Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1250–1260. doi:10.1111/biom.12923.

Analysis of multiple survival events in generalized case-cohort designs

Soyoung Kim¹, Donglin Zeng², and Jianwen Cai^{2,*}

¹Division of Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, U.S.A

²Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A

Summary

Generalized case-cohort design has been proposed to assess the effects of exposures on survival outcomes when measuring exposures is expensive and events are not rare in the cohort. In such design, expensive exposure information is collected from both a (stratified) randomly selected subcohort and a subset of individuals with events. In this paper, we consider extension of such design to study multiple types of survival events by selecting a proportion of cases for each type of event. We propose a general weighting scheme to analyze data. Furthermore, we examine the optimal choice of weights and show that this optimal weighting yields much improved efficiency gain both asymptotically and in simulation studies. Finally, we apply our proposed methods to data from the Atherosclerosis Risk in Communities study.

Keywords

Case-cohort study; Multiple events; Multiple disease outcomes; Non-rare diseases; Proportional hazards; Stratified sampling; Survival analysis

1. Introduction

Case-cohort study design is an economical means for large cohort studies with rare survival events when it is expensive to assemble covariate information for all cohort members (Prentice, 1986). In such design, a random sample from the full cohort, namely subcohort, is selected via simple random sampling, then all subjects having events of interest outside this subcohort are sampled. The covariate information on the expensive exposure is obtained for the subcohort members as well as all sampled cases.

Extensive work has been done for the case-cohort studies with a single event. Prentice (1986) and Self and Prentice (1988) proposed a pseudo-likelihood approach for inference. In order to improve efficiency, Barlow (1994) developed a robust estimator using a time-varying weight. Later, Borgan et al. (2000) considered the subcohort selected via a stratified

* cai@bios.unc.edu.

7. Supplementary Materials

Web Appendix, referenced in Section 3, is available with this paper at the *Biometrics* website on Wiley Online Library.

random sampling and showed that the stratification leads to more powerful and efficient estimators than the unstratified case-cohort study. Kulich and Lin (2004) and Samuelsen et al. (2007) proposed efficient estimation for a stratified case-cohort design by using auxiliary covariate data.

In many applications, the same subject can experience multiple types of survival events. When these survival outcomes are all of interest, the case-cohort design has also been recommended to study the effects of risk factors on multiple diseases simultaneously, where the information on expensive exposures from a subcohort and the cases of all event types is collected. Using data from this design, Kang and Cai (2009; 2010) developed estimation procedures based on the joint analysis in the unstratified and stratified case-cohort studies, respectively. However, when one particular event is of interest, their methods did not use all available exposure information collected on the cases of the other types of events. More recently, Kim et al. (2013) proposed estimating equations with a new weight function to incorporate this information in order to improve efficiency for estimation.

All the aforementioned methods considered the classical case-cohort study design, which samples all event cases for exposure assessment. However, in many cohort studies, the number of cases can be large, because the event is relatively common, the cohort size is large, or the follow-up duration is long. For example, in the Atherosclerosis Risk in Communities (ARIC) study (Duncan et al., 2003; Ballantyne et al., 2004), 15,792 subjects were recruited from 1987 to 1989 and followed up since then. It was of interest to examine the effect of high-sensitivity C-reactive protein (hs-CRP) on incident diabetes events. In the ARIC study, the rate of diabetes is 11.2%, resulting in a large number of cases. Since measuring hs-CRP from blood sample was expensive at the time, it was not feasible to measure the expensive covariates from all cases due to limited resources.

When there are a large number of cases, instead of collecting exposure information from all cases, a generalized case-cohort design was proposed where only a fraction of the non-subcohort cases were sampled for exposure assessment. Cai and Zeng (2007) provided sample size and power calculation for this generalized case-cohort design. They demonstrated that when the event was not rare, such a design could perform as well as a classical case-cohort design even if a small fraction of the cases were sampled. Kim et al. (2016) extended Kim et al. (2013)'s classical case-cohort design to generalized case-cohort design for additive hazard models but they considered only two disease. In this paper, we extend the idea of the generalized case-cohort design to study multiple survival events. Specifically, in addition to a randomly chosen subcohort, a subsample of each type of event cases is selected to assemble expensive covariate information. The sampling fractions may differ for different event types. Furthermore, we allow stratified sampling in this design which is typical in biomedical research. The strata are usually formed based on participants' characteristics at baseline and sampling probabilities may vary across different strata in order to oversample low-prevalence subpopulations for study purpose. We then develop an efficient approach to analyze data arising from such design. Particularly, we propose a general weighting scheme to account for the fact that only fractions of the cases are sampled in this generalized case-cohort design. The proposed general weighting includes the weights in Kim et al. (2013) as a special case.

The paper is organized as following. In Section 2, we describe models, estimation procedures, and their asymptotic properties for the proposed methods. Section 3 provides optimally weighted estimators and Section 4 reports simulation results. In Section 5, we apply our proposed method to data from the Atherosclerosis Risk in Communities (ARIC) study. Some concluding remarks are given in Section 6.

2. Generalized Case-Cohort Design for Multiple Events

Suppose that there are n independent subjects and K survival endpoints of interest in a cohort. In order to ensure proper representation of certain subgroups in the sampling of the subcohort, the entire cohort can be divided into mutually exclusive strata. These strata are usually defined by participants's baseline characteristics. Assume that there are L strata. Let T_{lik} be the failure time, C_{lik} be the potential censoring time, and $Z_{lik}(t)$ be a $p \times 1$ possibly time-dependent covariates vector for disease k of subject i in stratum l , $l = 1, \dots, L$, $k = 1, \dots, K$, $i = 1, \dots, n_l$ where n_l is defined as the number of subjects in stratum l . Let $X_{lik} = \min(T_{lik}, C_{lik})$ denote the observed time of type k in the full cohort and $I_{lik} = \mathbb{I}(T_{lik} < C_{lik})$ be the indicator for event k . We use V_{lik} to denote the stratum that the participant belongs to. In order to study the effects of covariates on each type of events, we consider the event-specific hazards model: for disease k of subject i in stratum l , the hazard function $\lambda_{lik}(\cdot)$ associated with $Z_{lik}(t)$ is assumed to be

$$\lambda_{lik}\{t|Z_{lik}(t)\} = \lambda_{0k}(t)e^{\beta_k^T Z_{lik}(t)}, \quad (1)$$

where $\lambda_{0k}(t)$ is a baseline hazard function and β_k is a p -vector unknown parameter for disease k . Note that V_{lik} can be part of Z_{lik} if it is of interest to adjust for the sampling strata for the exposure effect. Finally, we assume that T_{lik} is independent of C_{lik} given Z_{lik} .

2.1 Generalized case-cohort design

In generalized case-cohort design, we select a fixed size $\tilde{n}_l$ subjects from n_l subjects in stratum l into the subcohort by using simple random sampling without replacement. After sampling the subcohort, another stratified random samples of cases outside of the subcohort for each disease outcome are drawn. For disease k in stratum l , we select $\tilde{m}_{lk}$ cases outside of the subcohort using simple random sampling without replacement. Let ξ_{li} indicate whether subject i in stratum l is selected into the subcohort and η_{lik} be the sampling indicator of selecting case of type k outside the subcohort in stratum l . Note that for $k \neq k'$, $(\eta_{1lk}, \dots, \eta_{l n_l k})$ is independent of $(\eta_{1lk'}, \dots, \eta_{l n_l k'})$ conditional on disease status. But the elements in $(\eta_{1lk}, \dots, \eta_{l n_l k})$ are correlated because of the sampling scheme.

Let $Z_{lik}(t) = \{Z_{lik}^E(t), Z_{lik}^C(t)\}$ where $Z_{lik}^E(t)$ represents the expensive covariates that are only available on subjects who are in the case-cohort sample, while $Z_{lik}^C(t)$ denotes the covariates information that are available on the entire sample, for example, age and sex. In the

generalized case-cohort design, the actual data for subject i consist of $(X_{lik}, \Delta_{lik}, Z_{lik}^E(t), Z_{lik}^C(t), 0 \leq t \leq X_{lik})$ when $\xi_{li} = 1$ or $\eta_{lik} = 1$ and $(X_{lik}, \Delta_{lik}, Z_{lik}^C(t))$ when $\xi_{li} = 0$ and $\eta_{lik} = 0$ ($k = 1, \dots, K$). Let τ denote the end of study time.

2.2 A class of weighted estimating equations

Let $N_{lik}(t) = I(X_{lik} \geq t, \eta_{lik} = 1)$ be the counting process for the observed failure time and $Y_{lik}(t) = I(X_{lik} \geq t)$ denote the at-risk indicator for disease k of subject i in stratum l , where $I(\cdot)$ is the indicator function. Let $n = \sum_{l=1}^L n_l$ be the total size of the cohort, $n_l = \sum_{i=1}^{n_l} \tilde{n}_l$ be the total size of the subcohort, d_{lk} and $\tilde{d}_{lk}$ denote the numbers of subjects with disease k in the cohort and in the subcohort in stratum l , respectively. Then $\Pr(\xi_{li} = 1) = \tilde{n}_l/n_l = \tilde{\alpha}_l$ and $\Pr(\eta_{lik} = 1 | \Delta_{lik} = 1, \xi_{li} = 0) = \tilde{m}_{lk}/(d_{lk} - \tilde{d}_{lk})$, denoted by $\tilde{\gamma}_{lk}$. The first probability is the selection probability of subjects for the subcohort and the second probability is the selection probability of subjects outside the subcohort with disease k in stratum l .

When exposure information is available for all subjects, estimating function based on the pseudo-likelihood in Prentice (1986) and Self and Prentice (1988) is given by

$$U_k(\beta) = \sum_{l=1}^L \sum_{i=1}^{n_l} \int_0^T \left\{ Z_{lik}(t) - \frac{S_k^{(1)}(\beta, t)}{S_k^{(0)}(\beta, t)} \right\} dN_{lik}(t),$$

where $S_k^{(d)}(\beta, t) = n^{-1} \sum_{l=1}^L \sum_{i=1}^{n_l} Y_{lik}(t) Z_{lik}(t)^{\otimes d} e^{\beta_k^T Z_{lik}(t)}$ for $d = 0, 1$ and 2 . Under generalized case-cohort design, the expensive exposure information is available only for subjects in the subcohort as well as sampled subjects with diseases of interest. Therefore, to use the data from this design for inference, our key idea is to use the subjects with available expensive exposure information to approximate each component on the right-hand side of $U_k(\beta)$. Specifically, we propose a class of weighted estimating functions as follows:

$$\tilde{U}_k^O(\beta) = \sum_{l=1}^L \sum_{i=1}^{n_l} \int_0^\tau \pi_{lik}(t) \left\{ Z_{lik}(t) - \frac{\tilde{S}_k^{(1)}(\beta, t)}{\tilde{S}_k^{(0)}(\beta, t)} \right\} dN_{lik}(t), \quad (2)$$

where $\tilde{S}_k^{(d)}(\beta, t) = n^{-1} \sum_{l=1}^L \sum_{i=1}^{n_l} \pi_{lik}(t) Y_{lik}(t) Z_{lik}(t)^{\otimes d} e^{\beta_k^T Z_{lik}(t)}$ for $d = 0, 1$ and 2 . Here, $\pi_{lik}(t)$ is a non-negative weight function that depends on ξ_{li} and η_{lik} 's such that $\pi_{lik}(t) = 0$ if $\xi_{li} = 0$ and $\eta_{lik} = 0$ for any k (i.e. subject i 's expensive exposure information is not available, and $E[\pi_{lik}(t)] = 1$. For any such weight π_{lik} 's, we solve $\tilde{U}_k^O(\beta) = 0$ and denote its solution as $\hat{\beta}_k$. Additionally, with the estimators for β_k 's, we can estimate the cumulative baseline hazard functions using the Breslow-Aalen type estimators given by

$$\tilde{\Lambda}_k(\tilde{\beta}, t) = \int_0^T \frac{\sum_{l=1}^L \sum_{i=1}^{n_l} \pi_{lik}(u) dN_{lik}(u)}{n\tilde{S}_k^{(0)}(\tilde{\beta}, u)}.$$

To construct the weight function $\pi_{lik}(t)$, we partition the whole cohort into disjoint parts, where each part consists of subjects who experience some events but not the others within each stratum, i.e.,

$$\mathcal{E}_{lv} \equiv \left\{ i: \prod_{w \in D_{lv}} \Delta_{liw} \prod_{q \in \bar{D}_{lv}} (1 - \Delta_{liq}) = 1 \right\}, l = 1, \dots, L, \text{ and } v = 1, \dots, 2^K - 1$$

where D_{lv} is the v -th nonempty subset of $S = \{1, \dots, K\}$ and $\bar{D}_{lv}$ is the complementary set of D_{lv} . We also use $\mathcal{E}_{l0}$ to denote the set of subjects with no event, i.e.,

$\mathcal{E}_{l0} = \left\{ i: \prod_{j=1}^K (1 - \Delta_{lij}) = 1 \right\}$. Thus, in the generalized case-cohort design, subject i in $\mathcal{E}_{l0}$ can only be selected if the subject is in the subcohort ($\xi_{li} = 1$). For subject i in $\mathcal{E}_{lv}$ for $v \geq 1$, the subject can be selected either because the subject is in the subcohort ($\xi_{li} = 1$) or because the subject is selected in the cases outside the subcohort ($\xi_{li} = 0$ but some $\eta_{lij} = 1$ where j indicates an event in $\mathcal{E}_{lv}$). Note that for the latter, subject i may be selected due to more than one event. Our proposed method is to assign different weights to subjects in each such possibility. Specifically, our proposed weight function $\pi_{lik}(t)$ takes the following form

$$= \pi_{lik}(t) \quad (3)$$

$$\begin{aligned} & I(i \in \mathcal{E}_{l0}) \xi_{li} \tilde{a}_{0lk}(t) + \sum_{v=1}^{2^K-1} I(i \in \mathcal{E}_{lv}) \xi_{li} \tilde{a}_{v lk}(t) \\ & + \sum_{v=1}^{2^K-1} I(i \in \mathcal{E}_{lv}) (1 - \xi_{li}) \left[\sum_{D \subset D_{lv}, D \neq \emptyset} \left\{ \left(\prod_{j \in D} \eta_{lij} \right) \left(\prod_{j \in \bar{D}_{lv}/D} (1 - \eta_{lij}) \right) \right\} \tilde{b}_{D, lk}(t) \right]. \end{aligned}$$

where the last summation sums over all nonempty subset of D_{lv} , D_{lv}/D denotes the set of indices in D_{lv} but not in D , and $\tilde{a}_{0lk}(t)$, $\tilde{a}_{v lk}(t)$, and $\tilde{b}_{D, lk}(t)$ are chosen to ensure $E[\pi_{lik}(t)] = 1$, for instance, the inverse probabilities of being sampled in each partitioned set.

To better illustrate the proposed approach, we use $K = 2$ as one example. Suppose there are two diseases of interest: diabetes and coronary heart disease (CHD). We can decompose the whole cohort into four groups within each stratum: Subjects a) with no disease, b) with only diabetes, c) with only CHD, and d) with both diabetes and CHD. Within each group that has at least one event, subjects are further divided into two subgroups: 1) those cases in the subcohort and 2) those cases who are outside the subcohort. These case subgroups and the

group with no events form the 9 partitioned disjoint sets. In this situation, the proposed weight is

$$\begin{aligned} \pi_{lik}(t) = & (1 - \Delta_{li1})(1 - \Delta_{li2})\xi_{li}\tilde{a}_{0lk}(t) \\ & + \Delta_{li1}(1 - \Delta_{li2})\left\{\xi_{li}\tilde{a}_{1lk}(t) + (1 - \xi_{li})\eta_{li1}\tilde{b}_{1lk}(t)\right\} \\ & + (1 - \Delta_{li1})\Delta_{li2}\left\{\xi_{li}\tilde{a}_{2lk}(t) + (1 - \xi_{li})\eta_{li2}\tilde{b}_{2lk}(t)\right\} \\ & + \Delta_{li1}\Delta_{li2}\left[\xi_{li}\tilde{a}_{3lk}(t) + (1 - \xi_{li})\left\{\eta_{li1}\tilde{b}_{3lk}(t) + \eta_{li2}\tilde{b}_{4lk}(t) + \eta_{li1}\eta_{li2}\tilde{b}_{5lk}(t)\right\}\right], \end{aligned} \quad (4)$$

where without confusion, we re-index $\tilde{b}_{D, lk}(t)$ as $\tilde{b}_{1lk}(t)$ to $\tilde{b}_{5lk}(t)$. Figure 1 illustrates all these partitions and the corresponding weights.

Note that the disjoint parts are defined within each stratum in order to calculate the proper weights. The strata are disjoint, so if two subjects belong to two strata, they will be in two separate disjoint parts.

Remark 1—The weights in (3) can be time-independent or time-varying. Prentice (1986) originally proposed constant weights. To improve efficiency, time-varying weights have been proposed by considering only subjects at risk at time t , not all subjects in the original cohort (Barlow, 1994; Borgan et al., 2000). The proportion of those at risk in the subcohort out of all those at risk in the entire cohort could be different at different time point. A time-varying weight function is more general than a time-constant weight function and it is shown that it produces better estimator (e.g. (Borgan et al., 2000)).

Remark 2—Our proposed method is equivalent to viewing type k' cases as non-cases when considering failure type k . However, even for those type k' cases, the probabilities of being selected for collecting expensive exposure information can be different for different k' in a generalized case-cohort design. Therefore, different weight functions may be necessary for those “non-cases”. The proposed class of the general weighted functions guarantees consistent estimation once the weights satisfy the condition $E[\pi_{lik}(t)] = 1$, as shown in Theorem 1.

Remark 3—In the estimating function for a particular disease k , Kang and Cai (2010)’s weight function ignores the covariate information collected on subjects who have other types of diseases and only uses individuals in the subcohort plus those sampled individuals with disease k for their weight function. Kim et al. (2016) in addition uses individuals with the other type of disease in their weight function in the set up when two diseases are considered. The basic idea for Kim et al. (2016)’s weight function is to divide the cohort into various strata defined by the status of the two diseases of interest. Then Kim et al. (2016)’s weight function is calculated within each of these strata by the inverse of the proportion of those who are at risk and are sampled among those who are at risk. Note that Kim et al. (2016)’s weight functions used only covariate information on subjects with the other disease, not the information on the disease status. Both existing weight functions proposed by Kang and Cai (2010) and Kim et al. (2016) are special cases of our proposed weight function (3). In particular, for Kang and Cai (2010)’s method, the weights used for disease k correspond to

$$\begin{aligned}\tilde{a}_{0lk}(t) &= \tilde{a}_{1lk}(t) = \tilde{a}_{2lk}(t) = \tilde{a}_{3lk}(t) = \frac{\sum_{i=1}^{n_l} (1 - \Delta_{lik}) Y_{lik}(t)}{\sum_{i=1}^{n_l} \xi_{li} (1 - \Delta_{lik}) Y_{lik}(t)} \\ \tilde{b}_{1lk}(t) &= \tilde{b}_{3lk}(t) = \frac{\sum_{i=1}^{n_l} \Delta_{lik} (1 - \xi_{li}) Y_{lik}(t)}{\sum_{i=1}^{n_l} \Delta_{lik} (1 - \xi_{li}) \eta_{lik} Y_{lik}(t)} \\ \tilde{b}_{2lk}(t) &= \tilde{b}_{4lk}(t) = \tilde{b}_{5lk}(t) = 0,\end{aligned}$$

while the weights proposed by Kim et al. (2016) correspond to

$$\begin{aligned}\tilde{a}_{0lk}(t) &= \frac{\sum_{i=1}^n Y_{lik}(t) \prod_{j=1}^2 (1 - \Delta_{lij})}{\sum_{i=1}^n Y_{lik}(t) \prod_{j=1}^2 (1 - \Delta_{lij}) \xi_{li}}, \tilde{a}_{1lk}(t) = \tilde{a}_{2lk}(t) = \tilde{a}_{3lk}(t) = 1 \\ \tilde{b}_{1lk}(t) &= \frac{\sum_{i=1}^n Y_{lik}(t) \Delta_{li1} (1 - \Delta_{li2})}{\sum_{i=1}^n Y_{lik}(t) \Delta_{li1} (1 - \Delta_{li2}) \eta_{li1}}, \tilde{b}_{2lk}(t) = \frac{\sum_{i=1}^n Y_{lik}(t) (1 - \Delta_{li1}) \Delta_{li2}}{\sum_{i=1}^n Y_{lik}(t) (1 - \Delta_{li1}) \Delta_{li2} \eta_{li2}} \\ \tilde{b}_{3lk}(t) &= \tilde{b}_{4lk}(t) = -\tilde{b}_{5lk}(t) = \frac{\sum_{i=1}^n Y_{lik}(t) \Delta_{li1} \Delta_{li2}}{\sum_{i=1}^n Y_{lik}(t) \Delta_{li1} \Delta_{li2} (\eta_{li1} + \eta_{li2} - \eta_{li1} \eta_{li2})}.\end{aligned}$$

Furthermore, when all cases outside the subcohort are selected (i.e. $\eta_{li1} = \eta_{li2} = 1$), the weight functions in (3) reduce to $\phi_{lik}(t) = \prod_{j=1}^K (1 - \Delta_{lij}) \xi_{li} \tilde{\alpha}_{lk}^{-1}(t) + \left\{1 - \prod_{j=1}^K (1 - \Delta_{lij})\right\}$, which was proposed by Kim et al. (2013) for the traditional case-cohort design.

2.3 Asymptotic properties

In this section, we provide the asymptotic properties for the proposed method for the generalized case-cohort studies. Let

$$\begin{aligned}Q_{lik}(\beta) &= \int_0^\tau \{Z_{lik} - e_k(\beta, t)\} dM_{lik}(t), \\ f_{1lk, D_{lv}} &= E[\xi_{li} \tilde{\alpha}_{lk}(t) + (1 - \xi_{li}) \sum_{D \subset D_{lv}, D \neq \emptyset} \left\{ \left(\prod_{j \in D} \eta_{lij} \right) \left(\prod_{j \in D_{lv}/D} (1 - \eta_{lij}) \right) \right\} \tilde{b}_{D, lk}(t)], \\ R_{lik}(\beta, t) &= Y_{lik}(t) [Z_{lik}(t) - e_k(\beta, t)] e^{\beta^T Z_{lik}(t)} \\ \tilde{R}_{lik}(\beta, t) &= R_{lik}(\beta, t) - \frac{Y_{lik}(t) E_l \left\{ \prod_{j=1}^K (1 - \Delta_{i1j}) R_{1lk}(\beta, t) \right\}}{E_l \left\{ \prod_{j=1}^K (1 - \Delta_{i1j}) Y_{1lk}(t) \right\}} \\ \tilde{Q}_{lik, \mathcal{E}_{lv}}(\beta) &= Q_{lik}(\beta) - \int_0^\tau \frac{Y_{lik}(t) E_l \{dQ_{1lk}(\beta, t) | \mathcal{E}_{lv}\}}{E_l \{Y_{1lk}(t) | \mathcal{E}_{lv}\}}.\end{aligned}$$

Theorem 1—Under the regularity conditions in the Supplementary material (Web Appendix A) and assuming $n_I/n \rightarrow q_I$ and $\tilde{\alpha}_I \rightarrow \alpha_I$ for $I = 1, \dots, L$, $\hat{\beta}_k$ converges in probability to β_k and $n^{1/2}(\hat{\beta}_k - \beta_k)$ is asymptotically normally distributed with mean zero and with the covariance matrix $A_k(\beta_k)^{-1} \Sigma_k(\beta_k) A_k(\beta_k)^{-1}$, where $A_k(\beta_k) = E[\partial \tilde{U}_k^O(\beta_k) / \partial \beta_k]$, and

$$\begin{aligned} \Sigma_k(\beta) &= \sum_{l=1}^L q_l [V_{I,lk}(\beta) + \frac{1-\alpha_l}{\alpha_l} V_{II,lk}(\beta) + V_{III,lk}(\beta)], \\ V_{I,lk}(\beta) &= E[Q_{IIk}(\beta)] \otimes^2, \\ V_{II,lk}(\beta) &= \text{Var} \left[\prod_{j=1}^K (1 - \Delta_{Ij}) \int_0^\tau \tilde{R}_{lik}(\beta, t) d\Lambda_{0k}(t) \right], \\ V_{III,lk}(\beta) &= \sum_{v=1}^{2^K-1} \Pr(\mathcal{E}_{lv}) \left\{ \frac{1}{f_{IIk} D_{lv}} - 1 \right\} \text{Var} \left\{ \tilde{Q}_{IIk, \mathcal{E}_{lv}}(\beta) | \mathcal{E}_{lv} \right\}. \end{aligned}$$

From Theorem 1, we note that $\Sigma_k(\beta)$ consists of three parts. The first part $V_{I,lk}(\beta)$ is a contribution to the variance from the full cohort, and the second part $V_{II,lk}(\beta)$ and the third part $V_{III,lk}(\beta)$ are due to sampling for the subcohort and for a portion of cases in non-subcohort, respectively. For studies based on the entire cohort, the second and third parts vanish, so the variance contains only the first part $V_{I,lk}(\beta)$. If traditional stratified case-cohort studies are conducted, then the third part equal to 0. Moreover, for unstratified generalized case-cohort studies (i.e. $L = 1$ and $q_I = 1$), the variance only consists of $V_{I,1k}(\beta)$, $V_{II,1k}(\beta)$, and $V_{III,1k}(\beta)$. The illustration of asymptotic covariance when $K = 2$ has been added in Supplementary material (Web Appendix C).

For the asymptotic property of the baseline cumulative hazard function estimators $\tilde{\Lambda}_{0k}(\hat{\beta}_k, t)$ we define $D[0, \tau]$ be a metric space consisting of right-continuous functions $f(t)$ with left-hand limits, where $f(t) : [0, \tau] \rightarrow R$ and $d(f, g) = \sup_{t \in [0, \tau]} \{|f(t) - g(t)|\}$ for $f, g \in D[0, \tau]$. The properties are summarized in the following theorem.

Theorem 2—Under the regularity conditions in the Supplementary material (Web Appendix A), $\tilde{\Lambda}_{0k}(\hat{\beta}_k, t)$ is a consistent estimator of $\Lambda_{0k}(t)$ in $t \in [0, \tau]$ and $n^{1/2}\{\tilde{\Lambda}_{0k}(t) - \Lambda_{0k}(t)\}$ converges weakly to a mean zero Gaussian process in $D[0, \tau]$ whose covariance function is given in the Supplementary material (Web Appendix A).

The proofs for Theorems 1 and 2 are provided in the Supplementary material (Web Appendix A).

3. Optimal Weighted Estimator

We aim to derive the optimal estimator among the class of generalized weighted estimating functions in Section 2.2. Equivalently, we wish to find the optimal weight for $\pi_{lik}(t)$ such that the asymptotic variance for each $\hat{\beta}_k$ is minimized.

From the expression in Theorem 1, the sandwich covariance matrix for $\hat{\beta}_k$ depends on the first derivative of the weighted estimating functions $A_k(\beta_k)$ and the asymptotic variance of the weighted estimating functions $\Sigma_k(\beta_k)$. The former is $A_k(\beta_k) = \int_0^\tau v_k(\beta_k, t) s_k^{(0)}(\beta_k, t) \lambda_{0k}(t) dt$ so is independent of the weights. Thus, only the asymptotic variance of the proposed weighted estimating functions $Var\{\tilde{U}_k^O(\beta)\}$ depends on the choice of weights. In order to find the optimal weights in the proposed weight function, we should minimize $Var\{\tilde{U}_k^O(\beta)\}$. Since this variance depends on the joint distribution of all outcomes in a complicated way, in the Supplementary material (Web Appendix B), we assume the weights at each partitioned region to be approximately constant yielding that the choice of $\pi_{lik}(t)$'s with the smallest variance subject to constraint $E\{\pi_{lik}(t)\} = 1$ is optimal.

For illustration, we consider $K = 2$. After some algebra, this optimization is equivalent to minimizing

$$\begin{aligned} & E\left\{\prod_{j=1}^2 (1 - \Delta_{lij})\right\} \alpha_{lk}(t) \tilde{a}_{0lk}^2(t) \\ & + E\left\{\Delta_{li1}(1 - \Delta_{li2})\right\} \left\{\alpha_{lk}(t) \tilde{a}_{1lk}^2(t) + (1 - \alpha_{lk}(t)) \gamma_{l1}(t) \tilde{b}_{1lk}^2(t)\right\} \\ & + E\left\{\Delta_{li1} \Delta_{li2}\right\} \left\{\alpha_{lk}(t) \tilde{a}_{3lk}^2(t) + (1 - \alpha_{lk}(t)) (\gamma_{l1}(t) \tilde{b}_{3lk}^2(t) + \gamma_{l2}(t) \tilde{b}_{4lk}^2(t) \right. \\ & \left. + \gamma_{l1}(t) \gamma_{l2}(t) [\tilde{b}_{5lk}^2(t) + 2\{\tilde{b}_{3lk}(t) \tilde{b}_{4lk}(t) + \tilde{b}_{3lk}(t) \tilde{b}_{5lk}(t) + \tilde{b}_{4lk}(t) + \tilde{b}_{5lk}(t)\}])\right\} \end{aligned}$$

subject to constraint

$$\begin{aligned} 1 = & E\left\{\prod_{j=1}^2 (1 - \Delta_{lij})\right\} \alpha_{lk}(t) \alpha_{0lk}(t) \\ & + E\left\{\Delta_{li1}(1 - \Delta_{li2})\right\} \left\{\alpha_{lk}(t) \tilde{\alpha}_{1lk}(t) + (1 - \alpha_{lk}(t)) \gamma_{l1}(t) \tilde{b}_{1lk}(t)\right\} \\ & + E\left\{(1 - \Delta_{li1}) \Delta_{li2}\right\} \left\{\alpha_{lk}(t) \tilde{\alpha}_{2lk}(t) + (1 - \alpha_{lk}(t)) \gamma_{l2}(t) \tilde{b}_{2lk}(t)\right\} \\ & + E\left\{\Delta_{li1} \Delta_{li2}\right\} \left\{\alpha_{lk}(t) \tilde{\alpha}_{3lk}(t) \right. \\ & \left. + (1 - \alpha_{lk}(t)) \left\{\gamma_{l1}(t) \tilde{b}_{3lk}(t) + \gamma_{l2}(t) \tilde{b}_{4lk}(t) + \gamma_{l1}(t) \gamma_{l2}(t) \tilde{b}_{5lk}(t)\right\}\right\}. \end{aligned}$$

Using the Lagrange multiplier (the detail is given in the Supplementary material (Web Appendix B)), we obtain the optimal weights as

$$\begin{aligned} \tilde{a}_{0lk}(t) &= \alpha_{lk}(t)^{-1}, \tilde{a}_{1lk}(t) = \tilde{b}_{1lk}(t) = [\alpha_{lk}(t) + \{1 - \alpha_{lk}(t)\} \gamma_{l1}(t)]^{-1}, \quad (5) \\ \tilde{a}_{2lk}(t) &= \tilde{b}_{2lk}(t) = [\alpha_{lk}(t) + \{1 - \alpha_{lk}(t)\} \gamma_{l2}(t)]^{-1}, \\ \tilde{a}_{3lk}(t) &= \tilde{b}_{3lk}(t) = \tilde{b}_{4lk}(t) = -\tilde{b}_{5lk}(t) \\ &= [\alpha_{lk}(t) + \{1 - \alpha_{lk}(t)\} \{\gamma_{l1}(t) + \gamma_{l2}(t) - \gamma_{l1}(t) \gamma_{l2}(t)\}]^{-1}. \end{aligned}$$

In other words, this proposed weight yields the smallest asymptotic variance. Using the observed data, the optimal weights can be estimated as

$$\hat{\alpha}_{lk}^{(t)} = \frac{\sum_{i=1}^{n_l} Y_{lik}(t)(1 - \Delta_{lik})\xi_{li}}{\sum_{i=1}^{n_l} Y_{lik}(t)(1 - \Delta_{lik})}, \quad \hat{\gamma}_{l1}^{(t)} = \frac{\sum_{i=1}^{n_l} Y_{lik}(t)\Delta_{li1}(1 - \xi_{li})\eta_{li1}}{\sum_{i=1}^{n_l} Y_{lik}(t)\Delta_{li1}(1 - \xi_{li})},$$

$$\hat{\gamma}_{l2}^{(t)} = \frac{\sum_{i=1}^{n_l} Y_{lik}(t)\Delta_{li2}(1 - \xi_{li})\eta_{li2}}{\sum_{i=1}^{n_l} Y_{lik}(t)\Delta_{li2}(1 - \xi_{li})}, \quad \hat{a}_{0lk}^{(t)} = \hat{\alpha}_{lk}^{(t)-1}$$

$$\hat{a}_{1lk}^{(t)} = \hat{b}_{1lk}^{(t)} = [\hat{\alpha}_{lk}^{(t)} + \{1 - \hat{\alpha}_{lk}^{(t)}\}\hat{\gamma}_{l1}^{(t)}]^{-1}, \quad \hat{a}_{2lk}^{(t)} = \hat{b}_{2lk}^{(t)} = [\hat{\alpha}_{lk}^{(t)} + \{1 - \hat{\alpha}_{lk}^{(t)}\}\hat{\gamma}_{l2}^{(t)}]^{-1},$$

$$\hat{a}_{3lk}^{(t)} = \hat{b}_{3lk}^{(t)} = \hat{b}_{4lk}^{(t)} = -\hat{b}_{5lk}^{(t)} = [\hat{\alpha}_{lk}^{(t)} + \{1 - \hat{\alpha}_{lk}^{(t)}\}\{\hat{\gamma}_{l1}^{(t)} + \hat{\gamma}_{l2}^{(t)} - \hat{\gamma}_{l1}^{(t)}\hat{\gamma}_{l2}^{(t)}\}]^{-1}.$$

If all information of covariates are available, all the sampling probabilities and the weights are equal to 1 (i.e. $\hat{\alpha}_{lk}^{(t)} = \hat{\gamma}_{lk}^{(t)} = 1$ for $k = 1, \dots, K$). Consequently, the optimal weighting function is the Cox score function in this extreme case.

4. Simulation Study

We conduct simulation studies to investigate the finite sample properties of the proposed methods. We also compare it with Kang and Cai (2010)'s and Kim et al. (2016)'s weights, and compare the performance of stratified sampling with unstratified sampling. Kang and Cai (2010)'s method ignores the exposure information of subjects with other disease, so we consider the results based on Kang and Cai (2010) as naïve analysis for comparison.

In the simulation study, we consider $K = 2$ and generate multivariate failure time data from Clayton-Cuzick model (Clayton and Cuzick (1985)). The bivariate survival function for the bivariate survival time (T_1, T_2) given (Z_1, Z_2) has the following form:

$$F(t_1, t_2 | Z_1, Z_2) = \{S_1(t_1; Z_1)^{-1/\theta} + S_2(t_2; Z_2)^{-1/\theta} - 1\}^{-\theta},$$

where $Z_1 = Z_2 = Z$ is generated from Bernoulli distribution with $\text{pr}(Z = 1) = 0.5$,

$S_k(t; Z_k) = \text{Pr}(T_k > t | Z_k) = e^{-\int_0^t \lambda_{0k}(t) e^{\beta_k Z_{lk}} dt}$, $\lambda_{0k}(t)$ and β_k ($k = 1, 2$) are the baseline hazard function and the covariate effect for disease k , respectively, and θ is the association parameter between the failure times of the two diseases. Exponential distribution with failure rate $\lambda_{0k} e^{\beta_k Z_{lk}}$ is considered for the marginal distribution of T_k ($k = 1, 2$). The relationship between Kendall's tau, τ_θ , and θ is $\tau_\theta = 1/(2\theta + 1)$, smaller Kendall's tau represents a less correlation between T_1 and T_2 . Values of 0.1, 0.67 and 4 are used for θ so the corresponding

Kendall's tau is 0.83, 0.43 and 0.11, respectively. We set $\beta_k = 0$ or $\log 2$, $\lambda_{01} = 2$ and $\lambda_{02} = 4$. Additionally, we generate sampling strata variable V where V has two strata: 0 and 1. We define two parameters: $\eta = \Pr(V = 1 | Z = 1)$ and $\nu = \Pr(V = 0 | Z = 0)$. Hence, an unstratified sampling is a special case with $\eta = 0.5$ and $\nu = 0.5$. The larger the values of η and ν than 0.5 the more V and Z are correlated. For stratified case-cohort studies, we set the values $[\eta, \nu] = [0.7, 0.7]$ and $[\eta, \nu] = [0.9, 0.9]$. Finally, the censoring time is simulated from uniform distribution $[0, u]$ where u depends on the specified level of the censoring probability resulting in the event rate of 10% and 12% for $k = 1$ and 18% and 23% for $k = 2$. Overall, the proportions of subjects who have both diseases are around 8%, 5% and 3% for $\theta = 0.1, 0.67$ and 4, respectively. The sample size of the full cohort is set to be $n = 4000$. For the generalized case-cohort design, we select the subcohort and a subset of cases by simple random sampling as well as stratified sampling and consider the subcohort size of 400 and 800. We select the subcohort $\tilde{n}_l = \tilde{n} \times q_l$ from each stratum. By using a simple random sampling, we select non-subcohort cases size of $\tilde{m}_{lk} = (d_{lk} - \tilde{d}_{lk}) \times r_k$ for $k = 1, 2$ and $l = 0, 1$. For each configuration, 2000 replications are conducted.

In the first set of simulations, we consider generalized case-cohort studies with simple random sampling of subcohort and cases (i.e. $L = 1$). Our main interests are to estimate the effect of Z on disease 1 (β_1) but covariate information for disease 2 is available from another generalized case-cohort study. We examine the performance of our proposed estimator based on (2) with optimal weights (5) which uses the additional information collected on the sampled subjects with disease 2. We set the selection probabilities of cases outside the subcohort for disease 1 and 2 with 0.1 and 0.2. Table 1 summarizes the results. For different combinations of true β_1 , case selection probabilities, the subcohort sample size, and correlation between two failure times, Table 1 shows the average of the estimates for β_1 , the average of the proposed estimated standard error (SE), empirical standard deviation (SD), and sample relative efficiency (SRE). The subscripts for SE, SD, and CR refer to the proposed method (o), Kim et al. (2016)'s method (k), and Kang and Cai (2009)'s method (c). The sample relative efficiency (SRE) relative to Kim et al. (2016)'s method and Kang and Cai (2009)'s method are defined as, $SRE_1 = SD_k^2 / SD_o^2$ and $SRE_2 = SD_c^2 / SD_o^2$, respectively.

From the results, we observe that the three estimators are approximately unbiased. The average of the proposed estimated standard error is close to the empirical standard deviation. As expected, larger case selection probability and subcohort size produce smaller standard deviations. The range of the 95% confidence interval coverage rate for the proposed optimal weight is between 94%-96%. All sample relative efficiency, defined as squared empirical standard deviations of the existing weight relative to those of the proposed optimal weight, are greater than 1. The results in Table 1 show that our proposed optimal weights are the most efficient compared to the other two weights. Specifically, the optimal weight increases the efficiency from 15% to 172% with higher efficiency gain associated with smaller case selection probability and larger subcohort size. Furthermore, the efficiency gain is larger when the dependence between the disease outcomes are more correlated.

In the second set of simulations, we also examine the performance for the proposed optimal weight under stratified case-cohort design and compare it with Kang and Cai (2010)'s and

Kim et al. (2016)'s weights. The population and subcohort sizes are 2000 and 400, respectively. We set the event proportion for disease 1 and disease 2 [12%, 23%] and case selection probabilities [0.3, 0.6]. Table 2 provides summary statistics for the estimate of $\beta_1 = 0$ and $\log(2)$. The conclusions are similar with those in Table 1. Note that empirical standard deviations are smaller when the correlation between stratum variables and covariates is larger. It suggests that stratified sampling produces efficiency gain when stratum variable is associated with covariate.

When there are more studies with other types of diseases, the number of subjects with expensive exposure information increases. Therefore, using information from more studies with other types of diseases could improve efficiency. We conducted some additional simulations including 3 diseases types. The results are summarized in the Supplementary material (Web Appendix D: Table S1). We compared the performance of the estimators for 4 different weights: 1) optimal weights with 3 disease types, 2) optimal weights with 2 disease types, 3) Kim et al. (2016)'s weight with 2 disease types, and 4) Kang and Cai (2010)'s weight with 2 disease types. The results suggest that the optimal weight with 3 disease types improved efficiency. We also provide information on computing time in the Supplementary material (Web Appendix D: Table S2). Computation time for using the optimal weight with 3 disease types is about 1.7 times of that for using the optimal weight with 2 disease types and the Kim et al. (2016)'s weight and it is about 3 times of that for using the Kang and Cai (2010)'s weight.

5. Application to the ARIC Study

We apply the proposed method to a data set from the ARIC study which is a population-based cohort study (Duncan et al., 2003; Ballantyne et al., 2004). This study consists of 15,792 men and women 45 - 64 years of age from four U.S. communities recruited during 1987 to 1989. All subjects are followed for incident diabetes. The incident diabetes are defined as a reported physician diagnosis, use of antidiabetes medications, a fasting (≥ 8 hours) glucose ≥ 7.0 mmol/l, or a nonfasting glucose of ≥ 11.1 mmol/l. Subjects are regarded as censored if they are alive and event-free at the end of 1998 or lost to follow-up.

Our interest is to investigate the association between high-sensitivity C-reactive protein (hs-CRP), which is a biomarker of inflammation, and incident diabetes events. In order to measure hs-CRP, a case-cohort study was conducted to reduce the cost and save blood specimen. Hs-CRP is also available on subjects for incident coronary heart disease (CHD) from another case-cohort study in the ARIC study (Ballantyne et al., 2004). We exclude subjects with prevalent CHD and prevalent diabetes at baseline, had transient ischemic attack or stroke, had missing follow-up visits, were in minority race group except for African-American or white, had no valid diabetes determination at follow-ups, or had missing CHD information and baseline measurements. The full cohort after exclusion consists of 10,279 subjects.

To preserve frozen biologic specimens and reduce costs, a generalized case-cohort design is conducted by selecting a subset of incident diabetes events since the rate of diabetes during follow-up is 11.2%. The subcohort and cases of incident diabetes are randomly selected via

stratified sampling where the strata variables are age at baseline (≤ 55 and > 55), sex, and race (black and white). Age, gender, race, parental history of diabetes, hypertension, and center are confounding factors and are adjusted in the model. The risk factor, hs-CRP, is used as a categorical variables with 4 levels based on quartiles. In Table 3, hs-CRP (C2), hs-CRP (C3), and hs-CRP (C4) are indicator variables for hs-CRP values in the second, third, and fourth quartiles, respectively. The hs-CRP values in the first quartile is used as the reference group in our analysis.

By using available hs-CRP information collected from both case-cohort studies, we apply our proposed method to this data set. The total sample size was 1,576 subjects including 572 noncases, 581 diabetes cases, 423 CHD cases. The subcohort size is 668 which consists of 96 diabetes cases and 572 non-cases. To study the effect of hs-CRP on diabetes, we fit the model using (1) and compare our proposed optimal estimator with that in Kang and Cai (2010) and Kim et al. (2016).

Table 3 presents the estimates, standard errors, hazard ratios, and the 95% confidence intervals for the three methods. First, we test overall effects for hs-CRP using our proposed method and they are statistically significant. The hazard ratio comparing the fourth with the first hs-CRP quartile group is 2.74 and confidence interval indicates that it is of statistical significance. Moreover, the hazard ratio comparing the third with the first hs-CRP quartile group is also statistically significant, but the hazard ratio for the second versus the first quartile group is not statistically significant. Race effect is statistically significant using the proposed method while it is not using Kang and Cai (2010)'s and Kim et al. (2016)'s methods. The regression coefficient estimates for the proposed method are similar with those for the existing method, but all the standard errors are smaller than those of the existing method and consequently the 95% confidence intervals are narrower.

6. Concluding Remarks

When multiple generalized case-cohort studies are conducted, some additional information for expensive covariates are available. In this paper, we proposed a more general approach for the generalized case-cohort study by using this additional information. Our proposed estimators are shown to be consistent and asymptotically normally distributed under some regularity conditions. We also examined the optimal choice of the weights within our proposed class of weights. In addition to simple random sampling for the subcohort and cases, we also considered stratified sampling to improve efficiency. The simulation results showed that our proposed optimal methods improve efficiency significantly compared to the existing methods especially in the situation when the case selection probability is very small.

In this paper, we allow for stratified sampling for the subcohort and cases selection. The sampling strata are formed to ensure proper representation of certain subgroups in the subcohort. Such stratified sampling will improve the estimation of stratum specific quantities if the stratum is relatively small in the whole cohort. It could also improve the overall estimation for the primary quantity of interest but that could depend on many factors such as the relationship between the strata and the disease of interest, the relationship

between the strata and the main exposure as well as other covariates in the model, the proportion of each stratum in the cohort, etc.

The model we considered in this paper has the baseline hazard function to be common across sampling strata. The effect of the sampling strata can be adjusted for by including the sampling strata variable as part of the covariates. This type of model is commonly used in epidemiological studies. An extension of this model is to allow the baseline function to be different across strata which is also commonly used in biomedical research. It is of interest to extend our approach to such stratified model.

The current method assumed the disease-specific effect model as was considered in Wei et al. (1989). If part of the covariate effects are expected to be common for different disease types, the model considered in Kang and Cai (2009) can be used. Under Kang and Cai (2009)'s model, one possibility to improve efficiency is to jointly model all the disease outcomes. Incorporating correlation between event times could further improve efficiency as was explored in Cai and Prentice (1995). This is worthy of future research.

In this paper, we only consider the situation where the diseases are non-competing and non-recurrent, for example, as in the situation for the ARIC study where coronary heart disease and diabetes are of interest and a person can have both coronary heart disease and diabetes. The ideas in this paper can be extend to other setting such as competing risks, semi-competing risks, or recurrent events. These extensions are worthy of future investigation.

In some applications, proportional hazard assumptions may not be appropriate or investigators may be interested in a different form of association between risk factor and disease outcomes. Hence, alternatives to proportional hazard models such as additive hazards models, proportional odds model, accelerated failure time model, and semiparametric transformation model could be of interest. Extending our approaches to such models warrants further investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the National Institutes of Health grants (P01CA142538 and R01ES021900) and Institutional Research Grant #14-247-29 from the American Cancer Society and the MCW Cancer Center. This manuscript was prepared using ARIC Research Materials obtained from the National Heart, Lung, and Blood Institute (NHLBI) Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the ARIC or the NHLBI.

References

- Ballantyne CM, Hoogeveen RC, Bang H, Coresh J, Folsom AR, Heiss G, Sharrett AR. 2004; Lipoprotein-associated phospholipase a2, high-sensitivity c-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the atherosclerosis risk in communities (aric) study. *Circulation*. 109:837–842. [PubMed: 14757686]
- Barlow W. 1994; Robust variance estimation for the case-cohort design. *Biometrics*. 50:1064–72. [PubMed: 7786988]

- Borgan O, Langholz B, Samuelsen SO, G L, Pogoda J. 2000; Exposure stratified case-cohort designs. *Lifetime Data Anal.* 6:39–58. [PubMed: 10763560]
- Cai J, Prentice R. 1995; Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika.* 82:15164.
- Cai J, Zeng D. 2007; Power calculation for case-cohort studies with nonrare events. *Biometrics.* 63:1288–95. [PubMed: 17608788]
- Clayton D, Cuzick J. 1985; Multivariate generalizations of the proportional hazards model(with discussion). *J R Statist Soc A.* 148:82–117.
- Duncan BB, Schmidt MI, Pankow JS, Ballantyne CM, Couper D, Vigo A, Hoogeveen R, Folsom AR, Heiss G. 2003; Low-grade systemic inflammation and the development of type 2 diabetes. *Diabetes.* 52:1799–1805. [PubMed: 12829649]
- Kang S, Cai J. 2009; Marginal hazard model for case-cohort studies with multiple disease outcomes. *Biometrika.* 96:887–901. [PubMed: 23946547]
- Kang S, Cai J. 2010; Asymptotic results for fitting marginal hazards models from stratified case-cohort studies with multiple disease outcomes. *J Korean Stat Soc.* 39:371–385. [PubMed: 22442642]
- Kim S, Cai J, Couper D. 2016; Improving the efficiency of estimation in the additive hazards model for stratified case-cohort design with multiple diseases. *Statistics in Medicine.* 35:282–293. [PubMed: 26310388]
- Kim S, Cai J, Lu W. 2013; More efficient estimators for case-cohort studies. *Biometrika.* 100:695–708. [PubMed: 24634519]
- Kulich M, Lin DY. 2004; Improving the efficiency of relative-risk estimation in caes-cohort study. *J Am Statist Assoc.* 99:832–44.
- Prentice R. 1986; A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika.* 73:1–11.
- Samuelsen SO, Anestad H, Skrandal A. 2007; Stratified case-cohort analysis of general cohort sampling designs. *Scan J Statist.* 34:103–19.
- Self SG, Prentice RL. 1988; Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann Statist.* 34:103–19.
- Wei LJ, Lin DY, Weissfeld L. 1989; Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Statist Assoc.* 84:1065–73.

Figure 1.
Example of generalized case-cohort data

Table 1

Simulation result for simple random sampling of subcohort and cases: $P[\gamma_1 = \gamma_2] = [10\%, 18\%]$

β_1	$\tilde{n}$	γ_1	τ_θ	Optimal weight				Kim et al. [2016]'s weight				Kang and Cai [2012]'s weight					
				β_1	SE_θ	SD_θ	CR_θ	β_1	SE_k	SD_k	CR_k	SRE_1	β_1	SE_c	SD_c	CR_c	SRE_2
0	400	0.1	0.83	-0.003	0.224	0.218	0.96	-0.005	0.302	0.264	0.97	1.47	0.002	0.365	0.309	0.96	2.01
			0.43	-0.006	0.233	0.239	0.94	-0.005	0.345	0.301	0.96	1.58	-0.008	0.360	0.316	0.95	1.75
		0.2	0.11	0.001	0.237	0.237	0.95	-0.001	0.366	0.303	0.97	1.64	0.000	0.361	0.315	0.95	1.77
			0.83	-0.002	0.191	0.187	0.95	-0.002	0.210	0.204	0.95	1.19	0.000	0.239	0.230	0.96	1.51
		800	0.43	-0.004	0.199	0.202	0.94	-0.004	0.227	0.220	0.95	1.19	-0.005	0.238	0.235	0.94	1.35
			0.11	0.003	0.202	0.201	0.95	0.003	0.234	0.221	0.96	1.21	0.004	0.238	0.229	0.96	1.30
	log(2) = 0.693	0.1	0.83	0.000	0.183	0.186	0.95	-0.007	0.277	0.253	0.96	1.86	-0.008	0.338	0.295	0.95	2.53
			0.43	0.002	0.189	0.192	0.95	0.009	0.323	0.271	0.97	2.00	0.008	0.328	0.291	0.94	2.29
		0.2	0.11	0.007	0.191	0.199	0.94	0.007	0.346	0.289	0.97	2.11	0.012	0.335	0.294	0.95	2.19
			0.83	0.001	0.162	0.163	0.95	-0.001	0.187	0.186	0.95	1.30	-0.002	0.215	0.212	0.95	1.68
		800	0.43	0.001	0.168	0.172	0.94	0.003	0.205	0.200	0.95	1.35	0.002	0.214	0.219	0.94	1.61
			0.11	0.005	0.171	0.176	0.94	0.004	0.213	0.208	0.95	1.41	0.006	0.215	0.214	0.95	1.48
0.2	400	0.1	0.83	0.708	0.234	0.234	0.95	0.716	0.322	0.281	0.96	1.45	0.711	0.388	0.335	0.95	2.05
			0.43	0.697	0.245	0.249	0.95	0.706	0.377	0.325	0.96	1.71	0.706	0.393	0.347	0.95	1.95
		0.2	0.11	0.708	0.250	0.258	0.95	0.719	0.391	0.334	0.96	1.67	0.723	0.388	0.347	0.94	1.80
			0.83	0.708	0.199	0.201	0.95	0.710	0.220	0.215	0.96	1.15	0.705	0.251	0.244	0.96	1.48
		800	0.43	0.693	0.208	0.209	0.95	0.696	0.239	0.232	0.95	1.23	0.694	0.250	0.250	0.95	1.43
			0.11	0.704	0.212	0.218	0.95	0.708	0.248	0.244	0.95	1.25	0.710	0.251	0.256	0.94	1.38
	log(2) = 0.693	0.1	0.83	0.695	0.191	0.192	0.95	0.702	0.300	0.267	0.96	1.93	0.699	0.354	0.317	0.94	2.72
			0.43	0.699	0.198	0.205	0.94	0.717	0.353	0.294	0.96	2.06	0.711	0.355	0.311	0.95	2.30
		0.2	0.11	0.704	0.201	0.206	0.94	0.711	0.370	0.304	0.97	2.17	0.713	0.362	0.312	0.94	2.29
			0.83	0.695	0.169	0.168	0.95	0.698	0.198	0.194	0.95	1.33	0.694	0.226	0.226	0.95	1.81
		800	0.43	0.701	0.176	0.181	0.94	0.711	0.218	0.212	0.95	1.38	0.709	0.227	0.222	0.95	1.51
			0.11	0.701	0.179	0.181	0.95	0.700	0.225	0.213	0.97	1.39	0.702	0.226	0.219	0.96	1.46

SE, the average of the estimates of standard error; SD, sample standard deviation; CR, the coverage rate of the nominal 95% confidence intervals; $SRE_1 = SD_k^2 / SD_o^2$ sample relative efficiency; $SRE_2 = SD_c^2 / SD_o^2$ sample relative efficiency.

Table 2

Simulation result for stratified sampling of subcohort and cases: $P [\tau_1, \tau_2] = [12\%, 23\%]$

β_1	$[\nu, \mu]$	τ_θ	Optimal weight					Kim et al. [2016]'s weight					Kang and Cai's weight				
			β	SE_o	SD_o	CR_o	β	SE_k	SD_k	CR_k	SRE_1	β	SE_c	SD_c	CR_c	SRE_2	
0	[0.5, 0.5]	0.83	0.008	0.192	0.187	0.95	0.008	0.199	0.197	0.95	1.10	0.006	0.240	0.242	0.94	1.67	
		0.43	0.006	0.202	0.198	0.95	0.005	0.222	0.215	0.95	1.18	0.004	0.240	0.237	0.95	1.43	
		0.11	0.003	0.205	0.209	0.94	0.005	0.232	0.229	0.95	1.21	0.007	0.239	0.249	0.93	1.43	
	[0.7, 0.7]	0.83	0.007	0.196	0.185	0.96	0.009	0.194	0.187	0.95	1.03	0.005	0.242	0.229	0.96	1.54	
		0.43	-0.006	0.200	0.198	0.95	-0.010	0.219	0.209	0.96	1.11	-0.006	0.242	0.235	0.95	1.41	
		0.11	-0.003	0.201	0.203	0.94	0.000	0.233	0.216	0.96	1.13	0.003	0.241	0.230	0.95	1.29	
	[0.9, 0.9]	0.83	0.001	0.209	0.180	0.97	0.002	0.177	0.180	0.95	1.00	0.001	0.247	0.230	0.95	1.63	
		0.43	0.000	0.194	0.191	0.95	0.002	0.211	0.195	0.95	1.04	0.000	0.246	0.222	0.96	1.36	
		0.11	0.002	0.187	0.193	0.94	0.003	0.234	0.198	0.97	1.06	0.005	0.246	0.213	0.96	1.21	
$\log(2) = 0.693$	[0.5, 0.5]	0.83	0.701	0.201	0.201	0.95	0.708	0.210	0.211	0.95	1.10	0.703	0.252	0.253	0.96	1.58	
		0.43	0.699	0.211	0.207	0.96	0.708	0.235	0.225	0.96	1.19	0.703	0.253	0.254	0.94	1.51	
		0.11	0.693	0.215	0.210	0.96	0.694	0.245	0.232	0.96	1.22	0.690	0.251	0.250	0.95	1.43	
	[0.7, 0.7]	0.83	0.697	0.205	0.196	0.96	0.704	0.205	0.203	0.95	1.07	0.706	0.255	0.250	0.95	1.62	
		0.43	0.697	0.209	0.213	0.95	0.706	0.233	0.229	0.95	1.15	0.702	0.254	0.253	0.95	1.41	
		0.11	0.699	0.211	0.213	0.94	0.704	0.246	0.234	0.96	1.21	0.704	0.254	0.249	0.95	1.37	
	[0.9, 0.9]	0.83	0.698	0.215	0.191	0.97	0.709	0.189	0.194	0.94	1.03	0.715	0.263	0.241	0.96	1.59	
		0.43	0.695	0.201	0.203	0.95	0.709	0.228	0.207	0.97	1.04	0.704	0.262	0.229	0.97	1.27	
		0.11	0.695	0.195	0.208	0.94	0.707	0.254	0.215	0.97	1.07	0.705	0.263	0.230	0.97	1.23	

SE, the average of the estimates of standard error; SD, sample standard deviation; CR, the coverage rate of the nominal 95% confidence intervals; $SRE_1 = SD_k^2 / SD_o^2$, sample relative efficiency; $SRE_2 = SD_c^2 / SD_o^2$, sample relative efficiency.

Table 3

Results for the effect of hs-CRP from the ARIC Study

Variables	Optimal weight				Kim et al. [2016]'s weight				Kang and Cai's weight			
	β_k	SE	HR	95% CI	β_k	SE	HR	95% CI	β_k	SE	HR	95% CI
hs-CRP(C4)	1.01	0.209	2.74	(1.82, 4.12)	1.01	0.213	2.74	(1.81, 4.16)	1.02	0.220	2.78	(1.80, 4.28)
hs-CRP(C2)	0.33	0.227	1.40	(0.89, 2.18)	0.20	0.238	1.22	(0.77, 1.95)	0.23	0.243	1.26	(0.78, 2.02)
hs-CRP(C3)	0.70	0.206	2.02	(1.35, 3.03)	0.72	0.212	2.06	(1.36, 3.12)	0.75	0.220	2.12	(1.38, 3.26)
Age	0.01	0.009	1.01	(0.99, 1.03)	0.01	0.011	1.01	(0.98, 1.03)	0.01	0.012	1.01	(0.98, 1.03)
African	0.66	0.258	1.94	(1.17, 3.22)	0.54	0.278	1.71	(0.99, 2.95)	0.55	0.287	1.73	(0.98, 3.03)
Male	0.28	0.092	1.33	(1.11, 1.59)	0.34	0.119	1.40	(1.11, 1.77)	0.33	0.131	1.40	(1.08, 1.81)
PHD	0.58	0.150	1.79	(1.33, 2.40)	0.60	0.153	1.82	(1.35, 2.46)	0.63	0.160	1.88	(1.37, 2.57)
HYP	0.55	0.151	1.74	(1.29, 2.34)	0.57	0.154	1.77	(1.31, 2.40)	0.56	0.161	1.75	(1.28, 2.40)
Center (F)	0.10	0.225	1.10	(0.71, 1.71)	0.14	0.226	1.15	(0.74, 1.79)	0.18	0.237	1.19	(0.75, 1.90)
Center (J)	-0.19	0.315	0.83	(0.45, 1.53)	-0.12	0.324	0.89	(0.47, 1.68)	-0.09	0.334	0.92	(0.48, 1.76)
Center (M)	0.02	0.220	1.02	(0.66, 1.57)	-0.06	0.223	0.95	(0.61, 1.46)	-0.02	0.233	0.98	(0.62, 1.56)

hs-CRP, high-sensitivity C-reactive protein; PHD, parental history of diabetes; HYP, hypertension; SE, standard error estimate; HR, hazard ratio estimate; CI, confidence interval