Biometrics 72, 1113–1122
December 2016
DOI: 10.1111/biom.12503

Marginal Regression Analysis of Recurrent Events with Coarsened Censoring Times

X. Joan Hu[1,*] and Rhonda J. Rosychuk[2]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada
[2]Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada
[*]*email:* joanh@stat.sfu.ca
[**]*email:* rhonda.rosychuk@ualberta.ca

Summary. Motivated by an ongoing pediatric mental health care (PMHC) study, this article presents weakly structured methods for analyzing doubly censored recurrent event data where only coarsened information on censoring is available. The study extracted administrative records of emergency department visits from provincial health administrative databases. The available information of each individual subject is limited to a subject-specific time window determined up to concealed data. To evaluate time-dependent effect of exposures, we adapt the local linear estimation with right censored survival times under the Cox regression model with time-varying coefficients (cf. Cai and Sun, *Scandinavian Journal of Statistics* 2003, **30**, 93–111). We establish the pointwise consistency and asymptotic normality of the regression parameter estimator, and examine its performance by simulation. The PMHC study illustrates the proposed approach throughout the article.

Key words: Administrative records; Doubly censored data; Event rate/mean function; Local linear estimation; Time-varying regression coefficient.

1. Introduction

Administrative databases provide sources of rich data but are often developed for non-research purposes. Issues of privacy or confidentiality may constrain the actual information released to researchers, particularly true for health related data. To answer specific health questions based on administrative records, researchers often encounter challenges to accommodate features of the released data in order to perform valid statistical analyses. Typical challenges include that the information availability varies from individual to individual since a scientifically meaningful time origin is usually subject-specific. Often each individual's information is both left- and right-censored, that is, doubly censored as referred to by Zhang and Li (1996) and Cai and Cheng (2004), for example. Further, the censoring times are missing with part or all of the study individuals in many applications.

The Ambulatory Care Classification System (ACCS) (Alberta Health and Wellness, 2004) is an example of a large, population-based database in the province of Alberta, Canada. It was initiated in 1999 by the provincial health ministry, now known as Alberta Health, who provides health services. All Alberta residents access health care at no personal cost in a uniform single-payer health system. The ACCS database includes all presentations made by Alberta residents to Alberta Emergency Departments (EDs) since 1999. With ethics approvals and data agreements in place, researchers can request an extraction from the ACCS database, although aspects of the delivered data may not be at the level of detail as the data stored. Several research studies have accessed the ACCS database for different health conditions. The pediatric mental health care study (PMHC for short) focuses on presentations to Alberta EDs made by children and youth (aged younger or equal to 17 years old at the time of the ED visit) for mental health reasons during April 1, 2002 to March 31, 2011. The extracted study information contains 41,159 ED visits made by 27,947 individuals and is referred to as the PMHC dataset in the rest of this article. The dataset includes demographic data (e.g., age at ED visit, age at fiscal year end, sex) and ED visit data (e.g., start and end dates and times of the ED visit, diagnosis).

The majority of mental illnesses begin in childhood (Leitch, 2007). The capacity of the pediatric mental health care system is exceeded by the need for care and there are limited options for mental health care. This situation may drive some families to seek help in EDs. The EDs may be the first point of contact with the health system for mental health for children in crisis (Halamandaris and Anderson, 1999). Newton et al. (2011) present a preliminary analysis of the PMHC dataset during 2002–2008. They observe a significant heterogeneity in mental health presentations and various patterns of repeated ED use. Additional findings include the impact of health system factors on patient outcomes, a lack of community-based care available to children and youth, and an ongoing need for mental health services. Demands to provide more insights into the heterogeneity and the impact of the factors/exposures motivated this article. We aim to assess the effects of risk factors/exposures and to evaluate the frequency of pediatric ED visits for mental health (hereafter MHED visits) in a population using the available administrative records.

This article formulates the MHED visit records in the PMHC dataset as recurrent event data (Cook and Lawless, 2007). The range of age differences among records within

[Correction added on July 12, 2017, after first online publication: Minor errors in a C routine corrected. Corrections described in text file supplied with the updated code.]

individual subjects and among them is rather wide. It motivated us to consider a marginal regression model for the counting process of an individual's cumulative MHED visits over age. The model is analogous to the extended Cox regression model with time-varying coefficients for a Poisson process, and allows for the exploration of age-dependent risk/exposure (covariate) effects. We adapt the local partial likelihood procedure of Cai and Sun (2003) and Tian et al. (2005) with the recurrent event data in the PMHC dataset. The records of the MHED visits were from individuals not older than 17 years of age during April 1, 2002, to March 31, 2011. This gives rise to doubly censored data in the sense analogous to the censoring of observations on a quantity such as an event time considered in Zhang and Li (1996) and Cai and Cheng (2004) if we assume the population is closed. Moreover, Alberta Health's privacy protocol prevents the release of the individuals' birthdates for this project. It results in incomplete information on censoring times of the PMHC dataset.

Considerably rich literature is available on statistical analysis of event history data under Cox regression models with time-varying coefficients, an extension of the popular proportional hazards model (Cox, 1972). For example, Zucker and Karr (1990) present a penalized partial likelihood approach, and Murphy and Sen (1991) adapt the method of sieves. More recently, in addition to the local linear estimation of the time-varying coefficients via a kernel-weighted partial likelihood function proposed by Cai and Sun (2003) and Tian et al. (2005), Nan et al. (2005) conduct a regression analysis using B-splines. Chen et al. (2012) advocate a global partial likelihood method in a more general setting of the Cox model with varying coefficients. Most of the published articles focus on analysis of right-censored event times; there are notable exceptions such as Amorim et al. (2008), which presents regression splines under a proportional rates model for recurrent event data.

Some researchers such as Kim et al. (1993) and Gomez and Lagakos (1994) refer to interval censored event times as doubly censored; see Chapter 8 of Sun (2006) for a comprehensive review of the inference procedures with doubly censored survival times in this sense. We adopt an alternative definition of doubly censored data, which is used in Zhang and Li (1996) and Cai and Cheng (2004), for example. Despite substantial research on incomplete data analysis, few articles deal with incomplete observations of censoring time; exceptions include Hu et al. (1998) and Wang (2011), which assume the censoring time distribution is known or can be estimated with an independent set of survey data. The coarsened censoring times in our application result from the unreleased individual birthdates together with the ED records of age information aggregated by years. We consider the commonly used assumption that the distribution of birthdays is uniform in a calendar year, and employ the available relevant information in the dataset to specify the distributions of different individuals. In fact, the frequency distribution of birthdates in the year 2011 provided by Alberta Health is in agreement with the uniform distribution.

The rest of this article is organized as follows. The notation and modeling are introduced in Section 2. Section 3 starts with an adaption of the procedure in Cai and Sun (2003) under the extended Cox regression model with recurrent events. The adaption is then used to motivate our proposed estimation procedures. Section 4 shows an analysis of the PMHC dataset applying the proposed approach. Final remarks are given in Section 5. Although this article is presented via the PMHC study that motivated the research, the approaches and discussions can apply more broadly.

2. Notation and Model

We focus on the individuals who have records of MHED visits in the PMHC dataset and assume they are independent. Let $N_i(a)$ represent the count of subject i's cumulative MHED visits since birth at age a, for $i = 1, \ldots, n$, $a > 0$; Z_i, subject i's external covariates. We assume the expected rate function conditional on Z_i is

$$E\{dN_i(a) \mid Z_i\} = \exp\{\beta(a)^{'} Z_i\} d\Lambda_0(a), \tag{1}$$

where $\Lambda_0(a)$ is the cumulative baseline rate function $\int_0^a \lambda_0(u)du$. Here the baseline rate function $\lambda_0(\cdot) > 0$ is unspecified, and the time-varying coefficients $\beta(\cdot)$ have continuous second derivatives.

The model specified by (1) is an extension of the proportional rates/means model considered by Pepe and Cai (1993), Lawless and Nadeau (1995), and Lin et al. (2000) in their marginal analysis with recurrent events. Our model includes it as a special case that $\{N_i(a) : a > 0\}$ is a Poisson process and its conditional intensity function satisfies

$$\lambda(a \mid \mathcal{H}_i(a)) = \lambda_0(a) \exp\{\beta(a)^{'} Z_i\}, \tag{2}$$

where $\mathcal{H}_i(a)$ denotes all the history information prior to age a of subject i. The Poisson process model (2), an extended Cox regression model, extends the Anderson–Gill model studied by Andersen and Gill (1982) to accommodate time-varying covariate effects. Our approaches are proposed for situations with the marginal model (1) and applicable to the situations with model (2) since model (1) is more general.

Denote the window in the calendar time of the data extraction by $[W_L, W_R]$ and subject i's birthdate in the calendar time by B_i. Since only the information on the MHED visits from Alberta residents aged younger than 18 years old is extracted, the MHED records of subject i are possibly available during his age (in years) over $(C_{Li}, C_{Ri}]$ in the PMHC dataset, where $C_{Li} = \max(0, W_L - B_i)$ and $C_{Ri} = \min(18, W_R - B_i)$. We assume that the population is closed, both W_L and W_R are independent of the MHED visit records, and subject i's birthdate B_i is independent of the counting process $N_i(\cdot)$ conditional on Z_i. Under the assumptions, the PMHC data are a collection of doubly censored counting processes coupled with the covariates, where subject i's left- and right-censoring times are C_{Li} and C_{Ri}, independent of the counting process $N_i(\cdot)$ conditional on Z_i.

Let T_{ij} be the calendar time of subject i's jth ED visit in the PMHC dataset and then the subject's age at the visit is $A_{ij} = T_{ij} - B_i$. Denote the total number of subject i's ED visits in the PMHC dataset by $N_i^* = N_i(C_{Ri}) - N_i(C_{Li})$. The available information includes T_{ij} and only the integer part

of the age A_{ij}, denoted by $\lceil A_{ij} \rceil$, for $j = 1, \ldots, N_i^*$. That is, the PMHC dataset contains only the change of the segment of $\{N_i(a) : a > 0\}$ over the interval $(C_{Li}, C_{Ri}]$ to $N_i(C_{Li})$, provided B_i is available: $N_i(a) - N_i(C_{Li}) = \sum_{j=1}^{N_i^*} I\big(A_{ij} \le a\big)$ for $a \in (C_{Li}, C_{Ri}]$.

The indicator of subject i's observation window $Y_i(a) = I(a \in (C_{Li}, C_{Ri}])$ can be presented as $Y(a\big|B_i)$, where $Y(a\big|B) = I\big(\max(0, W_L - B) < a \le \min(18, W_R - B)\big)$ with a given data extraction window $[W_L, W_R]$. Provided with B_i, the derived counting process $\big\{N_i^*(a) = \int_0^a Y(u\big|B_i) dN_i(u),\ a > 0\big\}$ is fully available and satisfies

$$E\{dN_i^*(a) \mid Z_i, B_i\} = Y(a\big|B_i) \exp\{\beta(a)' Z_i\} d\Lambda_0(a) \tag{3}$$

under our model assumption. This allows us to adapt well-established inference approaches in event history data analysis.

However, the PMHC dataset does not include the individual birthdates B_i's due to the privacy protocol. Hence, the censoring times C_{Li} and C_{Ri} are missing and only coarsened information on them is available. The available information on $dN_i(a)$ for $C_{Li} < a \le C_{Ri}$ is consequently coarsened. In the following section, we start with a procedure for estimating the time-varying regression coefficient $\beta(\cdot)$ when the birthdates B_i's are available under the extended Cox model with a Poisson process. Subsequently presented are procedures for estimating $\beta(\cdot)$ developed under the marginal model (1) with either available or missing B_i's.

3. Estimation Procedures

We begin with a procedure for estimating $\beta(\cdot)$ under (2), an extended Cox model with a Poisson process. It motivates the proposed estimation procedures in Section 3.2 under the marginal regression model (1).

3.1. *Estimation under Extended Cox Model with Poisson Process (2) when Birthdates B_i's are Available*

With the PMHC dataset plus the individuals' birthdates B_i's, the likelihood function of the unknown parameter functions $\lambda_0(\cdot)$ and $\beta(\cdot)$ involved in the intensity functions of the counting processes $N_i(\cdot)$ under model (2) in Section 2 is

$$\prod_{i=1}^{n} \prod_{a \in (C_{Li}, C_{Ri}]} \big[\lambda_0(a) \exp\{\beta(a)' Z_i\}\big]^{dN_i(a)} \times \big\{1 - \exp\{\beta(a)' Z_i\} d\Lambda_0(a)\big\}^{1 - dN_i(a)}.$$

In principle, one may obtain the nonparametric maximum likelihood estimators of $\lambda_0(\cdot)$ and $\beta(\cdot)$ jointly in the sense of Kiefer and Wolfowitz (1956) by maximizing the likelihood function or its log transformation. Extending the local linear partial likelihood approach in Cai and Sun (2003) and Tian et al. (2005) with right-censored survival times, one may consider the local linear maximum partial likelihood estimator (MPLE) of $\beta(\cdot)$ as follows.

Since $C_{Li} > 0$ and $C_{Ri} \le 18$ in years, the log-partial likelihood function of $\beta(\cdot)$ under model (2) is

$$\sum_{i=1}^{n} \int_0^{18} Y(u\big|B_i) \Big\{ \beta(u)' Z_i - \log\Big(\sum_{l=1}^{n} Y(u\big|B_l) \exp\{\beta(u)' Z_l\} \Big) \Big\} dN_i(u). \tag{4}$$

Choose constants $0 < \tau_L, \tau_R < 18$ such that the left and right censoring times satisfy $P(C_L < \tau_L) > 0$ and $P(C_R > \tau_R) > 0$ to avoid the boundary problem in the local likelihood estimation. For a fixed $a \in [\tau_L, \tau_R]$, approximate $\beta(u)$ by the Taylor expansion to the first order: $\beta(a) + \dot{\beta}(a)(u - a)$. Let $\gamma = \big(\beta(a)', \dot{\beta}(a)'\big)'$ and $Z_i^*(u, a) = \big(Z_i', (u - a) Z_i'\big)'$. Using a kernel function $K(\cdot)$ and substituting $\beta(u)$ in (4) with its linear approximation yields the local linear partial likelihood function of γ:

$$l_n(\gamma; a\big|\mathbf{B}) = \sum_{i=1}^{n} \int_0^{18} K_h(u - a) Y(u\big|B_i) \Big\{ \gamma' Z_i^*(u, a) - \log\Big(\sum_{l=1}^{n} Y(u\big|B_l) \exp\{\gamma' Z_l^*(u, a)\} \Big) \Big\} dN_i(u), \tag{5}$$

where $K_h(\cdot) = K(\cdot/h)/h$. Following the arguments of Cai and Sun (2003), we can show that the Hessian matrix $\partial^2 l_n(\gamma; a\big|\mathbf{B})/\partial\gamma^2$ divided by n converges to a negative definite matrix a.s. under some mild conditions. The concavity of $l_n(\gamma; a\big|\mathbf{B})$ ensures its unique maximum point, denoted by $\hat{\gamma}_n$. The first component vector of $\hat{\gamma}_n$, denoted by $\hat{\beta}_n(a\big|\mathbf{B})$, is an adapted version of Cai and Sun's local linear maximum partial likelihood estimator (MPLE) for $\beta(a)$, $a \in [\tau_L, \tau_U]$. Extending the arguments of Cai and Sun (2003), we can establish the pointwise consistency and weak convergence of the local linear MPLE $\hat{\beta}_n(a\big|\mathbf{B})$ for $a \in [\tau_L, \tau_U]$ with the adaption of the required conditions for the current situation. The asymptotics properties require $h = O(n^{-\nu})$ with $1/2 < \nu < 1$, as pointed out in Tian et al. (2005).

3.2. *Proposed Estimation Procedures*

3.2.1. Estimation with available birthdates B_i's. Let $\bar{Z}_n^*(\gamma; u, a) = S_n^{(1)}(\gamma; u, a)\big/S_n^{(0)}(\gamma; u, a)$, where $S_n^{(q)}(\gamma; u, a) = \sum_{i=1}^{n} Y(u\big|B_i)\big[Z_i^*(u, a)\big]^{\otimes q} \exp\{\gamma' Z_i^*(u, a)\}\big/n$ with $A^{\otimes q} = 1, A, AA'$ for $q = 0, 1, 2$, respectively. Then,

$$U_n(\gamma; a\big|\mathbf{B}) = \frac{1}{n} \frac{\partial l_n(\gamma; a\big|\mathbf{B})}{\partial\gamma} = \frac{1}{n} \int_0^{18} K_h(u - a) \times \sum_{i=1}^{n} Y(u\big|B_i) \big\{ Z_i^*(u, a) - \bar{Z}_n^*(\gamma; u, a) \big\} dN_i(u). \tag{6}$$

Note that the solution of $U_n(\gamma; a\big|\mathbf{B}) = 0$ is the maximum point of $l_n(\gamma; a\big|\mathbf{B})$ in (5). In fact, (6) defines an unbiased estimating function of $\beta(a)$ under the marginal model (1).

Further, denote the limit of $S_n^{(q)}(\gamma; u, a)$ as $n \to \infty$ by $s^{(q)}(\gamma; u, a)$ for $q = 0, 1, 2$, and $s^{(1)}(\gamma; u, a)/s^{(0)}(\gamma; u, a)$ by $\bar{z}^*(\gamma; u, a)$. Substituting $\bar{Z}_n^*(\gamma; u, a)$ with $\bar{z}^*(\gamma; u, a)$ in (6) leads to another estimating function $\tilde{U}_n(\gamma; a|\mathbf{B})$, which yields an estimator of $\beta(a)$ asymptotically equivalent to the local linear MPLE $\hat{\beta}_n(a|\mathbf{B})$ under the extended Cox model (2) with usual regularity conditions. We can show that $\tilde{U}_n(\gamma; a|\mathbf{B})$ is also asymptotically unbiased under the marginal model (1).

In Web Appendix A.1, we outline a proof of the pointwise consistency and weak convergence of $\hat{\beta}_n(a|\mathbf{B})$ with the bandwidth $h = O(n^{-\nu})$ where $1/2 < \nu < 1$ under model (1). We also present a consistent estimator of $\hat{\beta}_n(a|\mathbf{B})$'s asymptotic variance. The asymptotics derivation employs the modern empirical process theory (Pollard, 1990), similar to the approaches in Bilias et al. (1997), Lin et al. (2000), and Hu et al. (2003), for example. It can be extended to derive the asymptotic properties of the estimator proposed in Section 3.2.2. The arguments of Cai and Sun (2003) using martingale theory may be followed to establish the consistency and weak convergence of $\hat{\beta}_n(a|\mathbf{B})$ as the local linear MPLE under model (2). They are not applicable in the situations with the marginal model (1).

Note that, with fixed $\beta(\cdot)$, the following estimating equation is unbiased under model (1):

$$\sum_{i=1}^{n} \left[Y(a|B_i) dN_i(a) - Y(a|B_i) \exp\{\beta(a)' Z_i\} d\Lambda_0(a) \right] = 0, \quad a \in (0, 18). \tag{7}$$

This yields an extended Breslow estimator (Lin et al., 2000) for the cumulative baseline rate function:

$$d\hat{\Lambda}_{0n}(a|\mathbf{B}) = \frac{\sum_{i=1}^{n} Y(a|B_i) dN_i(a)}{\sum_{l=1}^{n} Y(a|B_l) \exp\{\hat{\beta}_n(a|\mathbf{B})' Z_l\}}, \quad a \in (0, 18). \tag{8}$$

Here we take the convention $0/0 = 0$, and $\hat{\beta}_n(a|\mathbf{B}) = \hat{\beta}_n(\tau_L|\mathbf{B})$ for $a \in (0, \tau_L)$ and $\hat{\beta}_n(a|\mathbf{B}) = \hat{\beta}_n(\tau_U|\mathbf{B})$ for $a \in (\tau_U, 18)$. With the consistency of $\hat{\beta}_n(a|\mathbf{B})$, we may establish the pointwise consistency of $\hat{\Lambda}_{0n}(a|\mathbf{B}) = \int_0^a d\hat{\Lambda}_{0n}(u|\mathbf{B})$ defined in (8). Further (8) can be used to estimate the baseline rate function $\lambda_0(\cdot)$ by the kernel smoothing:

$$\tilde{\lambda}_{0n}(a|\mathbf{B}) = \int_0^{18} K_g^*(u - a) d\hat{\Lambda}_{0n}(u|\mathbf{B}), \quad a \in (0, 18),$$

where $K_g^*(\cdot) = K^*(\cdot/g)/g$ with $K^*(\cdot)$ a kernel function. Integrating the arguments in Tian et al. (2005) and Lin et al. (2000), one may establish the uniform consistency and the weak convergence of the baseline estimator and the regression parameter estimator.

3.2.2. Estimation with missing B_i's. When the birthdates B_i's are unavailable, the censoring times C_{Li} and C_{Ri} are missing and the information on the change of the segment of $\{N_i(a) : a > 0\}$ over the interval $(C_{Li}, C_{Ri}]$ is coarsened if the ages at the visits are recorded in years. The estimator $\hat{\beta}_n(\cdot|\mathbf{B})$ in Section 3.2.1 is then not evaluable. The PMHC dataset exemplifies the situation. We propose the following approach to accommodate it.

It is a commonly acceptable assumption that birthdates follow the uniform distribution over a calendar year. In addition, the birthdate of subject i is the difference of the calendar time and his/her age at his/her visits, $B_i = T_{ij} - A_{ij}$ for all j. We infer that the missing B_i is contained in all the intervals of $\left(T_{ij} - \lceil A_{ij} \rceil - 1, T_{ij} - \lceil A_{ij} \rceil\right]$ for $j = 1, \ldots, N_i^*$. This inference, together with $B_i \le W_R$ and $B_i + 18 > W_L$, yields that, conditional on the subjects' available records in the PMHC dataset, the missing B_i's can be viewed as independent random variables from the uniform distributions over the intervals

$$I_i = \left(W_L - 18, W_R\right] \bigcap \left\{ \bigcap_{j=1}^{N_i} \left(T_{ij} - \lceil A_{ij} \rceil - 1, T_{ij} - \lceil A_{ij} \rceil\right] \right\},$$

for $i = 1, \ldots, n$, denoted by $B_i \sim G_i(\cdot) = \text{Unif}(I_i)$.

Let $\tilde{Y}_i(a) = \int_0^\infty Y(a|b) dG_i(b)$ and $d\tilde{N}_i^*(a) = \int_0^\infty \{Y(a|b) dN_i(a)\} dG_i(b)$. They are the expectations of $Y(a|B_i)$ and $dN_i^*(a)$ conditional on the available data, respectively, attained by integrating out B_i's with $B_i \sim \text{Unif}(I_i)$. Here $d\tilde{N}_i^*(a) = E\{dN_i^*(a) \mid \text{available data}\}$ is not in general the product of $\tilde{Y}_i(a)$ and $E\{dN_i(a) \mid \text{available data}\}$ since $Y(a|B_i)$ and $dN_i(a)$ are not necessarily independent conditional on the available information. We consider the following estimating function of γ to estimate $\beta(a)$ for $a \in [\tau_L, \tau_R]$:

$$E_n(\gamma; a) = \frac{1}{n} \int_0^{18} K_h(u - a) \sum_{i=1}^{n} \left\{ Z_i^*(u, a) - \tilde{\bar{Z}}_n^*(\gamma; u, a) \right\} d\tilde{N}_i^*(u), \tag{9}$$

where $\tilde{\bar{Z}}_n^*(\gamma; u, a) = \tilde{S}_n^{(1)}(\gamma; u, a)/\tilde{S}_n^{(0)}(\gamma; u, a)$ with $n\tilde{S}_n^{(q)}(\gamma; u, a) = \sum_{i=1}^{n} \tilde{Y}_i(u) Z_i^*(u, a)^{\otimes q} \exp\{\gamma' Z_i^*(u, a)\}$ for $q = 0, 1, 2$. The limit of $\tilde{\bar{Z}}_n^*(\gamma; u, a)$ is $\bar{z}(\gamma; u, a)$, the same as the limit of $\bar{Z}_n^*(\gamma; u, a)$ in Section 3.1. Plugging $\bar{z}(\gamma; u, a)$ in (9) gives a slight variation of $E_n(\gamma; a)$, denoted by $\tilde{E}_n(\gamma; u|\mathbf{B})$, which is the projection of $\tilde{U}_n(\gamma; u|\mathbf{B})$ in Section 3.1 to the space of the available data.

Clearly $E_n(\gamma; a)$ in (9) is evaluable with the PMHC dataset. In fact, it is asymptotically unbiased since both $E\{d\tilde{N}_i^*(u)|Z_i\}$ and $E\{\tilde{Y}_i(u)|Z_i\} \exp(\beta(u)' Z_i) d\Lambda_0(a)$ are $E\{dN_i^*(u)|Z_i\} = E\{Y(u|B_i) dN_i(u)|Z_i\}$. Denote by $\hat{\beta}_n(a)$ the estimator of $\beta(a)$ derived from the solution of $E_n(\gamma; a) = 0$ for $a \in [\tau_L, \tau_R]$. We outline a proof in Web Appendix A.2 for the pointwise consistency and asymptotic normality of the estimator $\hat{\beta}_n(\cdot)$. Note that the difference of $d\tilde{N}_i^*(a)$ and $\tilde{Y}_i(a) \exp(\beta(a)' Z_i) d\Lambda_0(a)$ does not define a martingale process with respect to the natural filtration. Our asymptotics derivation adapts the arguments of Cai and Sun (2003) by applying the functional central limit theorems (Pollard, 1990).

Moreover, projecting the estimating equation (7) to the space of the PMHC dataset yields an estimator of $\Lambda_0(a)$ with

Table 1
Demographic variable summary for the PMHC dataset

	Sex		Age (in years)			Socio-economic proxy (pSES)				Residence	
						Human					
	Female	Male	$0-5$	$6-12$	$13-17$	Other	Aboriginal	Services	Welfare	Urban	Rural
subjects[a] ($n = 27,947$)	15852	12095	579	3938	23430	18260	3785	4098	1804	21293	6654
visits ($n = 41{,}159$)	24160	16999	603	5105	35451	25391	5856	7003	2909	31524	9635

[a] based on the information collected at all the subjects' first MHED visits in the PMHC dataset.

fixed $\beta(\cdot)$:

$$\hat{\Lambda}_{0n}(a;\beta(\cdot)) = \int_0^a \frac{\sum_{i=1}^n d\tilde{N}_i^*(u)}{\sum_{l=1}^n \tilde{Y}_l(u)\exp\{\beta(u)'Z_l\}}, \quad a \in (0,18).$$

We may obtain by the kernel smoothing an estimator of $\lambda_0(\cdot)$ based on $\hat{\Lambda}_{0n}(a;\beta(\cdot))$ and $\hat{\beta}_n(\cdot)$ obtained above as

$$\tilde{\lambda}_{0n}(a) = \int_0^{18} K_g^*(u-a)\hat{\Lambda}_{0n}(u;\hat{\beta}_n(u))du, \quad a \in (0,18),$$

where $K_g^*(\cdot) = K^*(\cdot/g)/g$ with $K^*(\cdot)$ a kernel function, and $\hat{\beta}_n(a) = \hat{\beta}_n(\tau_L)$ for $0 < a < \tau_L$ and $\hat{\beta}_n(a) = \hat{\beta}_n(\tau_U)$ for $\tau_U < a < 18$.

Computing the estimator $\hat{\beta}_n(\cdot)$ requires expectations calculated with respect to the missing birthdates B_i's. The expectations can be well-approximated by the corresponding admissible sample means with a large number of independent sets of generated B_i's from the distributions of $G_i(\cdot)$. Moreover, with given B_i's, there are available computing codes/software packages to evaluate the estimator $\hat{\beta}_n(\cdot|\mathbf{B})$ in Section 3.2.1. These considerations lead to the following two easy-to-implement procedures for estimating $\beta(\cdot)$.

Conditional on the available data, generate W independent sets of $\mathbf{B} = \{B_i : i = 1,\ldots,n\}$ with $B_i \sim G_i(\cdot) = \text{Unif}(I_i)$ independent with each other, denoted by $\mathbf{B}^{(w)}$, $w = 1,\ldots,W$.

Procedure A. Approximate the estimating function $E_n(\gamma;u)$ in (9) by $\bar{U}_{n,W}(\gamma;a) = \sum_{w=1}^W U_n(\gamma;a \mid \mathbf{B}^{(w)})/W$, and solve the equation $\bar{U}_{n,W}(\gamma;a) = 0$ to attain an estimator of $\beta(a)$ for $a \in [\tau_L,\tau_U]$. Denote the derived estimator by $\tilde{\beta}_{nW}(\cdot)$.

Procedure B. Obtain $\hat{\beta}_n(\cdot|\mathbf{B}^{(w)})$ by solving $U_n(\gamma;a \mid \mathbf{B}^{(w)}) = 0$ as in (6), and estimate $\beta(\cdot)$ with the sample mean $\bar{\tilde{\beta}}_{nW}(\cdot) = \sum_{w=1}^W \hat{\beta}_n(\cdot|\mathbf{B}^{(w)})/W$.

We justify in Web Appendix A.3 that the two alternative estimators $\tilde{\beta}_{nW}(\cdot)$ and $\bar{\tilde{\beta}}_{nW}(\cdot)$ are asymptotically equivalent to $\hat{\beta}_n(\cdot)$ as $W \to \infty$ and $n \to \infty$. Procedure A needs to implement only once a local estimator with the augmented data using W sets of generated B_i's. It adapts the idea of the Monte Carlo implementation of the EM algorithm (Wei and Tanner, 1990) in the estimation procedure proposed in this section, a type of the expectation-solution algorithm (Rosen et al., 2000). Procedure B is similar to one of the two proposed approaches in Schaubel and Zhang (2010), applying the idea of multiple imputation (Rubin, 1987). It requires implementing the local estimator with each of the W generated B_i's sets. Procedure B is expected to take more computing time than Procedure A. On the other hand, since the resulting estimates by Procedure B are sample means of W realizations of the local estimator, we anticipate the estimate curves by Procedure B are more smooth than the ones by Procedure A.

4. Analysis of the PMHC Dataset

Almost 75% individuals had only one visit record in the aforementioned PMHC dataset; 15.2%, two visits; 5.4%, three visits; 4.7%, more than three visits. Table 1 summarizes the demographic information both according to the individual subjects and to the individual MHED visits in the PMHC dataset. The nonnegligible recurrence of MHED visits guided us to formulate them as recurrent events.

Three risk factors/exposures were considered as the external covariates: *sex* (male vs. female), *pSES* (socio-economic proxy: Aboriginal/human service recipient/welfare vs. others), and *residence* (indicator of residence region: rural vs. urban). The PMHC data show that the variables of *pSES* and *residence* are rather stable over time within an individual subject (Wang, 2014). Thus, we assumed all three covariates are time-independent and used the covariate information at each individual's first MHED visit in the PMHC dataset.

We started with an analysis under the proportional rates/means model with time-independent coefficients (Lin et al., 2000). Table 2 presents three sets of estimates and standard errors for the coefficients, using one set of generated birthdates and by Procedures A and B. The three sets of estimates are very similar to each other. This analysis confirms the significance of all the three factors in MHED visit.

We then conducted an analysis under the marginal regression model with time-varying coefficients as given in (1). The approach in Section 3.2.2 was applied using the "Epanechnikov kernel" function $K(u) = 3(1-u^2)/4, -1 \le u \le 1$. We used two months as a time unit, set the bandwidth to be $h = 3$ units, and set the window of age to be $\tau_L = 6$ to $\tau_R = 102$ time units (i.e., one to 17 years of age) to evaluate the estimators for the regression parameter function $\beta(a)$ of the model in (1) and then the estimators for the baseline cumulative rate function. Each of the local constant and local linear estimators was implemented for one set of generated birthdates, and for both Procedures A and B in Section 3.2.2.

Figures 1 and 2 present the estimates by Procedures A and B with $W = 100$ together with approximate 95% pointwise confidence intervals, respectively. The variance estimates were obtained using the consistent variance estimators given

Table 2
Coefficient estimates and standard error estimates with the proportional rates model

	Procedure with generated B_i's		Procedure A $W = 100$		Procedure B $W = 100$	
Covariate	$\hat{\beta}(\mathbf{B})$	$SE(\hat{\beta}(\mathbf{B}))$	$\tilde{\beta}$	$SE(\tilde{\beta})$	$\bar{\tilde{\beta}}$	$SE(\bar{\tilde{\beta}})$
sex (male vs. female)	−.0544	.0156	−.0556	.0156	−.0540	.0155
pSES (government sponsored vs. other)	.1563	.0173	.1556	.0173	.1564	.0173
residence (rural vs. urban)	−.0483	.0185	−.0495	.0185	−.0490	.0185

in Web Appendix A.3. The left panel in each figure presents the local constant estimates; the right panel, the local linear estimates. The local constant estimates are very similar to the local linear estimates, and the corresponding estimates in Figures 1 and 2 agree with each other strongly. The estimated curves by Procedure B are expectedly more smooth than the ones by Procedure A. Our analysis confirmed that Procedure B requires more time than Procedure A.

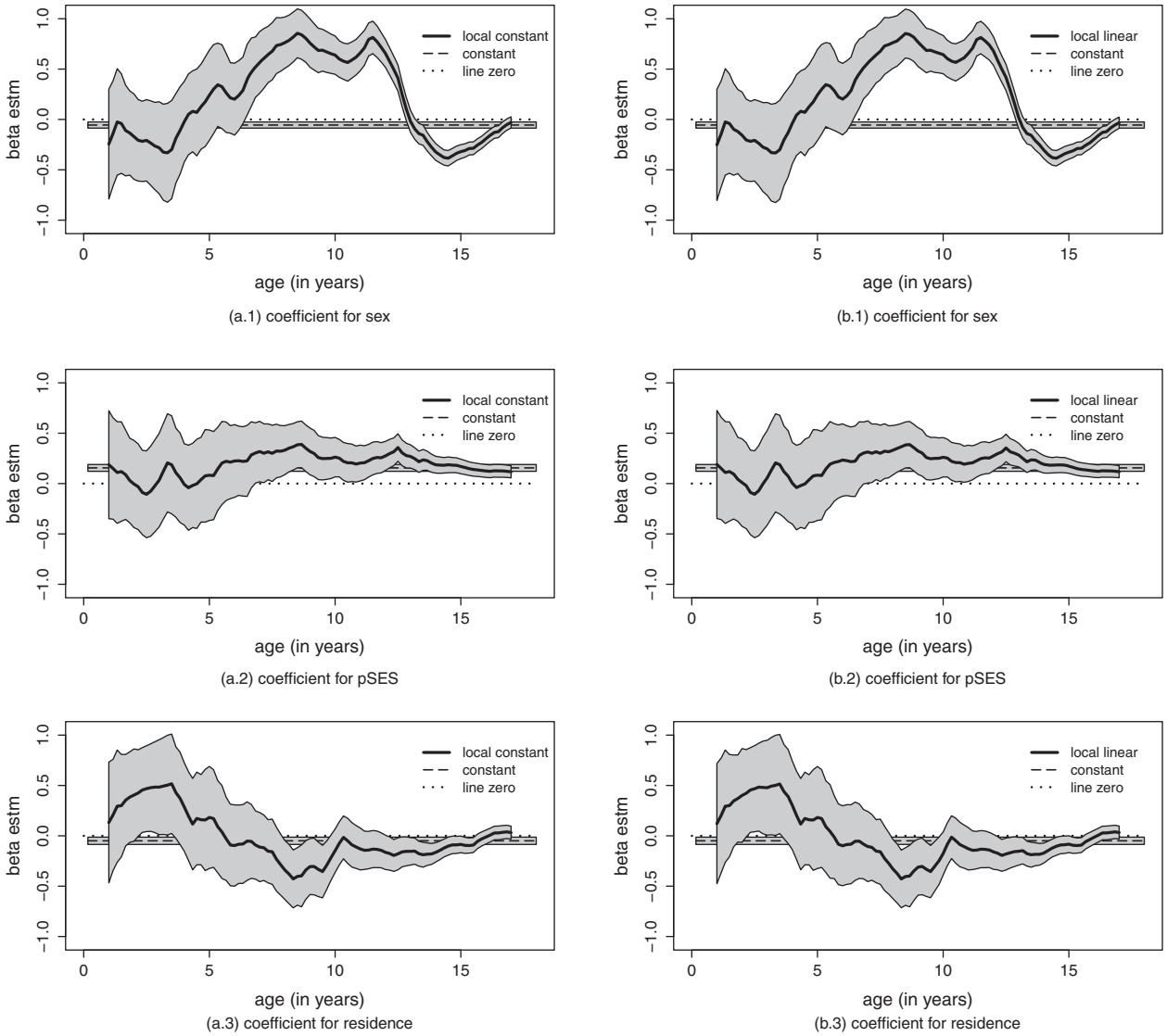

Figure 1. Regression parameter estimates with PMHC data: the thick solid curves are the regression estimates under the marginal model (1) by Procedure A of Section 3.2.2, the local constant/linear estimates in the left/right panels, and shaded with the limits of approximate 95% piecewise confidence intervals; the dashed lines, under the proportional rates model.

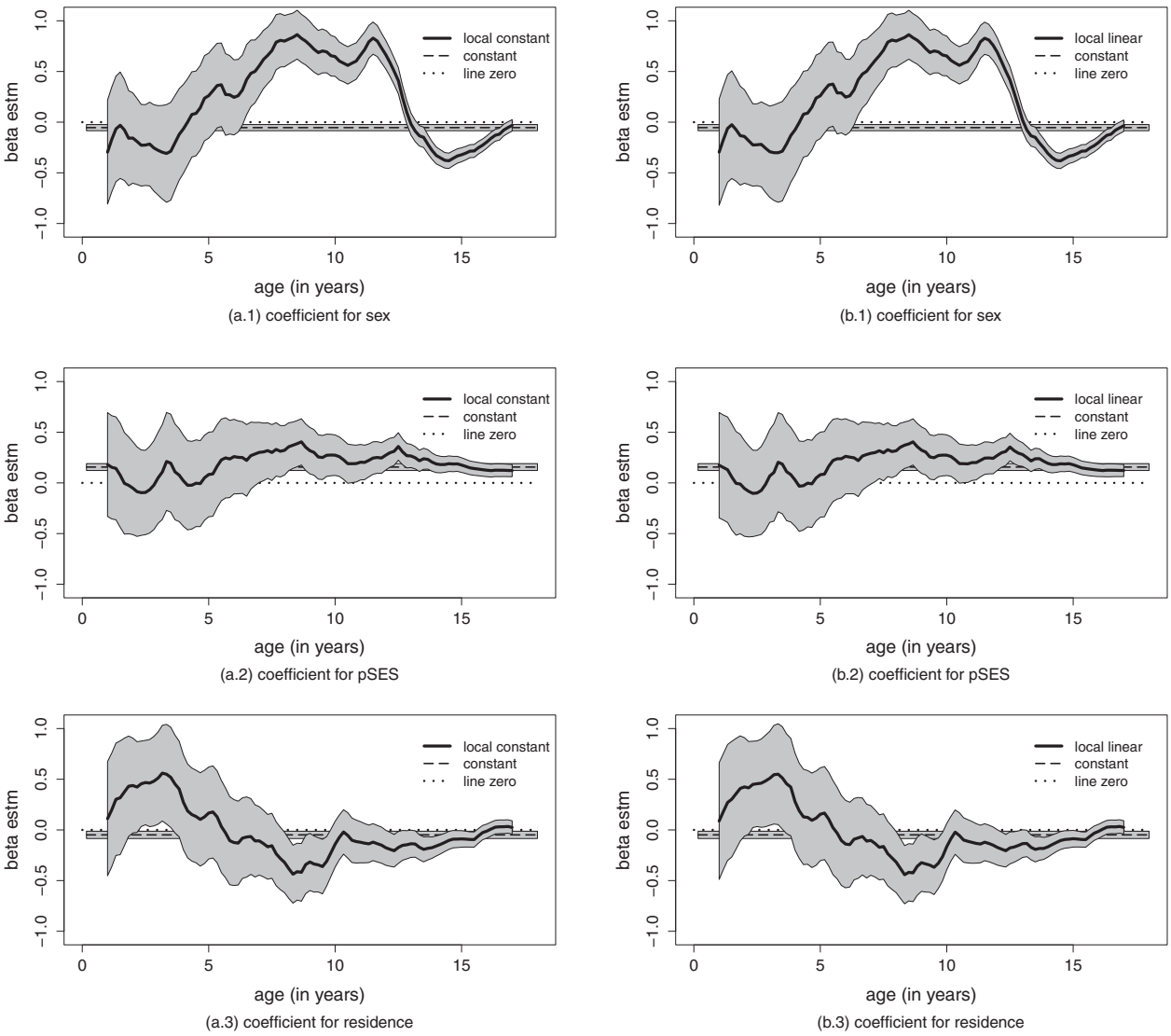

Figure 2. Regression parameter estimates with PMHC data: the thick solid curves are the regression estimates under marginal model (1) by Procedure B of Section 3.2.2, the local constant/linear estimates in the left/right panels, and shaded with the limits of approximate 95% piecewise confidence intervals; the dashed lines, under the proportional rates model.

For comparison, we add in each plot the zero line, and the estimate and 95% confidence intervals for the time-independent regression coefficient in the proportional rates model from Table 2. The proportional rates model allows the estimation of the regression parameters by pooling information throughout the whole interval of 0–17 years of age, and thus leads to very narrow confidence intervals. The pointwise confidence intervals associated with the local estimate curves are rather wide for younger ages and become narrower for older ages. This feature may be explained by the counts of ED visits from different age groups. The regression estimates with the proportional rates model appear to be the overall averages of the corresponding estimate curves.

The curve estimates interestingly reveal age-varying effects of all the three factors. For example, the four sets of local estimate curves for the coefficient to sex in our marginal regression model show the following: (i) boys and girls have similar ED visit frequencies before starting school; (ii) younger, school aged boys tend to have significantly more ED visits than girls; and (iii) teenage girls have significantly higher ED visit frequencies than teenage boys. These findings are in agreement with the general understanding of sex differences in ED presentations for mental health reasons. As another example, the estimates for the pSES-related effect over time indicate significantly higher frequencies of ED visits from children of families supported by social programs during school ages. Moreover, there appears a decreasing trend in the ED visit frequency associated with children from rural regions in early school ages. This leads to the significantly higher frequency in the rural group before school ages to become rather comparable with the frequency in the urban group during the teenage years.

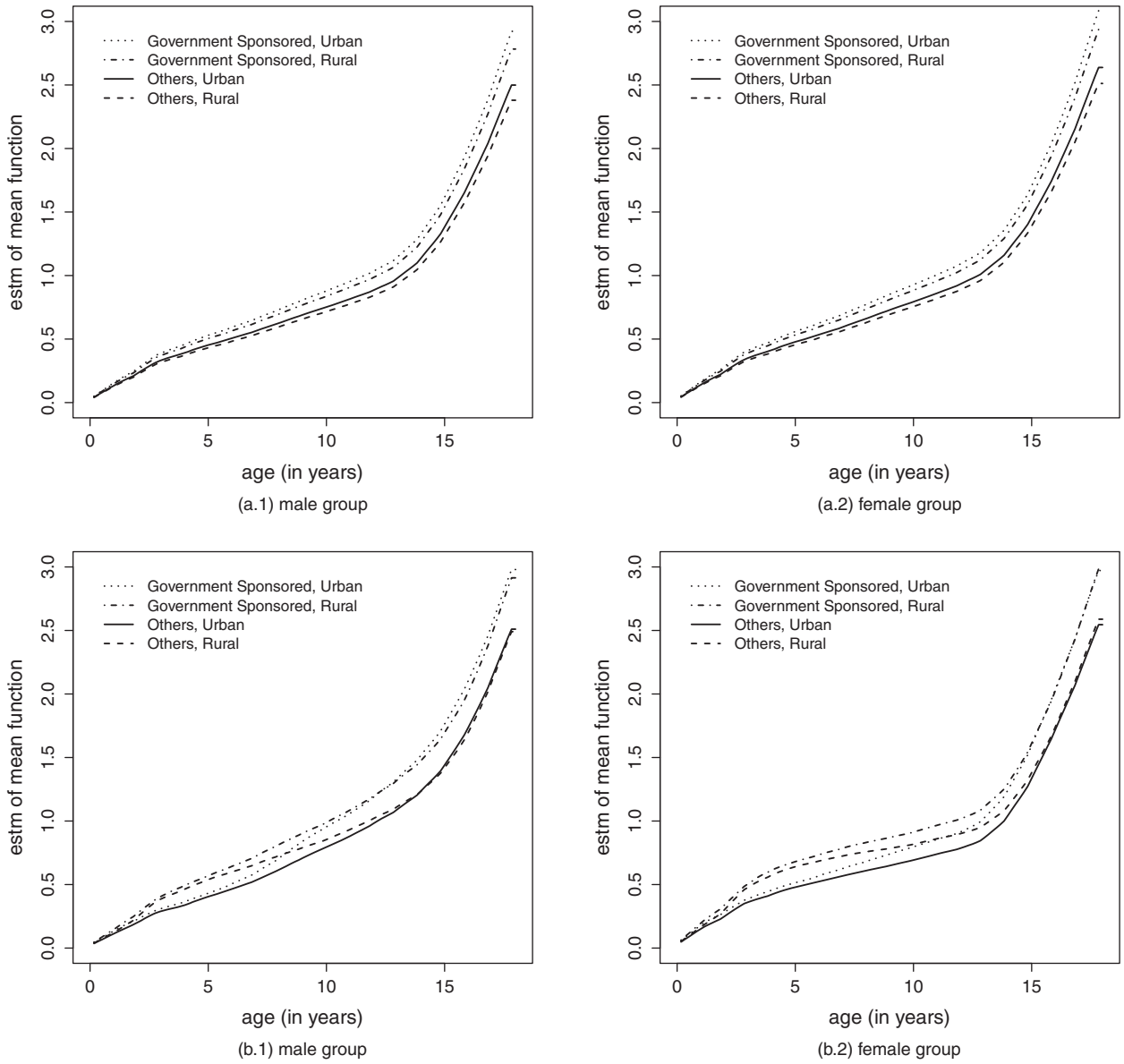

Figure 3. Mean estimates with PMHC data: plots of the top row present the local linear estimates under the marginal model (1) by Procedure B of Section 3.2.2; the bottom row, under the proportional rates model.

We combined the regression parameter estimates and the estimates of the baseline function with the local constant/linear estimators by Procedures A and B, and obtained estimates for the cumulative rate (mean) functions of eight groups, determined by the three binary factors: sex, pSES, and region. The four sets of mean estimates are very similar to each other. As an example, we present the rate estimate curves based on the local linear estimator by Procedure B in Figure 3 together with the rate estimates under the proportional rates model for comparison. All of the estimate curves show distinct age eras. At about 13 years of age, the mean (cumulative rate) estimates increase sharply overall, especially in the female groups. This finding would agree with empirical evidence. Adolescents are more frequent ED visitors for mental health reasons than younger children. We also know from earlier work that older female youth have higher numbers of ED visits than males (Newton et al., 2011). The curves by the proposed approach cross over with each other and indicate a non-proportional relationship between the ED visit frequencies of the different groups, different from what the proportional rates model assumes.

We varied the time unit in the analysis to compare the resulting estimates of the time-varying covariate effects. Web Figures 1 and 2 in Web Appendix B display the estimated curves by the local constant and local linear approaches under Procedures A and B with the time unit being 1, 6, and 12 months in addition to 2 months. A longer time unit leads to smoother estimates and provides less detail on the shape of

the curve. One may apply the commonly used plug-in or cross-validation selection method in kernel smoothing to determine the bandwidth corresponding to the chosen time unit.

A small simulation was conducted to examine how well our approach handles the missing birthdates. We considered the model (1) with one covariate (sex), took one set of generated birthdates as the real birthdates, and evaluated the local constant and local linear estimators of the regression parameter given in Section 3.2.1. The curve estimates are presented in Web Figure 3 together with the realizations of the local constant and local linear estimators in Section 3.2.2 by Procedures A and B with $W = 100$. Both procedures yielded estimates that captured the general shape of the curve estimate by Section 3.2.1. Further details are presented in Web Appendix B.

5. Discussion

This article proposes an approach to analyzing recurrent event data extracted from an administrative database. Motivated by the PMHC dataset, we address missing start times (e.g., birthdates) of the underlying counting processes, which results in missing censoring times and coarsening information of the recurrent events. We adapt the local linear/constant estimation procedures with survival times to evaluate covariate effects over time. The application of the approach with the PMHC data verifies earlier findings of the PMHC study, and provides new insights into pediatric mental health care in general. Strictly speaking, the target population would be any individual with an MHED visit before the age of 18 and the study population is comprised of Alberta residents with an MHED visit in Alberta before the age of 18. Our findings from the data analysis in this article may apply to the target population, since the individuals in the PMHC dataset (during April 1, 2002, to March 31, 2011) can reasonably be taken as a representative sample of the population.

The proposed approach is rather flexible. We can adapt it to accommodate time-dependent covariates, and to explore seasonal and spatial patterns in the ED visits. Although the approach assumes the study subjects are independent, we may modify it to accommodate subjects geographically clustered. Moreover, we may develop a procedure based on the proposed estimator and its weak convergence for testing whether the effect of a covariate differs significantly from a constant.

There are other practical issues worthy of further investigation. For example, we assume the study population is closed. In fact, study subjects may enter into and exit out of the population because of relocation, hospitalization, or death, say. If not a resident or in hospital, the subject is not eligible to have a recorded ED visit. This feature results in "later entry," "earlier exit," or "gapped observation times" associated with the study subjects and violates the assumption. Not accounting for it could bias the inference. As another example, generations turn over every 10–15 years and the pattern of ED visits may change with the generation. This aspect may yield informative censoring in the extracted information. Moreover, one may wish to draw conclusions about all of Alberta's children and youth, and that would require data beyond the individuals with ED visit records.

6. Supplementary Materials

Web Appendices referenced in Sections 3.2 and 4 are available with this article at the *Biometrics* website on Wiley Online Library. Preliminary R/C++ code is also available at the website.

Acknowledgements

The study was funded by an operating grant from the Canadian Institutes of Health Research (CIHR, Ottawa, Canada) and research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). Professor Rosychuk is salary supported by Alberta Innovates-Health Solutions (AI-HS, Edmonton, Canada; formerly the Alberta Heritage Foundation for Medical Research) as a Health Scholar. We thank Jueyu Gao for his assistance in computer code development and an AE and two referees for their careful review and constructive comments and suggestions.

Disclaimer: This article is based in part on data provided by Alberta Health. The interpretation and conclusions are contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta. Neither the government nor Alberta Health expresses any opinion in relation to this study.

References

Alberta Health and Wellness (2004). *Ambulatory Care in Alberta Using Ambulatory Care Classification System Data*, Edmonton: Alberta.

Amorim, L. D., Cai, J., Zeng, D., and Barreto, M. L. (2008). Regression splines in the time-dependent coefficient rates model for recurrent event data. *Statistics in Medicine* **28**, 5890–5906.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1100–1120.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1992). *Statistical Models Based on Counting Processes*, New York: Springer.

Bilias, Y., Gu, M., and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics* **25**, 662–682.

Cai, T. and Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika* **91**, 277–290.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics* **30**, 93–111.

Chen, K., Lin, H. and Zhou, Y. (2012). Efficient estimation for the Cox model with varying coefficients. *Biometrika* **99**, 379–392.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*, New York: Springer.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Gomez, G. and Lagakos, S. W. (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics* **50**, 204–212.

Halamandaris, P. V. and Anderson, T. R. (1999). Children and adolescents in the psychiatric emergency setting. *The Psychiatric Clinics of North America* **22**, 866–875.

Hu, X. J. and Lawless, J. F. (1997). Pseudolikelihood estimation in a class of problems with response-related missing covariates. *Canadian Journal of Statistics* **25**, 125–142.

Hu, X. J., Lawless, J. F. and Suzuki, K. (1998). Nonparametric estimation of a lifetime distribution when censoring times are missing. *Technometrics* **40**, 3–13.

Hu, X. J., Sun, J. and Wei, L. J. (2003). Regression analysis of panel count data. *Scandinavian Journal of Statistics* **25**, 25–43.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**, 887–906.

Kim, M. Y., De Gruttola, V. G. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49**, 13–22.

Leitch, K. K. (2007). *Reaching for the top: A report by the advisor on healthy children and youth*. Technical report. Ottawa, Ontario: Health Canada.

Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158–168.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62**, 711–730.

Murphy, S. and Sen, P. (1991). Time-dependent coefficients in a Cox type regression model. *Stochastic Processes and Their Applications* **39**, 153–180.

Nan, B., Lin, X., Lisabeth, L. D. and Harlow, S. D. (2005). A varying-coefficient Cox model for the effect of age at a marker event on age at menopause. *Biometrics* **61**, 576–583.

Newton, A. S., Rosychuk, R. J., Ali, S., Cawthorpe, D., Curran, J., Dong, K., et al. (2011). *The Emergency Department Compass: Children's Mental Health. (Pediatric mental health emergencies in Alberta, Canada: Emergency department visits by children and youth aged 0 to 17 years, 2002-2008)*. Edmonton, Alberta.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of American Statistical Association* **88**, 811–820.

Pollard, D. (1990). *Empirical processes: Theory and applications. Regional conference series in probability and statistics 2*. Institute of Mathematical Statistics, Hayward, CA.

Rosen, O., Jiang, W. and Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika* **87**, 391–404.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schaubel, D. E. and Zhang, M. (2010). Estimating treatment effects on the marginal recurrent event mean in the presence of a terminating event. *Lifetime Data Analysis* **16**, 451–477.

Shorack, G. R. and Wellner, J. A. (1990). *Empirical processes with applications to statistics, Wiley Series in Probability and Mathematical Statistics*. New York: Wiley.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. New York: Springer.

Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*, New York: Springer.

Tian, L., Zucker, D. M. and Wei, L. J. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association* **100**, 172–183.

Wang, F. (2014). *Exploring Mental Health Related Emergency Department Visits: Frequency of Recurrence and Risk Factors*. MSc thesis, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada.

Wang, J. (2011). Estimation of lifetime distribution with missing censoring. *Journal of Data Science* **9**, 331–343.

Zhang, C. H. and Li, X. (1996). Linear regression with doubly censored data. *The Annals of Statistics* **24**, 2720–2743.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

Zhao, L. P., Lipsitz, S. and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165–1182.

Zucker, D. and Karr, A. (1990). Nonparametric survival analysis with time dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics* **18**, 329–353.

Received March 2015. *Revised January* 2016.
Accepted January 2016.