NIH Public Access
Author Manuscript
Biometrics. Author manuscript; available in PMC 2010 December 1.

Published in final edited form as:
Biometrics. 2009 December ; 65(4): 1184. doi:10.1111/j.1541-0420.2009.01198.x.

Estimated Pseudo-Partial-Likelihood Method for Correlated Failure Time Data with Auxiliary Covariates

Yanyan Liu,
School of Mathematics and Statistics, Wuhan University, P. R. of China

Haibo Zhou[*], and
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, N.C. 27599-7420, U.S.A.

Jianwen Cai[**]
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, N.C. 27599-7420, U.S.A.

Summary

As biological studies become more expensive to conduct, statistical methods that take advantage of existing auxiliary information about an expensive exposure variable are desirable in practice. Such methods should improve the study efficiency and increase the statistical power for a given number of assays. In this paper, we consider an inference procedure for multivariate failure time with auxiliary covariate information. We propose an estimated pseudo-partial likelihood estimator under the marginal hazard model framework and develop the asymptotic properties for the proposed estimator. We conduct simulation studies to evaluate the performance of the proposed method in practical situations and demonstrate the proposed method with a data set from the Studies of Left Ventricular Dysfunction (SOLVD,1991).

Keywords

Auxiliary covariate; Marginal hazard model; Multivariate data; Pseudo-partial likelihood; Validation sample

1. Introduction

Statistical methods are usually developed assuming that the exposure variable is fully observed. In many studies, due to financial limitations or technical difficulties, the true exposure may only be measured precisely in a subset of the study cohort. This subset is often referred to as the validation set. With the continuing advancement in the use of biological markers in epidemiology and genetic studies, which often involve expensive assays, there is a growing incentive to further improve study efficiency and power by optimally incorporating into the statistical analysis the available auxiliary covariate. For example, in the Studies of Left Ventricular Dysfunction (SOLVD,1991) prevention trial, it is of interest to assess the effects of covariates (e.g. ejection fraction, intervention, and gender) on the risk of heart failure and

email: liuyy@whu.edu.cn.
[*]*email:* zhou@bios.unc.edu
[**]*email:* cai@bios.unc.edu

6. Supplementary Materials The Web Appendix and Web Table referenced in Sections 1, 3 and 5, are available under the Paper Information link at the Biometrics web site http://www.biometrics.tibs.org.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

on the first myocardial infarction. The gold standard of the ejection fraction (LVEF) measurement was to use a standardized radionucleotide technique. Since this standardized radionucleotide technique is too expensive to be used on every patient, the LVEF was only measured for a subset of 108 out of a total of the 4228 SOLVD patients. A cheaper and easily obtained measure of ejection fraction (EF) was, however, ascertained for all the patients using a nonstandardized technique. EF was considered as an auxiliary covariate of LVEF. The auxiliary covariate is defined as the surrogate information (the nonstandardized EF measure in the SOLVD study) that relates to the true exposure variable (LVEF) but provides no additional information to the regression model when the true covariates are known. Some proposed methods have been developed for the univariate survival time data in the areas of mismeasured covariates, missing data, and auxiliary covariate problems. This includes but is not limited to Pepe and Fleming (1991), Carroll and Wang (1991), Lin and Ying (1993), Zhou and Pepe (1995), Wang, Lin, and Gutierrez (1998), Hu, Tsiatis and Davidian (1998), Tsiatis and Davidian (2001), Huang and Wang (2000), Zhou and Wang (2000), Hu and Lin (2002) and Wang and Zhou (2006).

All the aforementioned studies assumed that each failure time is taken from independent subjects. In practice, multivariate or correlated failure time data with auxiliary data is just as likely to be encountered. For example, in the SOLVD study, the heart failure time and the first myocardial infarction time from the same subject could be correlated. Models dealing with multivariate failure time data where the true covariates of interest are fully available for all subjects have been well studied. In particular, if the correlation among the observations is not of interest, the marginal proportional hazards model is widely used, e.g., Wei, Lin and Weissfeld (1989); Lee, Wei and Amato (1992); Liang, Self and Chang (1993); Lin (1994); Cai and Prentice (1995, 1997); Spiekerman and Lin (1998); Clegg, Cai and Sen (2000). There has been limited progress on the methods for dealing with covariate measurement error for multivariate failure time. Greene and Cai (2004) proposed using the SIMEX approach for handling measurement errors in the marginal hazards model for multivariate failure time data, when a validation set is not available.

In this paper, assuming a validation set is available, we develop an estimated pseudo partial likelihood method for handling auxiliary covariates in the presence of a validation sample under the framework of the marginal hazards model with distinguishable baseline hazards (Wei *et. al.*, 1989). The auxiliary covariate could be a mismeasured surrogate to the true covariate, or any covariate which is informative about the true covariate. The proposed method is nonparametric with respect to the conditional distribution of the exposure variable conditional on the auxiliary covariate.

The rest of the paper is organized as follows. In Section 2, we outline the model and present the estimated pseudo-partial likelihood estimator. We develop the asymptotic properties of the proposed estimator and propose a variance estimator in Section 3. In Section 4, we evaluate the proposed methodology through simulation studies. We apply the proposed estimator to study the effect of ejection fraction on the risk of heart failure and first myocardial infarction using the data from the SOLVD study. Final remarks are given in Section 5. Outline of the proof for theoretical results are given in the Web Appendix.

2. Model and Estimation

2.1 *Notation and Data Structure*

Suppose that there is a random sample of n independent subjects from an underlying population and that there are K different types of failures of interest for each subject. Let (i, k) denote the kth failure type for the ith subject. Let T_{ik} and C_{ik} denote the potential failure time and censoring time, respectively. With censoring, we observe $X_{ik} = \min(T_{ik}, C_{ik})$. Let $\Delta_{ik} = I(X_{ik} \leq C_{ik})$ be the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

failure indicator and $Y_{ik}(t) = I(X_{ik} \geq t)$ denote the at-risk indicator process. Let (E_{ik}, Z_{ik}) denote a set of covariates, where E_{ik} is the primary exposure subjecting to missing and $Z_{ik} = (Z_{ik1}, \cdots, Z_{ikp})'$ is the remaining covariate vector that is always observed. We denote variable A as an auxiliary variable for the exposure variable E, assuming that conditional on E, A provides no additional information to the regression model, i.e. $\lambda(t; E(t), Z(t), A(t)) = \lambda(t; E(t), Z(t))$.

Suppose that there is a simple random validation sample with sample size n_v, denoted by V, such that subjects belonging to V have their (E, A) measured. Similarly, let $\bar{V}$ denote the remaining subjects, the non-validation set, and assume that the subjects in $\bar{V}$ will only have their A measured. Hence, the observed data structure for (i, k) is:

$$\{X_{ik}, \Delta_{ik}, Z_{ik}, A_{ik}, E_{ik}\} \quad \text{if} \quad i \in V$$

$$\{X_{ik}, \Delta_{ik}, Z_{ik}, A_{ik}\} \quad \text{if} \quad i \in \bar{V}$$

2.2 Models and Estimated Pseudo-Partial Likelihood Function

Assume that, the marginal hazard function for the kth failure type of subject i takes the form:

$$\lambda_{ik}(t; Z_{ik}(t), E_{ik}(t)) = Y_{ik}(t)\lambda_{0k}(t)\exp\left\{\beta_2' Z_{ik}(t) + \beta_1' E_{ik}^*(t)\right\} \tag{1}$$

where E_{ik}^* is an m-vector consisting of E_{ik} and possibly interaction terms between E_{ik} and some fully observed covariates, $\beta = (\beta_1', \beta_2')'$ is the relative risk parameter to be estimated, and $\lambda_{0k}(t)$ is an unspecified marginal baseline hazard function pertaining to the type k failure.

If subject i belongs to the validation set, then Z_{ik} and E_{ik} are observed and the marginal model takes the form as in (1). If subject i belongs to the non-validation set $\bar{V}$, we only observe $Z_{ik}(t)$ and $A_{ik}(t)$. Under this situation, we can show, using the argument of Prentice (1982) and Zhou and Pepe (1995), that the hazard function for $\lambda_{ik}(t; Z_{ik}(t), A_{ik}(t))$ satisfied the induced model

$$\begin{aligned}\lambda_{ik}(t; Z_{ik}(t), A_{ik}(t)) = &\quad Y_{ik}(t)\lambda_{0k}(t)e^{\beta_2' Z_{ik}(t)} E\left\{e^{\beta_1' E_{ik}^*(t)} | Y_{ik}(t) = 1, A_{ik}(t), Z_{ik}(t)\right\} \\ = &\quad Y_{ik}(t)\lambda_{0k}(t)e^{\beta_2' Z_{ik}(t)} E\left\{e^{\beta_1' E_{ik}^*(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)\right\}\end{aligned} \tag{2}$$

where A^* includes auxiliary variable A and the part of the information in covariate Z that, given A, are still related to E. That is, A^* satisfying the following conditional dependence $f(E_{ik}(t) | X_{ik} \geq t, Z_{ik}(t), A_{ik}(t)) = f\left(E_{ik}(t) | X_{ik} \geq t, A_{ik}^*(t)\right)$. 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., $\lambda(t; Z(t), E(t), A^*(t)) = \lambda(t; Z(t), E(t))$.

Equation (2) implies that this induced hazard model is also a proportional hazard model with the relative risk function $\exp\left(\beta_2' Z_{ik}(t)\right)\phi_{ik}(\beta_1; t)$, where $\phi_{ik}(\beta_1, t) = E\left\{e^{\beta_1' E_{ik}^*(t)} | Y_{ik}(t) = 1, A_{ik}^*(t)\right\}$. Based on (1) and (2), the relative risk function can be written in general as

$$r_{ik}(\beta, t) = R_{ik}(\beta_1, t)\exp\left(\beta_2' Z_{ik}(t)\right)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where

$$R_{ik}(\beta_1, t) = \exp\left[\beta_1' E_{ik}^*(t)\right]\eta_i + \phi_{ik}(\beta_1, t)(1-\eta_i),$$

and the binary variable $\eta_i = 1$ or 0 denote whether subject i is in validation set V or not. In addition to the unspecified baseline hazard function $\lambda_{0k}(t)$, the expectation above also depends on the underlying distributions of $E_{ik}(t)$ and $Z_{ik}(t)$. If $f\left(E_{ik}(t)|X_{ik} \geq t, A_{ik}^*(t)\right)$ is a known function up to a parameter θ, then the inference about β and θ can be drawn from a pseudo partial likelihood based on the model for multivariate failure times (Wei *et. al.*, 1989; Cai and Prentice, 1995). However, misspecification of such parameterization may lead to biased estimates. We develop an estimated pseudo-partial likelihood approach for correlated failure time data that avoids making undesirable parametric assumptions on the conditional distribution.

If all the observations were independent, we could write the partial likelihood as

$$PPL(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left[\frac{r_{ik}(\beta, X_{ik})}{\sum_{l=1}^{n} Y_{lk}(X_{ik})\, r_{lk}(\beta, X_{ik})}\right]^{\Delta_{ik}}. \tag{3}$$

When the failure times within a subject are not independent, the above function is referred to as the pseudo-partial likelihood (Wei *et. al.*, 1989; Cai and Prentice, 1995). Since the induced relative risk function $r_{ik}(\beta, t)$ is unknown, we will estimate it by using the validation set. Without loss of generality, we assume that $\{A_{ik}^*\}$ are identically distributed categorical variables with the distribution $Pr(A^* = a_m) = p_m$, $m = 1, \cdots, L$, $\sum_{m=1}^{L} p_m = 1$. Hence, if subject i is in the non-validation set $\bar{V}$, we will estimate the induced hazard function for the kth type failure, $\phi_{ik}(\beta_1, t)$, as:

$$\widehat{\phi}_{ik}(\beta_1, t) = \frac{\sum_{l\in V} Y_{lk}(t)\, I\left(A_{lk}^*(t) = A_{ik}^*(t)\right)\exp\left(\beta_1' E_{lk}^*(t)\right)}{\sum_{l\in V} Y_{lk}(t)\, I\left(A_{lk}^*(t) = A_{ik}^*(t)\right)}. \tag{4}$$

It follows that the estimated relative risk function is: $\widehat{r}_{ik}(\beta, t) = \widehat{R}_{ik}(\beta_1, t)\exp\left[\beta_2' Z_{ik}(t)\right]$, where $\widehat{R}_{ik}(\beta_1, t) = \exp\left(\beta_1' E_{ik}^*(t)\right)\eta_i + \widehat{\phi}_{ik}(\beta_1, t)(1-\eta_i)$.

Replacing $r_{ik}(\beta, t)$ by $\hat{r}_{ik}(\beta, t)$ in (3), we obtain an estimated pseudo-partial likelihood function:

$$EPPL(\beta) = \prod_{k=1}^{K}\prod_{i=1}^{n}\left[\frac{\widehat{r}_{ik}(\beta, X_{ik})}{\sum_{l=1}^{n} Y_{lk}(X_{ik})\,\widehat{r}_{lk}(\beta, X_{ik})}\right]^{\Delta_{ik}}. \tag{5}$$

We define our proposed estimator $\hat{\beta}_E$ as the maximizer of (5). $\hat{\beta}_E$ can be obtained by solving the estimated pseudo-partial likelihood score equation, $\hat{U}(\beta) = (\partial/\partial\beta)\{\log EPPL(\beta)\} = 0$, where

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\widehat{U}(\beta)=\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{\tau}\frac{\widehat{r}_{ik}^{(1)}(\beta,u)}{\widehat{r}_{ik}(\beta,u)}dN_{ik}(u)-\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{\tau}\frac{\sum_l Y_{lk}(u)\widehat{r}_{lk}^{(1)}(\beta,u)}{\sum_l Y_{lk}(u)\widehat{r}_{lk}(\beta,u)}dN_{ik}(u), \tag{6}$$

and $N_{ik}(t) = I(X_{ik} \le t, \Delta_{ik} = 1)$ is the counting process corresponding to failure time T_{ik}. For a function $g(\beta, u)$, $g^{(j)}(\beta, u)$ denotes the jth derivative of $g(\beta, u)$ with respect to β. A Newton-Raphson iterative procedure can be invoked to obtain $\hat{\beta}_E$.

3. Asymptotic Properties

To investigate the asymptotic properties of the estimated pseudo-partial likelihood estimator $\hat{\beta}_E$, we define the following notations. For a vector a, define $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa'$, $\|a\| = \sup_i |a_i|$. For a matrix A, define $\|A\| = \sup_{i,j} |a_{ij}|$. We also define

$$s_k^{(0)}(\beta,t)=E\left(Y_{ik}(t)\,r_{ik}(\beta,t)\right),$$

$$s_k^{(j)}(\beta,t)=E\left(Y_{ik}(t)\,r_{ik}^{(j)}(\beta,t)\right),(j=1,2),$$

$$e_{1k}(\beta,t)=E\left(Y_{ik}(t)\left(\frac{r_{ik}^{(1)}(\beta,t)}{r_{ik}(\beta,t)}\right)^{\otimes 2}r_{ik}(\beta_0,t)\right),$$

$$e_{2k}(\beta,t)=E\left(Y_{ik}(t)\,\frac{r_{ik}^{(2)}(\beta,t)}{r_{ik}(\beta,t)}\,r_{ik}(\beta_0,t)\right).$$

Assume that the study duration is from 0 to τ. Suppose that $\beta_0=\left(\beta_{10}',\beta_{20}'\right)'$ is the true hazards parameter. Our asymptotic results rely on the following assumptions:

(i) (Finite interval): $\int_0^{\tau}\lambda_{0k}(t)\,dt<\infty$ ($k = 1, \cdots, K$);

(ii) $Pr\left(Y_{ik}(t)=1|A_{ik}^{*}(t)=a_m\right)>0$, for $m = 1, 2, \cdots, L$;

(iii) For any $k = 1, \cdots, K$, there exits a neighborhood B_2 of β_{20} such that

$$E\left(\sup_{B_2\times[0,\tau]}\|Z_{ik}(t)\|^2 e^{\beta_2' Z_{ik}(t)}\right)<\infty$$

(iv) There exists an open set B_1, containing β_{10}, such that $\phi_{ik}(\beta_1, t)$ is bounded away from 0 on $B_1 \times [0, \tau]$. $\Sigma(\beta_0)$, as defined in Theorem 2, is positive definite.

(v) For any $k = 1, \cdots, K$,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$E\left\{\sup_{B_1\times[0,\tau]}\left[Y_{ik}(t)R_{ik}^{(d)}(\beta,t)\right]\right\}<\infty \quad (d=0,1,2)$$

$$E\left\{\sup_{B_1\times[0,\tau]}\left[Y_{ik}(t)\left\|\left(\frac{R_{ik}^{(1)}(\beta,t)}{R_{ik}(\beta,t)}\right)^{\otimes 2}\right\|^{d}R_{ik}(\beta_0,t)\right]\right\}<\infty \quad (d=1,2)$$

$$E\left\{\sup_{B_1\times[0,\tau]}\left[Y_{ik}(t)\left\|\frac{R_{ik}^{(2)}(\beta,t)}{R_{ik}(\beta,t)}\right\|^{d}R_{ik}(\beta_0,t)\right]\right\}<\infty \quad (d=1,2)$$

(vi) $\sup_{t\in[0,\tau]}|L_k^{(d)}(t)|=O_p(1)$, d = 0, 1, where

$$L_k^{(d)}(t)=\sqrt{n_v}\left[\frac{1}{n_v}\sum_{j=1}^{n_v}I_{\left(Y_{jk}(t)=1,A_{jk}^{*}=a\right)}\gamma_{jk}^{(d)}(\beta_1,t)-E\left(I_{\left(Y_{ik}(t)=1,A_{ik}^{*}(t)=a\right)}\gamma_{ik}^{(d)}(\beta_1,t)\right)\right]$$

and $\gamma_{ik}(\beta_1,t)=\exp\left(\beta_1' E_{ik}\right)$.

Following closely the argument of Foutz (1977), we can show that $\hat{\beta}_E$ is consistent for β_0. To show the asymptotic normality of $\hat{\beta}_E$, we use the Taylor expansion of the score equation which, using martingale representation and theory, can be shown to be asymptotically equivalent to a sum of two independent terms. Each of the terms can be shown to be a sum of independent vectors. The multivariate central limit theorem is then applied. We summarize the results in the following theorems and give the outline of the proofs in the Web Appendix.

Theorem 1

(Consistency) $\hat{\beta}_E$ is a consistent estimator of β_0 under assumptions (i)-(vi).

Theorem 2

(Asymptotic Normality) Under the assumptions (i)-(vi), we have that $n^{1/2}(\hat{\beta}_E - \beta_0)$ is asymptotically normally distributed with mean zero and variance matrix $\Sigma_{EPPL}(\beta_0)=\Sigma^{-1}(\beta_0)\left[\Sigma_1(\beta_0)+\Sigma_2(\beta_0)\right]\Sigma^{-1}(\beta_0)^T$, where

$$\Sigma(\beta_0)=-\int_0^{\tau}\sum_{k=1}^{K}\left[\left(\frac{s_k^{(1)}(\beta_0,t)}{s_k^{(0)}(\beta_0,t)}\right)^{\otimes 2}s_k^{(0)}(\beta_0,t)-e_{1k}(\beta_0,t)\right]\lambda_{0k}(t)\,dt$$

$$\Sigma_1(\beta_0)=(1-q)\left(\sum_{k=1}^{K}E\left(g_{ik}(\beta_0)g_{ik}'(\beta_0)\right)+\sum_{1\le l\ne j\le K}E\left(g_{il}(\beta_0)g_{ij}'(\beta_0)\right)\right)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\Sigma_2(\beta_0) = q\left(\sum_{k=1}^{K} E\left(h_{ik}(\beta_0)\, h'_{ik}(\beta_0)\right) + \sum_{1\le l\ne j\le K} E\left(h_{il}(\beta_0)\, h'_{ij}(\beta_0)\right)\right)$$

and

$$g_{ik}(\beta_0) = \int_0^\tau \left[\begin{pmatrix} \frac{\phi_{ik}^{(1)}(\beta_{10},t)}{\phi_{ik}(\beta_{10},t)} \\ Z_{ik}(t) \end{pmatrix} - \frac{s_k^{(1)}(\beta_0,t)}{s_k^{(0)}(\beta_0,t)}\right] dM_{ik}(t)$$

$$h_{ik}(\beta_0) = \int_0^\tau \left[\begin{pmatrix} \frac{\phi_{ik}^{(1)}(\beta_{10},t)}{\phi_{ik}(\beta_{10},t)} \\ Z_{ik}(t) \end{pmatrix} - \frac{s_k^{(1)}(\beta_0,t)}{s_k^{(0)}(\beta_0,t)}\right] dM_{ik}(t) - \frac{1-q}{q}\begin{pmatrix} Q_{ik}(\beta_0) \\ H_{ik}(\beta_0) \end{pmatrix}$$

$$Q_{ik}(\beta_0) = \int_0^\tau \left(\frac{\phi_{ik}^{(1)}(\beta_{10},t)}{\phi_{ik}(\beta_{10},t)} - \frac{s_k^{(11)}(\beta_0,t)}{s_k^{(0)}(\beta_0,t)}\right) Y_{ik}(t)\left(e^{\beta'_{10}E^*_{ik}} - \phi_{ik}(\beta_{10},t)\right)\delta_k^*(\beta_0,t)\,\lambda_{0k}(t)\,dt$$

$$H_{ik}(\beta_0) = \int_0^\tau Y_{ik}(t)\left(e^{\beta'_{10}E^*_{ik}} - \phi_{ik}(\beta_{10},t)\right)\delta_k^{**}(\beta_0,t)\,\lambda_{0k}(t)\,dt$$

Here $s_k^{(11)}(\beta_0,t)$ is the first m elements of $s_k^{(1)}(\beta_0,t)$ and $s_k^{(12)}(\beta_0,t)$ contains the remaining p elements, so $s_k^{(1)}(\beta_0,t) = \begin{pmatrix} s_k^{(11)}(\beta_0,t) \\ s_k^{(12)}(\beta_0,t) \end{pmatrix}$, and

$$\delta_k^*(t,\beta_0) = E\left(e^{\beta'_{20}Z_{ik}(t)} | Y_{ik}(t)=1, A^*_{ik}(t)\right),$$

$$\delta_k^{**}(t,\beta_0) = E\left(\left[Z_{ik}(t) - \frac{s_k^{(12)}(\beta_0,t)}{s_k^{(0)}(\beta_0,t)}\right] e^{\beta'_{20}Z_{ik}(t)} | Y_{ik}(t)=1, A^*_{ik}(t)\right),$$

$q = \lim_{n\to\infty}(n_v/n)$ denotes the validation fraction and $M_{ik}(t) = N_{ik}(t) - \int_0^t \lambda_{ik}(u)\,du$ is the marginal martingale.

Remark 1

Observe that, when the validation fraction $q = 1$, the variance matrix Σ_{EPPL} is the same as that of the usual pseudo-partial likelihood estimator (Wei *et. al.*, 1989; Cai and Prentice, 1995), as it should be.

The variance estimator for $\hat{\beta}_E$ can be consistently estimated by replacing the population quantities in the asymptotic covariance matrix $\Sigma_{EPPL}(\beta_0)$ with their corresponding sample

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

quantities. The cumulative hazard $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(u)\,du$ can be estimated by Aalen-Breslow type of estimator:

$$\widehat{\Lambda}_{0k}(t) = \int_0^t \frac{\sum_{i=1}^{n} dN_{ik}(s)}{\sum_{i=1}^{n} Y_{ik}(s)\widehat{r}_{ik}\left(\widehat{\beta}_E, s\right)} = \int_0^t \frac{1}{\widehat{S}_k^{(0)}\left(\widehat{\beta}_E, s\right)} \frac{1}{n} \sum_{i=1}^{n} dN_{ik}(s),$$

where $\widehat{S}_k^{(0)}(\beta, s)$ is the empirical estimator of $s_k^{(0)}(\beta, s)$ which is defined as:

$$\widehat{S}_k^{(0)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_{ik}(t)\widehat{r}_{ik}(\beta, t).$$

4. Numerical Studies

4.1 *Simulation Studies*

In this section, we examine the finite sample properties of the proposed estimator $\hat{\beta}_E$ via simulation studies. We compare $\hat{\beta}_E$ with two naive estimators, $\hat{\beta}_V$, which is the pseudo-partial likelihood estimator (Wei *et. al.*, 1989) based only on the validation set, and $\hat{\beta}_N$, which is the pseudo-partial likelihood estimator using the auxiliary variable in place of the true exposure variable. We evaluate these estimators under various situations with different levels of censoring proportions, validation fractions, dependence between failure times within a subject, and the informativeness of the auxiliary covariate.

We generated the multivariate failure times from the popular multivariate Clayton and Cuzick (1985) distribution with exponential marginals. The joint survival distribution function takes the form

$$S(t_1, \cdots, t_K; Z_1, \cdots, Z_k, E_1, \cdots, E_k) = \left\{ \sum_{k=1}^{K} \exp\left(\theta^{-1} \lambda_{0k} t_k e^{\beta'_{(k)} D_k}\right) - (K-1) \right\}^{-\theta}$$

where $D_k = \left(E_k, Z_k'\right)'$ and $\beta_{(k)}$ is the corresponding coefficient for D_k. The parameter θ, where $\theta > 0$, controls the degree of dependence between T_j and T_l ($j, l = 1, \cdots, K$), with $\theta \to \infty$ corresponding to independence and $\theta \to 0$ corresponding to increasing positive correlation. We take λ_{0k} to be an arbitrary constant baseline hazard. We simulate failure times that follow Clayton-Cuzick distribution by transforming independent uniform (0, 1) variables ($u_1, \cdots, u_K$) as follows: $t_1 = -(1/\lambda_{01})\, e^{-\beta'_{(1)} D_1} \times \ln(1 - u_1)$ and $t_k = (1/\lambda_{0k})\, e^{-\beta'_{(k)} D_k} \ln\left[(k-1) - \sum_{i=1}^{k-1} a_i + \left(\sum_{i=1}^{k-1} a_i - (K-2)\right)(1-u_k)^{-(\theta+k-1)^{-1}}\right]$, for $k = 2, \cdots, K$, where $a_l = e^{-\theta^{-1} t_l e^{\beta'_{(l)} D_l}}$ for $l = 1, \cdots, k-1$. We considered an exposure variable E which could have different effect for different failure types, and an adjustment covariate Z which has the same effect for different failure types. The generated failure times satisfy the following marginal hazards model:

$$\lambda_{ik}(t; E_{ik}, Z_{ik}) = \lambda_{0k} \exp\{\beta_{k1} E_{ik} + \beta_2 Z_{ik}\}, \quad k = 1, \cdots, K, \tag{7}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where β_{k1} denotes the exposure (E) effect on the kth failure type and β_2 is the common effect of Z.

We simulate $K = 2$ failure types with baseline hazard $\lambda_{0k} = 1$. We consider $(\beta_{11}, \beta_{21}) = (0, 0)$ or $(\beta_{11}, \beta_{21}) = (\ln(2), \ln(1.3)) = (0.693, 0.262)$ and $\beta_2 = -0.2$. We set $\theta = 0.25$, 1.5, or 5.7, which respectively represents a strong, modest, or weak positive dependence between failure times within a subject. When $\beta_{11} = \beta_{21} = 0$ and $\beta_2 = -0.2$ with 20% censoring, these values of θ correspond to the correlation coefficients (for the failure times within a subject) of 62%, 30% and 10%, respectively. Censoring times were generated from uniform distribution over $(0, c)$, where c was chosen to yield the censoring percentage of 20% or 50%.

Mimicking the SOLVD study data structure, where $f(E|A, Z) = f(E|A)$, we create our exposure and auxiliary variables for each subject from the following scheme. We generate the partly observed covariate (E_1, E_2) from a multivariate normal distribution with marginal mean of 0, standard deviation of 1, and the correlation between E_1 and E_2 being $r = 0$ and 0.8, which represent cases of independence and strong dependence between E_1 and E_2. The fully observed covariate (Z_1, Z_2) are generated from independent normal distribution $N(0, 1)$. To generate auxiliary variable A, we first generate $W_k = E_k + e_k$, $(k = 1, 2)$, where $e_k \sim N(0, \sigma^2)$, and σ is the parameter that controls the strength of the association between E_k and W_k. The auxiliary covariate A is then assigned the value of 1, 2, 3 or 4 based on whether W_k is in the interval $(-\infty, Q_1]$, $(Q_1, Q_2]$, $(Q_2, Q_3]$, or $(Q_3, +\infty)$, where Q_1, Q_2, Q_3 are the quartiles of W. The parameter σ also controls the strength of the association between E and A: as σ^2 increases, A becomes less informative about E. We set $\sigma = 0.2$ or 0.8.

For each specified set of parameters, the number of independent subjects is $n = 200$ and each simulation is repeated 1000 times. The sample standard deviation of the 1000 estimates are given in the corresponding SD columns. The SE columns give the average of the estimated standard errors and "95% CI" is the nominal 95% confidence interval coverage of the true parameter using the estimated standard error.

Table 1 displays the simulation results when the exposure E has no effect on failures, i.e. $\beta_{11} = \beta_{21} = 0$, under a variety of configurations of the parameters when validation fraction is 50% and censoring rate is 20%. Under this situation, failure times satisfy the following model:

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

Table 2 provides results under the situation when E has different effect on different failure types. From Tables 1 and 2, we make the following observations. (i) In Table 1, i.e. when $\beta_1 = 0$, all the estimates are approximately unbiased. In Table 2, i.e. when $\beta_{11} \neq 0$, $\beta_{21} \neq 0$, $\hat{\beta}_N$ is biased towards 0. Both the validation estimator $\hat{\beta}_V$ and the proposed estimator $\hat{\beta}_E$ are approximately unbiased. (ii) The proposed estimator $\hat{\beta}_E$ is more efficient than the validation set only estimator $\hat{\beta}_V$. (iii) When σ is large, i.e. when A is not as informative about E, $\hat{\beta}_E$ is less accurate in estimating the true β, e. g. the bias exibited in $\hat{\beta}_E$ is larger for $\sigma = 0.8$. This bias, however, decreases as we increase the sample size to $n = 500$ (results not shown). (iv) The proposed variance estimator provides a good estimation of $\Sigma_{EPPL}(\beta)$. (v) The 95% confidence intervals based on the proposed estimated standard errors provide good coverage for most of the situations studied when $\sigma = 0.2$. The coverage rates are lower when $\sigma = 0.8$. However, when we increased n to 500 (results not shown), the 95% CI coverage rates increased and were close to 95%.

Table 3 compares the estimated relative efficiency of $\hat{\beta}_E$ vs. $\hat{\beta}_V$ under different censoring proportions (20%, 50%), validation fractions (30%, 50%, 70%) and strength of correlations between failure times within a subject ($\theta = 0.25$, 1.5, 5.7). The estimated relative efficiency

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

(RE) are calculated as $(SD(\hat{\beta}_V)/SD(\hat{\beta}_E))^2$. From Table 3, we had the following observation: when the validation fraction decreases, the efficiency gain of $\hat{\beta}_E$ relative to $\hat{\beta}_V$ increases. For example, when $\theta = 0.25$ and 20% censoring, the REs for $(\hat{\beta}_{11}, \hat{\beta}_{21}, \hat{\beta}_2)$ increase from (1.452, 1.399, 1.423) to (3.466, 3.551, 3.894) when validation fraction decreases from 70% to 30%. We observe the same trend when censoring rate is 50% or θ=1.5 or 5.7. This suggests that when the validation fraction is small, using our proposed method is even more beneficial compared to using the estimator based on the validation set only.

4.2 *Analysis of the SOLVD Study Data Set*

We illustrated the proposed method with a data set from the SOLVD study to evaluate the effect of ejection fraction on the time of heart failure and the time to first nonfatal myocardial infarction (nonfatal MI). The SOLVD study was a randomized, double-masked, placebo-controlled trial between 1986 and 1991. The trial had a three year recruitment and a two year follow-up. The basic inclusion criteria for the prevention trial were: age between 21 and 80 years, inclusive, no overt symptoms of congestive heart failure, and left ventricular ejection fraction less than 35 percent. Ejection fraction is a number between 0 and 100 that measures the efficiency of the heart in ejecting blood. A total of 4228 patients with asymptomatic left ventricular dysfunction were randomly assigned to receive either enalapril or placebo at one of the 83 hospitals linked to 23 centers in the United States, Canada, and Belgium.

The correlated outcomes of interest are time to heart failure and time to the first nonfatal MI after the randomization. The primary clinical issues of interest are the effects of covariates on the risk of heart failure and on the nonfatal MI after adjusting for the confounding variables. In the SOLVD study, only 108 among the total of 4228 patients have their ejection fraction accurately measured using a standardized radionucleotide technique (LVEF). A related nonstandardized measure (EF) was, however, ascertained for all the patients. Therefore, the nonstandardized measure (EF) is a surrogate measure for the standardized measure for ejection fraction (LVEF) in this case. Both LVEF and EF were measured in percentage.

The average LVEF in the validation set is 28.3% ranging from 12.3% to 45.4%. The average EF for the entire cohort is 19.37% ranging from 1% to 32%. We create a new auxiliary variable VEF taking values $a_1, \cdots, a_4$ depending on whether EF belongs to the interval $[\min(EF), Q_1]$, $(Q_1, Q_2]$, $(Q_2, Q_3]$ and $(Q_3, \max(EF)]$ respectively, where $a_1, \cdots, a_4$ are the values of the midpoint of the aforementioned intervals, and Q_1, Q_2, and Q_3 are the quartiles of EF. We use VEF as the auxiliary covariate for LVEF. Other covariates we considered here are patient's gender (SEX), which is coded 1 for male and 0 for female; treatment (TRT), which is coded as 1 for enalapril and 0 for placebo, and patient's age (AGE), which was measured in years. The average age of the patients is 59 years old with a standard deviation of 10 years. Hence, in terms of the notation in the previous sections, we have $E = LVEF$, $A = VEF$, $Z = (TRT, SEX, AGE)'$. To check whether, for given VEF, LVEF is dependent on TRT, SEX, and AGE, we added TRT, SEX, and AGE to the linear model of LVEF on VEF. The results showed that the TRT, SEX, and AGE effects are not statistically significant for given VEF. We also examined this relationship for subjects who are at risk at some selected time points and the same conclusion was arrived. Hence we took $A^* = VEF$ in this case.

We fit the following marginal hazards model to the SOLVD data:

$$\begin{aligned}&\lambda_{ik}(t|\mathrm{SEX}_{ik}, TRT_{ik}, LVEF_{ik}, VEF_{ik}, \mathrm{AGE}_{ik})\\&=\lambda_{0k}(t)\,Y_{ik}(t)\,R_{ik}(\beta_1, \gamma_1, t)\exp\{\beta_2 \cdot TRT_{ik}+\beta_3 \cdot \mathrm{SEX}_{ik}+\beta_4 \cdot \mathrm{AGE}_{ik}\}\\&\times\exp\{\gamma_2 I(k=2)\,TRT_{ik}+\gamma_3 I(k=2)\,\mathrm{SEX}_{ik}+\gamma_4 I(k=2)\,\mathrm{AGE}_{ik}\}\end{aligned} \quad (8)$$

where

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$R_{ik}(\beta_1, t) = \begin{cases} \exp(\beta_1 \cdot LVEF_{ik} + \gamma_1 I(k=2) LVEF_{ik}) & \text{when } LVEF \text{ is observed} \\ \phi_{ik}(\beta_1, \gamma_1, t) & \text{when } LVEF \text{ is missing} \end{cases}$$

and

$$\phi_{ik}(\beta_1, \gamma_1, t) = \frac{\sum_{l \in V} Y_{lk}(t) I(VEF_{lk} = VEF_{ik}) \exp\{\beta_1 \cdot LVEF_{lk} + \gamma_1 I(k=2) LVEF_{lk}\}}{\sum_{l \in V} Y_{lk}(t) I(VEF_{lk} = VEF_{ik})}$$

where k denotes failure type with $k = 1$ for heart failure and $k = 2$ for nonfatal MI and i denotes the patient with $i = 1, \cdots, 4228$.

Table 4 presents the data analysis results. The columns under "Proposed method" list results using the proposed method. The columns under "Validation method" are the results from fitting model (1), using the pseudo-partial likelihood approach, based only on the validation set. The proposed method utilizes the auxiliary information while the validation analysis relies only on the available true ejection fraction.

An inspection of the point estimates of the covariate effects reveals that the estimates from the proposed method are very close to those from the validation set only analysis, except for SEX. However, the proposed estimator is much more precise than the validation set only analysis, e.g, for the effect of LVEF on heart failure, the standard error is 0.008 for the proposed estimator and 0.038 for the validation only estimator. Consequently, the proposed method provided a tighter 95% confidence intervals, e.g., the 95% CI for LVEF for heart failure is (-0.071, -0.039) for the proposed method and (-0.149, -0.001) for the validation set only analysis. The *p*-values in Table 4 indicate that at 0.05 significance level, LVEF, TRT, SEX and AGE are all statistically significant for heart failure using the proposed method, while only LVEF is marginally significant for heart failure for the validation analysis. Note that this real data set has an extremely low validation fraction with a moderate validation sample size. From our additional simulations (results not shown), our proposed variance estimator could over-estimate the true variance when the validation fraction is low (see Concluding Remarks). Hence these confidence intervals should be interpreted on the conservative side.

The results from the proposed method also indicate that the effects of ejection fraction, treatment, sex, and age are different for the heart failure and for the non-fatal MI. Specifically, ejection fraction, treatment, sex and age do not seem to affect the risk of non-fatal MI, but is related to the risk of heart failure. The risk of heart failure increases by 2.5% (95% CI: (1.3%, 3.7%)) per year increase in age. With 1% decrease in ejection fraction, the risk of heart failure is about 5.3% (95% CI: (3.8%, 6.8%)) higher. Females are at 25.3% (95% CI: (0, 44.6%)) higher risk for heart failure than males. Enalapril reduces the risk of heart failure by 35.5% (95% CI: (17.6%, 49.5%)).

In conclusion, we found that the pseudo-partial likelihood method using only the validation data yielded no significant covariate effect and the estimated standard errors were much bigger than those from the proposed method. This is because the validation set is only a very small subset ($n = 108$) of the entire cohort ($n = 4228$). Utilizing the auxiliary information in the proposed method, we had in effect regained the statistical power of the study that would have been lost had one conducted the analysis using only the validation data set.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

5. Concluding Remarks

In this paper, we studied an estimated pseudo-partial likelihood method for multivariate failure time data with an auxiliary covariate. A key feature of this method is that it is nonparametric with respect to the association between the missing covariate and the observed auxiliary covariate. The estimated pseudo-partial likelihood estimator asymptotically follows a normal distribution. Simulation studies demonstrate that the proposed estimator approaches the large sample properties with moderate sample size. These results also show that the proposed estimator outperforms the estimator which uses only data from the validation sample. The proposed variance estimator based on the approximated asymptotic variance performs well. When the auxiliary covariate A is more informative about the partly observed covariate E, the proposed estimator is more efficient.

The real data example also demonstrates that a much more precise estimator can be obtained by incorporating the auxiliary covariate information. The proposed method shows improved statistical power over what would be achieved using only the validation set.

We have a couple of cautionary notes on the limitations of the proposed method. First, the auxiliary variable A^* is assumed to be discrete with the number of categories fixed. The asymptotic properties were developed under this assumption. One way to deal with a continuous auxiliary variable is to discretize it into categories and then apply the proposed method. In practice, the number of categories of the auxiliary variable cannot be too large (no more than 6), especially when the sample size is small. In our simulation, we have run into convergence problems when the validation size is less than 60 and the number of categories is greater than 8. We recommend reducing the number of categories of the auxiliary variable if it is greater than 8. Second, if the validation size is small and the validation fraction is also very low, the resulting variance estimator tends to over-estimate the true variance. Increasing the validation sample size helps to alleviate this problem (see Web Table 1).

To fully take advantage of a continuous auxiliary covariate, a nonlinear smoothing version of (4) can be developed. Cai and Prentice (1995) showed that more efficient β-estimators could be obtained by introducing weights into the pseudo-partial likelihood score equations. Future work that introduces suitable weights to our proposed method to improve the efficiency of estimators is certainly warranted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by SRF for ROCS. SEM. and the National Natural Science Foundation of China 10771163 (Liu) and NIH grants R01 CA79949 (Zhou) and R01 HL57444 (Cai). The authors thank the co-editor, the associate editor, and the two referees for their valuable suggestions which have led to significant improvement of this paper.

References

Andersen PK, Gill RD. Cox's regression model for counting processed: A large sample study. Ann.Statist 1982;10:1100–20.

Breslow NE. Covariate analysis of censored survival data. Biometrics 1974;30:89–99. [PubMed: 4813387]

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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]

Clegg, LX.; Cai, J.; Sen, PK. Handbook of Statistics. Vol. 18. 2000. Modeling multivariate Failure time data; p. 804-838.

Carroll RJ, Wang MP. Semiparametric estimation in logistic measurement error models. J. Roy. Statist. Soc 1991;53:573–585.Series B

Cox DR. Regression Models and Life-Tables. J. Roy. Statist. Soc 1972;34:187–202.B

Foutz RV. On the Unique consistent solution to the likelihood equations. Journal of the American Statistical Association 1977;72:147–148.

Greene WF, Cai J. Measurement Error in Covariate in the Marginal Hazards Model for Multivariate Failure Time Data. Biometrics 2004;60:987–996. [PubMed: 15606419]

Huang Y, Wang CY. Cox Regression with Accurate Covariates Uncertainable-A Noparametric Approach. Journal of the American Statistical Association 2000;95:1209–1219.

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

Hu P, Tsiatis AA, Davidian M. Estimating the parameters in the Cox model when covariates variables are measured with error. Biometrics 1998;54:1407–1419. [PubMed: 9883541]

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; 1992. p. 237-247.

Liang KY, Self SG, Chang Y. Modeling Marginal Hazards in Multivariate Failure Time Data. J. Royal Statist. Soc 1993;55:441–453.B

Lin DY. Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach. Statist. Med 1994;13:2233–2247.

Lin DY, Ying Z. Cox regression with incomplete covariate measurements. Journal of the American Statistical Association 1993;88:1341–1349.

Pepe MS, Fleming TR. A nonparametric method for dealing with mismeasured covariate data. Journal of the American Statistical Association 1991;86:108–113.

Prentice RL. covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 1982;69:331–342.

Spiekerman CF, Lin DY. Marginal regression models for multivaraite failure time data. J. Amer. Statist. Assoc 1998;93:1164–1175.

SOLVD Investigators. Studies of Left Ventricular Dysfunction (SOLVD)—Rationale, Design and Methods: Two Trials that Evaluate the Effect of Enalapril in Patients with Reduced Ejection Fraction'. The American Journal of Cardiology 1990;66:315–322. [PubMed: 2195865]

SOLVD Investigators. Effect of Enalapril on Survival in Patients with Reduced Left Ventricular Ejection Fractions and Congestive Heart Failure. New England Journal of Medicine 1991;325:293–302. [PubMed: 2057034]

Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika 2001;88:447–458.

Wang N, Lin X, Gutierrez RG, Carroll RJ. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. Journal of the American Statistical Association 1998;93:249–261.

Wang X, Zhou H. A semiparametric Empirical Likelihood Method for Biased Sampling Schemed with Auxiliary Covariates. Biometrics 2006;62:1149–1160. [PubMed: 17156290]

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;84:1065–73.

Zhou, H. University of Washington Ph.D. Thesis. 1992. Auxiliary and Missing Covariate Problems in Failure Time Regression Analysis.

Zhou H, Pepe MS. Auxiliary covariate data in failure time regression analysis. Biometrika 1995;82:139–149.

Zhou H, Wang C-Y. Failure time regression with continuous covariates measured with error. J. R. Statist. Soc. B 2000;62:657–665.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Simulation results for $\beta_1 = 0$ under model: $\lambda_{ik}(t; E_{ik}, Z_{ik}) = \lambda_{0k}(t)\, exp\{\beta_1 E_{ik} + \beta_2 Z_{ik}\}$ (k = 1, 2). Validation fraction is 50% and censoring rate is 20%

				$\hat{\beta}_1(\cdot)$				$\hat{\beta}_2(\cdot)$			
θ	r	σ	$\hat{\beta}$	$\hat{\beta}_1$	SD	SE	95% CI	$\hat{\beta}_2$	SD	SE	95% CI
0.25	0		$\hat{\beta}_V$	0.003	0.081	0.080	0.949	-0.204	0.085	0.081	0.935
		0.2	$\hat{\beta}_N$	0.000	0.068	0.065	0.943	-0.203	0.059	0.057	0.942
		0.8	$\hat{\beta}_N$	-0.001	0.053	0.052	0.942	-0.201	0.057	0.057	0.955
		0.2	$\hat{\beta}_E$	0.000	0.062	0.059	0.924	-0.203	0.059	0.062	0.926
		0.8	$\hat{\beta}_E$	0.004	0.063	0.065	0.949	-0.201	0.058	0.062	0.934
	0.8		$\hat{\beta}_V$	-0.002	0.083	0.080	0.935	-0.202	0.089	0.082	0.920
		0.2	$\hat{\beta}_N$	0.001	0.066	0.065	0.946	-0.201	0.060	0.057	0.938
		0.8	$\hat{\beta}_N$	0.002	0.056	0.052	0.934	-0.203	0.059	0.057	0.943
		0.2	$\hat{\beta}_E$	0.001	0.060	0.057	0.922	-0.201	0.060	0.063	0.925
		0.8	$\hat{\beta}_E$	0.000	0.066	0.067	0.939	-0.202	0.059	0.064	0.929
1.5	0		$\hat{\beta}_V$	-0.002	0.084	0.079	0.937	-0.203	0.084	0.081	0.936
		0.2	$\hat{\beta}_N$	0.001	0.065	0.065	0.951	-0.200	0.059	0.057	0.945
		0.8	$\hat{\beta}_N$	0.001	0.054	0.052	0.931	-0.203	0.058	0.057	0.944
		0.2	$\hat{\beta}_E$	-0.001	0.059	0.056	0.930	-0.200	0.059	0.061	0.922
		0.8	$\hat{\beta}_E$	0.002	0.070	0.066	0.914	-0.203	0.059	0.061	0.923
	0.8		$\hat{\beta}_V$	0.000	0.084	0.080	0.934	-0.204	0.082	0.081	0.945
		0.2	$\hat{\beta}_N$	0.000	0.066	0.065	0.943	-0.203	0.059	0.057	0.939
		0.8	$\hat{\beta}_N$	-0.001	0.054	0.052	0.934	-0.199	0.058	0.057	0.948
		0.2	$\hat{\beta}_E$	0.000	0.061	0.058	0.926	-0.203	0.059	0.062	0.935
		0.8	$\hat{\beta}_E$	-0.002	0.068	0.065	0.922	-0.199	0.057	0.061	0.928
5.7	0		$\hat{\beta}_V$	-0.004	0.085	0.080	0.935	-0.201	0.083	0.081	0.930
		0.2	$\hat{\beta}_N$	0.000	0.066	0.065	0.941	-0.203	0.058	0.057	0.951

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

				$\hat{\beta}_1(\cdot)$				$\hat{\beta}_2(\cdot)$			
θ	r	σ	$\hat{\beta}$	$\hat{\beta}_1$	SD	SE	95% CI	$\hat{\beta}_2$	SD	SE	95% CI
		0.8	$\hat{\beta}_N$	0.001	0.055	0.052	0.930	-0.201	0.059	0.057	0.949
		0.2	$\hat{\beta}_E$	-0.002	0.060	0.056	0.928	-0.203	0.058	0.061	0.921
		0.8	$\hat{\beta}_E$	0.000	0.071	0.068	0.929	-0.200	0.058	0.060	0.931
	0.8		$\hat{\beta}_V$	-0.002	0.084	0.080	0.939	-0.203	0.082	0.081	0.941
		0.2	$\hat{\beta}_N$	-0.001	0.070	0.066	0.944	-0.202	0.058	0.057	0.945
		0.8	$\hat{\beta}_N$	0.001	0.052	0.052	0.954	-0.201	0.059	0.057	0.945
		0.2	$\hat{\beta}_E$	-0.001	0.063	0.059	0.933	-0.202	0.058	0.062	0.927
		0.8	$\hat{\beta}_E$	0.001	0.068	0.066	0.928	-0.200	0.058	0.061	0.930

$\hat{\beta}_V$ is the WLW estimator by using only the validation set; $\hat{\beta}_N$ is a naive estimator, which uses auxiliary covariate A instead of the true exposure variable E; $\hat{\beta}_E$ is our proposed estimator. Large θ corresponds to weak dependence between group members, large σ corresponds to that the auxiliary covariate is less informative, and large r corresponds to the strong dependence between Z_1 and Z_2.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Simulation results for different covariate effects across failure types under the model: $\lambda_{ik}(t; E_{ik}(t), Z_{ik}(t)) = \lambda_{0k}(t)\, exp\{\beta_{k1}E_{ik}(t) + \beta_2 Z_{ik}(t)\}$ (k = 1, 2). $\beta_{11} = log(2) = 0.693$, $\beta_{21} = log(1.3) = 0.262$, $\beta_2 = -0.2$. Validation fraction is 50% and censoring rate is 20%

				$\beta_{11} = \log(2) = 0.693$				$\beta_{21} = \log(1.3) = 0.262$				$\beta_2 = -0.2$			
θ	r	σ	$\hat{\beta}$	$\hat{\beta}_{11}$	SD	SE	95% CI	$\hat{\beta}_{21}$	SD	SE	95% CI	$\hat{\beta}_2$	SD	SE	95% CI
0.25	0		$\hat{\beta}_V$	0.7062	0.1348	0.1292	0.937	0.2722	0.1226	0.1156	0.935	-0.2031	0.0891	0.0816	0.923
		0.2	$\hat{\beta}_N$	0.0069	0.0914	0.0934	0.000	-0.0030	0.0992	0.0922	0.194	-0.0024	0.0570	0.0562	0.064
		0.8	$\hat{\beta}_N$	0.0054	0.0754	0.0745	0.000	0.0013	0.0772	0.0734	0.060	-0.0025	0.0571	0.0562	0.059
		0.2	$\hat{\beta}_E$	0.6964	0.0967	0.1033	0.930	0.2646	0.0849	0.0837	0.906	-0.1975	0.0606	0.0628	0.925
		0.8	$\hat{\beta}_E$	0.6913	0.0961	0.0947	0.894	0.2565	0.0858	0.0918	0.919	-0.1915	0.0600	0.0626	0.908
0.25	0.8		$\hat{\beta}_V$	0.6977	0.1358	0.1293	0.937	0.2714	0.1273	0.1162	0.922	-0.2040	0.0900	0.0817	0.913
		0.2	$\hat{\beta}_N$	0.0061	0.0941	0.0935	0.000	-0.0003	0.0988	0.0927	0.201	-0.0028	0.0575	0.0562	0.063
		0.8	$\hat{\beta}_N$	0.0048	0.0743	0.0742	0.000	0.0020	0.0773	0.0735	0.072	-0.0027	0.0575	0.0562	0.064
		0.2	$\hat{\beta}_E$	0.6952	0.0959	0.1063	0.934	0.2638	0.0877	0.0851	0.910	-0.1980	0.0597	0.0629	0.924
		0.8	$\hat{\beta}_E$	0.6896	0.0964	0.0975	0.906	0.2562	0.0889	0.0943	0.911	-0.1931	0.0601	0.0634	0.917
1.5	0		$\hat{\beta}_V$	0.7060	0.1349	0.1291	0.935	0.2726	0.1215	0.1152	0.937	-0.2017	0.0870	0.0813	0.928
		0.2	$\hat{\beta}_N$	0.0071	0.0914	0.0934	0.000	-0.0011	0.0953	0.0919	0.188	-0.0026	0.0573	0.0564	0.057
		0.8	$\hat{\beta}_N$	0.0054	0.0754	0.0744	0.000	0.0014	0.0757	0.0733	0.068	-0.0024	0.0573	0.0564	0.056
		0.2	$\hat{\beta}_E$	0.6955	0.0960	0.1069	0.938	0.2646	0.0859	0.0858	0.920	-0.1961	0.0600	0.0630	0.930
		0.8	$\hat{\beta}_E$	0.6912	0.0965	0.0959	0.895	0.2563	0.0868	0.0921	0.924	-0.1904	0.0598	0.0627	0.920
1.5	0.8		$\hat{\beta}_V$	0.7075	0.1355	0.1291	0.935	0.2751	0.1241	0.1161	0.927	-0.2016	0.0877	0.0813	0.926
		0.2	$\hat{\beta}_N$	0.0062	0.0941	0.0935	0.000	0.0002	0.0942	0.0925	0.194	-0.0028	0.0574	0.0564	0.065
		0.8	$\hat{\beta}_N$	0.0049	0.0743	0.0742	0.000	0.0014	0.0746	0.0734	0.062	-0.0027	0.0573	0.0564	0.064
		0.2	$\hat{\beta}_E$	0.6947	0.0958	0.1089	0.944	0.2647	0.0868	0.0875	0.932	-0.1962	0.0595	0.0636	0.935
		0.8	$\hat{\beta}_E$	0.6899	0.0960	0.0976	0.903	0.2567	0.0875	0.0944	0.922	-0.1914	0.0597	0.0640	0.930
5.7	0		$\hat{\beta}_V$	0.7060	0.1351	0.1291	0.938	0.2720	0.1213	0.1152	0.946	-0.2012	0.0876	0.0812	0.932
		0.2	$\hat{\beta}_N$	0.0071	0.0914	0.0934	0.00	-0.0006	0.0935	0.0918	0.18	-0.0024	0.0576	0.0564	0.06

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

				β_{11} = log(2) = 0.693				β_{21} = log(1.3) = 0.262				β_2 = -0.2			
θ	r	σ	$\hat{\beta}$	$\hat{\beta}_{11}$	SD	SE	95% CI	$\hat{\beta}_{21}$	SD	SE	95% CI	$\hat{\beta}_2$	SD	SE	95% CI
		0.8	$\hat{\beta}_N$	0.0055	0.0754	0.0744	0.000	0.0014	0.0748	0.0732	0.058	-0.0021	0.0576	0.0564	0.056
		0.2	$\hat{\beta}_E$	0.6951	0.0958	0.1104	0.934	0.2627	0.0864	0.0861	0.903	-0.1959	0.0600	0.0627	0.920
		0.8	$\hat{\beta}_E$	0.6900	0.0945	0.0993	0.907	0.2552	0.0871	0.0924	0.918	-0.1902	0.0597	0.0627	0.908
5.7	0.8		$\hat{\beta}_V$	0.7076	0.1353	0.1291	0.934	0.2748	0.1231	0.1163	0.936	-0.2012	0.0886	0.0811	0.923
		0.2	$\hat{\beta}_N$	0.0063	0.0941	0.0935	0.000	0.0009	0.0922	0.0925	0.196	-0.0025	0.0579	0.0564	0.067
		0.8	$\hat{\beta}_N$	0.0049	0.0743	0.0742	0.000	0.0019	0.0727	0.0733	0.064	-0.0023	0.0578	0.0564	0.066
		0.2	$\hat{\beta}_E$	0.6955	0.0954	0.1112	0.942	0.2635	0.0861	0.0868	0.922	-0.1960	0.0595	0.0639	0.930
		0.8	$\hat{\beta}_E$	0.6892	0.0952	0.1003	0.903	0.2549	0.0880	0.0948	0.921	-0.1912	0.0598	0.0647	0.912

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Relative efficiency for $\hat{\beta}_E$ vs. $\hat{\beta}_V$ with $\beta_{11} = log(2) = 0.693$, $\beta_{21} = log(1.3) = 0.262$, $\beta_2 = -0.2$, r = 0.8, and σ = 0.2

			20% censoring					
			$\hat{\beta}_{11}(\cdot)$		$\hat{\beta}_{21}(\cdot)$		$\hat{\beta}_2(\cdot)$	
Validatin Fraction	θ	$\hat{\beta}$	**SD**	**RE**	**SD**	**RE**	**SD**	**RE**
30%	0.25	$\hat{\beta}_V$	0.1822		0.1691		0.1183	
		$\hat{\beta}_E$	0.0979	3.466	0.0898	3.551	0.0599	3.894
	1.5	$\hat{\beta}_V$	0.1821		0.1704		0.1152	
		$\hat{\beta}_E$	0.0971	3.512	0.0877	3.779	0.0596	3.737
	5.7	$\hat{\beta}_V$	0.1820		0.1705		0.1146	
		$\hat{\beta}_E$	0.0964	3.567	0.0876	3.786	0.0596	3.701
50%	0.25	$\hat{\beta}_V$	0.1358		0.1273		0.0900	
		$\hat{\beta}_E$	0.0959	2.004	0.0877	2.107	0.0597	2.270
	1.5	$\hat{\beta}_V$	0.1355		0.1241		0.0877	
		$\hat{\beta}_E$	0.0958	2.001	0.0868	2.041	0.0595	2.173
	5.7	$\hat{\beta}_V$	0.1353		0.1231		0.0886	
		$\hat{\beta}_E$	0.0954	2.013	0.0861	2.045	0.0595	2.214
70%	0.25	$\hat{\beta}_V$	0.1154		0.1007		0.0723	
		$\hat{\beta}_E$	0.0958	1.452	0.0851	1.399	0.0606	1.423
	1.5	$\hat{\beta}_V$	0.1151		0.1008		0.0729	
		$\hat{\beta}_E$	0.0954	1.456	0.0834	1.459	0.0602	1.463
	5.7	$\hat{\beta}_V$	0.1149		0.1029		0.0732	
		$\hat{\beta}_E$	0.0951	1.460	0.0848	1.470	0.0604	1.469
			50% censoring					
30%	0.25	$\hat{\beta}_V$	0.2267		0.2194		0.1437	

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

			20% censoring					
			$\hat{\beta}_{11}(\cdot)$		$\hat{\beta}_{21}(\cdot)$		$\hat{\beta}_{2}(\cdot)$	
Validatin Fraction	θ	$\hat{\beta}$	**SD**	**RE**	**SD**	**RE**	**SD**	**RE**
50%	0.25	$\hat{\beta}_E$	0.1183	3.672	0.1103	3.959	0.0747	3.700
		$\hat{\beta}_V$	0.2261		0.2129		0.1407	
	1.5	$\hat{\beta}_E$	0.1174	3.713	0.1110	3.676	0.0756	3.463
		$\hat{\beta}_V$	0.2265		0.2071		0.1402	
	5.7	$\hat{\beta}_E$	0.1172	3.734	0.1122	3.408	0.0756	3.437
		$\hat{\beta}_V$	0.1677		0.1610		0.1118	
	0.25	$\hat{\beta}_E$	0.1160	2.088	0.1091	2.179	0.0746	2.245
		$\hat{\beta}_V$	0.1674		0.1566		0.1138	
	1.5	$\hat{\beta}_E$	0.1161	2.079	0.1101	2.023	0.0764	2.217
		$\hat{\beta}_V$	0.1674		0.1566		0.1125	
	5.7	$\hat{\beta}_E$	0.1161	2.078	0.1105	2.007	0.0753	2.230
		$\hat{\beta}_V$	0.1371		0.1301		0.0915	
70%	0.25	$\hat{\beta}_E$	0.1147	1.428	0.1079	1.454	0.0758	1.460
		$\hat{\beta}_V$	0.1367		0.1313		0.0931	
	1.5	$\hat{\beta}_E$	0.1144	1.427	0.1088	1.456	0.0770	1.462
		$\hat{\beta}_V$	0.1367		0.1319		0.0930	
	5.7	$\hat{\beta}_E$	0.1146	1.425	0.1076	1.502	0.0758	1.493

RE is the estimated relative efficiency of $\hat{\beta}_E$ to $\hat{\beta}_V$, which is calculated as $(SD(\hat{\beta}_V)/SD(\hat{\beta}_E))^2$.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 4

SOLVD data analysis results

	Proposed method				Validation method			
factors	**coef**	**exp(coef)**	**se**	***p*-value**	**coef**	**exp(coef)**	**se**	***p*-value**
for heart failure								
LVEF (β_1)	-0.055	0.947	0.008	< 0.001	-0.075	0.928	0.038	0.051
TRT (β_2)	-0.438	0.645	0.125	< 0.001	-0.835	0.434	0.564	0.140
SEX (β_3)	-0.292	0.747	0.152	0.028	0.405	1.499	1.089	0.710
AGE (β_4)	0.025	1.025	0.006	< 0.001	0.035	1.035	0.033	0.300
for non-fatal MI								
LVEF ($\beta_1 + \gamma_1$)	-0.008	0.992	0.011	0.510	-0.012	0.988	0.042	0.780
TRT ($\beta_2 + \gamma_2$)	-0.391	0.676	0.723	0.294	-0.703	0.495	0.875	0.420
SEX ($\beta_3 + \gamma_3$)	0.042	1.043	0.794	0.479	-0.838	0.433	1.058	0.430
AGE ($\beta_4 + \gamma_4$)	0.004	1.004	0.019	0.407	0.003	1.003	0.0323	0.920