Published in final edited form as:

Biometrics. 2011 March ; 67(1): 18–28. doi:10.1111/j.1541-0420.2010.01445.x.

Proportional hazards regression for the analysis of clustered survival data from case-cohort studies

Hui Zhang, Douglas E. Schaebel, and John D. Kalbeisch

Hui Zhang: huizh@umich.edu

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.

Summary

Case-cohort sampling is a commonly used and efficient method for studying large cohorts. Most existing methods of analysis for case-cohort data have concerned the analysis of univariate failure time data. However, clustered failure time data are commonly encountered in public health studies. For example, patients treated at the same center are unlikely to be independent. In this article, we consider methods based on estimating equations for case-cohort designs for clustered failure time data. We assume a marginal hazards model, with a common baseline hazard and common regression coefficient across clusters. The proposed estimators of the regression parameter and cumulative baseline hazard are shown to be consistent and asymptotically normal, and consistent estimators of the asymptotic covariance matrices are derived. The regression parameter estimator is easily computed using any standard Cox regression software that allows for offset terms. The proposed estimators are investigated in simulation studies, and demonstrated empirically to have increased efficiency relative to some existing methods. The proposed methods are applied to a study of mortality among Canadian dialysis patients.

Keywords

Case-cohort study; Clustered data; Cox model; Estimating equation; Robust variance; Survival analysis

1. Introduction

The case-cohort design is commonly used in large cohort studies. The design entails collecting covariate data for all subjects who experienced the event of interest (cases) in the full cohort, and for a random sample (the subcohort) from the entire cohort. Therefore, the most important advantage of this design is cost savings, especially when the disease is rare. A second advantage of the case-cohort design is that the subcohort can be used as the comparison group for multiple disease outcomes. A number of methods have been proposed for regression analysis of case-cohort data under the proportional hazards model. Prentice (1986) proposed a pseudo-likelihood method for estimating the regression parameter. Self

Correspondence to: Hui Zhang, huizh@umich.edu.

Supplementary Materials

The Web Appendix, referenced in Section 3, is available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>

and Prentice (1988) and Lin and Ying (1993), using different approaches, derived large sample properties of the pseudo-likelihood related estimators. Wacholder et al. (1989) presented variance estimators for the log relative hazard through a bootstrap resampling plan. Barlow (1994) proposed a computationally convenient robust variance estimator. Chen and Lo (1999) suggested a class of estimating functions which in many cases offered improved efficiency. Therneau and Li (1999) and Langholz and Jiao (2007) described the computation of parameter and variance estimates using common software packages, such as SAS and R/S-PLUS. Borgan et al. (2000), Chen (2001) and Samuelsen, Anestad and Skrdondal (2007) obtained more efficient estimators by different approaches. Sorensen and Andersen (2000) considered competing risks analysis of case-cohort data.

The case-cohort design has also been studied in the context of other regression models. For example, Kulich and Lin (2000), Sun, Sun and Flournoy (2004) and Ma (2007) studied the case-cohort design under an additive hazards regression model. Chen (2001) and Kong, Cai and Sen (2004, 2006) considered semiparametric transformation models in the case-cohort design. Nan, Yu and Kalbfleisch (2006) and Nan, Kalbfleisch and Yu (2009) considered accelerated failure time models and rank based analyses in case-cohort designs.

Each of the studies in the preceding paragraphs focused on univariate failure time data. However, clustered failure time data are commonly encountered in biomedical research. For example, in a family disease study, members from the same family may be correlated due to shared genetic and/or environmental factors. Similarly, outcomes of patients treated at the same center may be correlated. In these cases, valid statistical inference requires that one account for the intra-cluster dependence. Methods proposed for handling clustered failure time data can generally be categorized into two approaches: conditional models and marginal models. As an example of a conditional approach, frailty models specify the correlation structure by postulating a random effect (frailty) that is common to individuals within the same cluster. The regression parameter for such models is interpreted conditional on the random effect. For example, Moger, Pawitan and Borgan (2008) proposed frailty based case-cohort methods for analyzing family survival data with families as the sampling unit. If the investigator is interested in population averaged covariate effects, a marginal model is appealing; such a model leaves the dependence structure unspecified in the model formulation, but adjusts for the dependence in the inference. Several methods have been proposed for fitting marginal proportional hazards models; e.g. Wei, Lin and Weissfeld (1989); Lee, Wei and Amato (1992); Cai and Prentice (1995); Spiekerman and Lin (1998); Lu and Wang (2005). Lu and Shih (2006) considered case-cohort designs adapted to clustered failure time data under a marginal model and developed inference procedures.

Our proposed method is motivated by a retrospective cohort study of a possible day-of-week effect on death rates among patients receiving hemodialysis to treat advanced kidney failure. Patients treated at the same renal center are likely to be correlated due to center-specific practice patterns as well as a tendency to share socio-economic and environmental characteristics. The dialysis schedule, Monday/Wednesday/Friday (M/W/F) or Tuesday/Thursday/Saturday (T/T/S), may put patients at higher risk of death on certain days. For example, patients may have higher risk of death on Monday and Tuesday since, on average, these days follow the longest intervals without dialysis.

In this article, we propose methods based on estimating equations for three case-cohort designs that are applicable to clustered survival data. We assume a marginal proportional hazards model with a common baseline hazard and common regression coefficient across clusters. The case-cohort sampling designs we consider are similar to those proposed by Lu & Shih (2006). However, the designs we propose feature Bernoulli sampling, which is convenient for establishing theoretical properties. More importantly, we construct the risk sets using not only the information in the subcohort, but also the information collected on future deaths, similar to Chen & Lo (1999). As a result, the proposed estimators have increased efficiency relative to those of Lu and Shih (2006).

The remainder of this article is organized as follows. In Section 2, we describe the proposed estimation procedures. In Section 3, we derive large sample properties for the proposed estimators. We conduct simulation studies in Section 4 to investigate the finite sample properties of the proposed estimators. In Section 5, we apply the proposed methods to a national organ failure database. The article concludes with some discussion in Section 6. All proofs are presented in the Web Appendix.

2. Proposed Methods

We first describe case-cohort designs with Bernoulli sampling for clustered failure time data. The full cohort consists of n independent clusters, and the i th cluster ($i = 1, \dots, n$) has m_i correlated subjects. We assume that subjects within the same cluster are exchangeable. In advance of follow-up, a random sample of the entire cohort, called the subcohort, is selected. Covariate data are then collected from individuals in the subcohort as well as those observed to fail in the entire cohort. Three designs are considered to obtain the subcohort:

- **Design A:** Randomly sample individuals from each cluster with Bernoulli sampling. That is, each individual in each cluster has an independent fixed probability of being selected to the subcohort.
- **Design B:** Randomly sample clusters from the full cohort with Bernoulli sampling.
- **Design C:** Randomly sample clusters from the full cohort with Bernoulli sampling, then randomly sample subjects with Bernoulli sampling from the selected clusters.

These are the same designs proposed by Lu & Shih (2006), except that we consider Bernoulli sampling, which greatly simplifies asymptotic derivations. Note that Design A and Design B are special cases of Design C.

Let T_{ij} and C_{ij} be the failure time and censoring time, where (i, j) represents the j th subject in the i th cluster. Let $\mathbf{Z}_{ij}(t)$ be the p -vector of possibly time-dependent covariates; with any time-dependent covariates assumed to be external (Kalbfleisch and Prentice, 2002). We assume that T_{ij} and C_{ij} are independent conditional on the observed covariates. Let $X_{ij} = T_{ij} \wedge C_{ij}$, $Y_{ij}(t) = I(X_{ij} > t)$, $\delta_{ij} = I(T_{ij} < C_{ij})$, and $N_{ij}(t) = I(T_{ij} \leq C_{ij} \wedge t)$, where $I(\cdot)$ is the indicator function and $a \wedge b = \min\{a, b\}$. We assume that $\{N_{ij}(\cdot), Y_{ij}(\cdot), \mathbf{Z}_{ij}(\cdot), m_i : j = 1, \dots, m_i\}$ are independently and identically distributed for $i = 1, \dots, n$. Let H_i indicate whether or not cluster i is selected into the subcohort, and let H_{ij} be the indicator for subject (i, j) being sampled as a potential individual in the subcohort. Subject (i, j) is selected into the subcohort

if and only if $H_i H_{ij} = 1$. The variates H_i and H_{ij} are assumed to be independent of $\{N_{ij}(\cdot), Y_{ij}(\cdot), \mathbf{Z}_{ij}(\cdot), m_i : j = 1, \dots, m_i\}$, for all i, j . Under Design A, B, and C, the H_i 's are independent Bernoulli variables with $\mathcal{E}(H_i) = \gamma$ for all $i = 1, \dots, n$, where $\mathcal{E}(\cdot)$ denotes expectation, and the H_{ij} 's are independent Bernoulli variables with $\mathcal{E}(H_{ij}) = \theta$, for all $i = 1, \dots, n$ and $j = 1, \dots, m_i$. Under Design A, $H_i = 1$, for all $i = 1, \dots, n$, i.e., $\gamma = 1$. Under Design B, $H_{ij} = 1$, for all $i = 1, \dots, n, j = 1, \dots, m_i$; i.e., $\theta = 1$.

Let the marginal hazard of failure of individual (i, j) be specified by a proportional hazards model (Cox, 1972),

$$\lambda_{ij}(t) = \lambda_0(t) e^{\beta_0^T \mathbf{Z}_{ij}(t)}, \quad (1)$$

where $\lambda_0(\cdot)$ is an unspecified marginal baseline hazard function and β_0 is a p -dimensional regression parameter. Since we are primarily interested in the estimation of β_0 , we leave the dependence structure of individuals within a cluster unspecified.

Many authors have studied the estimation of the regression parameters under model (1). Under a working independence assumption, Lee, Wei and Amato (1992) proposed the estimating function

$$\mathbf{U}_{LWA}(\beta) = \sum_{i=1}^n \sum_{j=1}^{m_i} \int_0^\tau \{ \mathbf{Z}_{ij}(u) - \mathbf{E}_{LWA}(\beta, u) \} dN_{ij}(u),$$

where $\tau < \infty$ equals the maximum follow-up time,

$\mathbf{E}_{LWA}(\beta, u) = \mathbf{S}_{LWA}^{(1)}(\beta, u) / \mathbf{S}_{LWA}^{(0)}(\beta, u)$, $\mathbf{S}_{LWA}^{(d)}(\beta, u) = n^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} Y_{ij}(u) e^{\beta^T \mathbf{Z}_{ij}(u)} \mathbf{Z}_{ij}(u)^{\otimes d}$, with $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a} \mathbf{a}^T$. Then, β_0 of model (1) can be estimated with $\hat{\beta}_{LWA}$, the solution to the estimating equation $\mathbf{U}_{LWA}(\beta) = 0$. Lu and Shih (2006) considered case-cohort designs for clustered failure time data under model (1) and proposed to estimate β_0 with $\hat{\beta}_{LS}$, the root of the estimating equation $\mathbf{U}_{LS}(\beta) = 0$, where

$$\mathbf{U}_{LS}(\beta) = \sum_{i=1}^n \sum_{j=1}^{m_i} \int_0^\tau \{ \mathbf{Z}_{ij}(u) - \mathbf{E}_{LS}(\beta, u) \} dN_{ij}(u),$$

where $\mathbf{E}_{LS}(\beta, u) = \mathbf{S}_{LS}^{(1)}(\beta, u) / \mathbf{S}_{LS}^{(0)}(\beta, u)$ and

$$\mathbf{S}_{LS}^{(d)}(\beta, u) = n^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} H_i H_{ij} Y_{ij}(u) \mathbf{Z}_{ij}(u)^{\otimes d}.$$

Lu and Shih (2006) used only subcohort subjects to construct the risk set. Since information on all failures in the full cohort are available, failures outside the subcohort can also contribute to the risk set, as proposed by Chen and Lo (1999) for independent subjects. We propose three procedures to estimate β_0 , the procedures differing with respect to their treatment of the marginal observed-event probability, $Pr(\delta_{ij} = 1)$, which we denote by p_0 . In the first proposed procedure, p_0 is assumed known, which follows the Chen & Lo (1999)

approach. Usually, p_0 is not known, but this gives a baseline to which other approaches can be compared. We estimate β_0 by $\hat{\beta}_s$, the solution to $U(\beta, p_0) = 0$, where

$$U(\beta, p) = \sum_{i=1}^n \sum_{j=1}^{m_i} \int_0^\tau \{Z_{ij}(u) - \bar{E}(\beta, p, u)\} dN_{ij}(u) \quad (2)$$

$$\bar{E}(\beta, p, u) = \frac{\bar{S}^{(1)}(\beta, p, u)}{\bar{S}^{(0)}(\beta, p, u)}$$

$$\bar{S}^{(d)}(\beta, p, u) = \sum_{i=1}^n \sum_{j=1}^{m_i} \left\{ \frac{p}{N_1} \delta_{ij} + \frac{1-p}{n_0} (1 - \delta_{ij}) H_i H_{ij} \right\} Y_{ij}(u) e^{\beta^T Z_{ij}(u)} Z_{ij}(u)^{\otimes d}$$

with $N_1 = \sum_{i=1}^n \sum_{j=1}^{m_i} \delta_{ij}$, and $n_0 = \sum_{i=1}^n \sum_{j=1}^{m_i} (1 - \delta_{ij}) H_i H_{ij}$. The motivation for building estimating equation (2) is that (β, p_0, u) is a consistent estimator of $\mathcal{E}\{Z_{ij}(u)|X_{ij} = u, \delta_{ij} = 1\}$, where

$$\begin{aligned} & \mathcal{E}\{Z_{ij}(u)|X_{ij}=u, \delta_{ij}=1\} \\ &= \frac{\mathcal{E}\{Y_{ij}(u)Z_{ij}(u)e^{\beta^T Z_{ij}(u)}\}}{\mathcal{E}\{Y_{ij}(u)e^{\beta^T Z_{ij}(u)}\}} \\ &= \frac{p_0 \mathcal{E}\{Y_{ij}(u)Z_{ij}(u)e^{\beta^T Z_{ij}(u)}|\delta_{ij}=1\} + (1-p_0) \mathcal{E}\{Y_{ij}(u)Z_{ij}(u)e^{\beta^T Z_{ij}(u)}|\delta_{ij}=0\}}{p_0 \mathcal{E}\{Y_{ij}(u)e^{\beta^T Z_{ij}(u)}|\delta_{ij}=1\} + (1-p_0) \mathcal{E}\{Y_{ij}(u)e^{\beta^T Z_{ij}(u)}|\delta_{ij}=0\}}. \end{aligned} \quad (3)$$

The first (second) conditional means in numerator and denominator can be estimated by their respective empirical counterparts from all failures in the whole cohort (controls in the subcohort). A derivation of (3) is given in the Web Appendix.

In almost all settings, the population failure probability, p_0 , is unknown but can be estimated using the subcohort case proportion, $\hat{p}_s$, or the full cohort case proportion, $\hat{p}_w$. These give rise to estimating functions $U(\beta, \hat{p}_s)$ and $U(\beta, \hat{p}_w)$, with solutions $\hat{\beta}_s$ and $\hat{\beta}_w$, respectively. In cases where the study cohort is well defined, $\hat{p}_w$ can be computed and used to obtain $\hat{\beta}_w$, which has the most practical value. When the study cohort is less well-defined, $\hat{\beta}_s$ is a suitable alternative. For example, if the study does not have a roster for the full cohort (such that the cohort size, N , is not known), then $\hat{\beta}_s$ can still be used.

Some simple algebra shows that

$$\bar{S}^{(d)}(\beta, \hat{p}_s, u) = \frac{\hat{q}_1}{n_0 + n_1} \sum_{i=1}^n \sum_{j=1}^{m_i} \left\{ \delta_{ij} + \frac{1}{\hat{q}_1} (1 - \delta_{ij}) H_i H_{ij} \right\} Y_{ij}(u) e^{\beta^T Z_{ij}(u)} Z_{ij}(u)^{\otimes d}$$

$$\bar{S}^{(d)}(\beta, \hat{p}_w, u) = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} \left\{ \delta_{ij} + \frac{1}{\hat{q}_0} (1 - \delta_{ij}) H_i H_{ij} \right\} Y_{ij}(u) e^{\beta^T \mathbf{Z}_{ij}(u)} \mathbf{Z}_{ij}(u)^{\otimes d},$$

where $N = N_0 + N_1$, $q_1 = n_1/N_1$, and $q_0 = n_0/N_0$, with $n_1 = \sum_{i=1}^n \sum_{j=1}^{m_i} H_i H_{ij} \delta_{ij}$ and $N_0 = \sum_{i=1}^n \sum_{j=1}^{m_i} (1 - \delta_{ij})$. The estimating equations are similar, therefore, to those arising from inverse sampling probability weighting (ISPW), such as that proposed by Kalbfleisch and Lawless (1988) and Borgan et al. (2000) for the Cox model; Kulich and Lin (2000) for the additive hazards model; and Nan, Kalbfleisch and Yu (2009) for the accelerated failure time model. These studies focused on univariate failure time data.

The cumulative baseline hazard function, $\Lambda_0(t) = \int_0^t \lambda_0(u) du$, can be consistently estimated by

$$\hat{\Lambda}_0(t; \hat{\beta}, \hat{p}) = \int_0^t \frac{d\bar{N}(u)}{\hat{\mu} \bar{S}^{(0)}(\hat{\beta}, \hat{p}, u)}, \quad (4)$$

where $\bar{N}(u) = n^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} N_{ij}(u)$, $\mu = \mathcal{E}(m_i)$, and $\hat{\mu} = n^{-1} \sum_{i=1}^n m_i$. In (4), either $p_s, \hat{p}_w$ or p_0 could be used.

The proportional hazards assumption may be violated for one or more covariates, which could be individual level covariates such as age or time since first dialysis, or cluster level covariates such as center size. Our proposed methods can be extended to allow for stratification on such covariates; details can be found in the Web Appendix Section A.5.

3. Asymptotic Properties of the Proposed Estimators

We make the following assumptions:

- a. $\{N_{ij}(\cdot), Y_{ij}(\cdot), \mathbf{Z}_{ij}(\cdot), m_i : j = 1, \dots, m_i\}, i = 1, \dots, n$ are independently and identically distributed.
- b. $P\{Y_{ij}(t) = 1\} > 0$ for $t \in (0, \tau]$, $i = 1, \dots, n, j = 1, \dots, m_i$, and all m_i .
- c. $|Z_{ijh}(0)| + \int_0^\tau |dZ_{ijh}(t)| < B_Z < \infty$ for $i = 1, \dots, n, j = 1, \dots, m_i$, and all m_i , where Z_{ijh} is the h th component of $\mathbf{Z}_{ij}$ and B_Z is a constant.
- d. There exists a neighborhood $\mathcal{B}$ of β_0 such that

$$\sup_{u \in [0, \tau], \beta \in \mathcal{B}} \|S^{(d)}(\beta, u) - s^{(d)}(\beta, u)\| \xrightarrow{P} 0 \text{ for } d = 0, 1, 2, \text{ where } s^{(d)}(\beta, u) = \mathcal{E}\{S^{(d)}(\beta, u)\}$$
 is absolutely continuous, for $\beta \in \mathcal{B}$, uniformly in $u \in (0, \tau]$. Moreover, $s^{(0)}(\beta, u)$ is assumed to be bounded away from zero.
- e.

$$\sup_{u \in [0, \tau], \beta \in \mathcal{B}} \|\bar{S}^{(d)}(\beta, p, u) - \mu^{-1} s^{(d)}(\beta, u)\| \xrightarrow{P} 0.$$
 For $d = 0, 1, 2$,
- f. The matrix $A(\beta_0)$ is positive definite, where

$$\mathbf{A}(\boldsymbol{\beta}) = \int_0^\tau \{s^{(2)}(\boldsymbol{\beta}, u)/s^{(0)}(\boldsymbol{\beta}, u) - e(\boldsymbol{\beta}, u)^{\otimes 2}\} dF(u)$$

with $e(\boldsymbol{\beta}, u) = s^{(1)}(\boldsymbol{\beta}, u)/s^{(0)}(\boldsymbol{\beta}, u)$, and $F(u) = \mathcal{E}\{N(u)\}$.

g. $\Lambda_0(\tau) < \infty$, and $\lambda_0(t)$ is absolutely continuous for $t \in (0, \tau]$.

Our main results are given in Theorems 1 – 4 below, the proofs of which are given in the Web Appendix. We provide only brief summary remarks about the proofs below.

Theorem 1

Under conditions (a) – (g), as $n \rightarrow \infty$, $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, p_0)$ converges to a mean zero Normal distribution with covariance $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0, p_0) = \mathcal{E}\{\mathbf{W}_1(\boldsymbol{\beta}_0, p_0)^{\otimes 2}\}$, with

$$\begin{aligned} \mathbf{W}_i(\boldsymbol{\beta}, p) &= \sum_{j=1}^{m_i} \int_0^\tau \{Z_{ij}(u) - e(\boldsymbol{\beta}, u)\} \\ &\quad \left[dN_{ij}(u) - \left\{ \frac{1}{\mu} \delta_{ij} + \frac{1}{\mu\gamma\theta} (1 - \delta_{ij}) H_i H_{ij} \right\} Y_{ij}(u) e^{\boldsymbol{\beta}^T \mathbf{Z}_{ij}(u)} \right. \\ &\quad \left. \times \{\mu^{-1} s^{(0)}(\boldsymbol{\beta}, u)\}^{-1} dF(u) \right] + \mathbf{D}_1(\boldsymbol{\beta}) G_{1i}(p) + \mathbf{D}_2(\boldsymbol{\beta}) G_{2i}(p) \\ \mathbf{D}_1(\boldsymbol{\beta}) &= \mathcal{E} \left[\sum_{j=1}^{m_1} \int_0^\tau \{Z_{1j}(u) - e(\boldsymbol{\beta}, u)\} \frac{\delta_{1j}}{\mu^2 p} Y_{1j}(u) e^{\boldsymbol{\beta}^T \mathbf{Z}_{1j}(u)} \frac{\sum_{k=1}^n \sum_{l=1}^{m_k} dN_{kl}(u)}{\mu^{-1} s^{(0)}(\boldsymbol{\beta}, u)} \right] \\ \mathbf{D}_2(\boldsymbol{\beta}) &= \mathcal{E} \left[\sum_{j=1}^{m_1} \int_0^\tau \{Z_{1j}(u) - e(\boldsymbol{\beta}, u)\} \frac{(1 - \delta_{1j}) H_1 H_{1j}}{(\mu\gamma\theta)^2 (1 - p)} Y_{1j}(u) e^{\boldsymbol{\beta}^T \mathbf{Z}_{1j}(u)} \frac{\sum_{k=1}^n \sum_{l=1}^{m_k} dN_{kl}(u)}{\mu^{-1} s^{(0)}(\boldsymbol{\beta}, u)} \right] \\ G_{1i}(p) &= n^{-1} \left(\sum_{j=1}^{m_i} \delta_{ij} - \mu p \right) \\ G_{2i}(p) &= n^{-1} \left\{ \sum_{j=1}^{m_i} (1 - \delta_{ij}) H_i H_{ij} - \mu\gamma\theta(1 - p) \right\}. \end{aligned}$$

In the Web Appendix, we show that $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, p_0) = n^{-1/2} \sum_{i=1}^n \mathbf{W}_i(\boldsymbol{\beta}_0, p_0) + o_p(1)$; hence, $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0, p_0)$ is essentially a scaled sum of n independent and identically distributed random quantities with mean zero and finite variance. The proof of asymptotic normality

follows from the Multivariate Central Limit Theorem (MCLT) and various results from empirical process theory. The result in Theorem 1 is used to derive the limiting distribution of the proposed estimators.

Theorem 2

Under conditions (a) – (g), $\hat{\beta}_t$ converges in probability to β_0 and $n^{1/2}(\hat{\beta}_t - \beta_0)$ converges in distribution to a mean zero normal distribution with covariance matrix $A(\beta_0)^{-1}\Sigma(\beta_0, p_0)A(\beta_0)^{-1}$.

The proof of the consistency of $\hat{\beta}_t$ follows by the Inverse Function Theorem (Foutz, 1977). The proof of asymptotic normality follows from a Taylor series expansion and the Cramér-Wold device.

Theorem 3

Under conditions (a) – (g), both $\hat{\beta}_s$ and $\hat{\beta}_w$ converge in probability to β_0 , and each of $n^{1/2}(\hat{\beta}_s - \beta_0)$ and $n^{1/2}(\hat{\beta}_w - \beta_0)$ converges in distribution to a zero-mean Normal with covariance matrices $A(\beta_0)^{-1}\Omega_s(\beta_0)A(\beta_0)^{-1}$ and $A(\beta_0)^{-1}\Omega_w(\beta_0)A(\beta_0)^{-1}$ respectively, where for $a = s$ or w , $\Omega_a(\beta) = \mathcal{E}\{\psi_1^a(\beta, p_0)^{\otimes 2}\}$, $\psi_i^a(\beta, p) = W_i(\beta, p) + B(\beta)Q_i^a(p)$, with $Q_i^s(p) = \{\mu\gamma\theta\}^{-1}\sum_{j=1}^{m_i} H_i H_{ij}(\delta_{ij} - p)$, $Q_i^w(p) = \mu^{-1}\sum_{j=1}^{m_i} (\delta_{ij} - p)$, and

$$B(\beta) = \int_0^\tau \left\{ \frac{s^{(1)}(\beta, u)r^{(0)}(\beta, u)}{s^{(0)}(\beta, u)^2} - \frac{r^{(1)}(\beta, u)}{s^{(0)}(\beta, u)} \right\} dF(u)$$

$$r^{(d)}(\beta, u) = \frac{1}{p_0} \mathcal{E} \left\{ \delta_{11} Y_{11}(u) e^{\beta^T \mathbf{Z}_{11}(u)} \mathbf{Z}_{11}(u)^{\otimes d} \right\} - \frac{1}{1 - p_0} \mathcal{E} \left\{ (1 - \delta_{11}) Y_{11}(u) e^{\beta^T \mathbf{Z}_{11}(u)} \mathbf{Z}_{11}(u)^{\otimes d} \right\}.$$

The results in Theorem 1, combined with two Taylor series expansions, the MCLT and Slutsky's Theorem, conclude the proof of asymptotic normality of $\hat{\beta}_s$ and $\hat{\beta}_w$ in Theorem 3. The covariance matrices in Theorems 2 and 3 can be consistently estimated from the observed case-cohort data, as described in the Web Appendix.

We now describe asymptotic results pertaining to the proposed baseline cumulative hazard estimator.

Theorem 4

Under conditions (a) – (g), $\hat{\Lambda}_0(\hat{\beta}, \hat{p}, \hat{\tau})$ converges in probability to $\Lambda_0(t)$ uniformly in $t \in [0, \tau]$, and $n^{1/2} \{\hat{\Lambda}_0(\hat{\beta}, \hat{p}, \hat{\tau}) - \Lambda_0(t)\}$ converges weakly to a Gaussian process with mean zero and covariance function at (s, t) given by $\mathcal{E} \{ \phi_1(\beta_0, p_0, s) \phi_1(\beta_0, p_0, t) \}$, where

$$\phi_i(\beta, p, t) = k(\beta, p, t)Q_i(p) + \mathbf{h}^T(\beta, p, t)A(\beta)\psi_i(\beta, p) + \chi_i(\beta, p, t)$$

$$\chi_i(\boldsymbol{\beta}, p, t) = \sum_{j=1}^{m_i} \int_0^t \frac{1}{s^{(0)}(\boldsymbol{\beta}, u)} dM_{ij}(u) + \sum_{j=1}^{m_i} \int_0^t \frac{1}{s^{(0)}(\boldsymbol{\beta}, u)^2} \left\{ 1 - \delta_{ij} - \frac{1}{\gamma\theta} (1 - \delta_{ij}) H_i H_{ij} \right\} Y_{ij} e^{\boldsymbol{\beta}^T \mathbf{Z}_{ij}(u)} dF(u)$$

$$k(\boldsymbol{\beta}, p, t) = - \int_0^t \frac{\mu r^{(0)}(\boldsymbol{\beta}, p, u)}{s^{(0)}(\boldsymbol{\beta}, u)} d\Lambda_0(u)$$

$$h(\boldsymbol{\beta}, p, t) = - \int_0^t e(\boldsymbol{\beta}, u) d\Lambda_0(u).$$

A sketch of the proof is given in Web Appendix.

4. Numerical Studies

We conducted simulation studies to investigate the finite sample properties of the estimators proposed in Section 2, and to compare the proposed methods with those of Lu and Shih (2006). We generated clustered failure time data from $n = 100$ clusters. Cluster sizes, m_i , were simulated from a Binomial (50, 0.8) distribution for $i = 1, \dots, n$, with $\mu = \mathcal{E}(m_i) = 40$. The covariate Z_{ij} took values 1 and 0, with probabilities 0.5 and 0.5 respectively. The failure time for the j th subject within the i th cluster was simulated from a distribution with conditional hazard function

$$\lambda_{ij}(t|Z_{ij}, Q_i) = Q_i h_0(t) \exp\{\xi_0 Z_{ij}\},$$

where Q_i is a frailty variable following a positive stable distribution with index $\alpha = 0.8$. The variate Q_i is generated following the method in Chambers et al. (1976),

$$Q_i = \frac{\sin(\alpha Q_{1i})}{\{\sin(Q_{1i})\}^{1/\alpha}} \left[\frac{\sin\{(1 - \alpha)Q_{1i}\}}{Q_{2i}} \right]^{(1-\alpha)/\alpha},$$

where Q_{1i} follows a $U(0, \pi)$ distribution, Q_{2i} follows an exponential distribution with mean 1, and Q_{1i} and Q_{2i} are independent. The baseline hazard function is given by $h_0(t) = \alpha^{-1} t^{\alpha-1}$, with ξ_0 set to $\log(0.5)/\alpha = -0.8664$ or 0. The resulting marginal hazard function is $\lambda_{ij}(t|Z_{ij}) = \lambda_0(t) \exp\{\beta_0 Z_{ij}\}$, and the marginal baseline hazard function is given by $\lambda_0(t) = 1, 0 \leq t < \infty, \beta_0 = \alpha \xi_0 = \log(0.5)$ or 0. The censoring times C_{ij} were constant and equal to 1, which led to average observed event probabilities of $p = 0.51$ or $p = 0.63$. For each data generation, for Design A, individuals within each cluster were selected into the subcohort by Bernoulli sampling with equal probability 0.2 or 0.15. For Design B, we selected clusters by Bernoulli sampling with probability 0.2 or 0.15. For Design C, we first sampled clusters by Bernoulli sampling with probability 0.4 or 0.3, then sampled individuals from those selected clusters by Bernoulli sampling with probability 0.5. Therefore for each design, we would expect approximately 800 or 600 individuals in the subcohort. In another data configuration,

$\beta_0 = \log(0.5)$, the marginal baseline hazard function is given by $\lambda_0 = 0.2$. The covariates Z_{ij} follows either a Bernoulli distribution, which takes value 1 with probability 0.5, or a Normal distribution with mean 0 and variance 1. The other settings were the same except that only approximately 800 individuals were sampled in the subcohort. In this data configuration, the average observed event probabilities are $p = 0.14$ and $p = 0.21$. Each data configuration was replicated 1000 times. The true case percentage, p_0 , would typically be unknown in real world settings; however, it is of course available in our simulation study and is evaluated for comparison purposes.

Tables 1 and 2 display the results of our proposed estimators and those of Lu and Shih (2006). For each data configuration, we list the empirical bias (BIAS) and standard deviation (ESD), average asymptotic standard error (ASE), asymptotic relative efficiency (ARE) with respect to the full cohort and empirical coverage probability (CP). Each of the estimators is approximately unbiased, and the variance estimators appear to be reasonably accurate. The 95% empirical coverage probabilities are generally close to the nominal value. In Table 1, for Design B, slight under-estimation of the standard error and under-coverage occur when $\beta_0 = 0$ and $n_s = 600$. This is due to the small number of clusters in the subcohort. For Design B, clusters are sampled and all individuals in the selected clusters are kept in the subcohort. Little extra information is gained when more subjects in the same cluster are included, since subjects within cluster are correlated. However, more information is available when the number of sampled clusters increases and, correspondingly, the under-coverage is reduced when n_s is increased to 800.

In Table 1, the proposed method appears to be more efficient than that of Lu and Shih (2006), at least for the examples considered. In comparing the proposed sampling designs, for approximately equal subcohort sizes, it appears that Design A is more efficient than Design C, which is more efficient than Design B. This can be attributed to differences in the number of clusters sampled and the resulting differences in the amount of independent information contained in the subcohort. This efficiency gain is more obvious when the covariate is cluster-specific (Web Table 6). In Table 2, the efficiency gain of the proposed methods over those of Lu and Shih (2006) is less evident in the presence of a lower event rate. This can be explained by there being fewer failures outside the subcohort to include in the risk sets.

Additional scenarios have been evaluated in order to examine various aspects, such as continuous covariates, stronger correlation among failure times, smaller number of clusters, smaller subcohort size, lower event rate, as well as the performance of the stratified methods. Results of several of these numerical studies are available in the Web Appendix. In the examples we evaluated, the proposed methods generally work well.

Also in the Web Appendix, the estimates based on the proposed methods are compared to those based on simple random sampling (SRS), and to an inverse sampling probability weighting (ISPW) method. The proposed methods do not lose efficiency relative to the SRS or ISPW methods, at least for the set-ups considered.

For the data settings with $\beta_0 = \log(0.5)$, $\lambda_0(t) = 1$ and $n_s = 800$, we calculated the average of the estimate of $\Lambda_0(t)$ at $t = 0.02, t = 0.04, \dots, t = 1.0$ based on 1000 replications. Figure 1 displays the average point estimate for the cumulative baseline hazards. The true cumulative baseline hazard is also included for comparison purposes. There appears to be no bias for our proposed estimators. We next assumed that the marginal baseline hazard function is given by $\lambda_0(t) = t$. Under this configuration, the proposed estimate is approximately unbiased.

5. Application

We applied the proposed methods to the estimation of the day-of-week effect among Canadian hemodialysis (HD) patients. The 1,276 patients who initiated HD between January 1, 1990 and December 31, 1990 were included in the analysis. Patients were followed from the time they first received HD until the time of death caused by cardiovascular disease (CVD), receiving transplantation, switching to peritoneal dialysis, loss to follow up, or last day of observation (December 31, 1998), whichever occurred first. Patients were clustered by center. In total, there were 70 centers yielding clusters with 1 to 75 patients and a mean of 18.2. Design A was chosen since, all else equal, it is generally at least as efficient as Designs B and C.

The primary outcome of interest is CVD death, and the covariate of interest is day of week (Sunday, Monday, ..., Saturday), which was coded using time-dependent covariates, where $Z_{ij1}(t) = I\{\text{day } t, \text{ for subject } (i, j), \text{ is a Monday}\}, \dots, Z_{ij6}(t) = I\{\text{day } t, \text{ for subject } (i, j), \text{ is a Saturday}\}$, with Sunday chosen as the reference day, where t is the time since initiation of HD for patient i, j . Adjustment covariates included age, gender, region, comorbid conditions and primary renal diagnosis. Age was categorized into 6 groups: <18, 18–39, 40–49, 50–59, 60–69, and 70, and was adjusted for through stratification. Patients from the same renal center may be correlated due to shared practice patterns. Therefore, one needs to account for such intra-cluster dependence for valid statistical inference. In total, there were 249 observed CVD deaths; hence, the event fraction for the full cohort was 0.195. In stratum 1 to 6, the numbers of CVD deaths were 0 (out of 24), 13 (out of 253), 14 (out of 179), 54 (out of 232), 87 (out of 313) and 81 (out of 275), respectively. We analyzed the data using Design A with sampling probability of 0.2. A total of 251 patients was selected into the subcohort. The point estimates were obtained using PROC PHREG in SAS with OFFSET terms, while the variance estimates were calculated using PROC IML. For comparison purposes, we also carried out a full cohort analysis and an analysis with the method of Lu & Shih (2006).

Results of the analysis are shown in Table 3. Using p_s , patients are estimated to have 1.36 and 1.68 times higher hazards of CVD death on Mondays and Tuesdays, respectively, compared to Sundays. Results based on p_w were similar. Results from the full cohort analysis were close to those from our case-cohort analyses, with smaller standard errors. Results based on the method of Lu & Shih (2006) were also similar to ours, with larger standard errors.

The cumulative baseline hazards for each age-specific stratum are exhibited in Figure 2. Each sub-figure contains 3 lines, which correspond to the cumulative baseline hazards for

Design A methods SC and WC, as well as the full cohort analysis. Since no CVD deaths occurred in stratum 1, cumulative baseline hazard estimation is not available for this stratum. In general, the proposed cumulative baseline hazard estimates are close to those for full cohort analysis. The exception was stratum 4, for which the SC and WC estimators are considerably above the full cohort estimator. To examine this phenomenon further, we reanalyzed the data several times (results not shown) which of course involves selecting different subcohorts. Based on this exercise, it appears that the disparity between the SC or WC estimator and the full cohort estimator in any stratum (including stratum 4) is due to sampling variation. In fact, when we drew several bootstrap samples and carried out full-cohort analyses of each, the variability in the estimates of the cumulative hazards was quite large. This suggests that the sampling variation we observed in the case-cohort cumulative hazard estimators was largely inherited from that in the full cohort analysis.

6. Discussion

The case-cohort design has been widely studied for univariate failure time data. Lu and Shih (2006) extended the case-cohort design to clustered failure time data. With respect to parameter estimation, compared to Lu and Shih's methods, the methods we propose feature risk sets which use future cases in addition to subcohort subjects. We demonstrate empirically that the proposed estimators have increased efficiency relative to the methods of Lu and Shih (2006), and that our asymptotic results are applicable to finite samples. The point estimates of our proposed methods are easily computed using standard Cox regression software.

Our simulation results suggest that the proposed methods gain efficiency relative to existing methods (Lu & Shih, 2006) when sampling a smaller number of subjects, or having longer censoring times. This is due to the inclusion of a larger number of failures in the risk sets which are outside the subcohort.

If subcohort sizes are approximately equal, it appears that Design A results in more efficient estimators than Design C, and that Design C has greater efficiency than Design B. This can be attributed to differences in the number of sampled clusters in the subcohort. This trend is stronger when the covariate is cluster-specific. However, the choice between Designs A – C also depends on the cluster size and the availability of data on all clusters.

For each of Designs A – C, we propose three estimation methods which differ based on their treatment of p_0 , the marginal probability of the observed event. When β_0 is away from zero, the general superiority of $\hat{\beta}_t$ over $\hat{\beta}_w$, and of $\hat{\beta}_w$ over $\hat{\beta}_s$, may be explained by the more accurate estimation of the marginal event probability, p_0 . Such superiority is more pronounced when β_0 is further from the null (data not shown), as in Chen & Lo (1999). If $\beta_0 = 0$, $\hat{\beta}_t$ and $\hat{\beta}_w$ should gain no efficiency over $\hat{\beta}_s$, since no information about β_0 is provided by p_0 . In most real-data applications, the true case percentage p_0 is unknown, and it is not feasible to use $\hat{\beta}_t$. However, in cases where the study cohort is well-defined, p_w can be computed and used to obtain $\hat{\beta}_w$, which has the most practical value. In other cases, p_0 can be estimated using the subcohort.

For set-ups with a smaller number of clusters and smaller subcohort size, the proposed methods generally work well, though there is some slight under-coverage for Designs B and C. The asymptotic properties are based on increasing the number of clusters, but Design B samples the smallest number of clusters. Correspondingly, this under-coverage is reduced as the number of clusters increases.

Studies with low event rates often motivate case-cohort sampling. As such, we carried out simulations where the marginal event rate was around $p_0 = 0.03$ (Web Table 4). With a reasonable subcohort sample size, $\hat{\beta}_s$ appears to work as well as other estimators. In the presence of a very low failure rate, the proposed methods do not gain much efficiency over those of Lu & Shih (2006). This would be expected since, in such settings, the subcohort would tend to contain fewer events; meaning that little efficiency gain would be expected by including future failures in the risk sets. Note that the case-cohort design may still be beneficial for studies with a frequently occurring event. For example, one may need to retrospectively collect additional information from a large database (e.g., disease registry). Case-cohort sampling could then result in substantial cost savings, especially when the collection of detailed covariate information is expensive. The design might also be altered to sample only a fraction of the cases.

The proposed stratified methods appear to perform well with a reasonable number of strata. The baseline cumulative hazard estimator was also examined and performs well.

Point estimates based on simple random samples (SRS) for some non-rare event settings are provided (Web Table 7). It appears that the ESDs of the point estimates based on SRS are very close to those based on Bernoulli sampling. Therefore, one would not gain much efficiency by using SRS, at least for the examples we considered.

Based on our analysis in Section 5, Canadian hemodialysis (HD) patients appear to be at increased risk of cardiovascular disease death on Monday and, in particular Tuesday. Peritoneal dialysis (PD) is an alternative to hemodialysis as a treatment method for kidney failure. A useful follow-up to our analysis would be to study the day-of-week effect on death among PD patients. Unlike HD patients who receive dialysis only 3 days per week, PD patients can get treatments daily at home, at work, or on trips. Therefore, we would expect that the risk of death would be constant from day to day within the week. For HD patients, days on which mortality is increased may depend on schedule (M/W/F or T/T/S), but the dialysis schedule information is not available in the CORR database.

We propose sampling designs which construct the subcohort by independent Bernoulli sampling; in contrast, Lu and Shih (2006) construct the subcohort through sampling without replacement. The subcohort from simple random sampling can only be constructed when accrual into the cohort has ended, while the subcohort from Bernoulli sampling can be formed concurrently. Therefore, case-cohort designs with Bernoulli sampling may be particularly appealing in a prospective study. However, with fixed sample size, case-cohort designs using simple random sampling can improve efficiency, although asymptotic derivations would be more delicate than those in this paper particularly because of the dependence between sampled clusters induced by Designs B and C.

The proposed methods are based on a marginal proportional hazards model, which does not formulate the within-cluster dependence structure. A proportional hazards frailty model specifies the dependence structure explicitly. Such a model, combined with maximum likelihood estimation, may result in increased efficiency and would be worth investigating.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported, in part, by National Institutes of Health grant R01 DK-70869. The authors thank the Canadian Organ Replacement Register (CORR) of the Canadian Institute for Health Information for the end stage renal disease data. They also thank Dr. Rajiv Saran for the suggestion to examine the day-of-week effect and for valuable comments on the data analysis. The authors are grateful to the Editor, Associate Editor and Referees for their several constructive comments.

References

- Andersen PK, Gill RD. Cox's regression model for counting processes: A large-sample study. *Annals of Statistics*. 1982; 10:1100–1120.
- Barlow WE. Robust variance estimation for the case-cohort design. *Biometrics*. 1994; 50:1064–1072. [PubMed: 7786988]
- Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis*. 2000; 6:39–58. [PubMed: 10763560]
- Cai JW, Prentice RL. Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*. 1995; 82:151–164.
- Chambers JM, Mallows CL, Stuck BW. A method for simulating stable random variables. *Journal of the American Statistical Association*. 1976; 71:340–344.
- Chen HY. Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. *Journal of the American Statistical Association*. 2001; 96:1446–1457.
- Chen HY. Fitting semiparametric transformation regression models to data from a modified case-cohort design. *Biometrika*. 2001; 88:255–268.
- Chen K, Lo SH. Case-cohort and case-control analysis with Cox's model. *Biometrika*. 1999; 86:755–764.
- Chen K. Generalized case-cohort sampling. *Journal of the Royal Statistical Society, B*. 2001; 63:791–809.
- Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B*. 1972; 34:187–220.
- Foutz RV. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*. 1977; 72:147–148.
- Hougaard, P. *Analysis of multivariate survival data*. New York: Springer; 2000.
- Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*. 1988; 7:149–160. [PubMed: 3353602]
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. New York: Wiley; 2002.
- Kong L, Cai J, Sen PK. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*. 2004; 91:305–319.
- Kong L, Cai J, Sen PK. Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies. *Statistica Sinica*. 2006; 16:135–151.
- Kulich M, Lin DY. Additive hazards regression for case-cohort studies. *Biometrika*. 2000; 87:73–87.
- Langholz B, Jiao J. Computational methods for case-cohort studies. *Computational Statistics & Data Analysis*. 2007; 51:3737–3748.

- Lee, EW.; Wei, LJ.; Amato, DA. Cox-type regression for large numbers of small groups of correlated failure time observations. In: Klein, JP.; Goel, PK., editors. *Survival Analysis: State of the Art*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1992. p. 237-247.
- Liang KY, Self SG, Bandeen-Roche KJ, Zeger SL. Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis*. 1995; 1:403–415. [PubMed: 9385112]
- Liang KY, Self SG, Chang YC. Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society, B*. 1993; 55:441–453.
- Lin DY, Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*. 1993; 88:1341–1349.
- Lu SE, Shih JH. Case-Cohort designs and analysis for clustered failure time data. *Biometrics*. 2006; 62:1138–1148. [PubMed: 17156289]
- Lu SE, Wang MC. Marginal analysis for clustered failure time data. *Lifetime Data Analysis*. 2005; 11:61–79. [PubMed: 15747590]
- Ma S. Additive risk model with case-cohort sampled current status data. *Statistical Papers*. 2007; 48:595–608.
- Moger TA, Pawitan Y, Borgan Ø. Case-cohort methods for survival data on families from routine registers. *Statistics in Medicine*. 2008; 27:1062–1074. [PubMed: 17639506]
- Nan B, Yu M, Kalbfleisch JD. Censored linear regression for case-cohort studies. *Biometrika*. 2006; 93:747–762.
- Nan B, Kalbfleisch JD, Yu M. Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Annals of Statistics*. 2009
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Samuelson SO, Anestad H, Skrandal A. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*. 2007; 34:103–119.
- Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics*. 1988; 16:64–81.
- Sorensen P, Anderson PK. Competing risks analysis of the case-cohort design. *Biometrika*. 2000; 87:49–59.
- Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*. 1998; 93:1164–1175.
- Sun J, Sun L, Flournoy N. Additive hazards model for competing risks analysis of the case-cohort design. *Communications in statistics: Theory and methods*. 2004; 33:351–366.
- Therneau TM, Li H. Computing the Cox model for case cohort designs. *Lifetime Data Analysis*. 1999; 5:99–112. [PubMed: 10408179]
- van der Vaart, A.; Wellner, JA. *Weak convergence and empirical processes*. New York: Springer; 1996.
- Wacholder S, Gail MH, Pee D, Brookmeyer R. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*. 1989; 76:117–123.
- 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.

Figure 1.
Simulation results to examine the cumulative baseline hazard estimators based on 1000 replications.

Figure 2.
Cumulative baseline hazard estimators for the study of CVD mortality among dialysis patients.

Table 1

Simulation results based on 1000 replications: $\beta_0 = \log(0.5)$.

Design & Method	$n = 100, n_s = 800$						$n = 100, n_s = 600$					
	Bias	ESD	ASE	ARE	CP		Bias	ESD	ASE	ARE	CP	
$\beta_0 = \log(0.5)$												
FC	-0.001	0.053	0.054	1.000	0.959		-0.001	0.053	0.054	1.000	0.959	
A SC	-0.003	0.083	0.082	0.434	0.954		-0.002	0.092	0.091	0.352	0.956	
WC	-0.003	0.082	0.082	0.434	0.953		-0.002	0.091	0.091	0.352	0.958	
T	-0.004	0.080	0.079	0.467	0.951		-0.003	0.089	0.088	0.377	0.950	
LS	-0.003	0.091	0.089	0.368	0.933		-0.003	0.103	0.100	0.292	0.941	
B SC	-0.001	0.083	0.084	0.413	0.946		-0.001	0.094	0.093	0.337	0.943	
WC	-0.003	0.086	0.086	0.394	0.940		-0.004	0.096	0.095	0.323	0.941	
T	-0.004	0.084	0.083	0.423	0.933		-0.004	0.094	0.093	0.337	0.929	
LS	-0.002	0.091	0.091	0.352	0.954		-0.004	0.104	0.102	0.280	0.943	
C SC	0.000	0.084	0.083	0.423	0.955		-0.001	0.093	0.092	0.345	0.941	
WC	-0.001	0.083	0.083	0.423	0.945		-0.002	0.093	0.092	0.345	0.942	
T	-0.002	0.082	0.080	0.456	0.942		-0.003	0.092	0.090	0.360	0.942	
LS	-0.002	0.090	0.090	0.360	0.944		-0.004	0.102	0.101	0.286	0.945	
$\beta_0 = 0$												
FC	0.000	0.040	0.040	1.000	0.942		0.000	0.040	0.040	1.000	0.942	
A SC	0.002	0.035	0.036	0.298	0.954		0.001	0.082	0.082	0.238	0.952	
WC	0.002	0.035	0.036	0.297	0.955		0.001	0.082	0.082	0.238	0.943	
T	0.002	0.035	0.036	0.297	0.952		0.001	0.083	0.082	0.238	0.943	
LS	0.006	0.038	0.041	0.225	0.967		0.003	0.095	0.095	0.177	0.951	
B SC	0.005	0.037	0.036	0.302	0.936		0.001	0.085	0.080	0.250	0.929	
WC	0.005	0.036	0.036	0.292	0.939		0.001	0.087	0.079	0.256	0.915	
T	0.005	0.036	0.036	0.293	0.939		0.001	0.087	0.079	0.256	0.913	
LS	0.002	0.041	0.041	0.231	0.946		0.002	0.099	0.093	0.185	0.922	
C SC	0.008	0.035	0.036	0.297	0.950		0.002	0.086	0.081	0.244	0.928	
WC	0.008	0.035	0.036	0.298	0.947		0.002	0.086	0.081	0.244	0.927	
T	0.008	0.035	0.036	0.298	0.947		0.002	0.086	0.081	0.244	0.922	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Design & Method	$n = 100, n_s = 800$					$n = 100, n_s = 600$				
	Bias	ESD	ASE	ARE	CP	Bias	ESD	ASE	ARE	CP
LS	0.012	0.040	0.041	0.225	0.939	0.003	0.097	0.095	0.177	0.937

Estimate of β_0 from 5 methods: Method FC = full cohort analysis; SC = estimating p_0 using the subcohort, p_S ; WC = estimating p_0 using whole cohort, p_W ; T = using true value, p_0 ; LS = Lu and Shih (2006) estimator.
The number of clusters $n = 100$, m_i follows a $\text{Bin}(50, 0.8)$ distribution, $\alpha = 0.8$, $\lambda_0 = 1$, censoring time $C = 1$, Z follows a $\text{Bernoulli}(0.5)$ distribution. The number of individuals in the subcohort is either $n_s = 800$ or $n_s = 600$.

Table 2

Simulation results with $p_0 = 0.14$ and $p_0 = 0.21$ based on 1000 replications.

Design & Method	$Z \sim \text{Bernoulli}(0.5)$						$Z \sim N(0, 1)$					
	Bias	ESD	ASE	ARE	CP		Bias	ESD	ASE	ARE	CP	
FC	-0.010	0.111	0.106	1.000	0.934		-0.004	0.055	0.054	1.000	0.932	
A												
SC	-0.011	0.127	0.123	0.743	0.937		-0.006	0.068	0.068	0.631	0.944	
WC	-0.011	0.127	0.123	0.743	0.935		-0.006	0.068	0.067	0.650	0.940	
T	-0.011	0.125	0.121	0.767	0.935		-0.005	0.063	0.063	0.735	0.944	
LS	-0.011	0.127	0.124	0.731	0.940		-0.008	0.074	0.072	0.563	0.937	
B												
SC	-0.010	0.129	0.123	0.743	0.928		-0.006	0.069	0.069	0.612	0.943	
WC	-0.011	0.130	0.124	0.731	0.922		-0.007	0.069	0.069	0.612	0.938	
T	-0.011	0.128	0.122	0.755	0.927		-0.006	0.065	0.065	0.690	0.941	
LS	-0.011	0.129	0.124	0.731	0.930		-0.007	0.073	0.074	0.533	0.948	
C												
SC	-0.011	0.129	0.123	0.743	0.929		-0.004	0.071	0.068	0.631	0.936	
WC	-0.011	0.129	0.124	0.731	0.932		-0.003	0.070	0.068	0.631	0.930	
T	-0.011	0.127	0.122	0.755	0.929		-0.002	0.065	0.064	0.712	0.932	
LS	-0.011	0.130	0.124	0.731	0.937		-0.004	0.077	0.072	0.563	0.930	

Estimate of β_0 from 5 methods: Method FC = full cohort analysis; SC = estimating p_0 using the subcohort, p_S ; WC = estimating p_0 using whole cohort, p_W ; T = using true value, p_0 ; LS = Lu and Shih (2006) estimator.

The number of clusters $n = 100$, m_i follows a $\text{Bin}(50, 0.8)$ distribution, $\alpha = 0.8$, $\lambda_0 = 0.2$, censoring time $C = 1$, $\beta = \log(0.5)$, Z follows either a $\text{Bernoulli}(0.5)$ distribution or a $N(0, 1)$ distribution, which corresponds to a marginal event rate of $p_0 = 0.14$ or $p_0 = 0.21$, respectively. The number of individuals in the subcohort is $n_S = 800$.

Table 3

Estimate of day-of-week effect on CVD mortality among dialysis patients.

Day	Design A					
	SC			WC		
	$\hat{\beta}$	SE	$\exp(\hat{\beta})$	$\hat{\beta}$	SE	$\exp(\hat{\beta})$
Sunday	0.00	0.00	1.00	0.00	0.00	1.00
Monday	0.31	0.27	1.36	0.33	0.26	1.39
Tuesday	0.52	0.28	1.68	0.51	0.28	1.67
Wednesday	-0.02	0.26	0.98	-0.004	0.25	1.00
Thursday	0.23	0.29	1.26	0.24	0.29	1.27
Friday	0.12	0.27	1.13	0.14	0.26	1.15
Saturday	-0.11	0.30	0.90	-0.09	0.29	0.91

Day	Full Cohort			LS		
	$\hat{\beta}$	SE	$\exp(\hat{\beta})$	$\hat{\beta}$	SE	$\exp(\hat{\beta})$
Sunday	0.00	0.00	1.00	0.00	0.00	1.00
Monday	0.39	0.21	1.48	0.33	0.38	1.39
Tuesday	0.55	0.24	1.73	0.79	0.37	2.20
Wednesday	-0.08	0.22	0.92	0.02	0.35	1.02
Thursday	0.27	0.25	1.31	0.23	0.40	1.26
Friday	0.14	0.22	1.15	0.16	0.33	1.17
Saturday	-0.02	0.25	0.98	-0.05	0.37	0.95