Published in final edited form as:

J Stat Plan Inference. 2016 November ; 178: 24–36. doi:10.1016/j.jspi.2016.05.001.

Outcome-Dependent Sampling Design and Inference for Cox's Proportional Hazards Model

Jichang Yu[a,b], **Yanyan Liu**[b], **Jianwen Cai**[c], **Dale P. Sandler**[d], and **Haibo Zhou**[c,*]

[a]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China

[b]School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

[c]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[d]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

Abstract

We propose a cost-effective outcome-dependent sampling design for the failure time data and develop an efficient inference procedure for data collected with this design. To account for the biased sampling scheme, we derive estimators from a weighted partial likelihood estimating equation. The proposed estimators for regression parameters are shown to be consistent and asymptotically normally distributed. A criteria that can be used to optimally implement the ODS design in practice is proposed and studied. The small sample performance of the proposed method is evaluated by simulation studies. The proposed design and inference procedure is shown to be statistically more powerful than existing alternative designs with the same sample sizes. We illustrate the proposed method with an existing real data from the Cancer Incidence and Mortality of Uranium Miners Study.

Keywords

Empirical process; optimal allocation; outcome-dependent sampling

1 Introduction

The cost of exposure assessment is often the deciding factor in determining the duration and size of many real studies. When the measurement of main exposure is expensive (such as the assessment of some biomarkers), it could be prohibitive to assess the main exposure on all the subjects. For example, assessing the effect of epidermal growth factor receptor (EGFR)

[*]Corresponding author. zhou@bios.unc.edu (H. Zhou).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

genetic mutations on tumor response (no response, partial response, or complete response) to EGFR-targeted therapy for patients with nonsmall cell lung cancer, investigators have to assay the EGFR mutation patients. However, study sizes are limited as the genetic assay on EGFR mutations is expensive (around $3,000 per person) (Wang and Zhou, 2006). Hence, new and efficient study designs which can reduce the overall cost and/or improve the efficiency under a given budget are desirable in this case.

For a continuous outcome, Zhou et al. (2007) had shown that outcome-dependent sampling (ODS) was a cost-effectiveness design for situation described above in large epidemiologic studies. The principal idea of ODS design is to select subjects who are believed to be more informative about the exposure-response relationship to enhance the study efficiency. A typical ODS design for a continuous outcome will have subject's exposure value assessed on a simple random sample (SRS) and additional supplemental samples selected with probability depending on the outcome variable (e.g., Zhou et al., 2002; Weaver and Zhou, 2005). Analysis based only on the SRS data would not be efficient because it does not utilize the information from the supplemental subjects. On the other hand, standard estimation that ignores the biased sampling structure of the ODS design will yield biased and inconsistent estimators. This general ODS design is rooted in the earlier development of biased sampling designs. For example, in the discrete outcome case, case-control design is the well-known outcome-dependent sampling scheme (e.g., Prentice and Pyke, 1979; Breslow and Cain, 1988; Weinberg and Wacholder, 1993; Breslow and Holubskov, 1997; Wang and Zhou, 2010). Case–cohort design (Prentice, 1986) is based on ODS idea in failure time response. The case–cohort design samples a simple random sample (SRS) from the underlying population and in addition collects all the failures out of SRS (e.g., Self and Prentice, 1988; Cai and Zeng, 2004; Scheike and Martinussen, 2004; Sun, Sun and Flournoy, 2004; Zhang, Schaubel and Kalbfleisch, 2011; Kim, Cai and Lu, 2013).

The case–cohort design can be viewed as a special case of ODS design with the selection probability of supplemental failure equal to 1. The case–cohort design is especially useful when the failure rate is low and the number of failures is small. However, in many large cohort studies, the failure rate may not be low or the number of failures could be large. Under such situations, investigators often are forced to decide how to assemble exposure information for only a subset of the failures instead of all the failures, so it will fit their overall budget. Variations of the case–cohort (Prentice, 1986) sampling scheme have been proposed to improve the efficiency of the design while reducing the overall experiment cost. For example, in the stratified case–cohort design (Borgan et al., 2000), the sample was drawn from the stratum defined by the covariate correlated with exposure. In the generalized case–cohort design (Cai and Zeng, 2007; Kang and Cai, 2009), a subset of failures are randomly sampled as the supplemental samples.

As Zhou et al. (2002) has demonstrated in a continuous outcome case, the certain segments of outcome variable are more informative than others in providing information on evaluating the association between an exposure and outcome. Specifically, for a given sample size, a sample composed subjects from the high and low region of response variables is more informative about the exposure-response relationship than a sample consisted with just a simple random sample. Inspired by the ODS design for continuous outcome, we propose an outcome-dependent sampling design for failure time data with right-censoring under Cox's proportional hazards model. The proposed failure time outcome-dependent sampling design is a retrospective design and the exposure value is only measured for the selected subjects. We use the weighted estimating equation method to estimate the interested regression parameters, which is easy to implement with the freely available R package "survival". To help investigators to design an optimal ODS study, we develop a computation formula for optimal subsamples allocations by evaluating the asymptotic relative efficiency between our proposed method and the simple random sampling design with the same sample size.

The rest of the paper is organized as following. In Section 2, we introduce the proposed failure time ODS design and the appropriate weighted estimating equation is given to estimate the regression parameters. In Section 3, we present the asymptotic properties of the proposed estimator. In Section 4, we establish a criteria and formula for optimal allocation of subsamples. In Section 5, we conduct the simulation studies to evaluate the finite sample performance of the proposed method. In Section 6, we illustrate the proposed method with a data set from the Cancer Incidence and Mortality of Uranium Miners Study. In Section 7, concluding remarks and discussions are given. Finally, the proofs for theoretical results are outlined in the Appendix.

2 Failure Time ODS Design and Proposed Estimator

2.1 Failure Time ODS Design and Data Structure

Assume there are m independent subjects in an underlying cohort. Let $\tilde{T}$ denote the failure time and C be the potential censoring time for $\tilde{T}$. Due to right-censoring, we only observes the vector (T, δ) with $T = \min\left(\tilde{T}, C\right)$ and $\delta = I\left(\tilde{T} \le C\right)$, where $I(\cdot)$ is the indicator function. Let $Z_e(t)$ be a possibly time-dependent one-dimensional exposure which is expensive or difficult to measure and $Z_c(t)$ be a possibly time-dependent $q - 1$-vector of covariates which are cheap or easily to measure, respectively. We assume that $\tilde{T}$ and C are independent conditional on $Z_e(\cdot)$ and $Z_c(\cdot)$. Assume the hazard function of the conditional distribution of failure time $\tilde{T}$ given $Z_e(t)$ and $Z_c(t)$ follows Cox's proportional hazards model (Cox, 1972):

$$\lambda\left(t | Z_e(s), Z_c(s), 0 \le s \le t\right) = \lambda_0(t) \exp\left\{\theta_0 Z_e(t) + \gamma'_0 Z_c(t)\right\}, \quad (2.1)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and $\beta_0 = \left(\theta_0, \gamma'_0\right)'$ is the q-dimensional unknown regression parameter and define $X(t) = \left(Z_e(t), Z'_c(t)\right)'$.

The sampling mechanism of the proposed failure ODS design is constituted by following two-stage sampling and the exposure value is only assessed on these selected subjects. First, a subcohort (SRS) of simple random sample is selected from the underlying cohort and the subjects in SRS is indexed by a binary indictor ξ (1, if belonging to SRS; otherwise, 0). Let n_0 denote the sample size of SRS and assume that n_0/m is convergent to p in probability.

Second, the domain of failure time $\tilde{T}$ are partitioned into K mutually exclusive intervals, $A_k = (a_{k-1}, a_k]$, $k = 1, \ldots, K$, where $\{a_k: k = 0, 1, \ldots, K\}$ are known constants with $a_0 = 0 < a_1 <$,

…, < $a_K = +\infty$. Let $\zeta_{ik}=I\left(\tilde{T}_i \in A_k\right)$ be the indictor of the i-th subject falling into the interval A_k. Let η_{ik} denote the indicator of failure subject i, which is from the stratum A_k and selected into supplemental sample. The size of supplemental sample from the k-th stratum is denoted by n_k. Let m_k and $n_{0,k}$ denote the size of the full cohort failure sample and the SRS failure sample falling into the stratum A_k, respectively. Assume that $n_k/(m_k - n_{0,k})$ is convergent to r_k in probability, for $k = 1, \ldots, K$. For most ODS applications, the $K = 3$ case is shown to be a practical and sufficient setting and usually assume $n_1 = n_3$ (Zhou et al., 2007). Therefore, we consider $K = 3$ and only select supplemental samples from the intervals A_1 and A_3 in this article. The sampling mechanism of ODS can be explained by the following figure:

The samples with exposure value observed are referred to as validation sample, and the set of remaining subjects whose exposure value are not assessed is referred to as nonvalidation sample. Therefore, the observed data for our failure ODS design is:

$$\text{Validation sample:}\begin{cases} \text{SRC:}(T_i,\delta_i,Z_{e,i}(t),Z_{c,i}(t)), i\in V_0,\\ (T_j,\delta_j,Z_{e,j}(t)|T_j\in A_k,\delta_j=1,Z_{c,j}(t)), j\in V_k, k=1,3;\end{cases}$$
$$\text{Nonvalidation sample:}(T_l,\delta_l,Z_{c,l}(t)), l\in\overline{V};$$

where V_0, V_k and $\overline{V}$ are the index for the SRS, supplemental sample from the stratum A_k and the nonvalidation sample, respectively. Note that the proposed failure time ODS design is coincided with classical case–cohort design if we let $K = 1$ and $r_1 = 1$; and coincided with generalized case–cohort design if we let $K = 1$ and $r_1 \in (0, 1)$.

2.2 A Weighted Estimating Equation For Regression Parameter

Define the counting processes $N_i(t) = I(T_i \le t, \delta_i = 1)$ and at risk processes $Y_i(t) = I(T_i \ge t)$, for $i = 1, \ldots, m$. Define $S^{(d)}(\beta,t)=m^{-1}\sum_{j=1}^{m}Y_j(t)X_j^{\otimes d}(t)\exp\{\beta' X_j(t)\}\quad(d=0,1,2)$, with $a^{\otimes 2}=aa'$, $a^{\otimes 1}=a$, $a^{\otimes 0}=1$, for a vector a and recall $X(t)=\left(Z_e(t),Z'_c(t)\right)'$. Let τ denote the study end time. If the covariate history are observable for all the study cohort, the regression parameters β_0 can be estimated by solving the standard partial likelihood score equation $U(\beta) = 0$ (Andersen and Gill 1982, referred as AG in the following), where

$$U(\beta)=\sum_{i=1}^{m}\int_0^\tau\left[X_i(t)-\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\right]dN_i(t). \tag{2.2}$$

Under the proposed failure time ODS design, since the exposure $Z_e(\cdot)$ is only observed for the selected subjects, therefore the estimator of β_0 cannot be calculated directly from (2.2). We propose to use the inverse probability weight (IPW)(e.g., Horvitz and Thompson, 1951) to inference the data from our proposed ODS design. Define $n = n_0+n_1+n_3$, $n/m \to \rho_V$ and $n_k/n \to \rho_k$, $k = 0, 1, 3$, respectively. Define $\pi_k = Pr(T \in A_k, \delta = 1)$, $k = 1, 2, 3$. From simple calculation, we can obtain the relationship between (p, r_k)(Section 2.1) and $(\rho_{V,} \rho_0, \rho_k)$ as following:

$$p{=}\rho_0 \times \rho_V, \quad r_k{=}\frac{\rho_k \times \rho_V}{\pi_k\left(1{-}\rho_0 \times \rho_V\right)}, \text{ for } k{=}1,2,3. \tag{2.3}$$

Due to biased sampling mechanism of proposed failure time ODS design, the inverse probability weight will have the following four characteristics: (i) nonvalidation samples are eliminated by setting $w = 0$; (ii) the sampled censored subjects are weighted by $(\rho_0\rho_V)^{-1}$; (iii) the sampled subcohort cases are weighted by 1, if their failure times belong to A_1 and A_3, and by $(\rho_0\rho_V)^{-1}$ otherwise; (iv) the sampled non-subcohort cases are weighted by $\pi_1(1 - \rho_0\rho_V)/(\rho_1\rho_V)$ and $\pi_3(1 - \rho_0\rho_V)/(\rho_3\rho_V)$, if their failure times belong to A_1 or A_3, respectively. Hence, the weight can be written as following formula:

$$w_i{=}\xi_i\delta_i\zeta_i{+}\frac{\xi_i\left(1-\delta_i\right)}{\rho_0\rho_V}{+}\frac{\xi_i\delta_i\left(1-\zeta_i\right)}{\rho_0\rho_V}{+}\left(1-\xi_i\right)\delta_i\sum_{k=\{1,3\}}\frac{\pi_k\left(1-\rho_0\rho_V\right)\zeta_{ik}\eta_{ik}}{\rho_k\rho_V}, \tag{2.4}$$

with $\zeta_i = \zeta_{i1} + \zeta_{i3}$.

The true regression coefficients, β_0, then can be estimated by $\hat{\beta}_{ODS}$ from solving the following weighted estimating equation: $U_W(\beta) = 0$, where

$$U_W\left(\beta\right){=}\sum_{i=1}^{m} w_i \int_0^\tau \left[X_i\left(t\right)-\frac{S_W^{(1)}\left(\beta,t\right)}{S_W^{(0)}\left(\beta,t\right)}\right] dN_i\left(t\right) \tag{2.5}$$

with $S_W^{(d)}\left(\beta,t\right){=}m^{-1}\sum_{j=1}^{m} w_j Y_j\left(t\right)\exp\{\beta' X_j\left(t\right)\} X_j^{\otimes d}\left(t\right), \quad (d{=}0,1,2)$.

The cumulative baseline hazard $\Lambda_0\left(t\right){=}\int_0^t \lambda_0\left(s\right)ds$ can be naturally estimated by:

$$\hat{\Lambda}_W\left(t\right){=}\int_0^t \frac{\sum_{i=1}^{m} w_i dN_i\left(s\right)}{m S_W^{(0)}\left(\hat{\beta}_{ODS},s\right)}.$$

3 Asymptotic Properties

To establish the asymptotic properties of $\hat{\beta}_{ODS}$ we first show that $m^{-1/2}U_W(\beta)$ can be approximated by two uncorrelated sums of independent random variables. Since the weights are not predictable, we employ empirical process theory for asymptotic properties, which do not require predictability. Define $s^{(d)}\left(\beta,t\right){=}E[Y_j\left(t\right)\exp\{\beta' X_j\left(t\right)\}X_j^{\otimes d}\left(t\right)], \ d{=}0,1,2$ and $e(\beta, t) = s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$. Straightforward calculation can show $E[S_W^{(d)}\left(\beta,t\right)]{=}s^{(d)}\left(\beta,t\right)$.

In order to estimate the asymptotic properties of proposed estimators, we impose the following regularity conditions:

(C1) $\int_0^\tau \lambda_0(t)dt < \infty$

(C2) $P_i(Y_i(t) = 1) > 0$ for $t \in (0, \tau]$.

(C3) $X_i(\cdot)$ ($i = 1, \ldots, m$) have bounded total variations, i.e. $|X_{ij}(0)| + \int_0^\tau |dX_{ij}(t)| \le Con_M$ for all $j = 1, \ldots, q$ and $i = 1, \ldots, m$, where X_{ij} is the j-th component of X_i and Con_M is a constant.

(C4) There exists a neighborhood B of β_0 such that $\sup_{t\in[0,\tau],\beta\in B} \|S_W^{(d)}(\beta,t) - s^{(d)}(\beta,t)\| \to_{a.s.} 0$ for d=0,1,2, where $s^{(d)}(\beta,t) = E[S_W^{(d)}(\beta,t)]$ is absolutely continuous, for $\beta \in B$, uniformly in $u \in (0, \tau]$. Moreover, $s^{(0)}(\beta, t)$ is assumed to be bounded away from zero for each $(\beta, t) \in B \times (0, \tau]$.

(C5) The matrix $\sum_A(\beta_0) = \int_0^\tau \left[\frac{s^{(2)}(\beta_0,t)}{s^{(0)}(\beta_0,t)} - e(\beta_0,t)^{\otimes 2}\right] s^{(0)}(\beta_0,t)\lambda_0(t)\,dt$ is positive definite.

Conditions (C1), (C2), (C4) and (C5) are analogous to those of Anderson and Gill (1982). Condition (C3) simplifies the derivation of the asymptotic results, but is not a practical limitation. Define

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\,exp(\beta'_0 X_i(s))\lambda_0(s)\,ds,$$

which is a martingale. Under some regularity conditions (see the Appendix), the asymptotic properties can be developed and summarized in following theorem.

Theorem 3.1

Under the conditions (C1)–(C5), (i)(consistency) $\hat{\beta}_{ODS} \to_p \beta_0$; (ii)(asymptotic normality) $m^{1/2}(\hat{\beta}_{ODS} - \beta_0)$ is asymptotically normally distributed with mean zero and variance matrix $\sum_{ODS}(\beta_0) = \sum_A^{-1}(\beta_0)\left(\sum_A(\beta_0) + \sum_B(\beta_0)\right)\left(\sum_A^{-1}(\beta_0)\right)'$, where $\Sigma_A(\beta_0)$ is defined as in assumption (C5) and

$$\begin{aligned}\sum_B(\beta_0) &= E[(w_1-1)^2 H_1^{\otimes 2}(\beta_0)]\\ &= \tfrac{1-\rho_0\rho_V}{\rho_0\rho_V}E[(1-\delta_1)H_1^{\otimes 2}(\beta_0)] + \tfrac{1-\rho_0\rho_V}{\rho_0\rho_V}E[\delta_1(1-\zeta_1)H_1^{\otimes 2}(\beta_0)]\\ &\quad + \sum_{k\in\{1,3\}} \frac{(1-\rho_0\rho_V)(\pi_k(1-\rho_0\rho_V)-\rho_k\rho_V)}{\rho_k\rho_V} E[\delta_1\zeta_{1k}H_1^{\otimes 2}(\beta_0)],\end{aligned}$$

where $H_1(\beta) = \int_0^\tau [Z_1(t) - e(\beta,t)]dM_1(t)$.

We note that, due to biased sampling, the asymptotic variance of $\hat{\beta}_{ODS}$ compared with full-cohort standard partial likelihood estimator has an extra variance term $\Sigma_B(\beta_0)$. The covariance matrix could be consistently estimated by replacing the means with their empirical counterparts.

Theorem 3.2

Under the assumptions (C1)–(C5), $\sup_{t\in[0,\tau]}|\hat{\Lambda}_W(t)-\Lambda_0(t)|\to_p 0$ and

$$\sqrt{m}\left(\hat{\Lambda}_W(t)-\Lambda_0(t)\right)+\sqrt{m}\left(\hat{\beta}_{ODS}-\beta_0\right)\int_0^t e(\beta_0,u)\,\lambda_0(u)\,du$$

converges to a Gaussian process with variance function

$$\sum\nolimits_{\Lambda}(t)=\frac{1}{\rho_0\rho_V}\int_0^t\frac{\lambda_0(u)}{s^{(0)}(\beta_0,u)}du+E\left[\delta_1\left(\frac{(\rho_0\rho_V)^2-1}{\rho_0\rho_V}\zeta_1+(1-\rho_0\rho_V)\sum_{k\in\{1,3\}}\frac{\pi_k(1-\rho_0\rho_V)\zeta_{1k}}{\rho_k\rho_V}\right)\times\left(\int_0^t\frac{dM_1(u)}{s^{(0)}(\beta_0,u)}\right)^2\right].$$

Similarly, $\Sigma_\Lambda(t)$ can be consistently estimated by

$$\frac{1}{\rho_0\rho_V}\int_0^t\frac{\hat{\lambda}_0(u)}{s_W^{(0)}\left(\hat{\beta}_{ODS},u\right)}du+\frac{1}{m}\sum_{i=1}^m\left[\delta_i\left(\frac{(\rho_0\rho_V)^2-1}{\rho_0\rho_V}\zeta_i+(1-\rho_0\rho_V)\sum_{k\in\{1,3\}}\frac{\pi_k(1-\rho_0\rho_V)\zeta_{1k}}{\rho_k\rho_V}\right)\left(\int_0^t\frac{w_i d\hat{M}_i(u)}{s_W^{(0)}\left(\hat{\beta}_{ODS},u\right)}\right)^2\right].$$

The proofs of Theorem 3.1 and 3.2 are provided in the Appendix.

4 The Optimal Failure Time ODS Design Under a Fixed Budget

The validation samples of ODS design are constituted by SRS and supplemental sample. Under the fixed budget, there are many options of n_0, n_1, n_3 satisfying the condition $n = n_0 + n_1 + n_3$ with n fixed. How to choose the allocation of (n_0, n_1, n_3) to improve efficiency is an important problem. The asymptotic relative efficiency between the standard partial likelihood estimator $(\hat{\beta}_{SRS})$ based on the same sample size as ODS design and the proposed estimator $(\hat{\beta}_{ODS})$ is

$$ARE\left(\hat{\beta}_{SRS},\hat{\beta}_{ODS}\right)=\rho v\, MI_q+\rho v\sum\nolimits_A^{-1}(\beta_0)\sum\nolimits_B(\beta_0), \qquad (4.1)$$

where MI_q is an identity matrix of size $q \times q$. The optimal failure time ODS design means the optimal allocation of n_0, n_1, n_3, which minimizes $ARE\left(\hat{\beta}_{SRS},\hat{\beta}_{ODS}\right)_{[1,1]}$ with n fixed, where $F_{[i,j]}$ denotes the (i, j) element of matrix F, and $ARE\left(\hat{\beta}_{SRS},\hat{\beta}_{ODS}\right)_{[1,1]}$ denotes the asymptotic relative efficiency of the exposure $Z_e(\cdot)$. We use the notation $ARE(\hat{\theta}_{SRS},\hat{\theta}_{ODS})$ for $ARE\left(\hat{\beta}_{SRS},\hat{\beta}_{ODS}\right)_{[1,1]}$. The formula of the $ARE(\hat{\theta}_{SRS},\hat{\theta}_{ODS})$ can be re-written as:

$$
\begin{aligned}
ARE(\hat{\theta}_{SRS},\hat{\theta}_{ODS})=&\rho_V+\tfrac{1-\rho_0\rho_V}{\rho_0}\left(\sum\nolimits_A^{-1}(\beta_0)\,E\left[(1-\delta_1)\,H_1^{\otimes 2}(\beta_0)\right]\right)_{[1,1]}\\
&+\tfrac{1-\rho_0\rho_V}{\rho_0}\left(\sum\nolimits_A^{-1}(\beta_0)\,E\left[\delta_1(1-\zeta_1)\,H_1^{\otimes 2}(\beta_0)\right]\right)_{[1,1]}\\
&+\sum_{k\in\{1,3\}}\frac{(1-\rho_0\rho_V)(\pi_k(1-\rho_0\rho_V)-\rho_k\rho_V)}{\rho_k}\\
&\times\left(\sum\nolimits_A^{-1}(\beta_0)\,E\left[\delta_1\zeta_{1k}H_1^{\otimes 2}(\beta_0)\right]\right)_{[1,1]}.
\end{aligned}
\quad (4.2)
$$

In practice, we usually assume $\rho_1 = \rho_3$. Therefore, $ARE(\hat{\theta}_{SRS},\hat{\theta}_{ODS})$ is a function of ρ_0 and the nonlinear program methods can be used to obtain the optimal failure time ODS design. In practice, we firstly select a subcohort (SRS) by simple random sampling. Then, we can obtain the estimator of $ARE(\hat{\theta}_{SRS},\hat{\theta}_{ODS})$ by estimating $\sum\nolimits_A^{-1}(\beta_0)$, $E[(1-\delta_1)H_1^{\otimes 2}(\beta_0)]$, $E[\delta_1(1-\zeta_1)\,H_1^{\otimes 2}(\beta_0)]$, $E[\delta_1\zeta_{1k}H_1^{\otimes 2}(\beta_0)]$ for $k = 1, 3$ by the samples from SRS, and $\beta_0=(\theta_0,\gamma'_0)'$ can be estimated by $\hat{\beta}_{SRS}=(\hat{\theta}_{SRS},\widehat{\gamma'}_{SRS})'$, respectively. Finally, the optimal ODS design can be obtained by the nonlinear program methods.

5 Simulation Studies

The simulation studies are conducted to evaluate the small sample performance of the proposed statistical method for failure time ODS design. First, we conduct the simulation I, where the underlying cohort has $m = 600$ independent subjects, whose failure times are generated by Cox's proportional hazards model:

$$
\lambda(t|Z_e,Z_c)=\lambda_0(t)\exp(\beta_1 Z_e+\beta_2 Z_c), \quad (5.1)
$$

with $\lambda_0(t) = 1$, the exposure $Z_e \sim N(0, 1)$, and covariate $Z_c \sim Bern(1, 0.5)$. We set $\beta_1 = 0.5$, $\beta_2 = 0$ and generate the corresponding censoring times from the mixture of uniform distribution over $[0, c_1]$ and uniform distribution over $[c_2, c_3]$ and the mixing probability is chosen to generate around 80%, 70% censoring respectively. The cutpoints (a_1, a_2) are set to be (30%, 70%) quartiles of the failure times. A subcohort with size $n_0 = 300$ is randomly sampled from the underlying cohort.

We compare the estimator from ODS design $\left(\hat{\beta}_{ODS}\right)$ with the estimators from generalized case–cohort design and SRS design with different sample size. The estimator, $\hat{\beta}_{GCC}$, is based on generalized case–cohort design, which randomly selects the SRS samples of size n_0 and the supplemental failures of size $n_1 + n_3$ out of SRS sample. The standard partial likelihood estimators, $\hat{\beta}_{Full}$, $\hat{\beta}_{R}$, $\hat{\beta}_{SRS}$, are based on the underlying cohort, SRS sample and SRS sample with the same sample size as the ODS design, respectively.

The results of simulation I are presented in Table 1. For each specified scenario, we generated 1000 simulated data sets. The Mean column gives estimator of the regression parameter. The SE column represents the sample standard deviation of the 1000 estimates. The $\widehat{\mathrm{SE}}$ column gives the average of the estimated standard error, which is calculated by the closed formula in Section 3 and the corresponding covariance matrix could be consistently estimated by replacing the means with their empirical counterparts. The column "CI" is the nominal 95% confidence interval coverage of the true parameter using the estimated standard error.

From the simulation results, we know the five estimators are all unbiased under all situations considered here. The proposed variance estimator provides a good estimation for the sample standard errors and the confidence intervals attain coverage close to nominal 95% level. When the censoring rate is increasing, the efficiency is decreasing. When the censoring rate is fixed, the efficiency is increasing with supplemental sampling sizes increasing. The proposed estimator $\hat{\beta}_{ODS}$ is more efficient than $\hat{\beta}_{GCC}$ and $\hat{\beta}_{SRS}$, which indicates sampling the supplemental samples from the tails of the failure time is more efficient than the supplemental samples selected by random sampling and all the subjects randomly sampled, respectively. Therefore, our proposed design is an effective way to enhance study efficiency.

Second, we conduct the simulation II to evaluate the performances of the proposed method under the different cutpoints where the failure time is generated from the Cox model (5.1) with $\beta_1 = 1$, $\beta_2 = -1$ and the censoring rate is set to be 70% and the cutpoints (a1,a2) are set to be (30%, 70%) and (20%, 80%) quartiles of the failure times, respectively. We set $m = 800$, $n_0 = 400$ and generate 1000 simulated data sets for each specified scenario. The results of simulation II are presented in Table 2.

The results from the Table 2 are almost the same as in Table 1. For example, the five estimators are all unbiased, the proposed variance estimator provides a good estimation for the sample standard errors and the confidence intervals attain coverage close to nominal 95% level. The result also confirm that the proposed estimator $\hat{\beta}_{ODS}$ is more efficient than $\hat{\beta}_{GCC}$ and $\hat{\beta}_{SRS}$.

Finally, We conduct the simulation III to evaluate the performance of the proposed optimal allocation method in Section 4. We consider the model which is the same as in the simulation II. The number of underlying cohort is $m = 2000$. There are 315 and 100 failures in the interval A_1 and A_3 under the cutpoints being (30%, 70%) quartiles of the failure times and there are 265 and 50 failures in the interval A_1 and A_3 under the cutpoints being (20%, 80%) quartiles of the failure times with the censoring rate being 70%. The fraction of the validation sample ρ_V is set to be 0.15. We select the same size of supplemental failures from the intervals A_1 and A_3. Simulation results based on 1, 000 data sets are presented in Figure 2.

From the results in Figure 2, we can obtain the optimal ρ_0 is 0.6 (relative efficiency 0.699) under the censoring rate being (30%, 70%) quartiles of the failure times. When the censoring rate is (20%, 80%) quartiles of the failure times, sampling less simple random sample will enhance the study's efficiency. The same number of supplemental samples from the intervals (20%, 80%) quartiles of the failure times will gain more efficiency than the intervals (30%, 70%) quartiles of the failure times.

6 The Cancer Incidence and Mortality of Uranium Miners Study

Lung cancer has been long recognized as an occupational disease in uranium miners and the miners therefore entitled for compensation 1926 in Germany and in 1932 in Czechoslovakia (BEIR VI, 1999; Sandler et al., 1998; Witschi, 2001). The Cancer Incidence and Mortality of Uranium Miners Study was conducted during January 1, 1977, to December 31, 1996 to evaluate the risk of developing radiation-related cancer among uranium miners. Uranium miners are chronically exposed to the alpha particles emitted by radon and its progeny (referred to as radon), which can cause random damage to the chromosomes and DNA molecules contained in the nucleus of the cell, and have a carcinogenic effect.

In the literature, association of radon and mortality had been studied by many authors, e.g., Tirmarche et al. (1993), Vacquier et al. (2008), Kreuzer et al. (2008, 2010) investigated mortality and radon, while others, such as, Řeřicha et al. (2006) and Kulich et al. (2011) pointed out that above studies would miss a substantial number of cases when the cancers have low fatality rates. However, they only considered a radon exposure in Cox’s model. Hence, we investigate incidence of non-lung solid cancers to test associations of radon exposures with cancers adjusting for age, smoking and airborne dust. So, in this article, we employ a failure time ODS design on the cancer incidence of Uranium Miner Study data set to investigate incidence of non-lung solid cancers to test associations of radon exposures with cancers adjusting for age, smoking and airborne dust.

The underlying cohort during follow-up includes 16, 434 miners and a total of 2, 330 subjects with incident cancers identified, of which 1, 444 had cancer types of interest (Sandler et al., 1998). We sampled the subcohort sample from each strata defined according to age on 1/1/1977 (5-year groups) so that the number in each stratum was approximately equal to the total number of all cancer cases in that stratum. There are total 12 stratums in the cohort. The sample size of SRS is n_0 = 1, 825 and the number of SRS from each stratum is (33, 55, 40, 134, 206, 462, 364, 222, 198, 78, 25, 8). The censoring rate is 91.2%. The cutpoints are a_2 and a_8, which are the 20% and 80% quantiles of the incidence time. We then randomly sample n_1 = 51 and n_3 = 51 supplemental samples from the intervals (0, a_2] and (a_8, ∞), respectively. So, the total size of ODS sample is 1, 927.

The Cox proportional hazards model is considered to illustrate the proposed method:

$$\lambda\left(t|Z\right)=\lambda_0\left(t\right)\exp\left\{\beta_1\text{Trad}+\beta_2\text{Age}+\beta_3\text{Tdust}+\beta_4\text{Smoking}\right\},$$

where Trad (total radon exposure) is measured as working level months (WLM, 1WLM = $3.5 \times 10^{-3}\text{Jhm}^{-3}$), Age is measured by year, Tdust (mg/m^3) represents total airborne dust and Smoking is defined as 0–1 variable (0 denotes non-smokers and light smokers who smoked less than 10 cigarettes a day for a period not exceeding 5 years; 1 denotes moderate and heavy smokers).

We consider three methods to evaluate the association between incident and Trad, such as $\hat{\beta}_{SRS}$, $\hat{\beta}_{GCC}$ and $\hat{\beta}_{ODS}$ based on the same sample size. We use the bootstrap method to obtain the variance estimation with the number of bootstrap being 300. The results are summarized in Table 3.

The three methods all confirm the Trad is not significantly related to the incidence of non-lung solid cancers. A more precise 95% confidence interval of Trad is (-0.570×10^{-3}, 1.075×10^{-3}) and it is achieved by $\hat{\beta}_{ODS}$. The standard deviations for Trad are 0.497×10^{-3}, 0.484×10^{-3} and 0.421×10^{-3} by $\hat{\beta}_{SRS}$, $\hat{\beta}_{GCC}$ and $\hat{\beta}_{ODS}$, respectively. The results also show Trad has a positive impact on the incidence of non-lung solid cancers.

7 Concluding Remarks and Discussions

We propose a weighted estimating equation approach for failure time ODS design with right censoring. Under the Cox proportional hazards model, we adopt the inverse probability weight (IPW) method to the standard partial likelihood score equation to estimate the regression parameters due to the biased sampling mechanism. The proposed estimators are shown to be consistent and asymptotically normality. One main advantage of the proposed estimator is that it is very easy to compute by existing R free package “survival”.

In this article, we consider the situation where the main exposure is a scaler variable. The proposed theory works when $Z_e(t)$ in (2.1) is a vector as well. However, in this case, we need to change the ARE formula in (4.2) to the trace of the corresponding matrix. Besides, there is only one exposure (radon) in the real data set.

To facilitate the practical use of the proposed design and method, we developed formula for the optimal study size allocation. The optimal allocation of subsamples is derived by evaluating the relative efficiency between our proposed estimator and the standard partial likelihood estimator from SRS design with the same sample size. This is especially useful tool in aiding investigators to design a cost-effective study. The simulation study suggests that our proposed methods can gain greater efficiency than other frequently used methods. We illustrate our proposed method by the data set from the Cancer Incidence and Mortality of Uranium Miners Study.

We use a simple random sampling as our subcohort. In the non ODS literature (Borgan et al., 2000; Samuelsen et al., 2007), it has been well established that stratified SRS sample is more efficient than the SRS alone. Exploring stratified failure time ODS design and inference could be interesting future work. This would be particularly useful in the case of auxiliary or surrogate covariate problems.

Acknowledgments

The authors are grateful for the valuable comments and suggestions from the associate editor and the referees which drastically improved the article. This work is supported by the National Science Foundation of China grant 11501578 (for Yu), 11571263 and 11371299 (for Liu), and NIH R01 ES021900, P01 CA142538 (for Zhou).

Appendix

We first introduce the following Lemma which will be useful in proving the weak convergence of processes and can be found in Lin et. al (2000).

Lemma 7.1

Let f_n and g_n be two sequences of bounded functions such that, for some constant τ,

- **a.** $\sup_{0 \le t \le \tau} |f_n(t) - f(t)| \to 0$, where f is continuous on $[0, \tau]$,
- **b.** $\{g_n\}$ are monotone on $[0, \tau]$ and
- **c.** $\sup_{0 \le t \le \tau} |g_n(t) - g(t)| \to 0$ for some bounded function g. Then

$$\sup_{0 \le t \le \tau} |\int_0^t f_n(s) dg_n(s) - \int_0^t f(s) dg(s)| \ \to 0,$$

$$\sup_{0 \le t \le \tau} |\int_0^t g_n(s) df_n(s) - \int_0^t g(s) df(s)| \ \to 0.$$

A. Proof of Theorem 1

First, we prove the consistency of $\hat{\beta}_{ODS}$. We use the theorem of Foutz (1977) to get the consistency of $\hat{\beta}_{ODS}$. $\hat{\beta}_{ODS}$ is consistent for β_0 provided:

- **I.** $m^{-1} \partial U_W(\beta) / \partial \beta$ exists and is continuous in an open neighborhood B of β_0;
- **II.** $m^{-1} \partial U_W(\beta_0) / \partial \beta_0$ is negative definite with probability going to 1;
- **III.** $m^{-1} \partial U_W(\beta) / \partial \beta$ converges in probability to a fixed function, say, $\Sigma(\beta)$, uniformly in an open neighborhood of β_0;
- **IV.** $m^{-1} U_W(\beta_0) \to 0$ in probability.

Specifically, one can write

$$\begin{aligned} -m^{-1} \frac{\partial U_W(\beta)}{\partial \beta} &= m^{-1} \sum_{i=1}^{m} \int_0^\tau \left[\frac{S_W^{(2)}(\beta,t)}{S_W^{(0)}(\beta,t)} - \left(\frac{S_W^{(1)}(\beta,t)}{S_W^{(0)}(\beta,t)} \right)^{\otimes 2} \right] w_i dN_i(t) \\ &= m^{-1} \sum_{i=1}^{m} \int_0^\tau \left[\frac{S_W^{(2)}(\beta,t)}{S_W^{(0)}(\beta,t)} - \left(\frac{S_W^{(1)}(\beta,t)}{S_W^{(0)}(\beta,t)} \right)^{\otimes 2} \right] w_i dM_i(t) \\ &+ m^{-1} \sum_{i=1}^{m} \int_0^\tau \left[\frac{S_W^{(2)}(\beta,t)}{S_W^{(0)}(\beta,t)} - \left(\frac{S_W^{(1)}(\beta,t)}{S_W^{(0)}(\beta,t)} \right)^{\otimes 2} \right] w_i Y_i(t) \lambda_0(t) \exp\{\beta' X_i(t)\} dt \end{aligned} \quad (7.1)$$

Because of $m^{-1} \sum_{i=1}^{m} w_i dM_i(t) = E(wdM(t)) = 0$ and $\sum_{i=1}^{m} w_i dM_i(t)$ is the difference of two nondecreasing processes, the first term of (7.1) converges to zero in probability by Lemma 7.1 and Conditions (C1)–(C4) (vander Vaart and Wellner, 1996). By Lemma 1 and conditions (C1) to (C5), we can prove that the second term of (7.1) converges in probability, uniformly for $\beta \in B$ to $\Sigma_A(\beta)$. Therefore (I) to (III) are satisfied by conditions (C1) to (C5) and Lemma 1.

For (IV), it can be show that:

$$m^{-1}U_W(\beta_0)=m^{-1}\sum_{i=1}^{m}\int_0^\tau\left[Z_i(t)-\frac{S_W^{(1)}(\beta_0,t)}{S_W^{(0)}(\beta_0,t)}\right]w_i dM_i(t),$$

which can be proved to converge in probability to zero like the first part of (7.1).

Second, we prove the asymptotic normality of $\hat{\beta}_{ODS}$. By Taylor expansion of $U_W(\beta)$ around β_0, we have

$$U_W(\beta)-U_W(\beta_0)=\frac{\partial U_W(\beta)}{\partial\beta^T}|_{\beta^*}(\beta-\beta_0)$$

Inserting $\hat{\beta}_{ODS}$ in above equation, we have

$$m^{-\frac{1}{2}}U_W(\beta_0)=\{-m^{-1}\frac{\partial U_W(\beta^*)}{\partial\beta^T}\}m^{\frac{1}{2}}(\hat{\beta}_{ODS}-\beta_0),$$

where β^* is between $\hat{\beta}_{ODS}$ and β_0. To prove the asymptotic normality of $\hat{\beta}_{ODS}$, it suffices to prove that $m^{-\frac{1}{2}}U_W(\beta_0)$ converges to a normal random variable in distribution and that $m^{-1}\frac{\partial U_W(\beta^*)}{\partial\beta^T}$ converges to an invertible matrix.

Since

$$\frac{1}{\sqrt{m}}U_W(\beta_0)=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\int_0^\tau\left[X_i(t)-\frac{S_W^{(1)}(\beta_0,t)}{S_W^{(0)}(\beta_0,t)}\right]w_i dM_i(t).$$

On the other hand, by the boundness of w and Lemma 7.1, we have

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\int_0^\tau\left[\frac{S_W^{(1)}(\beta_0,t)}{S_W^{(0)}(\beta_0,t)}-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right]w_i dM_i(t)$$

converges to zero in probability. Therefore, $m^{-\frac{1}{2}}U_W(\beta_0)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\int_0^\tau\left[X_i(t)-\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right]w_i dM_i(t)=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}w_i H_i(\beta_0),$$

Where

$$H_i(\beta)=\int_0^{\tau}\left[X_i(t)-\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right]dM_i(t).$$

So,

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}w_iH_i(\beta_0)=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}H_i(\beta_0)+\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(w_i-1)H_i(\beta_0). \tag{7.2}$$

The first part of (7.2) is the same as partial likelihood in Cox (1972), the second part due to the ODS design. The expectation of $H_i(\beta_0)$ is zero and $E[(w_i-1)H_i(\beta_0)] = E[H_i(\beta_0)E[(w_i-1)/\zeta_{ik}, T_i, \delta_i, X_i(t), 0 \le t \le \tau, 1 \le k \le K]] = 0$. We can get $E[(w_i-1)H_i^{\otimes 2}(\beta_0)]=0$ by the same way.

The variance of $(w_i-1)H_i(\beta_0)$ is finite because w is bounded and the variance of $H_i(\beta_0)$ exists. Therefore $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(w_i-1)H_i(\beta_0)$ converges to a mean zero Gaussian distribution with covariance equivalent to $E[(w_1-1)^2H_1^{\otimes 2}(\beta_0)]$.

Simple calculation show that

$$(w_i-1)^2=(1-\delta_i)\left(\frac{\xi_i}{\rho_0\rho_V}-1\right)^2+\frac{(1-\rho_0\rho_V)^2}{(\rho_0\rho_V)^2}(1-\zeta_i)\xi_i\delta_i+(1-\xi_i)\delta_i$$

$$\sum_{k\in\{1,3\}}\frac{\pi_k^2(1-\rho_0\rho_V)^2-\pi_k\rho_k\rho_V(1-\rho_0\rho_V)}{(\rho_k\rho_V)^2}\eta_{ik}\zeta_{ik}.$$

Therefore,

$$\begin{aligned}E[(w_1-1)^2H_1^{\otimes 2}(\beta_0)]=&\frac{1-\rho_0\rho_V}{\rho_0\rho_V}E[(1-\delta_1)H_1^{\otimes 2}(\beta_0)]+\frac{1-\rho_0\rho_V}{\rho_0\rho_V}E[\delta_1(1-\zeta_1)H_1^{\otimes 2}(\beta_0)]\\&+\sum_{k\in\{1,3\}}\frac{(1-\rho_0\rho_V)(\pi_k(1-\rho_0\rho_V)-\rho_k\rho_V)}{\rho_k\rho_V}E[\delta_1\zeta_{1k}H_1^{\otimes 2}(\beta_0)].\end{aligned}$$

B. Proof of Theorem 3

First, we prove the consistency of $\hat{\Lambda}_W(t)$. Define $d\overline{N}_W(u)=\sum_{i=1}^{m}w_iN_i(u)$. By the definition of $\hat{\Lambda}_W(t)$, we can show that

$$\begin{aligned}\left|\hat{\Lambda}_W(t)-\Lambda_0(t)\right| \le & \left|\int_0^t \frac{d\overline{N}_W(u)}{\sum_{j=1}^m w_j Y_j(u) e^{\widehat{\beta'}_{ODS} Z_j(u)}}\right. \\ & -\int_0^t \frac{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)} d\Lambda_0(u)}{\sum_{j=1}^m w_j Y_j(u) e^{\widehat{\beta'}_{ODS} X_j(u)}} \\ & +\left|\int_0^t \frac{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)} d\Lambda_0(u)}{\sum_{j=1}^m w_j Y_j(u) e^{\widehat{\beta'}_{ODS} X_j(u)}} - \int_0^t \frac{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)} d\Lambda_0(u)}{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)}}\right|. \end{aligned} \tag{7.3}$$

The first part of (7.3) is equal to

$$\int_0^t \frac{\sum_{j=1}^m d w_j M_j(u)}{\sum_{j=1}^m w_j Y_j(u) e^{\widehat{\beta'}_{ODS} X_j(u)}},$$

and the second part of (7.3) equals

$$\int_0^t \left[\left(\frac{1}{\sum_{j=1}^m w_j Y_j(u) e^{\widehat{\beta'}_{ODS} X_j(u)}} - \frac{1}{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)}}\right) \sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)}\right] d\Lambda_0(u).$$

We can show them uniform converge to zero by the Condition (C1) to (C5) and Lemma 1, so is the conclusion.

Secondly, we prove the asymptotic normality of $\hat{\Lambda}_W(t)$. We can show that

$$\begin{aligned} & m^{1/2}(\hat{\Lambda}_W(t)-\Lambda_0(t)) \\ = & m^{1/2}\int_0^t \left[\frac{1}{\sum_{j=1}^m w_j Y_j(u) e^{\widehat{\beta'}_{ODS} X_j(u)}} - \frac{1}{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)}}\right] d\overline{N}_W(u) \\ & + m^{1/2}\left[\int_0^t \frac{d\overline{N}_W(u)}{\sum_{j=1}^m w_j Y_j(u) e^{\beta'_0 X_j(u)}} - \Lambda_0^*(t)\right] \\ & + m^{1/2}[\Lambda_0^*(t)-\Lambda_0(t)], \end{aligned}$$

where $\Lambda_0^*(t)=\int_0^t \lambda_0(u) I(\sum_{i=1}^m Y_i(u)>0) du$ is defined as in Anderson and Gill (1982).

By the argument of Anderson and Gill (1982), $\Lambda_0^*(t)=\Lambda_0(t)$, *a.s.* for all $t \in [0, \tau]$. Therefore the third term of above equality is negligible.

The second term equals $m^{1/2}\int_0^t \frac{m^{-1}\sum_{i=1}^m w_i dM_i(u)}{S_W^{(0)}(\beta_0,u)}$, which asymptotically equals

$$W_m(t)=\int_0^t \frac{m^{-\frac{1}{2}}\sum_{i=1}^m w_i dM_i(u)}{s^{(0)}(\beta_0,u)} \qquad (7.4)$$

by Lemma 1. Since $m^{-\frac{1}{2}}\sum_{i=1}^m w_i dM_i(u)$ is the difference of two nondecreasing bounded processes, it converges to a Gaussian process. The linear functions of the Gaussian processes are Gaussian implies that $W_m(t)$ converges to a Gaussian process with mean zero and variance function

$$\sum_\Lambda(t)=\frac{1}{\rho_0\rho_V}\int_0^t \frac{\lambda_0(u)}{s^{(0)}(\beta_0,u)}du+E\left[\delta_1\left(\frac{(\rho_0\rho_V)^2-1}{\rho_0\rho_V}\zeta_1+(1-\rho_0\rho_V)\sum_{k\in\{1,3\}}\frac{\pi_k(1-\rho_0\rho_V)\zeta_{1k}}{\rho_k\rho_V}\right)\left(\int_0^t\frac{dM_1(u)}{s^{(0)}(\beta_0,u)}\right)^2\right].$$

A Taylor expansion of the first term yields the quantity $H(\beta^*,t)m^{1/2}(\hat{\beta}_{ODS}-\beta_0)$, where

$$H(\beta,t)=-\int_0^t \frac{\sum_{i=1}^m Y_i(u)X_i(u)e^{\beta' X_i(u)}}{\left[\sum_{i=1}^m Y_i(u)e^{\beta' X_i(u)}\right]^2} d\overline{N}_W(u)$$

and β^* is on the line segment between $\hat{\beta}_{ODS}$ and β_0. Similar arguments as the proof of consistency of $\hat{\Lambda}_W(t)$ can show that

$$\sup_{t\in[0,\tau]} \|H(\beta^*,t)+\int_0^t e(\beta_0,u)\lambda_0(u)du\|\to_p 0.$$

Therefore Theorem 3.2 is proved.

References

Andersen PK, Gill RD. Cox's regression model for counting processes: A large samle study. Annals of Statistics. 1982; 10:1100–1120.

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case–cohort designs. Lifetime Data Analysis. 2000; 6:39–58. [PubMed: 10763560]

Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988; 75:11–20.

Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regressiion parameters under two-phase, outcome-dependent sampling. Journal of the Royal Statistical Society. 1997; 59:447–461.Series B

Breslow NE, McNeney B, Wellner JA. Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. The annals of Statistics. 2003; 31:1110–1139.

Cai J, Zeng D. Sample size/power calculation for case–cohort studies. Biometrics. 2004; 60:1015–1024. [PubMed: 15606422]

Cai J, Zeng D. Power calculation for case–cohort studies with nonrare events. Biometrics. 2007; 63:1288–1295. [PubMed: 17608788]

Chen K. Generalized case–cohort sampling. Journal of the Royal Statistical Society. 2001; 63:791–809.Series B

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

Foutz R. On the unique consistent solution to the likelihood equations. Journal of The American Statistical Association. 1977; 72:147–148.

Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1951; 47:663–685.

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]

Kreuzer M, Grosche B, Schnelzer M, Tschense A, Dufey F, Walsh L. Radon and risk of death from cancer and cardiovascular diseases in the German uranium miners cohort study : follow-up 1946–2003. Radiation and Environmental Biophysics. 2010; 49:177–185. [PubMed: 19855993]

Kreuzer M, Walsh L, Schnelzer M, Tschense A, Grosche B. Radon and risk of extrapulmonary cancers: results of the German uranium miner's cohort study. British Journal of Cancer. 2008; 99:1945–1953.

Kulich M, Řeřicha V, Řeřicha R, Shore DL, Sander D. Incidence of non-lung solid cancers in Czech uranium miners: a case–cohort study. Enviromental Health. 2011; 111:400–405.

Lin D, Wei L, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society. 2000; 62:711–730.Series B

National Research Council. Committee on the Biological Effects of lionizing Radiation (BEIR VI), Health effects of exposure to radon. National Academy Press; Washington DC: 1999.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–412.

Řeřicha V, Kulich M, Řeřicha R, Shore DL, Sander D. Incidence of leukemia, lymphoma, and multiple myeloma in Czech uranium miners: a case–cohort study. Environmental Health Perspective. 2006; 114:818–822.

Samuelsen S, Ȧnestad H, Skrondal A. Stratified case–cohort analysis of general cohort sampling designs. Scandinavian Journal of Statistics. 2007; 34:103–119.

Sandler DP, Shore DL, Solansky I, Rericha V, Hnizdo E, Sram R. Lung cancer in Czech Uranium miners. Epidemiology. 1998; 9:S44.

Sandler DP, Shore DL, Solansky I, Rericha V, Hnizdo E, Sram R. Lung cancer incidence in Czech uranium miners with low-level radon exposure. Am J Epidemiology. 1998; 147:S86.

Scheike T, Martinussen T. Maximum likelihood estimation in Cox's regression model under case–cohort sampling. Scandinavian Journal of Statistics. 2004; 31:283–293.

Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case–cohort studies. Annals of Statistics. 1988; 16:64–81.

Song R, Zhou H, Kosorok M. A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. Biometrika. 2009; 96:221–228. [PubMed: 20107493]

Sun J, Sun L, Flournoy N. Additive hazards model for competing risks analysis of the case–cohort design. Communications in Statistics — Theory and Methods. 2004; 33:351–366.

Tirmarche M, Raphalen A, Allin F, Bredon P. Mortality of a cohort of French uranium miners exposure to relatively low radon concentrations. British Journal of Cancer. 1993; 67:1090–1097. [PubMed: 8494704]

Vacquier B, Caer S, Rogel A, Feurprier M, Tirmarche M, Luccioni C, Quesne B, Acker A, Laurier D. Mortality risk in the French cohort of uranium miners: extended follow-up 1964–1999. Occupational Environmental Medicine. 2008; 65:597–604. [PubMed: 18096654]

vander Vaart, AW.; Wellner, JA. Weak convergence and empirical processes. Springer-Verlag; New York: 1996.

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]

Weaver MA, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. Journal of The American Statistical Association. 2005; 100:459–469.

Weinberg CR, Wacholder S. Prospective analysis of case-control data under general multiplicative intercept risk models. Biometrika. 1993; 80:461–465.

Zhang H, Schaubel DE, Kalbfleisch J. Proportional hazards regression for the analysis of clustered survival data from case–cohort studies. Biometrics. 2011; 67:18–28. [PubMed: 20560939]

Zhou H, Chen J, Rissnen T, Korrick S, Hu H, Salonen J, Longnecker MP. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. Epidemiology. 2007; 18:461–468. [PubMed: 17568219]

Zhou H, Qin G, Longnecker M. A partial linear model in the outcome-dependent sampling setting to evaluate the effect of prenatal PCB exposure on cognitive function in children. Biometrics. 2011; 67:876–885. [PubMed: 21039397]

Zhou H, Weaver M, Qin J, Longnecker M, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 2002; 58:413–421. [PubMed: 12071415]

Highlights

1. Proposing the outcome-dependent sampling design for the survival time subjecting to right censor
2. To account for the biased sampling scheme, we derive estimators from a weighted partial likelihood estimating equation.
3. A criteria that can be used to optimally implement the ODS design in practice is proposed and studied.

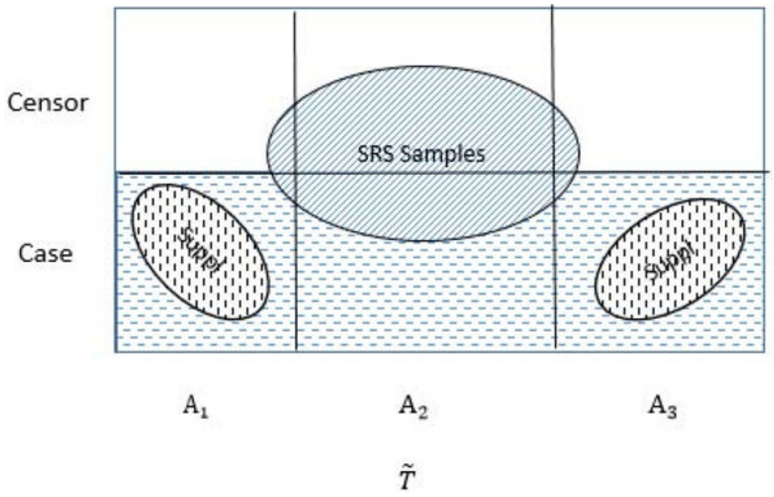

Figure 1.

Outcome-dependent sampling mechanism, SRS samples: subcohort by simple random sample, Suppl: supplemental failure sample, interval A_1: $(0, a_1]$, interval A_2: $(a_1, a_2]$, interval A_3: (a_2, ∞), a_1 and a_2 are the cutpoints.

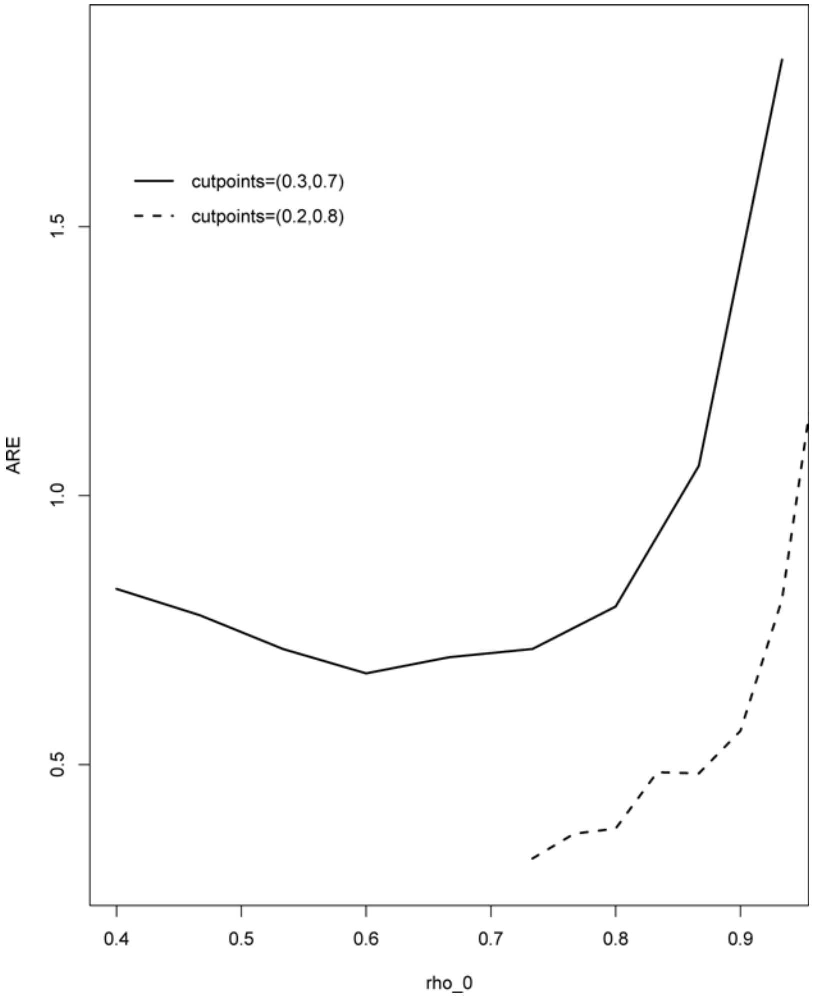

Figure 2.

Asymptotic relative efficiency between $\hat{\beta}_{SRS}$ and $\hat{\beta}_{ODS}$ with $\rho_V = 0.15$.

Table 1

Results of simulation I

			$\beta_1 = 0.5$				$\beta_2 = 0$			
Censoring	**(n_1, n_3)**	**Method**	**Mean**	**SE**	$\widehat{\mathrm{SE}}$	**CI**	**Mean**	**SE**	$\widehat{\mathrm{SE}}$	**CI**
80%	(19,4)	$\hat{\beta}_{Full}$	0.505	0.095	0.096	0.960	−0.001	0.189	0.184	0.945
		$\hat{\beta}_{R}$	0.513	0.143	0.138	0.939	−0.007	0.266	0.263	0.945
		$\hat{\beta}_{SRS}$	0.507	0.128	0.133	0.952	0.004	0.257	0.253	0.953
		$\hat{\beta}_{GCC}$	0.509	0.135	0.134	0.946	−0.005	0.260	0.253	0.945
		$\hat{\beta}_{ODS}$	0.506	0.127	0.128	0.958	−0.005	0.245	0.244	0.951
	(27,5)	$\hat{\beta}_{Full}$	0.505	0.095	0.096	0.960	−0.001	0.189	0.184	0.945
		$\hat{\beta}_{R}$	0.513	0.143	0.138	0.939	−0.007	0.266	0.263	0.945
		$\hat{\beta}_{SRS}$	0.508	0.128	0.131	0.950	0.004	0.251	0.249	0.947
		$\hat{\beta}_{GCC}$	0.505	0.123	0.122	0.944	−0.002	0.235	0.231	0.942
		$\hat{\beta}_{ODS}$	0.510	0.121	0.121	0.956	−0.001	0.234	0.229	0.950
70%	(23,7)	$\hat{\beta}_{Full}$	0.506	0.079	0.078	0.951	−0.002	0.152	0.150	0.954
		$\hat{\beta}_{R}$	0.507	0.115	0.112	0.948	−0.009	0.220	0.214	0.951
		$\hat{\beta}_{SRS}$	0.507	0.110	0.106	0.947	0.000	0.200	0.203	0.957
		$\hat{\beta}_{GCC}$	0.509	0.110	0.106	0.936	0.001	0.204	0.204	0.948

			$\beta_1 = 0.5$				$\beta_2 = 0$			
Censoring	**(n_1, n_3)**	**Method**	**Mean**	**SE**	$\widehat{\text{SE}}$	**CI**	**Mean**	**SE**	$\widehat{\text{SE}}$	**CI**
		$\hat{\beta}_{ODS}$	0.508	0.106	0.104	0.948	−0.003	0.197	0.199	0.957
	(37,12)	$\hat{\beta}_{Full}$	0.506	0.079	0.078	0.951	−0.002	0.152	0.150	0.954
		$\hat{\beta}_{R}$	0.507	0.115	0.112	0.948	−0.009	0.220	0.214	0.951
		$\hat{\beta}_{SRS}$	0.507	0.107	0.102	0.947	−0.001	0.195	0.196	0.960
		$\hat{\beta}_{GCC}$	0.510	0.099	0.097	0.943	−0.004	0.191	0.186	0.949
		$\hat{\beta}_{ODS}$	0.509	0.098	0.097	0.953	0.003	0.185	0.185	0.954

- Mean: point estimate, SE: standard error, $\widehat{\text{SE}}$: estimated standard error, CI: Coverage for nominal 95% confidence intervals.

- $\hat{\beta}_{Full}$, $\hat{\beta}_{R}$ and $\hat{\beta}_{SRS}$ are the standard pseudo-score estimator based on full cohort, SRS sub-cohort and SRS sample with same size as ODS design, respectively.

- $\hat{\beta}_{GCC}$ and $\hat{\beta}_{ODS}$ denote the proposed estimator based on the GCC design and our proposed failure time ODS, respectively.

- n_1 and n_3 are the number of supplemental failures from A_1 and A_3, respectively.

Table 2

Results of simulation II

			$\beta_1 = 1$				$\beta_2 = -1$			
Cutpoints	**(n_1, n_3)**	**Method**	**Mean**	**SE**	$\widehat{\mathrm{SE}}$	**CI**	**Mean**	**SE**	$\widehat{\mathrm{SE}}$	**CI**
(20%, 80%)	(25,5)	$\hat{\beta}_{Full}$	1.005	0.080	0.079	0.950	−1.010	0.134	0.139	0.959
		$\hat{\beta}_{R}$	1.007	0.112	0.112	0.951	−1.008	0.201	0.197	0.943
		$\hat{\beta}_{SRS}$	1.008	0.111	0.108	0.952	−1.015	0.191	0.190	0.952
		$\hat{\beta}_{GCC}$	1.005	0.123	0.120	0.944	−1.003	0.204	0.210	0.953
		$\hat{\beta}_{ODS}$	1.009	0.105	0.107	0.957	−1.011	0.189	0.189	0.948
	(35,8)	$\hat{\beta}_{Full}$	1.005	0.080	0.079	0.950	−1.010	0.134	0.139	0.959
		$\hat{\beta}_{R}$	1.007	0.112	0.112	0.951	−1.008	0.201	0.197	0.943
		$\hat{\beta}_{SRS}$	1.007	0.108	0.107	0.956	−1.016	0.187	0.188	0.960
		$\hat{\beta}_{GCC}$	1.008	0.111	0.109	0.957	−1.007	0.190	0.193	0.955
		$\hat{\beta}_{ODS}$	1.007	0.102	0.103	0.955	−1.013	0.181	0.183	0.959
(30%, 70%)	(30,12)	$\hat{\beta}_{Full}$	1.002	0.077	0.078	0.957	−1.005	0.137	0.138	0.955
		$\hat{\beta}_{R}$	1.008	0.108	0.111	0.960	−1.005	0.201	0.196	0.947
		$\hat{\beta}_{SRS}$	1.007	0.106	0.106	0.954	−1.012	0.188	0.186	0.954
		$\hat{\beta}_{GCC}$	1.009	0.110	0.110	0.947	−1.007	0.197	0.194	0.941

			$\beta_1 = 1$				$\beta_2 = -1$			
Cutpoints	**(n_1, n_3)**	**Method**	**Mean**	**SE**	$\widehat{\mathrm{SE}}$	**CI**	**Mean**	**SE**	$\widehat{\mathrm{SE}}$	**CI**
		$\hat{\beta}_{ODS}$	1.007	0.104	0.104	0.950	−1.006	0.184	0.185	0.958
	(45,15)	$\hat{\beta}_{Full}$	1.002	0.077	0.078	0.957	−1.005	0.137	0.138	0.955
		$\hat{\beta}_{R}$	1.008	0.108	0.111	0.960	−1.005	0.201	0.196	0.947
		$\hat{\beta}_{SRS}$	1.007	0.102	0.104	0.961	−1.012	0.186	0.183	0.948
		$\hat{\beta}_{GCC}$	1.009	0.101	0.102	0.944	−1.004	0.183	0.180	0.937
		$\hat{\beta}_{ODS}$	1.012	0.099	0.100	0.947	−1.007	0.178	0.178	0.948

Notations are the same as in Table 1.

Table 3

Analysis results for Cancer Incidence and Mortality of Uranium Miners Study: the listed values are the original values ×10^{-2}

Methods	$\hat{\beta}$	$\mathrm{SE}\left(\hat{\beta}\right)$	**95%CI**
$\hat{\beta}_{SRS}$			
Trad	0.006	0.050	(−0.091, 0.104)
Age	4.600	0.502	(3.617, 5.583)
Tdust	0.002	0.001	(0.001, 0.004)
Smoking	76.879	13.839	(49.754, 104.004)
$\hat{\beta}_{GCC}$			
Trad	0.027	0.048	(−0.068, 0.122)
Age	4.516	0.440	(3.654, 5.377)
Tdust	0.002	0.001	(0.001, 0.004)
Smoking	74.055	13.468	(47.658, 100.452)
$\hat{\beta}_{ODS}$			
Trad	0.025	0.042	(−0.057 0.108)
Age	4.835	0.458	(3.937, 5.733)
Tdust	0.002	0.001	(0.001, 0.004)
Smoking	73.313	12.886	(48.056, 98.570)

Note: Trad is the total radon exposure, Tdust is the total airborne dust. $\hat{\beta}_{SRS}$: the estimator obtained by simple random sampling; $\hat{\beta}_{GCC}$: the estimator obtained by generalized Case-Cohort sampling; $\hat{\beta}_{ODS}$: the estimator obtained by ODS sampling. The three methods base on the same size of the sample.