Published in final edited form as:

Biometrics. 2010 June ; 66(2): 374–381. doi:10.1111/j.1541-0420.2009.01277.x.

Cox Regression in Nested Case-Control Studies with Auxiliary Covariates

Mengling Liu^{1,*}, Wenbin Lu², and Chi-hong Tseng³

¹Division of Biostatistics, School of Medicine, New York University, New York, NY

²Department of Statistics, North Carolina State University, Raleigh, NC

³Department of Medicine, University of California at Los Angeles, Los Angeles, CA

SUMMARY

Nested case-control (NCC) design is a popular sampling method in large epidemiologic studies for its cost effectiveness to investigate the temporal relationship of diseases with environmental exposures or biological precursors. Thomas' maximum partial likelihood estimator is commonly used to estimate the regression parameters in Cox's model for NCC data. In this paper, we consider a situation that failure/censoring information and some crude covariates are available for the entire cohort in addition to NCC data and propose an improved estimator that is asymptotically more efficient than Thomas' estimator. We adopt a projection approach that, heretofore, has only been employed in situations of random validation sampling and show that it can be well adapted to NCC designs where the sampling scheme is a dynamic process and is not independent for controls. Under certain conditions, consistency and asymptotic normality of the proposed estimator are established and a consistent variance estimator is also developed. Furthermore, a simplified approximate estimator is proposed when the disease is rare. Extensive simulations are conducted to evaluate the finite sample performance of our proposed estimators and to compare the efficiency with Thomas' estimator and other competing estimators. Moreover, sensitivity analyses are conducted to demonstrate the behavior of the proposed estimator when model assumptions are violated, and we find that the biases are reasonably small in realistic situations. We further demonstrate the proposed method with data from studies on Wilms' tumor.

Keywords

Counting process; Cox proportional hazards model; Martingale; Risk set sampling; Survival analysis

1. Introduction

Due to its quality of being cost-effective in studying temporal relationship between disease and exposures, nested case-control (NCC) sampling (Thomas, 1977; Oakes, 1981) has been considered a useful alternative to cohort design and case-control design. The most commonly used analytical approach for NCC data is Thomas' maximum partial likelihood estimation approach (Thomas, 1977; Oakes, 1981) under Cox proportional hazards model (Cox, 1972) assumption. The consistency and asymptotic normality of Thomas' estimator have been formally established using counting process and martingale theory (Goldstein and Langholz, 1992). Recently, Chen (2004) proposed a partial likelihood based local-averaging estimator that is more efficient than Thomas' estimator away from the null.

*email: mengling.liu@nyu.edu.

Furthermore, in the presence of *extended* NCC data (Chen, 2004) which consist of failure/censoring times and indices for the full cohort and entire covariate histories for the cases and selected controls, a number of methods have been proposed to improve the estimation efficiency: e.g. the inverse probability weighted (IPW) method (Robins et al., 1994; Samuelsen, 1997); the local-average estimation approach (Chen, 2004); the likelihood-based approaches (Chen and Little, 1999; Scheike and Juul, 2004; Zeng et al., 2006). Since parent cohorts of NCC studies are usually well-characterized, carefully-followed epidemiological cohorts, the failure/censoring information on the entire cohort is often available. In many studies, however, the true exposure covariates may be difficult or expensive to be assembled for the full cohort or for their entire history to be measured for the cases and selected controls. Instead, some auxiliary covariates, such as crude measurements of the exposure or inferred covariates from questionnaire, can be easily or cheaply assembled for the entire cohort. The aims of this paper are to incorporate the information of failure/censoring and auxiliary covariates from the entire cohort into the analysis of NCC data and to propose an easily computed estimator that is asymptotically more efficient than Thomas' estimator.

Towards this goal, we propose to adopt a projection technique which has been used to improve the efficiency of various models in cohort studies with random validation sampling, such as general linear regression models (Chen and Chen, 2000), Cox's model (Chen, 2002), and the additive hazards model (Jiang and Zhou, 2007). To the best of our knowledge, the projection method heretofore has been only studied for the random validation sampling and its adaptation to the NCC sampling entails new challenges, primarily due to the non-independent sampling scheme of NCC design. Statistical inference thus can not rely on the conventional independent central limit theory. In this paper, we show that the projection method can be well adapted to the NCC design under certain conditions and will lead to an improved estimator that is guaranteed to achieve an asymptotic variance no bigger than that of Thomas' estimator.

The rest of this article is organized as follows. In Section 2, we derive the proposed estimator and its asymptotic properties and present a practical computation procedure. A rare-disease approximate estimator is also provided and some inference remarks are discussed. In Section 3, extensive simulation studies are conducted to evaluate the performance of our proposed estimators under various practical settings. An illustration with a real dataset from Wilms' tumor studies is also provided. We conclude with some discussions in Section 4 and provide all the technical details in Supplementary Material.

2. Projection Estimator and Statistical Inference

Consider a full cohort of size n . Let $\{T_i^*, C_i, Z_i(\cdot), i=1, \dots, n\}$ denote n i.i.d. triplets of failure times, censoring times, and p -dimensional covariate processes of interest. Define $T_i = \min(T_i^*, C_i)$, $\delta_i = I(T_i^* \leq C_i)$, $N_i(t) = \delta_i I(T_i \leq t)$ and $Y_i(t) = I(T_i \geq t)$, where $I(\cdot)$ denotes the indicator function throughout. An NCC study identifies cases as subjects of $\delta_i = 1$ and randomly samples $(m-1)$ controls without replacement from the risk set at each failure time, excluding the failed subject itself. For a given case i , let R_i^* denote the indices of the $(m-1)$ selected controls and define $\tilde{R}_i = R_i^* \cup \{i\}$. The true covariates are then ascertained for all the cases and selected controls. Therefore, for a standard NCC design, the observed data consist of $\{T_i, Z_i(T_i), Z_j(T_i); \delta_i=1, j \in R_i^*, i=1, \dots, n\}$. As we discussed in the introduction section, in addition to the data collected by the NCC sampling, we consider the situation that the failure/censoring information and some auxiliary covariates, i.e. $\{T_i, \delta_i, X_i(t) : 0 \leq t \leq T_i, i=1, \dots, n\}$, are also collected for the entire cohort, where $X_i(t)$ denotes the q -dimensional auxiliary covariate processes of subject i .

Assume that, given the true covariate $Z(\cdot)$, T^* follows a Cox proportional hazards model

$$\lambda\{t|\bar{Z}(t)\} = \lambda_0(t) \exp\{\beta'_0 Z(t)\}, \quad (1)$$

where $\bar{Z}(t) = \{Z(s) : 0 \leq s \leq t\}$, $\lambda_0(t)$ is an unspecified baseline hazard function and β_0 is a p -dimensional parameters of interest. Furthermore, we assume that the censoring time C is independent of the failure time T^* given Z .

2.1 Thomas' estimator under the true model

Thomas' estimator, denoted by $\hat{\beta}$, is the solution to

$$U_Z(\beta) = \sum_{i=1}^n \int_0^\tau \{Z_i(t) - E_{Z,w}(t; \beta)\} dN_i(t) = 0, \quad (2)$$

where $\tau = \inf\{t : \text{pr}(T > t) = 0\}$ and $E_{Z,w}(t; \beta) = \frac{\sum_{j \in w} e^{\beta' Z_j(t)} Z_j(t)}{\sum_{j \in w} e^{\beta' Z_j(t)}}$ for a set w . Oakes (1981) showed that Thomas' estimator maximizes the partial likelihood, and Goldstein and Langholdz (1992) proved that, under certain regularity conditions,

$$n^{1/2}(\hat{\beta} - \beta_0) \rightarrow N(0, \Gamma^{-1}), \quad (3)$$

as $n \rightarrow \infty$, where $\Gamma = - \lim_{n \rightarrow \infty} n^{-1} \frac{\partial U_Z(\beta)}{\partial \beta} |_{\beta=\beta_0}$.

2.2 Estimators under a working model

To utilize the auxiliary information available on the full cohort, we assume a working Cox's model specified by $\alpha_0(t) \exp\{\gamma' X(t)\}$. We first introduce extra notation:

$$S^{(k)}(t; \gamma) = n^{-1} \sum_{i=1}^n Y_i(t) e^{\gamma' X_i(t)} X_i^{\otimes k}(t) \quad \text{and} \quad S^{(k)}(t) = n^{-1} \sum_{i=1}^n Y_i(t) \lambda_i(t) X_i^{\otimes k}(t),$$

where $k = 0, 1, 2$, and for a vector a , $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa'$; $\lambda_i(t)$ generically denotes the true hazard function of subject i . Let $s^{(k)}(t; \gamma) = E\{S^{(k)}(t; \gamma)\}$ and $s^{(k)}(t) = E\{S^{(k)}(t)\}$ where the expectation is taken with respect to the joint distribution of (T, δ, X) .

Let $\tilde{\gamma}$ denote the full-cohort maximum partial likelihood estimator under the working model, defined as the solution to

$$\tilde{U}(\gamma) = \sum_{i=1}^n \int_0^\tau \{X_i(t) - \bar{X}(t, \gamma)\} dN_i(t) = 0, \quad (4)$$

where $\bar{X}(t, \gamma) = \frac{S^{(1)}(t; \gamma)}{S^{(0)}(t; \gamma)}$. Lin and Wei (1989) showed that $\tilde{\gamma}$ converges in probability to a constant vector γ_* , which is the unique solution to

$$\int_0^\tau \left\{ s^{(1)}(t) - \frac{s^{(1)}(t; \gamma)}{s^{(0)}(t; \gamma)} s^{(0)}(t) \right\} dt = 0,$$

provided that the matrix

$$A = - \left\{ \lim_{n \rightarrow \infty} n^{-1} \frac{\partial \tilde{U}(\gamma)}{\partial \gamma} \Big|_{\gamma=\gamma_*} \right\} = \int_0^\tau \left\{ \frac{s^{(2)}(t; \gamma)}{s^{(0)}(t; \gamma)} - \left(\frac{s^{(1)}(t; \gamma)}{s^{(0)}(t; \gamma)} \right)^{\otimes 2} \right\} s^{(0)}(t) dt$$

is positive definite. Furthermore, Lin and Wei (1989) showed that, as $n \rightarrow \infty$,

$$n^{1/2}(\tilde{\gamma} - \gamma_*) \rightarrow N(0, A^{-1}BA^{-1}), \quad (5)$$

under certain regularity conditions, where $B = \lim_{n \rightarrow \infty} \text{Var} \{n^{-1/2} \tilde{U}(\gamma_*)\}$.

Next, we derive another consistent estimator for γ_* based on the auxiliary covariates of those subjects selected by the NCC sampling. To achieve this, we impose the following conditions on the auxiliary covariates X :

C1. given the true covariates $Z(\cdot)$, $X(\cdot)$ is independent of T^* and C

C2. there exist $\check{\alpha}_0(\cdot)$ and $\check{\gamma}$ such that the induced hazard function of T^* given $X(\cdot)$ has a proportional form, i.e.

$$\lambda \{t | \bar{X}(t)\} = \check{\alpha}_0(t) \exp \{\check{\gamma}' X(t)\}. \quad (6)$$

Condition C1 indicates that X is a true surrogate of Z , which is commonly assumed in many studies of surrogacy. Condition C2 ensures that Thomas' estimator based on the auxiliary covariates can estimate the same quantity as the full cohort estimator γ under the working model (Xiang and Langholz, 1999). Therefore, let γ be the solution to

$$U_X(\gamma) = \sum_{i=1}^n \int_0^\tau \{X_i(t) - E_{X, R_i}(t; \gamma)\} dN_i(t) = 0, \quad (7)$$

where $E_{X,w}(t; \gamma) = \frac{\sum_{j \in w} e^{\gamma' X_j(t)} X_j(t)}{\sum_{j \in w} e^{\gamma' X_j(t)}}$ for a set w . Xiang and Langholz (2003) showed that

$$n^{1/2}(\hat{\gamma} - \gamma_*) \rightarrow_{n \rightarrow \infty} N(0, I^{-1} V I^{-1}), \quad (8)$$

in distribution, where $I = - \lim_{n \rightarrow \infty} n^{-1} \frac{\partial U_X(\gamma)}{\partial \gamma} \Big|_{\gamma=\gamma_*}$ and $V = \lim_{n \rightarrow \infty} \text{Var} \{n^{-1/2} U_X(\gamma_*)\}$.

The assumption C2 is required for rigorous theoretical justification but in general it may not hold exactly (Prentice, 1982). Note that the primary interest here is how well the NCC estimator $\hat{\gamma}$ approximates the full-cohort estimator γ under the working model rather than how the working model deviates from the true model. Although the limiting difference of $\hat{\gamma} - \gamma$ may not exactly be zero, such difference in general does not occur unless the magnitude

of the misspecification is unreasonably large as noted in Xiang and Langholz (1999). In addition, under the rare-disease assumption that is often true in NCC studies, the induced hazard function can be adequately approximated by $\lambda_0(t)E\{\exp(\beta'Z(t)) | X(t)\}$ that can further relax the assumption. We will further investigate the impact of condition C2 on the parameter estimation in our simulation studies.

2.3 Projection estimator and its asymptotic properties

Following the similar projection idea used in Chen and Chen (2000), Chen (2002) and Jiang and Zhou (2007), we incorporate the information available on the entire cohort, i.e. $\{(T_i, \delta_i, X_i), i = 1, \dots, n\}$, into the estimation of β by considering the joint limiting distribution of $\{n^{1/2}(\hat{\beta} - \beta_0)', n^{1/2}(\hat{\gamma} - \gamma_0)'\}$. We introduce some notation. Let $r = \{1, \dots, m\}$, $Y_r(t) = \prod_{i \in r} Y_i(t)$,

$P_Y(t) = \text{pr}\{Y_1(t) = 1\}$, and $\bar{x}(t; \gamma) = \frac{s^{(1)}(t; \gamma)}{s^{(0)}(t; \gamma)}$. Moreover, define

$$\begin{aligned} K_1 &= \int_0^\tau P_Y(t) E \left[m^{-1} \sum_{i \in r} \{Z_i(t) - E_{Z,r}(t; \beta_0)\} \{X_i(t) - E_{X,r}(t; \gamma_*)\}' \lambda_i(t) | Y_r(t) = 1 \right] dt, \\ K_2 &= \int_0^\tau P_Y(t) E \left[m^{-1} \sum_{i \in r} \{Z_i(t) - E_{Z,r}(t; \beta_0)\} \{X_i(t) - \bar{x}(t; \gamma_*)\}' \lambda_i(t) | Y_r(t) = 1 \right] dt, \\ \Sigma_1 &= \int_0^\tau P_Y(t) E \left[m^{-1} \sum_{i \in r} \{X_i(t) - E_{X,r}(t; \gamma_*)\} \{X_i(t) - \bar{x}(t; \gamma_*)\}' \lambda_i(t) | Y_r(t) = 1 \right] dt, \\ \Sigma_2 &= \int_0^\tau \int_0^\tau P_Y(t) \text{cov} \left[\sum_{i \in r} \{X_i(t) - E_{X,r}(t; \gamma_*)\} \lambda_i(t), \right. \\ &\quad \left. Y_1(s) \{X_1(s) - \bar{x}(s; \gamma_*)\}' \left\{ \lambda_1(s) - \frac{e^{\gamma_*' X_1(s)} s^{(0)}(s)}{s^{(0)}(s; \gamma_*)} \right\} | Y_r(t) = 1 \right] dt ds. \end{aligned}$$

PROPOSITION 1—Under conditions C1 and C2, and the regularity conditions given in the Web Appendix A,

$$n^{1/2} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \rightarrow N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Gamma^{-1} & \Delta \\ \Delta' & \Omega \end{pmatrix} \right),$$

in distribution as $n \rightarrow \infty$, where

$$\Delta = \Gamma^{-1} K_1 I^{-1} - \Gamma^{-1} K_2 A^{-1}, \quad (9)$$

$$\Omega = I^{-1} V I^{-1} + A^{-1} B A^{-1} - 2 I^{-1} (\Sigma_1 + \Sigma_2) A^{-1}. \quad (10)$$

The proof of Proposition 1 is given in the Web Appendix A. By Proposition 1 and the multivariate normal distribution theory, we have,

$$E \{ n^{1/2} (\hat{\beta} - \beta_0) | (\hat{\gamma} - \gamma_0) \} = n^{1/2} \Delta \Omega^{-1} (\hat{\gamma} - \gamma_0).$$

It is easy to see that Γ , I and A can be consistently estimated by

$$\hat{\Gamma} = -n^{-1} \frac{\partial U_Z(\hat{\beta})}{\partial \beta}, \hat{I} = -n^{-1} \frac{\partial U_X(\hat{\gamma})}{\partial \gamma} \text{ and } \hat{A} = -n^{-1} \frac{\partial \hat{U}(\hat{\gamma})}{\partial \gamma}, \text{ respectively. Furthermore, let}$$

$$\begin{aligned}\widehat{K}_1 &= n^{-1} \sum_{i=1}^n \int_0^\tau \{Z_i(t) - E_{Z, \widehat{R}_i}(t; \widehat{\beta})\} \{X_i(t) - E_{X, \widehat{R}_i}(t; \widehat{\gamma})\}' dN_i(t), \\ \widehat{K}_2 &= n^{-1} \sum_{i=1}^n \int_0^\tau \{Z_i(t) - E_{Z, \widehat{R}_i}(t; \widehat{\beta})\} \{X_i(t) - \bar{X}(t; \widehat{\gamma})\}' dN_i(t).\end{aligned}$$

Under the regularity conditions, the consistencies of $\widehat{K}_1$ and $\widehat{K}_2$ easily follow Lemma 1 in the supplementary material of Xiang and Langholz (2003). Therefore, the covariance component Δ can be consistently estimated by $\widehat{\Delta} = \widehat{\Gamma}^{-1} \widehat{K}_1 \widehat{\Gamma}^{-1} - \widehat{\Gamma}^{-1} \widehat{K}_2 \widehat{A}^{-1}$.

Next, examining the components of Ω in (10), we note that Σ_2 has a very complicated exposition and it is not straightforward to construct a consistent estimator in general. Thus, we propose to use the bootstrap method (Efron, 1979) to estimate Ω . The bootstrapping approach is feasible here because the auxiliary covariates are available on the entire cohort. More specifically, in the j th run of bootstrap, $j = 1, \dots, J$, where J is a large number, we first randomly sample n subjects from the full cohort with replacement. Then for each case in this bootstrapped sample, we randomly select $m-1$ controls from the risk-set at this case failure time excluding case itself and thus obtain a new NCC dataset. Next, we estimate $\widehat{\gamma}(j)$ and $\widehat{\gamma}(j)$ by fitting the working model to the j th bootstrapped full cohort data and NCC data, respectively. The empirical variance-covariance matrix of $[n^{1/2}\{\widehat{\gamma}(j) - \widehat{\gamma}(j)\}, j = 1, \dots, J]$ yields a consistent estimator for Ω , denoted by $\widehat{\Omega}$. The algorithm does not require any complex variance formula or much programming effort and can be easily implemented in many existing statistical software.

After obtaining the estimates of $\widehat{\Delta}$ and $\widehat{\Omega}$, an improved estimator for β can be constructed as $\widetilde{\beta} = \widehat{\beta} - \widehat{\Delta} \widehat{\Omega}^{-1} (\widehat{\gamma} - \widehat{\gamma})$. Based on Proposition 1, it is easy to show that $n^{1/2}(\widetilde{\beta} - \beta_0)$ is asymptotically normal with mean zero and variance-covariance matrix $\Gamma^{-1} - \Delta \Omega^{-1} \Delta'$. Therefore, the asymptotic variance of $\widetilde{\beta}$ is guaranteed to be no bigger than that of Thomas' estimator and can be consistently estimated by $\widehat{\Gamma}^{-1} - \widehat{\Delta} \widehat{\Omega}^{-1} \widehat{\Delta}'$.

2.4 Inference remarks and rare-disease approximation

It is worthy of making two observations when comparing the projection approach under the random validation sampling and under the NCC sampling. First, in the methods proposed for random validation sampling, all estimating equations can be rewritten as sums of independent mean-zero terms asymptotically. But in our procedure, estimating functions $U_Z(\beta)$ and $U_X(\gamma)$ based on NCC data do not have such independent presentations, and thus entail new technical challenges to establish the asymptotic properties of the proposed estimator. Second, although $\widehat{\gamma}$ and $\widehat{\gamma}$ converge in probability to the same limit γ_* under certain conditions, they do not have the same limiting distribution unless $m \rightarrow \infty$, see (5) and (8). However, in random validation sampling, the validation-set estimator and the full-cohort estimator always converge to the same limiting distribution and have the same asymptotic covariance with the validation-set estimator based on the true model. In the context of NCC sampling, the asymptotic variance of $n^{1/2}(\widehat{\gamma} - \widehat{\gamma})$ has much more complicated form as shown in Proposition 1 and we thus propose to estimate it using the bootstrap method. In summary, these complications root in the non-independent sampling scheme of NCC design.

When the disease is rare, as in many NCC studies, the proposed projection estimator can be well approximated by a plug-in type estimator because the estimation on the variance component V can be greatly simplified (Xiang and Langholz, 2003) and Σ_2 is approximately

negligible. More specifically, we first propose a rare-disease estimator for Ω given by $\hat{\Omega}_r = \hat{F}^{-1} \hat{V}_r \hat{F}^{-1} + \hat{A}^{-1} \hat{B} \hat{A}^{-1} - 2\hat{F}^{-1} \hat{\Sigma}_1 \hat{A}^{-1}$, where

$$\begin{aligned}\hat{V}_r &= n^{-1} \sum_{i=1}^n \int_0^\tau \{X_i(t) - E_{x, \hat{R}_i}(t, \hat{\gamma})\}^{\otimes 2} dN_i(t), \\ \hat{B} &= n^{-1} \sum_{i=1}^n \left[\int_0^\tau \{X_i(t) - \bar{X}(t, \hat{\gamma})\} \left\{ dN_i(t) - Y_i(t) e^{\hat{\gamma}' X_i(t)} \frac{\sum_j dN_j(t)}{nS^{(0)}(t; \hat{\gamma})} \right\} \right]^{\otimes 2}, \\ \hat{\Sigma}_1 &= n^{-1} \sum_{i=1}^n \int_0^\tau \{X_i(t) - E_{x, \hat{R}_i}(t, \hat{\gamma})\} \{X_i(t) - \bar{X}(t, \hat{\gamma})\}' dN_i(t).\end{aligned}$$

Therefore, the rare-disease approximate estimator is defined $\tilde{\beta}_r = \hat{\beta} - \hat{\Delta} \hat{\Omega}_r^{-1} (\hat{\gamma} - \tilde{\gamma})$ and its variance estimator is given by $\hat{\Gamma}^{-1} - \hat{\Delta} \hat{\Omega}_r^{-1} \hat{\Delta}'$.

3. Numerical Studies

3.1 Simulations under correct model conditions

We first investigate the finite-sample performance of the proposed estimator and the rare-disease estimator by extensive simulations. We compare the efficiency of the proposed estimator with Thomas' estimator, a local-averaging estimator (Chen, 2004), and an IPW estimator (Samuelsen, 1997). We consider the following scenarios:

S1. independent auxiliary covariate: Z and X are independently and identically distributed

S2. normal auxiliary covariate is measured with normal error: $X = Z + \varepsilon$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, The true covariate $Z \sim N(2, 0.5^2)$ and $\sigma_\varepsilon = 0.5$ or 0.2 . We generate the failure time T^* from a Cox's model $\lambda(t|Z) = \lambda_0 e^{\beta Z}$, where three different values of β as $0, -0.5$ and -1 are considered. We examine two censoring scenarios: the random censoring, where $C \sim U(0, 5)$, and the covariate-dependent censoring by generating the censoring time C uniformly from $\{0, \min(3|Z|, 5)\}$. The value of λ_0 is chosen to control the disease incidence rate at $6\% \sim 7\%$. Under S1, we examine the robustness of the proposed estimator with completely independent/wrong surrogate covariate. Scenario S2 is a classical measurement error model and it is easy to see that conditions C1 and C2 are satisfied (Xiang and Langholz, 1999). We consider the cohort size of 2000 and the NCC study with 2 or 4 controls. For Chen's estimator, we set the local-average bandwidth to be $2n^{-1/3}$. For the IPW estimator, the weight function is defined as $\pi_i = \delta_i + (1 - \delta_i)V_{0i}/p_{0i}$ where V_{0i} is the indicator of subject i ever being selected as a control and

$p_{0i} = 1 - \prod_{r_j \leq T_i} \left(1 - \frac{m-1}{\sum_k Y_k(T_j) - 1} \delta_j \right)$. We run 500 simulations for each setting and the number of bootstrap samples is set to be 500.

Simulation results under the random censoring are summarized in Table 1 and those under the covariate-dependent censoring are presented in Table 2. In all the scenarios, the proposed estimator shows negligible biases. The estimated standard errors (SE), using the proposed bootstrap method, are close to the sample standard deviation (SD) of the estimates. Thus, the 95% Wald-type confidence intervals all have reasonable coverage probabilities (CP). Moreover, under this rare-disease situation, the rare-disease approximate estimator performs well as it yields reasonable coverage probabilities (CP*).

To compare the efficiency of various approaches, we calculate the empirical relative efficiency for each estimator, defined as the ratio of sample variances of the estimator and the full-cohort maximum partial likelihood estimator under the true model with the latter one being a reference. The efficiency results are summarized in Table 1 and Table 2 (see the last four columns). In scenario S1, where the surrogate covariate is completely independent of the true covariate, the proposed estimator shows very comparable efficiency as Thomas' estimator because independent surrogate covariates can hardly provide any information to improve Thomas' estimator $\hat{\beta}$. Under this scenario, the IPW estimator outperforms others as the selection probability is accurately estimated and used to recover the original full-cohort.

In scenario S2, where a true surrogate covariate is available, the proposed general estimator shows efficiency gain over Thomas' estimator and the magnitude of gain is more obvious when the number of controls is small and the measurement error is small. For example, when $\beta = -1$, $\sigma_\varepsilon = 0.2$ and $m-1 = 2$, the gain of the proposed method over Thomas's estimator, calculated as $(RE/RE_1 - 1)\%$, reaches 61% with covariate-independent censoring and 49% with covariate-dependent censoring. On the other hand, the efficiency of the proposed method approaches the full-cohort efficiency as the measurement error decreases or the number of controls increases. For example, when $\beta = 0$, $\sigma_\varepsilon = 0.2$ and $m-1 = 4$, the relative efficiency of our estimator achieves 99.1% with covariate-independent censoring and 98.1% with covariate-dependent censoring. Moreover, in all simulations under scenario S2, the proposed estimator always outperforms other competing estimators.

Additional simulations when the disease is common with 15% and 25% incidence rates are presented in the Web Appendix B. We observe that the proposed estimator still performs well but the rare-disease approximation may fail with unsatisfactory coverage probabilities.

3.2 Simulations of sensitivity analysis

In this subsection, we further investigate the properties of the proposed estimator $\tilde{\beta}$ when conditions C1 and C2 are violated. Primarily, we focus on the violation of condition C2 that has some practical implications and consider the following scenarios:

- S3. non-normal covariate and error: $X = Z + u$ where $Z \sim U(1, 3)$ and $u \sim U(-1, 1)$
- S4. multiplicative error, i.e. $X = \varepsilon Z$ where $\varepsilon \sim \exp(1)$ and $Z \sim N(2, 0.5^2)$
- S5. working model with a missing covariate, i.e. $Z = (Z_1, Z_2)$ but $X = Z_1$ only
- S6. informative auxiliary covariate, i.e. $X = Z + \alpha \log T^* + \varepsilon$.

Under scenarios S3 and S4, the induced hazard function $\lambda(t|X)$ defined in (6) does not have a proportional exponential form unless $\beta_0 = 0$. Under S5 of missing covariate, $\lambda(t|X)$ generally will not be proportional. We generate dichotomous variables (Z_1, Z_2) from a multinomial distribution with $\pi_{lk} = \text{pr}(Z_1 = l, Z_2 = k)$, where $k, l = 0, 1$ as in Xiang and Langholz (1999). We consider an extreme situation of the odds-ratio being 5 by setting $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.2, 0.2, 0.1, 0.5)$. Note that, under scenarios S3–S5, the surrogate covariate X violates condition C2 only. Finally, under scenario S6, $Z \sim N(2, 0.5^2)$, $\varepsilon \sim N(0, 0.2^2)$ and we explore the informative auxiliary covariate as in Jiang and Zhou (2007). When $\alpha \neq 0$, the auxiliary covariate X in S6 clearly violates both C1 and C2.

In the sensitivity analyses, we consider the random censoring of $U(0, 5)$ and use 2 controls only. Table 3 gives some representative results for different values of β_0 . Under scenarios S3, S4 and S5 where only condition C2 is violated, we observe that the proposed estimator is reasonably robust with small biases and satisfactory coverage probabilities. All mean squared errors are also reasonably small. The results agree with the observations by Xiang and Langholz (1999) that the difference between NCC estimates and the full-cohort

estimates is often small for moderate violation of condition C2 due to measurement error or covariate omission. Moreover, for the missing covariate situation (S5), the efficiency gain of the proposed estimator on the parameter corresponding to the covariate without missing (Z_1) is clearly more obvious than that for the covariate with missing (Z_2). For scenario S6, we observe that biases become more obvious when the magnitude of β_0 increases, and thus the coverage probabilities deteriorate; indicating that condition C1 is an important assumption for the validity of the projection method.

3.3 Wilms' tumor studies

We demonstrate the proposed approach by utilizing of a full-cohort study collected from studies conducted by the National Wilms Tumor Study Group (D'Angio et al., 1989; Green et al., 1998). Wilms' tumor is a malignant tumor of the kidney and typically occurs in children. This dataset contains full information of 3,915 subjects participating the third and fourth Wilms' tumor studies and 669 (17.09%) patients who had disease relapse are considered as cases. We also compare the proposed estimator with Thomas' estimator and the IPW estimator under the NCC sampling with various numbers of controls.

To estimate the effects of unfavorable histology status and other covariates on patients' relapse-free survival, we follow Kulich and Lin (2004) to assume that the relapse time follows model (1) with eight covariates: Age1 (age of diagnosis if less than 1 year old); Age2 (age of diagnosis of 1 year and older); UH (unfavorable central histology); Age1×UH; Age2×UH; Stage (3–4 vs 1–2); Diameter; and Stage×Diameter. We simulate NCC studies from this full cohort with the number of controls ranging from 1 to 3. The evaluation of tumor histology by central pathologists is considered the true histology assessment and pretended to be available only for the cases and selected controls; the reading by pathologists in the local institutions is considered a surrogate measurement and available for the entire cohort.

The results are summarized in Table 4. Under all situations, we observe that the estimates from the proposed method, the IPW method and Thomas' approach are all similar to the full-cohort estimates. The standard error estimates from the proposed method are uniformly smaller than those from Thomas' estimator indicating that the proposed estimator is more efficient by incorporating auxiliary information from the full cohort. In addition, as observed in our simulations, the empirical efficiency gain of the proposed method over Thomas' estimator is evident when the number of controls in the NCC study is small. When the number of controls increases, the efficiencies of all estimators approach to the full cohort estimator. Moreover, for those covariates whose true values are available for the entire cohort, the proposed estimator approaches the full-cohort efficiency very fast and achieves higher relative efficiency even with a small number of controls compared to Thomas' estimator. For example, for the covariate of tumor stage (Stage), the relative efficiency of our proposed estimator with respect to the full-cohort estimator already achieves 78% with just one control while it is only 33% for Thomas' estimator.

4. Concluding remarks

We show that the projection idea can be well employed in NCC studies with auxiliary covariates and can lead to an improved estimator for the regression parameters in Cox's model. The efficiency gains of our proposed estimator over Thomas' estimator are large when the number of controls in NCC study is small and the correlation between true covariates and auxiliary covariates is strong. When condition C2 (6) is violated, the proposed projection estimator is theoretically biased but the bias is usually small in realistic situations. In addition, our simulation studies showed that the bias was often negligible compared to the variance. The proposed approach is computationally convenient and can be

implemented using common statistical software with a little programming effort. The R-code for implementing the proposed approach can be obtained from the authors.

In this paper, the proposed estimator builds on Thomas' estimator, which is the most commonly used method in practice for analyzing NCC data and only requires the true covariates to be measured for the cases and selected controls at case failure times rather than the entire history of the true covariate process. But when extended NCC data are available, Thomas' estimator is not semiparametrically efficient (Robins, et al. 1994). In fact, as we observed in simulation studies, the IPW estimator (Samuelsens, 1997) often performed well. Therefore, statistical methods that can make use of the auxiliary covariate information to further improve the efficiency of the IPW estimator are also of our great interest. More specifically, replace the estimating equations used in (2) and (7) by

$$U_A^*(\theta) = \sum_{i=1}^n \int_0^\tau \pi_i \{A_i(t) - \bar{A}_w(t; \theta)\} dN_i(t) = 0,$$

where $\bar{A}_w(t; \theta) = \frac{\sum_j \pi_j Y_j(t) e^{\gamma' A_j(t)} A_j(t)}{\sum_j \pi_j Y_j(t) e^{\gamma' A_j(t)}}$ with $(A, \theta) = (Z, \beta)$ for the true model and $(A, \theta) = (X, \gamma)$ for the working model. It is easy to show that the solution to the above estimating equation and the corresponding full cohort estimator always converge to the same limit under either the true model or the working model. Thus, the IPW-based projection method may relax the proportional assumption on the induced hazard function in condition C2. This research will be investigated elsewhere.

5. Supplementary Materials

The Web Appendices referenced in Section 2.3 and Section 3.1 are available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

Acknowledgments

The authors thank the associate editor and two referees for their comments that substantially improved the presentation of the paper. This work was partially supported by NIEHS Pilot Project (M. Liu) and National Science Foundation Grant DMS-0504269 (W. Lu).

References

- Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of The American Statistical Association*. 1999; 94:896–908.
- Chen KN. Statistical estimation in the proportional hazards model with risk set sampling. *Annals of Statistics*. 2004; 32:1513–1532.
- Chen YH. Cox regression in cohort studies with validation sampling. *Journal of the Royal Statistical Society, Series B*. 2002; 64:51–62.
- Chen YH, Chen H. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society, Series B*. 2000; 62:449–460.
- Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
- D'Angio GJ, Breslow N, Beckwith JB, Evans A, Baum H, de Lorimier A, Ferbach D, Hrabovsky E, Jones G, Kelalis P. Results of the Third National Wilms Tumor Study. *Treatment of Wilms' tumor. Cancer*. 1989; 64:349–360. [PubMed: 2544249]
- Efron B. Bootstrap methods - another look at the jackknife. *Annals of Statistics*. 1979; 7:1–26.

- Green DM, Breslow NE, Beckwith JB, Finklestein JZ, Grundy PG, Thomas PRM, Kim T, Shochat S, Haase GM, Ritchey ML, Kelalis PP, D'Angio GJ. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: A report from the National Wilms' Tumor Study Group. *Journal of Clinical Oncology*. 1998; 16:237–245. [PubMed: 9440748]
- Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in the cox regression-model. *Annals of Statistics*. 1992; 20:1903–1928.
- Jiang JC, Zhou HB. Additive hazard regression with auxiliary covariates. *Biometrika*. 2007; 94:359–369.
- Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of The American Statistical Association*. 2004; 99:832–844.
- Lin DY, Wei LJ. The robust inference for the cox proportional hazards model. *Journal of The American Statistical Association*. 1989; 84:1074–1078.
- Oakes D. Survival times - aspects of partial likelihood. *International Statistical Review*. 1981; 49:235–252.
- Prentice R. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*. 1982; 69:331–342.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression-coefficients when some regressors are not always observed. *Journal of The American Statistical Association*. 1994; 89:846–866.
- Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*. 1997; 84:379–394.
- Scheike TH, Juul A. Maximum Likelihood Estimation for Cox's Regression Model under Nested Case-Control Sampling. *Biostatistics*. 2004; 5:193–206. [PubMed: 15054025]
- Thomas, DC. Addendum to "Methods of Cohort Analysis - Appraisal by Application to Asbestos Mining". In: Liddell, FDK.; McDonald, JC.; Thomas, DC., editors. *Journal of the Royal Statistical Society, Series A*. Vol. 140. 1977. p. 469-491.
- Xiang AH, Langholz B. Comparison of case-control to full cohort analyses under model misspecification. *Biometrika*. 1999; 86:221–226.
- Xiang AH, Langholz B. Robust variance estimation for rate ratio parameter estimates from individually matched case-control data. *Biometrika*. 2003; 90:741–746.
- Zeng D, Lin DY, Avery CL, North KE, Bray MS. Efficient Semi-parametric Estimation of Haplotype-Disease Associations in Case-Cohort and Nested Case-Control Studies. *Biostatistics*. 2006; 7:486–502. [PubMed: 16500923]

Table 1

Simulation results with independent censoring

Model	β	σ_e	$m-1$	Bias	SD	SE	CP	CP*	RE	RE ¹	RE ²	RE ³
S1	0		2	0.014	0.240	0.227	94.4	94.0	63.9	64.0	61.0	66.7
			4	0.010	0.216	0.205	93.6	93.8	78.8	78.6	74.8	81.4
			2	0.009	0.229	0.224	95.0	95.0	60.0	60.3	61.0	63.9
			4	0.005	0.203	0.202	96.0	96.0	76.1	76.6	74.2	81.7
	-1		2	0.010	0.243	0.236	94.8	94.8	53.9	53.7	58.7	58.8
			4	0.005	0.212	0.209	94.6	94.4	70.6	71.2	71.8	76.7
	-0.5		2	0.014	0.219	0.207	93.4	94.0	76.6	64.0	61.0	66.7
			4	0.011	0.204	0.195	93.4	93.8	88.1	78.6	74.8	81.4
			0.2	0.014	0.199	0.189	93.6	94.0	92.8	64.0	61.0	66.7
			4	0.012	0.193	0.186	93.4	93.8	99.1	78.6	74.8	81.4
S2	-0.5	0.5	2	0.015	0.203	0.204	94.8	95.0	75.9	60.3	61.0	63.9
			4	0.013	0.193	0.191	95.8	96.0	84.0	76.6	74.2	81.7
			0.2	0.021	0.184	0.187	95.0	95.0	92.9	60.3	61.0	63.9
			4	0.018	0.183	0.181	95.0	96.0	93.5	76.6	74.2	81.7
	-1	0.5	2	0.020	0.221	0.214	93.2	94.8	65.1	53.7	58.7	58.8
			4	0.012	0.203	0.197	94.0	94.4	77.1	71.2	71.8	76.7
			0.2	0.031	0.192	0.187	93.4	94.8	86.6	53.7	58.7	58.8
			4	0.020	0.190	0.184	92.8	94.4	88.1	71.2	71.8	76.7

SD: sample standard deviation of the proposed estimates from 500 runs; SE: average standard error estimates of 500 runs; CP: coverage probability of 95% Wald-type confidence interval using the bootstrap method; CP*: coverage probability of 95% Wald-type confidence interval using the rare-disease asymptotic variance estimator; empirical relative efficiency of each estimator is calculated by the ratio of sample variances of the estimator with that of the full cohort maximum partial likelihood estimator under the true model; RE: relative efficiency of the proposed estimator; RE¹ relative efficiency of Thomas' estimator; RE² relative efficiency of Chen (2004) estimator; RE³ relative efficiency of Samuelsen (1997) estimator.

Table 2

Simulation results with covariate-dependent censoring

Model	β	σ_e	$m-1$	Bias	SD	SE	CP	CP*	RE	RE ¹	RE ²	RE ³
S1	0	2	4	0.018	0.237	0.224	94.0	93.8	65.7	65.1	63.1	64.2
				0.013	0.217	0.204	93.6	93.4	78.2	77.4	75.3	80.6
	-0.5	2	4	0.008	0.236	0.231	95.4	95.2	64.5	64.4	62.5	68.5
				0.007	0.205	0.210	96.2	96.2	85.4	85.9	83.4	89.3
	-1	2	4	-0.008	0.250	0.250	95.0	94.8	57.6	58.1	60.0	61.7
				-0.003	0.222	0.223	95.8	95.4	72.6	72.7	72.8	74.8
S2	0	0.5	4	0.015	0.217	0.206	95.2	95.2	78.0	65.1	63.1	64.2
				0.014	0.203	0.194	94.8	93.2	89.5	77.4	75.3	80.6
	0.2	2	4	0.013	0.199	0.189	94.2	94.2	92.8	65.1	63.1	64.2
				0.014	0.194	0.184	95.0	92.6	98.1	77.4	75.3	80.6
	-0.5	0.5	4	0.010	0.225	0.214	93.8	93.6	71.0	64.4	62.5	68.5
				0.013	0.197	0.199	94.2	93.6	92.7	85.9	83.4	89.3
	0.2	2	4	0.020	0.204	0.195	93.6	94.2	86.6	64.4	62.5	68.5
				0.019	0.190	0.188	93.8	92.4	99.4	85.9	83.4	89.3
	-1	0.5	4	0.011	0.223	0.230	94.8	95.6	72.6	58.1	60.0	61.7
				0.010	0.209	0.211	95.2	95.0	81.7	72.7	72.8	74.8
	0.2	2	4	0.029	0.204	0.204	94.6	95.8	86.6	58.1	60.0	61.7
				0.023	0.196	0.196	94.4	95.4	93.3	72.7	72.8	74.8

SD: sample standard deviation of the proposed estimates from 500 runs; SE: average standard error estimates of 500 runs; CP: coverage probability of 95% Wald-type confidence interval using the bootstrap method; CP*: coverage probability of 95% Wald-type confidence interval using the rare-disease asymptotic variance estimator; empirical relative efficiency of each estimator is calculated by the ratio of sample variances of the estimator with that of the full cohort maximum partial likelihood estimator under the true model; RE: relative efficiency of the proposed estimator; RE¹ relative efficiency of Thomas' estimator; RE² relative efficiency of Chen (2004) estimator; RE³ relative efficiency of Samuelsen (1997) estimator.

Table 3

Simulation results of sensitivity analysis

Model	β	Bias	SD	SE	MSE	CP	RE	RE ¹
$X = Z + u$	0	0.005	0.171	0.169	0.029	95.4	81.1	66.0
	-1	0.007	0.190	0.190	0.036	95.2	72.4	58.8
	-2	0.026	0.274	0.264	0.076	94.0	53.7	45.7
$X = \epsilon Z$	0	-0.000	0.235	0.225	0.055	94.0	64.4	63.8
	-1	-0.014	0.242	0.235	0.059	94.4	55.5	55.2
	-2	-0.040	0.318	0.294	0.103	94.2	32.5	32.0
$X = Z_1$	(0,0)	β_1	0.175	0.176	0.031	94.8	93.5	63.9
		β_2	0.014	0.228	0.224	0.052	95.2	67.1
	(-1,-1)	β_1	0.011	0.205	0.209	0.042	94.6	95.6
		β_2	-0.018	0.257	0.249	0.066	95.2	54.1
	(-2,-2)	β_1	0.062	0.398	0.305	0.162	93.3	63.2
		β_2	-0.052	0.358	0.350	0.131	96.6	40.1
$X = Z + 0.2 \log(T^*) + \epsilon$	0	0.046	0.230	0.219	0.055	94.2	69.1	64.0
	-0.5	0.055	0.205	0.210	0.045	94.8	74.4	60.3
	-1	0.098	0.217	0.213	0.057	90.0	67.6	53.7
	-1.5	0.117	0.238	0.237	0.070	89.8	62.1	45.7

SD: sample standard deviation of the proposed estimates from 500 runs; SE: average standard error estimates of 500 runs; CP: coverage probability of 95% Wald-type confidence interval using the bootstrap method; CP*: coverage probability of 95% Wald-type confidence interval using the rare-disease asymptotic variance estimator; empirical relative efficiency of each estimator is calculated by the ratio of sample variances of the estimator with that of the full cohort maximum partial likelihood estimator under the true model; RE: relative efficiency of the proposed estimator; RE¹: relative efficiency of Thomas' estimator; RE²: relative efficiency of Chen (2004) estimator; RE³: relative efficiency of Samuelsen (1997) estimator.
MSE: mean squared error defined as $E(\hat{\beta} - \beta_0)^2$.

Table 4

Wilms' tumor study: parameter estimates and standard errors

	$m - 1 = 1$			$m - 1 = 2$			$m - 1 = 3$			Full
	β	$\tilde{\beta}$	β_s	β	$\tilde{\beta}$	β_s	β	$\tilde{\beta}$	β_s	
Age1	-0.682 (0.431)	-0.632 (0.282)	-0.717 (0.440)	-0.504 (0.377)	-0.687 (0.331)	-0.564 (0.346)	-0.981 (0.370)	-0.755 (0.333)	-0.788 (0.326)	-0.661 (0.321)
Age2	0.144 (0.029)	0.096 (0.021)	0.121 (0.029)	0.072 (0.022)	0.097 (0.018)	0.100 (0.020)	0.105 (0.021)	0.107 (0.017)	0.093 (0.018)	0.104 (0.015)
UH	4.757 (1.512)	4.381 (1.037)	4.219 (0.632)	4.496 (1.071)	4.617 (0.893)	3.977 (0.706)	3.885 (0.911)	4.033 (0.843)	3.565 (0.614)	4.042 (0.503)
Age1×UH	-3.380 (1.615)	-2.988 (1.093)	-2.819 (0.751)	-2.880 (1.155)	-3.221 (0.953)	-2.334 (0.770)	-2.394 (0.978)	-2.551 (0.900)	-2.165 (0.672)	-2.635 (0.552)
Age2×UH	-0.108 (0.076)	-0.089 (0.069)	-0.095 (0.075)	-0.050 (0.065)	-0.027 (0.059)	-0.079 (0.054)	-0.113 (0.048)	-0.109 (0.045)	-0.079 (0.041)	-0.058 (0.033)
Stage	1.428 (0.450)	1.331 (0.293)	1.219 (0.446)	1.427 (0.346)	1.397 (0.282)	1.399 (0.349)	1.384 (0.317)	1.266 (0.274)	1.338 (0.297)	1.346 (0.259)
Diameter	0.081 (0.022)	0.072 (0.016)	0.074 (0.024)	0.078 (0.018)	0.078 (0.015)	0.076 (0.019)	0.062 (0.017)	0.065 (0.015)	0.065 (0.016)	0.069 (0.015)
Sig×Diam	-0.090 (0.036)	-0.078 (0.023)	-0.067 (0.036)	-0.074 (0.027)	-0.084 (0.022)	-0.086 (0.027)	-0.078 (0.025)	-0.071 (0.021)	-0.075 (0.023)	-0.076 (0.020)

β : Thomas' maximum partial likelihood estimator; $\tilde{\beta}$: the proposed projection estimator; β_s Samuelsen's estimator; the numbers in the parentheses are the standard errors for the estimates above.