NIH Public Access
Author Manuscript
Biometrika. Author manuscript; available in PMC 2014 June 10.

Published in final edited form as:
Biometrika. 2013 ; 100(3): 695–708. doi:10.1093/biomet/ast018.

More efficient estimators for case-cohort studies

S. KIM,
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

J. CAI, and
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

W. LU
Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A

S. KIM: kimso@live.unc.edu; J. CAI: cai@bios.unc.edu; W. LU: lu@stat.ncsu.edu

Summary

The case-cohort study design, used to reduce costs in large cohort studies, is a random sample of the entire cohort, named the subcohort, augmented with subjects having the disease of interest but not in the subcohort sample. When several diseases are of interest, several case-cohort studies may be conducted using the same subcohort, with each disease analyzed separately, ignoring the additional exposure measurements collected on subjects with the other diseases. This is not an efficient use of the data, and in this paper, we propose more efficient estimators. We consider both joint and separate analyses for the multiple diseases. We propose an estimating equation approach with a new weight function, and we establish the consistency and asymptotic normality of the resulting estimator. Simulation studies show that the proposed methods using all available information gain efficiency. We apply our proposed method to the data from the Busselton Health Study.

Some key words

Case-cohort study; Multiple disease outcomes; Multivariate failure time; Proportional hazards; Survival analysis

1. Introduction

For large epidemiologic cohort studies, assembling some types of covariate information, e.g. measuring genetic information or chemical exposures from stored blood samples, for all cohort members may entail enormous cost. With cost in mind, Prentice (1986) proposed the case-cohort study design, which requires covariate information only for a random sample of the cohort, named the subcohort, as well as for all subjects with the disease of interest. One important advantage of the case-cohort study design is that the same subcohort can be used for studying different diseases, whereas for designs such as the nested case-control design, new matching of cases and controls is needed for different diseases (Langholz & Thomas, 1990; Wacholder et al., 1991).

Many methods have been proposed for case-cohort data under the proportional hazards model. Prentice (1986) and Self & Prentice (1988) studied a pseudo-likelihood approach, which is a modification of the partial likelihood method (Cox, 1975) that weights the contributions of the cases and subcohort differently. To improve the efficiency of the pseudo-likelihood estimator, Chen & Lo (1999) and Chen (2001b) studied different classes

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of estimating equations and used a local type of average as weight, respectively. Borgan et al. (2000) proposed using time-varying weights, and Kulich & Lin (2004) developed a class of weighted estimators by using all available covariate data for the full cohort. Breslow & Wellner (2007) considered the semiparametric model using inverse probability weighted methods with two-phase stratified samples. Various other semiparametric survival models have also been modified to accommodate case-cohort studies (e.g. Chen, 2001a; Chen & Zucker, 2009; Kong et al., 2004; Kulich & Lin, 2000; Lu & Tsiatis, 2006).

Taking advantage of the case-cohort design, several diseases are often studied using the same subcohort. In such situations, the information on the expensive exposure measure is available on the subcohort as well as any subjects with any of the diseases of interest. For example, in the Busselton Health Study, two case-cohort studies were conducted to investigate the effect of serum ferritin on coronary heart disease and on stroke, respectively (Knuiman et al., 2003). Serum ferritin was measured on the subcohort, a random sample of the cohort, as well as in all subjects with coronary heart disease and/or stroke. Typically, the coronary heart disease analysis would not include any exposure information collected on stroke patients not in the subcohort, and vice versa. In this paper, we develop more efficient estimators for a single disease outcome, which can effectively use all available exposure information. Because it is often of interest to compare the effect of a risk factor on different diseases, we propose a more efficient version of the Kang & Cai (2009) test of association across multiple diseases.

2. Model and Estimation

2·1. Model definitions and assumptions

Suppose that there are n independent subjects in a cohort study with K diseases of interest. Let T_{ik} denote the potential failure time and C_{ik} denote the potential censoring time for disease k of subject i. Let $X_{ik} = \min(T_{ik}, C_{ik})$ denote the observed time, $\Delta_{ik} = I(T_{ik} \le C_{ik})$ the indicator for failure, and $N_{ik}(t) = I(X_{ik} \le t, \Delta_{ik} = 1)$ and $Y_{ik}(t) = I(X_{ik} \ge t)$ the counting and at-risk processes for disease k of subject i, respectively, where $I(\cdot)$ is the indicator function. Let $Z_{ik}(t)$ be a $p \times 1$ vector of possibly time-dependent covariates for disease k of subject i at time t. The time-dependent covariates are assumed to be external (Kalbfleisch & Prentice, 2002). Let τ denote the end of study time. We assume that T_{ik} is independent of C_{ik} given the covariates Z_{ik} and follows the multiplicative intensity process (Cox, 1972)

$$\lambda_{ik}\{t|Z_{ik}(t)\}=Y_{ik}(t)\lambda_{0k}(t)e^{\beta_0^T Z_{ik}(t)}, \quad (1)$$

where $\lambda_{0k}(t)$ is an unspecified baseline hazard function for disease k of subject i and β_0 is p-dimensional vector of fixed and unknown parameters. Model (1) can incorporate disease-specific effect model, $\lambda_{ik}\{t|Z_{ik}^*(t)\}=Y_{ik}(t)\lambda_{0k}(t)e^{\beta_k^T Z_{ik}^*(t)}$, as a special case. Specifically, we define $\beta_0^T=(\beta_1^T,\ldots,\beta_k^T,\ldots,\beta_K^T)$ and $Z_{ik}(t)^T=[0_{i1}^T,\ldots,0_{i(k-1)}^T,\{Z_{ik}^*(t)\}^T,0_{i(k+1)}^T,\ldots,0_{iK}^T]$, letting 0^T be a $1 \times p$ zero vector. Then we have $\beta_0^T Z_{ik}(t)=\beta_k^T Z_{ik}^*(t)$.

Assume that there are $\tilde{n}$ subjects in the subcohort. Let ξ_i be an indicator for subcohort membership, i.e. $\xi_i = 1$ denotes that subject i is selected into the subcohort and $\xi_i = 0$ denotes otherwise. Let $\alpha = \text{pr}(\xi_i = 1) = \tilde{n}/n$ denote the selection probability of subject i into the subcohort. The covariates $Z_{ik}(t)$ $(0 \le t \le \tau)$ are measured for subjects in the subcohort and those with any disease of interest.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2·2. Estimation for univariate failure time

First, we consider the situation in which only one disease is of interest, but covariate information is available for subjects with other diseases. In the Busselton Health study, for example, this corresponds to the situation in which we are interested in the effect of serum ferritin on coronary heart disease with additional serum ferritin measurements available on subjects outside the subcohort who had stroke.

In this situation, the observable information is $\{X_{ik},\Delta_{ik},\xi_i, Z_{ik}(t), 0 \le t \le X_{ik}\}$ when $\xi_i = 1$ or $\Delta_{ik} = 1$, and is $(X_{ik},\Delta_{ik},\xi_i)$ when $\xi_i = 0$ and $\Delta_{ik} = 0$ ($k = 1,\ldots, K$). If we are interested in disease k and ignore the covariate information collected on subjects with other diseases, we can use Borgan et al. (2000)'s estimator with time-varying weights. Specifically, the estimator is the solution to

$$\widehat{U}_k(\beta) \equiv \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_{ik}(t) - \frac{\widehat{S}_k^{(1)}(\beta,t)}{\widehat{S}_k^{(0)}(\beta,t)} \right\} dN_{ik}(t) = 0, \quad (2)$$

where $\widehat{S}_k^{(d)}(\beta,t) = n^{-1}\sum_{i=1}^{n} \rho_{ik}(t) Y_{ik}(t) Z_{ik}(t)^{\otimes d} e^{\beta^T Z_{ik}(t)}$ for $d = 0,1$ and 2 with $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$, and the time-varying weight $\rho_{ik}(t) = \Delta_{ik} + (1-\Delta_{ik})\xi_i \widehat{\alpha}_k^{-1}(t)$ with $\widehat{\alpha}_k(t) = \sum_{i=1}^{n} \xi_i (1-\Delta_{ik}) Y_{ik}(t) / \{\sum_{i=1}^{n} (1-\Delta_{ik}) Y_{ik}(t)\}$. Here $\widehat{a_k}(t)$, an estimator for the true selection probability a, is the proportion of the sampled censored subjects for disease k among censored subjects who remain in the risk set at time t for disease k. This estimator does not use the covariate information from subjects outside the subcohort who had other diseases.

To use the collected covariate information on subjects who are outside the subcohort and have other diseases, we consider the pseudo-partial likelihood score equations

$$\tilde{U}_k(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_{ik}(t) - \frac{\tilde{S}_k^{(1)}(\beta,t)}{\tilde{S}_k^{(0)}(\beta,t)} \right\} dN_{ik}(t) = 0; \quad (3)$$

where

$$\tilde{S}_k^{(d)}(\beta,t) = n^{-1} \sum_{i=1}^{n} \psi_{ik}(t) Y_{ik}(t) Z_{ik}(t)^{\otimes d} e^{\beta^T Z_{ik}(t)} \quad (d = 0,1,2),$$
$$\psi_{ik}(t) = \left\{ 1 - \prod_{j=1}^{K} (1-\Delta_{ij}) \right\} + \prod_{j=1}^{K} (1-\Delta_{ij}) \xi_i \tilde{\alpha}_k^{-1}(t),$$

and $\tilde{\alpha}_k(t) = \sum_{i=1}^{n} \xi_i \{\prod_{j=1}^{K} (1-\Delta_{ij})\} Y_{ik}(t) / \sum_{i=1}^{n} \{\prod_{j=1}^{K} (1-\Delta_{ij})\} Y_{ik}(t)$. Here $\tilde{a_k}(t)$ is the proportion of sampled subjects among subjects who do not have any diseases and are remaining in the risk set at time t. Our proposed weight for disease k is $\psi_{ik}(t) = 1$ when $\Delta_{ij} = 1$ for some j, and $\psi_{ik}(t) = \tilde{\alpha}_k^{-1}(t)$ when $\xi_i = 1$ and $\Delta_{ij} = 0$ for all j ($j = 1,\ldots, k$). This weight takes the failure status of the other diseases into consideration, and thus our proposed estimator will use the available covariate information for other diseases.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2·3. Estimation for multivariate failure time

For multivariate failure time data in case-cohort studies, Kang & Cai (2009) proposed the pseudo-likelihood score equations

$$\widehat{U}^M(\beta) \equiv \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \left\{ Z_{ik}(t) - \frac{\widehat{S}_k^{(1)}(\beta,t)}{\widehat{S}_k^{(0)}(\beta,t)} \right\} dN_{ik}(t) = 0, \quad (4)$$

with the corresponding solution denoted $\hat{\beta}^{\hat{M}}$.

As with Borgan et al. (2000)'s estimator, when calculating the contribution of disease k in the estimating equation, the quantity $\widehat{S}_k^{(d)}(\beta,t)$ does not use the covariate information collected on subjects with other diseases outside the subcohort. In order to improve efficiency, we consider the pseudo-likelihood score equations with new weights

$$\tilde{U}^M(\beta) \equiv \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \left\{ Z_{ik}(t) - \frac{\tilde{S}_k^{(1)}(\beta,t)}{\tilde{S}_k^{(0)}(\beta,t)} \right\} dN_{ik}(t) = 0. \quad (5)$$

When there is only a single disease of interest, i.e. $K = 1$, (5) reduces to (3). Let $\hat{\beta}^{\tilde{M}}$ denote the solution of equation (5). We estimate the baseline cumulative hazard function for disease k using a Breslow–Aalen type estimator $\tilde{\Lambda}_{0k}^M(\tilde{\beta}^M, t)$, where

$$\tilde{\Lambda}_{0k}^M(\beta,t) = \int_0^t \frac{\sum_{i=1}^n dN_{ik}(u)}{n\tilde{S}_k^{(0)}(\beta,u)}. \quad (6)$$

3. Asymptotic properties

Because the estimators for the univariate failure time are special cases of those for the multivariate failure time, we present results only for the multivariate case. We make the following assumptions:

- **a.** $(T_i, C_i, Z_i, i = 1,\ldots, n)$ are independently and identically distributed, where $T_i = (T_{i1}, \ldots, T_{iK})^T$, $C_i = (C_{i1},\ldots, C_{iK})^T$, and $Z_i = (Z_{i1},\ldots, Z_{iK})^T$;
- **b.** $\mathrm{pr}\{Y_{ik}(t) = 1\} > 0$ for $t \in [0, \tau]$, $i = 1,\ldots, n$ and $k = 1,\ldots, K$;
- **c.** $|Z_{ik}(0)| + \int_0^\tau |dZ_{ik}(t)| < D_z < \infty$ for $i = 1,\ldots, n$ and $k = 1,\ldots, K$ almost surely, where D_z is a constant;
- **d.** for $d = 0, 1, 2$, there exists a neighborhood $\mathcal{B}$ of β_0 such that $s_k^{(d)}(\beta,t)$ are continuous functions and $\sup_{t\in(0,\tau),\beta\in\mathcal{B}} \| S_k^{(d)}(\beta,t) - s_k^{(d)}(\beta,t) \| \to 0$ in probability, where $S_k^{(d)}(\beta,t) = n^{-1}\sum_{i=1}^n Y_{ik}(t) Z_{ik}(t)^{\otimes d} e^{\beta^T Z_{ik}(t)}$;
- **e.** the matrix $A_k(\beta_0) = \int_0^\tau v_k(\beta_0,t) s_k^{(0)}(\beta_0,t) \lambda_{0k}(t) dt$ is positive definite for $k = 1,\ldots, K$, where $v_k(\beta,t) = s_k^{(2)}(\beta,t)/s_k^{(0)}(\beta,t) - e_k(\beta,t)^{\otimes 2}$ and $e_k(\beta,t) = s_k^{(1)}(\beta,t)/s_k^{(0)}(\beta,t)$;

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

f. for all $\beta \in \mathcal{B}$, $t \in [0, \tau]$, and $k = 1,\ldots, K$, $S_k^{(1)}(\beta, t) = \partial S_k^{(0)}(\beta, t)/\partial\beta$, and $S_k^{(2)}(\beta, t) = \partial^2 S_k^{(0)}(\beta, t)/(\partial\beta\partial\beta^T)$, where $S_k^{(d)}(\beta, t), d = 0, 1, 2$, $d = 0, 1, 2$ are continuous functions of $\beta \in \mathcal{B}$ uniformly in $t \in [0, \tau]$ and are bounded on $\mathcal{B} \times [0, \tau]$, and $s_k^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0, \tau]$;

g. for all $k = 1,\ldots, K$, $\int_0^\tau \lambda_{0k}(t)dt < \infty$; and

h. $\lim_{n\to\infty} \tilde{\alpha} = \alpha$, where $\tilde{\alpha} = \tilde{n}/n$ and α is a positive constant.

Theorem 1—Under regularity conditions (a)–(h), $\beta^{\tilde{M}}$ converges in probability to β_0 and $n^{1/2}(\beta^{\tilde{M}} - \beta_0)$ converges in distribution to a mean zero normal distribution with covariance matrix $A(\beta_0)^{-1}\Sigma(\beta_0)A(\beta_0)^{-1}$, where

$$A(\beta) = \sum_{k=1}^{K} A_k(\beta), \quad \Sigma(\beta) = V_I(\beta) + \frac{1-\alpha}{\alpha} V_{II}(\beta),$$

$$V_I(\beta) = E\left\{\sum_{k=1}^{K} W_{1k}(\beta)\right\}^{\otimes 2}, \quad V_{II}(\beta) = E\left\{\sum_{k=1}^{K} \int_0^\tau \Omega_{1k}(\beta, t)d\Lambda_{0k}(t)\right\}^{\otimes 2},$$

$$W_{ik}(\beta) = \int_0^\tau \{Z_{ik}(t) - e_{ik}(\beta, t)\}dM_{ik}(t),$$

$$\Omega_{ik}(\beta, t) = \prod_{j=1}^{K} (1 - \Delta_{ij}) \left[Q_{ik}(\beta, t) - \frac{Y_{ik}(t)E\{\prod_{j=1}^{K}(1-\Delta_{1j})Q_{1k}(\beta, t)\}}{E\{\prod_{j=1}^{K}(1-\Delta_{1j})Y_{1k}(t)\}}\right],$$

$$Q_{ik}(\beta, t) = Y_{ik}(t)\{Z_{ik}(t) - e_k(\beta, t)\}e^{\beta^T Z_{ik}(t)}.$$

The outline of the proof is given in the Appendix. The covariance matrix $\Sigma(\beta_0)$ consists of two parts: $V_I(\beta_0)$ is a contribution to the variance from the full cohort, and $V_{II}(\beta_0)$ is due to sampling the subcohort from the full cohort.

We summarize the asymptotic properties of the proposed baseline cumulative hazard estimator $\tilde{\Lambda}_{0k}^M(\tilde{\beta}^M, t)$ in the next theorem.

Theorem 2—Under regularity conditions (a)–(h), $\tilde{\Lambda}_{0k}^M(\tilde{\beta}^M, t)$ is a consistent estimator of $\Lambda_{0k}(t)$ in $t \in [0, \tau]$ and

$$H(t) = \{H_1(t), \ldots, H_K(t)\}^T = [n^{1/2}\{\tilde{\Lambda}_{01}^M(\tilde{\beta}^M, t) - \Lambda_{01}(t)\}, \ldots, n^{1/2}\{\hat{\Lambda}_{0K}^M(\tilde{\beta}^M, t) - \Lambda_{0K}(t)\}]^T$$

converges weakly to the Gaussian process $\mathcal{H}(t) = \{\mathcal{H}_1(t),\ldots, \mathcal{H}_K(t)\}^T$ in $D[0, \tau]^K$ with mean zero and the following covariance function $\mathcal{R}_{jk}(t, s)$ between $\mathcal{H}_j(t)$ and $\mathcal{H}_k(s)$ for $j \neq k$

$$\mathcal{R}_{jk}(t, s)(\beta_0) = E\{\eta_{1j}(\beta_0, t)\eta_{1k}(\beta_0, s)\} + \frac{1-\alpha}{\alpha} E\{\zeta_{1j}(\beta_0, t)\zeta_{1k}(\beta_0, s)\},$$

where

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\eta_{ik}(\beta,t)=l_k(\beta,t)^T A(\beta)^{-1}\sum_{m=1}^{K} W_{im}(\beta,t)+\int_0^t \frac{1}{s_k^{(0)}(\beta,u)}dM_{ik}(u),$$

$$\zeta_{ik}(\beta,t)=l_k(\beta,t)^T A(\beta)^{-1}\sum_{m=1}^{K}\int_0^\tau \Omega_{im}(\beta,u)d\Lambda_{0m}(u)$$

$$+\prod_{j=1}^{K}(1-\Delta_{ij})\int_0^t Y_{ik}(u)\left[e^{\beta^T Z_{ik}(u)}-\frac{E\{\prod_{j=1}^{K}(1-\Delta_{1j})e^{\beta^T Z_{1k}(u)}Y_{1k}(u)\}}{E\{\prod_{j=1}^{K}(1-\Delta_{1j})Y_{1k}(u)\}}\right]\frac{d\Lambda_{0k}(u)}{s_k^{(0)}(\beta,u)},$$

$$and\ l_k(\beta,t)^T=-\int_0^t e_k(\beta,u)d\Lambda_{0k}(u).$$

The proof of Theorem 2 is outlined in the Appendix.

4. Simulations

We conducted simulation studies to examine the performance of the proposed methods and to compare them with the Borgan et al. (2000) method for univariate outcomes and the Kang & Cai (2009) method for multiple outcomes. We also compared separate analysis with joint analysis. Suppose case-cohort studies have been conducted for diseases 1 and 2. Then covariate information is collected for the subcohort and all the subjects with disease 1 and/or 2. We generated bivariate failure times from the Clayton–Cuzick model (Clayton & Cuzick, 1985) with the conditional survival function

$$S(t_1,t_2|Z_1,Z_2)=\left[\exp\{\int_0^{t_1}\lambda_{01}(t)e^{\beta_1 Z_1}dt/\theta\}+\exp\{\int_0^{t_2}\lambda_{02}(t)e^{\beta_2 Z_2}dt/\theta\}-1\right]^{-\theta},$$

where $\lambda_{0k}(t)$ and β_k (k = 1, 2) are the baseline hazard function and the effect of a covariate for disease k, respectively, and θ is the association parameter between the failure times of the two diseases. Kendall's tau is $\tau_\theta = (2\theta + 1)^{-1}$. Smaller Kendall's tau values represent lower correlation between T_1 and T_2. Values of 0·1, 4, and 10 are used for θ, with corresponding Kendall's tau values 0·83, 0·11, and 0·05, respectively. We set the baseline hazard functions $\lambda_{01}(t) \equiv 2$ and $\lambda_{02}(t) \equiv 4$. We consider the situation $Z_1 = Z_2 = Z$, where Z is generated from a Bernouilli distribution with pr(Z = 1) = 0·5. Censoring times are simulated from a uniform distribution [0, u], where u depends on the specified level of the censoring probability. We set the event proportions of approximately 8% and 20% for k = 1, and 14% and 35% for k = 2. The corresponding u values are 0·08 and 0·22, respectively, for β_1 = 0·1; they are 0·06 and 0·16 for β_1 = log 2. The sample size of the full cohort is set to be n = 1000. We create the subcohort by simple random sampling and consider subcohort sizes of 100 and 200. For each configuration, 2000 simulations were conducted.

In the first set of simulations, we consider the case that disease 1 is of primary interest. We compare the performance of our proposed estimator with the estimator of Borgan et al. (2000). Table 1 summarizes the results. We see that both methods are approximately unbiased. The average of the estimated standard error of the proposed estimator is close to the empirical standard deviation, and the coverage rate of the 95% confidence interval is close to the nominal level. As expected, the variation of the estimators in general decreases as the subcohort size increases. Our proposed estimators have smaller variance relative to the estimators of Borgan et al. (2000) in all cases. This shows that the extra information collected on subjects with the other disease helps to increase efficiency. The efficiency gain is larger in situations with larger event proportions, smaller subcohort sizes and lower correlation. We also considered disease 2 with β_2 = log 2 and conducted additional simulations to compare our proposed estimator with those of Prentice (1986), Self &

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Prentice (1988), Kalbfleisch & Lawless (1988), and Barlow (1994). Similar results were obtained but are not presented in the paper due to space limitations.

In the second set of simulations, we are interested in the joint analysis of the two diseases. We fit the following models:

$$\lambda_{ik}(t|Z_i)=Y_{ik}(t)\lambda_{0k}(t)e^{\beta_k Z_i} \quad (k=1,2;i=1,\ldots,n).$$

We compare the performance of the proposed estimator with the estimator of Kang & Cai (2009). Table 2 provides summary statistics for the estimator of β_1 for different combinations of event proportion, subcohort sample size, and correlation. The estimates from both methods are nearly unbiased, and their estimated standard errors are close to the empirical standard deviations. Our method is more efficient than that of Kang & Cai (2009). The efficiency gain is very limited when the event proportion is small. Higher efficiency gains are associated with smaller subcohort sizes. Estimates for β_2 are not shown in Table 2, but the overall performance is similar to that of β_1.

We also compared separate analysis of the two diseases with the joint analysis using the proposed method. Data were generated satisfying the following model:

$$\lambda_k(t|Z_1,Z_2)=\lambda_{0k}(t)e^{\beta_k Z+\beta_3 Z^*} \quad (k=1,2),$$

where β_1 represents the effect of Z on the risk of disease 1, β_2 represents the effect of Z on the risk of disease 2, and β_3 represents the common effect of Z^* for both diseases. We set $\beta_1 = \beta_2 = \log 2$ and $\beta_3 = 0{\cdot}1$. Table 3 summarizes the results for β_1. The sample standard deviations of Kang & Cai's estimator in the joint analysis are slightly smaller than Borgan's estimator in the separate analysis. The sample standard deviations of the proposed estimators are similar in the joint and separate analyses, and they are smaller than Kang & Cai's and Borgan's estimators, respectively. Conclusions for the estimator of β_2 are similar. We also conducted hypothesis tests for $H_0 : \beta_1 = \beta_2$. Table 4 presents the Type I error rates and power of the tests at the 0·05 significance level. The tests under the separate analysis treat the two estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$, as from two independent samples. Type I error rates from separate analyses are much lower than 5% while those from the joint analysis are close to 5%. The settings for power analysis are the same as before except that $\beta_1 = 0{\cdot}1$ and $\beta_2 = 0{\cdot}7$. Tests based on the proposed methods are more powerful than those based on Kang & Cai's and Borgan's methods, and the joint analysis produces more powerful tests than the separate analysis.

5. Data analysis

We apply the proposed method to analyze data from the Busselton Health Study (Cullen, 1972; Knuiman et al., 2003), conducted in the south-west of Western Australia, and intended to evaluate the association between coronary heart disease and stroke and their risk factors. General health information for adult participants was obtained by questionnaire every three years from 1966 to 1981. This study population consists of 1612 men and women aged 40–89 who participated in 1981 and were free of coronary heart disease or stroke at that time. Coronary heart disease event is defined as hospital admission, any procedure, or death related to coronary heart disease. Stroke event is defined as hospital admission, any procedure, or death from stroke. The outcomes of interest were time to the first coronary heart disease event and time to the first stroke event. The event time for a subject was

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

considered censored if the subject was free of that event type by December 31, 1998 or lost to follow-up during the study period.

One of the main interests of the study was to compare the effect of serum ferritin on coronary heart disease with its effect on stroke. To reduce cost and preserve stored serum, case-cohort sampling was used. Serum ferritin was measured for all the subjects with coronary heart disease and/or stroke as well as those in the subcohort. We conduct a joint analysis of the two diseases. In our analysis, the full cohort consists of 1210 subjects with viable blood serum samples, which includes 174 subjects with only coronary heart disease, 75 with only stroke, and 43 with both diseases. The subcohort consisted of 334 disease-free subjects, 61 with only coronary heart disease, 36 with only stroke, and 19 with both diseases. The total number of assayed sera samples was 626. If a subject was censored and free of both events at the censoring time, then the censoring times for the two disease events were the same. Two subjects died due to both coronary heart disease and stroke, for whom the times for both events were the same. No other subjects died at the first diagnosis of either disease. For this study, it is reasonable to assume, as in the original study (Knuiman et al., 2003), that censoring was conditionally independent of the event processes.

We fit the following model

$$\lambda_k(t|Z_1,Z_2,Z_3,Z_4)=\lambda_{0k}(t)e^{\beta_{1k}Z_1+\beta_{2k}Z_2+\beta_{3k}Z_3+\beta_{4k}Z_4} \quad (k=1,2),$$

where Z_1, Z_2, Z_3, and Z_4 denote the logarithm of serum ferritin level, age in years, triglycerides in millimoles per liter, and whether subjects had blood pressure treatment, respectively. We then tested $H_0 : \beta_{21} = \beta_{22}, \beta_{31} = \beta_{32}, \beta_{41} = \beta_{42}$ based on the proposed method, and the p-value is 0·138. Therefore, we fit the final model

$$\lambda_k(t|Z_1,Z_2,Z_3,Z_4)=\lambda_{0k}(t)e^{\beta_{1k}Z_1+\beta_2Z_2+\beta_3Z_3+\beta_4Z_4} \quad (k=1,2).$$

Table 5 summarizes the results of the final fit. With a 1 unit increase in the logarithm of the serum ferritin level, the hazard ratio for coronary heart disease risk is increased by 16% and for stroke risk by 19%. When we tested $H_0 : \beta_{11} = \beta_{12}$, H_0 was not rejected with the p-value = 0·823. We also fit the same model using Kang & Cai (2009)'s method. The standard errors for the effects of the logarithm of the serum ferritin level are slightly larger, 0·0949 for coronary heart disease and 0·1304 for stroke.

6. Concluding Remarks

When disease rates are low, the efficiency gain of the proposed method is not large. When the event rates are low, the number of cases is small, and consequently, the amount of extra information is small. In the case of common diseases, sampling all cases in the traditional case-cohort design with multiple diseases limits applications (Breslow & Wellner, 2007). Instead, a generalized case-cohort design (Cai & Zeng, 2007) in which cases are sampled can be considered. Extending the proposed weights to this general case merits further investigation.

In our proposed estimation framework, time-dependent covariates can be allowed. However, estimation generally requires one to know the entire history of time-dependent covariates. In many follow-up studies, this may not be true. One commonly used approach for handling time-dependent covariates is to consider the last-value-carry-forward, but this could

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

introduce bias. A more sensible approach is to consider the joint modeling of survival times and longitudinal covariates via shared random effects, which has not been studied for case-cohort data.

When studying multiple diseases, different diseases may be competing risks for the same subject. In a competing risks situation, a subject can only experience at most one event; in the situation we considered, a subject can still experience the other events. Consequently, in the competing risks situation, a subject is at risk for all types of events simultaneously and will not be at risk for any other events as soon as one event occurs. Our approach in this paper can be adapted to competing risks by modifying the at-risk process and the weight function, but analysis will be based on the cause-specific hazards as studied in Sorensen & Andersen (2000).

The current method is based on estimating equations, which improves the estimation efficiency by incorporating a refined weight function for the risk set. However, it is not semiparametric efficient. To derive the most efficient estimator, we need to specify the joint distribution of the correlated failure times from the same subject and consider nonparametric maximum likelihood estimation based on the joint likelihood function for case-cohort sampling. This may be very challenging, especially when expensive covariates are continuous. This is an interesting topic which warrants future research.

Acknowledgments

We thank the editor, the associate editor, and two referees for the careful reading and the constructive comments which have led to great improvement of our manuscript. We thank Professor Matthew Knuiman and the Busselton Population Medical Research Foundation for permission to use their data. We also thank Professor Amy Herring and Forrest DeMarcus for their editorial assistance. This work was partially supported by grants from the National Institutes of Health.

References

Barlow W. Robust variance estimation for the case-cohort design. Biometrics. 1994; 50:1064–72. [PubMed: 7786988]

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. Lifetime Data Anal. 2000; 6:39–58. [PubMed: 10763560]

Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand J Statist. 2007; 34:86–102.

Cai J, Zeng D. Power calculation for case-cohort studies with nonrare events. Biometrics. 2007; 63:1288–95. [PubMed: 17608788]

Chen HY. Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. J Am Statist Assoc. 2001a; 96:1446–57.

Chen K. Generalized case-cohort sampling. J R Statist Soc B. 2001b; 63:791–809.

Chen K, LOS. Case-cohort and case-control analysis with Cox's model. Biometrika. 1999; 86:755–64.

Chen Y, Zucker DM. Case-cohort analysis with semiparametric transformation models. J Statist Plan Inf. 2009; 139:3706–17.

Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model. J R Statist Soc A. 1985; 148:82–117.

Cox DR. Regression models and life-tables (with discussion). J R Statist Soc B. 1972; 34:187–220.

Cox DR. Partial likelihood. Biometrika. 1975; 62:269–76.

Cullen KJ. Mass health examinations in the Busselton population, 1996 to 1970. Aust J Med. 1972; 2:714–8.

Fourtz RV. On the unique consistent solution to the likelihood equations. J Am Statist Assoc. 1977; 72:147–8.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Hájek J. Limiting distributions in simple random sampling from a finite population. Publ Math Inst Hungar Acad Sci. 1960; 5:361–74.

Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. Statist Med. 1988; 7:149–60.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2. New York: John Wiley; 2002.

Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. Biometrika. 2009; 96:887–901. [PubMed: 23946547]

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. Am J Epidemiol. 2003; 158:144–9. [PubMed: 12851227]

Kong L, Cai J, Sen PK. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. Biometrika. 2004; 91:305–19.

Kulich M, Lin DY. Additive hazards regression for case-cohort studies. Biometrika. 2000; 87:73–87.

Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Statist Assoc. 2004; 99:832–44.

Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison. Am J Epidemiol. 1990; 131:169–76. [PubMed: 2403467]

Lin DY. On fitting Cox's proportional hazards models to survey data. Biometrika. 2000; 87:37–47.

Lu W, Tsiatis AA. Semiparametric transformation models for the case-cohort study. Biometrika. 2006; 93:207–14.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11.

Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Statist. 1988; 16:64–81.

Sorensen P, Andersen 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. J Am Statist Assoc. 1998; 93:1164–75.

van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes. New York: Springer; 1996.

Wacholder S, Gail M, Pee D. Efficient design for assessing exposure-disease relationships in an assembled cohort. Biometrics. 1991; 47:63–76. [PubMed: 2049514]

Appendix. Outline of the Proofs of Theorems 1 – 2

Under the assumptions in Section 3, we outline the proofs for the main theorems. To prove the asymptotic properties for the proposed estimators, the following lemmas are used. The proof of Lemma 1 is in Lin (2000) and Lemma 2 is in Lemma A1 in Kang & Cai (2009).

Lemma 1

Let $\mathcal{H}_n(t)$ and $\mathcal{W}_n(t)$ be two sequences of bounded processes. If we assume that the following conditions (i), (ii), and (iii) hold for some constant τ, for which (i) $\sup_{0\le t\le\tau}\| \mathcal{H}_n(t) - \mathcal{H}(t) \| \to 0$ in probability for some bounded process $\mathcal{H}(t)$; (ii) $\mathcal{H}_n(t)$ is monotone on $[0,\tau]$; and (iii) $\mathcal{W}_n(t)$ converges to a zero-mean process with continuous sample paths, then

$\sup_{0\le t\le\tau}\| \int_0^t \{\mathcal{H}_n(s)-\mathcal{H}(s)\} d\mathcal{W}_n(s)\| \to 0$ in probability, and

$\sup_{0\le t\le\tau}\| \int_0^t \mathcal{W}_n(s) d\{\mathcal{H}_n(s)-\mathcal{H}(s)\}\| \to 0$ in probability.

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

Lemma 2

Let $B_i(t)$ ($i = 1,\ldots, n$) be independent and identically distributed real-valued random process on $[0, \tau]$, and denote random process vector, $B(t) = \{B_1(t),\ldots, B_n(t)\}$ with $E\{B_i(t)\} \equiv \mu_B(t)$, $\text{var}\{B_i(0)\} < \infty$, and $\text{var}\{B_i(\tau)\} < \infty$. Let $\xi = [\xi_1,\ldots,\xi_n]$ be random vector containing $\tilde{n}$ ones and $n - \tilde{n}$ zeros with each permutation equally likely. Let ξ be independent of $B(t)$. Suppose that almost all paths of $B_i(t)$ have finite variation. Then, $n^{-1/2}\sum_{i=1}^{n}\xi_i\{B_i(t)-\mu_B(t)\}$ converges weakly in $l^\infty[0, \tau]$ to a zero-mean Gaussian process, and $n^{-1}\sum_{i=1}^{n}\xi_i\{B_i(t)-\mu_B(t)\}$ converges in probability to zero uniformly in t.

Proof of Theorem 1

First, the proof of consistency of $\beta^{\tilde{M}}$ can be shown by the extension of Fourtz (1977): (I) $\partial\tilde{U}_n^M(\beta)/\partial\beta^T$ exists and is continuous in an open neighborhood $\mathcal{B}$ of β_0; (II) $\partial\tilde{U}_n^M(\beta)/\partial\beta^T$ is negative definite with probability going to one as n $\to \infty$; (III) $-\partial\tilde{U}_n^M(\beta)/\partial\beta^T$ converges to $A(\beta_0)$ in probability uniformly for β in an open neighborhood about β_0; (IV) $\tilde{U}_n^M(\beta)$ converges to 0 in probability, where $\tilde{U}_n^M = n^{-1}\tilde{U}^M$. Clearly, (I) is satisfied. If we show that $\|\{-\partial\tilde{U}_n^M(\beta)/\partial\beta^T\}-A(\beta)\|$ converges to zero in probability uniformly in $\beta\in\mathcal{B}$ as $n \to \infty$, then (II) and (III) are satisfied. We have

$$\|\{-\partial\tilde{U}_n^M(\beta)/\partial\beta^T\}-A(\beta)\| \le \|\sum_{k=1}^{K}\int_0^\tau\{\tilde{V}_k(\beta,t)-v_k(\beta,t)\}n^{-1}d\sum_{i=1}^{n}N_{ik}(t)\|$$
$$+\|\sum_{k=1}^{K}\int_0^\tau v_k(\beta,t)n^{-1}d\sum_{i=1}^{n}M_{ik}(t)\|$$
$$+\|\sum_{k=1}^{K}\int_0^\tau v_k(\beta,t)\{S_k^{(0)}(\beta,t)-s_k^{(0)}(\beta,t)\}\lambda_{0k}(t)dt\|$$

. Each of the three parts converges to zero in probability by Lemma 2, the Lenglart inequality, and conditions (d), (e), (f), and (g). Convergence of $\tilde{U}_n^M(\beta)$ to zero in probability shows that (IV) is satisfied. Therefore, $\beta^{\tilde{M}}$ converges to β_0 in probability and is a consistent estimator of β_0.

To establish the asymptotic normality of $n^{-1/2}\tilde{U}_n^M(\beta)$, we decompose it into two parts:

$$n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau\{Z_{ik}(u)-S_k^{(1)}(\beta,t)/S_k^{(0)}(\beta,t)\}dN_{ik}(t)$$
$$+n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau\{S_k^{(1)}(\beta,t)/S_k^{(0)}(\beta,t)-\tilde{S}_k^{(1)}(\beta,t)/\tilde{S}_k^{(0)}(\beta,t)\}dN_{ik}(t).$$

The first term is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}W_{ik}(\beta_0)$ by Spiekerman & Lin (1998). The second term can be decomposed into two parts

$$\sum_{k=1}^{K}\int_0^\tau D_k(\beta,t)d\{n^{-1/2}\sum_{i=1}^{n}M_{ik}(t)\}+n^{-1/2}\sum_{k=1}^{K}\int_0^\tau D_k(\beta,t)\{\sum_{i=1}^{n}Y_{ik}(t)e^{\beta_0 Z_{ik}(t)}d\Lambda_{0k}(t)\}$$

, where $D_k(\beta,t)=\{S_k^{(1)}(\beta,t)/S_k^{(0)}(\beta,t)-\tilde{S}_k^{(1)}(\beta,t)/\tilde{S}_k^{(0)}(\beta,t)\}$. The first term converges in probability uniformly in t to zero by van der Vaart & Wellner (1996), the Kolmogorov–Centsov Theorem, conditions (c), (d), and (f), and Lemma 1. The second term is asymptotically equivalent to

$$n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^\tau(1-\xi_i\tilde{\alpha}^{-1})\prod_{j=1}^{K}(1-\Delta_{ij})\,(Q_{ik}(\beta,t)$$
$$-Y_{ik}(t)E\{\prod_{j=1}^{K}(1-\Delta_{1j})Q_{1k}(\beta,t)\}[E\{\prod_{j=1}^{K}(1-\Delta_{1j})Y_{1k}(t)\}]^{-1})d\Lambda_{0k}(t)$$

by Lemma 1. Hence, $n^{-1/2}\tilde{U}_n^M(\beta)$ is asymptotically equivalent to

$$n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}W_{ik}(\beta_0)+n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau(1-\xi_i\tilde{\alpha}^{-1})\Omega_{ik}(\beta_0,t)d\Lambda_{0k}(t).$$

By Spiekerman & Lin (1998), the first term converges weakly to a zero-mean normal vector

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

with covariance matrix $V_I(\beta_0)=E\{\sum_{k=1}^K W_{1k}(\beta_0)\}^{\otimes 2}$. The second term is asymptotically a zero-mean normal vector with covariance matrix $\{1-\alpha\}\alpha^{-1}V_{II}(\beta_0)=\{1-\alpha\}\alpha^{-1}E\{\sum_{k=1}^K\int_0^\tau\Omega_{ik}(\beta_0,t)d\Lambda_{0k}(t)\}^{\otimes 2}$ by Hájek (1960)'s central limit theorem for finite sampling. In addition, $n^{-1/2}\sum_{i=1}^n\sum_{k=1}^K W_{ik}(\beta_0)$ and $n^{-1/2}\sum_{i=1}^n\sum_{k=1}^K\int_0^\tau(1-\xi_i\tilde{\alpha}^{-1})\Omega_{ik}(\beta_0,t)d\Lambda_{0k}(t)$ are independent. Thus $n^{1/2}(\beta^{\tilde{M}}-\beta_0)$ converges weakly to a zero-mean normal vector with covariance matrix $A(\beta_0)^{-1}\Sigma(\beta_0)A(\beta_0)^{-1}$. This completes the proof of Theorem 1.

Proof of Theorem 2

We decompose $\tilde{\Lambda}_{0k}^M(\tilde{\beta}^M,t)-\Lambda_{0k}(t)$ as

$$\begin{aligned}&n^{1/2}\int_0^t\left\{\frac{1}{n\tilde{S}_k^{(0)}(\tilde{\beta}^M,u)}-\frac{1}{n\tilde{S}_k^{(0)}(\beta_0,u)}\right\}d\sum_{i=1}^n M_{ik}(u)\\&+n^{1/2}\int_0^t\left\{\frac{1}{\tilde{S}_k^{(0)}(\tilde{\beta}^M,u)}-\frac{1}{\tilde{S}_k^{(0)}(\beta_0,u)}\right\}S_k^{(0)}(\beta_0,u)d\Lambda_{0k}(u)\\&+n^{-1/2}\int_0^t\frac{1}{\tilde{S}_k^{(0)}(\beta_0,u)}d\sum_{i=1}^n M_{ik}(u)+n^{1/2}\int_0^t\left\{\frac{S_k^{(0)}(\beta_0,u)-\tilde{S}_k^{(0)}(\beta_0,u)}{\tilde{S}_k^{(0)}(\beta_0,u)}\right\}d\Lambda_{0k}(u).\end{aligned}\tag{A1}$$

The first term here converges to zero in probability uniformly in t by Taylor expansion and Lemma 1. The second term can be written as $n^{1/2}l_k(\beta,t)^T(\beta^{\tilde{M}}-\beta_0)+o_p(1)$, where $l_k(\beta,t)^T=\int_0^t\{-e_k(\beta,u)\}\,d\Lambda_{0k}(u)$ by Taylor expansion, uniform convergence of $\tilde{S}_k^{(d)}(\beta^*,u)$ and $\tilde{S}_k^{(0)}(\beta_0,u)$, d=0,1, and boundedness of $d\Lambda_{0k}(u)$, where β^* is on the line segment between $\beta^{\tilde{M}}$ and β_0. Because $\tilde{S}_k^{(0)}(\beta_0,u)^{-1}$ can be written as a sum of two monotone functions in t and converges uniformly to $s_k^{(0)}(\beta_0,u)^{-1}$, in which $s_k^{(0)}(\beta_0,u)$ is bounded away from 0, and $n^{-1/2}d\sum_{i=1}^n M_{ik}(u)$ converges to a zero-mean Gaussian process with continuous sample path, the third term in (A1) can be written as $\int_0^t\{s_k^{(0)}(\beta_0,u)\}^{-1}\{n^{-1/2}\,d\sum_{i=1}^n M_{ik}(u)\}+o_p(1)$. Due to the uniform convergence of $\tilde{S}_k^{(0)}(\beta_0,u)^{-1}$ to $s_k^{(0)}(\beta_0,u)^{-1}$, where $s_k^{(0)}(\beta_0,u)$ is bounded away from 0, the last term in (A1) is asymptotically equivalent to

$$n^{-1/2}\sum_{i=1}^n(1-\xi_i\tilde{\alpha}^{-1})\prod_{j=1}^K(1-\Delta_{ij})$$

$$\int_0^t Y_{ik}(u)[e^{\beta^T Z_{ik}(u)}-E\{\prod_{j=1}^K(1-\Delta_{1j})e^{\beta^T Z_{1k}(u)}Y_{1k}(u)\}$$

$$[E\{\prod_{j=1}^K(1-\Delta_{1j})Y_{1k}(u)\}]^{-1}]\;d\Lambda_{0k}(u)/s_k^{(0)}(\beta_0,u)+o_p(1).$$

Using a decomposition of $n^{1/2}(\beta^{\tilde{M}}-\beta_0)$, we have

$$n^{1/2}\{\tilde{\Lambda}_{0k}^M(\tilde{\beta}^M,t)-\Lambda_{0k}(t)\}=n^{-1/2}\sum_{i=1}^n\eta_{ik}(\beta_0,t)+n^{-1/2}\sum_{i=1}^n(1-\xi_i\tilde{\alpha}^{-1})\zeta_{ik}(\beta_0,t)+o_p(1).$$

Let $H(t)=\{H^{(1)}(t)+H^{(2)}(t)\}$, where

$H^{(a)}(t)=\{H_1^{(a)}(t),\ldots,H_K^{(a)}(t)\}^T$, $a=1,2$, $H_k^{(1)}(t)=n^{-1/2}\sum_{i=1}^n\eta_{ik}(\beta_0,t)$, and $H_k^{(2)}(t)=n^{-1/2}\sum_{i=1}^n(1-\xi_i\tilde{\alpha}^{-1})\zeta_{ik}(\beta_0,t)$. By Spiekerman & Lin (1998), $H^{(1)}(t)=\{H_1^{(1)}(t),\ldots,H_K^{(1)}(t)\}^T$ converges weakly to a Gaussian process

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$\mathscr{H}^{(1)}(t)=(\mathscr{H}_1^{(1)}(t),\ldots,\mathscr{H}_K^{(1)}(t))^T$ whose mean is zero and covariance function between $\mathscr{H}_j^{(1)}(t)$ and $\mathscr{H}_k^{(1)}(s)$ is $E\{\eta_{1j}(\beta_0, t),\eta_{1k}(\beta_0, s)\}$ for $t, s \in [0,\tau]$ in $D[0,\tau]^K$. By Lemma 1, Lemma 2, boundedness conditions, and the Cramer–Wold device, it can be shown that $H^{(2)}(t)=\{H_1^{(2)}(t),\ldots,H_K^{(2)}(t)\}^T$ converges weakly to a Gaussian process $\mathscr{H}^{(2)}(t)=\{\mathscr{H}_1^{(2)}(t),\ldots,\mathscr{H}_K^{(2)}(t)\}^T$ whose mean is zero and covariance function between $\mathscr{H}_j^{(2)}(t)$ and $\mathscr{H}_k^{(2)}(s)$ is $\{1-a\}a^{-1}E\{\zeta_{1j}(\beta_0, t),\zeta_{1k}(\beta_0, s)\}$ for $t, s \in [0,\tau]$ in $D[0,\tau]^K$. It can easily be shown that $H^{(1)}(t)$ and $H^{(2)}(s)$ are independent. Therefore the conclusion in Theorem 2 holds. This completes the proof of Theorem 2.

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

Table 1

Comparison of two methods with a single disease outcome: $\beta_1 = \log 2 = 0{\cdot}693$

			The proposed method				Borgan et al.'s method				
Event proportion	**Size of subcohort**	τ_θ	$\tilde{\beta_1}$	**SE**	**SD**	**CR**	$\hat{\beta_1}$	**SE**	**SD**	**CR**	**SRE**
8%	100	0·83	0·706	0·32	0·32	94	0·705	0·33	0·33	94	1·04
		0·11	0·718	0·31	0·32	94	0·719	0·33	0·33	94	1·07
		0·05	0·708	0·32	0·32	94	0·705	0·33	0·33	94	1·06
	200	0·83	0·715	0·28	0·28	95	0·716	0·28	0·28	95	1·02
		0·11	0·704	0·28	0·28	95	0·705	0·28	0·29	95	1·03
		0·05	0·697	0·28	0·27	95	0·698	0·28	0·28	95	1·05
20%	100	0·83	0·703	0·25	0·25	94	0·704	0·26	0·27	95	1·13
		0·11	0·694	0·23	0·23	94	0·694	0·26	0·27	95	1·31
		0·05	0·700	0·23	0·23	94	0·701	0·26	0·26	95	1·29
	200	0·83	0·693	0·20	0·20	95	0·692	0·21	0·21	95	1·10
		0·11	0·696	0·19	0·19	95	0·699	0·21	0·21	95	1·17
		0·05	0·694	0·19	0·19	95	0·695	0·21	0·21	95	1·26

SE, average standard errors; SD, sample standard deviation; CR, coverage rate (%) of the nominal 95% confidence intervals; $\mathrm{SRE}=\mathrm{SD}_c^2/\mathrm{SD}_p^2$, sample relative efficiency, where SD_c and SD_p are the sample standard deviation for the Borgan et al. (2000)'s method and the proposed method, respectively.

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

Table 2

Comparison of two methods with multiple disease outcomes: $[\beta_1, \beta_2] = [0{\cdot}1, 0{\cdot}7]$

			The proposed method				Kang & Cai's method				SRE
Event proportion	**Size of subcohort**	τ_θ	$\tilde{\beta}_1^M$	**SE**	**SD**	**CR**	$\hat{\beta}_1^M$	**SE**	**SD**	**CR**	$\hat{\beta_1}$
[8%, 14%]	100	0·83	0·099	0·31	0·30	95	0·101	0·32	0·31	95	1·07
		0·11	0·101	0·30	0·30	95	0·098	0·32	0·32	95	1·13
		0·05	0·109	0·30	0·31	94	0·111	0·32	0·33	94	1·11
	200	0·83	0·106	0·26	0·27	95	0·105	0·27	0·27	95	1·04
		0·11	0·096	0·26	0·26	94	0·096	0·27	0·27	94	1·05
		0·05	0·098	0·26	0·27	94	0·098	0·27	0·27	94	1·05
[20%, 35%]	100	0·83	0·098	0·23	0·24	94	0·094	0·26	0·27	94	1·24
		0·11	0·099	0·22	0·22	94	0·097	0·26	0·26	95	1·42
		0·05	0·095	0·22	0·22	94	0·101	0·26	0·27	95	1·44
	200	0·83	0·103	0·19	0·19	94	0·104	0·20	0·21	95	1·19
		0·11	0·098	0·18	0·18	95	0·097	0·20	0·20	95	1·29
		0·05	0·098	0·18	0·18	95	0·100	0·20	0·20	96	1·31

SE, average standard errors; SD, sample standard deviation; CR, coverage rate (%) of the nominal 95% confidence intervals; $\mathrm{SRE}=\mathrm{SD}_e^2/\mathrm{SD}_p^2$, sample relative efficiency, where SD_e and SD_p are the sample standard deviation for the Kang & Cai (2009)'s method and the proposed method, respectively.

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

Table 3

Comparison between separate and joint analysis: $\beta_1 = \log 2$ with event proportion 20%

		Separate analysis					
		The proposed weight			Borgan at al.'s method		
Size of subcohort	τ_θ	$\tilde{\beta}_1$	SE	SD	$\hat{\beta}_1$	SE	SD
100	0·83	0·713	0·244	0·245	0·716	0·263	0·265
	0·11	0·702	0·226	0·236	0·705	0·262	0·270
	0·05	0·700	0·226	0·232	0·710	0·263	0·268
200	0·83	0·703	0·196	0·194	0·704	0·206	0·206
	0·11	0·697	0·186	0·193	0·699	0·205	0·213
	0·05	0·698	0·186	0·187	0·702	0·206	0·209

		Joint analysis					
		The proposed weight			Kang and Cai's method		
Size of subcohort	τ_θ	$\tilde{\beta}_1^M$	SE	SD	$\hat{\beta}_1^M$	SE	SD
100	0·83	0·711	0·243	0·245	0·713	0·262	0·264
	0·11	0·701	0·226	0·235	0·701	0·261	0·267
	0·05	0·700	0·225	0·231	0·707	0·262	0·266
200	0·83	0·703	0·195	0·194	0·703	0·205	0·205
	0·11	0·696	0·186	0·193	0·697	0·205	0·212
	0·05	0·698	0·186	0·187	0·700	0·205	0·209

SE, average standard errors; SD, sample standard deviation.

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

Table 4

Type I error and power (%) in separate and joint analyses with event proportion 20%

		Type I error ($\beta_1 = \beta_2 = \log 2$)				Power ($\beta_1 = 0{\cdot}1, \beta_2 = 0{\cdot}7$)			
		Separate analysis		Joint analysis		Separate analysis		Joint analysis	
Size of subcohort	τ_θ	P	BR	P	KC	P	BR	P	KC
100	0·83	0·6	0·6	6·3	6·7	49	42	90	78
	0·11	0·8	1·7	5·9	5·9	56	42	83	61
	0·05	1·2	2·1	5·1	5·6	59	43	81	61
200	0·83	0·2	0·3	5·2	5·8	80	72	98	94
	0·11	1·6	1·9	5·4	5·4	77	65	89	78
	0·05	1·8	2·5	5·3	5·4	79	68	90	79

P, the proposed weight; BR, the method of Borgan et al. (2000); KC, the method of Kang & Cai (2009).

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

Table 5

Analysis results for the Busselton Health Study

	Proposed method				Kang & Cai method			
Variables	$\tilde{\beta}_M$	**SE**	**HR**	**95% CI**	$\hat{\beta}_M$	**SE**	**HR**	**95% CI**
log(ferritin) on CHD	0·145	0·0897	1·16	(0·97, 1·38)	0·092	0·0949	1·10	(0·91, 1·32)
log(ferritin) on Stroke	0·172	0·1219	1·19	(0·93, 1·51)	0·186	0·1304	1·20	(0·93, 1·56)
Age	0·071	0·0069	1·07	(1·06, 1·09)	0·069	0·0070	1·07	(1·06, 1·09)
Triglycerides	0·239	0·0484	1·27	(1·16, 1·40)	0·232	0·0541	1·26	(1·13, 1·40)
Blood pressure treatment	0·423	0·1633	1·53	(1·11, 2·10)	0·408	0·1727	1·50	(1·07, 2·11)

CHD, coronary heart disease; SE, standard error; HR, hazard ratio; CI, confidence interval.