Published in final edited form as:
*Lifetime Data Anal*. 2012 January ; 18(1): 116–138. doi:10.1007/s10985-011-9209-x.

# Marginal Hazard Regression for Correlated Failure Time Data with Auxiliary Covariates

**Yanyan Liu**[1], **Zhongshang Yuan**[1,2], **Jianwen Cai**[3], and **Haibo Zhou**[3,*]

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

[2] Department of Epidemiology and Health Statistics, Shandong University, Jinan, Shandong 250012, China

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, U.S.A.

## Abstract

In many biomedical studies, it is common that due to budget constraints, the primary covariate is only collected in a randomly selected subset from the full study cohort. Often, there is an inexpensive auxiliary covariate for the primary exposure variable that is readily available for all the cohort subjects. Valid statistical methods that make use of the auxiliary information to improve study efficiency need to be developed. To this end, we develop an estimated partial likelihood approach for correlated failure time data with auxiliary information. We assume a marginal hazard model with common baseline hazard function. The asymptotic properties for the proposed estimators are developed. The proof of the asymptotic results for the proposed estimators is nontrivial since the moments used in estimating equation are not martingale-based and the classical martingale theory is not sufficient. Instead, our proofs rely on modern empirical theory. The proposed estimator is evaluated through simulation studies and is shown to have increased efficiency compared to existing methods. The proposed methods are illustrated with a data set from the Framingham study.

### Keywords

Marginal hazard model; Correlated failure time; Validation set; Auxiliary covariate

## 1 Introduction

Exposure assessment like assays for biomarker or genetic traits can be prohibitively expensive in modern biomedical studies. Due to budget constraints, the main exposure in many studies can only be assembled on a subset of the full study cohort. This subset is refereed to as the validation set. Meanwhile, an inexpensive auxiliary variable for the main exposure is often readily available for all the cohort subjects. It is desirable to improve the study efficiency through properly utilizing these auxiliary information in the statistical inference.

In failure time studies, some methods have been proposed. For example, Zhou and Pepe (1995), Zhou and Wang (2000) and Wang et al. (1997) studied the auxiliary covariates problem for a multiplicative semiparametric hazard model using regression calibration techniques. Kulich and Lin (2000) and Jiang and Zhou (2007) proposed a corrected pseudo-

[*]zhou@bios.unc.edu *Tel*: 919-966-3885 *Fax*: 919-966-3804.

score estimator for the additive risks model of Lin and Ying (1994). While the aforementioned methods are focused on univariate failure time data, correlated failure time data are commonly encountered in practice. For example, the data arise from family studies where individuals within a family may be correlated due to shared genetic or environmental factors. Two major classes of models have been proposed for correlated failure time data: the frailty models (Clayton and Cuzick, 1985; Nielsen et al., 1992; Hougaard, 2000; Gorfine, Zucker, and Hsu, 2006) and the marginal hazard models (Wei, Lin, and Weissfeld, 1989; Lee, Wei, and Amato, 1992; Cai and Prentice, 1995, 1997; Spiekerman and Lin, 1998). When the intracluster correlation is not of interest, marginal hazard models are preferred approach since they avoid strong assumptions about the dependencies among correlated failure times.

In the literature of marginal hazard models, two types of models have been extensively studied: the different baseline hazard model (Wei et al. 1989, referred to as the WLW model) and the common baseline model (Lee et al. 1992, hereafter referred to as the CBM model). When the main exposure is observed only on a validation set, several methods have been proposed to fit marginal hazard model by taking use the auxiliary information. For example, Hu and Lin (2002) developed a corrected score approach to provide a class of consistent estimators assuming that the auxiliary and the true covariate have the same mean. Liu et al. (2009) proposed an estimated pseudo-partial likelihood method assuming a discrete auxiliary covariate. For continuous auxiliary covariate, Liu et al. (2010) proposed to correct the partial likelihood through a kernel estimation procedure. All these methods are based on a marginal hazard model with different baseline hazards(WLW model). However, in many practical situations, including studies of disease occurrence patterns of twins or siblings, or in litter mate experiments, or the clustered failure time data in which the subjects within clusters are exchangeable, it will be natural to restrict the baseline hazard functions to be common for some or all members of a cluster. To the best of our knowledge, no methods are available for analysis of correlated failure time data with auxiliary information under the framework of CBM model.

In this paper, we propose a new inference method based on a pseudo-partial likelihood for the clustered failure time data where the main exposure is only observed for the validation set. We assume a marginal proportional hazards model with common baseline hazard and discrete auxiliary variable. The relative risk function is estimated by a weighted average of all the observations from the validation set. Consequently, the resulted estimating equation is not a marginal martingale and the classical martingale theory, the key to the theoretical development of CBM model, is not sufficient to derive the asymptotic results in this case. We employ results from the modern empirical process theory to derive the asymptotic properties of the proposed estimators. Simulation studies show that our proposed estimator is more efficient than the simple estimator based on the validation data, while not much less efficient than that from the full data. The merit of our approach is that it does not require the specification of the association between the main exposure and its auxiliary.

The rest of the paper is organized as follows. In Section 2, we outline the data structure and propose an estimating procedure for the regression coefficients and cumulative hazard function. The large sample properties of the proposed estimators for regression coefficients and baseline hazard function are given in section 3. Extensive simulation studies are conducted to examine the finite-sample properties and robustness properties of the proposed methods in section 4. We illustrate the proposed method through the analysis of a real data set from the Framingham study in section 5. Concluding remarks are given in Section 6. All proofs are outlined in the Appendix.

# 2 Model and Estimation

## 2.1 Notation and Models

We first set up the requisite notation. Suppose that the full cohort consists of $n$ independent clusters, and the $i$-th cluster has $n_i$ correlated subjects. We assume that subjects within the same cluster are exchangeable conditional on covariates (Hougaard, 2000). Suppose that each individual has a fixed probability to have the main exposure covariate being measured. The set for individuals who have their main exposure covariate and other covariates being observed is refereed as validation set.

Let $T_{ik}$ and $C_{ik}$ be the failure time and censoring time for $(i, k)$, where $(i, k)$ represents the $k$-th subject in the $i$-th cluster. The observed time is $X_{ik} = \min(T_{ik}, C_{ik})$. Let $Y_{ik}(t) = I(X_{ik} \geq t)$ be at-risk indicator process, $\Delta_{ik} = I(T_{ik} \leq C_{ik})$ denotes the failure indicator and $N_{ik}(t) = I(X_{ik} \leq t, \Delta_{ik} = 1)$ is the standard counting process, where $I(\cdot)$ is the indicator function. Let $E_{ik}(t)$ and $Z_{ik}(t)$ denote the possibly time-dependent covariates, where $E_{ik}(t)$ is the main exposure subjecting to missing and $Z_{ik}(t)$ is the remaining covariate vector which is fully observed. Let $\mathbf{E}_{ik} = \{E_{ik}(t), t \geq 0\}$, and $\mathbf{Z}_{ik}$ is defined similarly. All the time-dependent covariates are assumed to be external, i.e. they are not affected by the disease processes (Kalbfleisch and Prentice, 2002). Suppose that the $n$ sets of clustered observations ($T$, $C$, $\mathbf{E}$, $\mathbf{Z}$) are independent and identically distributed. Within each cluster, the observed vectors ($T$, $C$, $\mathbf{E}$, $\mathbf{Z}$) maybe dependent on each other, but are identically distributed. The number of subjects in each cluster, $n_i$, does not depend on the observations of ($T$, $C$, $\mathbf{E}$, $\mathbf{Z}$). In addition, the clustered observations of $T$ and $C$ are assumed to be independent conditional on the clustered observations of covariates $\mathbf{E}$ and $\mathbf{Z}$ (i.e. independent censoring).

Suppose that the complete covariate histories ($E_{ik}(\cdot)$, $Z_{ik}(\cdot)$) are available for the subjects in validation set and only $Z_{ik}(\cdot)$ available for the subjects in non-validation set. Let $\eta_{ik}$ be the indicator for subject $(i, k)$ being selected into the validation set, and $\eta_{ik}$ is assumed to be independent of $\{N_{ik}(\cdot), Y_{ik}(\cdot), E_{ik}(\cdot), Z_{ik}(\cdot), n_i : k = 1, \cdots, n_i\}$. In addition, some auxiliary information for $E_{ik}(\cdot)$ are observed for the whole cohort subjects and are denoted by $A_{ik}(\cdot)$. In this paper, we assume that $A_{ik}(\cdot)$ is categorical. Therefore, the observed data can be represented by $(X_{ik}, \Delta_{ik}, \mathbf{Z}_{ik}, \eta_{ik}\mathbf{E}_{ik}, \mathbf{A}_{ik})$. For $i = 1, \cdots, n$ and $k = 1, \cdots, n_i$, we assume that $\eta_{ik}$'s are independent Bernoulli variables with distribution $Pr(\eta_{ik} = 1) = \rho$.

We assume that conditional on $(\mathbf{E}_{ik}, \mathbf{Z}_{ik})$, $\mathbf{A}_{ik}$ provides no additional information to the regression model, *i.e.*

$$\lambda_{ik}(t|\mathbf{E}_{ik}, \mathbf{Z}_{ik}, \mathbf{A}_{ik}) = \lambda_{ik}(t|\mathbf{E}_{ik}, \mathbf{Z}_{ik}), \tag{1}$$

where $\lambda_{ik}(\cdot)$ denotes the corresponding conditional marginal hazards function.

Suppose that the marginal hazard function of $T_{ik}$ follows the proportional hazards model (Cox 1972):

$$\lambda_{ik}(t|\mathbf{Z}_{ik}, \mathbf{E}_{ik}) = \lambda_0(t) \exp\left\{\beta_1' E_{ik}(t) + \beta_2' Z_{ik}(t)\right\}, \tag{2}$$

where $\lambda_0(t)$ is an unspecified common baseline hazard function, and $\beta = \left(\beta_1', \beta_2'\right)'$ are the parameters to be estimated.

When subject $(i, k)$ belongs to the non-validation set, we only observe $Z_{ik}$ and $A_{ik}$. Under this situation, we can derive an induced hazard function as follows:

$$\begin{aligned}\lambda_{ik}(t|\mathbf{Z}_{ik},\mathbf{A}_{ik}) &= \lambda_0(t)\, e^{\beta_2' Z_{ik}(t)} \mathcal{E}\left\{ e^{\beta_1' E_{ik}(t)} | Y_{ik}(t)=1, A_{ik}(t), Z_{ik}(t) \right\} \\ &= \lambda_0(t)\, e^{\beta_2' Z_{ik}(t)} \mathcal{E}\left\{ e^{\beta_1' E_{ik}(t)} | Y_{ik}(t)=1, A^*_{ik}(t) \right\},\end{aligned} \tag{3}$$

where $A^*$ includes auxiliary variable $A$ and the part of the information in covariate $Z$ that, given $A$, are still related to $E$ and $\mathcal{E}(\cdot)$ denotes the expectation. That is, $A^*$ satisfying the following conditional dependence $f(E_{ik}(t)|X_{ik}(t) \geq t, Z_{ik}(t), A_{ik}(t)) = f(E_{ik}(t)|X_{ik}(t) \geq t, A^*_{ik}(t))$, where $f$ denotes the conditional density function. Notice that under this formulation, $A^*$ still satisfies the auxiliary assumption that given $E$ and $Z$, $A^*$ does not contribute to the regression model, *i.e.*, $\lambda(t|\mathbf{Z}, \mathbf{E}, \mathbf{A}^*) = \lambda(t|\mathbf{Z}, \mathbf{E})$. In this paper, we assume that $A^*_{ik}$ is categorical.

## 2.2 Proposed estimators

By (2) and (3), we derive the induced relative risk function to the baseline as:

$$r_{ik}(\beta, t) = \left[ \exp\left\{ \beta_1' E_{ik}(t) \right\} \eta_{ik} + \phi_{ik}(\beta_1, t)(1-\eta_{ik}) \right] \exp\left( \beta_2' Z_{ik}(t) \right), \tag{4}$$

where $\phi_{ik}(\beta_1, t) = \mathcal{E}\left\{ e^{\beta_1' E_{ik}(t)} | Y_{ik}(t)=1, A^*_{ik}(t) \right\}$. Note that $\phi$ is a conditional expectation. $\phi(\beta_1, t)\, e^{\beta_2' Z(t)}$ can be interpreted as the induced relative risk for a subject with a missing $E$. If the data were completely observed, then the regression parameter $\beta$ of model (2) could be estimated by solving the estimating equation $U(\beta) = 0$ (Lee et al. 1992), where

$$U(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{n_i} \int_0^{\tau} \frac{r_{ik}^{(1)}(u;\beta)}{r_{ik}(u;\beta)} dN_{ik}(u) - \int_0^{\tau} \frac{S^{(1)}(\beta, u)}{S^{(0)}(\beta, u)} d\bar{N}(u), \tag{5}$$

with $\tau$ denote the study end time, $S^{(d)}(\beta, u) = n^{-1} \sum_{j=1}^{n} \sum_{l=1}^{n_j} Y_{jl}(u)\, r_{jl}^{(d)}(u;\beta)$, $r_{ik}^{(d)}(u;\beta)$ is the $d$-th derivative of $r_{ik}(\beta, u)$ with respect to $\beta$ ($d = 0, 1$) and $\bar{N}(u) = \sum_{i=1}^{n} \sum_{k=1}^{n_i} N_{ik}(u)$. Since the data is not complete, (5) cannot be calculated. We need to estimate $r_{ik}$ first. It is sufficient to estimate $\phi_{ik}(\beta_1, t)$. Before we give the estimation formula, we first define some necessary notations. Suppose $A^*_{ik}(t)$ is finite discrete with the distribution $P(A^*(t) = a_m) = p_m$, $m = 1 \cdots L$. Let

$$\begin{aligned}\theta_{jl}(t, a_m) &= Y_{jl}(t)\, \eta_{jl} I\left( A^*_{jl}(t) = a_m \right), \\ \theta_{j\cdot}(t, a_m) &= \sum_{l=1}^{n_j} Y_{jl}(t)\, \eta_{jl} I\left( A^*_{jl}(t) = a_m \right), \\ w_{jl}(t, a_m) &= \frac{\theta_{jl}(t, a_m)}{\theta_{j\cdot}(t, a_m)} I\left( \theta_{j\cdot}(t, a_m) > 0 \right), \\ \bar{\phi}_{j, a_m}(\beta_1, t) &= \sum_{l=1}^{n_j} w_{jl}(t, a_m)\, e^{\beta_1' E_{jl}(t)}.\end{aligned}$$

Define $\phi_{a_m}(t) = \mathcal{E}\left\{ e^{\beta_1' E_{ik}(t)} | Y_{ik}(t)=1, A^*_{ik}(t) = a_m \right\}$. It can be shown that

$$\phi_{a_m}(t) = \mathcal{E}\left\{\bar{\phi}_{j,a_m}(\beta_1, t) \,|\, Y_{jl}(t) = 1, A^*_{jl}(t) = a_m, l = 1, \cdots, n_j\right\}.$$

Therefore, it is natural to estimate $\phi_{am}(t)$ empirically by taking average over those non-zero $\bar{\phi}_{j,a_m}(\beta_1, t)$ as following:

$$\widehat{\phi}_{a_m}(\beta_1, t) = \frac{\sum_{j=1}^{n} I\left(\bar{\phi}_{j,a_m}(\beta_1, t) > 0\right) \bar{\phi}_{j,a_m}(\beta_1, t)}{\sum_{j=1}^{n} I\left(\bar{\phi}_{j,a_m}(\beta_1, t) > 0\right)}.$$

$\phi_{ik}(\beta_1, t)$ can be estimated by

$$\widehat{\phi}_{ik}(\beta_1, t) = \widehat{\phi}_{a_m}(\beta_1, t)\,|_{a_m = A^*_{ik}(t)}. \tag{6}$$

Replace $\phi_{ik}(\beta_1, t)$ in (4) with its estimator $\widehat{\phi}_{ik}(\beta_1, t)$, we obtain the estimator for relative risk $r_{ik}(\beta, t)$,

$$\widehat{r}_{ik}(\beta, t) = \exp\left\{\beta_1' E_{ik}(t)\right\} \exp\left\{\beta_2' Z_{ik}(t)\right\} \eta_{ik} + \widehat{\phi}_{ik}(\beta_1, t) \exp\left\{\beta_2' Z_{ik}(t)\right\} (1 - \eta_{ik}).$$

Define $\widehat{S}^{(d)}(\beta, t) = n^{-1} \sum_{j=1}^{n} \sum_{l=1}^{n_j} Y_{jl}(u) \widehat{r}^{(d)}_{jl}(u;\beta)$. We can estimate $\beta_0$, the true parameter, by $\widehat{\beta}_E$ the solution of $\widehat{U}(\beta) = 0$, where

$$\widehat{U}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{n_i} \int_0^\tau \frac{\widehat{r}^{(1)}_{ik}(u;\beta)}{\widehat{r}_{ik}(u;\beta)} dN_{ik}(u) - \int_0^\tau \frac{\widehat{S}^{(1)}(\beta, u)}{\widehat{S}^{(0)}(\beta, u)} d\bar{N}(u), \tag{7}$$

with $\widehat{r}^{(1)}_{ik}(\beta, u)$ ($d$=0, 1, 2) be the first derivative of $\widehat{r}_{ik}(\beta, u)$ with respect to $\beta$

By plugging the estimator of relative risk $r_{ik}$ in the commonly-used Breslow estimator for the cumulative baseline function, we obtain a natural estimator for cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$:

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{d\bar{N}(s)}{\sum_{j=1}^{n} \sum_{l=1}^{n_j} Y_{jl}(s) \widehat{r}_{jl}\left(\widehat{\beta}_E, s\right)}.$$

## 3 Asymptotic Properties

To investigate the asymptotic properties of the proposed estimator, we introduce some notations first. Let $\beta_0 = \left(\beta_{10}', \beta_{20}'\right)'$ be the true regression parameter. For a vector $a$ = ($a_1$, · · ·,

$a_p)'$, let $a^{\otimes 0}=1, \quad a^{\otimes 1}=a, \quad a^{\otimes 2}=aa'$. Unless otherwise stated, all the limits are taken as $n \to \infty$.

Let $M_{ik}(t)=N_{ik}(t)-\int_0^t \lambda_{ik}(u)\,du$ be the marginal martingale and

$$\begin{aligned} s^{(d)}(\beta,t) &= \mathcal{E}\left(\sum_{k=1}^{n_i}\left\{Y_{ik}(t)\, r_{ik}^{(d)}(\beta,t)\right\}\right), (d=0,1,2) \\ s^{(3)}(\beta,t) &= \mathcal{E}\left(\sum_{k=1}^{n_i}\left[Y_{ik}(t)\left(\frac{r_{ik}^{(1)}(\beta,t)}{r_{ik}(\beta,t)}\right)^{\otimes 2} r_{ik}(\beta_0,t)\right]\right), \\ s^{(4)}(\beta,t) &= \mathcal{E}\left(\sum_{k=1}^{n_i}\left[Y_{ik}(t)\left(\frac{r_{ik}^{(2)}(\beta,t)}{r_{ik}(\beta,t)}\right) r_{ik}(\beta_0,t)\right]\right). \end{aligned}$$

Our main results are given in Theorem 1-2 below, the regularity conditions and the proofs of which are given in the Appendix. We provide only brief remarks about the proofs below.

**Theorem 1.** *Under the regularity conditions in the Appendix, $\widehat{\beta}_E$ is a consistent estimator of $\beta_0$. Also, $n^{1/2}\left(\widehat{\beta}_E-\beta_0\right)$ is asymptotically normally distributed with mean zero and variance matrix in the form $\Sigma_E(\beta_0)=\Sigma^{-1}(\beta_0)\left[\Sigma_1(\beta_0)+\Sigma_2(\beta_0)\right]\Sigma^{-1}(\beta_0)'$, where*

$$\begin{aligned} \Sigma(\beta) &= -\int_0^\tau \left[\frac{s^{(1)}(\beta,t)^{\otimes 2}}{s^{(0)}(\beta,t)} - s^{(3)}(\beta,t)\right]\lambda_0(t)\,dt, \\ \Sigma_1(\beta) &= \mathcal{E}\left\{\sum_{k=1}^{n_i}(1-\eta_{ik})\, g_{ik}(\beta)\right\}^{\otimes 2}, \Sigma_2(\beta)=\mathcal{E}\left\{\sum_{k=1}^{n_i}\eta_{ik} h_{ik}(\beta)\right\}^{\otimes 2}, \end{aligned}$$

*with*

$$\begin{aligned} g_{ik}(\beta) &= \int_0^\tau \left[\begin{pmatrix} \frac{\phi_{ik}^{(1)}(\beta_1,t)}{\phi_{ik}(\beta_1,t)} \\ Z_{ik}(t) \end{pmatrix} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right] dM_{ik}(t), \\ h_{ik}(\beta) &= \int_0^\tau \left[\begin{pmatrix} E_{ik}(t) \\ Z_{ik}(t) \end{pmatrix} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right] dM_{ik}(t) - (1-\rho)\begin{pmatrix} Q_{ik}(\beta) \\ H_{ik}(\beta) \end{pmatrix}, \\ Q_{ik}(\beta) &= \int_0^\tau \frac{w_{ik}\left(t,A_{ik}^*(t)\right)}{Pr\left(\theta_i.\left(t,A_{ik}^*(t)\right)>0\right)}\left(\frac{\phi_{ik}^{(1)}(\beta_1,t)}{\phi_{ik}(\beta_1,t)} - \frac{s^{(11)}(\beta,t)}{s^{(0)}(\beta,t)}\right)\left(e^{\beta_1' E_{ik}} - \phi_{ik}(\beta_1,t)\right)\delta_{ik}^*(\beta,t)\,\lambda_0(t)\,dt \\ H_{ik}(\beta) &= \int_0^\tau \frac{w_{ik}\left(t,A_{ik}^*(t)\right)}{Pr\left(\theta_i.\left(t,A_{ik}^*(t)\right)>0\right)}\left(e^{\beta_1' E_{ik}} - \phi_{ik}(\beta_1,t)\right)\delta_{ik}^{**}(\beta,t)\,\lambda_0(t)\,dt, \\ \delta_{ik}^*(t,\beta) &= \mathcal{E}\left[\sum_{l=1}^{n_i}\left(Y_{il}(t)\, e^{\beta_2' Z_{il}(t)} I\left(A_{il}^*(t)=A_{ik}^*(t)\right)\right)\right], \\ \delta_{ik}^{**}(t,\beta) &= \mathcal{E}\left[\sum_{l=1}^{n_i} Y_{il}(t)\left(Z_{il}(t) - \frac{s^{(12)}(\beta,t)}{s^{(0)}(\beta,t)}\right) e^{\beta_2' Z_{il}(t)} I\left(A_{il}^*(t)=A_{ik}^*(t)\right)\right]. \end{aligned}$$

*where $s^{(11)}(\beta, t)$ and $s^{(12)}(\beta, t)$ equals the first derivative of $s^{(0)}(\beta, t)$ to $\beta_1$ and $\beta_2$ respectively.*

It is worth pointing out that when there is no individual subjecting to missing, the asymptotic variance $\Sigma_E$ of proposed estimator is equal to that of partial likelihood estimator in Lee et al. (1992). The proof of consistency of $\widehat{\beta}_E$ follows by the inverse Function Theorem (Foutz 1977). The asymptotic normality follows from the asymptotic normality of $n^{-1/2}\widehat{U}(\beta_0)$, a Taylor expansion and the Cramèr-Wold device. The asymptotic normality of

$n^{-1/2}\widehat{U}(\beta_0)$ is derived by multiple central limit theorem and results from empirical process theory.

To study the asymptotic properties of $\widehat{\Lambda}_0(t)$, we define the following metric space. Let $\mathcal{D}[0,\tau]$ be a space consisting of right-continuous functions $\{f(t)\}$ with left limits, where $f(t): [0, \tau] \to R$, make $\mathcal{D}[0,\tau]$ a metric space by equipping it with the metric $\rho(f,g) = \max_{t\in[0,\tau]} |f(t) - g(t)|$ for $f, g \in \mathcal{D}[0,\tau]$. The essential asymptotic results for the baseline cumulative hazard function estimator are summarized by the following theorem.

**Theorem 2.** *Under the regularity conditions in the Appendix, $\widehat{\Lambda}_0(t)$ converges in probability to $\Lambda_0(t)$ uniformly in $t \in [0,\tau]$, $\sqrt{n}\left(\widehat{\Lambda}_0(t) - \Lambda_0(t)\right)$ converges weakly to a zero mean Gaussian random field $\mathcal{W}(t)$ in $\mathcal{D}[0,\tau]$, the covariance function between $\mathcal{W}(s)$ and $\mathcal{W}(t)$ is C(s, t), where*

$$
\begin{aligned}
C(s,t)(\beta_0) &= E\left(\left[\sum_{k=1}^{n_i}(1-\eta_{ik})u_{ik}(s,\beta_0)\right]\times\left[\sum_{k=1}^{n_i}(1-\eta_{ik})u_{ik}(t,\beta_0)\right]\right)+E\left(\left[\sum_{k=1}^{n_i}\eta_{ik}v_{ik}(s,\beta_0)\right]\times\left[\sum_{k=1}^{n_i}\eta_{ik}v_{ik}(t,\beta_0)\right]\right),\\
u_{ik}(t,\beta_0) &= \int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0,s)} - \left[\int_0^t \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\lambda_0(s)\,ds\right]'\Sigma^{-1}(\beta_0)\,g_{ik}(\beta_0)\\
v_{ik}(t,\beta_0) &= \int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0,s)} - \left[\int_0^t \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\lambda_0(s)\,ds\right]'\Sigma^{-1}(\beta_0)\,h_{ik}(\beta_0) - (1-\rho)\int_0^t \frac{w_{ik}\left(s,A_{ik}^*(s)\right)\delta_{ik}^*(\beta_0,s)}{s^{(0)}(\beta_0,s)Pr\left(\theta_i.\left(t,A_{ik}^*(t)\right)>0\right)}\left[e^{\beta_{10}'E_{ik}(t)} - \phi_{ik}(\beta_{10},s)\right]\lambda_0(s)\,ds.
\end{aligned}
$$

The asymptotic variance can be consistently estimated by replacing the population quantities with its empirical counterparts.

## 4 Simulation Studies

Simulation studies are conducted to evaluate the finite sample performance of the proposed estimator and to compare the proposed methods with existing methods.

We generated failure time data from $n = 300$ clusters. We allow the cluster size $n_i$ to range from 1 to 6 with equal probability for each integer value in the range. The partially observed covariate $E$ is generated from a $n_i$ multivariate normal distribution $N_{ni}(0, V)$, where

$$
V = \begin{pmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{pmatrix}_{n_i\times n_i}.
$$

and $Z_{ik}$'s to be standard normal random variables.

For each cluster $i$, we use the method in Cai and Shen (2000) to generate the $n_i$ correlated failure times with $\lambda_0 = 1$, which is an extension of the commonly used multivariate failure time distribution of Clayton and Cuzick (1985). The joint survival function of the $n_i$ correlated failure times take the form:

$$
S(t_{i1},\cdots,t_{in_i}|Z_{i1},\cdots,Z_{in_i},E_{i1},\cdots,E_{in_i}) = \left\{\sum_{k=1}^{n_i}\exp\left(\theta^{-1}\lambda_0 t_{ik}e^{\beta_1'E_{ik}+\beta_2'Z_{ik}}\right) - (n_i-1)\right\}^{-\theta}. \tag{8}
$$

The positive parameter $\theta$ represents the intra-cluster association. $\theta$ is chosen to be 0.25, 1.5 and 5.7, which represents a strong, moderate and weak intra-cluster association, respectively. We assume an uniform distribution over $(0, \tau)$ for the censoring time, where $\tau$ = 5.96, 1.57 and 0.39 corresponding to censoring proportion of approximately 20%, 50% and 80%.

The auxiliary $A_{ij}$ is generated as follows: we first generate $W_{ij} = E_{ij} + e_{ij}$, where $e_{ij} \sim N(0, \sigma^2)$, then we assign $A_{ij}$ the value of 0, 1, 2, or 3 based on whether $W_{ij}$ is in the interval $(-\infty, a_1]$, $(a_1, a_2]$, $(a_2, a_3]$, or $(a_3, \infty)$, respectively, where $a_1$, $a_2$, $a_3$ are the 25%, 50%, 75% quantiles of $W$. Here $\sigma$ is the parameter that controls the strength of the association between $E_{ij}$ and $W_{ij}$, then between $E_{ij}$ and $A_{ij}$. Smaller $\sigma$ induces stronger correlation between $E$ and $A$. Individuals are selected into the validation set by Bernoulli sampling with equal probability. Simulation results are based on independent runs of 1000 for each data configuration.

We compare the proposed estimator $\widehat{\beta}_E$ with three alternative estimators: $\widehat{\beta}_F$, $\widehat{\beta}_V$ and $\widehat{\beta}_N$. The first two are standard partial likelihood estimators (solution of (5)) by using the full data and the validation data, respectively. $\widehat{\beta}_N$ is the standard partial likelihood using the auxiliary variable to replace the true exposure variable. In real data settings where $E$ is observed only for a validation set, $\widehat{\beta}_F$ can not be calculated.

Table 1 summarizes the simulation results for $\beta_1 = 0$ and 0.693 with validation fraction 30% and censoring rate 50%. We list the empirical mean (mean) and standard error (SD), average of estimated standard error (SE), the empirical coverage rate of nominal 95% confidence interval and the asymptotic relative efficiency (RE) with respect to the validation set. When $\beta_1 = 0$, all estimators are unbiased. When $\beta_1 \neq 0, \widehat{\beta}_N$ under estimated $\beta_1$, both $\widehat{\beta}_V$ and $\widehat{\beta}_E$ are approximately unbiased. For $\beta_1$, $\widehat{\beta}_E$ is more efficient than validation data estimator $\widehat{\beta}_V \left( SD_{\widehat{\beta}_E} < SD_{\widehat{\beta}_V} \right)$; $\widehat{\beta}_E$ is much more efficient when the auxiliary provides more information (*i.e.* smaller $\sigma$) or when the intracluster association is weaker (*i.e.* larger $\theta$); In all the simulated settings, the proposed estimator is not much less efficient than that from the full data case $\left(\widehat{\beta}_F\right)$. For $\beta_2$, $\widehat{\beta}_E$ is almost as efficient as the full data estimator under all settings we considered (results do not show here). The proposed estimated standard errors provide a very good estimate of the true variability of $\beta_1$ and $\beta_2$ and the corresponding 95% confidence intervals have reasonable coverage rates.

Table 2 summarizes the relative efficiency of $\widehat{\beta}_E$ to the validation estimator $\widehat{\beta}_V$ for $\beta_1$ under various validation fractions and censoring rate. We fix $\beta_1 = 0.693$, $\theta = 1.5$ and $\sigma = 0.1$. The relative efficiency increases when the censoring rate increases and when the validation fraction decreases. This suggests that, with low validation fraction and high censoring rate, our proposed estimator performs even more efficient when compared to the validation set method.

Furthermore, as suggested by the referees, we compare the proposed methods with the estimated pseudo-partial likelihood estimator (EPPLE, denoted by $\widehat{\beta}_L$ proposed by Liu et al (2009), who constructed the estimator based on the marginal hazard model with different baseline hazard function for different clusters. The failure time satisfies model (8) with fixed cluster size being $K = 2$ and baseline hazard being $\lambda_{01} = \lambda_{02} = 1$. We set the size of clusters as $n = 300$. The parameter settings are: $\beta_1 = 0$ and 0.693, $\beta_2 = -0.2$, $\theta = 0.25$ and 1.5, $\sigma = 0.2$. Table 3 show the results for the estimators of $\beta_1$ by listing the empirical mean (mean),

standard error (SD) and relative efficiency of $\widehat{\beta}_E$ with respect to $\widehat{\beta}_L$. As can be seen from Table 3, our proposed estimator $(\widehat{\beta}_E)$ for $\beta_1$ tends to be a little more efficient than that of Liu et al. $(\widehat{\beta}_L)$ this is natural since we use more information to estimate the relative risk in the proposed method. The relative efficiency of $\widehat{\beta}_E$ to $\widehat{\beta}_L$ becomes smaller for larger $\theta$. For the estimator of $\beta_2$, both these two methods are as efficient as the partial likelihood estimator based on the full data (results are not listed here).

We also conducted some simulation studies to test the robustness of our approach. The failure times are generated from marginal hazard model with different baseline functions:

$$\lambda_{ik}(t;E_{ik},Z_{ik})=\lambda_{0k}\exp\{\beta_1 E_{ik}+\beta_2 Z_{ik}\},\quad k=1,2,i=1,\cdots,n.$$

with $\beta_1 = 0.693$, $\beta_2 = -0.2$ and $n = 300$. The censoring rate is around 50% and the validation fraction is set to 30%. We fix one of the baseline $\lambda_{01} = 1$ and $\lambda_{02}$ varies from 1 to 2.4 with jump 0.2. The working model remains to be marginal hazard model with common baseline function. Table 4 listed the results. It can be seen that when the working model are not too far away from the true model (e.g. $\lambda_{02} \leq 2$), the proposed estimator still works well.

## 5 Analysis of Framingham study

We illustrate our proposed method to estimate the effiect of cholesterol level on the risk of Coronary Heart disease (CHD) using a data set from the Framingham study (Dawber 1980). The Framingham Heart Study was the first prospective study on cardiovascular disease. The study began in 1948 in the United States. Participants from the town of Framingham, Massachusetts were randomly sampled. The full cohort includes 2336 men and 2873 women aged between 30 and 62 years. Examination of participants has taken place every two years and the patients were followed for morbidity and mortality. Since the primary sampling unit was the family, failure times are likely to be correlated for the individuals within a family.

For simplicity, our analysis consists of data for patients who had no history of hyper-tension or glucose intolerance and no previous experience of coronary heart disease or a cerebrovascular accident around age “45”. The data we used consists of 1571 patients from 1401 families, among which 1261 families have only 1 patient, 113 families have 2 patients, 24 families have 3 patients and 3 families have 4 patients. Among the full cohort, 250 patients experienced CHD. The time is originated at age “45”(Age45) and the follow-up information is up to the year 1980. In our analysis, the failure time was defined as the time from Age45 to the onset of CHD, and all observations were censored either at the time loss to follow up, or at the end of the study.

In addition to the cholesterol level (Chol45), as the exposure variable of interest, other potential confounding variables available for all subjects under study include age at first exam Framingham (Agev1), body mass index (Bmi45), systolic blood pressure (Sbp45), gender(Sex, 1 for female and 0 for male), waiting time from first exam to age “45” (Wait), smoking status (Smoke, 0= no, 1=yes). Since the patients were clustered by family and the family members are exchangeable within each family, therefore it is reasonable to assume marginal hazard model (2) with common baseline function. We consider model (2) that include the above mentioned seven risk factors at Age45 as the covariates.

In the Framingham study, the covariates were measured for all patients, and therefore, this provide us an opportunity to evaluate our proposed method using various validation sampling fractions against not only a validation set analysis but also a full data analysis.

Measurement of the cholesterol level (Chol45) is one example of a variable that maybe moderately expensive to obtain and therefore represents a candidate main exposure variable which is observed in a sub-cohort. In terms of practical consideration, smokers usually have higher Chol45. Hence, we use the smoking status as the auxiliary variable for Chol45. We consider all seven covariates in the model. The fitted model is:

$$\lambda_{ik}(t|Z_{ik}(t), E_{ik}(t)) = \lambda_0(t) \exp\left\{\beta_1 Chol45_{ik}(t) + \beta_2' Z_{ik}\right\},$$

where $Z$ = (*Agev*1, *Bmi*45, *Sbp*45, *gender, Wait, Smoke*).

We sampled a sequence of validation sets, with validation fraction $\rho$ being 10%, 20%, 30% and 40%, from the full cohort of 1571 patients and analyzed them using the proposed method by assuming that the main covariate, cholesterol level, is only available for the validation set.

Table 5 listed the results from the Framingham study for the factor “chol45”. It is noted that the cholesterol level is significant in the full data analysis (95% CI: [0.001, 0.007], p-value: 0.007). Comparing the proposed method and the validation method, we see that at small validation fractions ($\rho$ < 40%), the proposed estimator does not achieve the significance of testing $\beta_{chol45} = 0$. Nevertheless, the proposed method is approaching the significant level of the full cohort analysis as we increase the validation fraction. At $\rho$ = 40%, the proposed method also reject the null hypothesis that $\beta_{chol} = 0$ at 0.05 significance level. Further inspection of Table 5 also reveals that the validation set analysis consistently produced smaller Z-scores than the proposed estimator and hence always yielded a larger p-value in testing $\beta_{chol45} = 0$. At $\rho$ = 40%, the validation estimator did not achieve the significance level of the full analysis or the proposed estimator.

Table 6 summarizes the results for all the factors in the Framingham study using the three methods with 615 ($\rho$ = 40%) sample as the validation set. The p-values indicate that, at 0.05 significance level, Chol45, Bmi45, Sbp45, Sex and Smoking status are all statistically significant for CHD using the proposed method, which is the same as the full data analysis. However, only Sbp45 and Sex are significant for CHD for the validation set analysis. The proposed estimates appeared to be closer to the full data analysis, and is more efficient than the validation set estimator. The standard error of the proposed estimator for all the covariates is similar at this $\rho$ level with that of the full data estimator.

This example confirmed that the proposed estimator is a more precise estimator. One would have improved the statistical power that would have been lost if one were only to analyze the validation set data without incorporating the auxiliary information.

## 6 Concluding Remarks

In this paper, we proposed an estimated likelihood approach for CBM model, where the main exposure is partly observed and a discrete auxiliary variable for the main exposure is available. An estimating equation based on the pseudo-partial likelihood is proposed. Our approach is nonparametric with respect to the association between the missing covariate and the observed auxiliary covariate. The proposed estimators are shown to be consistent and asymptotically normal. The theoretical proof is nontrivial because the classical martingale theory is not sufficient. Instead, we rely on the results from modern empirical process theory. Simulation studies and real data example demonstrate that the proposed method works well in moderate-size sample and shows an improved statistical efficiency over what would be achieved using only the validation set. Simulation studies also show that the

statistical efficiency of the proposed method also depend on the validation fraction. Sampling more individuals result in more efficient proposed estimators.

We have a few recommendations on the applications of the proposed method. First, one can discretize a continuous auxiliary variable into categories and then apply the proposed method. To fully take advantage of a continuous auxiliary covariate, a nonlinear smoothing version of equation (6) will need to be developed. Secondly, the number of categories of the auxiliary variable cannot be too large (*e.g.* no more than 6) if the validation sample size is small (< 60). Additional simulations showed that there could be convergence problems when the validation size is less than 60 and the number of categories is greater than 6. We recommend to reduce the number of categories of the auxiliary variable if the sample is small. Thirdly, the estimating equation of Lee, et al (1992) did not take into consideration of the potential correlation in the multivariate failure times. Cai and Prentice (1995) and Xue et al.(2010) showed that more efficient $\beta$-estimators could be obtained by introducing weights into the estimating equations for small and large cluster size respectively. In modeling panel count data, which involves taking account of the dependence of the successive counts, Wellner and Zhang (2000) showed that the full non-parametric maximum likelihood estimator (NPMLE) improved the study efficiency compared to the pseudo likelihood estimator which ignores the potential correlation between counts. Therefore, introducing suitable weights to our proposed equation could further improve efficiency. Future work that improve the efficiency of estimators is certainly warranted.

## Acknowledgments

The authors are very grateful for the valuable comments and suggestions from the editor and the two referees. This work was partly supported the National Natural Science Fund of China grant 11171263 (Liu and Yuan) and U.S. NIH grants R01 HL 57444 (Cai) and U.S. NIH grants R01 CA 79949 (Zhou).

## APPENDIX

## Assumptions and Outline of the Proofs of Theorem 1 and 2

We assume that the following conditions hold:

### Conditions

(C1) (Finite interval): $\int_0^\tau \lambda_0(t)\,dt<\infty$;

(C2) $Pr(\tau_j.(t, a_m) > 0) > 0$;

(C3) For any $k=1,\cdots,n_i, i=1,\cdots,n, \left(E_{ik}(t), Z_{ik}(t), A^*_{ik}(t)\right)'$ has uniformly bounded variation almost surely over $[0, \tau]$;

(C4) For $d = 0, 1, 2$, there exits a neighborhood $\mathcal{B}=\mathcal{B}_1 \times \mathcal{B}_2$ of $\beta_0$ such that $s^{(d)}(\beta, t)$ are continuous function of $\beta$, uniformly in $t \in (0, \tau]$, bounded on $\mathcal{B} \times (0, \tau]$. $s^{(0)}(\beta, t)$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$ and $\Sigma(\beta_0)$ is positive definite.

(C5) For $d=0, 1, \quad \sup_{t\in[0,\tau]} |L^{(d)}(t)|=O_p(1)$, where

$$L^{(d)}(t)=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{l=1}^{n_j} w_{jl}(t, a_m)\left[e^{\beta_1' E_{jl}(t)}E_{jl}^{\otimes d}(t) - \mathcal{E}\left(e^{\beta_1' E_{jl}(t)}E_{jl}^{\otimes d}(t)\,|\,Y_{jl}(t)=1, A^*_{jl}(t)=a_m\right)\right].$$

The following lemma (Lemma 4.2, p. 54) in Kosorok (2008) will be often used in proving the theorem:

**Lemma A1.** *$B_n \in D[a, b]$ and $A_n \in l^\infty[a, b]$ be either cadlag or caglad, where $l^\infty[a, b]$ is the collection of all bounded functions on $[a, b]$, and assume* $\sup_{t\in[0,\tau]} |A_n(t)| \to_p 0$. $A_n$ has uniformly bounded variation and $B_n$ converges weakly to a tight, mean zero process with sample paths in $D[a, b]$. *Then* $\int_a^b A_n(s)\, dB_n(s) \to_p 0$.

Define $\phi_{a_m}(\beta_1, t) = E\left(e^{\beta_1' E_{jk}(t)} | Y_{jk}(t) = 1, A_{jk}^* = a_m\right)$. For $d$=0, 1, 2, $\phi_{a_m}^{(d)}(\beta_1, t)$ be the $d$-th derivative of $\phi_{am}(\beta_1, t)$ respect to $\beta_1$, $\phi_{a_m}^{(0)}(\beta_1, t) = \phi_{a_m}(\beta_1, t)$. Define $b^{\otimes 2} = bb'$, $\| b \| = \sup_{1\le i\le p} |b_i|$ for a vector $b = (b_1, \cdots, b_p)'$ and $\| B \| = \sup_{i,j} |b_{ij}|$ for a matrix $B = (b_{ij})$. The following asymptotic property plays important role in proving the theorems.

**Lemma A2.** *For $m = 1, \cdots, L$, $d = 0, 1, 2$,*

$$\sup_{B\times[0,\tau]} \| \widehat{\phi}_{a_m}^{(d)}(\beta_1, t) - \phi_{a_m}^{(d)}(\beta_1, t) \| \to_p 0 \tag{A1}$$

Proof: For $d = 0$,since $I\left(\bar{\phi}_{j,a_m}(\beta_1, t) > 0\right) = I\left(\theta_{j\cdot}(t, a_m) > 0\right)$, we have

$$\widehat{\phi}_{a_m}(\beta_1, t) = \frac{\frac{1}{n}\sum_{j=1}^n I\left(\theta_{j\cdot}(t, a_m) > 0\right)\bar{\phi}_{j,a_m}(\beta_1, t)}{\frac{1}{n}\sum_{j=1}^n I\left(\theta_{j\cdot}(t, a_m) > 0\right)}.$$

Therefore, the nominator of $\widehat{\phi}_{a_m}(\beta_1, t) - \phi_{a_m}(\beta_1, t)$ equals :

$$\frac{1}{n}\sum_{j=1}^n \sum_{l=1}^{n_i} w_{jl}(t, a_m)\left(e^{\beta_1' E_{jl}(t)} - \phi_{a_m}(\beta_1, t)\right),$$

which is $o_p(1)$ by condition (C5). Combine condition (C2), we can prove (A1) for $d = 0$. The same argument works for $d = 1, 2$.

Define

$$
\begin{aligned}
\widehat{S}^{(d)}(\beta,t) &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{n_i} Y_{ik}\widehat{r}_{ik}^{(d)}(\beta,t),\ (d=0,1,2)\\
\widehat{S}^{(3)}(\beta,t) &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{n_i} Y_{ik}(t)\left(\frac{\widehat{r}_{ik}^{(1)}(\beta,t)}{\widehat{r}_{ik}(\beta,t)}\right)^{\otimes 2} r_{ik}(\beta_0,t)\\
\widehat{S}^{(4)}(\beta,t) &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{n_i} Y_{ik}(t)\left(\frac{\widehat{r}_{ik}^{(2)}(\beta,t)}{\widehat{r}_{ik}(\beta,t)}\right) r_{ik}(\beta_0,t)\\
\Delta\widehat{r}_{ik}(\beta,t) &= \frac{\widehat{r}_{ik}^{(2)}(\beta,t)}{\widehat{r}_{ik}(\beta,t)} - \left(\frac{\widehat{r}_{ik}^{(1)}(\beta,t)}{\widehat{r}_{ik}(\beta,t)}\right)^{\otimes 2},\\
\Delta r_{ik}(\beta,t) &= \frac{r_{ik}^{(2)}(\beta,t)}{r_{ik}(\beta,t)} - \left(\frac{r_{ik}^{(1)}(\beta,t)}{r_{ik}(\beta,t)}\right)^{\otimes 2},\\
\Delta\widehat{S}(\beta,t) &= \frac{\widehat{S}^{(2)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)} - \left(\frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}\right)^{\otimes 2},\\
\Delta s(\beta,t) &= \frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} - \left(\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right)^{\otimes 2},
\end{aligned}
$$

Similar as the argument by Liu et al.(2009), we can prove that, for $d$ = 0, 1, 2, 3, 4,

$$\sup_{B\times[0,\tau]} \|\widehat{S}^{(d)}(\beta,t) - s^{(d)}(\beta,t)\| \to_p 0, \tag{A2}$$

$$\sup_{B\times[0,\tau]} \|\Delta\widehat{S}(\beta,t) - \Delta s(\beta,t)\| \to_p 0. \tag{A3}$$

Since $M_{ik}(t)$ is a Donsker class and $n^{-1/2}\sum_{i=1}^{n} M_{ik}(t)$ converges weakly to a tight, mean zero process, we can prove the following useful property by (A3), condition (C4) and Lemma A1,

$$n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \Delta\widehat{S}(\beta,t)\,dM_{ik}(t) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \Delta s(\beta,t)\,dM_{ik}(t) + o_p(1). \tag{A4}$$

Before we prove theorems, we prove the asymptotic normality of $n^{-\frac{1}{2}}\widehat{U}(\beta_0)$ in the following lemma:

**Lemma A3.** *Under conditions* $(C1)-(C5)$, $n^{-\frac{1}{2}}\widehat{U}(\beta_0)$ converges to a mean zero Normal distribution with covariance $\Sigma_1(\beta_0) + \Sigma_2(\beta_0)$.

Proof: By simple algebraic manipulation, we have

$$n^{-1/2}\widehat{U}(\beta) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \left[\frac{\widehat{r}_{ik}^{(1)}(\beta,t)}{\widehat{r}_{ik}(\beta,t)} - \frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}\right] dM_{ik}(t) + n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \left[\frac{\widehat{r}_{ik}^{(1)}(\beta,t)}{\widehat{r}_{ik}(\beta,t)} - \frac{\widehat{S}^{(1)}(\beta,t)}{\widehat{S}^{(0)}(\beta,t)}\right] r_{ik}(\beta_0,t)\,Y_{ik}(t)\,\lambda_0($$

(A5)

By (A1)-(A3), we can show that the first term of (A5) evaluated at $\beta_0$ equals

$$n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^{\tau}\left[\frac{r_{ik}^{(1)}(\beta_0,t)}{r_{ik}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right]dM_{ik}(t)+o_p(1).$$

Define $S^{(d)}(\beta, t)$ as the corresponding functions with $r_{ik}(\beta, t)$ substituted for $\widehat{r}_{ik}(\beta,t)$ in $\widehat{S}^{(d)}(\beta,t)$, $d$=0, 1, ⋯ , 4. For the second term of (A5) evaluated at $\beta_0$, we apply the first order expansion $x/y=x_0/y_0+(x-x_0)/y_0-(y-y_0)\,x_0/y_0^2+o\left\{(x-x_0)^2+(y-y_0)^2\right\}$ to $\widehat{r}_{ik}^{(1)}/\widehat{r}_{ik}$ and $\widehat{S}^{(1)}/\widehat{S}^{(0)}$ at $r_{ik}^{(1)}/r_{ik}$ and $S^{(1)}/S^{(0)}$, respectively, we can rewrite the second term of (A5) evaluated at $\beta_0$ equals asymptotically:

$$\begin{aligned}
&-n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^{\tau}\left[\frac{r_{ik}^{(1)}(\beta_0,t)}{r_{ik}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right]Y_{ik}(t)\left(\widehat{r}_{ik}(\beta_0,t)-r_{ik}(\beta_0,t)\right)\lambda_0(t)\,dt\\
&\quad=-n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{k=1}^{n_i}(1-\eta_{ik})\int_0^{\tau}\left[\begin{pmatrix}\frac{\phi_{ik}^{(1)}(\beta_{10},t)}{\phi_{ik}(\beta_{10},t)}-\frac{s^{(11)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\\ Z_{ik}(t)-\frac{s^{(12)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\end{pmatrix}\right]\\
&\quad\times Y_{ik}(t)\,e^{\beta_{20}'Z_{ik}(t)}\left(\widehat{\phi}_{ik}(\beta_{10},t)-\phi_{ik}(\beta_{10},t)\right)\lambda_0(t)\,dt\\
&\quad=\begin{pmatrix}\Psi^{(1)}(t)\\ \Psi^{(2)}(t)\end{pmatrix},
\end{aligned}\tag{A6}$$

where $\Psi^{(1)}(t)$ and $\Psi^{(2)}(t)$ are defined as the first line and second line of (A6).

By condition (C1)-(C5) and the definition of $\widehat{\phi}_{ik}(\beta_{10},t)$, $\Psi^{(1)}(t)$ can be rewritten as

$$\begin{aligned}
\Psi^{(1)}(t)=&-n^{-1/2}\int_o^{\tau}\sum_{m=1}^{L}\left(\frac{\phi_{a_m}^{(1)}(\beta_{10},t)}{\phi_{a_m}(\beta_{10},t)}-\frac{s^{(11)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)\left(\widehat{\phi}_{a_m}(\beta_{10},t)-\phi_{a_m}(\beta_{10},t)\right)\sum_{i=1}^{n}\sum_{k=1}^{n_i}Y_{ik}(t)\,e^{\beta_{20}'Z_{ik}(t)}(1-\eta_{ik})\,I\left(A_{ik}^{*}(t)=a_m\right)\lambda_0(t)\,dt\\
=&-\int_0^{\tau}\sum_{m=1}^{L}\left(\frac{\phi_{a_m}^{(1)}(\beta_{10},t)}{\phi_{a_m}(\beta_{10},t)}-\frac{s^{(11)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)\frac{1}{Pr\left(\theta_{j.}(t,a_m)>0\right)}\\
&\times\sqrt{n}\frac{1}{n}\sum_{j=1}^{n}\left[\bar{\phi}_{j,a_m}(\beta_1,t)-\phi_{a_m}(\beta_{10},t)\right]I\left(\theta_{j.}(t,a_m)>0\right)\\
&\times\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n_i}Y_{ik}(t)\,e^{\beta_{20}'Z_{ik}(t)}I\left(A_{ik}^{*}(t)=a_m\right)(1-\eta_{ik})\,\lambda_0(t)\,dt\\
+o_p(1)=&-\int_0^{\tau}\sum_{m=1}^{L}\left(\frac{\phi_{a_m}^{(1)}(\beta_{10},t)}{\phi_{a_m}(\beta_{10},t)}-\frac{s^{(11)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)\frac{1-\rho}{Pr\left(\theta_{j.}(t,a_m)>0\right)}\\
&\times\mathcal{E}\left(\sum_{k=1}^{n_i}Y_{ik}(t)\,e^{\beta_{20}'Z_{ik}(t)}I\left(A_{ik}^{*}(t)=a_m\right)\right)\\
&\times n^{-1/2}\sum_{j=1}^{n}\sum_{l=1}^{n_j}w_{jl}(t,a_m)\left(e^{\beta_1'E_{jl}(t)}-\phi_{a_m}(\beta_{10},t)\right)\lambda_0(t)\,dt\\
+o_p(1)=&-n^{-1/2}(1-\rho)\sum_{j=l}^{n}\sum_{l=1}^{n_j}Q_{il}(\beta_0)+o_p(1)
\end{aligned}$$

where $Q_{jl}(\beta_0)$ is defined as in Theorem 1. Similarly, we can prove that $\Psi^{(2)}(t)$ is asymptotically $-n^{-\frac{1}{2}}\sum_{j=1}^{n}\sum_{l=1}^{n_j}(1-\rho)H_{jl}(\beta_0)$ where $H_{jl}(\beta_0)$ is defined as in Theorem 1.

It follows that $n^{-\frac{1}{2}}\widehat{U}(\beta_0)$ is asymptotically equivalent to

$$n^{-\frac{1}{2}}\sum_{j=1}^{n}\sum_{l=1}^{n_j}\left(1-\eta_{jl}\right)\int_0^\tau\left(\frac{r_{jl}^{(1)}(\beta_0,t)}{r_{jl}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{10}(\beta_0,t)}\right)dM_{jl}(t) \tag{A7}$$

$$+n^{-\frac{1}{2}}\sum_{j=1}^{n}\sum_{l=1}^{n_j}\eta_{jl}\left[\int_0^\tau\left(\frac{r_{jl}^{(1)}(\beta_0,t)}{r_{jl}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)dM_{jl}(t)-(1-\rho)\begin{pmatrix}Q_{jl}(\beta_0)\\ H_{jl}(\beta_0)\end{pmatrix}\right]. \tag{A8}$$

(A7) and (A8) are independent. By martingale central limit theorem, (A7) converges weekly to a continuous normal process with covariance $\Sigma_1(\beta_0)$. (A8) is a summation of iid terms from the validation sample. By central limit theorem, it converges to a normal distribution with mean

$$\mathcal{E}\left(n^{-\frac{1}{2}}\sum_{l=1}^{n_j}\eta_{jl}\left[\int_0^\tau\left(\frac{r_{jl}^{(1)}(\beta_0,t)}{r_{jl}(\beta_0,t)}-\frac{s(1)(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)dM_{jl}(t)-(1-\rho)\begin{pmatrix}Q_{jl}(\beta_0)\\ H_{jl}(\beta_0)\end{pmatrix}\right]\right) \tag{A9}$$

and covariance

$$Var\left(\sum_{l=1}^{n_j}\eta_{jl}\left[\int_0^\tau\left(\frac{r_{jl}^{(1)}(\beta_0,t)}{r_{jl}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)dM_{jl}(t)-(1-\rho)\begin{pmatrix}Q_{jl}(\beta_0)\\ H_{jl}(\beta_0)\end{pmatrix}\right]\right)$$

Since $\eta_{ik}$ and $n_i$ are independent of covariates $\{N_{ik}(\cdot), Z_{ik}(\cdot)\}$ and $\int_0^\tau\left(\frac{r_{ik}^{(1)}(\beta_0,t)}{r_{ik}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)dM_{ik}(t)$ is a local martingale, we have the expected value of the first term in the mean expression (A9) is 0. It is easy to show that $\mathcal{E}\left(Q_{jl}(\beta_0)\right)=0$ and $\mathcal{E}\left(H_{jl}(\beta_0)\right)=0$. Therefore the second term is 0.

The covariance matrix can be expressed as $\Sigma_2(\beta_0$, which is defined in Theorem1. Therefore the limiting distribution of $n^{-1/2}\widehat{U}(\beta_0)\to_D N(0,\Sigma_1(\beta_0)+\Sigma_2(\beta_0))$

Proof of theorem 1:

**(1) Consistency—**To prove the consistency of $\widehat{\beta}_E$, we use the Inver Function Theorem (Foutz 1977) by verifying the following conditions: (I) $n^{-1}\partial\widehat{U}(\beta)/\partial\beta$ exists and is continuous in an open neighborhood $\mathcal{B}$ of $\beta_0$; (II) $n^{-1}\partial\widehat{U}(\beta_0)/\partial\beta_0$ is negative definite with probability going to 1; (III) $n^{-1}\partial\widehat{U}(\beta)/\partial\beta$ converges in probability to $A(\beta)$, uniformly for $\beta$ in an open neighborhood of $\beta_0$; (IV) $n^{-1}\widehat{U}(\beta_0)\to 0$ in probability.

Clearly (I) is satisfied due to the continuity of $\widehat{r}_{ik}^{(d)}(\beta,t)$ and $\widehat{S}^{(d)}(\beta,t)$.

Similar as Andersen and Gill (1982), we can decompose $n^{-1}\partial\widehat{U}(\beta)/\partial\beta$ into several parts:

$$n^{-1}\partial\widehat{U}(\beta)/\partial\beta = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \left[\Delta\widehat{r}_{ik}(\beta,t) - \Delta\widehat{S}(\beta,t)\right] dM_{ik}(t) + \widehat{A}(\beta), \tag{A10}$$

where $\widehat{A}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \left[\Delta\widehat{r}_{ik}(\beta,t) - \Delta\widehat{S}(\beta,t)\right] Y_{ik}(t)\, r_{ik}(\beta_0,t)\,\lambda_0(t)\,dt.$

By (A1)-A(4), we can prove that the first term of the right side of (A10) equals asymptotically

$$n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\tau \left[\Delta r_{ik}(\beta,t) - \Delta s(\beta,t)\right] dM_{ik}(t),$$

which is martingale and converges to zero in probability by Lenglart inequality.

By condition (C1) and (A1), we can prove that $\widehat{A}(\beta)$ converges in probability to

$$A(\beta) = \int_0^\tau \left[ s^{(4)}(\beta,t) - \frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} s^{(0)}(\beta_0,t) - s^{(3)}(\beta,t) + \left(\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right)^{\otimes 2} s^{(0)}(\beta_0,t)\right]\lambda_0(t)\,dt.$$

When $\beta = \beta_0$, we have $s^{(4)}(\beta_0, t) = s^{(2)}(\beta_0, t)$, and $A(\beta_0) = -\Sigma(\beta_0)$ is negative by condition (C4). Thus (II) and (III) are satisfied.

Using the result in the proof of Lemma A3, (IV) hold by Chebyshev's inequality. Having now verified (I)-(IV), we conclude that $\widehat{\beta}_E$ converges in probability to $\beta_0$.

**(2) Asymptotic Normality—**By a Taylor expansion of $\widehat{U}(\beta_0)$ with respect to $\beta$ and around $\beta_0$, we have

$$n^{-\frac{1}{2}}\widehat{U}(\beta_0) = \left\{-n^{-1}\frac{\partial}{\partial\beta^T}\widehat{U}(\beta^*)\right\} n^{\frac{1}{2}}\left(\widehat{\beta}_E - \beta_0\right), \tag{A11}$$

where $\beta^*$ is between $\widehat{\beta}_E$ and $\beta_0$.

By (III) and the asymptotical normality of $n^{-\frac{1}{2}}\widehat{U}(\beta_0)$, we proved Theorem 1.

**Proof of theorem 2**: Note that

$$\sup_{t\in[0,\tau]}|\widehat{\Lambda}_0(t) - \Lambda_0(t)| \le \sup_{t\in[0,\tau]}\left|\int_0^t \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n_i} dM_{ik}(s)}{\widehat{S}^{(0)}\left(\widehat{\beta}_E, s\right)}\right| + \sup_{t\in[0,\tau]}\left|\int_0^t \frac{S^{(0)}(\beta_0,s) - \widehat{S}^{(0)}\left(\widehat{\beta}_E,s\right)}{\widehat{S}^{(0)}\left(\widehat{\beta}_E,s\right)}\lambda_0(s)\,ds\right|.$$

With the consistency of $\widehat{\beta}_E$, we can show the first term on the right-hand side of the above inequality converges to zero by (A2) and the martingale properties, and the second term is also asymptotically negligible by (A1). Then we prove the uniform consistency of $\widehat{\Lambda}_0(t)$.

We can decompose $\sqrt{n}\left(\widehat{\Lambda}_0(t) - \Lambda_0(t)\right)$ into the following three parts:

$$\sqrt{n}\left(\widehat{\Lambda}_0(t) - \Lambda_0(t)\right) = \frac{1}{\sqrt{n}}\int_0^t \frac{\sum_{i=1}^{n}\sum_{k=1}^{n_i} dM_{ik}(s)}{\widehat{S}^{(0)}\left(\widehat{\beta}_E, s\right)} + \sqrt{n}\int_0^t \frac{\widehat{S}^{(0)}(\beta_0, s) - \widehat{S}^{(0)}(\beta_E, s)}{\widehat{S}^{(0)}\left(\widehat{\beta}_E, s\right)}\lambda_0(s)\,ds + \sqrt{n}\int_0^t \frac{S^{(0)}(\beta_0, s) - \widehat{S}^{(0)}(\beta_0, s)}{\widehat{S}^{(0)}\left(\widehat{\beta}_E, s\right)}\lambda_0(s)\,ds$$

By (A2), the first term equals $n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0, s)} + o_p(1)$. The second term is equal to $-\left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)}\lambda_0(s)\,ds\right]' \sqrt{n}\left(\widehat{\beta}_E - \beta_0\right) + o_p(1)$ by Taylor expansion and (A1), and to

$$-\left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)}\lambda_0(s)\,ds\right]' \Sigma^{-1}(\beta_0)\, n^{-\frac{1}{2}}\widehat{U}(\beta_0) + o_p(1)$$

by (A11) and (III).

By similar arguments as in the approximation proof of $\Psi^{(1)}(t)$, we can show the third term is asymptotically equal to

$$-n^{-1/2}(1-\rho)\sum_{i=1}^{n}\sum_{k=1}^{n_i}\eta_{ik}\int_0^\tau \frac{w_{ik}\left(s, A_{ik}^*(s)\right)\delta_{ik}^*(\beta_0, s)}{s^{(0)}(\beta_0, s)\,Pr\left(\theta_i.\left(t, A_{ik}^*(t)\right) > 0\right)}\left[e^{\beta_{10}' E_{ik}(t)} - Q_{ik}(\beta_{10}, s)\right]\lambda_0(s)\,ds.$$

Thus

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\Lambda}_0(t) - \Lambda_0(t)\right) \\
&\quad = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^t \frac{dM_{ik}(s)}{s^{(0)}(\beta_0, s)} \\
&\quad - \left[\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)}\lambda_0(s)\,ds\right]' \Sigma^{-1}(\beta_0)\, n^{-\frac{1}{2}}\widehat{U}(\beta_0) \\
&\quad - n^{-1/2}(1-\rho)\sum_{i=1}^{n}\sum_{k=1}^{n_i}\eta_{ik}\int_0^t \frac{w_{ik}\left(s, A_{ik}^*(s)\right)\delta_{ik}^*(\beta_0, s)}{s^{(0)}(\beta_0, s)\,Pr\left(\theta_i.\left(t, A_{ik}^*(t)\right) > 0\right)}\left[e^{\beta_{10}' E_{ik}(t)} - \phi_{ik}(\beta_{10}, s)\right]\lambda_0(s)\,ds + o_p(1).
\end{aligned}$$

Therefore, we can rewrite the above sums into two independent items as

$$\sqrt{n}\left(\widehat{\Lambda}_0(t) - \Lambda_0(t)\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{k=1}^{n_i}(1-\eta_{ik})\,u_{ik}(t, \beta_0) + n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{k=1}^{n_i}\eta_{ik}v_{ik}(t, \beta_0) + o_p(1),$$

where $u_{ik}(t, \beta_0)$ and $v_{ik}(t, \beta_0)$ are defined as in theorem 2. We can easily show $E(u_{ik}(t, \beta_0)) = 0$ and $E(v_{ik}(t, \beta_0)) = 0$. By multivariate central limit theorem, Theorem 2 can be proved.

## References


Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Anna Stat. 1982; 10:1100–1120.

Cai J, Prentice RL. Estimating equations for hazard ratio parameters based on correlated failure time data. Biometrika. 1995; 82:151–164.

Cai J, Prentice RL. Regression analysis for correlated failure time data. Lifetime Data Analysis. 1997; 3:197–213. [PubMed: 9384652]

Cai J, Shen Y. Permutation tests for comparing marginal survival functions with clustered failure time data. Stat Med. 2000; 19:2963–2973. [PubMed: 11042626]

Clayton D, Cuzick J. Multivariate generalizations of the proportional hazard model (with discussion). J R Stats Soc, Series A. 1985; 54:168–184.

Cox DR. Regression Models and Life-Tables. J R Stats Soc, Series B. 1972; 62:187–202.

Dawber, TR. The Epidemiology of Atherosclerotic Disease. Harvard University Press; Cambridge, MA: 1980. The Framingham Study..

Foutz RV. On the unique consistent solution to the likelihood equations. J Am Stats Assoc. 1977; 72:147–148.

Gorfine M, Zucker DM, Hsu L. Prospective survival analysis with a general semi-parametric shared frailty model: A pseudo full likelihood approach. Biometrika. 2006; 93:735C741.

Hu C, Lin DY. Cox Regression with Covariate Measurement Error. Scand J Stat. 2002; 29:637–655.

Hougaard, P. Analysis of Multivariate Survival Data. Springer-Verlag; New York: 2000.

Jiang J, Zhou H. Additive hazards regression with auxiliary covariates. Biometrika. 2007; 94:359–369.

Kalbfleisch, JD.; Prentice, RL. The statistical analysis of failure time data. 2nd ed.. Wiley; New York: 2002.

Kosorok, MR. Introduction to empirical process and semiparametric inference. Springer; New York: 2008.

Kulich M, Lin DY. Additive hazards regression with covariate measurement error. Journal of the American Statistical Association. 2000; 95:238–248.

Lee, EW.; Wei, LJ.; Amato, DA. Cox-type regression analysis for large numbers of small groups of correlated failure time observations.. In: Klein, JP.; Goel, PK., editors. Survival Analysis: State of the Art. Kluwer Academic Publishers; Netherlands: 1992. p. 237-247.

Lin DY, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994; 81:61–71.

Liu Y, Wu Y, Zhou H. Multivariate failure time regression with a continuous auxiliary covariates. Journal of Multivariate Analysis. 2010; 101:679–691. [PubMed: 21966052]

Liu Y, Zhou H, Cai J. Estimated pseudo-partial-likelihood method for correlated failure time data with auxiliary covariates. Biometrics. 2009; 65:1184–1193. [PubMed: 19432779]

Nielsen GG, Gill RD, Andersen PK, Sørensen TI. A counting process approach to maximum likelihood estimation in frailty models. Scandinavian Journal of Statistics. 1992; 19:25–43.

Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. Journal of the American Statistical Association. 1998; 93:1164–1175.

Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. Biometrics. 1997; 53:131–145. [PubMed: 9147589]

Wellner JA, Zhan Y. Two estimators of the mean of a counting process with panel count data. Annals of Statistics. 2000; 28:779C814.

Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association. 1989; 84:1065–1073.

Xue L, Wang L, Qu A. Incorporating Correlation for Multivariate Failure Time Data When Cluster Size Is Large. Biometrics. 2010; 66:393–404. [PubMed: 19673860]

Zhou H, Pepe MS. Auxiliary covariate data in failure time regression analysis. Biometrika. 1995; 58:352–360.

Zhou H, Wang CY. Failure time regression with continuous covariates measured with error. J R Stats Soc, Series B. 2000; 62:657–665.