

Published in final edited form as:

*Biometrika*. 2013 ; 100(2): 339–354. doi:10.1093/biomet/ass091.

## Estimating time-varying effects for overdispersed recurrent events data with treatment switching

**QINGXIA CHEN,**

Department of Biostatistics, Vanderbilt University, 1161 21st Avenue South, S-2323 Medical Center North, Nashville, Tennessee 37232, U.S.A

**DONGLIN ZENG,**

Department of Biostatistics, University of North Carolina, 3105-D McGavran-Greenberg Hall, Campus Box 7420, Chapel Hill, North Carolina 27516, U.S.A

**JOSEPH G. IBRAHIM,**

Department of Biostatistics, University of North Carolina, 3105-D McGavran-Greenberg Hall, Campus Box 7420, Chapel Hill, North Carolina 27516, U.S.A

**MOUNA AKACHA,** and

Statistical Methodology, Novartis Pharma AG, POB CH-4002, Basel, Switzerland

**HEINZ SCHMIDLI**

Statistical Methodology, Novartis Pharma AG, POB CH-4002, Basel, Switzerland

QINGXIA CHEN: cindy.chen@vanderbilt.edu; DONGLIN ZENG: dzeng@bios.unc.edu; JOSEPH G. IBRAHIM: ibrahim@bios.unc.edu; MOUNA AKACHA: mouna.akacha@novartis.com; HEINZ SCHMIDLI: heinz.schmidli@novartis.com

### Summary

In the analysis of multivariate event times, frailty models assuming time-independent regression coefficients are often considered, mainly due to their mathematical convenience. In practice, regression coefficients are often time dependent and the temporal effects are of clinical interest. Motivated by a phase III clinical trial in multiple sclerosis, we develop a semiparametric frailty modelling approach to estimate time-varying effects for overdispersed recurrent events data with treatment switching. The proposed model incorporates the treatment switching time in the time-varying coefficients. Theoretical properties of the proposed model are established and an efficient expectation-maximization algorithm is derived to obtain the maximum likelihood estimates. Simulation studies evaluate the numerical performance of the proposed model under various temporal treatment effect curves. The ideas in this paper can also be used for time-varying coefficient frailty models without treatment switching as well as for alternative models when the proportional hazard assumption is violated. A multiple sclerosis dataset is analysed to illustrate our methodology.

### Keywords

B-spline; Expectation-maximization algorithm; Maximum likelihood estimate; Recurrent event; Time-varying coefficient; Treatment switching

## 1. Introduction

This research was motivated by the need to better evaluate treatment effects in a phase II/III clinical trial with multiple sclerosis patients, in which the primary endpoint, a clinical relapse, is a recurrent event (Kappos et al., 2010). As multiple sclerosis is a heterogeneous disease, relapse rates differ among patients, and the number of relapses shows overdispersion compared to a Poisson model (Friede & Schmidli, 2010a). The most commonly used model for overdispersed count data is the negative binomial regression model (Wang et al., 2009), which assumes constant event rates as well as constant covariate effects over time. This latter assumption may be too stringent in practice. Depending on the mode of action and the pharmacokinetics of the drugs, treatment effects may vary with time (Nicholas et al., 2012). Time-varying treatment effects are of particular interest in clinical trials that consist of a core phase and an extension phase. In a core phase, patients are randomized to the experimental treatment and the control treatment, and are then followed for a fixed time period. Once a patient has reached the end of the follow-up period, the patient enters the extension phase and then receives the experimental treatment. Hence patients who were randomized to the control treatment in the core phase are switched to the experimental treatment, while patients randomized to the experimental treatment in the core phase stay on the experimental treatment in the extension phase. Such designs are frequently used in clinical trials for ethical reasons. Similar in spirit are multiple sclerosis clinical trials where information on patient relapses before randomization is available. For example in the randomized placebo-controlled trial described in Lycke et al. (1996), relapses that occurred in the two years preceding randomization were recorded. Such settings can also be seen to correspond to a switch of treatment, where before randomization, patients receive standard treatment, and after randomization, some patients are switched to the experimental treatment. Although we focus here on clinical trials in multiple sclerosis, the proposed methods are also applicable to other diseases where recurrent events are important endpoints, and where study designs with a core and extension phase are common. Examples of such diseases are asthma and chronic obstructive pulmonary disease (Friede & Schmidli, 2010b) where the recurrent event is an exacerbation, or gout (Akacha & Benda, 2010) where the recurrent event is a relapse.

Better evaluation of the treatment effects in the aforementioned studies relies on advances in recurrent events, time-varying coefficients, and treatment switching. There has been a vast literature on each individual area, but little has been done on time-varying coefficients in recurrent event models that accommodate treatment switching. A counting process with a Cox-type of intensity function has been commonly used to analyse recurrent events data. This model assumes that the underlying counting process is a time-transformed counting process and that the covariates have multiplicative effects on the mean and rate functions of the counting process. In order to adjust for the correlations among the recurrent events, research efforts have been focused on marginal hazards models and frailty or random effects models. For marginal hazards models, Pepe & Cai (1993) proposed semiparametric procedures for graphical displays as well as for making inferences; Lin et al. (2000) provide asymptotic justifications of these models; Wei et al. (1989) examine regression methods for multivariate survival data by modelling marginal distributions and obtain asymptotic results for the proposed estimators; Cai & Prentice (1995, 1997) proposed weighted partial likelihood estimating equations for hazard models with distinct or common baseline hazard rate functions. Frailty or random effects models were studied by Clayton & Cuzick (1985), Oakes & Jeong (1998), Murphy (1995), Spiekerman & Lin (1998), Anderson & Louis (1995), Fan & Li (2002) among others. An excellent article reviewing previous work as well as building a very general framework for transformation models with random effects is Zeng & Lin (2007).

Statistical methods developed for recurrent events typically assume constant regression coefficients. This assumption, however, is likely to be violated in practice and methods allowing for nonconstant coefficients are of interest. Varying-coefficient models for survival data, which assume the regression coefficients to be unknown functions of observed covariates,  $U$ , have been studied by Murphy (1993) and Cai et al. (2007, 2008). In particular, the latter two papers developed varying-coefficient models for multivariate failure time data under marginal hazards models and under partial linear regression models, respectively. In this paper, we focus on time-varying coefficient models. For nonfailure time data settings, the time-varying coefficient model is only a special case of varying-coefficient models by letting the observed covariate  $U$  be time. This is not true for failure time data settings because the time is subject to censoring. Time-varying survival models have been studied by Zucker & Karr (1990), Gray (1992), Cai et al. (2000), Martinussen et al. (2002), Cai & Sun (2003), Tian et al. (2005) among others. However, all have focused on the multiplicative models with univariate failure time data. Recently, Martinussen et al. (2011) developed the methodology to handle time-varying coefficients in the context of multivariate event times, but using the Aalen additive hazards model. Extending time-varying coefficient models to the multiplicative models with multivariate event times remains unexplored.

There has also been related work on treatment switching in clinical trials (Robins & Tsiatis, 1991; Branson & Whitehead, 2002; Shao et al., 2005; Zeng et al., 2012). The literature, however, has been focused on univariate time-to-event data with treatment switching caused by drop-in or drop-out in clinical trials. The treatment switching considered in this paper is instead determined by the study design, as in the motivating study of a trial with a core phase and an extension phase. No literature, to the best of our knowledge, has studied time-varying treatment effects with design-based treatment switching in the context of multivariate failure time data.

In this paper, we develop a multiplicative semiparametric frailty modelling approach to estimate time-varying effects for overdispersed recurrent events data with treatment switching. The proposed model incorporates the treatment switching time into the time-varying coefficients. This paper deals with new methodological challenges arising from within-cluster dependence and time-varying nonparametric effects of the covariates. For the parameter space, we have a mixture of parametric components, nonparametric components, and time-varying components. An expectation-maximization algorithm based on B-spline functions over the sieve space is derived to obtain the maximum likelihood estimates. Although the new model is tailored to handle treatment switching in recurrent event times with time-varying coefficients, the inference procedure and theorems provided in this paper can easily be used in a non-treatment-switching framework by letting the switching time be zero for the experimental group and be the study duration length for the control group. Since the regression coefficients depend on time, our model is not a proportional hazards model and can be used when the proportional hazards assumption is violated. In fact, the proposed model is preferred to the Cox proportional hazards model with time-dependent covariates because it has nonparametric time-varying coefficients while artificial time transformation functions have to be created for the Cox model.

## 2. Inference procedures

### 2.1. Statistical models

Let  $N(t)$  be the counting process of relapses and let  $X$  denote the baseline information. If a given patient stays on the placebo treatment all the time since time zero, we assume that the patient's intensity function is  $\lambda(t)$  ( $t > 0$ ). However, if the patient has treatment initiated at time  $c^*$ , then we assume that at any time  $t > c^*$  the intensity at time  $t$  is  $\lambda(t) \exp\{\beta(t - c^*)\}$ , where  $\beta(s)$  reflects the treatment effect depending on the length of time since treatment. We

further incorporate the baseline characteristics and within-subject frailty, and hence propose the following intensity models: for patients treated from the beginning,

$$E\{dN(t)|H_{t-}, A(0)=A(c^*)=1, X, \xi\}=\xi \exp\{\gamma^T X+\beta(t)\} d\Lambda(t), \quad (1)$$

where  $H_{t-}$  denotes all the past history before  $t$ ,  $X$  denotes all baseline covariates,  $\xi$  is a gamma-frailty with mean one, and  $d\Lambda(t) = \lambda(t) dt$ . For patients untreated at the beginning but switched to treatment at time  $c^*$ ,

$$E\{dN(t)|H_{t-}, A(0)=0, A(c^*)=1, X, \xi\}=\xi \exp\{\gamma^T X+\beta(t-c^*)I(t>c^*)\} d\Lambda(t). \quad (2)$$

When  $\beta(s) = 0$  ( $s = 0$ ), this implies that there is no treatment effect.

We can integrate models (1) and (2) into one expression

$$E\{dN(t)|H_{t-}, A(c^*)=1, X, \xi\}=\xi \exp\{\gamma^T X+\beta(t-c^*)I(t>c^*)\} d\Lambda(t), \quad (3)$$

where  $c^*$  denotes the time of initiating treatment. Therefore,  $c^* = 0$  for the experimental treatment arm and  $c^* > 0$  for the control arm. In this paper, we assume that conditioning on  $\xi$ , the current event rate is independent of the past, therefore, equation (3) is an intensity.

## 2.2. Nonparametric maximum likelihood estimation

The observed data from  $n$  subjects are

$$\{N_i(t)I(t \leq Y_i), Y_i, X_i, C_i\} \quad (i=1, \dots, n),$$

where  $Y_i$  is the minimum of the censoring time and study duration  $\tau$ , and  $C_i = C_i^* I(C_i^* < Y_i) + \infty I(C_i^* \geq Y_i)$ . Here,  $C_i^*$  is the potential initiation time of the treatment and the definition of  $C_i$  assumes that if the patient does not initiate the experimental treatment before  $Y_i$ , the observed initiation time is taken to be infinity for mathematical convenience. We also assume that the switching times and censoring times are noninformative. In other words,  $C_i^*$  is independent of recurrent events and  $Y$  given  $X$ , and  $Y$  is independent of  $N(t)$  and  $\xi$  given  $X$ . Similar to Zeng & Lin (2007) with identity transformation, the observed data likelihood function is given by

$$\prod_{i=1}^n \left( \int \prod_{t \leq Y_i} [\lambda(t) \xi_i \exp\{\gamma^T X_i + \beta(t - C_i)I(t > C_i)\}]^{\Delta N_i(t)} \right. \\ \left. \times \exp \left[ - \int_0^{Y_i} \xi_i \exp\{\gamma^T X_i + \beta(t - C_i)I(t > C_i)\} d\Lambda(t) \right] f(\xi_i) d\xi_i \right)$$

where  $\gamma$  is a  $p \times 1$  vector of unknown parameters,  $\Delta N_i(t)$  denotes the jump size of  $N_i$  at time  $t$  and  $f(\xi)$  is a gamma density with mean 1 and variance  $\theta$ , denoted by  $\text{Ga}(\theta^{-1}, \theta^{-1})$ . The parameters are  $\{\beta(\cdot), \gamma, \Lambda(\cdot), \theta\}$ .

We use a sieve estimator of  $\beta$  but use nonparametric maximum likelihood to estimate  $\Lambda$ . Specifically, we approximate  $\beta(t)$  via a sequence of B-splines for  $t \in [0, \tau]$ ; that is, we

assume that  $\beta(t) = \sum_{j=1}^{K_n+m} \alpha_j B_j(t)$ , where  $B_1(t), \dots, B_{K_n+m}(t)$  are the B-splines based on knots  $x_1 = \dots = x_m = 0 < x_{m+1} < \dots < \tau = x_{m+K_n} = \dots = x_{K_n+2m}$ , where  $x_{m+j} = j K_n^{-1}$ , and the B-splines are  $m$ th order piecewise polynomials. To construct the nonparametric maximum

likelihood estimator of  $\Lambda$ , we assume that  $\Lambda(t)$  is a step function with jumps at observed events of the observed counting process.

We use the expectation-maximization algorithm for parameter estimation (Klein, 1992; Nielsen et al., 1992). In the maximization step, we maximize the conditional expectation of the complete loglikelihood function

$$\sum_{i=1}^n \left[ \int I(Y_i \geq t) \left\{ \log \Lambda(t) + \hat{E}(\log \xi_i | D) + \gamma^T X_i + \sum_{j=1}^{K_n+m} \alpha_j B_j(t - C_i) I(t > C_i) \right\} dN_i(t) - \hat{E}(\xi_i | D) \int_0^{Y_i} \exp \left\{ \gamma^T X_i + \sum_{j=1}^{K_n+m} \alpha_j B_j(t - C_i) I(t > C_i) \right\} d\Lambda(t) + \hat{E} \{ \log f(\xi_i) | D \} \right],$$

where  $\hat{E}(\cdot | D)$  denotes the conditional expectation given the observed data and the estimates at the previous iteration. The maximization is equivalent to solving the equations

$$0 = \sum_{i=1}^n \int I(Y_i \geq t) \left\{ Z_i(t) - \frac{\sum_{l=1}^n I(Y_l \geq t) \hat{E}(\xi_l | D) Z_l(t) G_l(t)}{\sum_{l=1}^n I(Y_l \geq t) \hat{E}(\xi_l | D) G_l(t)} \right\} dN_i(t)$$

for  $\gamma$  and the quantities  $\alpha$ , where  $Z_i(t) = \{X_i, I(t > C_i)B_1(t - C_i), \dots, I(t > C_i)B_{K_n+m}(t - C_i)\}^T$  and  $G_i(t) = \exp \{ \gamma^T X_i + \sum_{j=1}^{K_n+m} \alpha_j I(t > C_i) B_j(t - C_i) \}$ . To solve for  $\theta$ , we maximize the quantity

$$\sum_{i=1}^n \{ (\theta^{-1} - 1) \hat{E}(\log \xi_i | D) - \hat{E}(\xi_i | D) / \theta - (\log \theta) / \theta - \log \Gamma(1/\theta) \}.$$

Both equations can be solved using the Newton–Raphson algorithm. We then update  $\Lambda(t)$  as

$$\hat{\Lambda}(t) = \int_0^t \frac{\sum_{i=1}^n I(Y_i \geq s) dN_i(s)}{\sum_{i=1}^n I(Y_i \geq s) \hat{E}(\xi_i | D) G_i(s)}.$$

In the expectation step of the expectation-maximization algorithm, we evaluate the conditional expectation of  $\hat{E}\{g(\xi_i) | D\}$  using  $\int g(\xi) h_i(\xi | D) d\xi$ , where  $h_i(\xi | D)$  is the density of

$$\text{Ga} \left\{ \theta^{-1} + N_i(Y_i), \quad \theta^{-1} + \int_0^{Y_i} G_i(t) d\Lambda(t) \right\},$$

which can be interpreted as the posterior density of  $\xi_i$  given the data.

At convergence, we can use the Louis (1982) formula to compute the observed information matrix for all the parameters including the quantities  $\alpha$ ,  $\gamma$ , and the jump size of  $\Lambda$ .

To test the treatment effect, we define the weighted average treatment effect as

$$w(g, \beta) = \int_0^\tau g(t) \hat{\beta}(t) dt \text{ for any smooth function } g(t), \text{ where } \hat{\beta}(t) = \sum_{j=1}^{K_n+m} \hat{\alpha}_j B_j(t) \text{ and } \int_0^\tau g(t) dt = 1. \text{ We can estimate } w(g, \beta) \text{ by}$$

$$\hat{w}(g, \beta) = \int_0^\tau g(t) \hat{\beta}(t) dt = \sum_{j=1}^{K_n+m} \hat{\alpha}_j \int_0^\tau g(t) B_j(t) dt.$$

The variance of  $w(g, \beta)$  can be consistently estimated using the delta method. In particular,

$$\text{var} \left\{ \int_0^\tau g(t) \hat{\beta}(t) dt \right\} = a(g, K_n, m)^T \text{cov}(\hat{\alpha}) a(g, K_n, m),$$

where  $a = (a_1, \dots, a_{K_n+m})$ ,  $a(g, K_n, m) = \left\{ \int_0^\tau g(t) B_1(t) dt, \dots, \int_0^\tau g(t) B_{K_n+m}(t) dt \right\}$ .

### 2.3. Asymptotic results

We establish the asymptotic properties of the nonparametric maximum likelihood estimators in this section. The following conditions are needed for the theorems in this paper.

*Condition 1.* The true parameters  $\beta_0(t)$  and  $\lambda_0(t)$  are  $r$  times continuously differentiable and their  $r$ th derivatives are bounded in  $[0, \tau]$  where  $r \geq 2$ . The true parameters  $\gamma_0$  and  $\theta_0$  belong to compact sets in their domains.

*Condition 2.* The conditional density of  $C^*$  given  $X$  has bounded  $r$ th derivatives and its support contains  $[0, \tau]$ . The conditional density of  $Y$  given  $X$  has a bounded  $r$ th derivative in its support, which contains  $\tau$ .

*Condition 3.* If  $\alpha_0 + \alpha_1^T X = 0$  with probability one, then  $\alpha_0 = 0$  and  $\alpha_1 = 0$ .

*Condition 4.* The number of the knots  $K_n$  satisfies  $K_n^3 n^{-1/2} \rightarrow 0$  and  $n^{1/2} K_n^{-2r} \rightarrow 0$ , as  $n \rightarrow \infty$ .

Conditions 1 and 2 are regularity and technical conditions. Condition 3 is an identifiability condition. Condition 4 is the regularity condition needed to construct the sieve estimate for the time-varying effect  $\beta(t)$ .

Under these conditions, the following theorems give the consistency and asymptotic distribution of the estimators.

**Theorem 1**—Under Conditions 1–4, when  $n \rightarrow \infty$ ,

$$\sup_{t \in [0, \tau]} |\hat{\beta}(t) - \beta_0(t)| + \|\hat{\gamma} - \gamma_0\| + \sup_{t \in [0, \tau]} |\hat{\Lambda}(t) - \Lambda_0(t)| + \|\hat{\theta} - \theta_0\| = o_p(1),$$

where  $\|\cdot\|$  is the Euclidean norm.

**Theorem 2**—Let  $F_\beta$  consist of all the functions in  $[0, \tau]$  whose  $r$ th derivatives are bounded by 1, where  $r$  is given in Condition 1. Let  $F_\Lambda$  consist of all the functions in  $[0, \tau]$  whose total variation is bounded by 1. Let  $O_\gamma$  be the unit ball in  $R^p$  and  $O_\theta$  be the unit interval in  $R$ . We treat  $\{\hat{\beta}_n(t) - \beta_0(t), \hat{\Lambda}_n(t) - \Lambda_0(t), \hat{\gamma} - \gamma_0, \hat{\theta} - \theta_0\}$  as a stochastic class in  $l^\infty(F_\beta \times F_\Lambda \times O_\gamma \times O_\theta)$  whose value for  $(g_\beta, g_\Lambda, a, b)$  is defined as

$$\int g_\beta(t) \{\hat{\beta}_n(t) - \beta_0(t)\} dt + \int g_\Lambda(t) d\{\hat{\Lambda}_n(t) - \Lambda_0(t)\} + a^T (\hat{\gamma} - \gamma_0) + b(\hat{\theta} - \theta_0),$$

where  $g_\beta$  and  $g_\Lambda$  are any functions on  $F_r$  and  $F_\Lambda$ , respectively,  $a$  is any vector on  $R^p$  and  $b$  is any interval on  $R$ . Then under Conditions 1–4,  $n^{1/2}\{\hat{\beta}_n(t) - \beta_0(t), \hat{\Lambda}_n(t) - \Lambda_0(t), \hat{\gamma} - \gamma_0, \hat{\theta} - \theta_0\}$  converges in distribution to a mean-zero Gaussian process in the metric space  $l^\infty(F_r \times F_\Lambda \times O_\gamma \times O_\theta)$ .

The proofs of the theorems are given in the Appendix. By choosing  $g_\beta(t) = 1$ ,  $g_\Lambda(t) = 0$ ,  $a$  as a vector with all  $p$  elements equal to 0, and  $b = 0$ , Theorem 2 establishes the weak convergence of  $n^{1/2} \int \{\hat{\beta}_n(t) - \beta_0(t)\} dt$ . Similarly, the weak convergence of other parameters can be obtained by choosing special values of  $(g_\beta, g_\Lambda, a, b)$ . With the approximation of B-splines for  $\beta(t)$ , the proposed model reduces to the usual frailty model and work has been done to show that the estimators based on the expectation-maximization algorithm have the same asymptotic properties as the maximum likelihood estimators in frailty models (Parner, 1998; Murphy, 1995).

### 3. Numerical results

#### 3.1. Simulation studies

We conducted extensive simulation studies to assess the finite sample performance of the proposed methods. We generated recurrent event times from the counting process with cumulative intensity  $\Lambda(t | X, C^*, A; \xi) = \xi \Lambda_0(t) \exp\{\gamma X + \beta(t - C^*)I(t > C^*)\}$ , where  $\gamma = 0.2$ , treatment assignment  $A$  was set to 0 or 1 with equal probability,  $X$  was simulated from a Poisson distribution with mean 2 and truncated at 5,  $\xi$  was simulated from the gamma distribution with mean 1 and variance  $\theta = 0.8$ , and the cumulative intensity function  $\Lambda_0(t) = 1.5t$ . We generated censoring times from the exponential distribution with hazard function  $\exp(-2 - 0.2A + 0.2X)$  and truncated by 1.2 and the largest follow-up time  $\tau = 3$ . We considered four scenarios for the treatment effects: (a)  $\beta(t) = 0$ , corresponding to the null hypothesis of no treatment effect; (b)  $\beta(t) = -1$ , corresponding to a constant treatment effect; (c)  $\beta(t) = \log\{1 - 0.9t/(0.6 + t)\}$ , corresponding to a monotone treatment effect; (d)  $\beta(t) = \log\{1 - 1.5t \exp(-0.75t^2)\}$ , corresponding to a nonmonotonic treatment effect. For scenarios (a)–(c), the switching times  $C^*$  were set to 0 for the treatment group with  $A = 1$ , and to 1 for the placebo group with  $A = 0$ . For scenario (d), we generated the switching times of the control group subjects from a uniform  $[0, 4]$  distribution to study staggered switching and partial switching. The treatment effect in scenario (c) is the Emax model with  $\beta(t) = \log\{1 - R_{\max}t/(T_{50} + t)\}$ , where  $R_{\max} = 0.9$  is the maximal intensity reduction and  $T_{50} = 0.6$  is the time to have a 50% intensity reduction.

We used the proposed expectation-maximization algorithm to calculate the estimates of  $\gamma$ ,  $\alpha$ ,  $\theta$ , and  $\Lambda\{t\}$ , where  $\Lambda\{t\}$  denotes the jump size of  $\Lambda$  at time  $t$ . The Louis formula was used to obtain the standard errors of the estimates. The cubic B-splines with  $m = 4$  and  $K_n = 6$  were used in all scenarios. The knots were equally spaced on  $[0, \tau]$ . For each scenario, we generated 1000 replicates with sample sizes  $n = 400$  and  $n = 200$ . Since  $\theta$  and  $\Lambda\{t\}$  are nonnegative, we used the log-transformation in constructing their confidence interval. In particular, the 95% confidence interval for  $\theta$  is  $\hat{\theta} \exp\{\pm 1.96 \times \text{se}(\hat{\theta})/\hat{\theta}\}$ , where  $\text{se}(\hat{\theta})$  denotes the standard error. A similar formula is obtained for  $\Lambda\{t\}$ .

Table 1 shows the summary statistics for the parameters of interest for  $n = 400$  and  $n = 200$ . For scenarios (a)–(d), the average event counts are roughly 5.42, 3.46, 3.94, 4.64, respectively, for the placebo group, and 5.63, 2.07, 2.54, 3.51, respectively, for the treatment group. The nonparametric maximum likelihood estimators for  $\gamma$ ,  $\theta$ ,  $\beta(t)$  with  $t = 0.5, 1.5$ , and 2.5, and the weighted average treatment effect  $\bar{\beta}(t) = \int_0^\tau g(t)\beta(t)dt$  with  $g(t) = 1/\tau$  are virtually unbiased, with the average standard errors close to the empirical standard errors, and the 95% confidence intervals achieve the nominal coverage probabilities. For each scenario, Figs. 1 and 2 show the true treatment effect  $\beta(t)$ , the estimates of the treatment

effect  $\hat{\beta}(t) = \sum_{j=1}^{K_n+m} \hat{\alpha}_j B_j(t)$ , the empirical 95% confidence intervals, and the average of the 95% confidence intervals for  $n = 400$  and  $n = 200$ , respectively. We used the 2.5% and 97.5% quantiles of  $\hat{\beta}(t)$  based on the 1000 replicates to construct the empirical 95% confidence interval, and used the average values of the upper and lower limits of the 95% confidence intervals calculated in each replicate to construct the average 95% confidence intervals. The estimated treatment effect curves are virtually unbiased and the average 95% confidence intervals are very close to the empirical 95% confidence intervals.

### 3.2. Application to acyclovir study

Viral infections have been suspected to initiate the disease process in multiple sclerosis (Gilden, 2005). Hence antiviral drugs may be able to reduce the relapse rate in multiple sclerosis patients. The first and largest study to investigate the effect of an antiviral drug, acyclovir, has been reported in Lycke et al. (1996). The acyclovir study was a randomized, placebo-controlled, double-blind clinical trial, where 60 multiple sclerosis patients were randomly assigned to acyclovir and placebo, and then followed for two years. The relapse rates were reduced by 34% in the acyclovir group as compared to placebo; however this reduction was not statistically significant with  $p = 0.083$ . Information on the relapses in the two years before randomization was also collected in this study, but not used in this statistical analysis.

We used the proposed methods to reanalyse the data from the acyclovir study, using relapse information before and after randomization, and investigating the time-dependence of the potential treatment effect. It was assumed here that the pre-randomization period is comparable to a placebo treatment. Since a few patients had pre-randomization information for less than two years, we staggered the switching times for the acyclovir group. According to the Akaike information criterion, the estimated treatment effect using quadratic B-spline functions with 9 knots were chosen and presented. Figure 3 suggests that acyclovir may at best temporarily reduce the relapse rate in the first few months after start of treatment.

## 4. Discussion

We have developed our model in the framework of a multiplicative model, but these issues may also be studied in an additive model framework (Cai & Zeng, 2001; Martinussen et al., 2011), for example, by assuming

$$d\Lambda_{ij}(t|X, C^*) = \xi_i dt + \phi^T X_{ij} + \alpha(t - C_i^*) I(t - C_i^*) dt + d\Lambda_0(t), \quad (4)$$

where  $\Lambda_{ij}(t|X)$  denotes the cumulative hazard function for the  $j$ th event of the  $i$ th individual at the given covariates  $X_{ij}$ ,  $\xi_i$  is the random effect following a gamma distribution,  $\phi$  is an unknown time-independent coefficient vector,  $\alpha(t - C^*)$  is an unknown time-dependent coefficient function reflecting the treatment effect depending on the time length of treatment, and  $\Lambda_0(t)$  is an unknown baseline cumulative hazard function. While the  $\beta(t)$  in (3) estimates the treatment effect in terms of hazard ratios, the  $\alpha(t)$  in (4) estimates the risk difference. When the model is correctly specified, the estimates in the multiplicative model can achieve their efficiency bound, but not the estimates in the additive model specified in (4).

The proposed procedure can be directly applied to the studies without treatment switching by assigning the switching time  $C^*$  to be either 0 or  $\tau$  for the experimental and control groups. The proposed model is developed using the gamma frailty mainly for numerical convenience. The estimation procedure and asymptotic properties can in principle be extended to other frailty or random effects models. As pointed out by one reviewer, the

proposed approach should be implemented with caution by recognizing that the analysis does not enjoy the protection against bias afforded by the randomization, and therefore must be viewed cautiously in comparison with the analysis based on the core phase data with randomization design only. Our motivating multiple sclerosis study has a design of this type and its detailed description and analysis are given in the Supplementary Material. Furthermore, in this paper, the switching time  $C^*$  is assumed to be noninformative, which is determined by the study design as in the motivating study with a core phase and an extension phase. The method can be generalized to allow  $C^*$  to be a drop-in or drop-out time; however, the informative selection of drop-in or drop-out should be adjusted to keep the randomization balanced between the two treatment arms. The latter is often done via inverse probability weighted estimating equations. Furthermore, we note that under the key assumption that  $C^*$  and  $A(C^*)$  are independent of the potential outcomes  $\{N(t; a, c^*): N(t; c^*) \text{ is the counting process when treatment is given at time } c^*\}$ ,  $\beta(t)$  has a causal interpretation.

To the best of our knowledge, obtaining a confidence band for a nonparametric estimator is a challenging and unresolved problem. One possible way to examine  $\beta(\cdot) = 0$  is to test  $\int_0^\tau g_l(t) \beta(t) dt = 0$  ( $l = 1, \dots, K$ ) based on Theorem 2, where  $g_l(t)$  are  $K$  basis functions. Even though the asymptotic properties are established only for  $\int_0^\tau g(t) \beta(t) dt$  in § 2-3, the numerical performance, including point estimates, standard errors, and coverage probabilities, of the treatment effect at a single time-point is quite satisfactory in all simulations considered in this paper.

The programs for the estimation procedure for the proposed model are currently written in R (R Development Core Team, 2013). The computations took only a few seconds for the acyclovir study. The computation time depends on the sample size of the dataset, the number of parameters, and the number of recurrent events per subject.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research was partially funded by the National Institutes of Health. We thank Simon Schneider for extracting the acyclovir relapse data. We thank the editor, the associate editor and two referees for their constructive comments that greatly improved this paper.

## Appendix

### Proof of invertibility of information operator

Let  $\psi = (\gamma, \theta)$  and  $\psi_0 = (\gamma_0, \theta_0)$ . The loglikelihood function from a single subject is

$$l(\Lambda, \psi, \beta) = \int \log \lambda(t) dN(t) + \int \{\gamma^T X + \beta(t - C)I(t \geq C)\} dN(t) + \log \Gamma\{\theta^{-1} + N(Y)\} - \{\theta^{-1} + N(Y)\} \log \left\{ \theta + \int_0^Y e^{\gamma^T X + \beta(t-C)I(t \geq C)} d\Lambda(t) \right\}.$$

By differentiating the loglikelihood function with respect to  $\Lambda$ ,  $\psi$ , and  $\beta$  along the submodel  $d\Lambda(1 + \varepsilon h)$ ,  $\psi + \varepsilon b$ , and  $\beta + \varepsilon \tilde{h}$ , respectively, where  $1$  in  $d\Lambda(1 + \varepsilon h)$  is an identical mapping,  $h \in L_2[0, \tau]$  and  $b \in L_2[0, \tau]$ , we obtain the score operators as

$$\begin{aligned}
l_{\Lambda}(\Lambda, \psi, \beta)[h] &= \int h(t) \{dN(t) - g(Y)I(Y \geq t)e^{\gamma^T X + \beta(t-C)I(t \geq C)}d\Lambda(t)\}, \\
l_{\beta}(\Lambda, \psi, \beta)[\tilde{h}] &= \int \tilde{h}(t-C)I(t \geq C) \{dN(t) - g(Y)I(Y \geq t)e^{\gamma^T X + \beta(t-C)I(t \geq C)}d\Lambda(t)\}, \\
&= \int \tilde{h}(t)I(t \geq 0) \{dN(t+C) - g(Y)I(Y \geq t+C)e^{\gamma^T X + \beta(t)I(t \geq C)}d\Lambda(t+C)\},
\end{aligned}$$

$$= \begin{bmatrix} l_{\psi}(\Lambda, \psi, \beta)^T b \\ -\theta^{-2}(\log \Gamma)' \{ \theta^{-1} + N(Y) \} + \theta^{-2} \log \left\{ \theta + \int_0^Y e^{\gamma^T X + \beta(t-C)I(t \geq C)} d\Lambda(t) \right\} - g(Y) \end{bmatrix}^T b,$$

where

$$g(Y) = \frac{\theta^{-1} + N(Y)}{\theta + \int_0^Y \exp\{\gamma^T X + \beta(t-C)I(t \geq C)\} d\Lambda(t)}.$$

We chose to differentiate the loglikelihood function with respect to  $\Lambda$  along the submodel  $d\Lambda(1 + \varepsilon h)$  to ensure that the  $\Lambda$  function along this submodel has jumps at the same time-points as  $\Lambda$ , and therefore, the derivative of the loglikelihood function is well defined.

We note that  $l_{\beta}$  has a similar integral form to  $l_{\Lambda}$ . Thus, following Murphy (1995) page 193 Lemma 2, we can show that the information operator,  $I(\Lambda, \beta, \psi) \equiv E\{(l_{\Lambda}, l_{\beta}, l_{\psi})^*(l_{\Lambda}, l_{\beta}, l_{\psi})\}$ , where  $(l_{\Lambda}, l_{\beta}, l_{\psi})^*$  is the dual operator of  $(l_{\Lambda}, l_{\beta}, l_{\psi})$ , can be written as one Fredholm operator of the first kind, which is the summation of an invertible operator and an integral operator when  $\Lambda = \Lambda_0$ ,  $\beta = \beta_0$  and  $\psi = \psi_0$ . Furthermore, using Conditions 1 and 2, the latter can be shown to be a compact operator from

$$H \equiv \{(h, \tilde{h}, b) : h \in BV[0, \tau], \tilde{h} \in C[0, \tau], b \in R^d, d = \dim(X) + 1\}$$

to itself, where  $BV[0, \tau]$  is the Banach space consisting of all the functions with bounded total variation in  $[0, \tau]$  and  $C[0, \tau]$  is the Banach space consisting of all the continuous functions in  $[0, \tau]$ .

We next show that  $I(\Lambda_0, \beta_0, \psi_0)$  is invertible. Following Rudin (1973) Theorem 4.18 and 4.24, it suffices to show that  $I(\Lambda_0, \beta_0, \psi_0)$  is one-to-one. Suppose  $I(\Lambda_0, \beta_0, \psi_0)[h, \tilde{h}, b] = 0$ ; that is,  $l_{\Lambda}[h] + l_{\beta}[\tilde{h}] + l_{\psi}^T b = 0$ . First, we let  $C = \tau$  so that  $l_{\Lambda}[h] + l_{\psi}^T b = 0$ . This is essentially the score function in the usual frailty model without  $\beta$  in the regression. Thus, from Murphy (1995) page 193 Lemma 2, we obtain  $h = 0$  and  $b = 0$ . Now with  $l_{\beta}[\tilde{h}] = 0$ , we set  $C = 0$  and  $Y = \tau$  and set  $N(t)$  to have a jump only at time  $t \in [0, \tau]$ , and therefore  $l_{\beta}[\tilde{h}] = g(Y) \int \tilde{h}(t) \exp\{\gamma_0^T X + \beta_0(t)\} d\Lambda_0(t)$ . Clearly,  $\tilde{h} = 0$ . Therefore,  $I(\Lambda_0, \beta_0, \psi_0)$  is invertible.

Finally, the same arguments apply if we consider a different Banach space,

$$H^* = \{(h, \tilde{h}, b) : h \in L_2[0, \tau], \tilde{h} \in L_2[0, \tau], b \in R^d, d = \dim(X) + 1\}.$$

Therefore, the invertibility of  $I(\Lambda_0, \beta_0, \psi_0)$  implies

$\|I(\Lambda_0, \beta_0, \psi_0)[h, \tilde{h}, b]\|_{L_2(P)}^2 \geq c\{\|h\|_{L_2}^2 + \|\tilde{h}\|_{L_2}^2 + \|b\|^2\}$ . Furthermore, we note that  $I(\Lambda, \beta, \psi)$  converges to  $I(\Lambda_0, \beta_0, \psi_0)$  uniformly in the norm  $\|\Lambda - \Lambda_0\|_\infty + \|\beta - \beta_0\|_\infty + \|\psi - \psi_0\|$  where  $\|f\|_\infty$  denotes the supreme norm in  $[0, \tau]$ . We conclude that there exists some  $\varepsilon_0$  such that whenever  $\|\Lambda - \Lambda_0\|_\infty + \|\beta - \beta_0\|_\infty + \|\psi - \psi_0\| < \varepsilon_0$ , the inequality

$$\|I(\Lambda, \beta, \psi)[h, \tilde{h}, b]\|_{L_2(P)}^2 \geq c/2\{\|h\|_{L_2}^2 + \|\tilde{h}\|_{L_2}^2 + \|b\|^2\} \quad (\text{A1})$$

holds. We will use this fact later.

## Proof of Theorem 1

We will show that there exists a local maximum of the observed data loglikelihood function over the sieve space

$$S_n = \left\{ (\Lambda, \beta, \psi) : \Lambda \text{ is the step function with jump sizes at the observed events, } \beta(t) = \sum_{j=1}^{K_n+m} \alpha_j B_j(t), \text{ the } B_j \text{ are B-spline bases with knots given in } \S 2, \psi \in R^d \right\},$$

such that the obtained estimator,  $(\hat{\Lambda}, \hat{\beta}, \hat{\psi})$ , converges to the true parameters in probability under the norm in Theorem 1.

To this end, by Condition 1 and page 229 of Schumaker (2007), there exists a function

$\hat{\beta}_0(t) = \sum_{j=1}^{K_n+m} \alpha_{j0} B_j(t)$  such that  $\|\hat{\beta}_0 - \beta_0\|_\infty = O(K_n^{-r})$ . Then we consider the following neighbourhood of  $\hat{\beta}_0$  in the sieve space

$$N_{\varepsilon_n} = \left\{ \beta(t) = \sum_{j=1}^{K_n+m} \alpha_j B_j(t) : \sum_{j=1}^{K_n+m} |\alpha_j - \alpha_{j0}|^2 \leq \varepsilon_n \right\},$$

where  $\varepsilon_n$  is to be chosen later. For each  $\beta \in N_{\varepsilon_n}$ , we define  $(\hat{\Lambda}_\beta, \hat{\psi}_\beta) = \operatorname{argmax} P_n l(\Lambda, \beta, \psi)$ , where  $P_n$  is the empirical measure and  $\Lambda$  is a step function with jumps at the observed events.

If we choose  $\varepsilon_n$  so that  $K_n^{3/2} \varepsilon_n \rightarrow 0$ , then for  $\beta \in N_{\varepsilon_n}$ ,

$$\|\beta - \hat{\beta}_0\|_{BV} \leq \sum_{j=1}^{K_n+m} |\alpha_j - \alpha_{j0}| \|B_j'\|_\infty = O(K_n) \{\varepsilon_n^2 (K_n+m)\}^{1/2} \rightarrow 0. \quad (\text{A2})$$

Therefore,  $\hat{\beta}$  has bounded total variation. Define

$$\hat{\Lambda}_0(t) = \frac{\sum_{j=1}^n I(Y_j \geq t) dN_j(t)}{\sum_{j=1}^n I(Y_j \geq t) \exp\{\gamma_0^T X_j + \beta_0(t - C_j) I(t \geq C_j)\}}.$$

Then it can be shown that  $\|\hat{\Lambda}_0 - \Lambda_0\|_{BV} = O_p(n^{-1/2})$ , so we have  $P_n l(\hat{\Lambda}_0, \hat{\beta}_0, \psi_0) = n^{-1} \log(n^{-1}) + O_p(1)$ . Because  $P_n l(\hat{\Lambda}_\beta, \hat{\psi}_\beta) \rightarrow P_n l(\hat{\Lambda}_0, \hat{\beta}_0, \psi_0)$ , similar algebra to that in Murphy (1994, pp. 718–23) yields the inequality

$$n^{-1} \sum_{j=1}^n \int \{n \Delta \hat{\Lambda}_\beta(t)\} dN_j(t) - c_1 \log\{1 + \hat{\Lambda}_\beta(\tau)\} \geq O(1)$$

for some constant  $c_1$  and  $O(1)$  that are independent of  $\beta \in N_{\varepsilon_n}$ . Hence, using the same partition arguments as in Murphy (1994), we conclude that  $\limsup_n \{\sup_{\beta \in N_{\varepsilon_n}} \Lambda_\beta(\tau)\}$  is finite with probability tending to one. Moreover, since  $P_n l_\Lambda(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)[h] = 0$ ,  $P_n l_\psi(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)^\top b = 0$ , we obtain

$$(P_n - P)l_\Lambda(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)[h] = Pl_\Lambda(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)[h], (P_n - P)l_\psi(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)^\top b = Pl_\psi(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)^\top b.$$

The left-hand sides of the equations are  $O_p(n^{-1/2})$  because both  $l_\Lambda$  and  $l_\psi$  are Donsker due to the fact that both  $\hat{\Lambda}_\beta$  and  $\beta$  belong to  $BV[0, \tau]$ . We apply Taylor expansion at the true  $(\Lambda_0, \beta_0, \psi_0)$  to the right-hand sides and obtain

$$\begin{aligned} O_p(n^{-1/2}) &= -\langle I_{11}(\Lambda_0, \beta_0, \psi_0)[h, b], [d\hat{\Lambda}_\beta - d\Lambda_0, \hat{\psi}_\beta - \psi_0] \rangle_{L_2(P)} \\ &\quad + o(\|\hat{\Lambda}_\beta - \Lambda_0\|_{BV} + \|\hat{\psi}_\beta - \psi_0\|) + O_p(\|\beta - \beta_0\|_{L_2}), \end{aligned}$$

where  $I_{11}$  is the operator in  $I$  corresponding to  $\Lambda$  and  $\psi$ . Using the invertibility of  $I_{11}$ , we have

$$\|\Lambda_\beta - \Lambda_0\|_{BV} + \|\hat{\psi}_\beta - \psi_0\| = A_n(n^{-1/2} + \|\beta - \beta_0\|_{L_2}), \quad (A3)$$

where  $\sup_{\beta \in N_{\varepsilon_n}} |A_n|$  is a bounded random variable.

We now consider  $B_n \equiv P_n l(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta) - P_n l(\hat{\Lambda}_0, \hat{\beta}_0, \hat{\psi}_0)$ . First,

$$B_n = (P_n - P)\{l(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta) - l(\hat{\Lambda}_0, \hat{\beta}_0, \hat{\psi}_0)\} + P\{l(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta) - l(\hat{\Lambda}_0, \hat{\beta}_0, \hat{\psi}_0)\}.$$

The first term on the right-hand side is equal to  $C_n n^{-1/2}$  where  $\sup_{\beta \in N_{\varepsilon_n}} |C_n| \rightarrow 0$  in probability. For the second term, we apply Taylor expansion at the true values  $(\lambda_0, \beta_0, \psi_0)$ . Since the first derivative of the second term at  $(\lambda_0, \beta_0, \psi_0)$  is zero, the expansion becomes

$$\begin{aligned} &-\langle I(\Lambda^*, \beta^*, \psi^*)(d\hat{\Lambda}_\beta/d\hat{\Lambda}_0 - \lambda_0, \beta - \beta_0, \hat{\psi}_\beta - \psi_0), (d\hat{\Lambda}_\beta/d\hat{\Lambda}_0 - \lambda_0, \beta - \beta_0, \hat{\psi}_\beta - \psi_0) \rangle_{L_2(P)} \\ &\quad + O(\|\hat{\Lambda}_0 - \Lambda_0\|_\infty^2 + \|\hat{\beta}_0 - \beta_0\|_\infty^2), \end{aligned}$$

where  $(\Lambda^*, \beta^*, \psi^*)$  is between  $(\hat{\Lambda}_\beta, \beta, \hat{\psi}_\beta)$  and  $(\Lambda_0, \beta_0, \psi_0)$ . Using the result in (A1), we obtain

$$B_n = C_n n^{-1/2} - c_1/2 \|\beta - \beta_0\|_{L_2}^2 + D_n(n^{-1} + K_n^{-2r}).$$

Therefore, if  $\beta \in N_{\varepsilon_n}$ , the result from de Boor (1978, p. 155) gives  $\|\beta - \beta_0\|_{L_2}^2 \geq c_2 \varepsilon_n^2$  so that

$$B_n \leq \sup_{\beta \in N_{\varepsilon_n}} \{ |C_n| n^{-1/2} + D_n (n^{-1} + K_n^{-2r}) \} - c_1 c_2 \varepsilon_n^2 / 2.$$

Consequently, if we choose  $\varepsilon_n^2 = 4 \sup_{\beta \in N_{\varepsilon_n}} \{ |C_n| n^{-1/2} + D_n (n^{-1} + K_n^{-2r}) \} / (c_1 c_2)$ , then  $B_n < 0$ ; note that such  $\varepsilon_n$  still satisfies  $K_n^{3/2} \varepsilon_n \rightarrow 0$  due to  $r \geq 2$  and Condition 4. Hence, there exists a local maximum  $\hat{\beta}$  within this neighbourhood. Furthermore,  $\|\hat{\beta} - \beta_0\|_{BV} \rightarrow 0$  by (A2) and

$$\|\hat{\beta} - \beta_0\|_{L_2}^2 \leq \|\hat{\beta} - \hat{\beta}_0\|_{L_2}^2 + O(K_n^{-2r}) \leq \varepsilon_n^2 + K_n^{-2r} = o_p(n^{-1/2})$$

according to Condition 4. By (A3), the corresponding  $(\hat{\Lambda} = \Lambda_{\hat{\beta}}, \hat{\psi} = \psi_{\hat{\beta}})$  satisfies

$$\|\hat{\Lambda} - \Lambda_0\|_{BV} + \|\hat{\psi} - \psi_0\| = O_p(n^{-1/2}) + \|\hat{\beta} - \beta_0\|_{L_2} = o_p(n^{-1/4}).$$

## Proof of Theorem 2

For any  $h \in BV[0, \tau]$ ,  $\tilde{h}$  with bounded  $r$ th derivative in  $[0, \tau]$  and  $b \in \mathbb{R}^d$ , we have  $P_n l_{\Lambda}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[h] = 0$ ,  $P_n l_{\beta}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[\tilde{h}_n] = 0$ , and  $P_n l_{\psi}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})^T b = 0$ . Here,  $h_n$  is the projection of  $h$  on  $S_n$  and  $\|\tilde{h}_n - \tilde{h}\|_{\infty} = O(K_n^{-r})$ . This gives

$$\begin{aligned} & G_n \{ l_{\Lambda}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[h] + l_{\beta}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[\tilde{h}_n] + l_{\psi}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})^T b \} \\ &= -n^{1/2} P \{ l_{\Lambda}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[h] + l_{\beta}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[\tilde{h}_n] + l_{\psi}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})^T b \}, \end{aligned} \quad (\text{A4})$$

where  $G_n = n^{1/2} (P_n - P)$ . It is straightforward to verify

$$\{ l_{\Lambda}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[h] + l_{\beta}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})[\tilde{h}_n] + l_{\psi}(\hat{\Lambda}, \hat{\beta}, \hat{\psi})^T b : \|h\|_{BV} \leq 1, \|\tilde{h}\|_{\infty} \leq 1, \|b\| \leq 1 \}$$

is P-Donsker. Thus, the left-hand side of (A4) is equal to

$$G_n \{ l_{\Lambda}(\Lambda_0, \beta_0, \psi_0)[h] + l_{\beta}(\Lambda_0, \beta_0, \psi_0)[\tilde{h}] + l_{\psi}(\Lambda_0, \beta_0, \psi_0)^T b \} + o_p(1),$$

where  $G_n = n^{1/2} (P_n - P)$  and  $o_p(1)$  here and in the sequel refers to some random element that converges in probability to zero uniformly in  $(h, \tilde{h}, b)$ . On the other hand, the right-hand side of (A4), after the Taylor expansion gives

$$\begin{aligned} & -n^{1/2} \{ 1 + o_p(1) \} \left\{ \int h^* d(\hat{\Lambda} - \Lambda_0) + \int \tilde{h}^* (\hat{\beta} - \beta_0) dt + (\hat{\psi} - \psi_0)^T b \right\} \\ & + n^{1/2} O(\|\hat{\Lambda} - \Lambda_0\|_{BV}^2 + \|\hat{\psi} - \psi_0\|^2 + \|\hat{\beta} - \beta_0\|_{L_2}^2 + \|\tilde{h}_n - \tilde{h}\|_{L_2}^2) \\ &= -n^{1/2} \{ 1 + o_p(1) \} \left\{ \int h^* d(\hat{\Lambda} - \Lambda_0) + \int \tilde{h}^* (\hat{\beta} - \beta_0) dt + (\hat{\psi} - \psi_0)^T b^* \right\} + o_p(1), \end{aligned}$$

where  $(h^*, \tilde{h}^*, b^*) = I(\Lambda_0, \beta_0, \psi_0)[h, \tilde{h}, b]$ . This yields

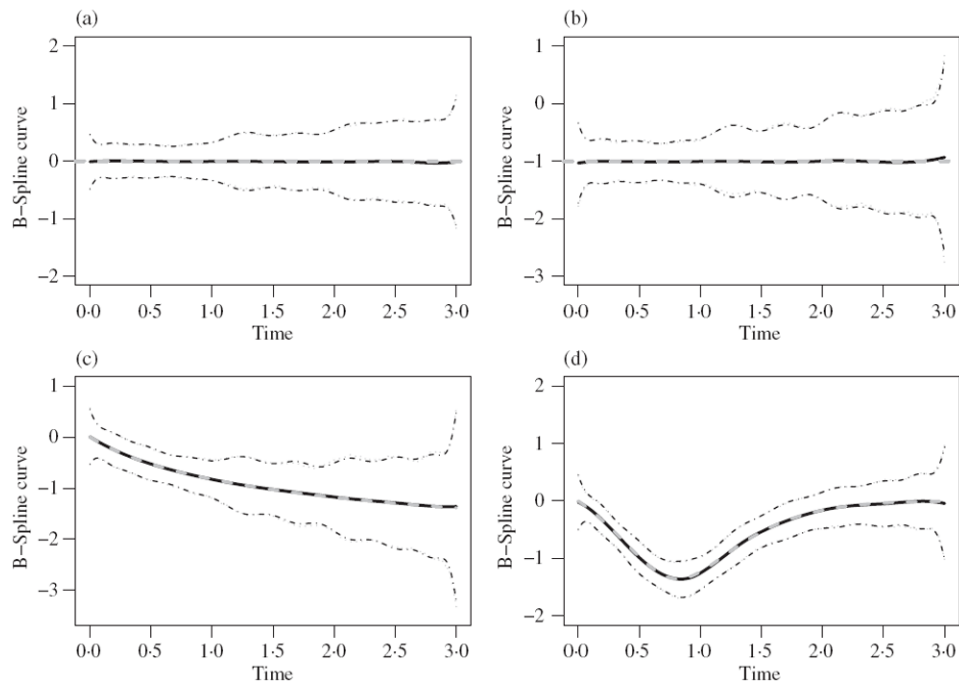
$$G_n\{l_\Lambda(\Lambda_0, \beta_0, \psi_0)[h^{**}] + l_\beta(\Lambda_0, \beta_0, \psi_0)[\tilde{h}^{**}] + l_\psi(\Lambda_0, \beta_0, \psi_0)^T b^{**}\} + o_p(1) \\ = -n^{1/2} \left\{ \int h d(\hat{\Lambda} - \Lambda_0) + \int \tilde{h}(\hat{\beta} - \beta_0) dt + (\hat{\psi} - \psi_0)^T b \right\}$$

where  $(h^{**}, \tilde{h}^{**}, b^{**}) = I(\Lambda_0, \beta_0, \psi_0)^{-1}[h, \tilde{h}, b]$ . Therefore, Theorem 2 holds.

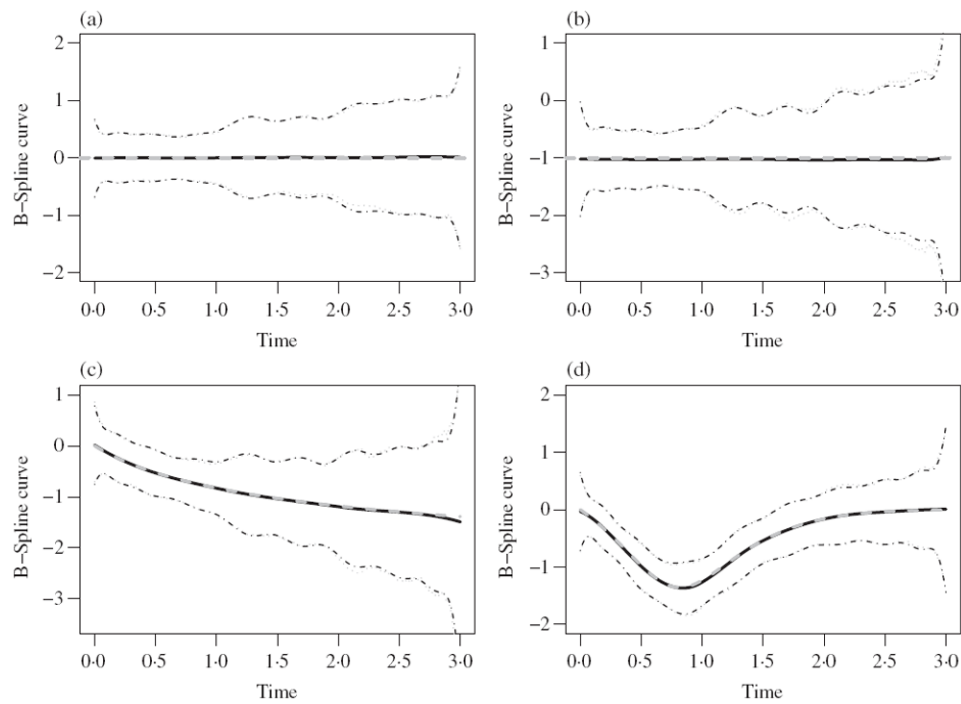
## References

- Akacha M, Benda N. The impact of dropouts on the analysis of dose-finding studies with recurrent event data. *Statist Med.* 2010; 29:1635–46.
- Anderson JE, Louis TA. Survival analysis using a scale change random effects model. *J Am Statist Assoc.* 1995; 90:669–79.
- Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Statist Med.* 2002; 21:2449–63.
- Cai J, Fan J, Jiang J, Zhou H. Partially linear hazard regression with varying coefficients for multivariate survival data. *J R Statist Soc B.* 2008; 70:141–58.
- Cai J, Fan J, Zhou H, Zhou Y. Hazard models with varying coefficients for multivariate failure time data. *Ann Statist.* 2007; 35:324–54.
- Cai J, Prentice RL. Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika.* 1995; 82:151–64.
- Cai J, Prentice RL. Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Anal.* 1997; 3:197–213. [PubMed: 9384652]
- Cai J, Zeng D. Additive mixed effect model for clustered failure time data. *Biometrics.* 2001; 67:1340–51. [PubMed: 21418052]
- Cai Z, Fan J, Li R. Efficient estimation and inferences for varying-coefficient models. *J Am Statist Assoc.* 2000; 95:888–902.
- Cai Z, Sun Y. Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand J Statist.* 2003; 30:93–111.
- Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model. *J R Statist Soc A.* 1985; 148:82–117.
- de Boor, C. A Practical Guide to Splines. Wroclaw: Springer; 1978.
- Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. *Ann Statist.* 2002; 30:74–99.
- Friede T, Schmidli H. Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. *Statist Med.* 2010a; 29:1145–56.
- Friede T, Schmidli H. Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Meth Info Med.* 2010b; 49:618–24.
- Gilden D. Infectious causes of multiple sclerosis. *Lancet Neurol.* 2005; 4:195–202. [PubMed: 15721830]
- Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J Am Statist Assoc.* 1992; 87:942–51.
- Kappos L, Radue E, O'Connor P, Polman C, Hohlfeld R, Calabresi P, Selmaj K, Agoropoulou C, Leyk M, Zhang-Auberson L, Burtin P. Freedoms Study Group. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *New Engl J Med.* 2010; 362:387–401. [PubMed: 20089952]
- Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics.* 1992; 48:795–806. [PubMed: 1420842]
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J R Statist Soc B.* 2000; 62:711–30.
- Louis TA. Finding the observed information matrix when using the EM algorithm. *J R Statist Soc B.* 1982; 13:2233–47.

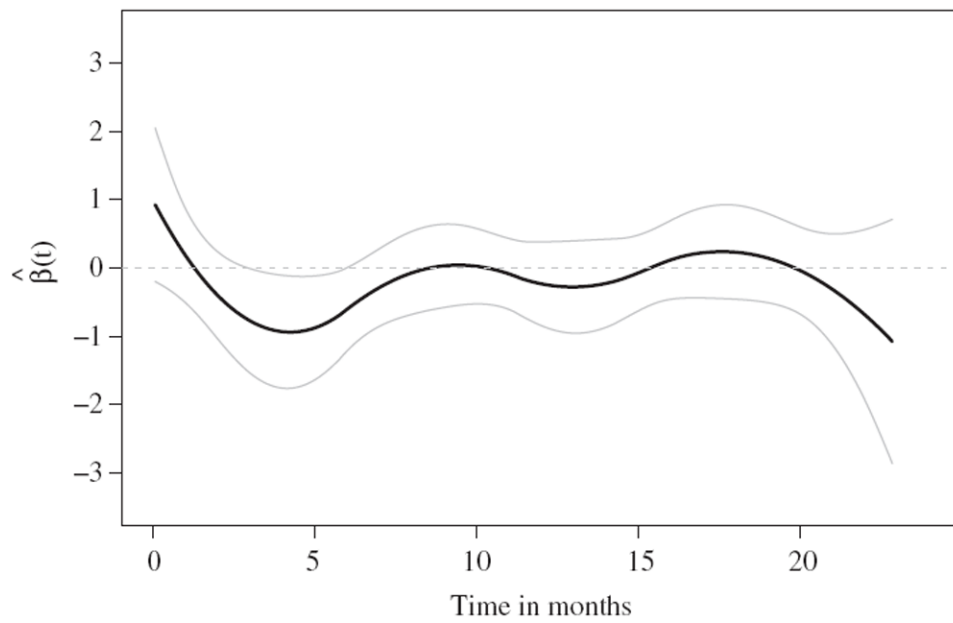
- Lycke J, Svennerholm B, Hjelmquist E, Frisén L, Badr G, Andersson M, Vahlne A, Andersen O. Acyclovir treatment of relapsing-remitting multiple sclerosis. A randomized, placebo-controlled, double-blind study. *J Neurol*. 1996; 243:214–24. [PubMed: 8936350]
- Martinussen T, Scheike T, Skovgaard IM. Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scand J Statist*. 2002; 29:57–74.
- Martinussen T, Scheike TH, Zucker DM. The Aalen additive gamma frailty hazards model. *Biometrika*. 2011; 98:831–43.
- Murphy SA. Testing for a time dependent coefficient in Cox's regression model. *Scand J Statist*. 1993; 20:35–50.
- Murphy SA. Consistency in a proportional hazards model incorporating a random effect. *Ann Statist*. 1994; 22:712–31.
- Murphy SA. Asymptotic theory for the frailty model. *Ann Statist*. 1995; 23:182–98.
- Nicholas R, Straube S, Schmidli H, Pfeiffer S, Friede T. Time-patterns of annualized relapse rates in randomized placebo-controlled clinical trials in relapsing multiple sclerosis: A systematic review and meta-analysis. *Mult Scler*. 2012; 18:1290–6. [PubMed: 22422808]
- Nielsen GG, Gill RD, Andersen PK, Sørensen TIA. A counting process approach to maximum likelihood estimation in frailty models. *Scand J Statist*. 1992; 19:25–43.
- Oakes D, Jeong J. Frailty models and rank tests. *Lifetime Data Anal*. 1998; 4:209–28. [PubMed: 9787603]
- Parner E. Asymptotic theory for the correlated gamma-frailty model. *Ann Statist*. 1998; 26:183–214.
- Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *J Am Statist Assoc*. 1993; 88:811–20.
- R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. <http://www.R-project.org>
- Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist A*. 1991; 20:2609–31.
- Rudin, W. *Functional Analysis*. New York: McGraw-Hill; 1973.
- Schumaker, L. *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press; 2007.
- Shao J, Chang J, Chow S-C. Statistical inference for cancer trials with treatment switching. *Statist Med*. 2005; 24:1783–90.
- Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. *J Am Statist Assoc*. 1998; 93:1164–75.
- Tian L, Zucker D, Wei LJ. On the Cox model with time-varying regression coefficients. *J Am Statist Assoc*. 2005; 100:172–83.
- Wang Y, Meyerson L, Tang Y, Qian N. Statistical methods for the analysis of relapse data in MS clinical trials. *J Neurol Sci*. 2009; 285:206–11. [PubMed: 19709676]
- Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Statist Assoc*. 1989; 84:1065–73.
- Zeng D, Chen Q, Chen M-H, Ibrahim JG. Amgen Research Group. Estimating treatment effects with treatment switching via semicompeting risks models: An application to a colorectal cancer study. *Biometrika*. 2012; 99:167–84. [PubMed: 23049136]
- Zeng D, Lin DY. Semiparametric transformation models with random effects for recurrent events. *J Am Statist Assoc*. 2007; 102:167–80.
- Zucker DM, Karr AF. Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Ann Statist*. 1990; 18:329–53.

**Fig. 1.**

Estimated treatment effect curves  $\beta(t)$  for simulation studies with  $n = 400$ . True (dashed grey line), estimates (solid black line), empirical 95% confidence interval (dotted grey line), and average of 95% confidence intervals (dashed-dotted black line) for  $\beta(t)$  specified as (a) 0, (b) -1, (c)  $\log\{1 - 0.9/(0.6 + t)\}$ , (d)  $\log\{1 - 1.5t \exp(-0.75t^2)\}$ .



**Fig. 2.** Estimated treatment effect curves  $\beta(t)$  for simulation studies with  $n = 200$ . True (dashed grey line), estimates (solid black line), empirical 95% confidence interval (dotted grey line), and average of 95% confidence interval (dashed-dotted black line) for  $\beta(t)$  specified as (a) 0, (b) -1, (c)  $\log\{1 - 0.9t/(0.6 + t)\}$ , (d)  $\log\{1 - 1.5t \exp(-0.75t^2)\}$ .



**Fig. 3.** Time-varying treatment effect in acyclovir study. The solid black curve is  $\hat{\beta}(t)$  with  $K_n = M = 3$ , the solid grey curves show the pointwise 95% confidence intervals, and the dashed grey curve is the null treatment reference line.

Table 1

Simulation results for four types of treatment effects

Scenario	Parameter	True value	$n = 400$				$n = 200$			
			Bias	SE	SEE	CP	Bias	SE	SEE	CP
(a)	$\gamma$	0.20	-0.1	3.9	3.7	94	0.1	5.6	5.3	93
	$\theta$	0.80	-0.9	7.4	8.9	96	-1.0	10.9	12.8	95
	$\beta(0.5)$	0.00	-0.4	14.6	14.8	94	-0.1	20.7	20.9	93
	$\beta(1.5)$	0.00	-0.8	22.0	22.9	96	0.8	32.0	32.2	96
	$\beta(2.5)$	0.00	-0.6	35.7	36.2	94	1.3	51.3	51.0	95
	$\beta^-$	0.00	-0.9	19.8	20.5	94	0.6	28.0	28.8	93
(b)	$\gamma$	0.20	0.2	4.1	4.1	95	0.1	5.8	5.8	95
	$\theta$	0.80	-0.7	9.4	8.6	91	-1.8	13.0	12.6	93
	$\beta(0.5)$	-1.00	-1.2	18.7	18.4	95	-2.7	26.1	26.3	95
	$\beta(1.5)$	-1.00	-1.3	27.1	27.4	94	-2.2	39.8	39.1	96
	$\beta(2.5)$	-1.00	-1.7	45.5	45.4	95	-2.7	67.0	64.9	95
	$\beta^-$	-1.00	-0.6	22.2	22.3	95	-2.7	32.5	31.7	95
(c)	$\gamma$	0.20	-0.1	4.0	3.9	94	-0.4	5.8	5.6	95
	$\theta$	0.80	-1.1	8.8	8.8	94	-1.3	12.6	12.7	92
	$\beta(0.5)$	-0.53	-0.2	16.3	16.3	96	-0.5	23.2	23.2	95
	$\beta(1.5)$	-1.03	0.3	26.1	25.9	94	-0.8	36.8	36.9	95
	$\beta(2.5)$	-1.29	0.6	45.2	45.6	95	-1.01	67.0	65.4	95
	$\beta^-$	-0.93	0.2	21.5	21.5	94	-1.1	30.8	30.7	95
(d)	$\gamma$	0.20	0.0	4.0	3.9	94	-0.1	5.5	5.5	94
	$\theta$	0.80	-0.7	8.5	8.8	94	-1.6	11.1	12.7	95
	$\beta(0.5)$	-0.97	-1.2	14.2	14.5	96	-1.5	20.2	20.6	96
	$\beta(1.5)$	-0.54	0.2	13.7	14.2	96	0.4	20.3	20.3	96
	$\beta(2.5)$	-0.04	-1.4	20.6	20.2	94	0.6	28.5	28.9	96
	$\beta^-$	-0.50	-0.5	10.5	10.5	94	-0.7	14.8	14.9	94

Scenarios (a)  $\beta(t) = 0$ , (b)  $\beta(t) = -1$ , (c)  $\beta(t) = \log(1 - 0.9t/(0.6 + t))$ , (d)  $\beta(t) = \log(1 - 1.5t \exp(-0.75t^2))$ ; Bias( $\times 100$ ), bias of the parameter estimate; SE( $\times 100$ ), standard error of the estimate; SEE( $\times 100$ ), mean of the standard error estimate; CP(%), coverage probability of the 95% confidence interval;  $\beta$ , average treatment effect defined as  $\int_0^T \beta(t) dt / \tau$ .