Biometrics 72, 30–38
March 2016
DOI: 10.1111/biom.12392

Augmented Case-Only Designs for Randomized Clinical Trials with Failure Time Endpoints

James Y. Dai,* Xinyi Cindy Zhang, Ching-Yun Wang,*** and Charles Kooperberg******

Fred Hutchinson Cancer Research Center and University of Washington, Seattle, Washington
**email:* jdai@fredhutch.org
***email:* xzhan2@fredhutch.org
****email:* cywang@fredhutch.org
*****email:* clk@fredhutch.org

Summary. Under suitable assumptions and by exploiting the independence between inherited genetic susceptibility and treatment assignment, the case-only design yields efficient estimates for subgroup treatment effects and gene-treatment interaction in a Cox model. However it cannot provide estimates of the genetic main effect and baseline hazards, that are necessary to compute the absolute disease risk. For two-arm, placebo-controlled trials with rare failure time endpoints, we consider augmenting the case-only design with random samples of controls from both arms, as in the classical case-cohort sampling scheme, or with a random sample of controls from the active treatment arm only. The latter design is motivated by vaccine trials for cost-effective use of resources and specimens so that host genetics and vaccine-induced immune responses can be studied simultaneously in a bigger set of participants. We show that these designs can identify all parameters in a Cox model and that the efficient case-only estimator can be incorporated in a two-step plug-in procedure. Results in simulations and a data example suggest that incorporating case-only estimators in the classical case-cohort design improves the precision of all estimated parameters; sampling controls only in the active treatment arm attains a similar level of efficiency.

Key words: Case-cohort design; Case-only estimator; Gene-treatment interaction; Nested case-control design; Pharmacogenetics.

1. Introduction

Individuals respond differently to treatment or prevention modalities, depending on their genetic background, environmental exposures, and clinical characteristics (Charlab and Zhang, 2013). In clinical trials, there is a growing interest to discover and characterize individual or subgroup treatment responses, supplementing primary intent-to-treat analyses. For instance, the emerging pharmacogenetics research aims to identify genetic susceptibility that contributes to inter-individual variability of treatment efficacy and safety, in scales ranging from several candidate genes to the whole genome (Evans and McLeod, 2003; Weinshilboum and Wang, 2004). These studies underscore the potential of personalized medicine, and may also elucidate mechanisms of treatment effect.

To this end, this article pertains to sampling designs for characterizing the influence of pre-treatment biomarkers, e.g., a panel of genetic variants, on treatment effects in randomized clinical trials. Ancillary studies of this nature are increasingly common in the genomic era. However, biomakers can be expensive to measure. To study the association of biomarkers with relatively uncommon study outcomes, including HIV infection, most cancers, and some cardiovascular events, it is cost-effective to adopt some form of outcome-dependent sampling. Popular outcome-dependent sampling schemes in cohort studies include the nested case-control design and the case-cohort design (Thomas, 1977; Prentice and Breslow, 1978; Prentice, 1986). Stratified versions of the two sampling designs to oversample certain groups have also been developed for better efficiency (Borgan et al., 2000; Langholz and Borgan, 1995). The properties and utilities of the two designs in cohort studies have been discussed (Self and Prentice, 1988; Langholz and Thomas, 1990).

Consider a two-arm, placebo-controlled randomized prevention trial with a rare failure event. The unique feature is that there is unequivocal design-imposed independence between the treatment assignment and pre-treatment biomarkers, e.g., germline genotypes. Exploiting this independence and assuming censoring being non-informative and independent of randomization arms, case-only methods are more efficient than the two aforementioned designs for estimating gene-treatment interactions and subgroup treatment effects on a rare disease endpoint (Vittinghoff and Bauer, 2006; Dai et al., 2012). These assumptions are better suited for phase III prevention trials where adverse effect is not of concern. Though computed from a logistic model, case-only estimators have the interpretation of hazard ratios in the Cox proportional hazards models. Sensitivity of case-only estimators toward violations of these assumptions has been investigated (Vittinghoff and Bauer, 2006). In recent years, use of case-only methods has started to permeate in prevention trials. See, e.g., trials in the Women's Health Initiative and the HIV

Vaccine Trial Network (Prentice et al., 2010; Dai et al., 2014; Li et al., 2014).

The case-only design, however, does not allow estimation of the full set of parameters in a Cox model. Specifically, neither the genotype main effect nor the cumulative baseline hazard function is estimable from cases alone. These parameters are needed to study the absolute risk of the endpoint for genotype groups in each arm. This limitation hinders interpretation and utility of the estimated gene-treatment interaction, because the estimate of individual absolute risk when treated or when not treated will inform medical counseling and guide treatment selection (Gail et al., 1989; Janes et al., 2011). On the other hand, the traditional case-cohort or nested case-control sampling provides estimates of the baseline hazard and absolute risk, but does not incorporate gene-treatment independence. Leveraging the strengths of both types of designs, we consider augmenting the case-only design to enable estimation of the full set of Cox model parameters. In particular, we focus on variations of the case-cohort design in this article, because it is easy to plan ahead a random subcohort for time-invariant genotypes in clinical trials, and because it has the advantage of accommodating multiple outcomes that may arise in an ancillary study.

Specifically, we consider two scenarios of adding controls to the case-only design, for both of which we can incorporate the case-only estimators in two-step plug-in estimation procedures:

Scenario I: Classical case-cohort design with controls drawn from both arms. In essence, this is one way of adding controls to the case-only design. In this scenario, we essentially propose a novel two-step estimation procedure for the classical case-cohort design: the case-only estimator is first used to estimate gene-treatment interaction and treatment main effect, these estimators will then be plugged into established case-cohort estimation methods as offsets. This method allows widely used case-cohort sampling to take advantage of efficient case-only estimators. Our contributions also include an explicit formula of variance estimates for this two-step procedure.

Scenario II: Augmented case-only (ACO) design with controls drawn from the active treatment arm only. This is a novel design motivated by vaccine trials, as we will elaborate next. Although primarily driven by scientific rationale, this design is of statistical interest, since only three of the four strata formed by case-control status and randomization arm are sampled. It violates the critical identifiability assumption of non-zero sampling probability for all strata in two-phase sampling (Robins et al., 1994; Breslow et al., 2003). The orthogonality between genotype and randomization arm has to be exploited in order to remedy this anomaly. We show that a similar two-step estimation procedure as for **Scenario I** will identify all parameters in a Cox model, and we show in the simulations that the estimators are nearly as efficient as those in **Scenario I**.

Scientifically, the motivation for selecting controls only from the active treatment arm (the ACO design in **Scenario II**) comes from studies on host genetics and immune correlates in HIV vaccine trials. It is common to study vaccine-specific immune responses in a pre-specified sample of trial participants in the vaccine arm, as no vaccine-induced immune responses are generated in the placebo arm. Case-cohort sampling is commonly used in this setting (McElrath et al., 2008). Take Li et al. (2014), e.g., if a genotype in the FcγR gene is associated with varying vaccine protection in the RV144 trial, it is useful to investigate whether specific vaccine-induced immune responses are associated with such genotype, in order to understand functionally why the vaccine effect varies by host genetics. Such relationship can only be studied in the vaccine arm. In this sense, concentrating controls in the vaccine arm is cost-effective when a pharmacogenetic study is a component of a systematic approach for understanding treatment effect. Similar rationale applies to high-throughput biomarker studies for better understanding hormone effect in clinical trials in the Women's Health Initiative (Pitteri et al., 2009).

This article is organized as follows. In Section 2.1 and Section 2.2, we review case-cohort sampling and case-only estimators, respectively. The latter section brings new insights on assumptions required for case-only estimators. In Section 2.3, we show that case-only estimators can be built into a two-step estimation procedure for the case-cohort design. The main parameter of interest we illustrate throughout the article is the genetic main effect. The asymptotic covariance matrix of estimators resulting from the two-step procedure is derived. Extending the results from Section 2.3, we show in Section 2.4 that sampling controls only in the active treatment arm is adequate to estimate all Cox model parameters. For completeness we briefly address the alternative ACO design and the nested case-control sampling in Section 2.5. In Section 3 we compare the efficiency of the proposed designs and estimation methods in simulations, where the standard estimation procedure for a case-cohort design with the same sample size is treated as the benchmark. We present in Section 4 a data example with the standard case-cohort sampling, and we compare standard error estimates resulted from the proposed estimation procedures to the original case-cohort methods. We close with a discussion of the utility of the ACO design and some future work.

2. Method

Consider a two-arm, placebo-controlled randomized prevention trial in which participants were followed for evaluating treatment effect on time to certain failure event. Let Z denote a binary treatment indicator taking the value 1 if the participant is assigned to the active treatment arm, and 0 if assigned to the placebo arm. Let G denote the baseline biomarker of interest, say an inherited genetic variant, and let V be a set of pre-treatment variables to be adjusted in risk association. Denote Y and C as the failure time and the right-censoring time since randomization, respectively. Given

Z, G, and V, C is assumed to be independent of Y. Data consist of independent and identically distributed (iid) vectors $(T_i, \Delta_i, Z_i, G_i, V_i)$, where $T_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \leq C_i)$ for $i = 1 \ldots n$, and $I(\cdot)$ is the indicator function. Cases are defined to be the participants who experienced the failure event ($\Delta = 1$) during follow-up. Using the usual counting process notation, we define $N_i(t) = I(T_i \leq t)$ and $R_i(t) = I(T_i \geq t)$.

Let $\lambda(t; G, Z, V)$ denote the hazard of the failure event occurring at time t for a subject with covariates (G, Z, V).

$$\Pr(Z = z|T = t, \Delta = 1, G, V) = \frac{\lambda(t|Z = z, G, V)\Pr(T \geq t|Z = z, G, V)\Pr(C \geq t|Z = z, G, V)\Pr(Z = z)}{\sum_l \lambda(t|Z = l, G, V)\Pr(T \geq t|Z = l, G, V)\Pr(C \geq t|Z = l, G, V)\Pr(Z = l)}. \quad (5)$$

Consider a proportional hazards model with gene-treatment interaction (Cox, 1972)

$$\lambda(t; G, Z, V) = \lambda_0(t)\exp(\beta_1 G_i + \beta_2 Z_i + \beta_3 G_i Z_i + \beta_4 V_i), \quad (1)$$

where $\lambda_0(t)$ is a baseline hazard function. Denote by $\mathbf{X}_i = (G_i, Z_i, G_i Z_i, V_i)^T$ the vector of baseline covariates included in the regression model and by $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ the vector of regression coefficients of interest. For the full cohort data, an estimate of $\boldsymbol{\beta}$ can be obtained by solving the usual Cox model score function (Cox, 1975),

$$U(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \{\mathbf{X}_i - \bar{\mathbf{X}}(\boldsymbol{\beta}, t)\} dN_i(t), \quad (2)$$

where $\bar{\mathbf{X}}(\boldsymbol{\beta}, t)$ is the weighted mean of covariates at t, expressed as

$$\bar{\mathbf{X}}(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^n R_i(t)\exp(\boldsymbol{\beta}^T \mathbf{X}_i)\mathbf{X}_i}{\sum_{i=1}^n R_i(t)\exp(\boldsymbol{\beta}^T \mathbf{X}_i)}. \quad (3)$$

2.1. *Standard Case-Cohort Sampling Design and Estimation*

The originally described case-cohort sampling design draws a random subcohort and all additional participants who experienced the clinical outcome (Prentice, 1986; Self and Prentice, 1988). With little loss of efficiency, the estimation procedures for the case-cohort design modify (3) using a subset of the entire cohort. Denote by $\mathcal{S}$ the random subcohort in the case-cohort design. The Self-Prentice estimator used at-risk participants in the subcohort (Self and Prentice, 1988),

$$\widehat{\bar{\mathbf{X}}}(\boldsymbol{\beta}, t) = \frac{\sum_{i \in \mathcal{S}} R_i(t)\exp(\boldsymbol{\beta}^T \mathbf{X}_i)\mathbf{X}_i}{\sum_{i \in \mathcal{S}} R_i(t)\exp(\boldsymbol{\beta}^T \mathbf{X}_i)}, \quad (4)$$

while the original Prentice estimator included one more observation, the event occurring at t. The difference of the two estimators is negligible when the sample size is large. More choices of (3) are available, such as the inverse-probability weighting (IPW) method in survey data to improve efficiency of case-cohort estimators (Binder, 1992; Barlow, 1994; Borgan et al., 2000).

2.2. *Case-Only Estimator for Gene-Treatment Interaction and Subgroup Effects*

The treatment effect parameters β_2 and β_3 in model (1) can be estimated from data in cases only (Vittinghoff and Bauer, 2006; Dai et al., 2012), under suitable assumptions about event rate and censoring mechanism. In our notation, the probability of the treatment being z given an event occurring at time t conditional on covariates (G, V) can be expressed as

Detailed derivation can be found in Dai et al. (2012). Equation (5) holds because of the independent censorship and the orthogonality between Z and (G, V). If the event is rare, i.e., $\Pr(T \geq t|Z, G, V) \approx 1$ for all t, and if the censoring time C is independent of treatment Z given (G, V), it follows that a simple logistic regression with an offset can estimate treatment effect parameters,

$$\log\left\{\frac{\Pr(Z = 1|T = t, \Delta = 1, G, V)}{\Pr(Z = 0|T = t, \Delta = 1, G, V)}\right\} = \log\left(\frac{p}{1-p}\right) + \gamma_1 + \gamma_2 G, \quad (6)$$

where $\gamma_1 \approx \beta_2$, $\gamma_2 \approx \beta_3$, and p is the probability of a trial participant being randomized to the treatment arm.

The models (1) and (6) can be more general than what is presented here. For example, the interaction between Z and V can be added into (1), perhaps also the interaction between Z and t, a time-varying hazard ratio. Similar derivations will lead to addition of V and t into (6) correspondingly.

Inspection of (5) and (6) sheds new insights on the assumptions. Arguably, the disease endpoint being rare and censoring being independent of treatment can be restrictive. The former assumption requires the cumulative probability of the event, not just the event probability in any given risk set, is nearly zero. In the context of binary outcomes and logistic regression, analysis of asymptotic bias suggested that the disease prevalence needs to be smaller than the order of $n^{-1/2}$ (Tchetgen and Robins, 2010). This may work for prevention trials with endpoints such as HIV infections and most cancers, but perhaps not for therapeutic endpoints such as tumor response rates. In the failure time setting, the cumulative disease probability is gradually increasing over time, and so the rare disease assumption is perhaps not as stringent as in the setting of binary outcomes. This may explain that in the extensive simulations conducted by Vittinghoff and Bauer (2006), case-only estimators perform surprisingly well even with 20% cumulative event probability. Strictly speaking, what is truly required for derivation of (6) is

$$\frac{\Pr(T \geq t|Z = 1, G, V)}{\Pr(T \geq t|Z = 0, G, V)} \approx 1,$$

which is indeed less restrictive than the rare disease assumption. Moreover, if G has no association with the failure time T and $\Pr(T \geq t|Z=1,G,V)/\Pr(T \geq t|Z=0,G,V)$ is a function of Z but not G, then estimation of the interaction will still work even if the disease is not so rare, because the intercept of (6) is affected in this case but not the slope.

Violation of censoring being independent of treatment given covariates is more amenable. One can directly estimate $\Pr(C \geq t|Z=1,G,V)/\Pr(C \geq t|Z=0,G,V)$ as a function of (t,G,V) from the data, assuming independent censorship. This quantity can be plugged into (6) as an offset to remove any bias induced by differential censoring. Furthermore, what is truly required for deriving (6) is

$$\frac{\Pr(C \geq t|Z=1,G,V)}{\Pr(C \geq t|Z=0,G,V)} \approx 1.$$

As long as G is not associated with C conditional on Z and V and $\Pr(C \geq t|Z=1,G,V)/\Pr(C \geq t|Z=0,G,V)$ is not a function of G, estimation of the interaction parameter in (6) is not affected.

2.3. **Scenario 1:** *Case-Cohort Estimation Incorporating the Case-Only Estimators*

Suppose a case-cohort sample has been drawn for measuring genetic factors, including controls from both arms and possibly stratified by arm and other baseline covariates. Let $\hat{\beta}_{2co}$ and $\hat{\beta}_{3co}$ denote the case-only estimators derived from (6). Plugging the case-only estimators into the Cox model (1),

$$\lambda(t;Z,G,V) = \lambda_0(t)\exp(\beta_1 G + \hat{\beta}_{2co}Z + \hat{\beta}_{3co}GZ + \beta_4 V). \tag{7}$$

The usual case-cohort estimation can be adapted to obtain estimates of β_1 and β_4. For example, the estimator in Self and Prentice (1988) can be obtained by tweaking the `coxph` function in **R** as examplified in Therneau and Li (1999), and adding the estimated offset $\hat{\beta}_{2co}Z + \hat{\beta}_{3co}GZ$. The estimated $\lambda_0(t)$ can be obtained from the Breslow estimator based on estimates of regression parameters as previously described (Prentice, 1986; Self and Prentice, 1988).

The variance estimate of the resulting $\hat{\beta}_1$ and $\hat{\beta}_4$ has to account for the fact that $\hat{\beta}_{2co}$ and $\hat{\beta}_{3co}$ were estimated first by the data in cases. The derivation entails modification of the Murphy–Topel variance estimate of two-step estimators widely used in the econometrics literature (Murphy and Topel, 1985). Here we provide a brief sketch, starting from asymptotic expansions of $(\hat{\beta}_{2co}, \hat{\beta}_{3co})$ and $(\hat{\beta}_1,\hat{\beta}_4)$.

Let $\boldsymbol{\gamma} = (\beta_2,\beta_3)^T$ and let $\boldsymbol{\beta}_g = (\beta_1,\beta_4)^T$. Let $\mathbf{U}_1 = \sum \mathbf{U}_{1i}$ be the estimating equation for $\widehat{\boldsymbol{\gamma}}=(\hat{\beta}_{2co},\hat{\beta}_{3co})^T$ based on the logistic model (6), and $\mathbf{U}_{1i}$ is the iid contribution from the i^{th} case. For all controls, we assume $\mathbf{U}_{1i}=0$. Suppose $\mathbf{A}_1 = \lim -\frac{1}{n}\partial\mathbf{U}_1/\partial\boldsymbol{\gamma}$. By first-order Taylor expansion at $\boldsymbol{\gamma}$,

$$\frac{1}{\sqrt{n}}\mathbf{U}_1 = \mathbf{A}_1\sqrt{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) + o_p(1).$$

The asymptotic linear expansion of the case-cohort estimator $\widehat{\boldsymbol{\beta}}_g$ after plugging in $\widehat{\boldsymbol{\gamma}}$ requires some algebra, an example of which is provided in Section 2.3.1 next. Suppose $\mathbf{U}_2$ is the estimating equation for $\widehat{\boldsymbol{\beta}}_g$, which it can be written as its asymptotically equivalent term $\sum \mathbf{W}_i$, the sum of the iid score contributions. We define $W_i = 0$ for controls that are not included in the case-cohort sample. Let $\mathbf{A}_2 = \lim -\frac{1}{n}\partial\mathbf{U}_2/\partial\boldsymbol{\beta}_g$ and $\mathbf{A}_3 = \lim -\frac{1}{n}\partial\mathbf{U}_2/\partial\boldsymbol{\gamma}$. The first-order Taylor expansion of $\mathbf{U}_2$ at both $\boldsymbol{\beta}_g$ and $\boldsymbol{\gamma}$ yields

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathbf{U}_2 &= \frac{1}{\sqrt{n}}\sum \mathbf{W}_i + o_p(1)\\ &= \mathbf{A}_2\sqrt{n}(\widehat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g) + \mathbf{A}_3\sqrt{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) + o_p(1).\end{aligned}$$

By the central limit theorem and under mild regularity conditions, $\sqrt{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) \to_d \mathcal{N}(0,\Sigma_1)$, where Σ_1 is its asymptotic variance matrix $\mathbb{E}(\mathbf{A}_1^{-1}\mathbf{U}_{1i}\mathbf{U}_{1i}^T\mathbf{A}_1^{-1})$, the robust variance estimator. Similarly, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g) \to_d \mathcal{N}(0,\Sigma_2)$, where the asymptotic variance matrix $\Sigma_2 = \mathbb{E}\{\mathbf{A}_2^{-1}(\mathbf{W}_i\mathbf{W}_i^T + \mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{U}_{1i}\mathbf{U}_{1i}^T\mathbf{A}_1^{-1}\mathbf{A}_3^T - \mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{U}_{1i}\mathbf{W}_i^T - \mathbf{W}_i\mathbf{U}_{1i}^T\mathbf{A}_1^{-1}\mathbf{A}_3^T)\mathbf{A}_2^{-1}\}$.

2.3.1. The Self-Prentice estimator. The asymptotic linear expansions for various case-cohort estimators were presented in respective works (Lin and Wei, 1989; Binder, 1992; Lin and Ying, 1993; Lin, 2000; Borgan et al., 2000). Here, we write out the expressions for the Self-Prentice estimator and leave the survey estimator using IPW to the Appendix. Both estimators are implemented in the R packages `cch` and `Survey`, two popular softwares for analyzing case-cohort data.

In our notation, the estimating function for the Self-Prentice estimator after plugging in the case-only estimators $(\hat{\beta}_{2co},\hat{\beta}_{3co})$ can be written as

$$\mathbf{U}_2 = \sum \mathbf{U}_{2i} = \sum \Delta_i \left\{\mathbf{X}_{2i} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}_g,T_i;\hat{\boldsymbol{\gamma}})}{\mathbf{S}^{(0)}(\boldsymbol{\beta}_g,T_i;\hat{\boldsymbol{\gamma}})}\right\}, \tag{8}$$

where $\mathbf{X}_{2i} = (G_i,V_i)^T$, $\mathbf{X}_{1i} = (Z_i,G_iZ_i)^T$, and

$$\mathbf{S}^{(r)}(\boldsymbol{\beta}_g,T_i;\hat{\boldsymbol{\gamma}}) = \frac{1}{n_{sc}}\sum_{i\in\mathcal{S}} R_i(t)\exp(\hat{\boldsymbol{\gamma}}^T\mathbf{X}_{1i} + \boldsymbol{\beta}_g^T\mathbf{X}_{2i})\mathbf{X}_{2i}^{\otimes r} \tag{9}$$

for $r=0,1$, where n_{sc} is the sample size of the random subcohort.

The asymptotic linearization of the Self-Prentice estimator can be expressed as $\mathbf{A}_2^{-1}\mathbf{W}_i$ (Lin and Wei, 1989; Lin and Ying,

1993), where $\mathbf{A}_2 = \lim -(1/n)(\partial \mathbf{U}_2/\partial \boldsymbol{\beta}_g)$,

$$\mathbf{W}_i = \mathbf{U}_{2i} - \sum_{l=1}^{n} \frac{\Delta_l R_i(T_l) I(i \in \mathcal{S}) \exp(\hat{\boldsymbol{\gamma}}^T \mathbf{X}_{1i} + \boldsymbol{\beta}_g^T \mathbf{X}_{2i})}{n_{sc}(\boldsymbol{\beta}_g, T_l; \hat{\gamma})} \times \left\{ \mathbf{X}_{2i} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}_g, T_l; \hat{\boldsymbol{\gamma}})}{\mathbf{S}^{(0)}(\boldsymbol{\beta}_g, T_l; \hat{\boldsymbol{\gamma}})} \right\}.$$

2.4. **Scenario II:** *Augmented Case-Only (ACO) Design for Vaccine Trials*

In the proposed ACO sampling design, the genotype is ascertained for a random subcohort from the active treatment arm only, denoted by $\mathcal{S}_1$, and all additional participants who developed the clinical outcome outside of $\mathcal{S}_1$. The set of sampled participants is, therefore, defined by $\{i : \Delta_i = 1 \text{ or } i \in \mathcal{S}_1\}$. Though the controls from the placebo arm are not sampled, we show next that a similar 3-step procedure estimates all parameters in the Cox model (1).

First, case-only estimators of β_2 and β_3 are obtained from (6). Second, based on the case-cohort sample in the active treatment arm, we estimate the parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$ in the Cox model

$$\lambda(t; Z = 1, G, V) = \lambda_0^*(t) \exp(\alpha_1 G + \alpha_2 V) \tag{10}$$

by standard case-cohort methods (Prentice, 1986; Self and Prentice, 1988; Lin and Ying, 1993; Barlow, 1994), where parameterization in models (1) and (10) dictate that $\alpha_1 = \beta_1 + \beta_3$, $\alpha_2 \equiv \beta_4$, and $\lambda_0^*(t) = \lambda_0(t) \exp(\beta_2)$. The estimated $\lambda_0^*(t)$ can be obtained from the Breslow estimator based on the estimate of $\boldsymbol{\alpha}$ as previously described (Prentice, 1986; Self and Prentice, 1988). Third, we compute the estimators of $(\beta_1, \lambda_0(t))$ using the estimators obtained in previous steps as follows:

$$\hat{\beta}_1 = \hat{\alpha}_1 - \hat{\beta}_3$$
$$\hat{\lambda}_0(t) = \frac{\hat{\lambda}_0^*(t)}{\exp(\hat{\beta}_2)}.$$

The full set of parameters in a Cox model are, therefore, estimated.

Since, the estimation of $(\hat{\alpha}_1, \hat{\alpha}_2)$ and $(\hat{\beta}_{2co}, \hat{\beta}_{3co})$ both used data from cases in the active treatment arm, the two sets of estimators are correlated. We estimate, the covariance matrix by the general estimating equation theory. As shown in Section 2.3, both estimators can be written as asymptotically linear estimators (Newey and Powell, 1990; Robins et al., 1994), that is, their asymptotic distribution can be expressed as $n^{-1/2} \sum_{i=1}^{n} B_i + o_p(1)$, where $B_i = \mathbf{A}^{-1}\mathbf{U}_i$ is the iid influence function from each subject, $\mathbf{A}$ is the expected information matrix and U_i is iid estimating function. Expressions of $\mathbf{A}$ and U_i for $(\hat{\alpha}_1, \hat{\alpha}_2)$ and $(\hat{\beta}_{2co}, \hat{\beta}_{3co})$ follow those in Section 2.3. Suppose the influence function for the case-cohort estimator $(\hat{\alpha}_1, \hat{\alpha}_2)$ is B_1, and suppose the influence function for the case-only estimator $(\hat{\beta}_2, \hat{\beta}_3)$ is B_2. Then by the central limit theorem, the covariance between $(\hat{\alpha}_1, \hat{\alpha}_2)$ and $(\hat{\beta}_2, \hat{\beta}_3)$ can be estimated as $\sum B_{1i}^T B_{2i}$.

2.4.1. The alternative ACO design and nested case-control design. For completeness, we briefly address an alternative ACO design, in which the subcohort is instead taken only from the placebo arm. Such design may be merely of theoretical interest to compare its efficiency to the ACO discussed in Section 2.4, as we will show in simulations. The estimation is simplified: the first step is the same, and in the second step β_1, β_4 and λ_0 are directly estimated using the case-cohort data in the placebo arm.

Frequently used in the biomarker research, the nested case-control design randomly selects a fraction of controls in the risk set at each failure time. This design is particularly suitable for time-varying biomarkers. Little added efficiency can be realized when selecting more than 5 controls per case (Breslow et al., 1983). Our augmented case-only design can be similarly modified by using all cases plus a nested case-control sample from one of the two arms. The estimation procedure follows closely to those in Section 2.2, but using the conditional logistic regression or the IPW partial likelihood method in the second step (Goldstein and Langholz, 1992; Samuelsen, 1997), possibly with stratification (Langholz and Borgan, 1995). The asymptotic linearization of these estimators required for estimating the covariance matrix can be found in the respective literature.

3. Simulation

The performance of the proposed estimation and designs is evaluated in a simulation study with 1000 simulated datasets, using the standard estimation for the case-cohort design and the full cohort analysis as benchmarks. For **Scenario I** where standard case-cohort sampling has taken place, the interest is to evaluate how much efficiency is gained for the genetic main effect when we incorporate the case-only estimators into case-cohort estimation. For **Scenario II**, the interest is to compare efficiency of the main effect estimator under different designs, adding controls in the active treatment arm only or adding controls by other ways, all of which use a similar hybrid estimation procedure that incorporated the case-only estimator.

Across all scenarios the sample size in the trial is 3000, and the participants are randomized in a 1:1 ratio to the active treatment arm or the placebo arm. The genotype G is assumed to be depending on V, a baseline covariate following a Bernoulli distribution with rate 0.5, as $\text{logit}\{\Pr(G = 1)\} = -1.6 + 1.4V$. The rate of variant allele is around 0.3. The cumulative probability of incident cases is set to be around 0.05. The event time is exponentially distributed, with a constant baseline hazard function of $\lambda_0(t) = 1$. The true regression parameters associated with the set of covariates are listed in Table 1 and 2, with the parameter associated with V set to $\log(1.5)$. The censoring time is exponentially distributed with mean 1, independent of the event time. Administrative censoring is set such that the cumulative event rate is around 5%. The ACO designs consist of a random subcohort of varying sample sizes in the active treatment arm only or in the placebo arm only, plus all cases outside the subcohort. The standard case-cohort design was devised to have almost identical sample size as the ACO designs in each simulated dataset, though the subcohort is drawn randomly from the entire trial population.

Table 1
Small-sample properties of estimators for the proposed designs and estimations in the Cox model (1) with a varying subcohort fraction

SC Fraction		$\beta_1=\beta_2=\beta_3=0$					$\beta_1=-\beta_2=\beta_3=\log 1.5$					$\beta_1=-\beta_2=\beta_3=\log 2$				
		β_{1a}	β_{1b}	β_{1c}	β_2	β_3	β_{1a}	β_{1b}	β_{1c}	β_2	β_3	β_{1a}	β_{1b}	β_{1c}	β_2	β_3
10%	Bias	0	0.005	−0.004	−0.006	−0.012	0.016	0.025	0.010	0.001	−0.017	0.020	0.034	0.012	0.001	−0.014
	Var	0.081	0.084	0.083	0.043	0.123	0.073	0.076	0.074	0.058	0.110	0.070	0.072	0.071	0.074	0.116
	$\widehat{\text{Var}}$	0.083	0.086	0.083	0.041	0.124	0.071	0.076	0.071	0.054	0.112	0.069	0.074	0.068	0.072	0.121
	CP	0.954	0.960	0.947	0.957	0.959	0.954	0.954	0.940	0.947	0.961	0.953	0.956	0.945	0.959	0.965
15%	Bias	−0.009	−0.009	−0.014	−0.003	−0.005	0.010	0.013	0.003	0.006	−0.014	0.016	0.023	0.008	0.003	−0.010
	Var	0.077	0.082	0.079	0.041	0.127	0.066	0.071	0.069	0.056	0.110	0.064	0.069	0.065	0.074	0.117
	$\widehat{\text{Var}}$	0.076	0.080	0.077	0.041	0.125	0.065	0.069	0.065	0.054	0.112	0.063	0.068	0.062	0.072	0.121
	CP	0.947	0.950	0.946	0.959	0.957	0.960	0.955	0.938	0.951	0.964	0.954	0.949	0.943	0.956	0.964
20%	Bias	−0.012	−0.010	−0.014	−0.002	−0.003	0.006	0.011	0.002	0.006	−0.011	0.011	0.019	0.006	0.005	−0.011
	Var	0.076	0.081	0.078	0.041	0.131	0.065	0.071	0.067	0.057	0.112	0.062	0.068	0.063	0.074	0.118
	$\widehat{\text{Var}}$	0.073	0.077	0.074	0.041	0.125	0.061	0.066	0.062	0.054	0.112	0.059	0.065	0.059	0.072	0.120
	CP	0.945	0.946	0.944	0.958	0.951	0.953	0.947	0.939	0.955	0.960	0.949	0.943	0.939	0.950	0.961
25%	Bias	−0.010	−0.009	−0.009	0.002	−0.005	0.010	0.015	0.008	0.011	−0.018	0.013	0.021	0.010	0.013	−0.017
	Var	0.073	0.078	0.074	0.041	0.130	0.062	0.067	0.063	0.055	0.108	0.059	0.065	0.060	0.069	0.113
	$\widehat{\text{Var}}$	0.071	0.075	0.072	0.041	0.125	0.059	0.064	0.060	0.054	0.112	0.057	0.063	0.057	0.072	0.120
	CP	0.954	0.950	0.947	0.956	0.954	0.956	0.945	0.948	0.958	0.960	0.954	0.943	0.947	0.954	0.961

Notation: SC Fraction, subcohort sampling fraction of the entire cohort; CP, 95% coverage probability; β_{1a}, β_{1b} and β_{1c}, the Self-Prentice estimator used to estimate β_1 in the standard case-cohort design incorporating the case-only estimators, the augmented case-only design while sampling controls from the active treatment arm only, or the augmented case-only design with controls from the placebo arm only.

Table 1 shows the small-sample properties of the proposed hybrid estimation procedures for incorporating case-only estimators into the case-cohort estimation (β_{1a}), and for the ACO design sampling controls from the active arm only (β_{1b}) or from the placebo arm only (β_{1c}). The Self-Prentice estimator was used in the second step for all three designs. The IPW estimator performs very closely to the Self-Prentice estimator and, thus, is omitted from this table and Table 2. Under the null hypothesis and under the moderate effect size for all three parameters, all estimators appear to be consistent as the biases of the estimators are all small relative to their empirical variability. The ACO design with controls from active arm tends to have bigger bias under the alternative. The estimated variances agree well with the empirical variances, and the coverage probabilities of 95% confidence intervals behave properly. Similar performance was observed for the simulations with a qualitative interaction model (Supplementary materials). We conclude that the hybrid estimation procedures detailed in Section 2.3 and 2.4 work well in the simulated datasets.

Table 2 shows the efficiency of the standard case-cohort design (with or without incorporating the case-only estimators) and the two ACO designs, relative to the full cohort analysis. The relative efficiency is calculated as the ratio of the sample variance of parameters estimated from the two designs. The results suggest that all designs incorporating case-only estimators lead to a major efficiency gain for all three parameters relative to the standard case-cohort estimation. The efficiency gains on β_2 and β_3 are not surprising as they are the case-only estimators (Vittinghoff and Bauer, 2006). More interestingly, over 10% efficiency gain is realized for the genetic main effect β_1 when the case-only estimators are incorporated into the case-cohort analysis (**Scenario I**), or when the random sampling fraction is 5-15% in the ACO designs (**Scenario II**). Sampling controls from both arms appears to outperform sampling controls from one arm only in estimating the genetic main effect: compared to the standard case-cohort design and analysis, the ACO design with controls from the active arm gains 5-15% efficiency due to the use of the case-only estimator, depending on the subcohort fraction; additional 5-10% can be gained by allocating the controls in both arms, a design benefit given the case-only estimator has been exploited. The ACO design with controls from the placebo arm only outperforms the ACO design with controls from the active arm only, presumably because the former has a simpler estimation procedure. However, when potential systematic studies such as immune correlates are of interest, the ACO design with controls from the active treatment arm may still be more cost-effective.

4. Data Application

We show a pedagogical example in HIV vaccine trials with a standard case-cohort sampling scheme. We analyzed the case-cohort data in four ways: standard case-cohort estimation, case-cohort estimation incorporating the case-only estimators (Section 2.3), and augmenting case-only data with controls from the vaccine arm only (Section 2.4). All estimators are compared to the standard case-cohort analysis in terms of standard errors for the genetic main effect.

Table 2
The efficiency of the proposed designs and estimations in the Cox model (1) with varying subcohort fraction, when compared to the full cohort analysis

SC Fraction		$\beta_1=\beta_2=\beta_3=\beta_4=0$			$\beta_1=-\beta_2=\beta_3=\log 1.5$			$\beta_1=-\beta_2=\beta_3=\log 2$		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
10%	Case-cohort	0.659	0.678	0.644	0.626	0.744	0.625	0.605	0.787	0.641
	Case-cohort + case-only	0.812	0.995	1.001	0.775	0.992	1.017	0.761	0.997	1.025
	ACO active	0.780	0.995	1.001	0.743	0.992	1.017	0.735	0.997	1.025
	ACO placebo	0.788	0.995	1.001	0.760	0.992	1.017	0.751	0.997	1.025
15%	Case-cohort	0.767	0.765	0.738	0.737	0.806	0.707	0.723	0.834	0.726
	Case-cohort + case-only	0.894	0.993	0.993	0.874	0.986	1.003	0.865	0.990	1.013
	ACO active	0.839	0.993	0.993	0.813	0.986	1.003	0.797	0.990	1.013
	ACO placebo	0.869	0.993	0.993	0.846	0.986	1.003	0.843	0.990	1.013
20%	Case-cohort	0.820	0.856	0.789	0.808	0.883	0.777	0.792	0.908	0.793
	Case-cohort + case-only	0.920	0.994	0.984	0.907	0.982	0.989	0.899	0.988	1.002
	ACO active	0.866	0.994	0.984	0.836	0.982	0.989	0.816	0.988	1.002
	ACO placebo	0.897	0.994	0.984	0.885	0.982	0.989	0.881	0.988	1.002
25%	Case-cohort	0.871	0.882	0.839	0.854	0.906	0.827	0.839	0.930	0.840
	Case-cohort + case-only	0.948	0.998	0.984	0.931	0.983	0.990	0.926	0.990	1.004
	ACO active	0.883	0.998	0.984	0.855	0.983	0.990	0.839	0.990	1.004
	ACO placebo	0.924	0.998	0.984	0.913	0.983	0.990	0.914	0.990	1.004

Notation: SC Fraction, subcohort sampling fraction of the entire cohort; Case-cohort, the standard case-cohort design and estimation with controls from both arms; Case-cohort + case-only, the case-cohort design incorporating the case-only estimators; ACO active, the augmented case-only design with controls from the active arm only; ACO placebo, the augmented case-only design with controls from the placebo arm only.

The Step trial, a test-of-concept study that evaluates the protection of a cell-mediated immune vaccine, was prematurely terminated because the risk of HIV infection was evidently elevated in the vaccine arm compared to the placebo arm (Buchbinder et al., 2008). In order to understand this disappointing result, a host genetics study was conducted to assess the association of several immune genes, namely GM, KM, and FcγR, with HIV infection and the vaccine effect. A case-cohort sample including 25% of study participants was pre-selected for storing blood samples, together with blood samples taken later from incident cases, and measuring immunogenecity (McElrath et al., 2008). The analysis of this genetic study has been reported elsewhere, and no significant gene-treatment interaction was found possibly due to the small sample size (Pandey et al., 2013). Following the strategy in Pandey et al. (2013), all four analyses were restricted to white males and the same set of covariates were adjusted for.

Table 3 shows the estimates for various designs and estimation procedures for the genetic variant in FcγR-2, coded by an additive genetic score (0/1/2). In the last three analyses, β_2 and β_3 were estimated by the case-only method, and thus all have smaller standard errors than the standard case-cohort estimates. The genetic main effect β_1 was estimated using Self-Prentice estimator in all analysis. The standard error of the ACO using only 60 controls from the vaccine arm is similar to that from the standard case-cohort estimation with all 169 controls. The standard error of the case-cohort analysis incorporating case-only estimators is even smaller because more controls were used.

Scientifically speaking, it is more cost-effective to store biology specimens of samples in the vaccine arm. If there was any genotype showing significant interactions with the vaccine, investigators could directly correlate the genotype with immune responses using samples in the vaccine arm (McElrath et al.,

Table 3
Comparison of various study designs and estimation procedures in a case-cohort genetic study in the STEP trial

	# cases	# controls	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)
Case-cohort	56	169	0.17 (0.40)	1.05 (0.66)	−0.32 (0.51)
Case-only	56	0	–	0.63 (0.49)	−0.26 (0.38)
Case-cohort + case-only	56	169	0.19 (0.31)	0.63 (0.49)	−0.26 (0.38)
ACO active	56	60	0.24 (0.38)	0.63 (0.49)	−0.26 (0.38)

Notation: Case-cohort + case-only, the case-cohort design incorporating the case-only estimators; ACO active, the augmented case-only design using controls from the vaccine arm

2008), to further explore the mechanism of differential vaccine protection.

5. Discussion

In prevention clinical trials with failure time endpoints, we investigated several ways of augmenting the case-only sampling design for studying the influence of pre-treatment biomarkers, such as genotypes, on treatment effect and the risk of the failure event. The goal is to be able to estimate all parameters in a Cox model, not just subgroup effects and the interaction, so that absolute risk can be computed. One way is to incorporate the case-only estimators into case-cohort estimation (**Scenario I**). We showed that such estimators and their variance estimates can be obtained by a hybrid estimation procedure. Motivated by vaccine trials, we also propose an augmented case-only design which builds on the efficient case-only method and adds controls from the active treatment arm only (**Scenario II**). Following a similar hybrid procedure, we showed that all parameters in a Cox model and the absolute risk can be estimated. Simulation results showed a sizable efficiency gain in estimating the genetic main effect over the standard case-cohort design, because of incorporating case-only estimators and exploiting gene-treatment independence. It is worthwhile to reiterate that the motivation of (**Scenario II**) is driven by scientific and cost-effective use of vaccine trial resources, not estimation efficiency, as the simulation showed that allocating controls in both arms achieved a better efficiency.

The assumptions required by case-only estimators can be restrictive. The rare disease and nondifferential censoring between arms may exclude many cancer therapeutic trials. The applications we consider in this article are primarily phase III prevention trials with a rare endpoint, e.g., HIV vaccine trials. These trials enroll healthy participants and evaluate the prevention effect of certain modality which should not induce severe adverse effects.

The benefit of concentrating controls in the active treatment arm is to have a bigger pool of active treatment recipients for systematically measuring both genotypes and a comprehensive profile of biological mediators. Ancillary studies of this nature are increasingly common in clinical trials (Pitteri et al., 2009; Li et al., 2014). The controls can be sampled from a random subcohort or sampled repeatedly from risk sets as in nested case-control sampling. For the ACO design, the efficiency of the estimators of the genetic main effect and the covariate can be further improved by using more efficient estimators of α_1 and α_2 in (10), e.g., the method of efficient score equations (Nan, 2004), though the variance of the resulting two-step estimators can be difficult to derive.

In the same vein, ACO designs can be devised for clinical trials with binary endpoints, where logistic regression models are used for analysis. For rare endpoints, the two-step estimation procedure proposed in this article applies with minor modification. The efficiency comparison with respect to the standard logistic regression under case-control sampling design is a bit complicated, since gene-treatment independence can also be exploited in the maximum likelihood estimation (Dai et al., 2009). Furthermore, it is not clear whether the ACO design can be extended to common endpoints, where the case-only estimator is no longer applicable. Use of additive interaction is another topic of interest, as it may be better to inform public health impact of gene-treatment interaction and the biological inter-dependence. Gene-treatment interaction has been exploited to improve efficiency (Han et al., 2012). But, it is not clear whether one can estimate parameters in an additive interaction model using a ACO design. We will pursue these topics in future work.

6. Supplementary Materials

Web Table referenced in Section 3 and the sample code to implement the estimation method are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This work was supported by the National Institutes of Health grants P01 CA53996, R01 HL114901, R01 HG006164, R01 ES017030 and R21 HL121347. The authors thanks two reviewers and the Associate Editor for their constructive comments.

References

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064–1072.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–47.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6**, 39–58.

Breslow, N. E., Lubin, J. H., Marek, P., and Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association* **78**, 1–12.

Breslow, N. E., Robins, J. M., and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome-dependent sampling. *Annals of Statistics* **31**, 1110–1139.

Buchbinder, S. P., Mehrotra, D. V., Duerr, A., Fitzgerald, D. W., Mogg, R., Li, D., et al. (2008). Efficacy assessment of a cell-mediated immunity hiv-1 vaccine (the step study): A double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet* **372**, 1881–1893.

Charlab, R. and Zhang, L. (2013). Pharmacogenomics: Historical persective and current status. *Methods in Molecular Biology* **1015**, 3–22.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Dai, J. Y., Kooperberg, C., LeBlanc, M., and Prentice, R. L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929–944.

Dai, J. Y., LeBlanc, M., and Koopberg, C. (2009). Semiparametric estimation exploiting covariate independence in two-phase randomized clinical trials. *Biometrics* **65**, 178–187.

Dai, J. Y., Li, S. S., and Gilbert, P. B. (2014). Case-only methods for competing risks models with application to assessing differential vaccine efficacy by viral and host genetics. *Biostatistics* **15(1)**, 196–203.

Dai, J. Y., Logsdon, B. A., Huang, Y., Hsu, L., Reiner, A. P., Prentice, R. L., et al. (2012). Simultaneously testing for

marginal genetic association and gene-environment interaction. *American Journal of Epidemiology* **176**, 164–173.

Evans, W. E. and McLeod, H. L. (2003). Pharmacogenomics- drug disposition, drug targets, and side effects. *The New England Journal of Medicine* **348**, 538–549.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., et al. (1989). Projecting individualized probabilities of developing breast cancer for while females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879–1886.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the cox regression model. *Annals of Statistics* **20**, 1903–1928.

Han, S. S., Rosenberg, P. S., Garcia-Closas, M., Figueroa, J. D., Silverman, D., Chanock, S. J., et al. (2012). Likelihood ratio test for detecting gene (g)-environment (e) interactions under an additive risk model exploiting g-e independence for case-control data. *American Journal of Epidemiology* **176**, 1060–7.

Janes, H., Pepe, M. S., Bossuyt, P. M., and Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine* **154**, 253–259.

Langholz, B. and Borgan, Y. (1995). Counter-matching: a stratified nested case-control sampling method. *Biometrika* **82**, 69–79.

Langholz, B. and Thomas, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology* **131**, 169–176.

Li, S. S., Gilbert, P. B., Tomaras, G. D., Kijak, G., Ferrari, G., Thomas, R., et al. (2014). Fcgr2c polyporphisms associate with hiv-1 vaccine protection in rv144 trial. *Cancer Epidemiology, Biomarkers & Prevention* **124**, 3879–3890.

Lin, D. Y. (2000). On fitting cox's proportional hazards models to survey data. *Biometrika* **87**, 37–47.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal American Statistical Association* **84**, 1074–1078.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal American Statistical Association* **88**, 1341–1349.

McElrath, M. J., De Rosa, S. C., Moodie, Z., Dubey, S., Kierstead, L., Janes, H., et al. (2008). Hiv-1 vaccine-induced immunity in the test-of-concept step study: a case-cohort analysis. *Lancet* **372**, 1894–1905.

Murphy, K. M. and Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics* **3**, 370–379.

Nan, B. (2004). Efficient estimation for case-cohort studies. *Canadian Journal of Statistics* **32**, 403–419.

Newey, W. K. and Powell, J. (1990). Efficient estimation of linear and type i censored regression models under conditional quantile restrictions. *Econometric Theory* **6**, 295–317.

Pandey, J. P., Namboodiri, A. M., Bu, S., Tapsoba, J. D., Sato, A., and Dai, J. Y. (2013). Immunoglobulin genes and the acquisition of hiv infection in a randomized trail of recombinant adenovirus hiv vaccine. *Virology* **441**, 70–4.

Pitteri, S. J., Hanash, S. H., Aragaki, A., Amon, L. M., Chen, L., Buson, T. B., et al. (2009). Postmenopausal estrogen and progestin effects on the serum proteome. *Genome Medicine* **1(12)**, 121.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.

Prentice, R. L., Huang, Y., Hinds, D. A., Peters, U., Cox, D. R., Beilharz, E., et al. (2010). Variation in the fgfr2 gene and the effect of a low-fat dietary pattern on invasive breast cancer. *Cancer Epidemiology, Biomarkers & Prevention* **19**, 74–9.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* **89**, 846–866.

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84**, 379–394.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* **16**, 64–81.

Tchetgen, E. J. and Robins, J. (2010). The semiparametric case-only estimator. *Biometrics* **66**, 1138–1144.

Therneau, T. M. and Li, H. (1999). Computing the cox model for case-cohort designs. *Lifetime Data Analysis* **5**, 99–112.

Thomas, D. C. (1977). Addendum to "methods of cohort analysis: appraisal by application to asbestos mining". *Journal of Royal Statistical Society, Serial A* **140,** 483–485.

Vittinghoff, E. and Bauer, D. C. (2006). Case-only analysis of treatment-covariate interactions in clinical trials. *Biometrics* **62**, 769–776.

Weinshilboum, R. and Wang, L. (2004). Pharmacogenomics: bench to bedside. *Nature Reviews Drug Discovery* **3**, 739–748.

Received November 2014. *Revised July* 2015. *Accepted July* 2015.

Appendix

The asymptotic linear expansion of the IPW estimator for case-cohort sampling

For computing the expected covariate values at each event time, those at-risk cases occurring outside of the random subcohort can be used with proper sampling weights. Specifically, in estimating function (8), the average term (9) is replaced by

$$\mathbf{S}^{(r)}(\boldsymbol{\beta}_g, T_i) = \frac{1}{n}\sum_{i\in\mathcal{S}\cup\mathcal{D}} \frac{1}{\pi_i} R_i(t)\exp(\hat{\boldsymbol{\gamma}}^T\mathbf{X}_{1i} + \boldsymbol{\beta}_g^T\mathbf{X}_{2i})\mathbf{X}_{2i}^{\otimes r},$$

where

$$\pi_i = \begin{cases} \frac{\sum I(\Delta_j=0, j\in\mathcal{S})}{\sum I(\Delta_j=0)} & \text{if } i \in \mathcal{S} \text{ and } \Delta_i = 0 \\ 1 & \text{if i} \in \mathcal{D} \end{cases},$$

and $\mathcal{D}$ is the set of cases.

The asymptotic expansion for the survey estimator using the inverse probability weights is modified as $B_{2i} = \mathbf{A}_2^{-1}\mathbf{W}_i$, where $\mathbf{A}_2 = \lim -(1/n)(\partial\mathbf{U}_2/\partial\boldsymbol{\beta}_g)$, and

$$\mathbf{W}_i = \mathbf{U}_{2i} - \sum_{l\in\mathcal{S}\cup\mathcal{D}} \frac{\frac{1}{\pi_l}\Delta_l R_i(T_l) I(i\in\mathcal{S}\cup\mathcal{D})\exp(\hat{\boldsymbol{\gamma}}^T\mathbf{X}_{1i} + \boldsymbol{\beta}_g^T\mathbf{X}_{2i})}{n\mathbf{S}^{(0)}(\boldsymbol{\beta}_g, T_l; \hat{\gamma})} \left\{\mathbf{X}_{2i} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}_g, T_l; \hat{\boldsymbol{\gamma}})}{\mathbf{S}^{(0)}(\boldsymbol{\beta}_g, T_l; \hat{\boldsymbol{\gamma}})}\right\}.$$

The computation of the covariance matrix of $\hat{\boldsymbol{\beta}}_g$ follows similarly as in Section 2.3.