Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 2, pp. 37-54

A RELATIVELY SIMPLE EFFICIENT ESTIMATOR FOR RELATIVE RISK IN CASE-COHORT STUDIES

EMMANUEL SAMPENE
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison
Email: sampene@biostat.wisc.edu

ABDUS S. WAHED
Department of Biostatistics, University of Pittsburgh, PA, USA
Email: WahedA@edc.pitt.edu

SUMMARY

A case-cohort study is a two-phase study where at the first phase a representative sample, referred to as the study cohort, is selected from the target population. At the second phase, a subsample is selected from the study cohort based on the case status. All cases are included in the subsample whereas only a random sample of controls is included. The endpoint of interest in such studies is usually the failure time. Several methods have been proposed to estimate the relative risk or hazard ratio from a case-cohort study. Many of these methods disregard the covariate information that is not included in the sampled study sub-cohort, and therefore, results in the loss of efficiency. While there have been attempts to derive the most efficient estimators, the resulting estimators are not easily implemented from the data analysis point of view. We propose a locally efficient estimator (LEE) based on Robins et al. (1994, J. AM Stat. Assoc. 89, 846-866) by restricting the estimator to a class of regular asymptotically linear estimators. The properties of this estimator are investigated through simulations, and application to the Wilm's tumor study.

Keywords and phrases: Case-cohort study; Influence Function; Martingale theory; Regular asymptotically linear estimator; Two-stage design

1 Introduction

Standard prospective cohort designs require assembly of all covariate (exposure) histories. In such studies, participants are followed prospectively, and subsequent status evaluations with respect to a disease or outcome are ascertained for exposure characteristics. Although subjects in a cohort design can be matched, which limits the influence of confounding variables, and since patients are followed prospectively for longer periods of time, the ascertainment of outcomes in this design is often delayed; as a result, this type of studies are generally expensive. Cohort studies typically measure exposure in many controls. This measurement of exposures in controls are wasteful and

thus inefficient. In addition, cohort designs require extended follow-up to observe the development of the condition of interest, e.g. death due to lung cancer.

Prentice (1986) introduced the case-cohort design as an economical way of studying large cohorts; this type of design is widely used in epidemiological studies with time-to-event data. The case-cohort study is a two-phase study where at the first phase, a representative sample referred to as the study cohort is selected from the target population. In practice, certain covariates such as treatment allocations, gender, age, and surrogate measurements of expensive covariates are obtained in all subjects in the cohort (Kulich & Lin, 2004). Considering that evaluation of all covariates on this cohort could be expensive, this list of covariates is usually kept to a minimum. At the second phase, a subsample is selected from the cohort based on the case status, and the most expensive covariates that were not measured in the first phase are evaluated for the subsample. All cases are included in the subsample, whereas only random samples of controls are included in the subcohort. The endpoint of interest in such studies is usually the failure time, e.g. time to death. Rothman (2002) described the advantages of using the case-cohort design. This design is very efficient, since controls can be used in all risk sets for which they qualify. Furthermore, this type of design is very flexible, and allows testing hypotheses that were not anticipated when the cohort was drawn from the subsample. That is, the subsample can be used to study multiple outcomes.

Analysis of case-cohort studies is very similar to the usual Cox regression approach with a few modifications. It is assumed that, if we had full data, then the standard Cox proportional hazard model would suffice. Prentice (1986) showed how to estimate the relative risk from a Cox proportional hazard model without necessarily obtaining the covariate information of all subjects in the cohort. His method used an estimating equation for the regression parameters through a pseudolikelihood approach that weighted the contributions from the cases and subsample members using the inverses of their true or estimated sampling probabilities.

Since Prentice's initial work, many authors have derived variations of his method by proposing different estimating equations. Most of these methods, however, fail to account for the covariate data collected outside the case-cohort sample, and hence incur loss of efficiency. In particular, the Kalbfleisch & Lawless (1988) estimator, which is based on the modifications of the full data partial likelihood score function, weights the contributions from the cases and subsample members with inverses of their true or estimated probabilities, called sampled fraction. However, it ignores the first phase covariate data. Similarly, the estimator proposed by Self & Prentice (1988) ignores all the first phase information. Only Borgan & Goldstein et al. (1995) method utilizes some of the first phase information by stratifying on the first phase covariates (Kulich & Lin, 2004).

To improve the efficiency of the case-cohort estimators, several authors have introduced different estimators. For example, Barlow (1994) introduced an estimator that incorporates time-varying weights by proposing different weighting schemes. This weighting scheme assigned a value of one for all cases inside or outside the subsample. On the contrary, the cases in the subsample prior to failure, and the subsample controls are weighted by the inverses of the sampling fraction. Also, Chen & Lo (1999) proposed an estimator based on the partial likelihood score function that improves the efficiency gain within the doubly weighted class of estimators that is stratified on all combinations of binary covariates (Kulich & Lin, 2004).

The efficiency of the relative risk estimation in analyzing case-cohort studies was further improved by Borgan & Langholz et al. (2000) where they proposed an exposure stratified case-cohort estimator, whereby complete covariate information is assembled for all failures (cases) and a stratified random subsample of the non-failures (controls). The stratification is usually based on an inexpensive easily observable covariate that is measured on all members in the cohort. By this approach, Borgan & Langholz et al. (2000) showed that stratified sampling designs can lead to substantial efficiency gain in case-cohort studies by employing the weighted versions of the pseudolikelihood methods. In addition, Kulich & Lin (2004) introduced the combined doubly weighted estimator (CDW), which is a linear combination of the class of augmented estimators described in Robins et al. (1994) and Borgan & Goldstein (1995) estimators. In their 2004 paper, Kulich and Lin compared their combined doubly weighted (CDW) estimator to the Borgan II estimator, hereafter referred to as BII-estimator. The BII-estimator is one of the most efficient estimators within the class of doubly weighted estimators. However, the BII estimator did not perform well against the CDW estimator in a head to head comparison. According to Kulich and Lin (2004, page 838), estimates of the CDW estimator for fully observed binary covariates show smaller efficiency gains compared with BII, but have the highest overall efficiency. They show that the efficiency gain of CDW for incompletely observed continuous covariates ranges from substantial to negligible, depending on the quality of the surrogate information. Thus, the CDW estimator is protected against a deterioration of efficiency below that of the Borgan et. al. estimators due to an incorrectly specified model (Kulich & Lin, 2004). Mark & Katki (2006) proposed a simple estimator based on inverse-probability weighting (Horvitz, 1952), which they refer to as the $\hat{\alpha}$-estimator, reflecting that the probability of being included in the sample, α, is estimated. In addition, Nan (2004) proposed an estimator for case-cohort designs with discrete covariates by solving the efficient score equations using a one-step Newton-Raphson approximation.

All the aforementioned estimators seek to efficiently estimate the association parameters from a case-cohort design by using the Cox proportional hazard model. Although the CDW estimator has appealing asymptotic properties, it may not always perform well in finite samples. It is computationally complex and difficult to implement. In addition, the efficiency gain of the CDW estimator depends on whether a fully observed continuous or binary covariate is observed at baseline. That is, fully observed continuous covariates achieve more efficiency compared to fully observed binary covariates.

The $\hat{\alpha}$-estimator proposed by Mark & Katki (2006) relaxes the requirement that the selection of subjects into the subsample be independent with known probabilities. It is similar to the usual inverse-probability weighted Horvitz-Thompson estimator (Breslow et al., 2009). The authors replaced the sampling probability α with its maximum likelihood estimate $\hat{\alpha}$. The efficiency gain of the $\hat{\alpha}$-estimator depends on the assumed correctness of the logistic model used to estimate α. The estimation procedure ignores the first phase information, and hence there is further room for efficiency gain. Since the asymptotic properties of the $\hat{\alpha}$-estimator is similar to the CDW, but without its complexities in terms of implementation, we anticipate the efficiencies of these two estimators are similar. As a result, the $\hat{\alpha}$-estimator will be used as a comparator for our proposed estimator.

In this article, we present a semiparametric locally efficient estimator for analyzing case-cohort

studies that do not require strong model assumptions and yet achieves efficiency gain over $\hat{\alpha}$-estimator. We derive the most efficient estimator along the lines of the semiparametric theory of Robins et. al (1994) by restricting ourselves to a subclass of estimators which are regular and asymptotically linear (Newey, 1990). Our proposed estimator is an asymptotically linear estimator which is not only easily computable, but also the most efficient within this subclass of estimators, and enjoys nice asymptotic properties.

In the next section, we present the model, notation, and assumptions used throughout this chapter, and we review a class of estimating equations for analyzing case-cohort studies. Section 3 presents our proposed locally efficient estimator. We demonstrate how to draw inference on the regression parameters on our proposed locally efficient estimator (LEE). Also, we show that LEE has the smallest asymptotic variance among all the class of restricted asymptotic linear (RAL) estimators. Section 4 shows simulation study results comparing our proposed estimator to the $\hat{\alpha}$-estimator. Finally, Section 5 discusses the analysis of the Wilm's tumor data.

2 Model, Notation, and Assumption

Let T be the failure time, C be the potential censoring time and Z be the vector of covariates. Suppose that T is conditionally independent of C given Z and that the conditional distribution of T given Z follows the Cox (1972) proportional hazards model

$$\lambda(t \mid Z) = \lambda_o(t) \exp(\beta^T Z), \tag{2.1}$$

where $\lambda(t \mid Z)$ is the conditional hazard for failure given the covariate history up to time t, β is a vector-valued parameter, and $\lambda_o(t)$ is an unspecified baseline hazard function. We will often write $\lambda(t \mid Z)$ as $\lambda(t)$ for brevity. Our goal is to draw inference about β from data observed through a case-cohort sampling scheme.

The case-cohort design evaluates some covariates on the overall cohort and measures expensive covariates on the subcohort of controls and all cases. Let the observed data be denoted by

$$\Big\{\Delta_i, U_i, (1-\Delta_i)\xi_i, \big\{\Delta_i + (1-\Delta_i)\xi_i\big\} Z_i, W_i\Big\},\ i = 1, \ldots, n,$$

where $\Delta_i = I(T_i \leq C_i)$, the indicator for cases, $U_i = \min(T_i, C_i)$ the observed time, ξ_i is the indicator for the controls $\Delta_i = 0$ who are in the subcohort, Z_i be the covariate of interest that is observed for all individuals in the subcohort $\Delta_i = 1$ or $(1-\Delta_i)\,\xi_i = 1$, or equivalently, $\Delta_i + (1-\Delta_i)\,\xi_i = 1$, and W_i is the other covariates observed for the i-th individual. We assume that the cohort study involves n individuals.

Inference for Cox model with simple random sampling follows the approach of a partial likelihood. Since our method primarily relies on the influence function from this process, we briefly describe the procedure in this section. Suppose the exposure Z_i is observed on all individuals in the sample. The partial likelihood score equation for estimating β in such case is given by

$$\sum_{i=1}^{n} \int_0^{\tau} \big\{Z_i - \mathbf{E}(\mathbf{u}, \beta)\big\} dN_i(u) = 0 \tag{2.2}$$

where

$$\mathbf{E}(\mathbf{u},\beta)=\frac{\sum\limits_{i=1}^{n}Y_i(u)Z_i\exp(\beta^TZ_i)}{\sum\limits_{i=1}^{n}Y_i(u)\exp(\beta^TZ_i)}$$

is the weighted average of the exposure vector Z_i among those who are at risk at time u, $Y_i(u)=I(U_i\geq u)$ be the at risk indicator at time u, and $N_i(u)=\Delta_iI(U_i\leq u)$, $i=1,\ldots,n$; τ is a fixed time chosen to limit the analysis to a follow-up time beyond which there is still a reasonable number at risk. The solution of β, $\hat{\beta}_{PL}$ from Equation (2.2) is known to follow an asymptotic normal distribution. Moreover, one can write (Tsiatis, Chapter 3, Page 21)

$$n^{1/2}(\hat{\beta}_{PL}-\beta_o)=n^{-1/2}\sum_{i=1}^{n}\varphi_i+o_p(1),$$

where

$$\varphi_i=I_i^{-1}(\beta_o)\int_0^{\tau}\left\{Z_i-\mathcal{E}(u,\beta_o)\right\}dM_i(u) \tag{2.3}$$

is the influence function of $\hat{\beta}_{PL}$, with

$$\mathcal{E}(u,\beta)=\frac{s^1(u,\beta)}{s^o(u,\beta)}=\frac{E\left[Y_i(u)Z_i\exp(\beta^TZ_i)\right]}{E\left[Y_i(u)\exp(\beta^TZ_i)\right]}$$

$$I_i(\beta)=E\left[\left\{\int_0^{\tau}\left\{Z_i-\mathcal{E}(u,\beta)\right\}dM_i(u)\right\}\times\left\{\int_0^{\tau}\left\{Z_i-\mathcal{E}(u,\beta)\right\}dM_i(u)\right\}^T\right],$$

where $M_i(u)=N_i(u)-\int\limits_0^u\lambda(u\,|\,Z_i)\,Y_i(u)\,du$ is the martingale process corresponding to the hazard function $\lambda(u\,|\,Z_i)$, and the death and at-risk processes $N_i(u)$ and $Y_i(u)$ respectively; $o_p(1)$ is a term that converges to zero in probability as n approaches infinity.

In the case-cohort sampling, however, not all the exposure variables are measured on all individuals. Therefore, it is not possible to estimate β from Equation (2.2). To account for the probability of such variables being included in the subsample, Equation (2.2) is usually weighted by the inverse of the probability of inclusion. In what follows, we first describe common approaches to estimating β from the case-cohort sampling, and their limitations, and then we describe our proposed estimator.

Almost all existing estimators for analyzing case-cohort studies are based on score equations similar to Equation (2.2). In a typical case-cohort setting, it takes the form

$$\sum_{i=1}^{n}\Big\{\Delta_i+(1-\Delta_i)\xi_i\Big\}\int_0^{\tau}\Big\{Z_i-\mathbf{E_c}(\mathbf{u},\beta)\Big\}\,dN_i(u)=0, \tag{2.4}$$

where

$$\mathbf{E_c}(\mathbf{u},\beta) = \frac{\sum_{i=1}^{n}(\xi_i/\alpha_i)Y_i(u)Z_i\exp(\beta^T Z_i)}{\sum_{i=1}^{n}(\xi_i/\alpha_i)Y_i(u)\exp(\beta^T Z_i)}$$

is the at-risk average of the exposure vector Z_i, and α_i is the sampling fraction which is estimated from true or estimated sampling probability. Various proposals for obtaining the sampling weight, α_i, have been published, yielding different case-cohort estimators (Kulich & Lin, 2004). In particular, the $\hat{\alpha}$-estimator is obtained by solving the equation

$$\sum_{i=1}^{n}\Big\{\Delta_i + (1-\Delta_i)(\xi_i/\hat{\alpha}_i)\Big\}\int_0^{\tau}\big\{Z_i - \mathbf{E_c}(\mathbf{u},\beta)\big\}\,dN_i(u) = 0. \tag{2.5}$$

This estimator replaces the sampling probability α_i by its maximum likelihood estimate $\hat{\alpha}_i$ in a correctly specified model. Since the estimation procedure in Equation (2.4) eliminates subjects with incomplete data, all existing estimators that incorporate the pseudoscore in Equation (2.2) for analyzing case-cohort study, including the $\hat{\alpha}$-estimator, use only the second phase subjects, while ignoring the first phase information. This leads to loss of efficiency.

To account for the missing covariates which result from ignoring the first phase information in the estimation in Equation (2.4), Robins et al. (1994), Laan & Robins (2003), and Nan (2004) have proposed estimators that augment the full data influence function, φ_i, by projecting it onto the orthogonal complement of the nuisance tangent space. In particular, the estimator proposed by Nan (2004), is obtained by a one-step Newton-Raphson approximation, solves the efficient score equation with initial values from existing estimators. In our notation, this estimator has influence function

$$\big\{\Delta_i + (1-\Delta_i)(\xi_i/\alpha_i)\big\}\varphi_i + \big\{1-\Delta_i-(1-\Delta_i)(\xi_i/\alpha_i)\big\}E\big\{\varphi_i \,|\, data\big\}, \tag{2.6}$$

where $E\big\{\varphi_i \,|\, data\big\}$ indicates the expectation of the full data influence function given the observed data. The expectation in Equation (2.6) contains population quantities which are often intractable, and without additional assumptions on the full data model and censoring mechanism, can not be be reasonably estimated with finite samples (Van der Laan et al., 2003 p.35). As a result, in the section that follows, we propose an estimator that restricts the influence function to a class of estimators that are regular and asymptotically normally distributed. Our proposed locally efficient estimator (LEE) is built on the semiparametric efficiency theory (Tsiatis, 2006), and contains quantities that are easy to calculate, and is more efficient than the $\hat{\alpha}$-estimator.

3 Proposed Locally Efficient Estimator

In this section, we describe the procedure in deriving our proposed estimator. We restrict ourselves to the class of estimators that are regular and asymptotically linear (RAL) (Newey, 1990). Following

the theory of inverse-weighting and the semiparametrics, all influence functions for case-cohort estimator of β can be written as

$$\big\{\Delta_i + (1-\Delta_i)(\xi_i/\alpha_i)\big\}\varphi_i + \big\{1-\Delta_i-(1-\Delta_i)(\xi_i/\alpha_i)\big\}g\big(\mathcal{F}(T_i)\big), \tag{3.1}$$

where φ_i is as defined in Equation (2.3), and $\mathcal{F}(t)$ is the history of covariates up to time t. The optimal influence function that gives rise to semiparametric efficient estimator is given by Equation (2.6). However, it is not easy to implement in practice due to its complex structure with intractable expectations. Alternatively, we restrict the class by setting $g(\cdot)$ as a linear function of the data indexed by a vector parameter γ, namely,

$$g\big(\mathcal{F}(T_i)\big) = \gamma^T H(T_i),$$

where $H(T_i)$ is a q-dimensional vector function of the covariates observed prior to T_i. In practice, $H(T_i)$ could represent age, gender and other expensive covariates that only need to be measured for the subcohort. In other words, we start with the influence function

$$\Psi_i = \big\{\Delta_i + (1-\Delta_i)(\xi_i/\alpha_i)\big\}\varphi_i + (1-\Delta_i)\big(1-(\xi_i/\alpha_i)\big)\,\gamma^T H(T_i). \tag{3.2}$$

The influence function (3.2) is indexed by the q-dimensional vector parameter γ. Choice of this vector parameter γ will determine how efficient the corresponding estimator will be. Therefore, the problem of finding the estimator with the minimum variance is equivalent to finding optimal γ for which the variance of Ψ_i in (3.2) is minimum. Also, the influence function for the $\hat{\alpha}$-estimator belongs to this class with $\gamma = 0$. Therefore, optimal influence function in this class will be more efficient than the influence function for $\hat{\alpha}$-estimator.

Let us define $K(u) = Pr(C_i > u)$ and $M_i^c(u) = N_i^c(u) - \int_0^u \lambda^c(t)Y_i(t)dt$ be the martingale associated with the censoring process, where $N_i^c(u) = I(U_i \le u, \Delta = 0)$, and $\lambda^c(u)$ is the hazard rate for the censoring distribution. Then plugging in φ_i from Equation (2.3) and using Gill (1980) equality

$$\frac{\Delta_i}{K(T_i)} = 1 - \int_0^{T_i} \frac{dM_i^c(u)}{K(u)},$$

we can express Equation (3.2) as

$$\begin{aligned}\left\{K(T_i) - K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right\}\big\{1+(\xi_i/\alpha_i)\big\}\,I_i^{-1}(\beta)\int_0^{\tau}\big\{Z_i - \mathcal{E}(u,\beta)\big\}\,dM_i(u)\\ +\ \left\{1-K(T_i)+K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right\}\big\{1-(\xi_i/\alpha_i)\big\}\,\gamma^T H(T_i).\end{aligned} \tag{3.3}$$

To find the optimal influence function, as in Robins et al. (1994), we consider the Hilbert space $\mathcal{H}$ consisting of all zero-mean random functions of the observed data with finite variance equipped with

the covariance inner product. Within this space we define the closed linear subspace $\mathcal{U}$ consisting of random functions

$$\mathcal{U}_i = \left\{1 - K(T_i) + K(T_i)\int_0^{T_i} \frac{dM_i^c(u)}{K(u)}\right\}\left\{1 - (\xi_i/\alpha_i)\right\}\gamma^T H(T_i). \tag{3.4}$$

Our aim is to find the γ which minimizes the variance of Equation (3.3), or equivalently, to find the element in $\mathcal{U}$ which is at the minimum distance from

$$\mathcal{V}_i = \left\{K(T_i) - K(T_i)\int_0^{T_i} \frac{dM_i^c(u)}{K(u)}\right\}\left\{1 + (\xi_i/\alpha_i)\right\} I_i^{-1}(\beta)\int_0^{\tau}\left\{Z_i - \mathcal{E}(u,\beta)\right\} dM_i(u). \tag{3.5}$$

By the projection theorem for Hilbert spaces (Tsiatis, 2006, pp.13-19), and the results presented in the Appendix, we deduce that the optimal γ is given by

$$\gamma^{opt} = E\big[\zeta_i H(T_i)H^T(T_i)\big]^{-1} E\big[H(T_i)\eta_i\big], \tag{3.6}$$

where

$$\zeta_i = -\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\left\{(1 - K(T_i))^2 - K^2(T_i)\int_0^{T_i}\frac{\lambda^c(u)S(u)}{K(u)}du\right\}$$

and

$$\begin{aligned}\eta_i = K(T_i)I_i^{-1}\Bigg[&\left\{(1 - K(T_i))\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\right.\\ &\left.-K(T_i)\left(1 - \frac{\xi_i}{\alpha_i}\right)\right\}\int_0^{T_i}(Z_i - \mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du\\ &+K(T_i)\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\int_0^{T_i}\frac{(Z_i - \mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg].\end{aligned}$$

Thus the optimal influence function,Ψ, is given by (3.2) with γ replaced by γ^{opt} from Equation (3.6). Consequently, we can obtain the optimal estimator of β, $\hat{\beta}_{LE}$, by solving

$$\sum_{i=1}^n \hat{\Psi}_i = \sum_{i=1}^n\left[\left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\hat{\varphi}_i + \left\{1 - \Delta_i - \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\hat{\gamma}^{opt^T}H(T_i)\right] = 0, \tag{3.7}$$

where we estimate the following quantities as follows:

$$\hat{\varphi}_i = \hat{I}_i^{-1}(\hat{\beta}_{LE}) \int_0^{\tau} \{Z_i - \mathbf{E_c}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}})\}\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\},$$

$$\hat{I}_i(\beta) = \left\{ \int_0^{T_i} (Z_i - \mathbf{E_c}(\mathbf{u}, \beta)\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\} \right\}$$

$$\times \left\{ \int_0^{T_i} (Z_i - \mathbf{E_c}(\mathbf{u}, \beta)\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\} \right\}^T$$

$$\hat{\gamma}^{opt} = E\big[\hat{\zeta}_i H(T_i) H^T(T_i)\big]^{-1} E\big[H(T_i)\hat{\eta}_i\big]$$

$$\hat{\zeta}_i = -\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\left\{(1 - \hat{K}(T_i))^2 - \hat{K}^2(T_i) \int_0^{T_i} \frac{\hat{\lambda}^c(u)\hat{S}(u)}{\hat{K}(u)} du \right\}$$

$$\hat{\eta}_i = \hat{K}(T_i)\hat{I}_i^{-1}(\beta)\Bigg[\left\{(1 - \hat{K}(T_i))\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\right.$$

$$\left. - K(T_i)\left(1 - \frac{\xi_i}{\alpha_i}\right)\right\} \int_0^{\tau} (Z_i - \mathbf{E_c}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}}))\hat{\lambda}^c(u)\hat{\lambda}(u)\hat{S}(u)du$$

$$+ \hat{K}(T_i)\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right) \int_0^{\tau} \frac{(Z_i - \mathbf{E_c}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}}))}{\hat{K}(u)} \hat{\lambda}^c(u)\hat{\lambda}(u)\hat{S(u)}du\Bigg]$$

$$\hat{\lambda}^c(u) = \frac{dN_i^c(u)}{Y_i(u)} \text{ and } \hat{\lambda}(u) = \frac{dN_i(u)}{Y_i(u)}$$

and $\hat{S}(u)$ is the survival function estimated by the product limit estimator. Note that even though the hazard $\lambda(u \mid Z_i)$ and the survival $S(u \mid Z_i)$ are dependent on covariates, for simplicity, in estimating ζ_i and η_i, we ignored the covariates and simply used Nelson-Aalen and product limit estimator to estimate them. However, the form of the baseline cumulative hazard function, $\Lambda(u \mid Z_i)$, can be estimated as

$$d\hat{\Lambda}_i(u, \hat{\beta}) = \hat{\lambda}(u \mid Z_i) = d\hat{\Lambda}_0(u, \beta_0) \exp(\hat{\beta}^T Z_i), \tag{3.8}$$

and the survival function, $S(u|Z_i)$, can be estimated as

$$\hat{S}(u \mid Z) = n^{-1} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(T_i)} \, I(U_i > t). \tag{3.9}$$

This leads to an estimator for β that is consistent and asymptotically normal. The variance of $\hat{\beta}_{LE}$ can be estimated by

$$\text{var}(\hat{\beta}_{LE}) = (1/n^2) \sum_{i=1}^{n} \hat{\Psi}_i^2. \tag{3.10}$$

4 Simulation Study

Simulation experiments have been carried out to evaluate the large sample properties of our proposed efficient estimator LEE. For comparisons, we also assessed the $\hat{\alpha}$-estimator, and then we used the full data Cox estimator as a reference estimator. The simulation set up is very similar to that described by Kulich & Lin (2004). Our simulation study involves three covariates, a binary covariate Z_1 with $Pr(Z_1 = 1) = p$, and two continuous covariates $Z_2 \sim N(0, 0.5^2)$, and $\log(Z_3) \sim N(cz_2, 0.5^2)$ conditional on $Z_2 = z_2$, and we set $p = 0.5$ and $c = 0.2$. Thus, our model contains three parameters: one for binary covariate Z_1, and two for continuous covariates Z_2 and Z_3. Also, the failure times are generated from the exponential distribution, and the censoring times are generated from a uniform distribution independent of the survival data. We choose 3,000 subjects in the study cohort and the subsample are drawn from the entire cohort.

Simulation results presented are for $\exp(\beta_1) = 1.3, \exp(\beta_2) = 1.2$, and $\exp(\beta_3) = 1.2$ corresponding to Z_1, Z_2 and Z_3 respectively. We assumed that Z_1 and Z_3 were observed at phase one, while Z_2 was only observed at phase two. We generated a surrogate variable $\tilde{Z}_2 \equiv Z_2 + \epsilon$ for every subject, where ϵ is normal with mean zero independent of Z_2. We note that Z_2 and $\tilde{Z}_2$ have correlation equal to either 0.71 or 0.93. We designated the vector function of covariates, $H(T_i)$, as $\tilde{Z}_2$ and Z_3 and $\alpha_i = (0.100, 0.105, 0.09)^T$. The choice of $H(T_i)$ in this setting was arbitrary, but could be tailored to the most expensive covariates that only need to be measured for the sub-cohort. Also, the sampling scheme for the sub-cohort is exactly the same as that described by Kulich and Lin (2004, page 834). Given the failure status, the probability that a case is included in the sub-cohort sample is 1. Once a failure status is known, one can form separate stratum consisting of the cases and consider the whole stratum sampled with probability 1, whereas the sub-cohort is sampled with probabilities $\alpha_1, \ldots, \alpha_k$ from the controls classified into the remaining K strata. In our simulation setup, stratified sampling was done so that sub-cohort subjects were about equally distributed among the strata. As a result, eight (8) strata were defined based on Z_1 and on the medians of Z_2 and Z_3. Simulation results are based on 1,000 replications. Also, the exact simulation setup was used to calculate the $\hat{\alpha}$-estimator. Then we used the Nested Cohort package in R, written by Mark and Katki (2006), to obtain the estimates. The simulation results are based on 1000 replicates, and so the means of the standard error estimates are based on Monte Carlo simulation samples. Full data estimates are based on all 3000 subjects and relative efficiencies (REs) are calculated by comparing the Monte Carlo variance of the full data estimate to the other estimates. Table 1 presents the estimators and relative efficiencies for the different estimators. The first row shows the results of 10% event rate for all three estimators. This means that after including all the cases, we obtained 1275 samples for analysis. We set the full data Cox as the reference category, and compared both our proposed locally efficient estimator (LEE), and the $\hat{\alpha}$-estimator to the full data Cox model. The relative efficiencies have been calculated by using the ratio of the Monte Carlo means-squared errors. For instance, the entry 1.320 (0.150), RE = 0.87 for $\hat{\alpha}$-estimator in row one of Table 1 refers to the case where with a sample size of 1275, Monte Carlo mean of relative risk estimates for the covariate Z_1 is 1.32, showing a bias of 0.03. The ratio of the Monte Carlo mean-squared error of the full data Cox model to that of the $\hat{\alpha}$-estimator is 0.87. Thus, the $\hat{\alpha}$-estimator in this case is 13% less efficient compared to the full data Cox estimator. In contrast, for the same scenario, the LE estimator has Monte Carlo mean of 1.339

Table 1: Estimator, Monte Carlo standard deviations (MCSE) and relative efficiencies (RE) of LEE and $\hat{\alpha}$-estimators with $Corr(Z_2, \tilde{Z}_2) = 0.71$. True Relative Risks are $\exp(\beta_1) = 1.3$, $\exp(\beta_2) = 1.2$, and $\exp(\beta_3) = 1.2$.

		Full data Cox	LE estimator		$\hat{\alpha}$-estimator	
Event Rate:(n)	Var	Est. (MCSE)	Est. (MCSE)	RE	Est. (MCSE)	RE
$10\% : 1275$	Z_1	$1.331(0.140)$	$1.339(0.145)$	0.93	$1.320(0.150)$	0.87
	Z_2	$1.201(0.099)$	$1.206(0.102)$	0.94	$1.197(0.109)$	0.82
	Z_3	$1.155(0.083)$	$1.158(0.085)$	0.95	$1.151(0.089)$	0.87
$20\% : 1537$	$Z1$	$1.868(0.096)$	$1.876(0.100)$	0.92	$1.887(0.104)$	0.85
	Z_2	$1.486(0.076)$	$1.483(0.078)$	0.95	$1.477(0.084)$	0.82
	$Z3$	$1.107(0.083)$	$1.111(0.086)$	0.93	$1.112(0.088)$	0.89

and a relative efficiency of 0.93, showing a bias of 0.01. This suggests that the LE estimator is only 7% less efficient compared to the full data Cox estimator. Furthermore, comparing the estimators show that the efficiency gain of the LE estimator over the $\hat{\alpha}$-estimator is approximately 7%.

For the variable Z_2 the LE estimator has Monte Carlo mean estimate of 1.206 and a relative efficiency of 0.94; the $\hat{\alpha}$-estimator, which has Monte Carlo mean estimate of 1.197 and relative efficiency of 0.82. In terms of efficiency gain for the Z_2 variable, the LE estimator gains 12% efficiency over the $\hat{\alpha}$-estimator. For the variable Z_3, the LE estimator has Monte Carlo mean estimate of 1.158 and relative efficiency of 0.95, compared to the $\hat{\alpha}$-estimator which has Monte Carlo mean estimate of 1.151 and a relative efficiency of 0.87. This suggests an efficiency gain of 9% of the LE estimator over the $\hat{\alpha}$-estimator. From the relative efficiencies of the two methods, it is evident that our proposed estimator always performs better. Also, when we increased the event rate by sampling 20% from the subsample and adding all cases in the cohort (n= 1537), a similar trend is observed. The LE estimator is more efficient than the $\hat{\alpha}$-estimator.

Likewise, Table 2 shows the relative efficiency of our proposed estimator compared with other estimators, while varying the correlation coefficient from 0.71 to 0.93. We notice that the differences in correlation has little effect on the overall results. This holds true for all three estimators. However, as the event rate increases (i.e. 10% to 20%), the relative efficiencies of case-cohort estimators are high compared to full-cohort estimator. The MCSEs are certainly smaller for higher event rate as the effective total sample size is larger compared to lower event rate.

5 Analysis of Wilm's Tumor Data

Wilm's tumor is a rare type of kidney cancer that occurs in children. Many factors contribute to the survival or relapse of this tumor. Some of the factors include: age at diagnosis, stage of diagnosis

Table 2: Estimator, Monte Carlo standard deviations (MCSE) and relative efficiencies (RE) of LEE and $\hat{\alpha}$-estimators with $Corr(Z_2, \tilde{Z}_2) = 0.93$. True Relative Risks are $\exp(\beta_1) = 1.3, \exp(\beta_2) = 1.2$, and $\exp(\beta_3) = 1.2$.

		Full data Cox	LE estimator		$\hat{\alpha}$-estimator	
Event Rate:(n)	Var	Est. (MCSE)	Est. (MCSE)	RE	Est. (MCSE)	RE
$10\% : 1201$	$Z1$	$1.328(0.147)$	$1.336(0.149)$	0.97	$1.318(0.151)$	0.95
	Z_2	$1.196(0.101)$	$1.198(0.104)$	0.94	$1.191(0.108)$	0.87
	$Z3$	$1.151(0.081)$	$1.154(0.084)$	0.93	$1.149(0.089)$	0.83
$20\% : 1425$	$Z1$	$1.866(0.095)$	$1.878(0.100)$	0.90	$1.891(0.103)$	0.85
	Z_2	$1.484(0.076)$	$1.490(0.078)$	0.95	$1.487(0.082)$	0.86
	$Z3$	$1.106(0.083)$	$1.112(0.085)$	0.95	$1.111(0.088)$	0.89

of disease (usually 4 stages), histological type of tumor, and the tumor diameter. We dichotomized the stage variable by combining stages I and II into one group and stages III and IV into another group. There are six parameters in our model excluding the intercept term. We have two interaction variables: histology and the continuous age variable, and the stage and tumor diameter variables. A total of 3915 patients were enrolled into the Wilm's Tumor Study [D'Angio et al, 1989]. We assume that all covariates except histological group are observed at phase one. Also, we assign the vector function of covariates, $H(T_i)$ as age and stage variables. After the second phase sampling we obtain 660 control subjects and 669 cases. As a result, a total of 1329 observations are analyzed.

Results in Table 3 show that the for each centimeter increase in tumor diameter, the relative risk of death is 1.06 for patients in stage I-II. This represents a 6% increase in risk of death per 1 unit increase in tumor diameter for individuals in this group. However, individuals in stage III-IV have a slightly lower risk (RR=0.98). Thus, an increase in tumor diameter for persons in stage I-II can be fatal but not for persons in stage III-IV. This is may seem counter-intuitive, however it is because individuals in stage III-IV are already at higher risk of dying from cancer and so an additional increase in diameter of the tumor would not necessarily determine the survival or otherwise death of the patient. The age effect is amplified among patients with unfavorable histology compared to those with favorable histology. For every year increase in age, the risk for a person with unfavorable histology increases by 10% (RR =1.10, CI =(1.07,1.13)) whereas for a person with favorable histology, age is not statistically significant.

As patients become older, the risk of death increases for patients in the unfavorable histology group compared to patients in the favorable histology group. The risk of death in Stage III-IV compared to that in Stage I-II depends on the diameter of tumor such that for smaller tumor diameter the relative risk is close to one; whereas, for larger tumor size, the Stage I-II patients are at greater risk of death compared to Stage III-IV patients.

Table 3: Analysis of Wilm's Tumor Data Using LE Estimator.

Variable	Relative Risk	SE	CI
Tumor Diameter (per cm):			
For Stage I-II patients	1.06	0.015	(1.03, 1.09)
For Stage III-IV patients	0.98	0.013	(0.95, 1.01)
Stage III-IV vs Stage I-II:			
At tumor diameter = 5cm	0.71	0.065	(0.58, 0.84)
At tumor diameter = 11.5cm	0.43	0.096	(0.24, 0.61)
At tumor diameter = 20cm	0.22	0.086	(0.05, 0.39)
Age effect (per year):			
With Favorable Histology	0.95	0.034	(0.89, 1.02)
With Unfavorable Histology	1.10	0.018	(1.07, 1.13)
Unfavorable vs Favorable Histology:			
At age = 1 year	0.95	0.001	(0.95, 0.96)
At age = 3.5 years	0.66	0.086	(0.49, 0.83)
At age = 10 years	0.26	0.037	(0.19, 0.32)

6 Discussion

In this article, we have derived an expression for the most efficient estimator for the restricted asymptotically linear estimators for analyzing case-cohort designs. The proposed LE estimator works well for both binary and continuous covariates. The LE estimator is guaranteed to gain efficiency over other available estimators such as the estimator proposed by Mark & Katki (2006). In particular, the LE estimator is more efficient than the $\hat{\alpha}$-estimator. Our proposed estimator uses all the first phase information, and therefore, for inferential procedures, our estimator will give more accurate results than existing ones. Furthermore, our proposed estimator contains quantities that are easy to calculate, with nice asymptotic properties.

On the contrary, the efficiency of the $\hat{\alpha}$-estimator depends on whether we have completely observed continuous covariates versus completely observed binary covariates. In addition, the efficiency of the $\hat{\alpha}$-estimator depends on the correctness of the assumed logistic model, while the efficiency of LEE does not depend on whether the observed covariates are binary or continuous. Through simulations and data analysis, we have shown that the efficiency gain of the LE estimator is substantial over the inverse probability weighted estimator. Also, the LE estimator is consistent and asymptotically normal.

Define the filtration $\mathcal{F}(u)$ as the increasing sequence of sub$-\sigma-$algebras

$$\mathcal{F}(u) = \sigma\{N_i(s), N_i^c(s), Z_i, 0 \leq s \leq u, i = 1, \ldots, n\},$$

and $H_{1i}(\cdot)$ and $H_{2i}(\cdot)$ are predictable functions with respect to $\mathcal{F}(u)$. Unless otherwise stated, this is the filtration with respect to which all the martingales are defined in this paper.

A Derivation of γ^{opt}

In order to derive the optimal gamma we use the following results which follow from theorem 2.4.4 in Fleming and Harrington (1991).

Under the assumption of independent censoring

$$\begin{aligned} E\Big[\int_0^{T_i} H_{1i}(u)dM_i(u) \times \int_0^{T_i} H_{2i}(u)dM_i^c(u)\Big] &= -E\Big[\int_0^{T_i} H_{1i}(u)H_{2i}(u)Y_i(u)\lambda(u)\lambda^c(u)du\Big] \\ &= -E\Big[\int_0^{T_i} H_{1i}(u)H_{2i}(u)\lambda^c(u)\lambda(u)S(u)du\Big], \end{aligned} \tag{A.1}$$

$$E\Big[H_{1i}(u)\int_0^{T_i} H_{2i}(u)dM_i^c(u)\Big] = E\Big[H_{2i}(u)\int_0^{T_i} H_{1i}(u)dM_i(u)\Big] = 0, \tag{A.2}$$

and

$$E\Big[\int_0^{T_i} H_{1i}(u)dM_i^c(u) \times \int_0^{T_i} H_{2i}(u)dM_i^c(u)\Big] = E\Big[\int_0^{T_i} H_{1i}(u)H_{2i}(u)\lambda^c(u)K(u)S(u)du\Big]. \tag{A.3}$$

From the discussion in Section 3, the optimal influence functions, or equivalently, the optimal γ is obtained by projecting $\mathcal{V}_i$ in Equation (3.5) on $\mathcal{U}$ defined by the elements $\mathcal{U}_i$ in Equation (3.4). By

the projection theorem, the optimal gamma (γ^{opt}) must satisfy

$$E\Bigg[\Bigg[\Bigg\{K(T_i)-K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}+\frac{\xi_i}{\alpha_i}\Bigg[1-K(T_i)+K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\Bigg]\Bigg\}$$
$$\times I_i^{-1}(\beta)\int_0^{\tau}(Z_i-\mathcal{E}(u,\beta))dM_i(u)+\Bigg\{1-K(T_i)+K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}-\frac{\xi_i}{\alpha_i}\Bigg[1-K(T_i)$$
$$+K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\Bigg]\Bigg\}\times\gamma^{opt^T}H(T_i)\Bigg]\times\Bigg[\Bigg\{1-K(T_i)+K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}$$
$$-\frac{\xi_i}{\alpha_i}\Bigg[1-K(T_i)+K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\Bigg]\Bigg\}\times\gamma^T H(T_i)\Bigg]\Bigg]=0 \quad \text{(A.4)}$$

Since Equation (A.4) has to be true for any γ, we set $\gamma=0$ and solve Equation (A.4) for γ^{opt} using iterative conditional expectation.

Consequently, we obtain

$$\gamma^{opt}E\Bigg[-\Bigg(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\Bigg)\Bigg\{(1-K(T_i))^2-K^2(T_i)\int_0^{T_i}\frac{\lambda^c(u)S(u)}{K(u)}du\Bigg\}\Bigg]$$
$$=E\Bigg[K(T_i)I_i^{-1}(\beta)\Bigg[\Bigg\{(1-K(T_i))\Bigg(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\Bigg)$$
$$-K(T_i)\Bigg(1-\frac{\xi_i}{\alpha_i}\Bigg)\Bigg\}\int_0^{\tau}(Z_i-\mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du$$
$$+K(T_i)\Bigg(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\Bigg)\int_0^{\tau}\frac{(Z_i-\mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg] \quad \text{(A.5)}$$

Therefore, by solving Equation (A.5) for γ^{opt} we obtain

$$\gamma^{opt}=E\big[\zeta_i H(T_i)H^T(T_i)\big]^{-1}E\big[H(T_i)\eta_i\big] \quad \text{(A.6)}$$

where

$$\zeta_i=-\Bigg(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\Bigg)\Bigg\{(1-K(T_i))^2-K^2(T_i)\int_0^{T_i}\frac{\lambda^c(u)S(u)}{K(u)}du\Bigg\}$$

and

$$\eta_i = K(T_i)I_i^{-1}\Bigg[\Bigg\{(1-K(T_i))\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right) - K(T_i)\left(1-\frac{\xi_i}{\alpha_i}\right)\Bigg\}\int_0^\tau (Z_i-\mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du + K(T_i)\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\int_0^\tau \frac{(Z_i-\mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg].$$

References

[1] Andersen, P.K. & Gill, R. (1983). Cox Regression Model for Counting Processes: A Large-Sample Study. *Ann. Stats.* **10**, 1100–1120.

[2] Barlow, W. E. (1994). Robust Variance Estimation for the Case-Cohort Design. *Biometrics* **50**, 1062–1072.

[3] Borgan, O. & Goldstein, L. et al. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics* **23**, 1749–1778.

[4] Borgan, O. & Langoholz, B. et al. (2000). Exposure Stratified Case-Cohort Designs. *Lifetime Data Analysis* **6**, 39–58.

[5] Breslow, N. E., Lumley, t., Ballantyne, C. M., Chambless, L. e. & Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Stat Biosci* **1**, 1–19.

[6] Breslow, N. E. & Holubkov, R. et al. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Stat. Soc.* B **59**, 447–461.

[7] Breslow, N. E. & Lumley, T. et al. (2009). Using the whole cohort in the analysis of case-cohort data. *Am J Epidem* **169**, 1–8.

[8] Breslow, N. E. & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scan J Stat* **34**, 86–102.

[9] Cai, J., Kim, S. & Lu, W. (2013). More efficient estimators for case-cohort studies. *Biometrika* **100**, 695–708.

[10] Chen, H. Y. (2002). Double-semiparametric method for missing covariates in Cox regression models. *J. Am Stat Assoc.* **97**, 565–576.

[11] Chen, H. Y. & Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *J. Am Stat Assoc.* **94**, 896–908.

[12] Chen, K. (2001). Generalized case-cohort sampling. *J. R. Stat. Soc.* B **63**, 791–809.

[13] Chen, K. & Lo, L. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **86**, 755–764.

[14] Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Stat. Soc.* B **34**, 187–220.

[15] D'Angio, G. J. & Breslow, N. (1989). Treatment of Wilm's Tumor: Results of the Third National Wilm's Tumor Study. *Cancer* **64**, 349–360.

[16] Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley Series, revised ed.

[17] Gill, R. D. (1980). Censoring and Stochastic Integrals, Mathematical Centre Tracts No. 124. *Amsterdam: Mathematisch Centrum*

[18] Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Am Stat Assoc.* **47**, 663–685.

[19] Kalbfleisch, J. D. & Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* **7**, 149–160.

[20] Kalbfleisch, J. D. & Prentice, R. L. (1978). Analysis of failure times in the presence of competing risks. *Biometrics* **43**, 541–554.

[21] Kong, L. & Cai, J. (2009). Case-Cohort Analysis with Accelerated Failure Time. *Biometrics* **65**, 135–142.

[22] Kong, L., Cai, J. & Sen, P. K(2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika* **91(2)**, 305–319.

[23] Kulich, M. & Lin, D. Y. (2004). Improving the efficiency of relative-risk in case-cohort studies. *J Am Stat Assoc.* **99**, 832–844.

[24] Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. New York: John Wiley & Sons, Inc.

[25] Mark, S. D. & Katki, H. A. (2006). Specifying and Implementing Nonparametric and Semiparametric Survival Estimators in Two-Stage(Nested)Cohort Studies With Missing Case Data. *J Am Stat Assoc.* **101**, 460–471.

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

[27] Nan, B., Emond, M. & Wellner, J. A. (2004). Information Bounds for Cox Regression Models With Missing Data. *Ann. Stats* **32**, 723–753.

[28] Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Economet.* **5**,99–135.

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

[30] Robins, J. M., Rotniztky, A. & Zhao, L.P (1994). Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* **89**, 846–866.

[31] Rothman, K. J. (2002). *Epidemiology: An Introduction*. New York: Oxford University Press, New York.

[32] Self, S. G. & Prentice, R. L. (1988). Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies. *Ann. Stats.* **16**, 64–81.

[33] Tsiatis, A. A. (2010). *Semiparametric Theory and Missing Data*. New York: Springer, revised ed.

[34] Van der Laan, M. J. & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer Series in Statistics, New York.

[35] Wahed, A. S & Tsiatis, A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. *B*iometrics **60**, 124–133.

[36] Wahed, A. S & Tsiatis, A. A. (2006). Semiparametric efficient estimation of survival distributions in two-stage randomized designs in clinical trials with censored data. *B*iometrika **93(1)**, 163–177.

[37] Wang, C. Y. & Chen, H. Y. (2001). Augmented Inverse Probability Weighted Estimator for Cox Missing Covariate Regression. *Biometrics* **57**, 414–419.

[38] Zhao, H. & Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84(2)**, 339–348.