Semiparametric Regression of Multivariate Panel Count Data with Informative Observation Times

Yang Li[1], Xin He[2], Haiying Wang[3], Bin Zhang[4] and Jianguo Sun[5]

University of North Carolina-Charlotte[1]
University of Maryland[2]
University of New Hampshire[3]
Cincinnati Children's Hospital[4]
University of Missouri[5]

Abstract

Multivariate panel count data occur in many fields such as medical and social science studies in which several outcomes of interest are measured simultaneously and repeatedly over time. When the observation times are not pre-specified, it is very likely that either the observation or follow-up times are informative about the response process. In such situations, most existing approaches either specify a dependence structure with some fixed distributions or assume independence given some covariates, which may not be true and result in misleading conclusions. In this paper, we present a joint modeling approach that allows the possible mutual correlations to be characterized by time-dependent random effects. Estimating equations are developed for the parameter estimation and the resulted estimators are shown to be consistent and asymptotically normal. The finite sample performance of the proposed estimators is assessed through a simulation study and an illustrative example from a maternal influenza immunization study on infant growth is provided.

Keywords: Estimating equation; Informative censoring; Informative observation process; Multivariate panel count data.

1 Introduction

Multivariate panel count data occur in many fields such as medical and social science studies in which several related recurrent events are of interest but the responses are recorded only at discrete times. In such situations, only the numbers of events that have occurred between the observation times are known, but the exact event occurrence times are not observable. The observed data consist of two parts, one being a sequence of discrete observation times which can be regarded as realizations from an observation process and the other being the sequences of counts that the events have occurred between discrete observation times.

One example of multivariate panel count data that motivated this research is the Mother's Gift study, a prospective, controlled, double blinded, randomized trial to assess the safety and immunogenicity in pregnant women of pneumococcal vaccines, as well as the clinical effectiveness of influenza vaccine in Bangladesh (Zaman et al., 2008). Research studies have shown that respiratory illness has significant negative impact on children's weight and height gains in Bangladesh (Torres et al., 2000). Therefore, reducing children's respiratory illness rates has become clinically important in enhancing children's growth in countries without sufficient food and health resources. One objective of the study is to evaluate the effectiveness of infant Pneumococcal conjugate vaccine (PCV7) in reducing the recurrences of febrile respiratory illness, pneumonia and difficulty with breathing. In addition to usual routine vaccines, all infants received either PCV7 (treatment) or Hib conjugate vaccine (comparator) at 6, 10 and 14 weeks of age. The experiences of febrile respiratory illness, pneumonia and difficulty with breathing of the infants were scheduled to be recorded weekly from 6 to 24 weeks of age, but as expected, the actual observation times varied among the infants. In particular, many infants were examined more frequently when the respiratory illness occurred and skipped a couple of clinical visits when they were healthy. Therefore, the actual observation times could be informative about the recurrence processes of respiratory illness and multivariate panel count data were generated as described above. In the original analysis of the study, the chi-square test and Fisher's exact test were used to compare the laboratory-confirmed influenza occurrence rates, defined as the total number of infant influenza cases divided by the total number of infants from each treatment group. It is easy to see that this is clearly not efficient.

For the analysis of multivariate panel count data, it is apparent that one may separately carry out univariate analysis for each type of outcomes by applying an existing procedure (Huang et al., 2006; Sun et al., 2007; Zhao et al., 2011a; Sun and Zhao, 2013; Zhao et al., 2013b). However, such practice ignores the mutual correlation between the related outcomes and would be less efficient than a joint or multivariate analysis. To overcome this, several authors have proposed multivariate regression models. For example, Chen et al. (2005) and He et al. (2008) considered parametric and semiparametric methods, respectively, by assuming that the response process and the observation process are independent of each other. Li et al. (2011) and Zhao et al. (2013a) proposed some marginal model-based approaches that allow the dependent observation processes. However, their models imply that the subjects with the same observation schedule are expected to have the same response rates for all types of recurrent events, which clearly may not be realistic in some applications. For instance, despite of observation times in the past, the proceeding observation times and longitudinal responses may both depend on individuals' current stages of disease progression. Also such correlation may vary over time and relate to the follow-up times.

In the following, we discuss regression analysis of multivariate panel count data when the response process, the observation process and the follow-up time may be mutually correlated. An easy-to-implement estimation approach is proposed and the asymptotic properties of the resulted estimators are established. The approach allows for time-dependent, arbitrary correlations. Before presenting the estimation approach, we will first introduce the notation and present the model in Section 2. Section 3 presents the estimation procedure and establishes the asymptotic properties of the resulted estimators, and Section 4 gives a procedure for model diagnostics. In Section 5, a simulation study is conducted to evaluate the finite-sample performance of the proposed estimators, and Section 6 applies the method to the data from the maternal influenza immunization study described above. The paper concludes in Section 7 with some discussion and remarks.

2 Notation and Models

Consider a recurrent event study that involves p types of events. For subject i, let $Y_{ik}(t)$ denote the total number of type-k events that have occurred up to time t, $i = 1, \cdots, n$, $k = 1, \ldots, p$. Suppose that $\mathbf{Y}_i(t) = \big(Y_{i1}(t), Y_{i2}(t), \ldots, Y_{ip}(t)\big)'$ is observed only at discrete time points $\{T_{i,1}, \cdots, T_{i,m_i}\}$, where m_i represents the total number of observations on subject i. Let $N_i(t)$ represent the observation process, which gives the cumulative numbers of observation times up to time t. In practice, there usually exists a censoring or follow-up time C_i and one observes $\widetilde{N}_i(t) = N_i(t \wedge C_i)$, where $a \wedge b = min(a, b)$. Let $\mathbf{Z_i}(t)$ denote a d-dimensional vector of covariates which is assumed to be continuously traceable in the study and denote $\mathcal{Z}_{it} = \{\mathbf{Z_i}(s), s \leq t\}$.

In the following, we assume that there exists an unobserved random vector $\mathbf{b}_i(t) = \big(b_{i1}(t), \ldots, b_{ip}(t), b_{i,p+1}(t), b_{i,p+2}(t)\big)'$ that will be used to model the correlation between $\mathbf{Y}_i(t)$, $N_i(t)$ and C_i. Define $\mathcal{B}_{it} = \{\mathbf{b}_i(s), s \leq t\}$ and assume that the $\mathbf{b}_i(t)$'s are independent and identically distributed, $\mathcal{B}_{it}$ is independent of $\mathcal{Z}_{it}$, and given $\mathcal{Z}_{it}$ and $\mathcal{B}_{it}$, $\mathbf{Y}_i(t)$, $N_i(t)$ and C_i are mutually independent. Also we will assume that the mean function of $Y_{ik}(t)$ has the form

$$E\{Y_{ik}(t)|\mathbf{Z_i}(t), \mathbf{b}_i(t)\} = \Lambda_{0k}(t) \exp\{\beta'\mathbf{Z_i}(t) + b_{ik}(t)\}, \; k = 1, \ldots, p, \tag{1}$$

where β denotes a vector of d-dimensional regression coefficients and $\Lambda_{0k}(t)$ is an unknown baseline mean function. Note that for the simplicity of presentation, we have assumed that regression parameters β are the same for different types of event, and it is straightforward to generalize the methodology proposed below to the situation where covariate effects may be different on different types of recurrent events.

For the observation process $N_i(t)$, it will be assumed that it follows the following marginal rate model

$$E\{dN_i(t)|\mathbf{Z_i}(t), \mathbf{b}_i(t)\} = \exp\{\gamma'\mathbf{Z_i}(t) + b_{i,p+1}(t)\}d\mu_0(t)\,, \tag{2}$$

where γ is a vector of unknown regression parameters and $d\mu_0(t)$ is an unknown baseline rate function. For the censoring time C_i, we will suppose that its hazard function is given by

$$\lambda_i(t|\mathbf{Z_i}(t), \mathbf{b}_i(t)) = \lambda_0(t) + \xi'\mathbf{Z_i}(t) + b_{i,p+2}(t). \tag{3}$$

Here ξ denotes the effect of covariates on the hazard function of $C_i's$ and $\lambda_0(t)$ is an unknown baseline hazard function. That is, the C_i's follow the additive hazards models (Lin et al., 1998; Kalbfleisch and Prentice, 2002; Zhang et al., 2005).

Note that models (1) - (3) can be viewed as natural generalizations of some existing models that have been commonly used in the literature. For example, when any of the $b_{il}(t)$'s is zero or independent of other $b_{ij}(t)$'s ($l, \; j = 1, 2, \ldots, p+2$ and $l \neq j$), the corresponding process is conditionally independent from others. In particular, if $b_{ik}(t) = 0$ ($k = 1, \ldots, p$), model (1) is equivalent to the proportional means model considered in Cheng and Wei (2000), Sun and Wei (2000), Zhang (2002), and Hu et al. (2003) among others. When both $b_{ik}(t)$ and $\mathbf{Z_i}(t)$ are time-independent, model (1) is equivalent to the model (1) discussed in Zhang et al. (2013). That is, the proposed joint modeling procedure also applies to special cases when either the observation or censoring times are noninformative or the longitudinal responses have a time-independent correlation structure. In general, since the form or distribution of $\mathbf{b}_i(t)$ is arbitrary and completely unspecified, the joint modeling procedure described above is quite flexible compared to many existing procedures. In contrast, the application of the existing procedures that assume independence to the situation considered here could give biased results or misleading conclusions.

3 Inference Procedure

In this section, we will present an inference procedure for estimation of β which is usually of the primary interest. For this, first note that the counting process $\widetilde{N}_i(t) = N_i(t \wedge C_i)$ jumps by one at time t if and only if $t \le C_i$ and $dN_i(t) = 1$. Based on the conditional independence assumption for $\mathbf{Y}_i(t)$, $N_i(t)$ and C_i, it can be shown that

$$\begin{aligned}
& E\{Y_{ik}(t)d\widetilde{N}_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\} \\
= & E\{I(t \le C_i)Y_{ik}(t)dN_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\} \\
= & E\{I(t \le C_i)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}E\{Y_{ik}(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}E\{dN_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\} \\
= & \exp\{-\Lambda_0^*(t) - B_i(t) - \xi'\mathbf{Z}_\mathbf{i}^*(t)\} \times \\
& \Lambda_{0k}(t)\exp\{\beta'\mathbf{Z}_\mathbf{i}(t) + b_{ik}(t)\}\exp\{\gamma'\mathbf{Z}_\mathbf{i}(t) + b_{i,p+1}(t)\}d\mu_0(t) \\
= & \exp\{(\beta+\gamma)'\mathbf{Z}_\mathbf{i}(t) - \xi'\mathbf{Z}_\mathbf{i}^*(t)\} \times \\
& \exp\{-\Lambda_0^*(t) + b_{ik}(t) + b_{i,p+1}(t) - B_i(t)\}\Lambda_{0k}(t)d\mu_0(t),
\end{aligned} \tag{4}$$

where

$$\mathbf{Z}_\mathbf{i}^*(t) = \int_0^t \mathbf{Z}_\mathbf{i}(s)ds, \;\; \Lambda_0^*(t) = \int_0^t \lambda_0(s)ds, \;\; B_i(t) = \int_0^t b_{i,p+2}(s)ds\,.$$

Thus we have

$$E\{Y_{ik}(t)d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} = \exp\{\beta'\mathbf{Z}_\mathbf{i}(t) + \eta'\mathbf{X}_\mathbf{i}(t)\}d\Lambda_{2k}^*(t)\,,$$

where $\eta = (\gamma', \;\; \xi')'$, $\mathbf{X}_\mathbf{i}(t) = (\mathbf{Z}_\mathbf{i}'(t), \;\; -\mathbf{Z}_\mathbf{i}'^*(t))'$ and

$$d\Lambda_{2k}^*(t) = \exp\{-\Lambda_0^*(t)\}\Lambda_{0k}(t)E[\exp\{b_{ik}(t) + b_{i,p+1}(t) - B_i(t)\}]d\mu_0(t).$$

Define

$$dM_{ik}(t;\beta,\eta) = Y_{ik}(t)d\widetilde{N}_i(t) - \exp\{\beta'\mathbf{Z}_\mathbf{i}(t) + \eta'\mathbf{X}_\mathbf{i}(\mathbf{t})\}d\Lambda_{2k}^*(t)$$

and $dM_{ik}(t) = dM_{ik}(t;\beta_0,\eta_0)$, where β_0 and η_0 denote the true values of β and η, respectively. Then $M_{ik}(t)$ is a mean-zero stochastic process. For estimation of β and $d\Lambda_{2k}^*(t)$, if η is known, this naturally suggests the following estimating equations

$$\sum_{i=1}^n \left[Y_{ik}(t)d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z}_\mathbf{i}(t)+\eta'\mathbf{X}_\mathbf{i}(t)}d\Lambda_{2k}^*(t)\right] = 0, \;\; 0 \le t \le \tau, \tag{5}$$

and

$$U_\beta(\beta;\eta) = \sum_{i=1}^n\sum_{k=1}^p \int_0^\tau W(t)\mathbf{Z}_\mathbf{i}(t)\left[Y_{ik}(t)d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z}_\mathbf{i}(t)+\eta'\mathbf{X}_\mathbf{i}(t)}d\Lambda_{2k}^*(t)\right] = 0, \tag{6}$$

where $W(t)$ is some possibly data-dependent function. Some simple and natural choices for $W(t)$ are $W(t) = 1$, constant for all t and $W(t) = \sum_{i=1}^n I(t \le C_i)/n$, proportional to the number of subjects under observation.

In reality, of course, η is unknown. However, it can be readily estimated based on the recurrent event data on $\widetilde{N}_i(t)$'s. Specifically, define

$$dM_i^*(t;\eta) = d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_\mathbf{i}(t)}d\Lambda_1^*(t),$$

where

$$d\Lambda_1^*(t) = \exp\{-\Lambda_0^*(t)\}E[\exp\{b_{i,p+1}(t) - B_i(t)\}]d\mu_0(t),$$

and $dM_i^*(t) = dM_i^*(t;\eta_0)$. It can be shown that

$$\begin{aligned} & E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} = E\{I(t \le C_i)dN_i(t)|\mathcal{Z}_{it}\} \\ = \ & E\left[E\{I(t \le C_i)dN_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}\Big|\mathcal{Z}_{it}\right] \\ = \ & E\left[E\{I(t \le C_i)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}E\{dN_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}\Big|\mathcal{Z}_{it}\right] \\ = \ & E\left[\exp\{-\Lambda_0^*(t) - B_i(t) - \xi'\mathbf{Z}_\mathbf{i}^*(t)\}\exp\{\gamma'\mathbf{Z}_\mathbf{i}(t) + b_{i,p+1}(t)\}d\mu_0(t)\Big|\mathcal{Z}_{it}\right] \\ = \ & \exp\{\gamma'\mathbf{Z}_\mathbf{i}(t) - \xi'\mathbf{Z}_\mathbf{i}^*(t)\}d\Lambda_1^*(t), \end{aligned} \tag{7}$$

and thus $M_i^*(t)$ is a mean-zero stochastic process. It follows that the estimators of η and $\Lambda_1^*(t)$ can be obtained by solving the following two estimating equations

$$U_\eta(\eta) = \sum_{i=1}^n \int_0^\tau \Big\{\mathbf{X}_\mathbf{i}(t) - \bar{X}(t;\eta)\Big\}d\widetilde{N}_i(t) = 0 \tag{8}$$

and

$$\sum_{i=1}^n \left[d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_\mathbf{i}(t)}d\Lambda_1^*(t)\right] = 0 \tag{9}$$

together. In the above, τ is a known time point often representing the length of the study, $\bar{X}(t;\eta) = S^{(1)}(t;\eta)/S^{(0)}(t;\eta)$ and $S^{(k)}(t;\eta) = n^{-1}\sum_{i=1}^n e^{\eta'\mathbf{X}_\mathbf{i}(t)}\mathbf{X}_\mathbf{i}(t)^{\otimes k}$ with $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $\bar{x}(t) = lim_{n\to\infty}\bar{X}(t;\eta_0)$ and $s^{(k)}(t) = lim_{n\to\infty}S^{(k)}(t;\eta_0)$, $k = 0, 1$.

To implement the estimation procedure presented above, we propose first to obtain the estimators $\hat{\eta}$ and $d\hat{\Lambda}_1^*(t)$ for η and $d\Lambda_1^*(t)$, respectively, by solving equations (8) and (9), then to estimate β and $d\Lambda_{2k}^*(t)$ by solving equations (5) and (6) with η and $d\Lambda_1^*(t)$ replaced by $\hat{\eta}$ and $d\hat{\Lambda}_1^*(t)$, respectively. The resulted estimators of β and $d\Lambda_{2k}^*(t)$ are denoted by $\hat{\beta}$ and $d\hat{\Lambda}_{2k}^*(t)$. To establish the asymptotic properties of $\hat{\beta}$ and $\hat{\eta}$, define

$$\widehat{M}_i^*(t) = \widetilde{N}_i(t) - \int_0^t e^{\hat{\eta}'\mathbf{X}_\mathbf{i}(s)}d\hat{\Lambda}_1^*(s;\hat{\eta}),$$

$$\widehat{M}_{ik}(t) = \int_0^t Y_{ik}(s)d\widetilde{N}_i(s) - \int_0^t e^{\hat{\beta}'\mathbf{Z}_\mathbf{i}(s)+\hat{\eta}'\mathbf{X}_\mathbf{i}(s)}d\hat{\Lambda}_{2k}^*(s;\hat{\beta},\hat{\eta}),$$

$$\widehat{E}_Z(t;\beta,\eta) = \frac{\sum_{i=1}^n \mathbf{Z}_\mathbf{i}(t)e^{\beta'\mathbf{Z}_\mathbf{i}(t)+\eta'\mathbf{X}_\mathbf{i}(t)}}{\sum_{i=1}^n e^{\beta'\mathbf{Z}_\mathbf{i}(t)+\eta'\mathbf{X}_\mathbf{i}(t)}} \quad \text{and} \quad e_z(t) = lim_{n\to\infty}\widehat{E}_Z(t;\beta_0,\eta_0).$$

The following theorem gives the consistency and asymptotic normality of $\hat{\beta}$ and $\hat{\eta}$.

Theorem 1. Assume that the conditions (C1)-(C5) given in Appendix A hold. Then $\hat{\eta}$ and $\hat{\beta}$ are consistent estimators of η_0 and β_0, respectively. Furthermore, the distributions of $n^{1/2}(\hat{\eta} - \eta_0)$ and $n^{1/2}(\hat{\beta} - \beta_0)$ are asymptotically normal with mean zero and covariance matrices that can be consistently estimated by $\widehat{\Sigma}_\eta = \widehat{\Omega}_\eta^{-1}\widehat{\Psi}\widehat{\Omega}_\eta^{-1}$ and $\widehat{\Sigma}_\beta = \widehat{A}_\beta^{-1}\widehat{\Sigma}\widehat{A}_\beta^{-1}$, respectively, where $a^{\otimes 2} = aa'$, $\widehat{\Psi} = n^{-1}\sum_{i=1}^n \hat{u}_i^{\otimes 2}$, $\widehat{\Sigma} = n^{-1}\sum_{i=1}^n(\hat{v}_{1i} - \hat{v}_{2i})^{\otimes 2}$,

$$\hat{u}_i = \int_0^\tau \Big(\mathbf{X}_\mathbf{i}(t) - \bar{X}(t;\hat{\eta})\Big)d\widehat{M}_i^*(t)\,,$$

$$\hat{v}_{1i} = \sum_{k=1}^{p} \int_0^{\tau} W(t)\Big(\mathbf{Z_i}(t) - \widehat{E}_Z(t;\hat{\beta},\hat{\eta})\Big) d\widehat{M}_{ik}(t)\,,$$

$$\hat{v}_{2i} = \int_0^{\tau} \widehat{A}_{\eta}\Omega_{\eta}^{-1}\Big(\mathbf{X_i}(t) - \bar{X}(t;\hat{\eta})\Big) d\widehat{M}_i^*(t)\,,$$

$$\widehat{A}_{\beta} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{p} \int_0^{\tau} W(t) e^{\hat{\beta}' Z_i(t) + \hat{\eta}'\mathbf{X_i}(t)}\Big(\mathbf{Z_i}(t) - \widehat{E}_Z(t;\hat{\beta},\hat{\eta})\Big)^{\otimes 2} d\hat{\Lambda}_{2k}^*(t;\hat{\beta},\hat{\eta}),$$

$$\widehat{A}_{\eta} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{p} \int_0^{\tau} W(t) e^{\hat{\beta}'\mathbf{Z_i}(t) + \hat{\eta}'\mathbf{X_i}(t)}\Big(\mathbf{Z_i}(t) - \widehat{E}_Z(t;\hat{\beta},\hat{\eta})\Big)\mathbf{X_i'}(t) d\hat{\Lambda}_{2k}^*(t;\hat{\beta},\hat{\eta})$$

and

$$\widehat{\Omega}_{\eta} = n^{-1}\sum_{i=1}^{n} \int_0^{\tau} \{\mathbf{X_i}(t) - \bar{X}(t;\hat{\eta})\}^{\otimes 2} e^{\hat{\eta}'\mathbf{X_i}(t)} d\hat{\Lambda}_1^*(t;\hat{\eta}).$$

The proof of the theorem above is sketched in Appendix A.

4 Model Diagnostics

For the implementation of the estimation procedure described above, one question of practical interest is the appropriateness of the regression models. In this section, we will consider the checking of the adequacy of models (1) - (3). For this, first note that one observes complete data for both models (2) and (3) and there exist several procedures to check their goodness-of-fit (Schoenfeld, 1982; Lin et al., 1993; Lin et al., 2000). For the diagnostics of model (1), a general approach is to employ the following supremum statistic (Lin et al., 1993; Lin et al., 2000; Zhao et al., 2013a)

$$\mathcal{F}(t,z) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{p} \int_0^{t} I(\mathbf{Z}_i(s) \le z) d\widehat{M}_{ik}(s)\,,$$

where the event $\{\mathbf{Z}_i(t) \le z\}$ means that each component of $\mathbf{Z}_i(t)$ is not larger than the corresponding component of z. In Appendix B, we will show that the null distribution of $\mathcal{F}(t,z)$ converges weakly to a mean-zero Gaussian process that can be approximated by

$$\widehat{\mathcal{F}}(t,z) = n^{-1/2}\sum_{i=1}^{n}\Big\{\hat{u}_{1i}(t,z) - \widehat{\Phi}_{\eta}(t,z)\widehat{\Omega}_{\eta}^{-1}\hat{u}_{2i} - \widehat{\Phi}_{\beta}(t,z)\widehat{A}_{\beta}^{-1}(\hat{v}_{1i} - \hat{v}_{2i})\Big\} e_i\,. \tag{10}$$

Here $e_1, \ldots, e_n$ are independent standard normal variables independent of the observed data,

$$\hat{u}_{1i}(t,z) = \sum_{k=1}^{p} \int_0^{t} \{I(\mathbf{Z}_i(s) \le z) - \widehat{E}_I(s,z;\hat{\beta},\hat{\eta})\} d\widehat{M}_{ik}(s),$$

$$\widehat{\Phi}_{\eta}(t,z) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{p} \int_0^{t} \{I(\mathbf{Z}_i(s) \le z) - \widehat{E}_I(s,z;\hat{\beta},\hat{\eta})\} e^{\hat{\beta}'\mathbf{Z}_i(s) + \hat{\eta}'\mathbf{X}_i(s)} \mathbf{X}_i'(s) d\hat{\Lambda}_{2k}^*(s;\hat{\beta},\hat{\eta}),$$

$$\widehat{\Phi}_{\beta}(t,z) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{p} \int_0^{t} \{I(\mathbf{Z}_i(s) \le z) - \widehat{E}_I(s,z;\hat{\beta},\hat{\eta})\} e^{\hat{\beta}'\mathbf{Z}_i(s) + \hat{\eta}'\mathbf{X}_i(s)} \mathbf{Z}_i'(s)\, d\hat{\Lambda}_{2k}^*(s;\hat{\beta},\hat{\eta}),$$

$$\widehat{E}_I(t,z;\beta,\eta) = \frac{\sum_{i=1}^n I(\mathbf{Z}_i(t) \leq z) e^{\beta'\mathbf{Z}_i(t)+\eta'\mathbf{X}_i(t)}}{\sum_{i=1}^n e^{\beta'\mathbf{Z}_i(t)+\eta'\mathbf{X}_i(t)}},\ e_I(t,z) = lim_{n\to\infty}\widehat{E}_I(t,z;\beta_0,\eta_0)$$

and $\hat{u}_{2i} = \hat{u}_i$, where $\hat{u}_i$, $\hat{v}_{1i}$ and $\hat{v}_{2i}$ are defined in Section 3. Therefore, one could obtain a large number of realizations from $\widehat{\mathcal{F}}(t,z)$ by repeatedly generating the standard normal random samples while fixing the observed data. Because $\mathcal{F}(t,z)$ is expected to fluctuate randomly around 0 under model (1), a formal lack-of-fit test can be constructed based on the statistic $sup_{0\leq t\leq\tau,z}|\mathcal{F}(t,z)|$. The corresponding p-value can be obtained by comparing the observed value of $sup_{0\leq t\leq\tau,z}|\mathcal{F}(t,z)|$ to a large number of realizations from $sup_{0\leq t\leq\tau,z}|\widehat{\mathcal{F}}(t,z)|$.

5 A Simulation Study

In this section, we present some results obtained from a simulation study conducted to assess the finite sample behavior of the estimation procedure proposed in Section 3. In the study, we considered the situation of $p = 2$ and the covariate Z_i was assumed to be a Bernoulli random variable with the probability of success being 0.5. Given Z_i and some unobserved random effects $\mathbf{b}_i(t) = (b_{i1}(t), b_{i2}(t), b_{i3}(t), b_{i4}(t))'$, the hazard function of the censoring time C_i was assumed to have the form

$$\lambda_i(t|Z_i,\mathcal{B}_{it}) = \lambda_0 + \xi Z_i + b_{i4}(t) \tag{11}$$

with $\tau = 1$. The observation process $N_i(t)$ was assumed to follow a Poisson process on $(0, C_i)$ with the rate function

$$E\{dN_i(t)|Z_i,\mathcal{B}_{it}\} = \exp\{\gamma Z_i + b_{i3}(t)\}d\mu_0(t)\,. \tag{12}$$

Note that in practice, the exact time of C_i may not be observable and $d\widetilde{N}_i(t)$ is observed instead of $dN_i(t)$, thus we considered $E\{\widetilde{N}_i(t)|Z_i,\mathcal{B}_{it}\}$ for the number of observations. From (11) and (12),

$$E\{d\widetilde{N}_i(t)|Z_i,\mathcal{B}_{it}\} = \exp\{\gamma Z_i - \xi Z_i t\}d\Lambda_1^*(t),$$

where $d\Lambda_1^*(t) = \exp\{-\lambda_0 t + b_{i3}(t) - B_i(t)\}d\mu_0(t)$ and $B_i(t) = \int_0^t b_{i4}(s)ds$. Given Z_i and $\mathbf{b}_i(t)$, $\widetilde{N}_i(t)$ was assumed to follow a nonhomogeneous Poisson process and the total number of observation times m_i was generated from a Poisson distribution with mean $E\{m_i\} = E\{\widetilde{N}_i(\tau)|Z_i,\mathcal{B}_{i\tau}\}$. Then the observation times $\{T_{i,1},\ldots,T_{i,m_i}\}$ were taken as the order statistics of a sample of size m_i from the density function

$$f_{\widetilde{N}}(t) = \frac{\exp\{\gamma Z_i - \xi Z_i t\}d\Lambda_1^*(t)}{\int_0^\tau \exp\{\gamma Z_i - \xi Z_i t\}d\Lambda_1^*(t)}.$$

Furthermore, the bivariate panel count data $Y_{ik}(t)$ were generated from mixed Poisson processes with the mean function

$$E\{Y_{ik}(t)|Z_i,\mathcal{B}_{it},Q_i\} = Q_i\Lambda_{0k}(t)\exp\{-\beta Z_i + b_{ik}(t)\},\ k = 1,2,$$

where Q_i was sampled independently from a gamma distribution with mean 1 and variance 0.5. The results given below are based on the sample size of 100 or 200 with 1000 replications.

First we considered the situations when all components of $\mathbf{b}_i(t)$ are time-independent. Table 1 shows the estimation results on β with the use of $W(t) = 1$. Note that $\xi = 0$ or $\gamma = 0$ represents the cases when either the censoring or observation times are independent of the covariate Z_i. For the random effects, we took $b_{i1} = b_{i2} = -b_{i3} = b_{i4} = u_i$, where the $u_i's$ were generated from a uniform distribution over $(-0.5,\ 0.5)$. It can be seen that the proposed estimators seem to be unbiased and the estimated standard errors (SEE) are close to the sample standard errors (SSE). As expected, the SSEs become smaller as the sample size increases. In addition, the empirical 95% coverage probabilities (CP) are quite accurate, suggesting that the normal distribution approximation seems to be reasonable.

Now we consider the scenarios when $\mathbf{b}_i(t)$ is time-dependent. For this, Table 2 shows the estimation results under the same set-ups as for Table 1 except that $\mu_0(t) = 10t$, $b_{i1}(t) = b_{i2}(t) = u_i\, t^{1/3}$, $b_{i3}(t) = -u_i\, t^{1/2}$ and $b_{i4} = u_i$. It can be easily seen that Table 2 gives similar results compared to Table 1. Furthermore, we also considered several other set-ups such as generating $b_{ik}(t)$ from different distributions and used several different weight functions for the estimation. For example, Table 3 presents the estimation results with the use of $W(t) = \sum_{i=1}^{n} I(t \leq C_i)/n$, and they gave similar conclusions to those described above and also suggest that the estimation procedure seems to be robust with respect to the weight function $W(t)$.

When both the observation and follow-up times are informative, one possible question of practical interest is whether their dependence with the multivariate panel count responses can be characterized only by modeling the observation process but assuming noninformative censoring. To investigate this numerically, we considered both the proposed estimation procedure and the one given by Zhang et al. (2013). Table 4 presents the estimation results for β obtained under the same set-ups as for Table 2 when $n = 200$, except that $\tau = 5$, $\lambda_0 = 0.5$, $\Lambda_{01}(t) = \Lambda_{02}(t) = 0.5t$, $b_{i1}(t) = b_{i2}(t) = v_i t$, $b_{i3}(t) = \exp(v_i)$ and $b_{i4}(t) = \exp[\exp(v_i)] + G_i$, with v_i and G_i generated from a uniform distribution over $(0, 1)$ and a gamma distribution with mean 2 and variance 4, respectively. They indicate that the proposed estimator for β seems still to be unbiased, but the estimator given by Zhang et al. (2013) seems to be biased. In other words, ignoring the informative C_i could lead to biased results and conclusions.

6 An Application

In this section, we will apply the methodology described in the preceding sections to the multivariate panel count data from the Mother's Gift study described in Section 1. As mentioned before, a main objective of the study is to evaluate the overall effectiveness of PCV7 received by infants in reducing the recurrences of febrile respiratory illness, pneumonia or difficulty with breathing which are important factors in infants' growth. Note that respiratory illness is an important factor in infants' growth especially in Africa and Asia, where many infants grow up in settings of inadequate food availability and increased exposure to infections. As an example, a recent observational report from southern India described a significant relationship between infant nasal carriage of pneumococci and decreased weight and length, as well as increased odds of stunting (Coles, et al., 2012). The Monther's Gift study involved 340 pregnant women who met the inclusion criteria and agreed to participate in the study. Among them, 168 women were randomly assigned to receive pneumococcal vaccine and the rest received influenza vaccine. The infants of each maternal vaccine group were randomly assigned to receive either the PCV7 or the Hib conjugate vaccine. For the analysis below, we will focus on the data of 331 infants who had at least one observation, 165 from the PCV7 group and 166 from the Hib group.

For the analysis, define the covariates $\mathbf{Z}_i = (Z_{i1},\ Z_{i2})'$, where $Z_{i1} = 1$ if the ith infant's mother was given the influenza vaccine during pregnancy and $Z_{i1} = 0$ otherwise; $Z_{i2} = 1$ if the ith infant was given the PCV7 vaccine after birth and $Z_{i2} = 0$ if not, $i = 1, \ldots, 331$. Let $\mathbf{Y}_i(t) = \Big(Y_{i1}(t), Y_{i2}(t), Y_{i3}(t)\Big)'$ be the outcome vector associated with the i-th infant observed up to time t in weeks, where $Y_{i1}(t)$, $Y_{i2}(t)$ and $Y_{i3}(t)$ represent the total numbers of occurrences of febrile

Table 1: Simulation results on the estimation of β with $W(t) = 1$, $\lambda_0 = 2$, $\mu_0(t) = 20t$, $\Lambda_{01}(t) = 3t$, $\Lambda_{02}(t) = 3t$, $b_{i1} = b_{i2} = -b_{i3} = b_{i4}$

	$n = 100$				$n = 200$		
β_0	0	0.2	0.5		0	0.2	0.5
				$(\xi_0, \gamma_0) = (0, 0)$			
Bias	-0.008	0.008	0.000		-0.001	0.003	0.001
SEE	0.202	0.207	0.213		0.144	0.147	0.152
SSE	0.216	0.204	0.226		0.148	0.155	0.153
CP	0.937	0.957	0.935		0.943	0.935	0.948
				$(\xi_0, \gamma_0) = (0.2, 0)$			
Bias	0.038	0.044	0.033		0.036	0.041	0.037
SEE	0.204	0.209	0.217		0.145	0.148	0.155
SSE	0.207	0.218	0.223		0.147	0.152	0.155
CP	0.943	0.945	0.930		0.933	0.944	0.941
				$(\xi_0, \gamma_0) = (0, 0.5)$			
Bias	0.002	-0.007	-0.003		0.012	0.003	0.004
SEE	0.199	0.203	0.211		0.143	0.145	0.150
SSE	0.201	0.210	0.223		0.147	0.148	0.149
CP	0.940	0.947	0.932		0.946	0.943	0.953
				$(\xi_0, \gamma_0) = (0.2, 0.5)$			
Bias	0.031	0.043	0.045		0.036	0.035	0.042
SEE	0.200	0.205	0.213		0.142	0.145	0.151
SSE	0.209	0.211	0.213		0.147	0.150	0.154
CP	0.928	0.934	0.941		0.938	0.938	0.934

Table 2: Simulation results on the estimation of β with $W(t) = 1$, $\lambda_0 = 2$, $\mu_0(t) = 10t$, $\Lambda_{01}(t) = 3t$, $\Lambda_{02}(t) = 3t$, $b_{ii}(t) = b_{i2}(t) = u_i t^{1/3}$, $b_{i3}(t) = -u_i\sqrt{t}$, $b_{i4}(t) = u_i$

	$n = 100$				$n = 200$		
β_0	0	0.2	0.5		0	0.2	0.5
				$(\xi_0, \gamma_0) = (0, 0)$			
Bias	0.006	-0.003	-0.003		-0.010	0.002	0.002
SEE	0.211	0.214	0.221		0.150	0.154	0.159
SSE	0.219	0.224	0.220		0.158	0.161	0.166
CP	0.940	0.934	0.945		0.932	0.934	0.928
				$(\xi_0, \gamma_0) = (0.2, 0)$			
Bias	0.046	0.049	0.036		0.032	0.032	0.040
SEE	0.212	0.218	0.226		0.152	0.156	0.161
SSE	0.216	0.223	0.228		0.162	0.162	0.168
CP	0.944	0.932	0.934		0.928	0.934	0.932
				$(\xi_0, \gamma_0) = (0, 0.5)$			
Bias	0.003	0.002	-0.010		0.000	-0.002	-0.001
SEE	0.206	0.210	0.217		0.147	0.150	0.155
SSE	0.217	0.229	0.230		0.156	0.152	0.159
CP	0.936	0.920	0.930		0.934	0.944	0.939
				$(\xi_0, \gamma_0) = (0.2, 0.5)$			
Bias	0.045	0.037	0.040		0.043	0.041	0.038
SEE	0.207	0.210	0.219		0.147	0.154	0.156
SSE	0.221	0.222	0.224		0.150	0.166	0.166
CP	0.932	0.918	0.932		0.928	0.919	0.930

Table 3: Simulation results on the estimation of β obtained under the same set-ups as for Table 1 but with $W(t) = \sum_{i=1}^{n} I(t \leq C_i)/n$

	$n = 100$				$n = 200$		
β_0	0	0.2	0.5		0	0.2	0.5
				$(\xi_0, \gamma_0) = (0, 0)$			
Bias	-0.003	0.000	0.002		-0.009	0.007	-0.003
SEE	0.208	0.212	0.221		0.149	0.151	0.159
SSE	0.217	0.210	0.224		0.150	0.155	0.161
CP	0.942	0.951	0.940		0.950	0.945	0.947
				$(\xi_0, \gamma_0) = (0.2, 0)$			
Bias	0.048	0.044	0.044		0.050	0.051	0.051
SEE	0.210	0.215	0.226		0.150	0.154	0.161
SSE	0.213	0.225	0.227		0.149	0.156	0.165
CP	0.936	0.931	0.945		0.936	0.936	0.932
				$(\xi_0, \gamma_0) = (0, 0.5)$			
Bias	-0.005	0.006	0.000		0.006	-0.001	0.007
SEE	0.204	0.208	0.217		0.145	0.149	0.154
SSE	0.207	0.216	0.221		0.151	0.143	0.152
CP	0.942	0.928	0.946		0.942	0.955	0.959
				$(\xi_0, \gamma_0) = (0.2, 0.5)$			
Bias	0.053	0.044	0.048		0.039	0.045	0.049
SEE	0.204	0.210	0.218		0.146	0.150	0.155
SSE	0.209	0.214	0.221		0.151	0.150	0.157
CP	0.937	0.942	0.944		0.936	0.936	0.930

Table 4: Estimated biases for β by the proposed method and that in Zhang et al. (2013)

β_0	Proposed	Zhang et al. (2013)
$(\xi_0, \gamma_0) = (0.2, 0.5)$		
0	0.035	0.129
0.2	-0.009	0.115
$(\xi_0, \gamma_0) = (0.2, 0.8)$		
0	0.067	0.143
0.2	-0.015	-0.138
$(\xi_0, \gamma_0) = (-0.2, 0.2)$		
0	0.059	0.173
0.2	0.026	0.147
$(\xi_0, \gamma_0) = (-0.2, 0.5)$		
0	0.080	0.129
0.2	0.008	0.127
$(\xi_0, \gamma_0) = (-0.2, 0.8)$		
0	0.047	0.156
0.2	-0.013	0.176

respiratory illness, pneumonia and difficulty with breathing, respectively. In the following, we assume that the illness recurrence processes, the observation process and the censoring times can be described by models (1)-(3), respectively. Corresponding to the model notation, we have $p = 3$, and for the possible dependence as mentioned in Section 1, $\mathbf{b}_i(t) = \Big(b_{i1}(t), b_{i2}(t), b_{i3}(t), b_{i4}(t), b_{i5}(t)\Big)'$ characterizes the correlation between the recurrence process of an infant's respirator illness $\mathbf{Y}_i(t)$ and his/her total number of clinical visits $N_i(t)$ subject to a censoring time C_i. The primary interest is to make inferences about $\beta = (\beta_1, \beta_2, \beta_3)'$ with respect to $\mathbf{Y}_i(t)$, which represent the respective effectiveness of maternal vaccine and PCV7 vaccine for infants in preventing their respiratory illness.

Table 5 presents the analysis results obtained by applying the proposed estimation procedure with $W(t) = 1$. The results include the point estimates (Est.), the estimated standard errors (SEE) and the estimated 95% confidence intervals (CI). The results suggest that the PCV7 received by infants significantly reduced the rates of febrile respiratory illness, pneumonia and difficulty with breathing. With respect to the effect of maternal vaccine, the point estimate also indicates an effect of reducing the occurrences of infants' respiratory illness; however, the effectiveness does not appear to be significant in addition to that of PCV7 received by infants. Compared with the original analysis of the study (Zaman et al., 2008) that concluded significant effectiveness of maternal vaccine in reducing laboratory-confirmed influenza, one possible reason for the discrepancy is that our response vector of interest is different here. Another reason can be that the effects of both maternal vaccine and PCV7 received by infants are correlated such that the inclusion of the latter masks the former. In addition, the original study applied chi-square test and Fisher's exact test based on the ratio between the numbers of infant influenza cases and of infants as discussed in Section 1, which does not consider the recurrence rates of responses and ignores the possibly informative observation times.

In order to evaluate the model adequacy of the proposed methodology, we also applied the model-checking procedure described in Section 4 and obtained the p-value of 0.416, which suggests that the proposed models seem to be appropriate for the influenza immunization data.

Table 5: Analysis results for the maternal influenza immunization data

	Est.	SEE	95% CI
β_1	-0.053	0.217	(-0.477, 0.372)
β_2	-0.583	0.252	(-1.077, -0.089)
γ_1	0.884	0.090	(0.708, 1.060)
γ_2	0.921	0.090	(0.744, 1.097)
ξ_1	0.054	0.005	(0.043, 0.064)
ξ_2	0.056	0.005	(0.045, 0.066)

7 Concluding Remarks

Regression analysis of multivariate panel count data has been studied for the situations where either the observation process or follow-up times are noninformative about the response process. As discussed above and shown in the example, this may not be true in reality and the ignorance of their informativeness could lead to biased results. To address this, we proposed a joint analysis method that allows the possible mutual correlations. For inference, an easy-to-implement estimating equation approach was developed and both finite and asymptotic properties of the resulting estimators were established. In particular, the numerical results showed that the proposed procedure works well for practical situations.

Compared to the existing models for dependent processes, the proposed joint model is flexible in that the shared vector of random effects can be time-dependent and neither of its structure nor distribution is prespecified. Note that for simplicity, however, it has been assumed that the random effects are multiplicative to all response processes and $\mathbf{b}_i(t)$ is covariates-independent. It is apparent that it would be helpful to relax or check the appropriateness of these assumptions. The same is true about how to choose the form of the regression model for the response process $\mathbf{Y}_i(t)$ among a class of models such as transformation models for a given set of data. Except for the situations when some extra information is available, a conventional method is to develop an omnibus goodness-of-fit test based on the cumulative summation of the residual process as presented in Section 4.

There exist several other possible directions for future work. For example, instead of the proposed model, one may be interested in some other models such as the additive model for $\mathbf{Y}_i(t)$. Although the idea discussed above still applies, both the estimating equation and the asymptotic properties of the resulting estimators can be quite different depending on the model. With respect to dependent processes, instead of employing joint modeling approach, one may consider marginal modeling approach in that the mean function of $\mathbf{Y}_i(t)$ can be estimated marginally. On the other hand, however, the drawback is that such estimation is subject to the form of correlations specified to characterize the marginal effects of $N_i(t)$. In order to provide flexibility and improve the estimation efficiency, we have employed a weight function $W(t)$ in the proposed method. For the selection of an optimal $W(t)$, however, this is clearly a difficult problem as it requires the specification of the covariance function of $Y_i(t)$ and $\widetilde{N}_i(t)$ as in most similar situations (Sun et al., 2012). It would be desirable to develop some procedures to solve the problem. In addition to dependent follow-up times due to a censoring event, sometimes a terminal event such as death may exist (Ghosh and Lin, 2002; Zhao et al., 2011b). In such complicated situations, the follow-up process is subject to competing risks and a new method is thus necessary for this challenge.

8 Acknowledgments

The authors wish to thank the Editor, the Associate Editor and three reviewers for their many helpful comments and suggestions that greatly improved the paper. The research was supported in part by the Faculty Research Grant (No. 1-11172) from the University of North Carolina at Charlotte to the first author.

A Proof of Theorem 1

To derive the asymptotic properties of the proposed estimators $\hat{\beta}$ and $\hat{\eta}$, we need the following regularity conditions. For $k = 1, \ldots, p$:

(C1). $\{\mathbf{Y}_i(t), \widetilde{N}_i(t), C_i, \mathbf{Z_i}(t)\}$ $(0 \leq t \leq \tau,\ i = 1, \ldots, n)$ are independent and identically distributed.

(C2). $P(C_i \geq \tau) > 0$.

(C3). $Y_{ik}(t)$ and $\widetilde{N}_i(t)$ $(0 \leq t \leq \tau,\ i = 1, \ldots, n)$ are all bounded.

(C4). $W(t)$ and $\mathbf{Z_i}(t)$, $i = 1, \ldots, n$, have bounded variations almost surely and $W(t)$ converges in probability to a deterministic function $w(t)$ uniformly in $t \in [0, \tau]$.

(C5). $A_\beta = E\{\sum_{k=1}^{p} \int_0^\tau W(t) e^{\beta_0' \mathbf{Z_i}(t) + \eta_0' \mathbf{X_i}(t)} [\mathbf{Z_i}(t) - e_z(t)]^{\otimes 2} d\Lambda_{2k}^*(t)\}$ and $\Omega_\eta = E\Big[\int_0^\tau \{\mathbf{X_i}(t) - \bar{x}(t)\}^{\otimes 2} e^{\eta_0' \mathbf{X_i}(t)} d\Lambda_1^*(t)\Big]$ are both positive definite.

For the conditions above, (C1) depends on the design of the study and the sampling method, which is generally true for randomized clinical trials. (C2) is true if not all subjects are censored. (C3), (C4) and (C5) ensure the existence of variance-covariance related matrices and are defined for estimation feasibility. Specifically, (C3) is true if the total numbers of events and observations are bounded in a fixed period of time for each subject and is usually reasonable for panel count data in medical studies. (C4) defines the choices of $W(t)$ and the condition on $\mathbf{Z}_i(t)$ is often true as $\mathbf{Z}_i(t)$ is often smooth enough or monotonic in practice; in particular, such condition is true if one observes time-independent covariates. (C5) is true as long as the components of $\mathbf{Z}_i(t)$ are not strongly correlated.

Define

$$U_1(\beta; \hat{\eta}) = \sum_{i=1}^{n} \sum_{k=1}^{p} \int_0^\tau W(t) \mathbf{Z_i}(t) \Big[Y_{ik}(t) d\widetilde{N}_i(t) - e^{\beta' \mathbf{Z_i}(t) + \hat{\eta}' \mathbf{X_i}(t)} d\hat{\Lambda}_{2k}^*(t; \beta, \hat{\eta}) \Big]$$

and note that $d\hat{\Lambda}^*_{2k}(t;\beta,\hat{\eta})$ satisfies

$$\sum_{i=1}^{n}\left[Y_{ik}(t)d\widetilde{N}_i(t) - e^{\beta'\mathbf{Z_i}(t)+\hat{\eta}'\mathbf{X_i}(t)}d\hat{\Lambda}^*_{2k}(t;\beta,\hat{\eta})\right] = 0,\ \ 0 \le t \le \tau. \qquad (A.1)$$

Let $\widehat{A}_\beta(\beta) = -n^{-1}\partial U_1(\beta,\hat{\eta})/\partial\beta$, $\widehat{A}_\eta(\eta) = -n^{-1}\partial U_1(\beta_0,\eta)/\partial\eta$, $A_\beta = \lim_{n\to\infty}\widehat{A}_\beta(\beta_0)$ and $A_\eta = \lim_{n\to\infty}\widehat{A}_\eta(\eta_0)$. The consistency of $\hat{\beta}$ and $\hat{\eta}$ follows from the facts that $U_\eta(\eta_0)$ and $U_1(\beta_0;\hat{\eta})$ both tend to 0 in probability as $n\to\infty$, and that $-n^{-1}\partial U_\eta(\eta)/\partial\eta$ and $\widehat{A}_\beta(\beta)$ both converge uniformly to the positive definite matrices Ω_η and A_β over η and β, respectively, in neighborhoods around the true values η_0 and β_0. Then the Taylor series expansions of $U_1(\hat{\beta};\hat{\eta})$ at $(\beta_0;\hat{\eta})$ and (β_0,η_0) yield $n^{1/2}(\hat{\beta}-\beta_0) = A_\beta^{-1}n^{-1/2}U_1(\beta_0;\hat{\eta}) + o_p(1) = A_\beta^{-1}\Big\{n^{-1/2}U_1(\beta_0;\eta_0) - A_\eta n^{1/2}(\hat{\eta}-\eta_0)\Big\} + o_p(1)$. The proof of the asymptotic normality in Theorem 1 is sketched as follows:

(1) First, using some derivation operation to $U_1(\beta;\hat{\eta})$ and (A.1), we get

$$\widehat{A}_\beta(\beta) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{p}\int_0^\tau W(t)\big\{\mathbf{Z_i}(t) - \widehat{E}_Z(t;\beta,\hat{\eta})\big\}^{\otimes 2}p_i^x(t)d\hat{\Lambda}^*_{2k}(t;\beta,\hat{\eta}),$$

where $p_i^x(t) = e^{\beta'\mathbf{Z_i}(t)+\hat{\eta}'\mathbf{X_i}(t)}$.

(2) Solving $d\hat{\Lambda}^*_{2k}(t;\beta,\hat{\eta})$ from (A.1), inserting it into $U_1(\beta;\hat{\eta})$ and evaluating the function at (β_0,η_0) yields

$$n^{-1/2}U_1(\beta_0;\eta_0) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{p}\int_0^\tau w(t)\Big(\mathbf{Z_i}(t) - e_z(t)\Big)dM_{ik}(t) + o_p(1).$$

(3) Differentiation of $U_1(\beta_0,\eta)$ and (A.1) with respect to η' yields

$$\widehat{A}_\eta(\eta) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{p}\int_0^\tau W(t)\big[\mathbf{Z_i}(t) - \widehat{E}_Z(t;\beta_0,\eta)\big]e^{\beta_0'\mathbf{Z_i}(t)+\eta'\mathbf{X_i}(t)}\mathbf{X}'_\mathbf{i}(t)d\hat{\Lambda}^*_{2k}(t;\beta_0,\eta)\,.$$

(4) According to equation (8) and the arguments similar as Lin et al. (2000), one can show that

$$n^{1/2}\{\hat{\eta}-\eta_0\} = \Omega_\eta^{-1}n^{-1/2}\sum_{i=1}^{n}\left[\int_0^\tau\Big(\mathbf{X_i}(t) - \bar{x}(t)\Big)dM_i^*(t)\right] + o_p(1), \qquad (A.2)$$

where $\Omega_\eta = E\Big[\int_0^\tau\big\{\mathbf{X_i}(t) - \bar{x}(t)\big\}^{\otimes 2}e^{\eta_0'\mathbf{X_i}(t)}d\Lambda_1^*(t)\Big]$.

Combining the results in steps (1)-(4), we have

$$U_1(\beta_0;\hat{\eta}) = \sum_{i=1}^{n}\left[\sum_{k=1}^{p}\int_0^\tau w(t)\big\{\mathbf{Z_i}(t) - e_z(t)\big\}dM_{ik}(t)\right] - A_\eta\Omega_\eta^{-1}\sum_{i=1}^{n}\left[\int_0^\tau\big\{\mathbf{X_i}(t) - \bar{x}(t)\big\}dM_i^*(t)\right] + o_p(n^{1/2}),$$

then it follows from the multivariate central limit theorem that $n^{1/2}(\hat{\eta}-\eta_0)$ and $n^{1/2}(\hat{\beta}-\beta_0)$ are both asymptotically normally distributed with mean zero and covariance matrices that can be consistently estimated by $\widehat{\Sigma}_\eta = \widehat{\Omega}_\eta^{-1}\widehat{\Psi}\widehat{\Omega}_\eta^{-1}$ and $\widehat{\Sigma}_\beta = \widehat{A}_\beta^{-1}\widehat{\Sigma}\widehat{A}_\beta^{-1}$, respectively.

B Proof of the null distribution of $\mathcal{F}(t,z)$ in Section 3

Let $V(\hat{\beta},\hat{\eta}) = \sum_{i=1}^{n}\sum_{k=1}^{p}\int_0^t I(\mathbf{Z}_i(s) \le z) d\widehat{M}_{ik}(s;\hat{\beta},\hat{\eta})$. By the Taylor expansion,

$$\mathcal{F}(t,z;\hat{\beta},\hat{\eta}) = n^{-1/2}V(\beta_0,\eta_0) + \frac{\partial V(\beta_0,\eta_0)}{n\partial\eta'}\sqrt{n}(\hat{\eta}-\eta_0) + \frac{\partial V(\beta_0,\hat{\eta})}{n\partial\beta'}\sqrt{n}(\hat{\beta}-\beta_0) + o_p(1).$$

Using the arguments and algebra manipulation similar to those in Appendix A, we have $V(\beta_0,\eta_0) = \sum_{i=1}^{n} u_{1i}(t,z) + o_p(n^{1/2})$, where $u_{1i}(t,z) = \sum_{k=1}^{p}\int_0^t \{I(\mathbf{Z}_i(s) \le z) - e_I(s,z)\} dM_{ik}(s)$. Also, $\frac{\partial V(\beta_0,\eta_0)}{n\partial\eta'}$ and $\frac{\partial V(\beta_0,\hat{\eta})}{n\partial\beta'}$ can be consistently estimated by $-\widehat{\Phi}_\eta(t,z)$ and $-\widehat{\Phi}_\beta(t,z)$, respectively.

In addition, it follows from (A.2) and Theorem 1 that

$$\sqrt{n}\{\hat{\eta}-\eta_0\} = \Omega_\eta^{-1} n^{-1/2}\sum_{i=1}^{n}\left[\int_0^\tau \Big(\mathbf{X}_i(t) - \bar{x}(t)\Big) dM_i^*(t)\right] + o_p(1),$$

and

$$\sqrt{n}\{\hat{\beta}-\beta_0\} = A_\beta^{-1} n^{-1/2}\sum_{i=1}^{n}(v_{1i} - v_{2i}) + o_p(1),$$

where $v_{1i} = \sum_{k=1}^{p}\int_0^\tau w(t)\Big(\mathbf{Z}_i(t) - e_z(t)\Big) dM_{ik}(t)$, and $v_{2i} = \int_0^\tau A_\eta\Omega_\eta^{-1}\Big(\mathbf{X}_i(t) - \bar{x}(t)\Big) dM_i^*(t)$. Therefore, $\mathcal{F}(t,z;\hat{\beta},\hat{\eta})$ can be expressed as a sum of i.i.d. mean-zero terms for fixed t. By the multivariate central limit theorem, $\mathcal{F}(t,z)$ converges in finite-dimensional distribution to a mean-zero Gaussian distribution. Since $\mathcal{F}(t,z)$ is tight based on the empirical process theory, $\mathcal{F}(t,z)$ converges weakly to a mean-zero Gaussian process that can be approximated by $\widehat{\mathcal{F}}(t,z)$ given by equation (10).

References

Chen, B. E., Cook, R. J., Lawless, J. F. and Zhan, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Statistics in Medicine*, **24**, 671-691.

Cheng, S. C. and Wei, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika*, **87**, 89-97.

Coles, C.L., Rahmathullah, L., Kanungo, R., Katz, J., Sandiford, D., Devi, S., Thulasiraj, R.D. and Tielsch, J.M. (2012). Pneumococcal carriage at age 2 months is associated with growth deficits at age 6 months among infants in South India. *The Journal of Nutrition*, **142** (6), 1088-1094.

Ghosh, D., Lin, D. Y. (2002). Marginal regression models for recurrent and terminal events. Statistica Sinica, 12, 663-688.

He, X., Tong, X. and Sun, J. and Cook, R. J. (2008). Regression analysis of multivariate panel count data. *Biostatistics*, **9**, 234-248.

Hu, X. J., Sun J. and Wei L. J. (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics*, **30**, 25-43.

Huang, C. Y., Wang, M. C., and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika*, **93**, 763-775.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Li, N., Park, D.-H., Sun, J. and Kim, K. (2011), Semiparametric transformation models for multivariate panel count data with dependent observation process. *Canadian Journal of Statistics*, **39**, 458-474.

Li, Y., Zhao, H., Sun, J. and Kim, K. M. (2014). Nonparametric tests for panel count data with unequal observation processes. *Computational Statistics & Data Analysis*, **73**, 103-111.

Lin, D. Y., Oaks, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometirka*, **85** (2), 289-298.

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.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, **80**, 557-572.

McCulluagh and Nelder (1989). Generalized linear models. Chapman and Hall, London.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

Sun, J., Tong, X. and He, X. (2007). Regression analysis of panel count data with dependent observation times. *Biometrics*, **63**, 1053-1059.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B*, **62**, 293-302.

Sun, J. and Zhao, X., (2013). The Statistical Analysis of Panel Count Data. Springer Science+Business Inc.

Sun, L., Song, X., Zhou, J. and Liu, L. (2012). Joint analysis of longitudinal data with informative observation times and a dependent terminal event. Journal of the American Statistical Association, 107(498), 688-700.

Torres, A. M., Peterson, K. E., de Souza, A. C., Orav, E. J., Hughes, M., and Chen, L. C. (2000). Association of diarrhoea and upper respiratory infections with weight and height gains in Bangladeshi children aged 5 to 11 years. Bulletin of the World Health Organization, **78** (11), 1316-1323.

Zaman, K., Roy, E., Arifeen, S.E., Rahman, M., Raqib, R., Wilson, E., Omer, S.B., Shahid, N.S., Breiman, R.F. and Steinhoff, M.C. (2008). Effectiveness of maternal influenza immunization in mothers and infants. *The New England Journal of Medicine*, **359**, 1555-1564.

Zhang, H., Zhao, H., Sun, J., Wang, D. and Kim, K.M. (2013). Regression analysis of multivariate panel count data with an informative observation process. *Journal of Multivariate Analysis*, **119** (C), 71-80.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika*, **89**, 39-48.

Zhang, Z., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, **24**, 1399-1407.

Zhao, H., Li, Y. and Sun, J. (2013a). Semiparametric analysis of multivariate panel count data with dependent observation processes and a terminal event. *nonparametrics*, **25** (2), 379-394.

Zhao, X. and Tong, X. (2011a). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics and Data Analysis*, **55** (1), 291-300.

Zhao, X., Tong, X. and Sun, J. (2013b). Robust estimation for panel count data with informative observation times. *Computational Statistics and Data Analysis*, **57**, 33-40.

Zhao, X., Zhou, J. and Sun, L. (2011b). Semiparametric Transformation Models with Time-Varying Coefficients for Recurrent and Terminal Events. *Biometrics*, **67**, 404-414.