theta, \phi; Y, X, \Delta, V) \end{pmatrix},$$

where

$$\ddot{l}_{011}(\theta, \phi; Y, X, \Delta, V) = \partial^2 l_0(\theta, \phi; Y, X, \Delta, V) / \{\partial \theta \partial \theta'\} = -\Delta D_2(\theta; V, X)' \sum^{-1} (\phi) D_2(\theta; V, X) + \Delta D_3(\theta, \phi; V, Y, X)$$

with

$$D_{3jk}(\theta, \phi; V, Y, X) = \begin{cases} 0, j \leq p_1 \text{ or } k \leq p_2 \\ \left(\frac{\partial^2 \mathbf{g}(V1 - t, \xi)}{\partial \xi_{j-p_1} \partial \xi_{k-p_2}} \right), \sum (\phi)^{-1} r(\theta; V, Y, X), & j > p_1 \text{ and } k > p_2, \end{cases} \quad (12)$$

$$\ddot{l}_{021}(\theta, \phi; Y, X, \Delta, V) = \partial^2 l_0(\theta, \phi; Y, X, \Delta, V) / \{ \partial \phi \partial \theta' \} = (\ddot{l}_{0211}(\theta, \phi; Y, X, \Delta, V), \dots, \ddot{l}_{021q}(\theta, \phi; Y, X, \Delta, V))'$$

with

$$\ddot{l}_{021j}(\theta, \phi; Y, X, \Delta, V) = -\Delta D_2(\theta; V, X)' \sum (\phi)^{-1} A_j(\phi) \sum (\phi)^{-1} r(\theta; V, Y, X),$$

and

$$\ddot{l}_{022}(\theta, \phi; Y, X, \Delta, V) = \partial^2 l_0(\theta, \phi; Y, X, \Delta, V) / \{ \partial \phi \partial \phi' \} = \begin{pmatrix} \ddot{l}_{02211}(\theta, \phi; Y, X, \Delta, V) & \dots & \ddot{l}_{0221q}(\theta, \phi; Y, X, \Delta, V) \\ \vdots & & \vdots \\ \ddot{l}_{022q1}(\theta, \phi; Y, X, \Delta, V) & \dots & \ddot{l}_{022qq}(\theta, \phi; Y, X, \Delta, V) \end{pmatrix}$$

with

$$\begin{aligned} \ddot{l}_{022jk}(\theta, \phi; Y, X, \Delta, V) = & -\Delta \text{tr} [- \sum (\phi)^{-1} A_j(\phi) \sum (\phi)^{-1} A_k(\phi) + \sum (\phi)^{-1} A_{jk}(\phi)] / 2 \\ & - \Delta r(\theta; V, Y, X)' \sum (\phi)^{-1} \{ A_j(\phi) \sum (\phi)^{-1} A_k(\phi) - A_{jk}(\phi) + A_k(\phi) \sum (\phi)^{-1} A_j(\phi) \} \sum (\phi)^{-1} r(\theta; V, Y, X) / 2. \end{aligned}$$

Condition A1 holds from consistency. Condition A2 holds since for any u ,

$$\int_{-\infty}^{\infty} f_{\theta_0, \phi_0}(y|u, x) r(\theta_0; u, y, x) dy = 0, \quad (13)$$

$$\int_{-\infty}^{\infty} f_{\theta_0, \phi_0}(y|u, x) r(\theta_0; u, y, x) r(\theta_0; u, y, x)' dy = \sum (\phi_0). \quad (14)$$

We have

$$P\ddot{l}_0(\theta_0, \phi_0; Y, X, \Delta, V) = \begin{pmatrix} D_4(\theta_0, \phi_0) & 0 \\ 0 & D_5(\phi_0) \end{pmatrix},$$

where

$$D_4(\theta_0, \phi_0) = -P\{\Delta D_2(\theta_0; V, X)' \sum (\phi_0)^{-1} D_2(\theta_0; V, X)\},$$

$$D_5(\phi_0) = \begin{pmatrix} D_{511}(\phi_0) & \dots & D_{51q}(\phi_0) \\ \vdots & & \vdots \\ D_{5q1}(\phi_0) & \dots & D_{5qq}(\phi_0) \end{pmatrix}$$

with

$$\begin{aligned} D_{5jk}(\phi_0) &= -P\left\{\Delta \text{tr}\left[\sum(\phi_0)^{-1}A_k(\phi_0)\sum(\phi_0)^{-1}A_j(\phi_0)\right]/2\right\} \\ &= -P\left\{\Delta \text{tr}\left[\sum(\phi_0)^{-1/2}A_k(\phi_0)\sum(\phi_0)^{-1}A_j(\phi_0)\sum(\phi_0)^{-1/2}\right]/2\right\}. \end{aligned}$$

Hence,

$$D_5(\phi_0) = -P\{\Delta D_1(\phi_0; X)'D_1(\phi_0; X)/2\}, \quad (15)$$

where

$$D_1(\phi_0; X)' = \begin{pmatrix} \text{vec}(\sum(\phi_0)^{-1/2}A_1(\phi_0)\sum(\phi_0)^{-1/2})' \\ \vdots \\ \text{vec}(\sum(\phi_0)^{-1/2}A_q(\phi_0)\sum(\phi_0)^{-1/2})' \end{pmatrix}.$$

Thus, $Pn\tilde{m}(\theta_0, \phi_0; Y, X, \Delta, V)$ is negative definite from Condition 2(a).

From (13), we have Condition A3 holds. Condition A4 holds automatically. Condition A5 holds if the class of functions $\{-\text{tr}[\Sigma(\phi)^{-1}A_j(\phi)]/2 + \text{tr}(\theta, V, Y, X)' \Sigma(\phi)^{-1}A_j(\phi) \Sigma(\phi)^{-1} \text{tr}(\theta, V, Y, X)/2 : j = 1, \dots, q, |\theta - \theta_0| < \delta, |\phi - \phi_0| < \delta\}$ is Donsker for some $\delta > 0$ and satisfies $P\|\tilde{m}(\theta, \phi; Y, X, \Delta, V) - \tilde{m}(\theta_0, \phi_0; Y, X, \Delta, V)\|^2 \rightarrow 0$ as $|\theta, \phi - (\theta_0, \phi_0)| \leq \delta_n \downarrow 0$. These two conditions hold from Conditions 1, 3–5, and Theorem 2.10.6 of van der Vaart and Wellner (1996). Condition A6 holds from Taylor expansion and Conditions 1 and 3–5. Hence,

$$\sqrt{n}(\tilde{\theta}_n, \tilde{\phi}_n) - (\theta_0, \phi_0) = -[P\tilde{l}_0(\theta_0, \phi_0; Y, X, \Delta, V)]^{-1}P\tilde{l}_n(\theta_0, \phi_0; Y, X, \Delta, V) + o_p^*(1),$$

which converges weakly to a mean zero normal random variable with variance $J_1^{-1}Q_1J_1^{-1}$, where $J_1 = -P\tilde{l}_0(\theta_0, \phi_0; Y, X, \Delta, V)$ and $Q_1 = P\{\dot{l}_0(\theta_0, \phi_0; Y, X, \Delta, V)^{\otimes 2}\}$. Furthermore,

$$\sqrt{n}(\tilde{\phi}_n - \phi_0) = D_5(\phi_0)^{-1}\sqrt{n}P_n C(\theta_0, \phi_0; Y, X, \Delta, V) + o_p(1), \quad (16)$$

where $D_5(\phi_0)$ and $C(\theta_0, \phi_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V})$ are defined in (15) and (10), respectively.

A.2.3 Proof of consistency in Theorem 4.2 for two-stage estimator

Proof—From Condition 2(b), we have

$$\sup_{d(\theta, \theta_0) > \delta} Pl(\theta, \phi_0, \eta_0; Y, X, \Delta, V) < Pl(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) \quad (17)$$

holds for every $\delta > 0$. By the definition of $\hat{\theta}_n$, we have

$$\mathbb{P}_n l(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) \geq \mathbb{P}_n l(\theta_0, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) = \mathbb{P}_n l(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1),$$

(18)

where the equality is obtained by Lemma A.4 and Lemma A.5. The class of functions $\{l(\theta, \phi, \eta; Y, X, \Delta, V) : \theta \in \Theta, \phi \in \Phi, \eta \in \mathcal{F}\}$ is Donsker from Lemma A.4. Hence it is Glivenko-Cantelli, and we then have

$$\begin{aligned} 0 &\leq Pl(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) - Pl(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) \\ &= \mathbb{P}_n l(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) - \mathbb{P}_n l(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1) \\ &\leq \mathbb{P}_n l(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) - \mathbb{P}_n l(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1) \quad (19) \end{aligned}$$

$$= Pl(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) - Pl(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1)$$

$$= o_p(1), \quad (20)$$

where (19) is obtained from (18) and (20) is obtained by Lemma A.5. By inequality (17), for every $\delta > 0$ we have

$$\{d(\hat{\theta}_n, \theta_0) \geq \delta\} \subset \{Pl(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) < Pl(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)\},$$

with the sequence of the events on the right going to a null event in view of inequality (20), which yields the almost sure (thus in probability) convergence of $\hat{\theta}_n$. This argument is taken from the proof of Theorem 5.8 in van der Vaart (2002) and the proof of Theorem 3 in Li and Nan (2011).

A.2.4 Proof of asymptotic normality in Theorem 4.2 for two-stage estimator

Proof—The proof follows Lemma A.2. Here

$$m(\theta, \phi, \eta; Y, X, \Delta, V) = l(\theta, \phi, \eta; Y, X, \Delta, V).$$

The partial derivative of $l(\theta, \phi, \eta; Y, X, \Delta, V)$ with respect to θ equals

$$\begin{aligned} i_1(\theta, \phi, \eta; Y, X, \Delta, V) &= \Delta D_2(\theta; V, X)' \sum (\phi)^{-1} r(\theta; V, Y, X) + (1 - \Delta) \left[\int_C^\tau f_{\theta, \phi}(Y|u, X) d\eta(u; X) \right]^{-1} \\ &\quad \left[\int_C^\tau f_{\theta, \phi}(Y|u, X) D_2(\theta; u, X)' \sum (\phi)^{-1} r(\theta; u, Y, X) d\eta(u; X) \right], \end{aligned}$$

where $D_2(\theta, u, X)$ is defined in (9).

The second order derivative of $\ell(\theta, \phi, \eta; Y, X, \Delta, V)$ with respect to θ equals

$$\begin{aligned} \ddot{l}_{11}(\theta, \phi, \eta; Y, X, \Delta, V) &= -\Delta D_2(\theta; V, X)' \sum (\phi)^{-1} D_2(\theta; V, X) + \Delta D_3(\theta, \phi; V, Y, X) + (1 - \Delta) \\ &\quad \times \left[\int_C^\tau f_{\theta, \phi}(Y|u, X) \left\{ -D_2(\theta; u, X)' \sum (\phi)^{-1} D_2(\theta; u, X) + D_3(\theta, \phi; u, Y, X) \right. \right. \\ &\quad \left. \left. + [D_2(\theta; u, X)' \sum (\phi)^{-1} r(\theta; u, Y, X)]^{\otimes 2} \right\} d\eta(u; X) \right] \left[\int_C^\tau f_{\theta, \phi}(Y|u, X) d\eta(u; X) \right]^{-1} \\ &\quad - \left[\int_C^\tau f_{\theta, \phi}(Y|u, X) D_2(\theta; u, X)' \sum (\phi)^{-1} r(\theta; u, Y, X) d\eta(u; X) \right]^{\otimes 2} \left[\int_C^\tau f_{\theta, \phi}(Y|u, X) d\eta(u; X) \right]^{-2}, \end{aligned}$$

where $D_3(\theta, \phi; V, Y, X)$ is defined in (12).

B1 holds from Theorem 4.1, Lemma A.3 and consistency of two-stage estimator. From (13) and (14),

$$\begin{aligned} P\ddot{l}_{11}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) &= -P \left\{ \Delta D_2(\theta_0; V, X)' \sum^{-1} (\phi_0) D_2(\theta_0; V, X) \right. \\ &\quad \left. + (1 - \Delta) \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) d\eta_0(u; X) \right]^{-2} \right. \\ &\quad \left. \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) D_2(\theta_0; u, X)' \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) d\eta_0(u; X) \right]^{\otimes 2} \right\}, \end{aligned} \quad (21)$$

which is negative definite from Condition 2(b); thus, B2 holds. From (13), we have B3 holds. B4 holds automatically.

Since

$$A_1/B_1 - A_2/B_2 = \{A_1(B_2 - B_1)\}/(B_1 B_2) + (A_1 - A_2)/B_2,$$

under Conditions 1, 3–5 and 8, we have

$$P\ddot{l}_1(\theta, \phi, \eta; Y, X, \Delta, V) - \ddot{l}_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) \rightarrow 0$$

as $|(\theta, \phi) - (\theta_0, \phi_0)| \leq \delta_n \downarrow 0$ by continuity and Condition 8. Similar to the proof of Lemma A.4, we have the class of functions $\{ \int_C^\tau f_{\theta, \phi}(Y|u, X) D_2(\theta; u, X)' \sum (\phi)^{-1} r(\theta; u, Y, X) d\eta(u; X) : \theta \in \Theta, \phi \in \Phi, \eta \in \mathcal{A} \}$ belongs to Donsker. Hence, $\{ \dot{l}_1(\theta, \phi, \eta; Y, X, \cdot, V) : \theta \in \Theta, \phi \in \Phi \}$ is Donsker from Section 2.10.2 of van der Vaart and Wellner (1996) and Condition 8. Furthermore, from Corollary 2.3.12 of van der Vaart and Wellner (1996), we have B5 holds. Under Conditions 3–5 and 8, similar to the proof of Theorem 1 in Kong and Nan (2016), we can show that B6 holds. Particularly in B6,

$$P\ddot{l}_{12}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) = (P\ddot{l}_{121}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V), \dots, P\ddot{l}_{12q}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V))$$

with

$$\begin{aligned} P\ddot{l}_{12j}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) = & -P \left((1 - \Delta) \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) D_2(\theta_0; u, X)' \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) d\eta_0(u; X) \right] \right. \\ & \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) \left\{ r(\theta_0; u, Y, X)' \sum (\phi_0)^{-1} A_j(\phi_0) \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) \right. \right. \\ & \left. \left. - \text{tr} \left[\sum (\phi_0)^{-1} A_j(\phi_0) \right] d\eta_0(u; X) \right] \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) d\eta_0(u; X) \right]^{-2} / 2 \right], \end{aligned}$$

and

$$\begin{aligned} P\ddot{l}_{13}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)[\tilde{\eta}_n - \eta_0] = & -P \left((1 - \Delta) \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) D_2(\theta_0; u, X)' \right. \right. \\ & \left. \left. \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) d\eta_0(u; X) \right] \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) d\{\tilde{\eta}_n(u; X) - \eta_0(u; X)\} \right. \right. \\ & \left. \left. \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) d\eta_0(u; X) \right]^{-2} \right) = -P \left((1 - \Delta) \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) D_2(\theta_0; u, X)' \right. \right. \\ & \left. \left. \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) d\eta_0(u; X) \right] \left[f_{\theta_0, \phi_0}(Y|\tau, X) \{\tilde{\eta}_n(\tau; X) - \eta_0(\tau; X)\} - f_{\theta_0, \phi_0}(Y|C, X) \{\tilde{\eta}_n \right. \right. \\ & \left. \left. (C; X) - \eta_0(C; X)\} - \int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) (\partial \mathbf{g}(u1 - t, \xi_0) / \partial u)' \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) \right. \right. \\ & \left. \left. \{\tilde{\eta}_n(u; X) - \eta_0(u; X)\} du \right] \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) d\eta_0(u; X) \right]^{-2} \right) \\ = & -\mathbb{G}_n\{G(\theta_0, \phi_0, \eta_0; \tilde{X}, \tilde{\Delta}, \tilde{V})\} + o_p(1), \end{aligned}$$

(22)

where

$$\begin{aligned}
G(\theta_0, \phi_0, \eta_0; \tilde{X}, \tilde{\Delta}, \tilde{V}) &= P\{E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)E_2(\theta_0, \phi_0, \eta_0; Y, X, \tau)A_1(\eta_0; \tau, X; \tilde{X}, \tilde{\Delta}, \tilde{V})\} \\
&- P\{E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)E_2(\theta_0, \phi_0, \eta_0; Y, X, C)A_1(\eta_0; C, X; \tilde{X}, \tilde{\Delta}, \tilde{V})\} \\
&+ P\left\{\int_C^\tau E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)E_2(\theta_0, \phi_0, \eta_0; Y, X, u)E_3(\theta_0, \phi_0; Y, X, u)A_1(\eta_0; u, X; \tilde{X}, \tilde{\Delta}, \tilde{V})du\right\}
\end{aligned}$$

with

$$\begin{aligned}
E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) &= (1 - \Delta) \left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) d\eta_0(u; X) \right]^{-2} \\
&\left[\int_C^\tau f_{\theta_0, \phi_0}(Y|u, X) D_2(\theta_0; u, X)' \sum (\phi_0)^{-1} r(\theta_0; u, Y, X) d\eta_0(u; X) \right],
\end{aligned}$$

$$E_2(\theta_0, \phi_0, \eta_0; Y, X, u) = f_{\theta_0, \phi_0}(Y|u, X)[1 - \eta_0(u; X)] \exp(\alpha'_0 X),$$

$$E_3(\theta_0, \phi_0; Y, X, u) = (\partial \mathbf{g}(u|1 - t, \xi_0) / \partial u)' \sum (\phi_0)^{-1} r(\theta_0; u, Y, X),$$

and $A_1(\eta_0; u, X; \tilde{X}, \tilde{\Delta}, \tilde{V})$ is defined in Lemma A.3.

Hence by Lemma A.2 and the central limit theorem,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &= J_2^{-1} \sqrt{n} \mathbb{P}_n \dot{l}_1(\theta_0, \phi_0, \eta_0; X) + J_2^{-1} \sqrt{n} P\{\ddot{l}_{12}(\theta_0, \phi_0, \eta_0; X)\}(\tilde{\phi}_n - \phi_0) \\
&+ J_2^{-1} \sqrt{n} P\{\ddot{l}_{13}(\theta_0, \phi_0, \eta_0; X)[\tilde{\eta}_n - \eta_0]\} + o_p(1),
\end{aligned}$$

which converges weakly to a mean zero normal random variable with variance $J_2^{-1} Q_2 J_2^{-1}$ from (16) and (22), where

$$J_2 = -P\{\ddot{l}_{11}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)\},$$

$$\begin{aligned}
&Q_2 \\
&= P\left[\dot{l}_1(\theta_0, \phi_0, \eta_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V}) + G(\theta_0, \phi_0, \eta_0; \tilde{X}, \tilde{\Delta}, \tilde{V}) + P\{\ddot{l}_{12}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)\} D_5(\phi_0)^{-1} C(\theta_0, \phi_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V})\right]^{\otimes 2}
\end{aligned}$$

with $D_5(\phi_0)$ and $C(\theta_0, \phi_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V})$ defined in (15) and (10), respectively.

References

- Albert PS, Shih JH. 2010; An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*. 4(3):1517–1532. [PubMed: 21938267]
- Andersen PK, Gill RD. 1982; Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*. 10(4):1100–1120.
- Breslow NE. 1972; Discussion of "Regression models and life-tables" by D. R. Cox. *Journal of the Royal Statistical Society, Series B*. 34(2):216–217.
- Chan K, Wang M. 2010; Backward estimation of stochastic processes with failure events as time origins. *The Annals of Applied Statistics*. 4(3):1602–1620. [PubMed: 21359167]
- Cox DR. 1972; Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*. 34(2):187–220.
- Ding J, Wang JL. 2008; Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*. 64(2):546–556. [PubMed: 17888040]
- Ghosh D, Lin DY. 2002; Marginal regression models for recurrent and terminal events. *Statistica Sinica*. 12:663–688.
- Harlow SD, Mitchell ES, Crawford S, Nan B, Little R, Taffe J. 2008; The restage collaboration: defining optimal bleeding criteria for onset of early menopausal transition. *Fertility and Sterility*. 89(1):129–140. [PubMed: 17681300]
- Hsieh F, Tseng YK, Wang JL. 2006; Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*. 62(4):1037–1043. [PubMed: 17156277]
- Huang CY, Wang MC. 2004; Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*. 99(468):1153–1165. [PubMed: 24068850]
- Kalbfleisch, JD, Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2. Hoboken: John Wiley & Sons, Inc; 2002.
- Kalbfleisch JD, Schaabel DE, Ye Y, Gong Q. 2013; An estimating function approach to the analysis of recurrent and terminal events. *Biometrics*. 69(2):366–374. [PubMed: 23651362]
- Kong S, Nan B. 2016; Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika*. 103(1):161–174.
- Lehmann, EL. *Theory of Point Estimation*. New York: Springer-Verlag; 1998.
- Li Z, Nan B. 2011; Relative risk regression for current status data in case-cohort studies. *The Canadian Journal of Statistics*. 39(4):557–577.
- Li Z, Tosteson TD, Bakitas MA. 2013; Joint modeling quality of life and survival using a terminal decline model in palliative care studies. *Statistics in Medicine*. 32(8):1394–1406. [PubMed: 23001893]
- Lin DY, Wei LJ, Ying Z. 1993; Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*. 80(3):557–572.
- Little, RJ, Rubin, DB. *Statistical Analysis with Missing Data*. 2. Hoboken: John Wiley & Sons, Inc; 2002.
- Liu L, Wolfe RA, Kalbfleisch JD. 2007; A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*. 26(1):139–155. [PubMed: 16526006]
- Lu X, Nan B, Song P, Sowers M. 2010; Longitudinal data analysis with event time as a covariate. *Statistics in Biosciences*. 2(1):65–80.
- Nan B, Wellner JA. 2013; A general semiparametric z-estimation approach for case-cohort studies. *Statistica Sinica*. 23:1155–1180. [PubMed: 24489449]
- Sowers M, Tomey K, Jannausch M, Eyvazzdh A, Crutchfield M, Nan B, Randolph J. 2007; Physical functioning and menopause states. *Obstet Gynecol*. 110(6):1290–1296. [PubMed: 18055722]
- Tsiatis AA, Davidian M. 2004; Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*. 14:809–834.

- van der Vaart, AW. Semiparametric Statistics. In Lectures on Probability Theory and Statistics, Ecole d'Ete de Probabilites de Saint-Flour XXIX99. Bernard, P, editor. Berlin Heidelberg: Springer-Verlag; 2002. 330–457.
- van der Vaart, AW, Wellner, JA. Weak Convergence and Empirical Processes. New York: Springer-Verlag; 1996.
- Wellner JA, Zhang Y. 2007; Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics*. 35(5):2106–2142.
- Zeng D, Lin DY. 2009; Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics*. 65(3):746–752. [PubMed: 18945267]

Figure 1.
Goodness of fit of the Cox model for the ESRD data.

Figure 2.

Monthly inpatient costs (log transformed). The solid line is the average of the estimated log transformed monthly cost. The shaded area is its 95% pointwise confidence band. The dotted lines are 30 randomly selected subjects with terminal event

Figure 3.
QQ plot for inpatient cost model.

Table 1

Simulation results for equally spaced time interval with sharp nonlinear term. var._B=bootstrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 1.2$
Full data	bias	-0.0064	0.0011	-0.0002	0.0001	-0.0032	0.0031
	var	0.0801	0.0115	0.0082	0.0002	0.0140	0.0032
Two-stage	bias	-0.0153	0.0045	0.0003	-0.0010	-0.0030	0.0039
	var	0.0973	0.0144	0.0098	0.0003	0.0166	0.0040
	var _B	0.1094	0.0161	0.0102	0.0003	0.0151	0.0043
90% CR		0.904	0.876	0.898	0.918	0.896	0.912
		0.966	0.944	0.960	0.972	0.952	0.946
95% CR		-0.0092	0.0010	0.0041	-0.0008	-0.0024	0.0024
		0.1217	0.0208	0.0130	0.0003	0.0242	0.0053

Table 2

Simulation results for equally spaced time interval with at nonlinear term. var $\hat{\mu}$ =bootstrap variance estimator; CR=coverage rate

	$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 0.2$
Full data	bias	0.0081	-0.0116	0.0209	-0.0007	0.0009
	var	0.1279	0.0130	0.0455	0.0009	0.0121
Two-stage	bias	0.0090	-0.0121	0.0229	0.0003	-0.0046
	var	0.1199	0.0161	0.0561	0.0014	0.0152
	var $\hat{\mu}$	0.1745	0.0166	0.0641	0.0015	0.0152
	90% CR	0.902	0.890	0.912	0.904	0.910
	95% CR	0.958	0.948	0.946	0.950	0.964
	bias	0.0044	-0.0196	0.0463	-0.0004	-0.0062
Complete case	var	0.2248	0.0239	0.0821	0.0017	0.0236
						0.0008

Table 3

Simulation results for non-equally spaced time interval with sharp nonlinear term $\text{var } \hat{\gamma} = \hat{\gamma}$ =bootstrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 1.2$
Full data	bias	-0.0045	0.0020	0.0068	0.0016	-0.0051	0.0059
	var	0.1474	0.0203	0.0187	0.0005	0.0177	0.0084
Two-stage	bias	-0.0238	0.0089	0.0111	0.0015	-0.0046	0.0090
	var	0.1675	0.0235	0.0253	0.0007	0.0229	0.0113
	var_b	0.1550	0.0220	0.0276	0.0007	0.0213	0.0135
90% CR		0.866	0.884	0.884	0.888	0.914	0.910
		0.942	0.936	0.938	0.944	0.950	0.960
95% CR		-0.0295	0.0093	0.0217	0.0004	0.0002	0.0120
		0.2480	0.0366	0.0353	0.0010	0.0340	0.0161

Table 4

Simulation results for non-equally spaced time interval with at nonlinear term. var $\hat{\mu}$ =bootstrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 0.2$
Full data	bias	-0.0141	-0.0044	0.0336	-0.0005	0.0015	0.0037
	var	0.1951	0.0189	0.0789	0.0036	0.0187	0.0013
Two-stage	bias	-0.0371	0.0035	0.0401	-0.0055	-0.0053	0.0048
	var	0.2471	0.0239	0.1019	0.0054	0.0225	0.0018
	var _b	0.3158	0.0227	0.1485	0.0063	0.0220	0.0019
90% CR		0.916	0.896	0.882	0.910	0.926	0.886
		0.966	0.952	0.946	0.954	0.956	0.940
95% CR		-0.0306	-0.0133	0.0772	-0.0059	-0.0021	0.0049
		0.4136	0.0388	0.1896	0.0078	0.0356	0.0026

Table 5
Longitudinal data analysis results for the inpatient cost paid by Medicare with death as a covariate.

	Complete Case			Two-Stage		
	estimate	var ($\times 10^{-3}$)	p-value	estimate	var ($\times 10^{-3}$)	p-value
Start1	-0.65	4.10	< 0.0001	-0.59	2.00	< 0.0001
Start2	-0.07	0.11	< 0.0001	-0.07	0.04	< 0.0001
Start3	-0.46	3.54	< 0.0001	-0.50	2.78	< 0.0001
Diabetes	0.06	2.51	0.20	0.09	1.46	0.01
Heart	0.12	3.31	0.04	0.11	1.80	0.01
Race	0.14	2.71	0.007	0.11	1.81	0.01
γ	1.54	3.67	< 0.0001	1.57	3.59	< 0.0001
ξ	0.99	42.90	< 0.0001	0.96	30.95	< 0.0001
μ	0.78	5.92	< 0.0001	0.76	5.30	< 0.0001