Published in final edited form as:

Biometrics. 2017 September ; 73(3): 1042–1052. doi:10.1111/biom.12657.

Improving Efficiency of Parameter Estimation in Case-Cohort Studies with Multivariate Failure Time Data

Ying Yan [Assistant Professor],

Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada, T2N 1N4
(ying.yan@ucalgary.ca).

Haibo Zhou [Professor], and

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A. (zhou@bios.unc.edu).

Jianwen Cai [Cary C. Boshamer Distinguished Professor and Interim Chair]

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A. (cai@bios.unc.edu).

Abstract

The case-cohort study design is an effective way to reduce cost of assembling and measuring expensive covariates in large cohort studies. Recently, several weighted estimators were proposed for the case-cohort design when multiple diseases are of interest. However, these existing weighted estimators do not make use of the covariate information available in the whole cohort.

Furthermore, the auxiliary information for the expensive covariates, which may be available in the studies, can not be incorporated directly. In this paper, we propose a class of updated-estimators. We show that, by making effective use of the whole cohort information, the proposed updated-estimators are guaranteed to be more efficient than the existing weighted estimators asymptotically. Furthermore, they are flexible to incorporate the auxiliary information whenever available. The advantages of the proposed updated-estimators are demonstrated in simulation studies and a real data analysis.

Keywords

Auxiliary Information; Correlated Data; Proportional Hazards Model; Marginal Methods; Updated-Estimators; Weighted Methods

1 Introduction

In large cohort studies, the cost of assembling and measuring expensive covariates for the whole cohort could be enormous. To reduce the cost of large cohort studies, Prentice (1986) proposed the case-cohort study design. A random sample, named a subcohort, from the

7 Supplementary Materials

Web Appendices referenced in Sections 3 and 5 are available with this paper at the Biometrics website on Wiley Online Library. The computer code Rcode_YZC2017_Biometrics.r for the simulation studies is also available at the Biometrics website.

original cohort is selected under this design. The expensive covariates are only assembled and measured for the subjects from the subcohort or who experienced failures.

The case-cohort design is widely used since its proposal. A large number of inference methods have been developed to analyze the case-cohort data under the Cox proportional hazards models (Cox 1972) with univariate failure time data. Prentice (1986) and Self and Prentice (1988) developed pseudolikelihood approaches. Barlow (1994) considered robust variance estimation. Chen and Lo (1999) and Chen (2001) proposed several estimating equation methods. Nan (2004) studied efficient estimation with discrete covariates. Scheike and Martinussen (2004) and Zeng and Lin (2014) proposed maximum likelihood estimators. Marti and Chavance (2011) and Keogh and White (2013) developed multiple imputation methods. Borgan et al. (2000), Kulich and Lin (2004), and Samuelsen, Anestad, and Skrandal (2007) studied stratified case-cohort designs. Breslow and Wellner (2007) and Breslow et al. (2009a, 2009b) proposed weighted likelihood estimators with two-phase stratified samples.

In many case-cohort studies, multivariate failure time data (or multiple disease outcomes) are available. For example, in the Busselton Health Study (Cullen 1972; Knuiman et al. 2003), two case-cohort studies were conducted, one for coronary heart disease, and the other for stroke. It was of interest to investigate the effect of serum ferritin on the risk of coronary heart disease and stroke, as well as compare the effect size. In contrast to the literature available for the case-cohort design with univariate failure time data, methods for multivariate failure time data from case-cohort studies are scarce. Kang and Cai (2009) proposed weighted estimators under marginal proportional hazards models. To improve efficiency, Kim, Cai, and Lu (2013) developed different weighted estimators by utilizing information from multiple disease outcomes. However, as we illustrate in Section 2.2, these existing weighted estimators do not fully incorporate information available in the whole cohort. Furthermore, they do not allow for the use of available auxiliary information for the main exposure.

Auxiliary information is often available in many studies. For example, the histologic type of tumor is one important risk factor for Wilms tumour patients in the study conducted by the National Wilms Tumor Study Group (D'Angio et al. 1989). The histologic type was measured cheaply and inaccurately for all the patients by local pathologists. Precise measurements of the histologic type were made by a central pathology centre for only a proportion of the patients. In this example, the auxiliary information consists of the inaccurate measurements of the histologic type obtained by local pathologists.

Chen (2002) proposed an “updated-method” for using the auxiliary covariate information in random samples under the proportional hazards models with univariate disease outcome. Jiang and Zhou (2007) and Lu (2013) extended this method to the additive hazards model and the accelerated failure time model, respectively. They assumed that the covariates are assembled and measured accurately in a random validation subsample, but some components of the covariates are not available outside this subsample. They showed that the updated-methods make effective use of the whole cohort information. Furthermore, the updated-methods are capable of incorporating the available auxiliary information. However,

their methods cannot be directly extended to the case-cohort design. Furthermore, they did not consider the scenario with multivariate failure time.

In this paper, we propose a class of updated-estimators with multivariate failure time data from case-cohort studies. Our methods can be regarded as a generalization of Chen (2002) from the validation sample design to the case-cohort design, and from the univariate setup to the multivariate setup. Our methods are novel in several aspects. First, developing procedures to effectively use the whole cohort information has been a research focus for the case-cohort studies in biostatistics and epidemiology (Breslow et al. 2009a, 2009b). This paper is the first attempt to develop the updated-estimators for the case-cohort studies, which greatly improve efficiency of the existing weighted estimators. Second, this paper is the first to develop updated-estimators for multivariate failure time data. Third, the generalization to the case-cohort design with multivariate disease outcomes entails new technical challenges because of the structure of the case-cohort design.

Our proposed updated-estimators in this paper make effective use of the whole cohort information available in the case-cohort studies. In particular, we show that they are guaranteed to be more efficient than the existing weighted estimators by Kang and Cai (2009) and Kim, Cai, and Lu (2013) asymptotically. Furthermore, they are flexible to incorporate the available auxiliary information. These updated-estimators are novel even for the case-cohort design with univariate failure time data. The advantages of the proposed updated-estimators are demonstrated in simulation studies and a real data analysis.

2 Notation and Model Setup

2.1 Cox Model

We assume that there is a cohort of N independent subjects and there are K disease outcomes of interest. Let T_{ik} be the failure time, C_{ik} be the censoring time, and $Z_{ik}(t)$ be a vector of covariates for disease k of subject i , $i = 1, \dots, N$, $k = 1, \dots, K$. Let

$Z_{ik}(t) = (X_{ik}^T(t), V_{ik}^T(t))^T$, where $X_{ik}(t)$ is a sub-vector of covariates that are expensive to assemble and measure, and $V_{ik}(t)$ is a sub-vector of covariates available for all the subjects in the cohort. We assume that $Z_{ik}(t)$ are external covariates (Kalbfleisch and Prentice 2002). In some real data sets, a vector of auxiliary variables $W_{ik}(t)$ is available for all the subjects in the cohort. We assume $W_{ik}(t)$ and (T_{ik}, C_{ik}) are independent conditional on $Z_{ik}(t)$. Let $Z_{ik}^*(t) = (W_{ik}^T(t), V_{ik}^T(t))^T$ if $W_{ik}(t)$ is available, and $Z_{ik}^*(t) = V_{ik}(t)$ if there is no auxiliary information.

We consider that the marginal hazard function of T_{ik} is related to $Z_{ik}(t)$ through the Cox-type proportional hazards model (Cox 1972)

$$\lambda_{ik}(t; Z_{ik}(t)) = \lambda_{0k}(t) \exp \left\{ \beta_x^T X_{ik}(t) + \beta_v^T V_{ik}(t) \right\}, \quad (1)$$

where $\lambda_{0k}(\cdot)$ is an unspecified baseline hazard function for disease k , and $\beta = (\beta_x^T, \beta_v^T)^T$ is a vector of unknown regression parameters. Model (1) can incorporate disease-specific

covariate effects. Specifically, let $\beta = (\beta_1^T, \dots, \beta_K^T)^T$ and

$Z_{ik}(t) = (0_{i1}^T, \dots, 0_{i(k-1)}^T, \tilde{Z}_{ik}^T(t), 0_{i(k+1)}^T, \dots, 0_{iK}^T)^T$, where 0_{ij} is a zero vector with the same length as that of $\tilde{Z}_{ik}(t)$, $j = 1, \dots, K$. Then we obtain $\beta^T Z_{ik}(t) = \beta_k^T \tilde{Z}_{ik}(t)$.

Let β_0 be the unknown true value of β . T_{ik} and C_{ik} are assumed to be conditionally independent given $Z_{ik}(t)$. Suppose individuals are observed over a common time interval $[0, \tau]$, where τ is a positive constant. Denote $S_{ik} = \min(T_{ik}, C_{ik})$, $\delta_{ik} = \mathbb{I}(T_{ik} = C_{ik})$, $N_{ik}(t) = \mathbb{I}(S_{ik} \geq t, \delta_{ik} = 1)$, and $Y_{ik}(t) = \mathbb{I}(S_{ik} \geq t)$.

If all the covariates and survival information are measured for each subject in the cohort, estimation of the regression parameter β can be based on solving the estimating equation $U(\beta) = 0$ (Wei, Lin, and Weissfeld 1989; Cai and Prentice 1995), where

$$U(\beta) = \sum_{i=1}^N \sum_{k=1}^K \int_0^\tau \left\{ Z_{ik}(t) - \frac{S_k^{(1)}(\beta, t)}{S_k^{(0)}(\beta, t)} \right\} dN_{ik}(t), \quad (2)$$

and $S_k^{(d)}(\beta, t) = N^{-1} \sum_{i=1}^N Y_{ik}(t) Z_{ik}^{\otimes d}(t) \exp\{\beta^T Z_{ik}(t)\}$, $d = 0, 1, 2$. Let $\bar{\beta}$ be the corresponding estimator.

2.2 The Case-Cohort Design

We consider the case-cohort design by Prentice (1986). The first phase information

$$\mathcal{V} \equiv \cup_{i=1}^N \mathcal{V}_i \equiv \cup_{i=1}^N \{Z_{ik}^*(t), S_{ik}, \delta_{ik}, 0 \leq t \leq S_{ik}, k=1, \dots, K\}$$

is available for all the subjects in the cohort. In the second phase, we use simple random sampling to select a random subcohort from the original cohort. Let ξ_i be an indicator function, such that $\xi_i = 1$ if the subject i is selected into the subcohort, and $\xi_i = 0$ otherwise. Let $\tilde{\alpha} = Pr(\xi_i = 1)$ denote the subcohort selection probability, and α be the limit of $\tilde{\alpha}$. We obtain the measurements of $X_{ik}(t)$ for the subjects in the subcohort, as well as the subjects who are outside the subcohort and experienced disease k , $k = 1, \dots, K$. Under the case-cohort design, the estimating function (2) and the estimator $\bar{\beta}$ are not available.

In the following, we briefly review the existing weighted estimators developed under the case-cohort design with multiple disease outcomes. Let $U(\beta)$ be a weighted pseudoscore function constructed by observed data with the following expression:

$$U_I(\beta) = \sum_{i=1}^N \sum_{k=1}^K \int_0^\tau \left\{ Z_{ik}(t) - \frac{S_{I,k}^{(1)}(\beta, t)}{S_{I,k}^{(0)}(\beta, t)} \right\} dN_{ik}(t), \quad (3)$$

where $S_{I,k}^{(d)}(\beta, t) = N^{-1} \sum_{i=1}^N w_{I,ik}(t) Y_{ik}(t) Z_{ik}^{\otimes d}(t) \exp\{\beta^T Z_{ik}(t)\}$, $d = 0, 1, 2$. Here,

$$\begin{aligned} w_{I,ik}(t) &= \delta_{ik} + (1 - \delta_{ik}) \xi_i \hat{\alpha}_{I,k}^{-1}(t), \\ \hat{\alpha}_{I,k}(t) &= \sum_{i=1}^N \xi_i (1 - \delta_{ik}) A_{ik}(t) / \sum_{i=1}^N (1 - \delta_{ik}) A_{ik}(t), \end{aligned}$$

and $A_{ik}(t)$ is a function of the observed first phase information $\mathcal{V}$. Let $\hat{\beta}_I$ be the solution of the equation $U_I(\beta) = 0$. The weighted estimator $\hat{\beta}_{KC}$ of Kang and Cai (2009) is a special case of $\hat{\beta}_I$ with $A_{ik}(t) = Y_{ik}(t)$. Kang and Cai (2009) showed that $\hat{\beta}_{KC}$ is consistent for β_0 and is asymptotically normal under certain regularity conditions.

Alternatively, consider the following weighted pseudoscore function:

$$U_{II}(\beta) = \sum_{i=1}^N \sum_{k=1}^K \int_0^\tau \left\{ Z_{ik}(t) - \frac{S_{II,k}^{(1)}(\beta, t)}{S_{II,k}^{(0)}(\beta, t)} \right\} dN_{ik}(t), \quad (4)$$

where $S_{II,k}^{(d)}(\beta, t) = N^{-1} \sum_{i=1}^N w_{II,ik}(t) Y_{ik}(t) Z_{ik}^{\otimes d}(t) \exp\{\beta^T Z_{ik}(t)\}$, $d = 0, 1, 2$. Here,

$$\begin{aligned} w_{II,ik}(t) &= \left\{ 1 - \prod_{j=1}^K (1 - \delta_{ij}) \right\} + \left\{ \prod_{j=1}^K (1 - \delta_{ij}) \right\} \xi_i \hat{\alpha}_{II,k}^{-1}(t), \\ \hat{\alpha}_{II,k}(t) &= \sum_{i=1}^N \xi_i \left\{ \prod_{j=1}^K (1 - \delta_{ij}) \right\} A_{ik}(t) / \sum_{i=1}^N \left\{ \prod_{j=1}^K (1 - \delta_{ij}) \right\} A_{ik}(t), \end{aligned}$$

and $A_{ik}(t)$ is a function of the observed first phase information. Let $\hat{\beta}_{II}$ be the solution of the equation $U_{II}(\beta) = 0$. The weighted estimator $\hat{\beta}_{KCL}$ of Kim, Cai, and Lu (2013) is a special case of $\hat{\beta}_{II}$ with $A_{ik}(t) = Y_{ik}(t)$. The weight $w_{II,ik}(t)$ involves the status of the other diseases, and thus the corresponding estimator $\hat{\beta}_{II}$ incorporates some information for subjects with other diseases. Kim, Cai, and Lu (2013) showed that $\hat{\beta}_{KCL}$ is consistent for β_0 and is asymptotically normal under certain regularity conditions. Furthermore, they showed in their numerical studies that $\hat{\beta}_{KCL}$ is more efficient than $\hat{\beta}_{KC}$.

Both of the weighted estimators $\hat{\beta}_I$ and $\hat{\beta}_{II}$ (including $\hat{\beta}_{KC}$ and $\hat{\beta}_{KCL}$) do not efficiently utilize the first phase information $\mathcal{V}$. To see this, note that the estimating equations $U_I(\beta)$ and $U_{II}(\beta)$ do not incorporate the covariates $V_{ik}(t)$ explicitly for censored subjects outside

the subcohort. Furthermore, $U_I(\beta)$ and $U_{II}(\beta)$ do not use the auxiliary information $W_{ik}(t)$ that may be available in the cohort. Our aim in this paper is to improve efficiency of the weighted estimators $\hat{\beta}_I$ and $\hat{\beta}_{II}$ by utilizing the available first-phase information $\mathcal{V}$.

In the following section, we propose a class of updated-estimators that effectively incorporate the first-phase information $\mathcal{V}$ from the whole cohort, and are guaranteed to be more efficient than the weighted estimators $\hat{\beta}_I$ and $\hat{\beta}_{II}$. To unify the notations, we denote $\{w_{I,ik}(t); i = 1, \dots, N, k = 1, \dots, K\}$ and $\{w_{II,ik}(t); i = 1, \dots, N, k = 1, \dots, K\}$ as type-I weights and type-II weights, respectively, and define the following general weighted estimating function:

$$U_W(\beta) = \sum_{i=1}^N \sum_{k=1}^K \int_0^\tau \left\{ Z_{ik}(t) - \frac{S_{W,k}^{(1)}(\beta, t)}{S_{W,k}^{(0)}(\beta, t)} \right\} dN_{ik}(t), \quad (5)$$

where $S_{W,k}^{(d)}(\beta, t) = N^{-1} \sum_{i=1}^N w_{W,ik}(t) Y_{ik}(t) Z_{ik}^{\otimes d}(t) \exp\{\beta^T Z_{ik}(t)\}$, $d = 0, 1, 2$. Here, $U_W(\beta)$ with type-I weights reduces to the weighted estimating function $U_I(\beta)$ in (3); similarly, $U_W(\beta)$ with type-II weights becomes the weighted estimating function $U_{II}(\beta)$ in (4). Let $\hat{\beta}_W$ be the solution of $U_W(\beta) = 0$. The weighted estimator $\hat{\beta}_W$ is identical to $\hat{\beta}_I$ or $\hat{\beta}_{II}$ if we set W to be I or II , respectively.

3 A Class of Updated-Estimators

Let $\bar{\gamma}$ be the solution of the estimating equation $U^*(\gamma) = 0$, where

$$U^*(\gamma) = \sum_{i=1}^N \sum_{k=1}^K \int_0^\tau \left\{ Z_{ik}^*(t) - \frac{S_k^{*(1)}(\gamma, t)}{S_k^{*(0)}(\gamma, t)} \right\} dN_{ik}(t), \quad (6)$$

and $S_k^{*(d)}(\gamma, t) = N^{-1} \sum_{i=1}^N Y_{ik}(t) Z_{ik}^{*\otimes d}(t) \exp\{\gamma^T Z_{ik}^*(t)\}$, $d = 0, 1, 2$. Equation (6) can be regarded as a modified version of equation (2) with $Z_{ik}(t)$ replaced by $Z_{ik}^*(t)$ and β replaced by γ . By Lin and Wei (1989), as $N \rightarrow \infty$, $\bar{\gamma}$ converges in probability to γ_0 , where γ_0 is the unique solution of

$$\sum_{k=1}^K \int_0^\tau \left\{ s_k^{*(1)}(t) - e_k^*(\gamma, t) s_k^{*(0)}(t) \right\} dt = 0.$$

Here, $s_k^{*(d)}(t) = E \left\{ Z_{1k}^{*\otimes d}(t) Y_{1k}(t) \lambda_{1k}(t) \right\}$, $\lambda_{ik}(t)$ is the true hazard function for disease k of subject i , $e_k^*(\gamma, t) = s_k^{*(1)}(\gamma, t) / s_k^{*(0)}(\gamma, t)$, and $s_k^{*(d)}(\gamma, t) = E \left\{ s_k^{*(d)}(\gamma, t) \right\}$, $d = 0, 1, 2$. Here, the expectation is taken under the true distribution of $\{Z_{ik}^*(t), V_{ik}(t), S_{ik}, \delta_{ik}\}$.

Let $\hat{\gamma}_W$ be the solution of the weighted equation $U_W^*(\gamma) = 0$, where

$$U_W^*(\gamma) = \sum_{i=1}^N \sum_{k=1}^K \int_0^\tau \left\{ Z_i^*(t) - \frac{S_{W,k}^{*(1)}(\gamma, t)}{S_{W,k}^{*(0)}(\gamma, t)} \right\} dN_{ik}(t), \quad (7)$$

and $S_{W,k}^{*(d)}(\gamma, t) = N^{-1} \sum_{i=1}^N w_{W,ik}(t) Y_{ik}(t) Z_{ik}^{*\otimes d}(t) \exp\{\gamma^T Z_{ik}^*(t)\}$, $d = 0, 1, 2$. Note that the equation (7) is a modified version of the weighted pseudoscore equation (5) with $Z_{ik}(t)$ and β replaced by $Z_{ik}^*(t)$ and γ , respectively.

For $k = 1, \dots, K$, let $s_k^{(d)}(\beta, t) = E[Y_{1k}(t) Z_{1k}^{\otimes d}(t) \exp\{\beta^T Z_{1k}(t)\}]$, $d = 0, 1, 2$,
 $e_k(\beta, t) = s_k^{(1)}(\beta, t) / s_k^{(0)}(\beta, t)$,

$$\mathcal{J}(\beta) = \sum_{k=1}^K \int_0^\tau \left\{ \frac{s_k^{(2)}(\beta, t)}{s_k^{(0)}(\beta, t)} - e_k^{\otimes 2}(\beta, t) \right\} s_k^{(0)}(\beta, t) \lambda_{0k}(t) dt,$$

$$M_{ik}(\beta, t) = N_{ik}(t) - \int_0^t Y_{ik}(u) \exp\{\beta^T Z_{ik}(u)\} \lambda_{0k}(u) du,$$

$$\tilde{M}_{ik}(\beta, t) = \int_0^t \{Z_{ik}(u) - e_k(\beta, u)\} dM_{ik}(\beta, u), \text{ and } \mathcal{J}(\beta) = E \left[\left\{ \sum_{k=1}^K \tilde{M}_{1k}(\beta, \tau) \right\}^{\otimes 2} \right].$$

Furthermore, define $B_{W,ik}$ for $W = I$ and II such that

$$B_{I,ik} = 1 - \delta_{ik}, \\ \text{and } B_{II,ik} = \prod_{j=1}^K (1 - \delta_{ij}).$$

Define $\mu_{W,k}(t) = E\{B_{W,1k} A_{1k}(t)\}$, $\psi_{W,k}(\beta, t) = E\{B_{W,1k} R_{1k}(\beta, t)\}$, and

$$U_{W,ik}(\beta) = B_{W,ik} \int_0^\tau \left\{ R_{ik}(\beta, t) - \mu_{W,k}^{-1}(t) A_{ik}(t) \psi_{W,ik}(\beta, t) \right\} \lambda_{0k}(t) dt,$$

where $R_{ik}(\beta, t) = \{Z_{ik}(t) - e_k(\beta, t)\} Y_{ik}(t) \exp\{\beta^T Z_{ik}(t)\}$.

For $k = 1, \dots, K$, let

$$\mathcal{J}^*(\gamma) = \sum_{k=1}^K \int_0^\tau \left\{ \frac{s_k^{*(2)}(\gamma, t)}{s_k^{*(0)}(\gamma, t)} - e_k^{*\otimes 2}(\gamma, t) \right\} s_k^{*(0)}(t) dt,$$

$$M_{ik}^*(\gamma, t) = N_{ik}(t) - \int_0^t Y_{ik}(u) \exp\{\gamma^T Z_{ik}^*(u)\} s_k^{*(0)}(u) / s_k^{*(0)}(\gamma, u) du,$$

$$\tilde{M}_{ik}^*(\gamma, t) = \int_0^t \{Z_{ik}^*(u) - e_k^*(\gamma, u)\} dM_{ik}^*(\gamma, u), \text{ and } \mathcal{J}^*(\gamma) = E \left[\left\{ \sum_{k=1}^K \tilde{M}_{1k}^*(\gamma, \tau) \right\}^{\otimes 2} \right].$$

Furthermore, let $\psi_{W,k}^*(\gamma, t) = E \left\{ B_{W,1k} R_{1k}^*(\gamma, t) \right\}$, and

$$U_{W,ik}^*(\gamma) = B_{W,ik} \int_0^\tau \left\{ R_{ik}^*(\gamma, t) - \mu_{W,k}^{-1}(t) A_{ik}(t) \psi_{W,k}^*(\gamma, t) \right\} s_k^{*(0)}(t) / s_k^{*(0)}(\gamma, t) dt,$$

where $R_{ik}^*(\gamma, t) = \{Z_{ik}^*(t) - e_k^*(\gamma, t)\} Y_{ik}(t) \exp \left\{ \gamma^T Z_{ik}^*(t) \right\}$.

We summarize the asymptotic joint distribution of $\sqrt{N}(\hat{\beta}_W - \beta_0)$ and $\sqrt{N}(\hat{\gamma}_W - \bar{\gamma})$ in the following theorem, and its proof is deferred to the Supplementary Material:

Theorem 1

Under the Regularity Conditions given in the Supplementary Material, as $N \rightarrow \infty$,

$$\sqrt{N} \begin{pmatrix} \hat{\beta}_W - \beta_0 \\ \hat{\gamma}_W - \bar{\gamma} \end{pmatrix} \xrightarrow{d} MVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{W,11}(\beta_0) & \Sigma_{W,12}(\beta_0, \gamma_0) \\ \Sigma_{W,21}(\beta_0, \gamma_0) & \Sigma_{W,22}(\gamma_0) \end{pmatrix} \right),$$

where

$$\begin{aligned} \Sigma_{W,11}(\beta) &= \mathcal{J}^{-1}(\beta) \left[\mathcal{J}(\beta) + \frac{1-\alpha}{\alpha} E \left\{ \left(\sum_{k=1}^K U_{W,1k}(\beta) \right)^{\otimes 2} \right\} \right] \mathcal{J}^{-1}(\beta), \\ \Sigma_{W,12}(\beta, \gamma) &= \frac{1-\alpha}{\alpha} \mathcal{J}^{-1}(\beta) E \left\{ \left(\sum_{k=1}^K U_{W,1k}(\beta) \right) \left(\sum_{k=1}^K U_{W,1k}^{*T}(\gamma) \right) \right\} \mathcal{J}^{*-1}(\gamma), \\ \Sigma_{W,21}(\beta, \gamma) &= \Sigma_{W,12}^T(\beta, \gamma), \\ \text{and } \Sigma_{W,22}(\gamma) &= \frac{1-\alpha}{\alpha} \mathcal{J}^{*-1}(\gamma) E \left\{ \left(\sum_{k=1}^K U_{W,1k}^*(\gamma) \right)^{\otimes 2} \right\} \mathcal{J}^{*-1}(\gamma). \end{aligned}$$

To simplify the notations, we let $\Sigma_{W,11} = \Sigma_{W,11}(\beta_0)$. The matrices $\Sigma_{W,12}$, $\Sigma_{W,21}$, and $\Sigma_{W,22}$ are similarly defined.

To improve efficiency of $\hat{\beta}_W$, we consider the following updated-estimator:

$$\hat{\beta}_U^{\hat{\Omega}} = \hat{\beta}_W - \hat{\Omega}(\hat{\gamma}_W - \bar{\gamma}),$$

where $\hat{\Omega}$ is a matrix constructed by observed information and satisfies that $\hat{\Omega}$ converges in probability to a deterministic matrix Ω as N goes to infinity. In the following, we specify $\hat{\Omega}$ so that the corresponding updated-estimator $\hat{\beta}_U^{\hat{\Omega}}$ is more efficient than the weighted estimator $\hat{\beta}_W$ asymptotically. By Theorem 1 and the multivariate delta method, we obtain that $\sqrt{N}(\hat{\beta}_U - \beta_0)$ converges in distribution to a mean-zero normal random vector with the asymptotic covariance matrix

$$\Sigma_{W,11} - \Omega \Sigma_{W,21} - \Sigma_{W,12} \Omega^T + \Omega \Sigma_{W,22} \Omega^T.$$

To ensure that the asymptotic covariance matrix of $\hat{\beta}_U^{\hat{\Omega}}$ is smaller than that of $\hat{\beta}_W$ in the sense that the difference of the two covariance matrices is negative definite, we set $\hat{\Omega} = \hat{\Sigma}_{W,12} \hat{\Sigma}_{W,22}^{-1}$, where $\hat{\Sigma}_{W,12}$ and $\hat{\Sigma}_{W,22}$ are the empirical versions of $\Sigma_{W,12}$ and $\Sigma_{W,22}$, respectively. We then obtain the updated-estimator

$$\hat{\beta}_U \equiv \hat{\beta}_U^{\hat{\Sigma}_{W,12} \hat{\Sigma}_{W,22}^{-1}} = \hat{\beta}_W - \hat{\Sigma}_{W,12} \hat{\Sigma}_{W,22}^{-1} (\hat{\gamma}_W - \bar{\gamma}), \quad (8)$$

with the asymptotic covariance matrix $\Sigma_{W,11} - \Sigma_{W,12} \Sigma_{W,22}^{-1} \Sigma_{W,21}$. The above argument is summarized in the following theorem.

Theorem 2

Under the Regularity Conditions given in the Supplementary Material, as $N \rightarrow \infty$,

$$\sqrt{N} (\hat{\beta}_U - \beta_0) \xrightarrow{d} MVN \left(0, \Sigma_{W,11} - \Sigma_{W,12} \Sigma_{W,22}^{-1} \Sigma_{W,21} \right).$$

Note that a class of pre-specified functions $A_{ik}(t)$ formulates a class of updated-estimators $\hat{\beta}_U$. Theorem 2 implies that for any pre-specified $A_{ik}(t)$, the corresponding updated-estimator $\hat{\beta}_U$ is asymptotically more efficient than the weighted estimator $\hat{\beta}_W$.

We now specify the random function $A_{ik}(t)$. Under the choice $A_{ik}(t) = Y_{ik}(t)$, the weighted estimator $\hat{\beta}_W$ with type-I weights reduces to $\hat{\beta}_{KC}$ by Kang and Cai (2009), and $\hat{\beta}_W$ with type-II weights becomes $\hat{\beta}_{KCL}$ by Kim, Cai, and Lu (2013), respectively. By Theorem 2, the updated-estimators $\hat{\beta}_U$ with type-I weights and type-II weights are more efficient than $\hat{\beta}_{KC}$ and $\hat{\beta}_{KCL}$, respectively. We consider $A_{ik}(t) = Y_{ik}(t)$ in the following numerical studies.

4 Simulation Studies

We conduct simulation studies to evaluate the finite sample performance of the proposed estimators under the case-cohort design with multiple disease outcomes. In the simulation studies, we consider the weighted estimator $\hat{\beta}_{KC}$ by Kang and Cai (2009) (KC), the weighted estimator $\hat{\beta}_{KCL}$ by Kim, Cai, and Lu (2013) (KCL), the updated-estimator $\hat{\beta}_U$ in (8) with type-I weights that updates the KC estimator (UpdatedKC), and $\hat{\beta}_U$ with type-II weights that updates the KCL estimator (UpdatedKCL).

We run 1000 simulations for each parameter configuration. The full cohort consists of $N=1000$ subjects. We choose a subcohort with size $n_s=200$ randomly from the cohort.

We set the number of disease outcomes K to be 2. We consider two cases for the marginal Cox proportional hazards models for the disease outcomes. In the first case, we set

$$\lambda_{ik}(t; Z_i) = \lambda_{0k}(t) \exp(\beta_x X_i + \beta_v V_i), \quad k=1, 2, \quad (9)$$

so that we assume common regression parameters for different disease outcomes. In the second case, we consider

$$\lambda_{ik}(t; Z_i) = \lambda_{0k}(t) \exp(\beta_{xk} X_i + \beta_{vk} V_i), \quad k=1, 2, \quad (10)$$

so that we allow different regression parameters for different disease outcomes. In both scenarios, we set the baseline hazard functions as $\lambda_{01}(t) = 1$ and $\lambda_{02}(t) = 2$, and the parameters as $(\beta_{x1}, \beta_{v1}) = (\beta_{x2}, \beta_{v2}) = (\beta_x, \beta_v) = (0.2, 0.5)$.

We generate the bivariate time-independent covariates $Z_i = (X_i, V_i)^T$ from the bivariate normal distribution:

$$\begin{pmatrix} X_i \\ V_i \end{pmatrix} \sim BVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

where we fix the correlation parameter $\rho = 0.5$. Bivariate survival times are generated using the Clayton-Cuzick (1985) model with the bivariate survival function

$$S(t_1, t_2 | Z_i) = \left(\exp \left[\int_0^{t_1} \lambda_{i1}(t; Z_i) dt \right] / \theta + \exp \left[\int_0^{t_2} \lambda_{i2}(t; Z_i) dt \right] / \theta - 1 \right)^{-\theta},$$

where the parameter θ represents the degree of dependence between the bivariate disease outcomes T_1 and T_2 . We set $\theta = 2$ so that the Kendall's tau $\tau_\theta = 1/(2\theta + 1) = 0.2$. Thus, the degree of dependence is moderate. The censoring times $C_{i1} = C_{i2}$ are generated from the uniform distribution $UNIF(0, C)$, where we set $C = 0.1, 0.2$, and 0.4 , leading to the bivariate event proportions [6%, 11%], [11%, 20%], and [20%, 33%], respectively.

We consider two scenarios for the auxiliary variable W_i . Let p^* be the length of W_i . In the first scenario, we set $p^* = 0$, i.e., there is no auxiliary variable W_i . In the second scenario, $p^* = 1$, i.e., there is one auxiliary variable W_i . We generate W_i from the classical measurement error model (Carroll et al. 2006): $W_i = X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. Here we set $\sigma^2 = 0.2$ or 1 to represent the situations where W_i is a strong or weak auxiliary variable, respectively.

We report the finite sample bias (Bias), the sample standard deviation (SSD), the average of the estimated standard error (ESE), the coverage probability of the nominal 95% confidence interval (CP), and the sample relative efficiency (RE) using the KC estimator as a reference.

The RE of an estimator E relative to the KC estimator is defined as SSD_{KC}^2 / SSD_E^2 , where SSD_E is the sample standard deviation of the E estimator, and E represents KC, KCL, UpdatedKC, or UpdatedKCL. Tables 1 and 2 provide summaries of the simulation studies.

We obtain the following conclusions from the tables. All the estimators have negligible biases. The proposed updated-estimators UpdatedKC and UpdatedKCL greatly improve efficiency compared to the weighted estimators KC and KCL. UpdatedKC tends to be more efficient than KC when estimating β_v ; similarly, UpdatedKCL is more efficient than KCL for β_v estimation. This phenomenon could be partly explained by that the estimating equations of KC and KCL do not incorporate the covariates V_{ik} explicitly for censored subjects outside the subcohort, whereas the proposed UpdatedKC and UpdatedKCL estimation procedures make effective use of this information through the updating step. For β_x estimation when there is no auxiliary information, UpdatedKC and UpdatedKCL have similar performance to KC and KCL, respectively. When auxiliary information is available, UpdatedKC and UpdatedKCL are capable of incorporating the information to improve efficiency for both β_x and β_v estimation. Furthermore, the efficiency gain of UpdatedKC and UpdatedKCL tends to be more significant with higher event proportions. The coverage probabilities of the 95% confidence intervals are close to the nominal level, and the SSD and ESE agree with each other in the simulations.

We consider a third scenario for the auxiliary variables with setting the bivariate event proportions to be [11%, 20%]. There are five cheaply measured risk factors, represented as a 5×1 variable $\tilde{W}_j$, available for the entire cohort. The components of $\tilde{W}_j$ are generated from the following models:

$$\begin{aligned}\tilde{W}_{i,1} &= 1 + 0.5I(X_i > 0) + \epsilon_{i,1}, \\ \tilde{W}_{i,2} &= 1 + 0.4\sin(X_i) + \epsilon_{i,2}, \\ \tilde{W}_{i,3} &= 0.12\exp(X_i) + \epsilon_{i,3}, \\ \tilde{W}_{i,4} &= -0.5 + 0.5X_i + 0.3X_iV_i + \epsilon_{i,4}, \\ \text{and } \tilde{W}_{i,5} &= 0.25X_i + 0.5X_i^2 + \epsilon_{i,5},\end{aligned}$$

where $\tilde{W}_{i,j}$ is the j th component of the vector $\tilde{W}_i$, and the $\epsilon_{i,j}$'s are independent standard normal variables, $j = 1, 2, 3, 4, 5$. By the generation setting, the $\tilde{W}_{i,j}$'s have weak correlations with the main exposure X_i (about 0.20).

In principle, we may set $W_j = \tilde{W}_j$ to be the auxiliary variable for the main exposure X_j and proceed to obtain the proposed updated-estimators. However, the dimension of $\tilde{W}_j$ is considerably higher than that of X_j , which may cause the instability issue of the resulting updated-estimators. This is because the estimation procedure and the variance formula involve the inverse of high dimensional matrices when the dimension of $\tilde{W}_j$ is high. To remedy the instability issue and make effective use of $\tilde{W}_j$, we propose the following procedure to “create” a suitable W_j for the main exposure X_j , which has the same dimension as that of X_j . To be specific, we fit a linear regression model

$$X_i = \alpha_0 + \alpha_w^T \tilde{W}_i + \alpha_v V_i + \sum_{k=1}^2 \alpha_{\delta,k} \delta_{ik} + \sum_{k=1}^2 \alpha_{s,k} S_{ik} + e_i,$$

based on all the subjects with the main exposure X_i measured (i.e., subjects with $\xi_i = 1$ or $\prod_{j=1}^2 (1 - \delta_{ij}) = 0$) to create the auxiliary variable W_i :

$$W_i = \hat{\alpha}_0 + \hat{\alpha}_w^T \tilde{W}_i + \hat{\alpha}_v^T V_i + \sum_{k=1}^2 \hat{\alpha}_{\delta,k} \delta_{ik} + \sum_{k=1}^2 \hat{\alpha}_{s,k} S_{ik}.$$

The sample mean of the auxiliary variable W_i (nearly zero) is almost the same as that of X_i . The sample correlation between W_i and X_i is near 0.60, much higher than the correlations between $\tilde{W}_{ij}$ and X_i (0.20), suggesting that W_i might be an appropriate auxiliary variable for the main exposure X_i . We use the “created surrogates” W_i 's to construct our updated-estimators, and the results are summarized in Table 3. UpdatedKC and Updated-KCL using the “created surrogates” have improved performance for β_x estimation, compared to UpdatedKC and UpdatedKCL without using auxiliary variable; in contrast, they share similar performance for β_v estimation.

5 Real Data Analysis

We applied the proposed methods to analyze the data arising from the Busselton Health Study (Cullen 1972; Knuiman et al. 2003). The interest of this study is to evaluate the effect of risk factors on coronary heart disease and stroke, and compare the effect size. The study consists of 1612 subjects aged 40-89 years who participated in the 1981 Busselton Health Survey and did not experience either coronary heart disease or stroke at the time that they entered the study. The event time was defined to be censored if the subject was free of that event type by December 31, 1998, or lost to follow-up during the study.

The main exposure for the two diseases is serum ferritin. The case-cohort design was adopted in the study to reduce costs of preserving stored serum and measuring serum ferritin. Consequently, serum ferritin was available only for the subcohort members and the subjects who experienced coronary heart disease or stroke or both. The risk factors age, triglycerides, blood pressure treatment were measured for all the subjects in the whole cohort.

We considered a subset of $N = 1210$ subjects in the cohort, to exclude subjects that did not have viable blood serum samples, or had missing values in the risk factors. There were 217 coronary heart disease cases and 118 stroke cases. The size of the random subcohort is 450 (that is, $\sum_{i=1}^{1210} \xi_i = 450$), and the random subcohort selection probability is $\alpha = 450/1210 = 0.37$. As described in Section 2.2, the serum ferritin measurements are obtained only for subjects who are in the random subcohort (i.e., $\xi_i = 1$), or experienced either coronary heart disease or stroke (i.e., $\prod_{j=1}^2 (1 - \delta_{ij}) = 0$). The total number of subjects with serum ferritin measurements is 626 (i.e., $\sum_{i=1}^{1210} \left[\left\{ 1 - \prod_{j=1}^2 (1 - \delta_{ij}) \right\} + \left\{ \prod_{j=1}^2 (1 - \delta_{ij}) \right\} \xi_i \right] = 626$).

We applied the methods KC, KCL, UpdatedKC, and Updated KCL to fit the marginal Cox proportional hazards model

$$\lambda_{ik}(t; Z_i) = \lambda_{0k}(t) \exp(\beta_{xk} X_i + \beta_{vk}^T V_i), \quad k=1, 2. \quad (11)$$

where X_i represents the log of serum ferritin, and V_i is the vector of age, triglycerides, and blood pressure treatment. Xue et al. (2013) developed three correlation tests based on Schoenfeld-type residuals to check the proportional hazards assumption of the Cox model with case-cohort data. We implemented these tests for the marginal Cox model (11) for each disease k separately, and the results are summarized in the Supplementary Material. The results indicated that there is not sufficient evidence for the violation of the proportional hazards assumption.

Seven risk factors smoking status, systolic blood pressure, cholesterol level, haemoglobin, BMI, diabetes treatment, and sex, were measured for all the subjects in the cohort. Analogous to the surrogate creation method in the simulations, we “created” the surrogates W_i by using the least square method to fit the linear regression model:

$$X_i = \alpha_0 + \alpha_w^T \tilde{W}_i + \alpha_v^T V_i + \sum_{k=1}^2 \alpha_{\delta,k} \delta_{ik} + \sum_{k=1}^2 \alpha_{s,k} S_{ik} + e_i,$$

based on the 626 subjects with serum ferritin measurements, where $\tilde{W}_i$ is the vector of the seven risk factors smoking status, systolic blood pressure, cholesterol level, haemoglobin, BMI, diabetes treatment, and sex. We make use of the “created surrogates” W_i to construct our UpdatedKC and UpdatedKCL estimators.

Table 4 provides the results from applying the methods KC, KCL, UpdatedKC, and UpdatedKCL to the data set. The updated methods are more efficient in the sense that the estimated standard errors are smaller and the 95% confidence intervals are narrower. The UpdatedKCL method is the most efficient one among the four methods included in the analysis. In particular, it is interesting to note that the updated methods show statistically significant increase in the risk of stroke associated with higher level of triglycerides, whereas the association is not statistically significant when using KC and KCL.

Our proposed method can be used to test whether the effect of covariates is the same for the different types of diseases under consideration. For the analysis of the Busselton Health Study, we further tested whether the effects of the confounding variables age, triglycerides, and blood pressure treatment are the same for both the coronary heart disease and stroke. We performed four Wald tests based on the methods KC, KCL, UpdatedKC, and UpdatedKCL for the null hypothesis: $H_0: \beta_{Age,1} = \beta_{Age,2}; \beta_{Tri,1} = \beta_{Tri,2}; \beta_{BPT,1} = \beta_{BPT,2}$, where “Tri” and “BPT” stand for triglycerides and blood pressure treatment, respectively. The p -value of the Wald tests was 0.313, 0.138, 0.140, and 0.168, respectively, suggesting there is little evidence against H_0 . Thus, we fitted a simplified marginal Cox proportional hazards model:

$$\lambda_{ik}(t; Z_i) = \lambda_{0k}(t) \exp \left(\beta_{LSF,k} LSF_i + \beta_{Age} Age_i + \beta_{Tri} Tri_i + \beta_{BPT} BPT_i \right), \quad k=1, 2, \quad (12)$$

where “LSF” stands for the expensive variable log of serum ferritin. The results are provided in Table 5, and are similar to those in Table 4.

6 Discussion

Breslow et al. (2009b) commented at the end of their paper: “...There is no longer any excuse for epidemiologists and statisticians to waste valuable information by inefficient analysis of data from stratified case-control and case-cohort studies”. In this paper, we proposed a class of updated-estimators that make effective use of valuable information in the case-cohort studies with multiple disease outcomes. Our proposed updated-estimators are novel even for the case-cohort studies with single disease outcomes. Furthermore, the proposed methods can be easily extended to the stratified case-cohort studies.

In the case-cohort studies, some covariates are prohibitively expensive and are measured only for the subcohort members and the cases. When there are auxiliary variables for these covariates, the updated-estimators are more efficient than the existing weighted estimators for estimating the effects of these expensive covariates. When there is no explicit auxiliary variable in the data set, we illustrated in the data analysis of Busselton Health Study that we can “create” auxiliary variables by a simple regression method based on the available information.

In contrast to the expensive covariates that are observed only in a proportion of the cohort, some covariates are measured for the whole cohort members. The updated-estimators are much more efficient compared to the existing weighted estimators when estimating the effects of these covariates.

In the Busselton Health Study, the blood samples were not viable for some cohort members, including those in the subcohort or who experienced failures. Consequently, measurements of serum ferritin for these cohort members were missing. In the data analysis in this paper, we disregarded the cohort members without viable blood samples. It is of future research interest to extend the proposed methods in this paper to handle missing data and make further use of the available information.

In this paper, we extended the updated method by Chen (2002) to the case-cohort setting with attention restricted to the Cox model. One may similarly extend the updated method to the additive hazards model (Jiang and Zhou 2007) or the accelerated failure time model (Yu 2013) under the case-cohort study design. Extension of the updated method to other survival models, including the transformation model, is of future research interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors thank the co-editor for his valuable comments which have led to significant improvement of the manuscript. The authors also thank Dr. Matthew Knuiman and the Busselton Population Medical Research Foundation for permission to use the data to illustrate the proposed methods. This work was supported by grants R01 ES 021900 and P01 CA 142538 from U.S. National Institutes of Health. Yan's research was partly supported by a Discovery Grant from Natural Sciences and Engineering Research Council of Canada (NSERC).

Reference

- 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]
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the Whole Cohort in the Analysis of Case-Cohort Data. *American Journal of Epidemiology*. 2009a; 169:1398–1405. [PubMed: 19357328]
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz-Thompson Estimation of Model Parameters from Two-Phase Stratified Samples: Application in Epidemiology. *Statistics in Biosciences*. 2009b; 1:32–49. [PubMed: 20174455]
- Breslow NE, Wellner JA. Weighted Likelihood for Semiparametric Models and Two-Phase Stratified Samples, with Application to Cox Regression. *Scandinavian Journal of Statistics*. 2007; 34:86–102.
- Cai J, Prentice RL. Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data. *Biometrika*. 1995; 82:151–164.
- Carroll, RJ., Ruppert, D., Stefanski, LA., Crainiceanu, CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed.. Chapman & Hall/CRC; Boca Raton: 2006.
- Chen K. Generalized Case-Cohort Sampling. *Journal of the Royal Statistical Society: Series B*. 2001; 63:791–809.
- Chen K, Lo S-H. Case-Cohort and Case-Control Analysis with Cox's Model. *Biometrika*. 1999; 86:755–764.
- Chen Y-H. Cox Regression in Cohort Studies with Validation Sampling. *Journal of the Royal Statistical Society: Series B*. 2002; 64:51–62.
- Clayton D, Cuzick J. Multivariate Generalizations of the Proportional Hazards Model. *Journal of the Royal Statistical Society: Series A*. 1985; 148:82–117.
- Cox DR. Regression Models and Life-Tables (with Discussion). *Journal of the Royal Statistical Society: Series B*. 1972; 34:187–220.
- Cullen KJ. Mass Health Examinations in the Busselton Population, 1966 to 1970. *Australian Journal of Medicine*. 1972; 2:714–718.
- D'Angio GJ, Breslow NE, Beckwith JB, Evans A, Baum H, De Lorimier A, Fernbach D, Hrabovsky E, Jones B, Kelalis P. Treatment of Wilms Tumor: Results of the Third National Wilms Tumor Study. *Cancer*. 1989; 64:349–360. [PubMed: 2544249]
- Jiang J, Zhou H. Additive Hazard Regression with Auxiliary Covariates. *Biometrika*. 2007; 94:359–369.
- Kalbfleisch, JD., Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2nd edn. Wiley; Hoboken: 2002.
- Kang S, Cai J. Marginal Hazards Model for Case-Cohort Studies with Multiple Disease Outcomes. *Biometrika*. 2009; 96:887–901. [PubMed: 23946547]
- Keogh RH, White IR. Using Full-Cohort Data in Nested Case-Control and Case-Cohort Studies by Multiple Imputation. *Statistics in Medicine*. 2013; 32:4021–4043. [PubMed: 23613433]
- Kim S, Cai J, Lu W. More Efficient Estimators for Case-Cohort Studies. *Biometrika*. 2013; 100:695–708. [PubMed: 24634519]
- Knuiman MW, Divitini ML, Olynyk JK, Cullen DJ, Bartholomew HC. Serum Ferritin and Cardiovascular Disease: A 17-year Follow-Up Study in Busselton, Western Australia. *American Journal of Epidemiology*. 2003; 158:144–149. [PubMed: 12851227]

- Kulich M, Lin DY. Improving the Efficiency of Relative-Risk Estimation in Case-Cohort Studies. *Journal of the American Statistical Association*. 2004; 99:832–844.
- Lin DY, Wei LJ. The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*. 1989; 84:1074–1078.
- Marti H, Chavance M. Multiple Imputation Analysis of Case-Cohort Studies. *Statistics in Medicine*. 2011; 30:1595–1607. [PubMed: 21351290]
- Nan B. Efficient Estimation for Case-Cohort Studies. *The Canadian Journal of Statistics*. 2004; 32:403–419.
- Prentice RL. A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika*. 1986; 73:1–11.
- Samuelsen SO, Anestad H, Skrondal A. Stratified Case-Cohort Analysis of General Cohort Sampling Designs. *Scandinavian Journal of Statistics*. 2007; 34:103–119.
- Scheike TH, Martinussen T. Maximum Likelihood Estimation for Cox's Regression Model under Case-Cohort Sampling. *Scandinavian Journal of Statistics*. 2004; 31:283–293.
- Self SG, Prentice RL. Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies. *The Annals of Statistics*. 1988; 16:64–81.
- Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*. 1989; 84:1065–1073.
- Xue X, Xie X, Gunter M, Rohan TE, Wassertheil-Smoller S, Ho GY, Cirillo D, Yu H, Strickler HD. Testing the Proportional Hazards Assumption in Case-Cohort Analysis. *BMC Medical Research Methodology*. 2013; 13:88. [PubMed: 23834739]
- Yu M. Adjusted Regression Estimation for Time-to-Event Data with Differential measurement Error. *Biometrika*. 2013; 100:757–763.
- Zeng D, Lin DY. Efficient Estimation of Semiparametric Transformation Models for Two-Phase Cohort Studies. *Journal of the American Statistical Association*. 2014; 109:371–383. [PubMed: 24659837]

Table 1

Simulation results under the marginal Cox model with common effect (9)

Event Proportion	Method	Auxiliary Variable	Estimating β_1					Estimating β_2				
			Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
[6%, 11%]	KC		0.006	0.140	0.130	93	1	0.016	0.136	0.131	94	1
	KCL		0.005	0.134	0.124	92	1.10	0.011	0.128	0.124	93	1.13
	UpdatedKC	None	0.006	0.141	0.129	93	0.98	0.003	0.103	0.102	95	1.72
		Weak	0.008	0.126	0.115	92	1.25	0.001	0.099	0.098	95	1.90
		Strong	0.007	0.109	0.101	91	1.64	0.000	0.094	0.094	94	2.08
	UpdatedKCL	None	0.006	0.135	0.124	91	1.07	0.002	0.102	0.099	94	1.75
		Weak	0.008	0.122	0.111	91	1.32	0.001	0.098	0.096	94	1.92
		Strong	0.007	0.108	0.099	92	1.69	0.000	0.094	0.093	94	2.09
	KC		0.009	0.113	0.107	93	1	0.009	0.108	0.107	95	1
	KCL		0.008	0.104	0.099	93	1.18	0.004	0.098	0.097	94	1.22
	UpdatedKC	None	0.010	0.115	0.106	92	0.97	-0.001	0.080	0.079	94	1.80
		Weak	0.012	0.100	0.092	92	1.28	-0.002	0.076	0.074	94	2.04
[11%, 20%]		Strong	0.010	0.084	0.078	92	1.83	-0.003	0.071	0.070	94	2.31
	UpdatedKCL	None	0.009	0.107	0.098	91	1.12	-0.001	0.078	0.076	93	1.90
		Weak	0.012	0.095	0.086	92	1.43	-0.003	0.075	0.072	94	2.09
		Strong	0.010	0.081	0.075	91	1.94	-0.003	0.071	0.069	94	2.33
	KC		0.004	0.093	0.089	94	1	0.010	0.091	0.088	93	1
	KCL		0.001	0.083	0.078	93	1.26	0.005	0.077	0.075	94	1.38
	UpdatedKC	None	0.005	0.094	0.088	93	0.97	0.001	0.065	0.062	93	1.95
		Weak	0.006	0.080	0.075	93	1.35	0.000	0.061	0.058	94	2.23
		Strong	0.005	0.064	0.061	94	2.11	0.000	0.056	0.054	93	2.63
	UpdatedKCL	None	0.001	0.084	0.077	92	1.22	0.001	0.063	0.059	93	2.10
		Weak	0.004	0.073	0.067	93	1.63	0.000	0.060	0.056	93	2.34
		Strong	0.004	0.060	0.057	94	2.37	0.000	0.056	0.053	93	2.68

Table 2

Simulation results under the marginal Cox model with disease-specific effect (10)

Disease Type		Event Proportion	Method	Auxiliary Variable		Estimating β_x				Estimating β_z					
						Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
Disease 1	[6%, 11%]	KC				0.007	0.193	0.184	92	1	0.027	0.195	0.188	93	1
			KCL				0.004	0.181	0.174	92	1.14	0.017	0.178	0.174	92
		UpdatedKC	None			0.008	0.195	0.183	91	0.98	0.010	0.164	0.163	93	1.41
			Weak			0.011	0.183	0.171	91	1.11	0.009	0.162	0.160	93	1.44
	Strong			0.008	0.168	0.161	91	1.32	0.008	0.158	0.158	93	1.52		
		UpdatedKCL	None			0.005	0.182	0.173	91	1.12	0.009	0.160	0.159	92	1.47
	Weak			0.007	0.175	0.164	91	1.21	0.009	0.160	0.156	92	1.49		
		Strong			0.006	0.165	0.157	91	1.37	0.008	0.157	0.154	92	1.54	
	[11%, 20%]	KC			0.007	0.154	0.148	93	1	0.021	0.155	0.150	94	1	
			KCL			0.003	0.136	0.132	94	1.29	0.012	0.131	0.131	95	1.39
		UpdatedKC	None			0.009	0.157	0.146	93	0.96	0.008	0.123	0.122	93	1.59
			Weak			0.011	0.142	0.133	93	1.18	0.007	0.120	0.119	93	1.67
		Strong			0.007	0.127	0.121	92	1.48	0.007	0.115	0.115	94	1.80	
			UpdatedKCL	None			0.003	0.138	0.130	93	1.25	0.008	0.119	0.117	93
		Weak			0.007	0.130	0.123	92	1.41	0.006	0.117	0.115	93	1.74	
			Strong			0.005	0.121	0.116	93	1.63	0.007	0.114	0.113	94	1.83
Disease 2	[20%, 33%]	KC			0.003	0.126	0.122	94	1	0.022	0.130	0.123	91	1	
			KCL			-0.003	0.103	0.099	94	1.50	0.012	0.099	0.097	94	1.70
		UpdatedKC	None			0.004	0.129	0.120	93	0.95	0.009	0.095	0.094	95	1.87
			Weak			0.004	0.114	0.106	93	1.22	0.008	0.092	0.090	94	2.00
	Strong			0.002	0.097	0.093	94	1.68	0.008	0.088	0.087	94	2.18		
		UpdatedKCL	None			-0.003	0.105	0.098	93	1.44	0.009	0.090	0.088	94	2.09
	Weak			0.000	0.097	0.092	94	1.68	0.008	0.089	0.087	93	2.14		
		Strong			0.001	0.090	0.086	94	1.97	0.008	0.086	0.085	94	2.25	
	[6%, 11%]	KC			0.007	0.160	0.148	93	1	0.015	0.157	0.150	95	1	
			KCL			0.007	0.152	0.143	93	1.11	0.009	0.148	0.142	94	1.13
		UpdatedKC	None			0.009	0.162	0.147	92	0.97	0.000	0.127	0.122	96	1.53

Disease Type	Event Proportion	Method	Auxiliary Variable	Estimating $\hat{\beta}_1$					Estimating $\hat{\beta}_2$				
				Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
			Weak	0.011	0.146	0.134	93	1.21	-0.002	0.121	0.119	95	1.69
			Strong	0.010	0.128	0.122	93	1.56	-0.003	0.116	0.116	95	1.83
		UpdatedKCL	None	0.009	0.154	0.141	92	1.08	-0.001	0.125	0.120	95	1.58
			Weak	0.011	0.141	0.130	93	1.29	-0.002	0.120	0.117	95	1.71
			Strong	0.009	0.127	0.119	93	1.61	-0.002	0.116	0.114	95	1.83
	[11%, 20%]	KC		0.013	0.131	0.122	93	1	0.006	0.126	0.123	95	1
		KCL		0.011	0.120	0.113	93	1.19	0.001	0.114	0.112	94	1.22
		UpdatedKC	None	0.016	0.133	0.121	92	0.97	-0.005	0.099	0.095	95	1.62
			Weak	0.018	0.117	0.107	93	1.23	-0.007	0.094	0.091	94	1.78
			Strong	0.015	0.100	0.094	93	1.69	-0.007	0.090	0.087	94	1.97
		UpdatedKCL	None	0.011	0.123	0.112	92	1.12	-0.005	0.096	0.092	94	1.72
			Weak	0.015	0.112	0.101	92	1.37	-0.006	0.093	0.089	94	1.85
			Strong	0.012	0.098	0.091	92	1.78	-0.006	0.089	0.086	95	2.01
	[20%, 33%]	KC		0.006	0.108	0.103	94	1	0.007	0.106	0.102	94	1
		KCL		0.003	0.095	0.091	93	1.28	0.003	0.090	0.088	94	1.38
		UpdatedKC	None	0.007	0.109	0.101	93	0.97	-0.002	0.080	0.076	94	1.74
			Weak	0.010	0.094	0.088	93	1.32	-0.007	0.075	0.072	94	1.99
			Strong	0.009	0.076	0.074	94	2.00	-0.004	0.070	0.068	95	2.29
		UpdatedKCL	None	0.004	0.097	0.089	93	1.22	-0.002	0.077	0.072	93	1.88
			Weak	0.007	0.087	0.079	92	1.54	-0.003	0.074	0.069	94	2.07
			Strong	0.007	0.073	0.070	94	2.15	-0.003	0.070	0.067	94	2.31

Table 3

Simulation results with “created surrogates”

Disease Type	Method	Auxiliary Variable	Estimating β_1					Estimating β_2				
			Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
Disease 1	KC		0.007	0.154	0.148	93	1	0.021	0.155	0.150	94	1
	KCL		0.003	0.136	0.132	94	1.29	0.012	0.131	0.131	95	1.39
	UpdatedKC	None	0.009	0.157	0.146	93	0.96	0.008	0.123	0.122	93	1.59
		“Created”	-0.009	0.143	0.136	94	1.16	0.009	0.126	0.122	93	1.50
	UpdatedKCL	None	0.003	0.138	0.130	93	1.25	0.008	0.119	0.117	93	1.70
Disease 2		“Created”	-0.004	0.134	0.125	92	1.31	0.011	0.119	0.117	93	1.67
	KC		0.013	0.131	0.122	93	1	0.006	0.126	0.123	95	1
	KCL		0.011	0.120	0.113	93	1.19	0.001	0.114	0.112	94	1.22
	UpdatedKC	None	0.016	0.133	0.121	92	0.97	-0.005	0.099	0.095	95	1.62
		“Created”	0.003	0.121	0.112	93	1.16	-0.004	0.101	0.096	94	1.56
	UpdatedKCL	None	0.011	0.123	0.112	92	1.12	-0.005	0.096	0.092	94	1.72
		“Created”	0.003	0.117	0.104	92	1.24	-0.002	0.097	0.091	94	1.69

Table 4

Analysis of Busselton Health Study under the marginal Cox model (11)

Disease	Variable	Method	Est	ESE	HR	95% CI
CHD	log(ferritin)	KC	0.097	0.093	1.10	(0.92,1.32)
		KCL	0.142	0.088	1.15	(0.97,1.37)
		UpdatedKC	0.096	0.090	1.10	(0.92,1.31)
		UpdatedKCL	0.119	0.083	1.13	(0.96,1.33)
	Age	KC	0.063	0.009	1.07	(1.05,1.08)
		KCL	0.066	0.007	1.07	(1.05,1.08)
		UpdatedKC	0.067	0.007	1.07	(1.06,1.09)
		UpdatedKCL	0.066	0.007	1.07	(1.05,1.08)
	Triglycerides	KC	0.248	0.052	1.28	(1.16,1.42)
		KCL	0.269	0.048	1.31	(1.19,1.44)
		UpdatedKC	0.284	0.041	1.33	(1.23,1.44)
		UpdatedKCL	0.286	0.039	1.33	(1.23,1.44)
	Blood pressure treatment	KC	0.262	0.206	1.30	(0.87,1.95)
		KCL	0.268	0.194	1.31	(0.89,1.91)
		UpdatedKC	0.326	0.170	1.39	(0.99,1.93)
		UpdatedKCL	0.320	0.169	1.38	(0.99,1.92)
Stroke	log(ferritin)	KC	0.193	0.135	1.21	(0.93,1.58)
		KCL	0.180	0.125	1.20	(0.94,1.53)
		UpdatedKC	0.178	0.131	1.19	(0.92,1.54)
		UpdatedKCL	0.165	0.121	1.18	(0.93,1.50)
	Age	KC	0.080	0.012	1.08	(1.06,1.11)
		KCL	0.085	0.013	1.09	(1.06,1.12)
		UpdatedKC	0.085	0.011	1.09	(1.07,1.11)
		UpdatedKCL	0.084	0.012	1.09	(1.06,1.11)
	Triglycerides	KC	0.193	0.107	1.21	(0.98,1.50)
		KCL	0.170	0.091	1.19	(0.99,1.42)
		UpdatedKC	0.191	0.096	1.21	(1.00,1.46)
		UpdatedKCL	0.190	0.087	1.21	(1.02,1.43)
	Blood pressure treatment	KC	0.663	0.249	1.94	(1.19,3.16)
		KCL	0.699	0.235	2.01	(1.27,3.19)
		UpdatedKC	0.764	0.217	2.15	(1.40,3.28)
		UpdatedKCL	0.741	0.217	2.10	(1.37,3.21)

Table 5

Analysis of Busselton Health Study under the simplified marginal Cox model (12)

Variable	Method	Est	ESE	HR	95% CI
log(ferritin) on CHD	KC	0.092	0.095	1.10	(0.91,1.32)
	KCL	0.145	0.090	1.16	(0.97,1.38)
	UpdatedKC	0.092	0.091	1.10	(0.92,1.31)
	UpdatedKCL	0.119	0.086	1.13	(0.95,1.33)
log(ferritin) on stroke	KC	0.186	0.130	1.20	(0.93,1.56)
	KCL	0.172	0.122	1.19	(0.94,1.51)
	UpdatedKC	0.172	0.127	1.19	(0.93,1.53)
	UpdatedKCL	0.148	0.120	1.16	(0.92,1.47)
Age	KC	0.069	0.007	1.07	(1.06,1.09)
	KCL	0.071	0.007	1.07	(1.06,1.09)
	UpdatedKC	0.072	0.006	1.08	(1.06,1.09)
	UpdatedKCL	0.072	0.006	1.07	(1.06,1.09)
Triglycerides	KC	0.232	0.054	1.26	(1.13,1.40)
	KCL	0.258	0.042	1.29	(1.19,1.41)
	UpdatedKC	0.258	0.042	1.29	(1.19,1.41)
	UpdatedKCL	0.259	0.039	1.30	(1.20,1.40)
Blood pressure treatment	KC	0.408	0.173	1.50	(1.07,2.11)
	KCL	0.423	0.163	1.53	(1.11,2.10)
	UpdatedKC	0.414	0.136	1.51	(1.16,1.98)
	UpdatedKCL	0.416	0.134	1.52	(1.17,1.97)