Published in final edited form as:

Biometrics. 2018 September ; 74(3): 1014–1022. doi:10.1111/biom.12838.

Regression Analysis for Secondary Response Variable in a Case-Cohort Study

Yinghao Pan¹, Jianwen Cai¹, Sangmi Kim², and Haibo Zhou^{1,*}

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

²Medical College of Georgia, GRU Cancer Center, Augusta University, Augusta, Georgia 30912, U.S.A

Summary

Case-cohort study design has been widely used for its cost-effectiveness. In any real study, there are always other important outcomes of interest beside the failure time that the original case-cohort study is based on. How to utilize the available case-cohort data to study the relationship of a secondary outcome with the primary exposure obtained through the case-cohort study is not well studied. In this paper, we propose a nonparametric estimated likelihood approach for analyzing a secondary outcome in a case-cohort study. The estimation is based on maximizing a semiparametric likelihood function that is built jointly on both time-to-failure outcome and the secondary outcome. The proposed estimator is shown to be consistent, efficient, and asymptotically normal. Finite sample performance is evaluated via simulation studies. Data from the Sister Study is analyzed to illustrate our method.

Keywords

Case-cohort design; Estimated likelihood; Secondary outcome; Semiparametric; Validation sample

1. Introduction

The case-cohort study design (Prentice, 1986) is a cost-effective sampling strategy that is used to study the association between an often expensive exposure variable and a time-to-event outcome. One example of a case-cohort study is the Sister Study conducted by NIEHS (Kim et al., 2011, 2013), where the time-to-event outcome of interest is time to breast cancer and the expensive exposure measure is a major prostaglandin E2 metabolite (PGE-M). The Sister Study targets US women who have a sister with breast cancer, but with no breast cancer themselves at the enrollment. The case-cohort study design can be viewed as a two-phased study. In the first phase, information on the time-to-event outcome and some relatively easy-to-obtain covariates are measured on all cohort members. In the second phase, the exposure variable of interest is measured in a random sample of the full cohort, plus those subjects who experienced the event of interest. The simple random sample here is

* zhou@bios.unc.edu.

referred as subcohort and those failures are referred as the cases. Many methods have been proposed to estimate the hazard ratio parameters with case-cohort data, e.g. Prentice (1986); Self and Prentice (1988); Chen and Lo (1999); Borgan et al. (2000); Kulich and Lin (2004); Kang and Cai (2009); Kim et al. (2013). For binary response outcome-dependent-sampling data, related work was done by Wang and Zhou (2006, 2010).

With tremendous cost and efforts involved in collecting the exposure variable in case-cohort study, interest rises to use the collected case-cohort data to study the relationship between other important responses to the exposure. Relative to the failure time used to design the case-cohort study, these important responses are referred as the secondary responses. For instance, in our Sister Study, investigators are interested in studying the relationship between BMI and PGE-M since recent research has indicated that there might be a positive association between obesity and prostaglandin E2 (Morris et al., 2011; Subbaramaiah et al., 2011). How to analyze the secondary outcome (BMI in this case) in a case-cohort study efficiently and correctly is not a straightforward exercise, as the original study data are obtained in an outcome-dependent way (depend on the primary outcome, i.e. the time-to-event variable). This is an issue that puzzled investigators who try to take advantage of the exposure variable measured in the original case-cohort study, yet not sure how to handle the biased sampling nature of the data based on the primary time-to-event outcome. Naively treating the case-cohort sample as a simple random sample could cause biased inference, as later shown in our simulation studies. An inefficient approach is to take the subcohort portion of the biased sampling data, and ignore those of cases. Clearly, this approach is discarding a big portion of the data. A significant amount of work was done on secondary analysis of case-control or nested case-control data. This includes the likelihood-based approach (Lee et al., 1997; Jiang et al., 2006; Saarela et al., 2008; Lin and Zeng, 2009; Salim et al., 2014); inverse probability weighting (Richardson et al., 2007; Monsees et al., 2009), and estimating equation (Wei et al., 2013; Ma and Carroll, 2016). Saarela et al. (2012) proposed a conditional likelihood method for secondary analysis under general two-phase cohort sampling designs, which includes case-cohort design. However, there has been a lack of research on the secondary response regression analysis in the case-cohort study in general.

Our research is motivated by the need in the Sister Study where we aim to establish the relationship between PGE-M and BMI. In this paper, we propose an estimated likelihood method for linear regression analysis of a continuous secondary response variable using case-cohort data. We jointly model the time-to-event outcome (time to breast cancer) and the continuous secondary response (BMI). The likelihood function involves the conditional distribution of the expensive exposure given other inexpensive covariates. We estimate it in a nonparametric fashion. We compare our proposed estimated likelihood estimator to two inverse probability weighting (IPW) type estimators we developed, and show that the estimated likelihood method has greater statistical efficiency. The advantage of our proposed method is that it is efficient, and yet require no strong parametric assumptions. The performance of our estimator is explored under a variety of conditions where complications could arrive.

The organization of the paper is as follows. In Section 2, we present some notations, data structure and model for case-cohort design. In Section 3, we outline the estimation algorithm for our proposed estimated likelihood estimator and establish its asymptotic properties. We further develop two new IPW type estimators in Section 4. In Section 5, we investigate the finite sample performance of our proposed estimators via simulation studies. In Section 6, we apply our method to Sister Study data. Final remarks are given in Section 7.

2. Data structure and model

We consider efficient inference of a continuous secondary response, denoted by Y_2 , with respect to an expensive exposure, denoted by X , in a case-cohort design. To fix notation, let $\tilde{T}$ denote the primary event time of interest and C the censoring time. Let $T = \min(\tilde{T}, C)$, and $\Delta = I(\tilde{T} \leq C)$, where T is the observation time, and Δ is the event indicator. Throughout the paper, we refer to individuals who have the event as cases ($\Delta = 1$) and censored individuals as non-cases ($\Delta = 0$). Furthermore, let $(X, \mathbf{Z})$ denote the vector of covariates with X being the expensive scalar covariate obtained only for the subcohort and the cases, and $\mathbf{Z}$ being the other first-phase covariates. $\mathbf{Z}$ can be either discrete or continuous variables. We assume that $\tilde{T}$ and C are conditionally independent given $(Y_2, X, \mathbf{Z})$, and the censoring time C does not depend on X but can depend on Y_2 and $\mathbf{Z}$.

We assume that the underlying data $\{(T_i, \Delta_i, Y_{2i}, X_i, \mathbf{Z}_i), i = 1, \dots, N\}$ are independent and identically distributed random vectors, where N denotes the size of the full cohort. Case-cohort studies can be considered as two-phase studies: in the first phase, information on observation time, event indicator, secondary response, and inexpensive covariates are gathered for each member of the full cohort. That is, we observe $\{(T_i, \Delta_i, Y_{2i}, \mathbf{Z}_i), i = 1, \dots, N\}$. In the second phase, covariate X is measured for subjects in the subcohort, and those who experienced the event of interest ($\Delta_i = 1$). Let V_0 be the index set of the simple random sample taken from the baseline cohort, and V_1 be the index set of the remaining cases in the full cohort. Then we observe $\{X_i, i \in V_0 \cup V_1\}$ in the second phase. Let n_{SRS} be the size of the subcohort, n_{V_1} be the size of the supplemental cases. Here, n_{SRS} is a pre-specified number and n_{V_1} is a random variable.

Let $n_V = n_{SRS} + n_{V_1}$ be the number of individuals for which we observed X , and let $n_{\bar{V}} = N - n_V$ be the number of individuals for whom X is not observed. Borrowing the terminology from measurement error literature, we will refer to the n_V complete observations as the validation sample, and $n_{\bar{V}}$ incomplete observations as non-validation sample. Let $V = V_0 \cup V_1$ be the index set of the validation sample, and $\bar{V}$ be the index set of the non-validation sample. In addition, let R_i indicates whether the i th subject is selected into the validation sample, then $V = \{i: R_i = 1\}$ and $\bar{V} = \{i: R_i = 0\}$. The data structure for the problem we consider is the following:

$$\begin{aligned}
\text{First phase:} & \{T_i, \Delta_i, Y_{2i}, \mathbf{Z}_i\}, \quad i = 1, \dots, N; \\
\text{Second phase: \{subchort sample\}} & \{X_i\}, \quad i \in V_0; \\
\text{\{supplemental cases\}} & \{X_i | \Delta_i = 1, i \notin V_0\}, \quad i \in V_1.
\end{aligned} \tag{1}$$

The main interest in this paper is to model the association between X and the secondary response Y_2 , adjusting for $\mathbf{Z}$ in the population from the following linear model:

$$Y_2 = \beta_0 + \beta_1 X + \beta_2 \mathbf{Z} + \varepsilon, \quad (2)$$

where $\varepsilon \sim N(0, \sigma^2)$, and ε 's are independent. We note that model (2) above, not model (3) below, is the main model for this paper. Our goal is to develop an efficient inference procedure on $\beta = (\beta_0, \beta_1, \beta_2)$.

Without loss of generality, we assume the following Cox model (Cox, 1972) for the primary response variable:

$$\lambda(t | X, Y_2, \mathbf{Z}) = \lambda_0(t) \exp(\gamma_1 X + \gamma_2 Y_2 + \gamma_3 \mathbf{Z}), \quad (3)$$

where $\lambda(t | X, Y_2, \mathbf{Z})$ is the conditional hazard function of $\tilde{T}$ given $(X, Y_2, \mathbf{Z})$, and $\lambda_0(t)$ is the baseline hazard function. Let $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ denote the baseline cumulative hazard function.

3. Estimated likelihood approach

Let $f(\cdot | \cdot)$ denote the conditional density function, and $G(\cdot | \cdot)$ denote the conditional distribution function. Let $\mathbf{W}$ be the informative components of $\mathbf{Z}$ about X , in the sense that $G_{X|\mathbf{Z}}(x | \mathbf{z}) = G_{X|\mathbf{W}}(x | \mathbf{w})$. If we are to jointly model (T, Y_2) , then the full-information likelihood for data in (1) is proportional to

$$L_N(\xi, \Lambda_0(t)) = \prod_{i \in V} f(T_i, \Delta_i, Y_{2i} | X_i, \mathbf{Z}_i) \times \prod_{j \in \bar{V}} \left\{ \int_x f(T_j, \Delta_j, Y_{2j} | x, \mathbf{Z}_j) dG_{X|\mathbf{W}}(x | \mathbf{W}_j) \right\}, \quad (4)$$

where $\xi = (\beta, \gamma, \sigma^2)$. The log-likelihood of (4) takes the form:

$$\begin{aligned}
l_N(\xi, \Lambda_0(t)) &= \sum_{i \in V} \left\{ \log f_{\gamma}(T_i, \Delta_i | X_i, Y_{2i}, \mathbf{Z}_i) + \log f_{\beta, \sigma^2}(Y_{2i} | X_i, \mathbf{Z}_i) \right\} \\
&+ \sum_{j \in \bar{V}} \log \left\{ \int_x f_{\gamma}(T_j, \Delta_j | x, Y_{2j}, \mathbf{Z}_j) f_{\beta, \sigma^2}(Y_{2j} | x, \mathbf{Z}_j) dG_{X|\mathbf{W}}(x | \mathbf{W}_j) \right\}
\end{aligned} \tag{5}$$

The parameters contained in the above equation involve $(\beta, \gamma, \sigma^2)$, though our main focus is on the inference of β while γ, σ^2 are nuisance parameters. Later, we will show that our proposed estimated likelihood estimator is more efficient than IPW type estimators because we took a joint likelihood approach and incorporated the failure outcome information (T, δ) in the full cohort. Note that the conditional distribution of X given $\mathbf{W}$, $G_{X|\mathbf{W}}$ is involved in (4) and (5).

Thus, we propose to work with the following estimated likelihood:

$$\begin{aligned} \hat{l}_N(\xi, \Lambda_0(t)) = & \sum_{i \in V} \left\{ \log f_{\gamma}(T_i, \Delta_i | X_i, Y_{2i}, \mathbf{Z}_i) + \log f_{\beta, \sigma^2}(Y_{2i} | X_i, \mathbf{Z}_i) \right\} \quad (6) \\ & + \sum_{j \in \bar{V}} \log \left\{ \int_x f_{\gamma}(T_j, \Delta_j | x, Y_{2j}, \mathbf{Z}_j) f_{\beta, \sigma^2}(Y_{2j} | x, \mathbf{Z}_j) d\hat{G}_{X|\mathbf{W}}(x | \mathbf{W}_j) \right\} \end{aligned}$$

where we nonparametrically estimate the conditional distribution $G_{X|\mathbf{W}}$. For discrete $\mathbf{W}$, let

$$\hat{G}_{X|\mathbf{W}}(x | \mathbf{w}) = \sum_{i \in V_0} I(X_i \leq x, \mathbf{W}_i = \mathbf{w}) / \sum_{i \in V_0} I(\mathbf{W}_i = \mathbf{w}).$$

For continuous $\mathbf{W}$, we use the kernel method to estimate $G_{X|\mathbf{W}}$.

$$\hat{G}_{X|\mathbf{W}}(x | \mathbf{w}) = \sum_{i \in V_0} I(X_i \leq x) K_{\mathbf{H}}(\mathbf{W}_i - \mathbf{w}) / \sum_{i \in V_0} K_{\mathbf{H}}(\mathbf{W}_i - \mathbf{w}).$$

where $K_{\mathbf{H}}(\cdot) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2} \cdot)$ is a multivariate kernel with a bandwidth matrix $\mathbf{H}$. The optimal bandwidth $\mathbf{H}$ can be selected using cross-validation, or using an ad-hoc value. For instance, when $\mathbf{W}$ is univariate, a simple bandwidth selection is $H = 2\hat{\sigma}_W \cdot n_{SRS}^{-1/3}$, where $\hat{\sigma}_W$ is the estimated standard deviation of W .

In the estimated likelihood expression (6), $f_{\gamma}(T_i, \Delta_i | X_i, Y_{2i}, \mathbf{Z}_i)$ and $f_{\beta, \sigma^2}(Y_{2i} | X_i, \mathbf{Z}_i)$ can be

written in explicit forms based on linear model (2) and Cox model (3). Under the assumption that $\tilde{T}$ and C are conditionally independent, $f_{\gamma}(T_i, \Delta_i | X_i, Y_{2i}, \mathbf{Z}_i)$ is equal to

$$\left\{ \lambda_0(T_i) \exp(\gamma_1 X_i + \gamma_2 Y_{2i} + \gamma_3 \mathbf{Z}_i) \right\}^{\Delta_i} e^{-\Lambda_0(T_i) \exp(\gamma_1 X_i + \gamma_2 Y_{2i} + \gamma_3 \mathbf{Z}_i)} \lambda_C^{1 - \Delta_i}(T_i | X_i, Y_{2i}, \mathbf{Z}_i) e^{-\Lambda_C(T_i | X_i, Y_{2i}, \mathbf{Z}_i)}$$

$\lambda_C(T | X, Y_2, \mathbf{Z})$ is the hazard function of C given $(X, Y_2, \mathbf{Z})$, and Λ_C is the cumulative hazard. Assuming that C is independent of X , $\log f_{\gamma}(T_i, \Delta_i | X_i, Y_{2i}, \mathbf{Z}_i)$ is the sum of

$$\Delta_i \log \lambda_0(T_i) + \Delta_i (\gamma_1 X_i + \gamma_2 Y_{2i} + \gamma_3 \mathbf{Z}_i) - \Lambda_0(T_i) \exp(\gamma_1 X_i + \gamma_2 Y_{2i} + \gamma_3 \mathbf{Z}_i)$$

and $(1 - \Delta_i) \log \lambda_C(T_i | Y_{2i}, \mathbf{Z}_i) - \Lambda_C(T_i | Y_{2i}, \mathbf{Z}_i)$. The latter term does not involve x or $(\xi, \Lambda_0(t))$, and thus irrelevant to estimating the parameters of interest. Also,

$$\log f_{\beta, \sigma^2}(Y_{2i}|X_i, \mathbf{Z}_i) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_{2i} - \beta_0 - \beta_1 X_i - \beta_2 \mathbf{Z}_i)^2.$$

The estimated log-likelihood $\hat{l}_N(\boldsymbol{\xi}, \Lambda_0(t))$ has a semi-parametric form, with $\boldsymbol{\xi}$ being the parametric part and $\Lambda_0(t)$ the nonparametric component. Because the baseline cumulative hazard $\Lambda_0(t)$ is unknown, we cannot directly maximize $\hat{l}_N(\boldsymbol{\xi}, \Lambda_0(t))$ with respect to $\boldsymbol{\xi}$. We propose to use an inverse probability weighted partial likelihood estimates from Chen and Lo (1999) to obtain $(\hat{\gamma}_1^0, \hat{\gamma}_2^0, \hat{\gamma}_3^0)$ as initial parameter estimates for the Cox model (3). We then estimate the baseline cumulative hazard using a weighted Breslow estimator,

$$\hat{\Lambda}_0(t) = \sum_{j \in V} \frac{I(T_j \leq t) \Delta_j}{\sum_{l \in V} \hat{\pi}_l^{-1} I(T_l \geq T_j) \exp(\hat{\gamma}_1^0 X_l + \hat{\gamma}_2^0 Y_{2l} + \hat{\gamma}_3^0 Z_l)},$$

where $\hat{\pi}_l = 1$ for cases, and $\hat{\pi}_l = n_{SRS}/N$ for non-cases. Furthermore, we use the inverse probability weighted linear regression on the validation sample to obtain $(\hat{\beta}_0^0, \hat{\beta}_1^0, \hat{\beta}_2^0, \hat{\sigma}^0)$ as initial parameter estimates for the linear regression model (2), where cases receive unit weight, and non-cases have weight N/n_{SRS} .

Finally, we plug in the baseline cumulative hazard estimate $\hat{\Lambda}_0(t)$ into the estimated log-likelihood, and maximize it with respect to $\boldsymbol{\xi}$. Let

$$\tilde{l}_N(\boldsymbol{\xi}) = \hat{l}_N(\boldsymbol{\xi}, \hat{\Lambda}_0(t)).$$

then our proposed nonparametric estimated likelihood estimator $\hat{\boldsymbol{\xi}}_{NPEL}$ is the solution to the following estimating equation (7):

$$\frac{1}{N} \frac{\partial \tilde{l}_N(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = 0. \quad (7)$$

Newton-Raphson algorithm can now be invoked, where $(\hat{\gamma}_1^0, \hat{\gamma}_2^0, \hat{\gamma}_3^0, \hat{\beta}_0^0, \hat{\beta}_1^0, \hat{\beta}_2^0, \hat{\sigma}^0)$ are the starting values.

Next we present the asymptotic property of our proposed estimator. The following results are provided for the case when $\mathbf{W}$ is discrete. Asymptotic results for $\mathbf{W}$ being continuous can be derived similarly. We assume that as $N \rightarrow \infty$, $n_V/N \rightarrow \bar{\pi} > 0$ and $n_{SRS}/n_V \rightarrow \rho_{SRS} > 0$. Therefore, $\bar{\pi}$ is the proportion of validation sample in the overall population. ρ_{SRS} represents the proportion of simple random sample in the validation sample. In addition, we let E_k denote the conditional expectation given $\mathbf{W} = k$. That is, $E_k\{h(T, Y_2, X, \mathbf{Z})\} =$

$E\{h(T, \Delta, Y_2, X, \mathbf{Z}) \mid \Delta = k\}$. Let ξ_* be the true underlying value of ξ . Under some general regularity conditions (see Web Appendix), we have the following theorems.

Theorem 1

(i) $\hat{\xi}_{NPEL}$ is a consistent estimator of ξ_* and (ii) $\hat{\xi}_{NPEL}$ has the following asymptotic distributional properties:

$$\sqrt{N}(\hat{\xi}_{NPEL} - \xi_*) \xrightarrow{D} N(\mathbf{0}, \Sigma_{NPEL}(\xi_*)).$$

where $\Sigma_{NPEL}(\xi_*) = \mathbf{I}^{-1}(\xi_*) + \rho_{SRS} \bar{\pi} \mathbf{I}^{-1}(\xi_*) \Sigma(\xi_*) \mathbf{I}^{-1}(\xi_*)$, $\Sigma(\xi) = \text{var}_{X, \mathbf{W}}\{(1 - \bar{\pi})\mathbf{Q}_{X, \mathbf{W}}(\xi, \Lambda_0)\}$, and

$$\begin{aligned} \mathbf{I}(\xi) = & -\rho_{SRS} \bar{\pi} E \left\{ \frac{\partial^2 \log f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \xi)}{\partial \xi \partial \xi^T} \right\} - (1 - \bar{\pi}) E_0 \left\{ \frac{\partial^2 \log f(T, \Delta, Y_2 \mid \mathbf{Z}; \xi)}{\partial \xi \partial \xi^T} \right\} - (1 - \rho_{SRS}) \bar{\pi} E_1 \left\{ \frac{\partial^2 \log f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \xi)}{\partial \xi \partial \xi^T} \right\}, \end{aligned} \quad (8)$$

$$\begin{aligned} \mathbf{Q}_{X, \mathbf{W}}(\xi, \Lambda_0) = & \sum_{\mathbf{w}} P(\mathbf{W} = \mathbf{w} \mid \Delta = 0) \frac{I(\mathbf{W} = \mathbf{w})}{P(\mathbf{W} = \mathbf{w}) \rho_{SRS} \bar{\pi}} E_{T, \Delta, Y_2, \mathbf{Z} \mid \Delta = 0} \{ \mathbf{M}_X(T, \Delta, Y_2, \mathbf{Z}, \mathbf{W} = \mathbf{w}; \xi, \Lambda_0) \}. \end{aligned} \quad (9)$$

where

$$\begin{aligned} \mathbf{M}_X(T, \Delta, Y_2, \mathbf{Z}, \mathbf{W}; \xi, \Lambda_0) = & \frac{\partial f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \Lambda_0) / \partial \xi}{\int f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \Lambda_0) dG_{X \mid \mathbf{W}}(X \mid \mathbf{W})} \\ & - \frac{f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \Lambda_0) \int \partial f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \Lambda_0) / \partial \xi dG_{X \mid \mathbf{W}}(X \mid \mathbf{W})}{\left| \int f(T, \Delta, Y_2 \mid X, \mathbf{Z}; \Lambda_0) dG_{X \mid \mathbf{W}}(X \mid \mathbf{W}) \right|^2}. \end{aligned} \quad (10)$$

Replacing the population quantities in Theorem 1 with the sample quantities, a consistent estimator for asymptotic covariance matrix $\Sigma_{NPEL}(\xi_*)$ is provided by the following theorem:

Theorem 2

A consistent estimator for variance matrix $\Sigma_{NPEL}(\xi_*)$ can be obtained as

$$\widehat{\Sigma}_{NPEL}(\widehat{\xi}_{NPEL}) = \widehat{I}^{-1}(\widehat{\xi}_{NPEL}) + \frac{n_{SRS}}{N} \times \widehat{I}^{-1}(\widehat{\xi}_{NPEL}) \widehat{\Sigma}(\widehat{\xi}_{NPEL}) \widehat{I}^{-1}(\widehat{\xi}_{NPEL}), \quad (11)$$

where $\widehat{I}(\xi) = -\frac{1}{N} \frac{\partial^2 \widetilde{l}_N(\xi)}{\partial \xi \partial \xi^T}$, and $\widehat{\Sigma}(\xi) = \widehat{var}_{\{X_i, \mathbf{W}_i; i \in V_0\}} \left\{ \frac{N-n_V}{N} \widehat{Q}_{X_i, \mathbf{W}_i}(\xi) \right\}$, with

$$\widehat{Q}_{X_i, \mathbf{W}_i}(\xi) = \sum_{j \in V} \frac{1}{N-n_V} \times \frac{N \cdot I(\mathbf{W}_i = \mathbf{W}_j)}{\sum_{i \in V_0} I(\mathbf{W}_i = \mathbf{W}_j)} \times \left[\frac{\partial f(T_j, \Delta_j, Y_{2j} | X_i, \mathbf{Z}_j; \widehat{\Lambda}_0) / \partial \xi}{\widehat{f}(T_j, \Delta_j, Y_{2j} | \mathbf{Z}_j; \widehat{\Lambda}_0)} - \frac{f(T_j, \Delta_j, Y_{2j} | X_i, \mathbf{Z}_j; \widehat{\Lambda}_0) \partial \widehat{f}(T_j, \Delta_j, Y_{2j} | \mathbf{Z}_j; \widehat{\Lambda}_0) / \partial \xi}{\left\{ \widehat{f}(T_j, \Delta_j, Y_{2j} | \mathbf{Z}_j; \widehat{\Lambda}_0) \right\}^2} \right],$$

where $\widehat{f}(T_j, \Delta_j, Y_{2j} | \mathbf{Z}_j; \widehat{\Lambda}_0) = \int f(T_j, \Delta_j, Y_{2j} | x, \mathbf{Z}_j; \widehat{\Lambda}_0) d\widehat{G}_{X|\mathbf{W}}(x | \mathbf{W}_j)$.

Outline of the proofs of these two theorems are provided in Web Appendix A.

4. Derivation of two new IPW type estimators

In this section, we derive two inverse probability weighting type estimators for the secondary outcome analysis. One is the basic weighted estimator and the other one is the augmented IPW estimator. Though the general ideas of probability weighted estimator is there (Horvitz and Thompson, 1952), we develop out these ideas explicitly as competing estimators for the proposed $\widehat{\xi}_{NPEL}$. Later in the simulation study, we will compare our proposed estimator to these two IPW type estimators.

The first IPW estimator $\widehat{\beta}_{IPW}$, is a basic weighted least squares estimator, where the weight is taken as the inverse of the selection probability into the validation sample. That is, we want to minimize

$$\sum_{i \in V} \frac{1}{\pi_i} (Y_{2i} - \beta_0 - \beta_1 X_i - \beta_2 \mathbf{Z}_i)^2 = \sum_{i=1}^N \frac{R_i}{\pi_i} (Y_{2i} - \beta_0 - \beta_1 X_i - \beta_2 \mathbf{Z}_i)^2,$$

where π_i is the selection probability for the i th subject. In the case-cohort design, $\pi_i = 1$ for cases, as all cases are included into the validation sample. For non-cases, $\pi_i = n_{SRS}/N$, the probability of being selected into the SRS portion of the validation sample. Let

$\widetilde{X}_i = (1, X_i, \mathbf{Z}_i)$ be the vector of covariates for the i th subject, including the intercept. It is easy

to see that the IPW estimator $\widehat{\beta}_{IPW}$ satisfies the following score equation:

$$\sum_{i=1}^N \frac{R_i}{\pi_i} \tilde{\mathbf{X}}_i^T (Y_{2i} - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) = 0. \quad (12)$$

We notice that in (12), all cases will receive unit weights, and all non-cases selected into the validation sample will have weights larger than one. The inverse probability weighting method will give an unbiased estimator of the parameters $\boldsymbol{\beta}$. However it may also result in reduced efficiency, since the cases would always receive smaller weights than non-cases regardless of whether the case status is linked to the secondary outcome/exposure relationship.

The second IPW estimator, $\hat{\boldsymbol{\beta}}_{AIPW}$, is an augmented IPW that incorporates the available information in the full cohort. Similar to Robins et al. (1994), we have the following estimating equation:

$$\sum_{i=1}^N \left[\frac{R_i}{\pi_i} \tilde{\mathbf{X}}_i^T (Y_{2i} - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) + \left(1 - \frac{R_i}{\pi_i}\right) E_{X_i|Y_{2i}, \mathbf{Z}_i} \left\{ \tilde{\mathbf{X}}_i^T (Y_{2i} - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \right\} \right] = 0, \quad (13)$$

where an augmented term is added. The second term in (13) involves $E(X_i / Y_{2i}, \mathbf{Z}_i)$ and $E(X_i^2 | Y_{2i}, \mathbf{Z}_i)$. We use a linear regression to approximate these moments, i.e. $E(X_i / Y_{2i}, \mathbf{Z}_i) = \phi_0 + \phi_1 Y_{2i} + \phi_2 \mathbf{Z}_i$, $Var(X_i / Y_{2i}, \mathbf{Z}_i) = \tau^2$. Based on SRS portion of the data, a linear regression can be fitted to obtain parameter estimates for $(\phi_0, \phi_1, \phi_2, \tau^2)$. After plugging $(\hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2, \hat{\tau}^2)$ into (13), Newton-Raphson algorithm can be invoked to obtain $\hat{\boldsymbol{\beta}}_{AIPW}$.

5. Simulation studies

We conducted a series of simulation studies. First, we evaluated the performance of our estimator when the underlying assumption is met. Then, we further broaden out to evaluate the proposed method under situations where models are misspecified: (i) If $\mathbf{W}$, the informative components of $\mathbf{Z}$ about X , is misspecified. (ii) The error term ε in linear model (2) is not normally distributed. (iii) The proportional hazards assumption does not hold.

Five competing estimators are compared: (i) $\hat{\boldsymbol{\beta}}_N$, linear regression of the secondary response Y_2 on $(X, \mathbf{Z})$ based on the validation sample (union of subcohort and cases). This naive estimator ignores the biased sampling nature of the data, and can be grossly misleading; (ii) $\hat{\boldsymbol{\beta}}_{SRS}$, fitting a linear regression of Y_2 on $(X, \mathbf{Z})$ based on the SRS portion of the validation sample; (iii) $\hat{\boldsymbol{\beta}}_{IPW}$ denotes the inverse probability weighting method using only the observations from the validation sample; (iv) $\hat{\boldsymbol{\beta}}_{AIPW}$ denotes the augmented inverse probability weighting method we derive in section 4 which incorporates the available information in the full cohort; (v) $\hat{\boldsymbol{\beta}}_{NPEL}$ denotes the estimated likelihood method we proposed.

In the first set of studies, we let the full cohort size to be $N = 1000$. The data were generated by the following models: $Y_2 = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$, where $X \sim N(0, 1)$, $Z \sim \text{Bernoulli}(0.45)$, and $\varepsilon \sim N(0, 1)$. We set $\beta_0 = 1$, $\beta_2 = -0.5$, and allow β_1 to take value 0 or 0.5. The event time $\tilde{T}$ follows an exponential distribution with hazard function: $\lambda(t/X, Y_2, Z) = \lambda_0(t) \exp(\gamma_1 X + \gamma_2 Y_2 + \gamma_3 Z)$, with $\lambda_0(t) = 1$, $\gamma_1 = 0.5$, $\gamma_3 = -0.5$, and γ_2 takes value 0 or $\log(3)$. We assume that the censoring time follows a uniform distribution on interval $(0, C)$. When $\gamma_2 = 0$, $C = 0.52$ or 0.15 corresponds to 80% or 95% censoring, respectively. When $\gamma_2 = \log(3)$, $C = 0.16$ or 0.02 corresponds to 80% or 95% censoring, respectively.

From the full cohort, we first select a simple random sample of size $n_{SRS} = 200$, then we add all cases for the case-cohort sample. As we are mainly interested in studying the relationship between the secondary outcome Y_2 and the expensive exposure X , we focus on the estimation of β_1 . We assume that the informative components of $\mathbf{W}$ is Z . Since Z is discrete, we use the empirical distribution function to estimate the conditional distribution of X given Z . We report the mean of the parameter estimates, empirical standard deviation, mean of the estimated standard deviation, and 95% confidence interval coverage. Simulation results based on 1000 replications are shown in Table 1.

We find that all estimators yield approximately unbiased estimates except $\hat{\beta}_N$. $\hat{\beta}_N$ performs well when the secondary outcome is not correlated with the time-to-event outcome ($\gamma_2 = 0$). However, it has bias and poor coverage rate when $\gamma_2 \neq 0$. The bias is even larger when censoring percent increases (rare event). Similar findings were reported by Lin and Zeng (2009) in secondary analysis of case-control data.

From Table 1, we draw several other conclusions. (i) The proposed estimated likelihood estimator $\hat{\beta}_{NPEL}$ is the most efficient in all settings. (ii) The estimated standard deviation is very close to the empirical standard deviation (i.e. $\hat{SD}$ is close to SD). (iii) The 95% confidence interval coverage is close to 0.95, which implies that the asymptotic normal approximation works well in these situations. (iv) We also notice that when β_1 is significantly different from 0 (i.e. $\beta_1 = 0.5$), $\hat{\beta}_{AIPW}$ outperforms $\hat{\beta}_{IPW}$ by a certain margin. The difference between these two estimators is negligible when $\beta_1 = 0$.

We conducted further simulation studies to assess the performance of our proposed estimator under a wide range of proportion of SRS sample in the full cohort, and under various censoring percentages. Simulation results are provided in Web Appendix B.

Misspecification of $\mathbf{W}$, the informative components of $\mathbf{Z}$ about X

To study the performance of $\hat{\beta}_{NPEL}$ when $\mathbf{W}$ is misspecified, we generate the data by the following models: $Y_2 = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \varepsilon$, where $Z_1 \sim \text{Bernoulli}(0.45)$, $Z_2 \sim \text{Bernoulli}(0.6)$, $\varepsilon \sim N(0, 1)$. The expensive exposure X is simulated as: $X \sim N(Z_1, 1)$. Let $\mathbf{Z} = (Z_1, Z_2)$, then $G(X|\mathbf{Z}) = G(X|Z_1)$, i.e. the true $\mathbf{W}$ is Z_1 . We set $\beta_0 = 1$, $\beta_2 = -0.5$, $\beta_3 = 2$, and allow β_1 to take value 0 or 0.5. The event time $\tilde{T}$ follows an exponential distribution with hazard function: $\lambda(t/X, Y_2, \mathbf{Z}) = \lambda_0(t) \exp(\gamma_1 X + \gamma_2 Y_2 + \gamma_3 Z_1 + \gamma_4 Z_2)$, with $\lambda_0(t) = 1$, $\gamma_1 = 0.5$, $\gamma_2 = \log(1.2)$, $\gamma_3 = -0.5$ and $\gamma_4 = 1$. The censoring time follows a uniform

distribution on interval $(0, 0.14)$, which leads to 80% censoring. Same as before, $N = 1000$ and $n_{SR5} = 200$. We consider three scenarios: (i) We estimate the conditional distribution $G(X|\mathbf{Z})$ by assuming that $\mathbf{W}$ is Z_2 . This is the case where $\mathbf{W}$ is misspecified. (ii) We assume that $\mathbf{W}$ is Z_1 . (iii) We assume that $\mathbf{W}$ is (Z_1, Z_2) .

Table 2 summarizes the simulation results. Under null hypothesis ($\beta_1 = 0$), $\hat{\beta}_{NPEL}$ has good performance even if we misspecify $\mathbf{W}$. However, when β_1 is significantly different from 0 (i.e. $\beta_1 = 0.5$), misspecification of $\mathbf{W}$ (using Z_2 only) would lead to substantial bias and poor coverage probability of 95% CI. In the Discussion Section, we will provide practical guidance on how to choose $\mathbf{W}$. The general idea is that we should select the components of $\mathbf{Z}$ that correlates to X the most.

Parametric vs Non-parametric estimation of X given $\mathbf{Z}$

We compare our proposed estimator $\hat{\beta}_{NPEL}$ to a parametric maximum likelihood estimator $\hat{\beta}_{PL}$, where we parametrically specify $G_{X|\mathbf{Z}}$ (i.e. conditional normal) to a set of additional parameters λ , and then directly maximize the log-likelihood function in (5) with respect to ξ and λ . The simulation set up is the same as Table 2, except for X . We consider the following three scenarios: (i) $X \sim N(0, 1)$. (ii) X follows a gamma distribution with shape parameter 0.5 and rate parameter 1. (iii) X follows an exponential distribution with rate parameter 0.5.

Table 3 displays the results. $\hat{\beta}_{PL}$ has large bias and can be grossly misleading when $G_{X|\mathbf{Z}}$ is misspecified. On the other hand, our proposed estimator $\hat{\beta}_{NPEL}$ works well in all settings. Another disadvantage of parametric MLE is that it requires numerical integration over X , and is hence computationally intensive.

Misspecification of the error distribution for the secondary outcome or misspecification of the survival model

We conducted additional simulations to examine the performance of our estimator for the secondary outcome parameters when the error distribution for the secondary outcome model is misspecified or when the survival model is misspecified. The simulation set up is the same as Table 2, except for the error term and the model for time to event outcome. We consider five scenarios: (i) The error term is a gamma distribution with shape parameter 2, rate parameter 1, then normalized to have mean 0 and variance 1. This error term is right-skewed. (ii) The error term depends on Z_1 , which is a standard normal distribution multiplied by $(1 + Z_1^2)^{3/4}/2$. (iii) The error term depends on X , which is a standard normal distribution multiplied by $(1 + X^2)^{3/4}/2$. (iv) The event time $\tilde{T}$ comes from an Accelerated Failure Time (AFT) model: $\log \tilde{T} = 0.5X + Y_2 - 0.5Z_1 + Z_2 + e$ where $e \sim N(0, 1)$. (v) The event time follows an Additive Hazards model: $\lambda(t) = 10 + 0.5X + Y_2 - 0.5Z_1 + Z_2$.

Simulation results are summarized in Table 4, where we use Z_1 to estimate $G(X|\mathbf{Z})$. We have the following conclusions: (a) When the error term is skewed (not symmetric around 0), $\hat{\beta}_{IPW}$ and $\hat{\beta}_{AIPW}$ performs relatively well, while $\hat{\beta}_{NPEL}$ has larger bias. This comes from

the fact that the estimated likelihood estimator is more reliant on the normality assumption. On the other hand, the IPW and AIPW estimator we developed do not explicitly use the normality assumption of the error term. (b) When the error term is heteroscedastic, but not depend on X , the proposed $\hat{\beta}_{NPEL}$ is unbiased and the most efficient among all estimators. (c) When the error term depends on X , $\hat{\beta}_{NPEL}$ is still unbiased, but the asymptotic covariance estimator tends to underestimate the empirical covariance. (d) Even though the true model for $\tilde{T}$ is AFT or Additive Hazards and we fit Cox model, our proposed estimator $\hat{\beta}_{NPEL}$ still works very well in terms of estimating the parameters in model (2).

In practice, we could first fit a linear regression of Y_2 on (X, Z) based on SRS portion of the data to check the normality and homogeneity assumption. If we find that the residual depends on X , we need to employ some variable transformation techniques to obtain a more homogeneous error term. After that, our estimated likelihood method can be used.

6. Analysis of Sister Study data

We applied our method to data from the Sister Study (Kim et al., 2011, 2013) to assess the relationship between BMI and PGE-M. The Sister Study is a cohort study conducted on US women aged 35-74. To be eligible, women cannot have breast cancer themselves at the start of the study but have a sister with breast cancer. The participants come from all 50 states and Puerto Rico, and were recruited by health professionals, internet, and a national campaign. The enrollment time is from 2003 to 2009. Twenty five thousand and eight hundred subjects completed the baseline activities by June 1, 2007. We further restrict the study population to women who are postmenopausal, age 50 or older, and not currently using hormone replacement therapies, which yields a full cohort of size 11338.

Information on subject's breast cancer status is followed up via annual and biennial questionnaires until September 2010. Other information such as age, BMI, medication use are collected for each member in the full cohort ($N=11338$). In the original study, the primary interest is to assess the effect of PGE-M on the risk of breast cancer. Because the measurement of PGE-M level from urine samples can be expensive, a case-cohort design is implemented. A simple random sample of 300 participants were selected from the eligible full cohort. Then 307 supplemental cases were added to create a case-cohort sample. PGE-M levels were quantified for the subcohort members and the supplemental cases.

Prior research indicate that obesity might be positively related to prostaglandin E2 (Morris et al., 2011; Subbaramaiah et al., 2011). In this analysis, we use normalized BMI as the secondary outcome and normalized PGE-M as the exposure which is only available on the case-cohort sample. Additional confounding variables include use of nonsteroidal anti-inflammatory drugs (NSAIDs), current drinking status, current smoking status, indicator variables of older than 60 and waist circumference larger than 35 inches. We fit the following two models:

$$BMI = \beta_0 + \beta_1 PGEM + \beta_2 NSAIDS + \beta_3 DRINKING + \beta_4 AGE + \varepsilon, \lambda(t) = \lambda_0(t) \exp(\gamma_1 PGEM + \gamma_2 BMI + \gamma_3 NSAIDS + \gamma_4 AGE + \gamma_5 WAIST + \gamma_6 SMOKE),$$

where $\lambda(t)$ is the conditional hazard of breast cancer given PGE-M, BMI and other covariates. The estimated conditional distribution of the expensive exposure PGE-M given other covariates is based on use of NSAIDS, current drinking status, and an indicator variable of older than 60.

The results are shown in Table 5. All three analyses confirm that PGE-M level is significantly related to BMI. The proposed estimator in general provides a more precise estimate of the effects. Most of the variance estimates for the proposed method are smaller than those from competing methods. All three analyses confirm that there is a positive relationship between BMI and use of NSAIDS, and that there is a negative relationship between current drinking, age and BMI.

7. Discussion

To study the hard-to-obtain exposure variable that is only measured in a case-cohort study with some other important responses is desired by all investigators. However, how to properly account for the biased sampling nature of a case-cohort study in a secondary outcome regression analysis, or how to best utilize the available data to achieve a more efficient inference is not a well thought out issue. Without loss of generality, we assume a linear model for the secondary response and our goal in this paper is to obtain unbiased and efficient estimates for the linear model parameters in model (2). We propose a new and efficient estimator for the secondary outcome regression analysis in a case-cohort study. Our estimation is based on jointly modeling the time-to-event outcome and the secondary response. The joint likelihood is not tractable as the conditional distribution of expensive exposure given other inexpensive covariates as well as the baseline cumulative hazard are not specified. We estimate this conditional distribution and the baseline cumulative hazard function nonparametrically. Our proposed method has the advantage of making no new parametric assumptions, and yet still able to improve efficiency relative to other competing estimators.

In the simulation studies, we find that misspecification of $\mathbf{W}$ would lead to certain degree of bias. We suggest the following practical guidelines on how to choose $\mathbf{W}$. Due to “curse of dimensionality”, the $\mathbf{W}$ we use should not be more than 3 components in general, unless the SRS sample size is very large. When covariates $\mathbf{Z}$ has low dimension (i.e. $3 < \dim(\mathbf{Z}) < 10$), we suggest to calculate the correlation between X and each covariate in $\mathbf{Z}$, and then select 3 most correlated components. When the dimension of $\mathbf{Z}$ is high (i.e. ≥ 10), we suggest to use dimension-reduction techniques such as principal component analysis to get a lower dimensional $\tilde{\mathbf{Z}}$, then our proposed estimated likelihood method can be carried out by using $\tilde{\mathbf{Z}}$ as the regression covariates and the informative components $\mathbf{W}$.

Our proposed estimated likelihood method requires specifying a model for the time-to-event outcome. However, in the simulation studies, we found that our proposed estimator works very well in estimating the parameters in the secondary outcome model, even if we completely misspecified the model for the time-to-event outcome. In the future, it is

interesting to see if we can altogether avoid specifying a model for the time-to-event outcome, and still make valid and efficient inference on the secondary response.

Even though we present our method in the setting where the subcohort is selected as a simple random sample from the full cohort, our method also works when the subcohort sampling probabilities depend on $\mathbf{Z}$ (e.g. stratified random sampling with respect to $\mathbf{Z}$). This is because our method only involves estimating the conditional distribution $G_{X|\mathbf{Z}}$. When the subcohort sampling depends on the survival outcome, our method cannot be directly applied and needs modification. This is another area of our future research.

Our approach can be extended in several other ways. Although we use a linear model (2) for the association between the secondary response Y_2 and all covariates $(X, \mathbf{Z})$, our method can be applied to other discrete or continuous secondary outcomes, as long as the conditional distribution of Y_2 given $(X, \mathbf{Z})$ is specified parametrically. Our models can also be easily generalized to the case in which the linear model (2) and Cox model (3) have different covariates. In our paper, we use the empirical distribution to estimate the conditional distribution $G_{X|\mathbf{W}}$ when $\mathbf{W}$ is discrete, and use the kernel method when $\mathbf{W}$ is continuous. When $\mathbf{W}$ has both discrete and continuous components, the product kernel estimator from Li and Racine (2008) could be used. In addition, instead of assuming that the error term ϵ in the linear model (2) follows a normal distribution, we are working on a more robust semiparametric estimation where the only assumption we make is $E(\epsilon / X, \mathbf{Z}) = 0$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editors and referees for very constructive comments that greatly improved the presentation of this article. This research is part of Yinghao Pan's Ph.D. dissertation. It is partially supported by grants R01 ES021900 from the National Institute of Environmental Health Sciences and P01 CA142538 from the National Cancer Institute.

References

- Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. Lifetime data analysis. 2000; 6:39–58. [PubMed: 10763560]
- Chen K, Lo SH. Case-cohort and case-control analysis with cox's model. Biometrika. 1999; 86:755–764.
- Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society. 1972; 34:187–220.
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association. 1952; 47:663–685.
- Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. Statistics in medicine. 2006; 25:1323–1339. [PubMed: 16220494]
- Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. Biometrika. 2009; 96:887–901. [PubMed: 23946547]
- Kim S, Cai J, Lu W. More efficient estimators for case-cohort studies. Biometrika. 2013; 100:695–708. [PubMed: 24634519]

- Kim S, Sandler DP, Carswell G, De Roo LA, Parks CG, Cawthon R, Weinberg CR, Taylor JA. Telomere length in peripheral blood and breast cancer risk in a prospective case-cohort analysis: results from the sister study. *Cancer Causes & Control*. 2011; 22:1061–1066. [PubMed: 21643930]
- Kim S, Taylor JA, Milne GL, Sandler DP. Association between urinary prostaglandin e2 metabolite and breast cancer risk: a prospective, case-cohort study of postmenopausal women. *Cancer Prevention Research*. 2013; 6:511–518. [PubMed: 23636050]
- Kulich M, Lin D. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*. 2004; 99:832–844.
- Lee A, McMURCHY L, Scott A. Re-using data from case-control studies. *Statistics in medicine*. 1997; 16:1377–1389. [PubMed: 9232759]
- Li Q, Racine JS. Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*. 2008; 26:423–434.
- Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genetic epidemiology*. 2009; 33:256–265. [PubMed: 19051285]
- Ma Y, Carroll RJ. Semiparametric estimation in the secondary analysis of case-control studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2016; 78:127–151. [PubMed: 26834506]
- Monsees GM, Tamimi RM, Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genetic epidemiology*. 2009; 33:717–728. [PubMed: 19365863]
- Morris PG, Hudis CA, Giri D, Morrow M, Falcone DJ, Zhou XK, Du B, Brogi E, Crawford CB, Kopelovich L, et al. Inflammation and increased aromatase expression occur in the breast tissue of obese women with breast cancer. *Cancer prevention research*. 2011; 4:1021–1029. [PubMed: 21622727]
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Richardson DB, Rzehak P, Klenk J, Weiland SK. Analyses of case-control data for additional outcomes. *Epidemiology*. 2007; 18:441–445. [PubMed: 17473707]
- 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.
- Saarela O, Kulathinal S, Arjas E, Läärä E. Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. *Statistics in medicine*. 2008; 27:5991–6008. [PubMed: 18792086]
- Saarela O, Kulathinal S, Karvanen J. Secondary analysis under cohort sampling designs using conditional likelihood. *Journal of Probability and Statistics*. 2012; 2012
- Salim A, Xiangmei M, Jialiang L, Reilly M. A maximum likelihood method for secondary analysis of nested case-control data. *Statistics in medicine*. 2014; 33:1842–1852. [PubMed: 24753004]
- Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*. 1988:64–81.
- Subbaramaiah K, Howe LR, Bhardwaj P, Du B, Gravaghi C, Yantiss RK, Zhou XK, Blaho VA, Hla T, Yang P, et al. Obesity is associated with inflammation and elevated aromatase expression in the mouse mammary gland. *Cancer prevention research*. 2011; 4:329–346. [PubMed: 21372033]
- Wang X, Zhou H. A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. *Biometrics*. 2006; 62:1149–1160. [PubMed: 17156290]
- Wang X, Zhou H. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics*. 2010; 66:502–511. [PubMed: 19508239]
- Wei J, Carroll RJ, Müller UU, Keilegom IV, Chatterjee N. Robust estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2013; 75:185–206. [PubMed: 23637568]

Table 1

Simulation results with full cohort size $N = 1000$, the SRS sample size is $n_{SRS} = 200$. $X \sim \mathcal{N}(0, 1)$, $Z \sim \text{Bernoulli}(0.45)$.

True β_1	True γ_2	Methods	Estimated quantities for $\hat{\beta}_1$			
			Bias	SD	$\widehat{SD}$	CI
Censoring percentage = 80%						
0	0	$\hat{\beta}_N$	0.001	0.050	0.052	0.950
		$\hat{\beta}_{SRS}$	0.000	0.070	0.071	0.948
		$\hat{\beta}_{IPW}$	0.002	0.063	0.063	0.950
		$\hat{\beta}_{AIPW}$	0.002	0.064	0.063	0.944
		$\hat{\beta}_{NPEL}$	0.002	0.051	0.052	0.945
		$\log(3)$	0.024	0.054	0.054	0.925
	0.5	$\hat{\beta}_N$	0.000	0.071	0.071	0.947
		$\hat{\beta}_{SRS}$	0.001	0.062	0.061	0.947
		$\hat{\beta}_{IPW}$	0.002	0.063	0.062	0.944
		$\hat{\beta}_{AIPW}$	0.001	0.053	0.052	0.945
		$\hat{\beta}_{NPEL}$	-0.003	0.051	0.052	0.959
		$\hat{\beta}_N$	-0.005	0.071	0.071	0.953
		$\hat{\beta}_{SRS}$	-0.005	0.063	0.063	0.939
		$\hat{\beta}_{IPW}$				

True β_1	True γ_2	Methods	Estimated quantities for $\hat{\beta}_i$			
			Bias	SD	$\widehat{SD}$	CI
$\log(3)$	$\hat{\beta}_{AIPW}$	$\hat{\beta}_{AIPW}$	-0.002	0.054	0.055	0.954
		$\hat{\beta}_{NPEL}$	-0.005	0.046	0.048	0.955
		$\hat{\beta}_N$	0.025	0.052	0.051	0.922
		$\hat{\beta}_{SRS}$	-0.001	0.071	0.071	0.945
	$\hat{\beta}_{IPW}$	$\hat{\beta}_{IPW}$	0.000	0.063	0.060	0.935
		$\hat{\beta}_{AIPW}$	0.004	0.056	0.054	0.944
		$\hat{\beta}_{NPEL}$	-0.001	0.047	0.047	0.950
		Censoring percentage = 95%				
	0.5	$\log(3)$	0.084	0.066	0.065	0.729
		$\hat{\beta}_N$	-0.001	0.073	0.071	0.940
		$\hat{\beta}_{SRS}$	0.002	0.069	0.066	0.928
		$\hat{\beta}_{IPW}$	0.005	0.062	0.059	0.933
	$\hat{\beta}_{AIPW}$	$\hat{\beta}_{AIPW}$	0.000	0.058	0.058	0.943
		$\hat{\beta}_{NPEL}$				

NOTE: the models are $Y_2 = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, $\lambda(t|X, Y_2, Z) = \lambda_0(t) \exp(\gamma_1 X + \gamma_2 Y_2 + \gamma_3 Z)$.

Performance of our proposed estimator $\hat{\beta}_{NPEL}$ when we misspecify $\mathbf{W} \cdot \mathbf{Z} = (Z_1, Z_2)$, where $Z_1 \sim \text{Bernoulli}(0.45)$, $Z_2 \sim \text{Bernoulli}(0.6)$. $X \sim N(Z_1, 1)$, which means that Z_1 is the informative components of $\mathbf{Z}$ about X . The true $\mathbf{W}$ is Z_1 .

Table 2

True β_1	W used	Estimated quantities for β_1			
		Bias	SD	$\widehat{\text{SD}}$	CI
0	Z_1	0.001	0.053	0.052	0.952
	Z_2	0.001	0.049	0.049	0.936
	(Z_1, Z_2)	0.001	0.053	0.052	0.954
0.5	Z_1	-0.002	0.046	0.047	0.958
	Z_2	-0.063	0.044	0.045	0.728
	(Z_1, Z_2)	-0.007	0.046	0.048	0.964

NOTE: Models are $Y_2 = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \epsilon$, $\lambda(i|X, Y_2, Z_2) = \lambda_0(i) \exp(\gamma_1 X + \gamma_2 Y + \gamma_3 Z_1 + \gamma_4 Z_2)$.

Table 3Parametric vs Non-parametric estimation of X given Z

Methods	Estimated quantities for β_1			
	Bias	SD	$\widehat{SD}$	CI
X is normal				
$\hat{\beta}_{PL}$	0.005	0.049	0.064	0.926
$\hat{\beta}_{NPEL}$	-0.005	0.050	0.050	0.957
X follows a gamma distribution				
$\hat{\beta}_{PL}$	-0.023	0.055	0.096	0.913
$\hat{\beta}_{NPEL}$	-0.005	0.059	0.060	0.962
X follows an exponential distribution				
$\hat{\beta}_{PL}$	-0.051	0.017	0.015	0.143
$\hat{\beta}_{NPEL}$	-0.001	0.017	0.017	0.956

NOTE: Models are $Y_2 = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + e$. $\lambda(t|X, Y_2, \mathbf{Z}) = \lambda_0(t) \exp(\gamma_1 X + \gamma_2 Y_2 + \gamma_3 Z_1 + \gamma_4 Z_2)$. $\hat{\beta}_{NPEL}$, proposed non-parametric estimated likelihood estimator. $\hat{\beta}_{PL}$, parametric maximum likelihood estimator assuming that X given Z is normal.

Table 4

Simulation results when the error term is misspecified or the survival model is misspecified. $\mathbf{Z} = (Z_1, Z_2)$, where $Z_1 \sim \text{Bernoulli}(0.45)$, $Z_2 \sim \text{Bernoulli}(0.6)$, $X \sim N(Z_1, 1)$, $\beta_1 = 0.5$.

Methods	Estimated quantities for β_1			
	Bias	SD	$\widehat{\text{SD}}$	CI
Misspecification of ε				
(i) gamma error misspecified as $N(0, \sigma^2)$				
$\hat{\beta}_{IPW}$	-0.001	0.063	0.061	0.936
$\hat{\beta}_{AIPW}$	0.007	0.065	0.061	0.916
$\hat{\beta}_{NPEL}$	-0.033	0.051	0.049	0.889
(ii) heteroscedastic error (related to Z_1) misspecified as $N(0, \sigma^2)$				
$\hat{\beta}_{IPW}$	0.001	0.044	0.042	0.937
$\hat{\beta}_{AIPW}$	0.004	0.037	0.036	0.942
$\hat{\beta}_{NPEL}$	-0.004	0.032	0.032	0.956
(iii) heteroscedastic error (related to X) misspecified as $N(0, \sigma^2)$				
$\hat{\beta}_{IPW}$	0.001	0.099	0.096	0.940
$\hat{\beta}_{AIPW}$	0.003	0.096	0.093	0.928
$\hat{\beta}_{NPEL}$	-0.013	0.074	0.051	0.818
Misspecification of survival model				
(iv) AFT misspecified as PH				
$\hat{\beta}_{IPW}$	0.005	0.064	0.063	0.940
$\hat{\beta}_{AIPW}$	0.007	0.057	0.056	0.930
$\hat{\beta}_{NPEL}$	-0.005	0.049	0.049	0.951
(v) Additive Hazards misspecified as PH				
$\hat{\beta}_{IPW}$	0.002	0.067	0.064	0.937
$\hat{\beta}_{AIPW}$	0.008	0.060	0.056	0.912
$\hat{\beta}_{NPEL}$	0.001	0.049	0.049	0.948

NOTE: Models for secondary outcome is $Y_2 = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \varepsilon$. AFT: Accelerated Failure Time. PH: Proportional Hazards.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Analysis of Sister Study.

	Proposed EL		SRS		IPW	
	Estimate	SD	Estimate	SD	Estimate	SD
Linear Model (outcome: BMI)	Int	0.020	0.061	0.158	0.337	0.135
	PGE-M	0.361	0.052	0.194	0.186	0.049
	NSAIDS	0.397	0.049	0.257	0.308	0.100
	DRINKING	-0.266	0.055	-0.434	-0.446	0.120
	AGE	-0.111	0.046	-0.266	-0.213	0.099
Cox Model	PGE-M	0.185	0.103	-0.183	0.190	0.250
	BMI	0.081	0.125	0.269	0.154	0.397
	NSAIDS	0.249	0.175	-0.493	0.580	0.666
	AGE	0.039	0.183	-0.227	-0.193	0.601
	WAIST	0.037	0.243	-0.061	-0.231	0.860
	SMOKE	0.469	0.331	-17.1	0.449	1.031