Biostatistics (2014), **15**, 4, *pp.* 636–650
doi:10.1093/biostatistics/kxu016
Advance Access publication on May 7, 2014

Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme

JIELI DING

School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

HAIBO ZHOU*

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
zhou@bios.unc.edu

YANYAN LIU

School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

JIANWEN CAI

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

MATTHEW P. LONGNECKER

National Institute of Environmental Health Sciences, National Institute of Health, Research Triangle Park, NC 27709, USA

SUMMARY

Motivated by the need from our on-going environmental study in the Norwegian Mother and Child Cohort (MoBa) study, we consider an outcome-dependent sampling (ODS) scheme for failure-time data with censoring. Like the case-cohort design, the ODS design enriches the observed sample by selectively including certain failure subjects. We present an estimated maximum semiparametric empirical likelihood estimation (EMSELE) under the proportional hazards model framework. The asymptotic properties of the proposed estimator were derived. Simulation studies were conducted to evaluate the small-sample performance of our proposed method. Our analyses show that the proposed estimator and design is more efficient than the current default approach and other competing approaches. Applying the proposed approach with the data set from the MoBa study, we found a significant effect of an environmental contaminant on fecundability.

Keywords: Biased-sampling; Empirical likelihood; Proportional hazards model; Survival analysis.

*To whom correspondence should be addressed.

1. Introduction

In many epidemiologic studies and disease prevention trials, much of the cost is spent on acquiring measurements of the main exposure variable. Large cohort studies with simple random sampling are too expensive to conduct for investigators with a limited budget. Alternative cost-efficient designs and procedures are therefore desirable and may play a critical role in reaching the prespecified power level for many studies with a limited budget. Outcome-dependent sampling (ODS) (e.g. the case–control study) is a retrospective sampling scheme that enhances the efficiency and reduces the cost of a study by allowing investigators observe the exposure with a probability that depends on the value of the outcome (e.g. Cornfield, 1951; Weinberg and Wacholder, 1993; Whittemore, 1997). Recent work has focused on a more general ODS design for continuous outcomes (Zhou *and others*, 2002; Chatterjee *and others*, 2003; Weaver and Zhou, 2005). The principle idea of such a design is to concentrate resources on a segment of the population that conveys the most information about the exposure–response relationship (Song *and others*, 2009; Zhou, Song *and others*, 2011; Zhou, Wu *and others*, 2011).

For the time-to-event data, the case-cohort design (Prentice, 1986) is a well-known biased-sampling scheme for censored failure-time data. The case-cohort design measures the covariates on a simple random sample (SRS) (subcohort) as well as on all the failures at the end of study (e.g. Sun *and others*, 2004; Lu and Tsiatis, 2006; Breslow and Wellner, 2007; Tsai, 2009). When the number of failures is large, a generalized case-cohort design has been proposed where, in addition to a random sample, the information on covariates is assembled only for a subset of the failures instead of all the failures to reduce the cost (e.g. Chen, 2001; Cai and Zeng, 2007; Kang and Cai, 2009). Case-cohort and generalized case-cohort designs are especially advocated when censoring of cases is frequent.

Our research is motivated by a recent substudy of the Norwegian Mother and Child Cohort (MoBa) about the potential health effects of perfluoroalkyl substances (PFASs) (Whitworth *and others*, 2012). PFASs are man-made chemicals that are widely used as industrial surfactants and emulsifiers and in a variety of consumer products. Two of the most widely detected and studied PFASs are perfluorooctane sulfonate (PFOS) and perfluorooctanoic acid (PFOA). Both PFOS and PFOA have shown the potential for toxicity in animal studies (e.g. Johansson *and others*, 2008). In human studies, several studies have linked PFOS and PFOA levels to lower birth weight, increased cholesterol, increased rates of cancer (e.g. Alexander *and others*, 2003), and reduced human fertility (e.g. Fei *and others*, 2009).

Our interests are focused on assessing the relationship between exposure to PFASs and women's subfecundity. Measurements to estimate fecundity were ascertained as time to pregnancy (TTP), reported by women around gestational week 17. Because of the expense measuring the PFAS levels, Whitworth *and others* (2012) chose two groups of women for measurement of PFAS levels: an overall SRS of 550 women from the cohort and a supplemental sample of 400 women sampled from those who delivered a child and had a TTP > 12 months. The MoBa substudy was designed to take advantage of the ODS scheme to yield more powerful and efficient inferences. In this paper, we consider a general failure-time ODS sampling scheme for the MoBa data.

One frequent approach in epidemiology for the above MoBa data is to dichotomize the TTP and apply logistic regression for a binary response. The odds ratios based on this logistic regression are then computed (e.g. per ng/ml of PFOA). Loss of information and bias may result. There is also the risk for misclassification and the estimations may not be comparable if different cutpoints are chosen to dichotomize the outcome. We assess the relationship between the exposure of interest and time-to-event response by analyzing the right-censored data obtained by the above ODS scheme under the framework of the proportional hazards model (Cox, 1972). We develop an estimated maximum semiparametric empirical likelihood approach where we replace the baseline cumulative hazard function and survival function of censoring time in the joint likelihood with some consistent estimators and then maximize it by an empirical approach without specifying the marginal distribution of covariates. We illustrate the proposed method through simulations and compare it with the results from different competing methods. Software and

practical recommendations are provided to researchers who will deal with biased-sampling failure-time data in practice.

The layout of the remainder of this article is as follows. In Section 2, we describe the proposed failure-time ODS design, present an estimated semiparametric empirical likelihood estimator, and develop the asymptotic properties of the proposed estimator. In Section 3, we conduct simulation studies to compare its efficiency with some alterative methods. In Section 4, we apply our proposed method to analyze a data set from the MoBa study. In Section 5, we give some final remarks.

2. Design and estimation

2.1 *ODS design and notations*

Suppose that there exists a large, but finite, study population of N independent individuals. Let $\tilde{T}_i$ denote the failure time and C_i denote the censoring time for subject i $(i = 1, \ldots, N)$. The observed time is $T_i = \min(\tilde{T}_i, C_i)$. Let $\Delta_i = I(\tilde{T}_i \leqslant C_i)$ denote the right-censoring indicator for subject i, $Y_i(t) = I(T_i \geqslant t)$ denote the at-risk process, and $N_i(t) = \Delta_i I(T_i \leqslant t)$ denote the counting process, where $I(\cdot)$ is an indicator function. We confine our attention to non-time-dependent covariates. Let Z_i be a p-dimensional covariate for subject i. Assume that $\tilde{T}_i$ and C_i are conditionally independent given Z_i. Let τ denote the end time for the study.

Suppose that the failure-time $\tilde{T}_i$ follows the following proportional hazards model (Cox, 1972):

$$\lambda(t|Z_i) = \lambda_0(t)\exp(\beta' Z_i), \quad i = 1, \ldots, N, \tag{2.1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, and β is a p-dimensional regression coefficient of primary interest. We assume that the range of observed failure time of all the cases is partitioned into K mutually exclusive and exhaustive strata: $A_k = (a_{k-1}, a_k]$, $k = 1, \ldots, K$, by some known constants $\{a_i, i = 1, \ldots, K\}$ which satisfy $0 = a_0 < a_1 < \cdots < a_{k-1} < a_k = \tau$. We consider the following ODS design where Z is observed: First, a random sample of size n_0 from the full cohort, denoted by the SRS sample, is selected. In addition, we select a supplemental sample of size n_k from each of the above kth stratum of cases. The samples from these two components constitute the ODS sample. We suppose that n_k is fixed by design for $k = 1, \ldots, K$. We denote $n = \sum_{k=0}^{K} n_k$ to be the total size of the ODS sample. Let V, S_0, and S_k be the index set of the total ODS sample, the SRS sample, and the supplemental sample from the kth stratum, respectively. Hence, the observed data for our ODS design can be summarized as

$$\begin{aligned} &\text{The SRS sample: } (T_i, \Delta_i, Z_i), \quad i \in S_0; \\ &\text{The supplemental sample: } (T_i, \Delta_i, Z_i|\Delta_i = 1,\ T_i \in A_k), \quad i \in S_k,\ k = 1, \ldots, K. \end{aligned} \tag{2.2}$$

The likelihood function corresponding to the observed data described in (2.2) is

$$L(\beta, Q_Z) = \left[\prod_{i \in S_0} f(T_i, \Delta_i|Z_i) q_Z(Z_i)\right] \cdot \left[\prod_{k=1}^{K} \prod_{i \in S_k} f(T_i, \Delta_i, Z_i|\Delta_i = 1,\ T_i \in A_k)\right], \tag{2.3}$$

where $Q_Z(\cdot)$ and $q_Z(\cdot)$ denote the cumulative distribution and density function of Z, respectively, $f(T_i, \Delta_i|Z_i)$ denotes the conditional distribution function of (T_i, Δ_i) given Z_i, and $f(T_i, \Delta_i, Z_i|\Delta_i = 1, T_i \in A_k)$ denotes the joint density function of (T_i, Δ_i, Z_i), conditional on the censoring indicator Δ_i being 1, and the failure-time T_i being in interval A_k. By applying Bayes' Law to the supplemental samples

in the second bracket of (2.3), we can rewrite (2.3) as

$$L(\beta, Q_Z) = \left[\prod_{i\in S_0} f(T_i, \Delta_i|Z_i)q_Z(Z_i)\right] \cdot \left[\prod_{k=1}^{K}\prod_{i\in S_k} \frac{f(T_i, \Delta_i|Z_i)q_Z(Z_i)I(\Delta_i = 1,\ T_i \in A_k)}{P(\Delta_i = 1,\ T_i \in A_k)}\right]. \tag{2.4}$$

Under random censorship, $f(T_i, \Delta_i|Z_i) = f_{\beta,\Lambda_0}(T_i|Z_i)S_C(T_i)$ when $\Delta_i = 1$ and $f(T_i, \Delta_i|Z_i) = \bar{F}_{\beta,\Lambda_0}(T_i|Z_i)s_C(T_i)$ when $\Delta_i = 0$, where $f_{\beta,\Lambda_0}(t|Z)$ and $\bar{F}_{\beta,\Lambda_0}(t|Z)$ are the conditional density function and survival function of $\tilde{T}$ given Z with the baseline cumulative hazard function $\Lambda_0(t)$, respectively, and $s_C(t)$ and $S_C(t)$ are the density function and survival function of the censoring time C, respectively. We assume that $S_C(t)$ is independent on the covariate Z. Thus, we have $P(\Delta_i = 1, T_i \in A_k) = \int_{\mathcal{Z}}\int_{A_k} f_{\beta,\Lambda_0}(t|Z)S_C(t)\,\mathrm{d}t\,\mathrm{d}Q_Z(Z)$. The likelihood function in (2.4) is proportional to

$$\begin{aligned} L(\beta, Q_Z, \Lambda_0, S_C) = &\left[\prod_{i\in S_0}(f_{\beta,\Lambda_0}(T_i|Z_i))^{\Delta_i}(\bar{F}_{\beta,\Lambda_0}(T_i|Z_i))^{1-\Delta_i}\right] \cdot \left[\prod_{k=1}^{K}\prod_{i\in S_k} f_{\beta,\Lambda_0}(T_i|Z_i)\right] \\ &\cdot \left[\prod_{k=0}^{K}\prod_{i\in S_k} q_Z(Z_i)\right] \cdot \left[\prod_{k=1}^{K}\left(\int_{\mathcal{Z}}\int_{A_k} f_{\beta,\Lambda_0}(t|Z)S_C(t)\,\mathrm{d}t\,\mathrm{d}Q_Z(Z)\right)^{-n_k}\right]. \end{aligned} \tag{2.5}$$

Note that the non-parametric portion (Q_Z, Λ_0, S_C) cannot be separated from the above likelihood function that combines both the conditional parametric likelihood and the marginal semiparametric likelihood. Clearly, the inference for the underlying parameters requires methods to deal with (Q_Z, Λ_0, S_C), which are effectively infinite-dimensional nuisance functions. For all these challenges, we develop next an estimated maximum semiparametric empirical likelihood approach, in which we replace (Λ_0, S_C) in the joint likelihood $L(\beta, Q_Z, \Lambda_0, S_C)$ with their estimators to get an estimated likelihood function $\hat{L}(\beta, Q_Z)$, and then maximize $\hat{L}(\beta, Q_Z)$ with respect to (β, Q_Z) by a semiparametric empirical approach without specifying Q_Z.

2.2 *An estimated maximum semiparametric empirical likelihood approach*

First, we estimate the baseline cumulative hazard function $\Lambda_0(t)$ by the Breslow–Aalen estimator

$$\hat{\Lambda}_0(t, \hat{\beta}^{(0)}) = \sum_{T_j\leqslant t,\ j\in S_0} \frac{\Delta_j}{\sum_{l\in S_0} Y_l(T_j)\,\mathrm{e}^{\hat{\beta}^{(0)\prime}Z_l}},$$

and the survival function $S_C(t)$ of censoring time C by the Nelson–Aalen estimator

$$\hat{S}_C(t) = \exp\left(-\sum_{T_j\leqslant t,\ j\in S_0} \frac{1-\Delta_j}{\tilde{n}_j}\right),$$

where $\tilde{n}_j$ denotes the number of subjects at risk at a time prior to T_j (for $j \in S_0$) and $\hat{\beta}^{(0)}$ is the estimate of β based only on the SRS portion. Replacing (Λ_0, S_C) in the likelihood function (2.5) with $(\hat{\Lambda}_0, \hat{S}_C)$, we

obtain the estimated log-likelihood function:

$$\hat{l}(\beta, Q_Z) = l(\beta, Q_Z, \hat{\Lambda}_0, \hat{S}_C) \propto \sum_{i \in S_0} \log h_\beta(T_i, \Delta_i, Z_i) + \sum_{k=1}^{K} \sum_{i \in S_k} \log f_{\beta, \hat{\Lambda}_0}(T_i | Z_i) + \sum_{k=0}^{K} \sum_{i \in S_k} \log q_Z(Z_i) - \sum_{k=1}^{K} n_k \log \pi_k, \tag{2.6}$$

where

$$h_\beta(T_i, \Delta_i, Z_i) = \left(\frac{\mathrm{e}^{\beta' Z_i}}{\sum_{l \in S_0} Y_l(T_i)\, \mathrm{e}^{\beta' Z_l}} \right)^{\Delta_i},$$

which is obtained by an extension of the result of Johansen (1983) for the Cox model, and

$$\pi_k \equiv \int_{\mathcal{Z}} P_k(Z; \beta) \mathrm{d}Q_Z(Z) \equiv \int_{\mathcal{Z}} \left(\int_{A_k} f_{\beta, \hat{\Lambda}_0}(t | Z) \hat{S}_C(t)\, \mathrm{d}t \right) \mathrm{d}Q_Z(Z), \quad k = 1, \ldots, K,$$

which are the stratum-specific estimated probabilities of the failure time across all cases.

Maximizing $\hat{l}(\beta, Q_Z)$ with respect to β without specifying Q_Z is not straightforward. We first profile the likelihood function in (2.6) by fixing β and replacing Q_Z with the empirical likelihood function (Vardi, 1982, 1985). To maximize $\hat{l}(\beta, Q_Z)$ over all distributions whose support contains the observed Z values, we only need to consider the discrete conditional distribution of Z with jumps at each of the observed points (Owen, 1990). Denote

$$p_i = q_Z(Z_i) = \mathrm{d}Q_Z(Z_i), \quad i = 1, \ldots, n.$$

For a fixed β, we have

$$\hat{l}(\beta, \{p_i\}) = \sum_{i \in S_0} \log h_\beta(T_i, \Delta_i, Z_i) + \sum_{k=1}^{K} \sum_{i \in S_k} \log f_{\beta, \hat{\Lambda}_0}(T_i | Z_i) + \sum_{i \in V} \log p_i - \sum_{k=1}^{K} n_k \log \left(\sum_{i \in V} P_k(Z_i; \beta) p_i \right). \tag{2.7}$$

We use the Lagrange multiplier argument to search for $\{\hat{p}_i\}$ that maximize (2.7) under the constraints $\{p_i \geqslant 0, \text{ for } i \in V, \ \sum_{i \in V} p_i = 1\}$. The Lagrange function can be written as

$$H(\beta, \{p_i\}, \rho) = \sum_{i \in S_0} \log h_\beta(T_i, \Delta_i, Z_i) + \sum_{k=1}^{K} \sum_{i \in S_k} \log f_{\beta, \hat{\Lambda}_0}(T_i | Z_i) + \sum_{i \in V} \log p_i - \sum_{k=1}^{K} n_k \log \left(\sum_{i \in V} P_k(Z_i; \beta) p_i \right) + \rho \left(1 - \sum_{i \in V} p_i \right),$$

where ρ denotes the Lagrange multipliers. It can be shown that the solutions to the score equation of H with respect to $\{p_i\}$ have the form:

$$\hat{p}_i = \left[n_0 \left(1 + \sum_{k=1}^{K} \frac{n_k}{n_0 \pi_k} P_k(Z_i; \beta) \right) \right]^{-1}, \quad i = 1, \ldots, n.$$

By plugging $\hat{p}_i$ back into $\hat{l}(\beta, \{p_i\})$ in (2.7), we have the resulting profile likelihood function

$$\begin{aligned} \hat{l}(\beta, \pi) = & \sum_{i \in S_0} \log h_\beta(T_i, \Delta_i, Z_i) + \sum_{k=1}^{K} \sum_{i \in S_k} \log f_{\beta, \hat{\Lambda}_0}(T_i | Z_i) \\ & - \sum_{i \in V} \log \left(n_0 \left(1 + \sum_{k=1}^{K} \frac{n_k}{n_0 \pi_k} P_k(Z_i; \beta) \right) \right) - \sum_{k=1}^{K} n_k \log \pi_k, \end{aligned} \tag{2.8}$$

where $\pi' = (\pi_1, \ldots, \pi_K)$. The proposed estimated maximum semiparametric empirical likelihood estimator (EMSELE) is the β that maximizes (2.8). Define $\xi' = (\beta', \pi')$, and denote the EMSELE for ξ to be $\hat{\xi}$ and the EMSELE for parameter β to be $\hat{\beta}_P$, the corresponding portion of $\hat{\xi}$. A Newton–Raphson algorithm can be used to obtain $\hat{\xi}$.

2.3 *Asymptotic properties of EMSELE*

To present the large-sample result, we introduce the following notations:

$$S_0^{(d)}(\beta, t) = \sum_{l \in S_0} Y_l(t) Z^{\otimes d} \, \mathrm{e}^{\beta' Z_l}, \quad s_0^{(d)}(\beta, t) = E(S_0^{(d)}(\beta, t)), \quad d = 0, 1, 2,$$

$$v_0(\beta, t) = \frac{s_0^{(2)}(\beta, t)}{s_0^{(0)}(\beta, t)} - \left(\frac{s_0^{(1)}(\beta, t)}{s_0^{(0)}(\beta, t)} \right)^{\otimes 2} \quad \text{and} \quad A_0(\beta) = \int_0^\tau v_0(\beta, t) s_0^{(0)}(\beta, t) \lambda_0(t) \, \mathrm{d}t.$$

Here, for a vector a, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, $a^{\otimes 2} = aa'$. We indicate the true values of a parameter by superscript "0". Let E_k denote expectation conditional on $\tilde{T} \in A_k$, so that, for any function $\gamma(T, Z)$,

$$E_k(\gamma(T, Z)) = \frac{1}{\pi_k^0} \int_{\mathcal{Z}} \int_{A_k} \gamma(t, Z) f_{\beta^0, \Lambda_0}(t | Z) \, \mathrm{d}t \, \mathrm{d}Q_Z(Z). \tag{2.9}$$

Under some general regularity conditions (see Appendix of supplementary material available at *Biostatistics* online) and assuming that $n_0/n \to \rho_0 > 0$ and $n_k/n \to \rho_k \geqslant 0$ for $k = 1, \ldots, K$, the following theorem establishes the asymptotic properties of the EMSELE $\hat{\xi}$ as well as a consistent estimator for the asymptotic variance matrix.

THEOREM 2.1 Under general regularity conditions, $\hat{\xi}$ converges in probability to ξ^0, while $\sqrt{n}(\hat{\xi} - \xi^0)$ has an asymptotic normal distribution with mean zero and with a variance matrix in the form $\Sigma(\xi^0) = J^{-1}(\xi^0)(\Sigma_1(\xi^0) + \Sigma_2(\xi^0) + \Sigma_3(\xi^0))J^{-1}(\xi^0)$, where $J(\xi)$ is the limiting Hessian matrix of the profile likelihood $\hat{l}(\xi)$, and

$$\Sigma_1(\xi) = \begin{pmatrix} \rho_0 A_0(\beta) & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma_2(\xi) = \begin{pmatrix} \sum_{k=1}^{K} \rho_k \operatorname{Var}_k(e(\beta)) & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma_3(\xi) = \sum_{k=0}^{K} \rho_k \operatorname{Var}_k(g(\xi)),$$

where $e(\beta) = \nabla_\beta f_{\beta,\hat{\Lambda}_0}(T|Z)/f_{\beta,\hat{\Lambda}_0}(T|Z)$, and

$$g(\xi) = \left(-\frac{\sum_{l=1}^{K}((\rho_l/\rho_0)/\pi_l)\nabla_\beta P_l(Z;\beta)}{1+\sum_{l=1}^{K}((\rho_l/\rho_0)/\pi_l)P_l(Z;\beta)}, \frac{((\rho_1/\rho_0)/\pi_1^2)P_1(Z;\beta)}{1+\sum_{l=1}^{K}((\rho_l/\rho_0)/\pi_l)P_l(Z;\beta)} - \frac{\rho_1}{\pi_1}, \ldots, \frac{((\rho_K/\rho_0)/\pi_K^2)P_K(Z;\beta)}{1+\sum_{l=1}^{K}((\rho_l/\rho_0)/\pi_l)P_l(Z;\beta)} - \frac{\rho_K}{\pi_K} \right)'.$$

A consistent estimator for the asymptotic covariance matrix $\Sigma(\xi^0)$ is $\hat{J}^{-1}(\hat{\xi})(\hat{\Sigma}_1(\hat{\xi}) + \hat{\Sigma}_2(\hat{\xi}) + \hat{\Sigma}_3(\hat{\xi}))\hat{J}^{-1}(\hat{\xi})$, where $\hat{J}$, $\hat{\Sigma}_1$, $\hat{\Sigma}_2$, and $\hat{\Sigma}_3$ are obtained by replacing the large-sample quantities in J, Σ_1, Σ_2, and Σ_3 with their corresponding small-sample quantities.

The proof for Theorem 2.1 is given in Appendix (see supplementary material available at *Biostatistics* online).

3. Simulation studies

We conducted simulation studies to assess the finite sample properties of our proposed method. We consider the following Cox's proportional hazards model:

$$\lambda(t|X, Z) = \lambda_0(t)\, e^{\beta_1 X+\beta_2 Z}. \tag{3.1}$$

We took the marginal distribution of failure-time $\tilde{T}$ to be exponential with failure rate $\lambda_0(t)\, e^{\beta_1 X+\beta_2 Z}$. The baseline hazard function $\lambda_0(t)$ was set to be 1. The covariate X was generated from a standard normal distribution and Z was generated from a Bernoulli distribution with $P(Z = 1) = 0.5$. We set $\beta_1 = \log(2)$ and $\beta_2 = -0.5$. The censoring time C was generated from a uniform distribution $[0, c]$ with c chosen to depend on the desired percentage of censoring. We considered censoring rates of approximately 60% and 90% with the corresponding c values 1.35 and 0.22.

For our ODS design, we first generated the SRS sample of n_0. We then partitioned all the cases into three strata, separated by quantiles q_1 and q_2 of failure times in the cases. We sampled the supplemental sample of n_1 and n_3 subjects from the low stratum and the high stratum, respectively. In addition to various configurations for the parameter values, we also chose two pairs of the cutpoints (0.30 and 0.70 quantiles, and 0.15 and 0.85 quantiles, respectively), to investigate the impact of different cutpoints for our ODS design for creating the supplemental samples.

Under each configuration, we compared the proposed estimator, $\hat{\beta}_P$, with three competing estimators: the maximum likelihood estimator based on a SRS of the same size as the ODS sample ($\hat{\beta}_R$); the weighted estimator under generalized case-cohort design developed by Kang and Cai (2009) ($\hat{\beta}_G$); the estimator under the case-cohort design developed by Prentice (1986) ($\hat{\beta}_C$). For calculating $\hat{\beta}_G$, we first selected a subcohort of n_0 by simple random sampling. We then selected a SRS of cases of n_c in the remaining cases, which we set to be the same size as the supplemental samples in ODS design, i.e. $n_c = n_1 + n_3$. We used the weighed estimating equation provided in Kang and Cai (2009) with the time-invariant weight function to obtain $\hat{\beta}_G$. For calculating $\hat{\beta}_C$, we randomly sampled a subcohort of n_0 and took all the remaining cases. In order to obtain an approximate sample size with the ODS samples, we adjusted the size of the full cohort according to different subcohort sizes and different censoring rates. For example, we set the full cohort size to be 1300 when the subcohort size n_0 was 800 and the censoring rate was 60%, and the mean of sample sizes for case-cohort design under 2000 simulations was 1000.2.

Table 1. *Results are based on the model* $\lambda(t|X, Z) = e^{\beta_1 X+\beta_2 Z}$, *where* $X \sim N(0, 1)$ *and* $Z \sim B(0.5)$

				$\beta_1 = \log(2)$				$\beta_2 = -0.5$			
(n_0, n_1, n_3)	Cutpoints	ρ		Mean	SD	SE	CP	Mean	SD	SE	CP
(900, 50, 50)	(0.30, 0.70)	0.60	$\hat{\beta}_R$	0.696	0.054	0.054	0.952	−0.500	0.101	0.102	0.949
			$\hat{\beta}_G$	0.694	0.090	0.104	0.981	−0.497	0.161	0.195	0.982
			$\hat{\beta}_C$	0.696	0.057	0.051	0.920	−0.498	0.103	0.095	0.932
			$\hat{\beta}_P$	0.695	0.053	0.053	0.946	−0.503	0.101	0.100	0.944
		0.90	$\hat{\beta}_R$	0.694	0.105	0.103	0.948	−0.501	0.213	0.208	0.947
			$\hat{\beta}_G$	0.698	0.101	0.099	0.932	−0.508	0.201	0.198	0.950
			$\hat{\beta}_C$	0.698	0.089	0.075	0.904	−0.503	0.171	0.150	0.914
			$\hat{\beta}_P$	0.693	0.085	0.087	0.953	−0.498	0.162	0.169	0.957
	(0.15, 0.85)	0.60	$\hat{\beta}_R$	0.696	0.055	0.055	0.956	−0.502	0.103	0.102	0.943
			$\hat{\beta}_G$	0.695	0.091	0.104	0.976	−0.503	0.160	0.195	0.976
			$\hat{\beta}_C$	0.695	0.057	0.051	0.919	−0.493	0.102	0.095	0.935
			$\hat{\beta}_P$	0.696	0.052	0.053	0.949	−0.495	0.101	0.100	0.943
		0.90	$\hat{\beta}_R$	0.695	0.104	0.104	0.954	−0.507	0.213	0.208	0.942
			$\hat{\beta}_G$	0.696	0.105	0.099	0.932	−0.504	0.199	0.198	0.950
			$\hat{\beta}_C$	0.697	0.084	0.075	0.922	−0.501	0.167	0.150	0.928
			$\hat{\beta}_P$	0.692	0.081	0.086	0.961	−0.516	0.162	0.168	0.954
(800, 100, 100)	(0.30, 0.70)	0.60	$\hat{\beta}_R$	0.696	0.055	0.054	0.950	−0.503	0.102	0.102	0.954
			$\hat{\beta}_G$	0.699	0.069	0.074	0.967	−0.500	0.131	0.138	0.965
			$\hat{\beta}_C$	0.695	0.060	0.048	0.876	−0.503	0.111	0.089	0.883
			$\hat{\beta}_P$	0.696	0.052	0.053	0.950	−0.500	0.103	0.102	0.943
		0.90	$\hat{\beta}_R$	0.693	0.103	0.103	0.949	−0.507	0.207	0.208	0.953
			$\hat{\beta}_G$	0.695	0.082	0.070	0.904	−0.500	0.156	0.141	0.918
			$\hat{\beta}_C$	0.696	0.084	0.062	0.860	−0.509	0.147	0.123	0.901
			$\hat{\beta}_P$	0.694	0.075	0.085	0.974	−0.504	0.138	0.164	0.982
	(0.15, 0.85)	0.60	$\hat{\beta}_R$	0.694	0.055	0.054	0.940	−0.502	0.104	0.102	0.948
			$\hat{\beta}_G$	0.696	0.068	0.074	0.968	−0.504	0.126	0.138	0.969
			$\hat{\beta}_C$	0.695	0.060	0.048	0.894	−0.500	0.112	0.089	0.881
			$\hat{\beta}_P$	0.695	0.050	0.052	0.956	−0.502	0.100	0.100	0.955
		0.90	$\hat{\beta}_R$	0.697	0.103	0.103	0.948	−0.509	0.214	0.208	0.946
			$\hat{\beta}_G$	0.696	0.082	0.071	0.908	−0.509	0.165	0.141	0.902
			$\hat{\beta}_C$	0.698	0.082	0.062	0.865	−0.504	0.155	0.123	0.884
			$\hat{\beta}_P$	0.690	0.073	0.084	0.975	−0.503	0.146	0.164	0.965

The cutpoints for the ODS design were 0.30 and 0.70 sample quantiles and 0.15 and 0.85 sample quantiles, respectively; $\hat{\beta}_R$ denotes the estimator from a simple random sample of the same size as the ODS sample; $\hat{\beta}_G$ denotes the generalized case-cohort estimator developed by Kang and Cai (2009); $\hat{\beta}_C$ denotes the case-cohort estimator developed by Prentice (1986), and the average sample size is 1000 by adjusting the full cohort sample size to different censoring rate ρ; $\hat{\beta}_P$ denotes the proposed EMSELE estimator. Simulation results are based on 2000 simulations with the total ODS sample size $n = 1000$.

The estimated means (Means), standard deviations (SDs), mean of the variance estimates (SEs), and 95% nominal confidence intervals coverages (CPs) for each estimator were obtained from 2000 independently generated data sets. The results are summarized in Table 1. Under all of the cases considered here, the four estimators for β_1 and β_2 are all unbiased. Our proposed variance estimator provides a good estimation for the sample standard errors and the confidence intervals attain coverage close to the nominal

Table 2. *Results are based on the model* $\lambda(t|X, Z) = e^{\beta_1 X + \beta_2 Z}$, *where* $X \sim N(0, 1)$ *and* $Z \sim B(0.5)$

			$\beta_1 = \log(2)$				$\beta_2 = -0.5$			
(n_0, n_1, n_3)	ρ		Mean	SD	SE	CP	Mean	SD	SE	CP
			$C_i \sim U[0, c]$							
(900, 50, 50)	0.60	$\hat{\beta}_P$	0.695	0.053	0.053	0.946	−0.503	0.101	0.100	0.944
(900, 20, 80)		$\hat{\beta}_P$	0.695	0.053	0.053	0.947	−0.505	0.105	0.101	0.939
(900, 80, 20)		$\hat{\beta}_P$	0.695	0.054	0.053	0.946	−0.496	0.101	0.100	0.947
(800, 100, 100)		$\hat{\beta}_P$	0.696	0.052	0.053	0.950	−0.500	0.103	0.102	0.943
(800, 50, 150)		$\hat{\beta}_P$	0.694	0.053	0.053	0.951	−0.506	0.104	0.103	0.948
(800, 150, 50)		$\hat{\beta}_P$	0.694	0.054	0.054	0.953	−0.499	0.099	0.102	0.952
			Censoring Scenario I: $C_i \sim U[0, c_1]I(Z_i = 0) + U[0, c_2]I(Z_i = 1)$							
(900, 50, 50)	0.60	$\hat{\beta}_P$	0.696	0.053	0.053	0.953	−0.472	0.106	0.097	0.921
	0.90	$\hat{\beta}_P$	0.687	0.081	0.368	0.971	0.056	0.203	2.213	0.345
			Censoring Scenario II: $C_i \sim E(1/c_1)I(Z_i = 0) + E(1/c_2\|X_i\|)I(Z_i = 1)$							
(900, 50, 50)	0.60	$\hat{\beta}_P$	0.679	0.050	0.049	0.931	−0.524	0.103	0.106	0.949
	0.90	$\hat{\beta}_P$	0.715	0.079	0.078	0.941	−0.803	0.181	0.184	0.649

The cutpoints for the ODS design were 0.30 and 0.70 sample quantiles. Censoring Scenario I: C_i was generated from the distribution $U[0, c_1]I(Z_i = 0) + U[0, c_2]I(Z_i = 1)$; Censoring Scenario II: C_i was generated from the distribution $E(1/c_1)I(Z_i = 0) + E(1/c_2|X_i|)I(Z_i = 1)$. The results are based on 1000 replicates for each setting. Simulation results are based on the total ODS sample size $n = 1000$.

95% level. We note that the estimation of the sample SDs become less stable at the very high censoring rate (e.g. 90%), which indicates a higher sample size may be needed. Further, the efficiency gains are higher when the cutpoint is further out ((0.15, 0.85) vs. (0.30, 0.70)). We also note that the proposed estimator $\hat{\beta}_P$ is the most efficient among all the estimators compared under all the different censoring rates. $\hat{\beta}_C$ is more efficient than $\hat{\beta}_R$ when the censoring rate is 90%. $\hat{\beta}_G$ is more efficient than $\hat{\beta}_R$ when the censoring rate is 90%. $\hat{\beta}_G$ and $\hat{\beta}_C$ are comparable under censoring rate of 90%. The fact that $\hat{\beta}_P$ is more efficient than $\hat{\beta}_C$ and $\hat{\beta}_G$ indicates that our ODS design for the survival analysis can be a more efficient alternative to the case-cohort design and the generalized case-cohort design. Further, comparing the results in Table 1, we note that, for a given total ODS sample size ($n = 1000$), the efficiency improves as we allocate more individuals in the supplemental samples (e.g. $n_0 = 800,\ n_1 = n_3 = 100$ vs. $n_0 = 900,\ n_1 = n_3 = 50$).

Table 2 provides additional simulation results on the sensitivity analysis and the unbalanced pattern of ODS supplemental samples allocations. We investigated the performance of the proposed estimator under unbalanced values of n_1 and n_3 by choosing $(n_0, n_1, n_3) = (900, 20, 80)$, $(900, 80, 20)$, $(800, 50, 150)$, and $(800, 150, 50)$, respectively. The results reported in Table 2 indicate that, overall, the observed properties of $\hat{\beta}_P$ under Table 1 is consistent for the balanced and unbalanced values of n_1 and n_3.

The second component in Table 2 was conducted to evaluate the performance of $\hat{\beta}_P$ when the censoring time depends on the covariates. We considered the following two scenarios: Scenario I: C_i was generated from the distribution $U[0, c_1]I(Z_i = 0) + U[0, c_2]I(Z_i = 1)$, and (c_1, c_2) was chosen to be $(1, 1.8)$ and $(0.1, 0.4)$; Scenario II: C_i was generated from the distribution $E(1/c_1)I(Z_i = 0) + E(1/c_2|X_i|)I(Z_i = 1)$, and (c_1, c_2) was chosen to be $(1, 1)$ and $(0.15, 0.1)$. The results in Table 2 indicate that the dependence of censoring time on the covariates will lead to biased estimates of β_1 and β_2. This suggests there is a need to check for censoring dependence on covariates before using the estimator in real analysis.

4. Analysis of the MoBa study data

The MoBa study is an ongoing pregnancy cohort study conducted by the Norwegian Institute of Public Health. Pregnant women in Norway were enrolled from 1999 to 2008 and completed questionnaires regarding demographic and lifestyle factors, and medical and reproductive history. Women were asked if their pregnancy was planned and reported TTP. Subfecundity was defined as having a TTP > 12 months. Our data set was based on women enrolled from 2003 to 2004 who delivered a live born child. Five hundred and fifty subjects were randomly sampled from the cohort who reported a TTP with 11 subjects excluded. Four hundred subjects were supplementally sampled from women whose TTP was > 12 months (Whitworth *and others*, 2012). In this data set, there was no censoring. Hence, there is no need to check the independence of censoring on covariates here.

Among these 939 eligible women, blood samples were collected around gestational week 17. Concentrations of PFOS and PFOA were measured from the maternal blood samples by high performance liquid

Table 3. *Demographics and characteristics of the MoBa study*

	All individuals	TTP $\leqslant 12$ months	TTP >12 months
PFOS, mean $\pm$ SD	14.05 ± 6.22	13.70 ± 5.38	14.46 ± 7.07
PFOA, mean $\pm$ SD	2.45 ± 1.17	2.34 ± 1.09	2.58 ± 1.24
BMI, mean $\pm$ SD	24.88 ± 4.83	24.39 ± 4.27	25.47 ± 5.36
FatherAge mean $\pm$ SD	33.65 ± 5.25	32.90 ± 5.03	34.55 ± 5.37
Alb, mean $\pm$ SD	3.90 ± 0.22	3.90 ± 0.22	3.90 ± 0.21
Oilyfish, mean $\pm$ SD	11.66 ± 13.86	12.50 ± 14.39	10.64 ± 13.14
Leanfish, mean $\pm$ SD	24.49 ± 15.68	24.86 ± 15.79	24.05 ± 15.56
MotherAge (%)			
<25	17.36 (163/939)	13.95 (71/509)	21.40 (92/430)
25–29	41.21 (387/939)	41.45 (211/509)	40.93 (176/430)
30–34	31.63 (297/939)	34.18 (174/509)	28.60 (123/430)
$\geqslant 35$	9.80 (92/939)	10.41 (53/509)	9.07 (39/430)
MotherEdu (%)			
<High school	8.86 (83/937)	6.69 (34/508)	11.42 (49/429)
High school and other	31.70 (297/937)	30.51 (155/508)	33.10 (142/429)
Some college	41.73 (391/937)	43.90 (223/508)	39.16 (168/429)
>College	17.72 (166/937)	18.90 (96/508)	16.32 (70/429)
FatherEdu (%)			
<High school	12.29 (112/911)	10.44 (52/498)	14.53 (60/413)
High school and other	44.13 (402/911)	44.18 (220/498)	44.07 (182/413)
Some college	26.89 (245/911)	27.71 (138/498)	25.91 (107/413)
>College	16.68 (152/911)	17.67 (88/498)	15.50 (64/413)
Smoke3 (%)			
None	69.86 (656/939)	73.48 (374/509)	65.58 (282/430)
Sometimes	10.12 (95/939)	9.63 (49/509)	10.70 (46/430)
Daily	20.02 (188/939)	16.90 (86/509)	23.72 (102/430)
Smoke17 (%)			
None	76.25 (716/939)	79.57 (405/509)	72.33 (311/430)
Stopped	15.76 (148/939)	14.73 (75/509)	16.98 (73/430)
Sometimes	1.28 (12/939)	1.38 (7/509)	1.16 (5/430)
Daily	6.71 (63/939)	4.32 (22/509)	9.53 (41/430)

Continued

Table 3. *Continued*

MotherDrink (%)			
>1 per week	7.22 (67/928)	7.17 (36/502)	7.28 (31/426)
1 per week	15.41 (143/928)	18.73 (94/502)	11.50 (49/426)
1–3 per month	32.87 (305/928)	34.86 (175/502)	30.52 (130/426)
<1 per month or never	44.50 (413/928)	39.24 (197/502)	50.70 (216/426)
SexFreq (%)			
<1 per week	18.45 (171/927)	17.33 (87/502)	19.76 (84/425)
1–2 per week	37.97 (352/927)	33.67 (169/502)	43.06 (183/425)
>2 per week	43.58 (404/927)	49.00 (246/502)	37.18 (158/425)
Endo (%)			
Yes	3.09 (29/939)	0.59 (3/509)	6.05 (26/430)
No	96.91 (910/939)	99.41 (506/509)	93.95 (404/430)
Ovary (%)			
Yes	2.66 (25/939)	2.36 (12/509)	3.02 (13/430)
No	97.34 (914/939)	97.64 (497/509)	96.98 (417/430)
Std (%)			
Yes	12.57 (118/939)	12.38 (63/509)	12.79 (55/430)
No	87.43 (821/939)	87.62 (446/509)	87.21 (375/430)
Diabete (%)			
Yes	1.60 (15/939)	0.39 (2/509)	3.02 (13/430)
No	98.40 (924/939)	99.61 (507/509)	96.98 (417/430)
YearDraw (%)			
2003	49.52 (465/939)	50.10 (255/509)	48.84 (210/430)
2004	50.48 (474/939)	49.90 (254/509)	51.16 (220/430)

chromatography/tandem mass spectrometry based on 150 μl of plasma. In the analysis, we included the following variables as potential confounders: pre-pregnancy body mass index (BMI), maternal plasma albumin concentration (Alb), maternal consumption of lean fish and oily fish (Leanfish, Oilyfish), maternal age (MotherAge), paternal age (FatherAge), maternal education (MotherEdu), paternal education (FatherEdu), maternal smoking (Smoke3 for smoking 3 months before pregnancy, Smoke17 for smoking at gestational week 17), maternal self-reported alcohol intake 3 months before pregnancy (MotherDrink), frequency of sexual intercourse 1 month before pregnancy (SexFreq), maternal diseases (endometriosis (Endo), ovary/fallopian tube infection (Ovary), sexually transmitted disease (Std), diabetes), and calendar year of blood draw (Yeardraw). Table 3 provides the demographic characteristics for all 939 women.

Odds ratio approach. We first implemented a standard epidemiologic approach by dichotomizing the subfecundity measurement as binary response, i.e. 0 for TTP $\leqslant$ 12 months and 1 for TTP $>$ 12 months, and used logistic regression to model the association between PFOA and subfecundity odds ratio adjusted for the confounders listed above. The result for this approach is summarized in the first column in Table 4. Due to missing values for covariates, the final sample size for the analysis included 910 women, 491 with TTP $\leqslant$ 12 months, and 419 with TTP $>$ 12 months. We note that the odds of subfecundity increases by $e^{0.2303} = 1.26$ times when PFOA level increases one unit. The second and third columns of Table 4 show results for the cutpoints 3 and 24 months. Comparing across the columns, we note that different choices of the cutpoints result in different inferences. The odds ratios for TTP $>$ 3, TTP $>$ 12, and TTP $>$ 24 are $e^{0.2646} = 1.30$, $e^{0.2303} = 1.26$, and $e^{0.114} = 1.12$, respectively. Further, the effect of PFOA becomes not significant when the cutpoint changes from 12 to 24 months ($p = 0.1527$).

Table 4. *Logistic regression analyses for covariate effects on subfecundity in the MoBa study with different cutpoints*

	TTP > 12 ($n = 910$)			TTP > 3 ($n = 910$)			TTP > 24 ($n = 910$)		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Intercept	−7.2631	1.0236	<0.0001*	−5.1018	1.0265	<0.0001*	−9.0628	1.2646	<0.0001*
PFOA	0.2303	0.0652	0.0004*	0.2646	0.0694	0.0001*	0.1114	0.0779	0.1527
BMI	0.0563	0.0153	0.0002*	0.0418	0.0158	0.0081*	0.0402	0.0182	0.0271*
MotherAge	0.7001	0.1078	<0.0001*	0.5182	0.1068	<0.0001*	0.9220	0.1404	<0.0001*
MotherEdu	−0.1428	0.0853	0.0939	−0.2258	0.0864	0.0090*	−0.3258	0.1124	0.0037*
FatherAge	0.1293	0.0183	<0.0001*	0.1123	0.0188	<0.0001*	0.1421	0.0209	<0.0001*
SexFreq	−0.3372	0.0983	0.0006*	−0.3080	0.0996	0.0020*	−0.1915	0.1274	0.1328
Endo	2.4183	0.6341	0.0001*	2.8997	1.0290	0.0048*	2.2673	0.4307	<0.0001*

*Parameter estimate is significant at 5% level.
Subfecundity = I (TTP > given month).

Table 5. *Cox regression analyses for covariate effects on TTP in the MoBa study*

	ODS design ($n_0 = 520,\ n_3 = 390$)			Naive design ($n_v = n_0 + n_3 = 910$)			SRS design ($n_0 = 520$)		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
PFOA	−0.0567	0.0253	0.0251*	−0.0479	0.0331	0.1473	−0.0633	0.0419	0.1304
BMI	−0.0148	0.0056	0.0090*	−0.0222	0.0063	0.0004*	−0.0110	0.0091	0.2282
MotherAge	−0.2155	0.0342	<0.0001*	−0.4066	0.0467	<0.0001*	−0.1662	0.0674	0.0137*
MotherEdu	0.0968	0.0333	0.0037*	0.1179	0.0396	0.0029*	0.0730	0.0523	0.1630
FatherAge	−0.0470	0.0059	<0.0001*	−0.0801	0.0087	<0.0001*	−0.0415	0.0117	0.0004*
SexFreq	0.1089	0.0396	0.0060*	0.1133	0.0488	0.0203*	0.0968	0.0649	0.1359
Endo	−1.2153	0.2906	<0.0001*	−0.7830	0.1879	<0.0001*	−0.8424	0.5027	0.0938

*Parameter estimate is significant at 5% level.

Proposed ODS design and analysis. Using the 539 women sampled randomly as the SRS portion and 400 women sampled additionally from those with TTP > 12 months as a supplemental sample, we implemented our EMSELE method adjusted for all potential confounders. Due to missing values of covariates, the sample sizes of SRS and supplemental samples for the final fitted model were 520 and 390, respectively. The result of the final fitted model is listed in the first column in Table 5.

We note that the estimate for PFOA is negative suggesting that PFOA level increases the risk of subfecundity (i.e. TTP > 12). Women with a higher PFOA level tend to have a longer TTP, and per unit increment in PFOA, the risk of subfecundity increases with hazard ratio $e^{0.0567} = 1.06$. Unsurprisingly, older mothers and fathers are more likely to have a longer TTP. Women who had endometriosis before pregnancy have a higher risk, hazard ratio $e^{1.2153} = 3.37$. One advantage of our proposed method is that, given covariates, we can predict the risk probability of TTP > T_0 for any TTP time T_0. In contrast, only one risk probability (i.e. only for TTP > 12) in the logistic method.

The analysis results using only the SRS portion of the ODS sample and treating the ODS sample as a SRS (Naive) are given in columns 3 and 2 of Table 5, respectively. Among the three estimation methods compared, the estimators are consistent and our proposed method is the most efficient one. However, the

significant impact of PFOA level shows in neither Naive Design nor SRS Design. The proposed estimator is unbiased and more efficient.

Following another standard epidemiologic approach, we used the discrete-time analog of the Cox model to estimate the fecundability odds ratio (FOR). The analyses were adjusted by the same confounders used in Tables 4 and 5. The resulted FOR is 0.83 with SE $= 0.04$, 95% CI $= (0.75, 0.91)$ and $p = 0.0001$. Note the FOR reflects the fecundability odds ratio for 1 month, which is different from the hazard ratio from the Cox model.

In summary, the proposed method provides an efficient and consistent estimate, utilizes fully the available survival data and takes advantage of the nature of the ODS design. It does not have the inconsistent issue of the existing odds ratio approach used by epidemiologists.

5. DISCUSSION

In this paper, motivated by the need to assess the relationship between the PFASs on women's subfecundability on our study of the MoBa study, we designed a general ODS sampling scheme for survival studies with a failure-time outcome. To reap in the benefit of such a survival ODS design, we developed a new inferential method and provide an EMSELE for the parameters of primary interest. Our proposed ODS method is an improvement over the current odds ratio approach used in epidemiology as well as improvement over the case-cohort and the generalized case-cohort designs. Because we allow the sample selection of cases to depend on the timing of disease endpoints, i.e. by oversampling subjects from the most informative regions, the proposed ODS design for survival data can enhance study efficiency and reduce study cost.

Simulation studies suggest that the small-sample performance of the proposed method approximates the asymptotic properties well. Our proposed estimator is the most efficient one among the four competing estimators: the maximum likelihood estimator based on the Cox's likelihood from a sample random sample of same size as the ODS sample, the estimator under case-cohort design developed by Prentice (1986), and the weighted estimator under generalized case-cohort design developed by Kang and Cai (2009). The efficiency gain shows that our proposed method is a cost-efficient alternative to case-cohort and generalized case-cohort designs.

A few comments on the behavior of the proposed design/estimator and some cautionary points on practical applications of the proposed method are presented. First, we note that the proposed ODS design and our estimator is more efficient than the SRS design when censoring rate is high. This is due to the fact that at high censoring cases, the SRS design will have substantial fewer failures than the ODS design. This suggests that ODS design is particularly useful when censoring is high. Secondly, we caution users that when censoring is extremely high (e.g. at 90% and 95%), our variance estimator based on the asymptotic properties could overestimate the true variance. One would need to either increase the sample size or employ some alternative variance estimator, such as the bootstrap estimator, in the small-sample situations. Thirdly, our estimator from (2.6) is based on assuming censoring is independent of covariates. Biases in effect estimation could result if this is violated (Table 2). A good practice is to check this assumption in real data analysis using the SRS sample. Finally, we estimated $\Lambda_0(t)$ with a simple consistent estimator from the SRS sample. Alternative estimators that use more available data could be used.

Application of the proposed ODS method to analyze the MoBa data suggested that women with higher PFOA levels tend to have a longer TTP. On the other hand, the default epidemiologic approach for the odds ratio under different TTP cutpoints yields inconsistent results. Comparing with competing designs, our proposed ODS method provides a feasible design and efficient estimates. Future study includes developing models and estimation procedures appropriate for studies with multiple disease outcomes. In some studies, researchers may need to consider several diseases or several subtypes of disease (Lu and Shih,

2006; Kang and Cai, 2009). For example, in the Busselton Health Study (Cullen, 1972), it was of interest to study the relationship between serum ferritin and coronary heart disease and stroke events.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We are grateful to all the participating families in Norway who take part in this ongoing cohort study. *Conflict of Interest*: None declared.

Funding

This research was supported in part by U.S. National Institute of Health grants (R01 ES021900, UL1 RR025747, P01 CA142538), and the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences. The Norwegian Mother and Child Cohort Study is supported by the Norwegian Ministry of Health and the Ministry of Education and Research, NIH/NIEHS (N01-ES-75558), NIH/NINDS (UO1 NS 047537-01, UO1 NS 047537-06A1), and the Norwegian Research Council/ FUGE (151918/S10). This research is also funded in part by National Science Foundation of China (11101314 to J.D. and 11171263, 11371299 to Y.L.).

References

Alexander, B. H., Olsen, G. W., Burris, J. M., Mandel, J. H. and Mandel, J. S. (2003). Mortality of employees of a perfluorooctanesulphonyl fluoride manufacturing facility. *Occupational and Environmental Medicine* **60**, 722–729.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *The Scandinavian Journal of Statistics* **34**, 86–102.

Cai, J. and Zeng, D. (2007). Power calculation for case-cohort studies with nonrare events. *Biometrics* **63**, 1288–1295.

Chatterjee, N., Chen, Y. H. and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* **98**, 158–168.

Chen, K. (2001). Generalized case-cohort sampling. *Journal of the Royal Statistical Society, Series B* **63**, 791–809.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of lung, breast, and cervix. *Journal of National Cancer Institute* **11**, 1269–1275.

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

Cullen, K. J. (1972). Mass health examinations in the busselton population, 1966 to 1970. *The Medical Journal of Australia* **2**, 714–718.

Fei, C. Y., McLaughlin, J. K., Lipworth, L. and Olsen, J. (2009). Maternal levels of perfluorinated chemicals and subfecundity. *Human Reproduction* **24**, 1200–1205.

Johansen, S. (1983). An extension of Cox's regression Model. *International Statistical Review* **51**, 165–174.

Johansson, N., Fredriksson, A., Eriksson, P. (2008). Neonatal exposure to perfluorooctane sulfonate (PFOS) and perfluorooctanoic acid (PFOA) causes neurobehavioural defects in adult mice. *Neurotoxicology* **29**, 160–169.

KANG, S. AND CAI, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika* **96**, 887–901.

LU, S. AND SHIH, J. H. (2006). Case-cohort designs and analysis for clustered failure time data. *Biometrics* **62**, 1138–1148.

LU, W. AND TSIATIS, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika* **93**, 207–214.

OWEN, A. B. (1990). Empirical likelihood for confidence regions. *The Annals of Statistics* **18**, 90–120.

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

SONG, R., ZHOU, H. AND KOSOROK, M. R. (2009). On semiparametric efficient inference for two-stage outcome dependent sampling with a continuous outcome. *Biometrika* **96**, 221–228.

SUN, J., SUN, L. AND FLOURNOYA, N. (2004). Additive hazards model for competing risks analysis of the case-cohort design. *Communications in Statistics—Theory and Methods* **33**, 351–366.

TSAI, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96**(3), 601–615.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* **10**, 616–620.

VARDI, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13**, 178–203.

WEAVER, M. A. AND ZHOU, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of The American Statistical Association* **100**, 459–469.

WEINBERG, C. R. AND WACHOLDER, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80**, 461–465.

WHITTEMORE, A. S. (1997). Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society, Series B* **59**, 589–602.

WHITWORTH, K. W., HAUG, L. S., BAIRD, D. D., BECHER, G., HOPPIN, J. A., SKJAERVEN, R., THOMSEN, C., EGGESBO, M., TRAVLOS, G., WILSON, R. AND LONGNECKER, M. P. (2012). Perfluorinated compounds and subfecundity in pregnant women. *Epidemiology* **23**, 257–263.

ZHOU, H., SONG, R. AND QIN, J. (2011). Statistical inference for a two-stage outcome dependent sampling design with a continuous outcome. *Biometrics* **67**, 194–202.

ZHOU, H., WEAVER, M., QIN, J., LONGNECKER, M. AND WANG, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome dependent sampling scheme with a continuous outcome. *Biometrics* **58**, 413–421.

ZHOU, H., WU, Y., LIU, Y. AND CAI, J. (2011). Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics* **12**, 521–534.

[*Received September 5, 2013; revised February 27, 2014; accepted for publication March 21, 2014*]