Statistica Sinica **27** (2017), 415-435
doi:http://dx.doi.org/10.5705/ss.202015.0294

END-POINT SAMPLING

Yuan Yao, Wen Yu and Kani Chen

Hong Kong Baptist University, Fudan University and Hong Kong University of Science and Technology

Abstract: Retrospective sampling designs, including case-cohort and case-control designs, are commonly used for failure time data in the presence of censoring. In this paper, we propose a new retrospective sampling design, called end-point sampling, which improves the efficiency of the case-cohort and case-control designs. The regression analysis is conducted using the Cox model. Under different assumptions, the maximum likelihood approach with computational aid from the EM algorithm, and the inverse probability weighting approach are developed respectively to estimate the regression parameters. The resulting estimators are shown to be consistent and asymptotically normal. Simulation and a real data study show favorable evidence for the proposed design in comparison with existing ones.

Key words and phrases: Case-control sampling, Cox model, EM algorithm, inverse probability weighting, maximum likelihood estimation, retrospective sampling.

1. Introduction

In many cohort studies, the failure or censoring times of all subjects, as well as the covariates for the cases, are observed in the first stage. In the second stage, one needs to determine the subjects to choose or sample for covariate ascertainment, which can be rather costly. Case-cohort, case-control, nested case-control designs or, more generally, the generalized case-cohort designs are common types of two-phase designs with censored data. Compared with prospective sampling designs, such retrospective sampling designs can save cost and time, and are particularly useful for studies of rare diseases. In a case-cohort design, all cases and a certain number of subcohort are sampled for covariate ascertainment; in a case-control design, all cases and a certain number of controls are sampled; and in a nested case-control design all cases and a fixed size of controls from each risk set at failure times are sampled. More general designs and analysis may be seen in Chen (2001), Nan (2004), Lu and Tsiatis (2006), Kang and Cai (2009), Kong and Cai (2009), Liu et al. (2010), Zeng and Lin (2014), and Yao (2015), among many others.

The problem is what type of design one should choose in order to maximize the information about the dependence of failure times on covariates at a

given size of sample with covariate ascertainment. In view of the importance of retrospective sampling, it is interesting to explore two-phase cohort designs for more efficient sampling for covariate ascertainment. We say design A can be viewed as more efficient than design B if the two designs have the same size of samples with covariate ascertainment, and the data based on design A contain more information about the dependence of the failure time on covariates, which can be characterized by the regression parameters. In this paper, we propose an end-point sampling design that aims at such efficiency and is demonstrated to be more efficient than the existing designs. The new design takes cases and controls with large censoring times for covariate ascertainment. In other words, among the censored individuals, we sample those with larger censoring times by higher probabilities. The motivation for it is that when censoring times are independent of the covariates and failure times, the larger censoring times tend to contain the most information about the regression parameters. In a cohort, the observed (censored) failure times are in distribution smaller than the actual failure times because of the nature of right censoring, so additional subjects with potentially large failure times should provide more additional information than those with small failure times. Moreover, failure times censored by large censoring times also tend to take values in smaller range than those censored by small censoring times.

The idea of the end-point sampling closely resembles that of efficient designs of covariates/inputs in simple linear regression, in which the variance of the least squares estimate of the slope is inversely proportional to the sample variance of the covariates. As a result, for a given sample size, a design would be more efficient if the covariates of the samples more spread out. The idea of the end-point sampling is precisely the same: design the sampling scheme so that the failure times, observed or unobserved, with covariate ascertainment spread out as much as possible.

The regression analysis is conducted under the Cox proportional hazards model (Cox (1972)). Let T be the failure time of interest and Z a p-dimensional covariate. The Cox model assumes that the conditional hazard of T given Z satisfies

$$\lambda_{T|Z}(t|Z) = \lambda(t)\exp(\beta^\top Z), \tag{1.1}$$

where $\lambda(t)$ is the baseline hazard function and β is the p-dimensional regression parameter of interest. Under (1.1), the partial likelihood for nested case-control and the pseudolikelihood for case-cohort designs are among the classical methods; see Thomas (1977) and Prentice (1986). Their estimates and inferences are straightforward, but they are not semiparametrically efficient. Various attempts have been made to improve the efficiency of the estimation methods; see, for example, Chen and Lo (1999), Nan (2004) and Chen (2004), among others.

More recently, Zeng and Lin (2014) and Yao (2015) proposed likelihood-based procedures under linear transformation models which include the Cox model as a special case. The resulting estimators are shown to be semiparametrically efficient. In this paper, under the assumption that censoring time is independent of all other variables, we develop a likelihood-based inference under the proposed sampling design, leading to semiparametrically efficient estimation. For the more conventional assumption that censoring time is conditionally independent of failure time given covariates, an inverse probability weighting approach is proposed to do estimation.

Case-control and case-cohort sampling do not aim at efficient sampling, but aim at reducing bias and produce relatively easy estimation methods, such as partial likelihood (Thomas (1977)) and pseudo-likelihood (Prentice (1986)). We caution that the efficiency of the proposed sampling design is different from the efficient utilization of samples for given data arising from a sampling design, to obtain better estimation; see, for example, Chen and Lo (1999), Kulich and Lin (2004), Breslow et al. (2009), Chen and Zucker (2009), Kim, Cai and Lu (2013), etc. Empirically, we find that, using the same estimation method, the proposed end-point sampling design is more efficient than case-cohort or case-control sampling with comparable number of controls.

The remainder of the paper is organized as follows. Section 2 presents the description of the proposed end-point sampling under independent censoring assumption. A nonparametric maximum likelihood approach is developed for estimation and justification of the asymptotic properties is provided. To overcome possible computational difficulties, we use a semiparametric expectation-maximization (EM) algorithm to obtain the maximum likelihood estimator (MLE). The MLE for the regression parameter is shown to be consistent, asymptotically normal, and it reaches the semiparametric efficiency bound. In Section 3, a slightly different end-point design is given under the conditionally independent censoring assumption, and an inverse probability weighting approach for estimation with related large sample properties is provided. Simulation studies and a data example are presented in Section 4 and 5, respectively. Some discussion is given in Section 6.

2. End-point Sampling Design with Maximum Likelihood Estimation

2.1. The proposed design and likelihood construction

Let C be the right censoring time. In this section, we assume that C is independent of (T, Z). Denote the observed event time by $Y = \min\{T, C\}$ and the censoring indicator by $\delta = I\{T \leqslant C\}$, where $I\{\cdot\}$ is the indicator function. The full cohort will be n independent and identically distributed (i.i.d.) copies of (Y, δ, Z), denoted by (Y_i, δ_i, Z_i), $i = 1, \ldots, n$. In a case-cohort sampling design,

the covariates of all the failures, the $n_1 = \sum_{i=1}^{n} \delta_i$ individuals with $\delta_i = 1$, and a simple random sample from the full cohort, known as the subcohort, are observed. In a case-control sampling design, we observe the covariates of all n_1 failures and a simple random sample of size n_2 from all the controls, the $n - n_1$ individuals with $\delta_i = 0$. For the ith individual, let Δ_i be the sampling indicator variable with 1 for covariates being observed and 0 otherwise. Then for the scenario of case-control sampling design, $\sum_{i=1}^{n} \Delta_i = n_1 + n_2$. In both sampling designs, the observed data is (Y_i, δ_i, Z_i) for $\Delta_i = 1$ and (Y_i, δ_i) for $\Delta_i = 0$, $i = 1, \ldots, n$. An important feature of the designs is that the distribution of $(\Delta_1, \ldots, \Delta_n)$ depends only on (Y_i, δ_i), $i = 1, \ldots, n$.

In the proposed end-point sampling design, we still collect the covariates information for all the failures. However, here the n_2 controls are drawn according to the values of the responses Y, say the individuals with largest n_2 censoring times are selected as the controls, and the corresponding covariates information of these controls are collected. Under the proposed design, the distribution of $(\Delta_1, \ldots, \Delta_n)$ still depends only on (Y_i, δ_i), $i = 1, \ldots, n$. The observed data is represented by the same notation as that used for case-cohort and case-control sampling.

Let f and F be the density and distribution of Z, respectively, and $\Lambda(t) = \int_0^t \lambda(u)du$. By using arguments similar to those in Yao (2015), the log-likelihood of the observed data is

$$
\begin{aligned}
&l(\beta, \Lambda, F) \\
&\quad = \sum_{i=1}^{n} \Delta_i \left[\delta_i \{\log \lambda(Y_i) + \beta^\top Z_i\} - \Lambda(Y_i) \exp(\beta^\top Z_i) + \log f(Z_i) \right] \\
&\qquad + \sum_{i=1}^{n} (1 - \Delta_i) \log \left[\int_{\mathcal{Z}} \left\{ \lambda(Y_i) \exp(\beta^\top z) \right\}^{\delta_i} \exp\left\{ -\Lambda(Y_i) \exp(\beta^\top z) \right\} f(z) dz \right] \\
&\qquad + \sum_{i=1}^{n} \log \left[\lambda_C(Y_i)^{1-\delta_i} \exp\{-\Lambda_C(Y_i)\} \right] \\
&\qquad + \log f_\Delta(\Delta_1, \ldots, \Delta_n | (Y_j, \delta_j), 1 \leqslant j \leqslant n), \qquad (2.1)
\end{aligned}
$$

where λ_c and Λ_C are the hazard and cumulative hazard of C, $\mathcal{Z}$ is the support of Z and f_Δ is the joint probability mass for $\Delta_1, \ldots, \Delta_n$. The observed likelihood takes the same form for case-cohort, case-control and the proposed end-point sampling design. The kernel part of the likelihood is the summation of the first two terms on the right-hand-side of (2.1), denoted by $l_k(\beta, \Lambda, F)$.

Here the maximum of l_k does not exist since Λ and f are infinite dimensional parameters. Thus, we restrict Λ to be a step function with jumps only at the Y_i's, and F to be a discrete distribution with mass only on observed Z_i's. Define

$\mathcal{O} = \{i : \Delta_i = 1\}$ and $\overline{\mathcal{O}} = \{i : \Delta_i = 0\}$. Let λ_i be the jump size of Λ at Y_i, $i = 1, \ldots, n$, and p_i be the probability mass at Z_i for $i \in \mathcal{O}$. Let $\vartheta = (\beta, \lambda_1, \ldots, \lambda_n, p_i, i \in \mathcal{O})$. Then we have

$$l_k(\vartheta) = \sum_{i\in\mathcal{O}} \Big[\delta_i \{\log \lambda_i + \beta^\top Z_i\} - \sum_{Y_j \leqslant Y_i} \lambda_j \exp(\beta^\top Z_i) + \log p_i \Big] + \sum_{i\in\overline{\mathcal{O}}} \log \Big[\sum_{j\in\mathcal{O}} p_j \Big\{ \lambda_i \exp(\beta^\top Z_j) \Big\}^{\delta_i} \exp\Big\{ - \sum_{Y_l \leqslant Y_i} \lambda_l \exp(\beta^\top Z_j) \Big\} \Big].$$

We maximize $l_k(\vartheta)$ with respect to ϑ subject to $\lambda_i \geqslant 0$, $i = 1, \ldots, n$, $p_i \geqslant 0$, $i \in \mathcal{O}$ and $\sum_{i\in\mathcal{O}} p_i = 1$. Denote the maximizer by $\hat{\vartheta}_{\mathsf{M}} = (\hat{\beta}_{\mathsf{M}}, \hat{\lambda}_1, \ldots, \hat{\lambda}_n, \hat{p}_i, i \in \mathcal{O})$. Write $\hat{\Lambda}_{\mathsf{M}}(t) = \sum_{Y_i \leqslant t} \hat{\lambda}_i$ and $\hat{F}_{\mathsf{M}}(z) = \sum_{Z_i \leqslant z, i\in\mathcal{O}} \hat{p}_i$. The MLE of β, Λ and F are $\hat{\beta}_{\mathsf{M}}$, $\hat{\Lambda}_{\mathsf{M}}$ and $\hat{F}_{\mathsf{M}}$, respectively.

2.2. Algorithm

Maximization of $l_k(\vartheta)$ can be difficult because the dimension of ϑ increases as the sample size increases. Since one can view Z_i with $\Delta_i = 0$ as missing data, it is quite natural to adopt an EM algorithm to calculate the MLE. Let z_i be the observed value of Z_i, $i \in \mathcal{O}$. Then the logarithm of the complete data likelihood is proportional to

$$l_C(\vartheta) = \sum_{i\in\mathcal{O}} \Big\{ \delta_i (\log \lambda_i + \beta^\top Z_i) - \sum_{Y_j \leqslant Y_i} \lambda_j \exp(\beta^\top Z_i) + \log p_i \Big\} + \sum_{i\in\overline{\mathcal{O}}} \sum_{j\in\mathcal{O}} I\{Z_i = z_j\} \Big[\Big\{ - \sum_{Y_k \leqslant Y_i} \lambda_k \exp(\beta^\top z_j) \Big\} + \log p_j \Big].$$

Our algorithm is similar to that devised in Zeng and Lin (2014).

1. *E-step*: Let $\vartheta^{(0)}$ be the starting value for ϑ. The conditional expectation of $l_C(\vartheta)$ at $\vartheta = \vartheta^{(0)}$ given the observations is

$$\sum_{i\in\mathcal{O}} \Big\{ \delta_i \Big(\log \lambda_i + \beta^\top Z_i \Big) - \sum_{Y_j \leqslant Y_i} \lambda_j \exp(\beta^\top Z_i) + \log p_i \Big\} + \sum_{i\in\overline{\mathcal{O}}} \sum_{j\in\mathcal{O}} w_{ij}(\vartheta^{(0)}) \Big[\Big\{ - \sum_{Y_k \leqslant Y_i} \lambda_k \exp(\beta^\top z_j) \Big\} + \log p_j \Big],$$

where

$$w_{ij}(\vartheta) = E_\vartheta \left(I\{Z_i = z_j\} | Y_i, \delta_i = 0 \right) = \frac{\exp\Big\{ - \sum_{Y_k \leqslant Y_i} \lambda_k \exp(\beta^\top z_j) \Big\} p_j}{\sum_{l\in\mathcal{O}} \exp\Big\{ - \sum_{Y_k \leqslant Y_i} \lambda_k \exp(\beta^\top z_l) \Big\} p_l}.$$

2. *M-step*: We update p_i's by maximizing $\sum_{i\in\mathcal{O}} \log p_i + \sum_{i\in\overline{\mathcal{O}}} \sum_{j\in\mathcal{O}} w_{ij}(\vartheta^{(0)}) \log p_j$ with respect to p_i's, subject to $\sum_{i\in\mathcal{O}} p_i = 1$. This gives the explicit solution

$$p_i^{(1)} = \frac{1 + \sum_{j\in\overline{\mathcal{O}}} w_{ji}(\vartheta^{(0)})}{n_1 + n_2 + \sum_{i\in\mathcal{O}} \sum_{j\in\overline{\mathcal{O}}} w_{ji}(\vartheta^{(0)})}.$$

The values of β and λ_i's are updated by maximizing

$$\sum_{i\in\mathcal{O}} \Big\{ \delta_i \Big(\log\lambda_i + \beta^\top Z_i \Big) - \sum_{Y_j \leqslant Y_i} \lambda_j \exp(\beta^\top Z_i) \Big\} + \sum_{i\in\overline{\mathcal{O}}} \sum_{j\in\mathcal{O}} w_{ij}(\vartheta^{(0)}) \Big\{ - \sum_{Y_k \leqslant Y_i} \lambda_k \exp(\beta^\top z_j) \Big\} \tag{2.2}$$

with respect to λ_i's and β. For fixed β, the λ_i's can be profiled out by setting

$$\lambda_i = \frac{\delta_i}{\sum_{j\in\mathcal{O}} I\{Y_j \geqslant Y_i\} \exp(\beta^\top Z_j) + \sum_{j\in\overline{\mathcal{O}}} I\{Y_j \geqslant Y_i\} \sum_{k\in\mathcal{O}} w_{jk}(\vartheta^{(0)}) \exp(\beta^\top z_k)}. \tag{2.3}$$

Plugging this into (2.2) results in maximizing

$$\sum_{i=1}^{n} \delta_i \Big[\beta^\top Z_i - \log \Big\{ \sum_{j\in\mathcal{O}} I\{Y_j \geqslant Y_i\} \exp(\beta^\top Z_j) + \sum_{j\in\overline{\mathcal{O}}} I\{Y_j \geqslant Y_i\} \sum_{k\in\mathcal{O}} w_{jk}(\vartheta^{(0)}) \exp(\beta^\top z_k) \Big\} \Big]$$

with respect to β. The objective function can be viewed as a weighted log-partial likelihood and is concave in β. Thus the maximization is easy to implement. Denote the maximizer by $\beta^{(1)}$. Replacing β in (2.3) by $\beta^{(1)}$ we obtain $\lambda_i^{(1)}$. The updated value of whole parameter is given by $\vartheta^{(1)} = (\beta^{(1)}, \lambda_1^{(1)}, \ldots, \lambda_n^{(1)}, p_i^{(1)}, i \in \mathcal{O})$.

The E-step is repeated using $\vartheta^{(1)}$, and the iteration continues until convergence. The resulting parameter values are treated as the MLE.

2.3. Large sample properties

The MLE has the expected large sample properties, under suitable regularity conditions, of consistency and asymptotic normality. Let β_0, Λ_0, and F_0 be the true value of β, Λ and F, respectively, $\|\cdot\|$ the Euclidean norm, τ the duration of the cohort study. Assume that $\mathcal{Z}$ is a bounded set.

Theorem 1. *Under Conditions* 1–3 *in the Appendix,* $\|\hat\beta_{\mathsf{M}} - \beta_0\| \to 0$, $\sup_{t\in[0,\tau]} |\hat\Lambda_{\mathsf{M}}(t) - \Lambda_0(t)| \to 0$ *and* $\sup_{z\in\mathcal{Z}} |\hat F_{\mathsf{M}}(z) - F_0(z)| \to 0$ *with probability one.*

Let sets of functions be $\mathcal{Q}_1 = \{l \in BV[0,\tau] : |l| \leq 1\}$ and $\mathcal{Q}_2 = \{l \in BV[\mathcal{Z}] : |l| \leqslant 1\}$, where $BV[D]$ is the set of functions on D with bounded total variation.

Theorem 2. *Under Conditions 1-4 in the Appendix,* $\sqrt{n}(\hat{\beta}_{\mathsf{M}} - \beta_0, \hat{\Lambda}_{\mathsf{M}} - \Lambda_0, \hat{F}_{\mathsf{M}} - F_0)$ *converges weakly to a zero-mean Gaussian process in* $\mathbb{R}^p \times l^\infty(\mathcal{Q}_1) \times l^\infty(\mathcal{Q}_2)$. *Moreover, the limiting covariance of* $\sqrt{n}(\hat{\beta}_{\mathsf{M}} - \beta_0)$ *attains the semiparametric efficiency bound.*

The regularity conditions and proofs of the two theorems are given in the Appendix. The limiting variances and covariances of the estimator can be estimated by inverting the observed information matrix derived from the log likelihood. Alternatively, one can use the bootstrap method to get variance estimates.

Remark 1. Assume the true value of the baseline cumulative hazard, Λ_0, is known. If one could observe the full cohort and fit the Cox model using maximum likelihood, the score function of β for the i-th individual, evaluated at the true value, takes the form $S_i(\beta_0) = \delta_i Z_i - Z_i \exp(\beta_0^\top Z_i)\Lambda_0(Y_i)$. As to the observed likelihood (2.1), the score function of β, evaluated at the true value, is

$$\frac{\partial l(\beta, \Lambda, F)}{\partial \beta}\Big|_{\beta=\beta_0} = \sum_{i=1}^{n} \{\Delta_i S_i(\beta_0) + (1-\Delta_i) E\left(S_i(\beta_0)|Y_i, \delta_i\right)\}.$$

Then the score function for the i-th individual is given by $S_i^R(\beta_0) = \Delta_i S_i(\beta_0) + (1-\Delta_i) E\left(S_i(\beta_0)|Y_i, \delta_i\right)$. A careful calculation yields that

$$E\left\{S_i^R(\beta_0)^{\otimes 2}\right\} = E\left\{S_i(\beta_0)^{\otimes 2}\right\} - E\left(I\{\Delta_i = 0\}\Lambda_0^2(Y_i) E\left[\left\{Z_i \exp(\beta_0^\top Z_i)\right\}^{\otimes 2} | Y_i, \delta_i = 0\right]\right),$$

where, for a column vector a, $a^{\otimes 2} = aa^\top$. If $\beta_0 = 0$, then Z_i is independent of (Y_i, δ_i), and $E[\{Z_i \exp(\beta_0^\top Z_i)\}^{\otimes 2} | Y_i, \delta_i = 0] = E(Z_i^{\otimes 2})$. Consequently,

$$E\left\{S_i^R(\beta_0)^{\otimes 2}\right\} = E\left\{S_i(\beta_0)^{\otimes 2}\right\} - E\left(Z_i^{\otimes 2}\right) E\left\{I\{\Delta_i = 0\}\Lambda_0^2(Y_i)\right\}. \quad (2.4)$$

Intuitively, to increase the information of i-th individual, one should decrease the second term on the right-hand-side of (2.4). Since Λ_0 is a non-decreasing function. one should set $\Delta_i = 0$ (not sampled) when Y_i is smaller.

Remark 2. The proposed end-point sampling design can be viewed as a special case of stratified case-cohort design. In most existing literature studying stratified case-cohort design, a stratum is decided by certain observed covariates, while in this design a stratum is decided by the observed failure time. More generally, this design can also be viewed as a special case of the class of generalized case-cohort design discussed in Chen (2001).

3. End-point Sampling Design with Inverse Probability Weighting

3.1. The proposed design and estimation approach

In survival analysis, a more conventional and realistic assumption is that C is conditionally independent of T given Z. When this is the case, we propose a slightly different end-point sampling design and develop an inverse probability weighting approach for estimating the regression parameters in (1.1). In this design, the controls for covariate ascertainment are sampled independently with a positive selection probability that is an increasing function of the observed time. Specifically, for subject i, let $\pi_i = \mathsf{P}(\Delta_i = 1|(Y_j, \delta_j), j = 1, \ldots, n)$. Here we set $\pi_i = \delta_i + (1 - \delta_i)p(Y_i)$, where $p(y) \in (0, 1)$ is a predetermined function increasing in y, and take the Δ_i's to be independent of each other.

Under this sampling with probabilities design, the inverse probability weighting approach is used to obtain the estimating equation for estimating β; cf., Chen and Lo (1999), Lu and Tsiatis (2006), Kong and Cai (2009), etc. For our case, the inverse probability weighted estimating equation is

$$S_{\mathsf{IPW}}(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} (\Delta_j/\pi_j) Z_j \exp(\beta^\top Z_j) I\{Y_j \geqslant Y_i\}}{\sum_{j=1}^{n} (\Delta_j/\pi_j) \exp(\beta^\top Z_j) I\{Y_j \geqslant Y_i\}} \right\} = 0.$$

Denote the solution of the above estimation equation by $\hat{\beta}_{\mathsf{I}}$, treated as the inverse probability weighted estimator of β. It is easy to see that $\hat{\beta}_{\mathsf{I}}$ can be obtained by maximizing

$$\sum_{i=1}^{n} \delta_i \Big[\beta^\top Z_i - \log \Big\{ \sum_{j=1}^{n} \frac{\Delta_j}{\pi_j} \exp\left(\beta^\top Z_j\right) I\{Y_j \geqslant Y_i\} \Big\} \Big]$$

with respect to β. This objective function is concave in β, so the maximization is easy to implement by standard software packages.

3.2. Large sample properties

It is expected that under suitable regularity conditions, the inverse probability weighted estimator $\hat{\beta}_{\mathsf{I}}$ is asymptotically normally distributed. To give specific results, we need some notation. Let $\mu(t) = E[Z \exp(\beta_0^\top Z) I\{Y \geqslant t\}]/E[\exp(\beta_0^\top Z) I\{Y \geqslant t\}]$, $N(t) = \delta I\{Y \leqslant t\}$ and $\pi = \delta + (1 - \delta)p(Y)$. Take

$$\Sigma_1 = E\left[\int_0^\infty \{Z - \mu(t)\}^{\otimes 2} \, dN(t) \right],$$

$$\Sigma_2 = E\left(\frac{1 - \pi}{\pi} \left[\int_0^Y \exp(\beta_0^\top Z) \{Z - \mu(t)\} \lambda_0(t) dt \right]^{\otimes 2} \right),$$

where λ_0 is the true value of λ.

Theorem 3. *Under Conditions* 1–4 *in the Appendix,* $\sqrt{n}(\hat{\beta}_{\mathsf{I}} - \beta_0)$ *converges weakly to a zero-mean normal vector with variance-covariance matrix* $\Sigma_1^{-1}(\Sigma_1 + \Sigma_2)\Sigma_1^{-1}$.

The proof is given in the Appendix. We estimate Σ_1 by

$$\hat{\Sigma}_1 = \frac{1}{n}\sum_{i=1}^{n} \delta_i \{Z_i - \hat{\mu}(Y_i)\}^{\otimes 2},$$

where $\hat{\mu}(t) = \sum_{j=1}^{n}(\Delta_j/\pi_j)Z_j \exp(\hat{\beta}_{\mathsf{I}}^{\top} Z_j) I\{Y_j \geqslant t\} / \sum_{j=1}^{n}(\Delta_j/\pi_j)\exp(\hat{\beta}_{\mathsf{I}}^{\top} Z_j) I\{Y_j \geqslant t\}$, and estimate Σ_2 by

$$\hat{\Sigma}_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-\pi_i)\Delta_i}{\pi_i^2} \left[\int_0^{Y_i} \exp(\hat{\beta}_{\mathsf{I}}^{\top} Z_i)\{Z_i - \hat{\mu}(t)\}\, d\hat{\Lambda}_{\mathsf{I}}(t)\right]^{\otimes 2},$$

where $d\hat{\Lambda}_{\mathsf{I}}(t) = \sum_{j=1}^{n} dN_j(t) / \sum_{j=1}^{n}(\Delta_j/\pi_j)\exp(\hat{\beta}_{\mathsf{I}}^{\top} Z_i) I\{Y_j \geqslant t\}$, with $N_i(t) = \delta_i I\{Y_i \leqslant t\}$, $i = 1, \ldots, n$.

Remark 3. When $\beta_0 = 0$, (Y, δ) is independent of Z. Then $\Sigma_2 = E\{(1-\pi)\pi^{-1}[\{Z - E(Z)\}\Lambda_0(Y)]^{\otimes 2}\}$. Since $\Lambda_0(y)$ is increasing in y, to decrease Σ_2 requires that $(1-\pi)\pi^{-1}$ to be decreasing so that $p(y)$ be an increasing function of y. However, since $p(y)$ is increasing, it is possible that $(1-\pi)/\pi$ might be very large when $y \to 0$. Still, $[\int_0^Y \exp(\beta_0^{\top} Z)\{Z - \mu(t)\}\lambda_0(t)dt]^{\otimes 2}$ is no larger than $O(Y^2)$, and if we choose $p(y)$ to be a linear or quadratic function, Σ_2 is finite.

Remark 4. For the sampling with probabilities design if $p(y)$ is strictly positive, the inverse probability weighting approach does not require the independence between censoring time and covariates. However, it is not easy to specify a sampling probability that leads to superior design. The end-point sampling design proposed in Section 2 can be viewed as a special case of the sampling with probabilities design: the sampling probability is 1 for the cases and the large censored observations, and 0 for small ones, but the inverse probability weighting method breaks down because of the existence of zero selection probability.

4. Simulation Studies

In this section, we report on some simulation studies to compare the estimation results under the proposed sampling designs with those under some existing retrospective designs.

4.1. Maximum likelihood approach

For model (1.1), we chose the dimension of Z to be 2. Two model setups were considered. In the first setup, Z_1 and Z_2 were generated from the Binomial

distribution with success probability 0.5 and uniform distribution on $(0, 1)$, respectively. The two covariates were independent of each other. We set $\beta_1 = 1$, $\beta_2 = -1$ and $\lambda(t) = 0.5$. The censoring time was $C = \min\{\tilde{C}, \tau\}$, where $\tilde{C}$ was exponential with mean 0.33 and independent of T, Z_1, and Z_2, and $\tau = 0.7$. The resulting censoring percentage was around 87%. In the second setup, the covariates were generated according to the same scheme. We set $\beta_1 = -1$, $\beta_2 = 0.5$, and $\lambda(t) = t$. A four-stage censoring scheme was considered: the first one-fourth of the individuals were censored at t_1, the second one-fourth at t_2, the third one-fourth at t_3, and the rest at t_4, where $t_1 = 0.1$, $t_2 = 0.5$, $t_3 = 0.9$, and $t_4 = 1.3$. The censoring percentage wa again about 87%.

The full cohort size n was set to be 2,000. We considered three retrospective sampling designs. The first one was case-cohort, where all the cases were sampled and a subcohort of size 235 was drawn by simple random sampling. With the censoring percentage at 87%, the subcohort contained about 200 censored individuals. The second one was case-control, where all the cases and a simple random sample of size 200 from the censored individuals were sampled. The third one was the proposed end-point sampling design, where all the cases and the 200 individuals with the largest censoring times from the censored individuals were sampled. Note that if more than 200 censored individuals are tied at the largest censoring time, we draw a simple random sample from the censored ones whose observed censoring times are the largest censoring time.

For the case-cohort design, the pseudolikelihood method proposed by Prentice (1986) and the maximum likelihood approach were applied to get estimators for β_1 and β_2. For the case-control design and the proposed design, the maximum likelihood approach was used. For the pseudolikelihood estimates, the standard errors were estimated by the plug-in method proposed by Chen and Lo (1999). To get the MLE, the EM algorithm was adopted. The initial values of β_1, β_2, and λ_i's were obtained by fitting the Cox model only using the individuals with covariates observed. The initial values of the p_i's were set to be equally distributed. The convergence criterion was set to be 10^{-2}. To implement the optimization in the M-step, we used the 'fminsearch' function in the 'Optimization toolbox' of MATLAB, which uses a simplex search method to find optimum. The variance estimates were obtained by inverting the observed information matrix at the MLE. We also calculated the maximum partial likelihood estimates based on the full cohort for comparison.

One thousand replicates were done. For each estimate, we report the average bias, the empirical standard error, the average of estimated standard error, and the coverage rate of 95% Wald-type confidence interval. To compare the efficiency of different designs, we used a quantity called efficiency gain per sample with covariates observed, denoted by EGPS. For each specific estimate, EGPS is

Table 1. Summarized simulation results for different retrospective designs with maximum likelihood estimation.

Setup	Method	Parameter	BIAS	SE	SEE	CP	RE
1	Full-cohort	β_1	−0.002	0.130	0.129	0.944	1.000
		β_2	0.001	0.198	0.209	0.963	1.000
	Case-cohort	β_1	−0.003	0.216	0.213	0.949	1.481
	Prentice	β_2	−0.021	0.364	0.377	0.957	1.207
	Case-cohort	β_1	−0.055	0.187	0.191	0.953	1.970
	MLE	β_2	0.042	0.293	0.308	0.961	1.866
	Case-control	β_1	−0.052	0.188	0.192	0.957	1.955
	MLE	β_2	0.046	0.312	0.307	0.936	1.645
	End-point	β_1	−0.021	0.154	0.149	0.944	2.940
	MLE	β_2	0.007	0.260	0.256	0.944	2.368
2	Full-cohort	β_1	0.003	0.143	0.137	0.935	1.000
		β_2	0.010	0.208	0.217	0.964	1.000
	Case-cohort	β_1	−0.004	0.236	0.225	0.938	1.573
	Prentice	β_2	0.009	0.402	0.399	0.954	1.148
	Case-cohort	β_1	0.067	0.195	0.204	0.932	2.321
	MLE	β_2	−0.028	0.296	0.319	0.948	2.125
	Case-control	β_1	0.073	0.198	0.210	0.950	2.274
	MLE	β_2	−0.024	0.305	0.320	0.954	2.019
	End-point	β_1	0.032	0.169	0.165	0.943	3.117
	MLE	β_2	0.001	0.256	0.267	0.954	2.859

BIAS: average bias of the estimates; SE: empirical standard error of the estimates; SEE: average of the estimated standard errors; CP: empirical coverage probabilities of Wald-type confidence intervals with 95% confidence level; RE: relative efficiency defined as the ratio of EGPS's; Full-cohort: maximum partial likelihood estimate using full cohort; Case-cohort Prentice: pseudo-likelihood estimate under case-cohort design; Case-cohort MLE: MLE under case-cohort design; Case-control MLE: MLE under case-control design; End-point MLE: MLE under the proposed end-point design.

the ratio between the inverse of its empirical variance (regarded as an efficiency estimate) and the average sample size used for covariate observation. Then for each estimate, we calculated a relative efficiency of that estimate to the corresponding maximum partial likelihood estimate, by taking the ratio of the two EGPS's. The results are summarized in Table 1.

Under both model setups, all the estimates were essentially unbiased. For the case-cohort design, the MLE had smaller empirical standard errors than the pseudolikelihood estimates. The MLE under case-cohort design and case-control design had almost the same empirical standard errors; as expected, since the two designs used almost the same amount of covariates ascertainment. The MLE under the proposed end-point sampling design was obviously less variable than

that under case-cohort and case-control designs, suggesting that using comparable size of covariates ascertainment, the proposed deign was more efficient than case-cohort and case-control designs. The MLE under the proposed design had the highest relative efficiency defined from the EGPS, implying the merit of the proposed design. All the estimated standard errors were close to the empirical standard errors, and the coverage probabilities were reasonably close to the nominal level. The proposed EM algorithm possesses reasonable computational efficiency. In our simulation, on average, the algorithm took around 9 steps to converge for the proposed design and, in each iteration, the maximization was solved efficiently since the objective function is concave.

4.2. Inverse probability weighting approach

Keeping the dimension of Z at 2, we considered two model setups. In the first, Z_1 and Z_2 were independently generated as Binomial with success probability 0.5 and uniform on $(0, 1)$, respectively. $\beta_1 = 1$, $\beta_2 = -1$ and $\lambda(t) = 0.5$. Here the covariates-dependent censoring was generated. Specifically, $C = \min\{\tilde{C}, \tau\}$ with $\tilde{C}$ exponential with mean $0.6Z_2$, and $\tau = 0.7$. This gave about 89% censoring percentage. In the second setup, we kept the same generating scheme for the two covariates. $\beta_1 = -0.5$, $\beta_2 = 1$ and $\lambda(t) = t$. We set $C = (1 - Z_1) \times \text{Uniform}(0, Z_2) + Z_1 \times \min\{\text{Uniform}(Z_2, 1.1), 1\}$, where $\text{Uniform}(a, b)$ stands for the uniform distribution from a to b. The censoring rate was also around 89%.

We set $n = 2,000$. For the proposed end-point sampling design, we set $p(y) = y$ for the first setup, and $p(y) = 0.7y^2$ for the second setup. For the purpose of comparison, we also considered the equal probability sampling with $p(y)$ a constant c. This can be treated as case-control sampling. In the first setup $c = 0.225$, and in the second $c = 0.245$. The two values were chosen to yield a comparable size of covariate ascertainment for the two sampling designs. When calculating the inverse probability weighted estimates, we used the 'fminsearch' function to find the optimum. Moreover, the maximum partial likelihood estimates based on the full cohort were obtained. One thousand replicates were done. For each estimate, we report the average bias, the empirical standard error, the average of estimated standard error, the coverage rate of 95% Wald-type confidence interval, and the relative efficiency based on EGPS. The results are given in Table 2.

We find from the results that all the estimates were essentially unbiased. The inverse probability weighted estimates under the proposed design were more efficient than those under the equal selection probability, as revealed by the standard error and the relative efficiency comparisons. The plug-in variance

Table 2. Summarized simulation results for different retrospective designs with the inverse probability weighting approach.

Setup	Method	Parameter	BIAS	SE	SEE	CP	RE
1	Full-cohort	β_1	0.003	0.149	0.150	0.955	1.000
		β_2	-0.009	0.266	0.283	0.966	1.000
	Equal-probability	β_1	0.007	0.182	0.186	0.958	2.149
	IPW	β_2	-0.024	0.351	0.361	0.956	1.855
	End-point	β_1	0.009	0.166	0.167	0.952	2.604
	IPW	β_2	-0.029	0.313	0.333	0.967	2.335
2	Full-cohort	β_1	0.006	0.185	0.179	0.943	1.000
		β_2	0.003	0.279	0.280	0.955	1.000
	Equal-probability	β_1	0.002	0.236	0.222	0.938	1.880
	IPW	β_2	-0.002	0.357	0.348	0.942	1.784
	End-point	β_1	-0.004	0.219	0.206	0.936	2.183
	IPW	β_2	-0.007	0.303	0.303	0.947	2.477

BIAS: average bias of the estimates; SE: empirical standard error of the estimates; SEE: average of the estimated standard errors; CP: empirical coverage probabilities of Wald-type confidence intervals with 95% confidence level; RE: relative efficiency defined as the ratio of EGPS's; Full-cohort: maximum partial likelihood estimate using full cohort; Equal-probability IPW: inverse probability weighted estimate under equal probability (case-control) design; End-point IPW: inverse probability weighted estimate under the proposed end-point design.

estimates were generally close to the empirical ones, and the confidence intervals had adequate coverage rates.

5. An Example

We applied the proposed sampling design to analyze the data from the South Welsh nickel refiners study. In this study, men employed in a nickel refinery in South Wales were investigated to determine the risk of developing carcinoma of the bronchi and nasal sinuses associated with the nickel refining. The cohort was identified by using the weekly payrolls of the company, and the full cohort was followed from 1934 to 1981. The full cohort consists of the complete records of 679 workers employed before 1925. There were 56 deaths from cancer of the nasal sinus until 1981. The details of the data can be found in Appendix VIII in Breslow and Day (1987).

Breslow and Day (1987) used the Cox model to fit the data. The survival time they considered was the years since first employment to the death from cancer of the nasal sinus. Three covariates, age at first employment (AFE), year at first employment (YFE), and exposure level (EXP) were found to be significant on the survival time. In the original data set reported in Breslow and Day (1987), all covariates information were collected. However, since the event rate

is only around 8%, the retrospective sampling designs is preferred. Here we first tried the case-cohort design and the end-point design with maximum likelihood estimation. For the end-point sampling design, we considered the 150 covariates ascertainment size of the censored individuals with the largest censoring times. In order to make the case-cohort sampling comparable, the subcohort size was set to 165. Then, for the end-point design with inverse probability weighing approach, we chose $p(y) = 1.3 \times 10^{-4} y^2$, yielding a covariate ascertainment proportion of about 24% (around 150 controls for covariates ascertainment). The so-called equal-probability sampling was used for comparison, and the selection probability of the controls was set to 0.24 to obtain a comparable covariate ascertainment percentage. We took the same covariates transformation as Breslow and Day (1987): $\log(\mathrm{AFE}-10)$, $(\mathrm{YFE}-1915)/10$, $[(\mathrm{YFE}-1915)/10]^2$, and $\log(\mathrm{EXP}+1)$ to fit the Cox model. For the MLE under the retrospective sampling designs, the EM algorithm was applied to get the MLE's, and the bootstrap method with 500 bootstrap samples was used to obtain the standard error estimates. For the inverse probability weighting approach, we did the sampling 500 times and report the average point estimates, standard error estimates and p-values. We also report the full cohort results under the Cox model using maximum partial likelihood. The results are summarized in Table 3.

Based on the full-cohort analysis results, three covariates, $\log(\mathrm{AFE}-10)$, $[(\mathrm{YFE}-1915)/10]^2$ and $\log(\mathrm{EXP}+1)$, had significant effect at the 0.05 significance level. Using the case-cohort design with subcohort size 165, the maximum likelihood estimation procedure failed to find the significant effect of $[(\mathrm{YFE}-1915)/10]^2$ at the 0.05 significance level, while the end-point sampling design with 150 controls gave the same inferential results as that of the full-cohort analysis. In most cases, the estimated standard errors of the MLE under the end-point design were smaller than their counterparts under the case-cohort design. For the inverse probability weighting approach, the average estimated standard errors under the end-point sampling were slightly smaller than the counterparts under the equal-probability sampling. Among the 500 samples, the percentages of times the end-point sampling found the four covariates were significant at the 0.05 level were 100%, 0.2%, 84.6% and 100%, respectively, while the percentages under the equal-probability sampling were 100%, 0.2%, 79.8% and 100%. These findings suggest that the proposed end-point designs is more efficient than the usual retrospective designs.

6. Concluding Remarks

Retrospective sampling designs, such as case-cohort and case-control designs, are cost effective for large epidemiological cohort studies when the event time is

Table 3. Summarized results for the South Welsh nickel refiners study.

Method	Parameter	EST	SE	p-Value
Full-cohort	$\log(\text{AFE}-10)$	2.2139	0.4319	$<0.0001^*$
	$(\text{YFE}-1915)/10$	0.0761	0.3074	0.8045
	$[(\text{YFE}-1915)/10]^2$	-1.3128	0.4942	0.0079^*
	$\log(\text{EXP}+1)$	0.7873	0.1752	$<0.0001^*$
Case-cohort MLE	$\log(\text{AFE}-10)$	1.9318	0.5109	0.0002^*
	$(\text{YFE}-1915)/10$	-0.3672	0.4074	0.3675
	$[(\text{YFE}-1915)/10]^2$	-1.1480	0.6665	0.0850
	$\log(\text{EXP}+1)$	0.8625	0.2210	0.0001^*
End-point MLE	$\log(\text{AFE}-10)$	3.1421	0.5937	$<0.0001^*$
	$(\text{YFE}-1915)/10$	0.2523	0.3970	0.5251
	$[(\text{YFE}-1915)/10]^2$	-1.4426	0.5975	0.0158^*
	$\log(\text{EXP}+1)$	0.7241	0.2025	0.0003^*
Equal-probability IPW	$\log(\text{AFE}-10)$	2.3638	0.5743	0.0005
	$(\text{YFE}-1915)/10$	0.0971	0.3923	0.6515
	$[(\text{YFE}-1915)/10]^2$	-1.3710	0.5300	0.0404
	$\log(\text{EXP}+1)$	0.8113	0.2097	0.0013
End-point IPW	$\log(\text{AFE}-10)$	2.3347	0.5653	0.0009
	$(\text{YFE}-1915)/10$	0.0785	0.3778	0.6849
	$[(\text{YFE}-1915)/10]^2$	-1.3428	0.5165	0.0330
	$\log(\text{EXP}+1)$	0.8078	0.2036	0.0006

EST: estimate of regression parameter; SE: estimate of standard error; p-value: p-value of the significance test; * significant at 0.05 significance level; Full-cohort: full data; Case-cohort MLE: MLE under the case-cohort design with subcohort of 165; End-point MLE: MLE under the end-point design with 150 largest censoring subjects; Equal-probability IPW: inverse probability weighting under the equal selection probability (case-control) design with $p(y) = 0.24$ (average for 500 repeated samples); Equal-probability IPW: inverse probability weighting under the end-point design with $p(y) = 1.3 \times 10^{-4} y^2$ (average for 500 repeated samples).

the outcome of interest. We propose a novel retrospective sampling design, end-point sampling, which samples all the event cases and controls with the larger censoring times. Under the Cox model and the independent censoring assumption, the maximum likelihood estimation is applied to estimate the regression parameters. We adopt an EM algorithm to obtain the MLE. The MLE is shown to be asymptotically normal and semiparametrically efficient. For conditionally independent censoring, we use the inverse probability weighting procedure for parameter estimation, and the resulting estimator is also asymptotically normal. Under the same estimation procedure, the proposed sampling design is more efficient than the ordinary case-cohort and case-control designs with comparable numbers of controls.

Data sampled from the proposed end-point sampling design can be easily analyzed under some other semiparametric survival models, such as linear transformation models and accelerated failure time models. Efficiency gain compared with case-cohort or case-control design is expected. The idea of the proposed sampling design is also possible to be extended to other two-phase cohort studies. These are topics of future research.

Acknowledgement

The authors thank the Associate Editor and two referees for their insightful comments and suggestions that have led to many improvements. W. Yu's research was supported by the National Natural Science Foundation of China (11101091, 11271081). K. Chen's research was supported by the Hong Kong Research Grant Council (601011, 600612, 600813, 16300714).

Appendix

Regularity conditions for Theorems 1, 2, and 3 are these.

1. The distribution function F is strictly increasing with derivative f absolutely continuous and β_0 lies in the interior of a known compact set in $\mathbb{R}^p$. The covariate Z lies in a bounded set $\mathcal{Z}$.
2. The baseline hazard function $\lambda_0(t) > 0$ for $t \in [0, \tau]$. There exists a positive constant c such that $P(C > \tau) > c$.
3. (First Identifiability) Let

$$\Psi_i(\beta, \Lambda, F) = \exp\left\{\delta_i \beta^\top Z_i - \Lambda(Y_i)\exp(\beta^\top Z_i)\right\}^{\Delta_i} \times \left[\int_{\mathcal{Z}} \left\{\lambda(Y_i)\exp(\beta^\top z)\right\}^{\delta_i} \exp\left\{-\Lambda(Y_i)\exp(\beta^\top z)\right\} f(z)dz\right]^{(1-\Delta_i)}.$$

   If $\Psi_i(\beta^*, \Lambda^*, F^*)\lambda^*(Y_i)^{\delta_i} f^*(Z_i)^{\Delta_i} = \Psi_i(\beta_0, \Lambda_0, F_0)\lambda_0(Y_i)^{\delta_i} f_0(Z_i)^{\Delta_i}$ almost surely, then $\beta^* = \beta_0$, $\Lambda^* = \Lambda_0$ and $F^* = F_0$.
4. (Second Identifiability) If

$$v^\top l_\beta(\beta_0, \Lambda_0, F_0) + l_\Lambda(\beta_0, \Lambda_0, F_0)\left\{\int p d\Lambda_0\right\} + l_F(\beta_0, \Lambda_0, F_0)\left\{\int q dF_0\right\} = 0$$

   almost surely for some $v \in \mathbb{R}^p$, $p \in BV[0, \tau]$ and $q \in BV[\mathcal{Z}]$, then $(v, p, q) = 0$, where $v^T l_\beta$, $l_H[g_1]$ and $l_F[g_2]$ denote the partial derivatives of l along directions of v, g_1, and g_2, respectively.

The proof of Theorems 1 and 2 are similar to that in Yao (2015), which is based on the argument on maximum likelihood estimators of Van der Vaart (1998, pp.419-424). We give the rough steps here.

Proof of Theorem 1. In the first step, the jump size of $\hat{\Lambda}_{\mathsf{M}}$ is shown to be finite almost surely, otherwise the log-likelihood would diverge to $-\infty$. In the second step, $\hat{\Lambda}_{\mathsf{M}}$ is bounded almost surely, otherwise if a new estimator $\bar{\Lambda}_{\mathsf{M}} = \hat{\Lambda}_{\mathsf{M}}/\hat{\Lambda}_{\mathsf{M}}(\tau)$ were considered, it would contradict the maximum property of $\hat{\Lambda}_{\mathsf{M}}$. Then Helly's selection theorem guarantees that for any subsequence of $\hat{\Lambda}_{\mathsf{M}}$, there exists a further subsequence that converges pointwise to some monotone function Λ^*. Without loss of generality, assume that $\hat{F}_{\mathsf{M}}$ converges to F^* and $\hat{\beta}_{\mathsf{M}}$ converges to β^* for the same subsequence. In the third step, we show that $\Lambda^* = \Lambda_0$, $F^* = F_0$ and $\beta^* = \beta_0$ with probability one. Note that

$$\hat{\Lambda}_{\mathsf{M}}(t) = \sum_{i=1}^{n} \int_0^t \hat{\lambda}_{\mathsf{M}}(u) dN_i(u) = -\sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\sum_{j=1}^{n} \frac{\Psi_{j\Lambda}(\hat{\beta}_{\mathsf{M}}, \hat{\Lambda}_{\mathsf{M}}, \hat{F}_{\mathsf{M}})[I\{Y_j \geqslant u\}]}{\Psi_j(\hat{\beta}_{\mathsf{M}}, \hat{\Lambda}_{\mathsf{M}}, \hat{F}_{\mathsf{M}})}},$$

where $\Psi_{j\Lambda}[I\{\cdot \geqslant u\}]$ is the partial derivative of Ψ_j with respect to Λ along $\Lambda + \epsilon[I\{\cdot \geqslant u\}]$.

Construct

$$\tilde{\Lambda}(t) = -\sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\sum_{j=1}^{n} \Psi_{j\Lambda}(\beta_0, H_0, F_0)[I\{Y_j \geqslant u\}]/\Psi_j(\beta_0, \Lambda_0, F_0)},$$

which converges to $\Lambda_0(t)$ almost surely by some careful calculation and repeated use of the Glivenko-Cantelli Theorem.

By the strictly increasing and smoothness condition of Λ_0, one can show that

$$\lim_{n\to\infty} \frac{d\hat{\Lambda}_{\mathsf{M}}(t)}{d\tilde{\Lambda}(t)} = \frac{\lambda^*(t)}{\lambda_0(t)}$$

uniformly, where λ^* is the derivative of Λ^*. Similar construction of $\tilde{F}$ and a similar argument yield

$$\lim_{n\to\infty} \frac{d\hat{F}_{\mathsf{M}}(z)}{d\tilde{F}(z)} = \frac{f^*(z)}{f_0(z)}$$

uniformly.

Letting $n \to \infty$, the inequality $l(\hat{\beta}_{\mathsf{M}}, \hat{\Lambda}_{\mathsf{M}}, \hat{F}_{\mathsf{M}}) \geqslant l(\beta_0, \tilde{\Lambda}, \tilde{F})$ yields

$$E\left(\log \frac{\Psi_j(\beta^*, \Lambda^*, F^*)\lambda^*(Y_j)^{\delta_j} f^*(Z_j)^{\Delta_j}}{\Psi_j(\beta_0, \Lambda_0, F_0)\lambda_0(Y_j)^{\delta_j} f_0(Z_j)^{\Delta_j}}\right) \geqslant 0.$$

The left-hand side here is the negative Kullback-Leibler distance, therefore Condition 3 requires that $\beta^* = \beta_0$, $\Lambda^* = \Lambda_0$ and $F^* = F_0$ with probability one.

Further, the continuity of Λ_0 and F_0 ensures that the convergence is uniform for Theorem 1.

Proof of Theorem 2. Let

$$\begin{aligned}\mathcal{L}(\beta,\Lambda,F) &= \log\Psi_j + \delta_j \log\lambda(Y_j) + \Delta_j \log f(Z_j),\\ \Phi_n(\beta,\Lambda,F) &= \mathcal{P}_n\Big[v^T\mathcal{L}_\beta + \mathcal{L}_H\Big\{\int p d\Lambda\Big\} + \mathcal{L}_F\Big\{\int q dF\Big\}\Big],\\ \Phi(\beta,\Lambda,F) &= \mathcal{P}\Big[v^T\mathcal{L}_\beta + \mathcal{L}_\Lambda\Big\{\int p d\Lambda\Big\} + \mathcal{L}_F\Big\{\int q dF\Big\}\Big],\end{aligned}$$

where $(v,p,q)\in\mathbb{R}^p\times l^\infty(\mathcal{Q}_1)\times l^\infty(\mathcal{Q}_2)$. We use $v^T\mathcal{L}_\beta$, $\mathcal{L}_\Lambda\{g_1\}$, and $\mathcal{L}_F\{g_2\}$ to denote the partial derivatives of $\mathcal{L}$ along direction of $\beta+\epsilon v$, $\Lambda+\epsilon g_1$, and $F+\epsilon g_2$, respectively, and let $\mathcal{P}_n$ denote the empirical measure based on n observations and $\mathcal{P}$ be its expectation.

By some careful calculation and Donsker's Theorem, we can show that

$$\sqrt{n}(\Phi_n-\Phi)(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}}) - \sqrt{n}(\Phi_n-\Phi)(\beta_0,\Lambda_0,F_0) = o_p(1)$$

when n is large enough and further calculation shows

$$\begin{aligned}&\sqrt{n}(\mathcal{P}_n-\mathcal{P})\Big[v^T\mathcal{L}_\beta(\beta_0,\Lambda_0,F_0)\\ &\quad+\mathcal{L}_\Lambda(\beta_0,\Lambda_0,F_0)\Big\{\int p d\Lambda_0\Big\} + \mathcal{L}_F(\beta_0,\Lambda_0,F_0)\Big\{\int q dF_0\Big\}\Big]\\ &= \sqrt{n}(\mathcal{P}_n-\mathcal{P})\Big[v^T\mathcal{L}_\beta(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}}) + \mathcal{L}_\Lambda(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}})\Big\{\int p d\hat\Lambda_{\mathsf{M}}\Big\}\\ &\quad+\mathcal{L}_F(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}})\Big\{\int q d\hat F_{\mathsf{M}}\Big\}\Big] + o_p(1)\\ &= -\sqrt{n}\mathcal{P}\Big[v^T\mathcal{L}_\beta(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}}) - v^T\mathcal{L}_\beta(\beta_0,\Lambda_0,F_0) + \mathcal{L}_\Lambda(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}})\Big\{\int p d\hat\Lambda_{\mathsf{M}}\Big\}\\ &\quad-\mathcal{L}_\Lambda(\beta_0,\Lambda_0,F_0)\Big\{\int p d\Lambda_0\Big\} + \mathcal{L}_F(\hat\beta_{\mathsf{M}},\hat\Lambda_{\mathsf{M}},\hat F_{\mathsf{M}})\Big\{\int q d\hat F_{\mathsf{M}}\Big\}\\ &\quad-\mathcal{L}_F(\beta_0,\Lambda_0,F_0)\Big\{\int q dF_0\Big\}\Big] + o_p(1). \qquad \text{(A.1)}\end{aligned}$$

One can show that there exists continuous invertible linear operator (B_1,B_{21},B_{22}) on the space $\mathbb{R}^p\times l^\infty(\mathcal{Q}_1)\times l^\infty(\mathcal{Q}_2)$, such that the right side of (A.1) has the same limit as

$$\begin{aligned}&= -\sqrt{n}\Big\{B_1[v,p,q]^T(\hat\beta_{\mathsf{M}}-\beta_0) + \int B_{21}[v,p,q]d(\hat\Lambda_{\mathsf{M}}-\Lambda_0)\\ &\quad+\int B_{22}[v,p,q]d(\hat F_{\mathsf{M}}-F_0)\Big\}\\ &\quad+o_p\Big(\sqrt{n}|\hat\beta_{\mathsf{M}}-\beta_0| + \sqrt{n}|\hat\Lambda_{\mathsf{M}}-\Lambda_0| + \sqrt{n}|\hat F_{\mathsf{M}}-F_0|\Big).\end{aligned}$$

Now with $(\tilde{v},\tilde{p},\tilde{q}) = (B_1, B_{21}, B_{22})^{-1}(v,p,q)$,

$$\begin{aligned}
&\sqrt{n}\Big\{v^T(\hat{\beta}_{\mathsf{M}}-\beta_0) + \int p d(\hat{\Lambda}_{\mathsf{M}}-\Lambda_0) + \int q d(\hat{F}_{\mathsf{M}}-F_0)\Big\}\\
&= -\sqrt{n}(\mathcal{P}_n-\mathcal{P})\Big[\tilde{v}^T\mathcal{L}_\beta(\beta_0,\Lambda_0,F_0) + \mathcal{L}_\Lambda(\beta_0,\Lambda_0,F_0)\Big\{\int \tilde{p}d\Lambda_0\Big\}\\
&\quad + \mathcal{L}_F(\beta_0,\Lambda_0,F_0)\Big\{\int \tilde{q}dF_0\Big\}\Big]\\
&\quad + o_p\Big(\sqrt{n}|\hat{\beta}_{\mathsf{M}}-\beta_0| + \sqrt{n}|\hat{\Lambda}_{\mathsf{M}}-\Lambda_0| + \sqrt{n}|\hat{F}_{\mathsf{M}}-F_0|\Big) + o_p(1). \qquad \text{(A.2)}
\end{aligned}$$

Since the first term on the right side of (A.2) is $O_p(1)$, one can see that

$$\sqrt{n}|\hat{\beta}_{\mathsf{M}}-\beta_0| + \sqrt{n}|\hat{\Lambda}_{\mathsf{M}}-\Lambda_0| + \sqrt{n}|\hat{F}_{\mathsf{M}}-F_0| = O_p(1).$$

Thus, the last two terms in the right side of (A.2) are $o_p(1)$.

Because the right side of (A.2) converges to a normal distribution by the Central Limit Theorem, we have proved that $\sqrt{n}(\hat{\beta}_{\mathsf{M}}-\beta_0, \hat{\Lambda}_{\mathsf{M}}-\Lambda_0, \hat{F}_{\mathsf{M}}-F_0)$ converges weakly to a zero-mean Gaussian process. Thus $\hat{\beta}_{\mathsf{M}}$ is an asymptotically linear estimator with influence function $\tilde{v}^T\mathcal{L}_\beta(\beta_0,\Lambda_0,F_0) + \mathcal{L}_\Lambda(\beta_0,\Lambda_0,F_0)\{\int \tilde{p}d\Lambda_0\} + \mathcal{L}_F(\beta_0,\Lambda_0,F_0)\{\int \tilde{q}dF_0\}$, which lies in the linear space spanned by the score functions $\{\tilde{v}^T\mathcal{L}_\beta + \mathcal{L}_\Lambda\{\int \tilde{p}d\Lambda\} + \mathcal{L}_F\{\int \tilde{q}dF\} : \tilde{v}\in\mathbb{R}^p, \tilde{p}\in\mathcal{Q}_1, \tilde{q}\in\mathcal{Q}_2\}$. By proposition 1 in Bickel et al. (1993, p.65), $\hat{\beta}_{\mathsf{M}}$ is semiparametrically efficient.

Proof of Theorem 3. The inverse probability weighted estimating equation can be written as

$$S_{\mathsf{IPW}}(\beta) = \sum_{i=1}^n \int \left\{ Z_i - \frac{\sum_{j=1}^n (\Delta_j/\pi_j) Z_j \exp(\beta^\top Z_j) I\{Y_j \geqslant t\}}{\sum_{j=1}^n (\Delta_j/\pi_j)\exp(\beta^\top Z_j) I\{Y_j \geqslant t\}} \right\} dN_i(t) = 0.$$

Let

$$\mu_n(t) = \frac{\sum_{j=1}^n Z_j \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\}}{\sum_{j=1}^n \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\}},\ a(t) = \sum_{j=1}^n Z_j \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\},$$

$$b(t) = \sum_{j=1}^n \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\},\ \hat{\mu}_n(t) = \frac{\sum_{j=1}^n (\Delta_j/\pi_j) Z_j \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\}}{\sum_{j=1}^n (\Delta_j/\pi_j)\exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\}},$$

$$\hat{a}(t) = \sum_{j=1}^n \frac{\Delta_j}{\pi_j} Z_j \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\},\ \hat{b}(t) = \sum_{j=1}^n \frac{\Delta_j}{\pi_j} \exp(\beta_0^\top Z_j) I\{Y_j \geqslant t\}.$$

For simplicity, we use $a, b, \hat{a}, \hat{b}$ instead of $a(t), b(t), \hat{a}(t), \hat{b}(t)$ in these calculations.

Now

$$\begin{aligned}
S_{\mathsf{IPW}}(\beta_0) &= \sum_{i=1}^{n} \int \{Z_i - \mu_n(t)\} dN_i(t) - \int \{\hat{\mu}_n(t) - \mu_n(t)\} dN_i(t) \\
&= \sum_{i=1}^{n} \int \{Z_i - \mu_n(t)\} dM_i(t) - \sum_{i=1}^{n} \int \left(\frac{\hat{a}}{\hat{b}} - \frac{a}{b} \right) dN_i(t) \\
&= \sum_{i=1}^{n} \int \{Z_i - \mu(t)\} dM_i(t) - \sum_{i=1}^{n} \int \left\{ \frac{1}{b}(\hat{a} - a) - \frac{a}{b^2}(\hat{b} - b) \right\} dM_i(t) \\
&\quad + o_p(\sqrt{n}) \\
&= \sum_{i=1}^{n} \int \{Z_i - \mu(t)\} dM_i(t) \\
&\quad - \sum_{i=1}^{n} \left(\frac{\Delta_i}{\pi_i} - 1 \right) \int \{Z_i - \mu(t)\} \exp(\beta_0^\top Z_i) I\{Y_i \geq t\} \lambda_0(t) dt + o_p(\sqrt{n}) \\
&= \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \eta_i + o_p(\sqrt{n}),
\end{aligned}$$

where $M_i(t) = N_i(t) - \int_0^t I\{Y_i \geqslant u\} \lambda_0(u) \exp(\beta_0^\top Z_i) du$, $\xi_i = \int \{Z_i - \mu(t)\} dM_i(t)$ and $\eta_i = (\Delta_i/\pi_i - 1) \int \{Z_i - \mu(t)\} \exp(\beta_0^\top Z_i) I\{Y_i \geq t\} \lambda_0(t) dt$, $i = 1, \ldots, n$.

Now $\hat{\beta}_{\mathsf{I}}$ satisfies the estimating equation $S_{\mathsf{IPW}}(\hat{\beta}_{\mathsf{I}}) = 0$, and it can be shown that $\hat{\beta}_{\mathsf{I}}$ is a consistent estimator of β_0. For any random vector ζ, let $Var(\zeta)$ denote its variance-covariance matrix. It follows by Taylor expansion, Slutsley's Theorem, and Central Limit Theorem that $\sqrt{n}(\hat{\beta}_{\mathsf{I}} - \beta_0)$ converges to a normal distribution with mean zero and variance-covariance matrix

$$Var(\xi_i)^{-1} \{Var(\xi_i) + Var(\eta_i)\} Var(\xi_i)^{-1} = \Sigma_1^{-1} (\Sigma_1 + \Sigma_2) \Sigma_1^{-1},$$

where Σ_1 and Σ_2 are as defined in Section 3.2.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins University Press, Baltimore.

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research, Volumne II: The Design and Analysis of Cohort Studies*. IARC, Lyon, France.

Breslow, N. E., Lumbley, T., Ballantyne, C. M. Chambless, L. E. and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort data. *Amer. J. Epidemiol.* **169**, 1398-1405.

Chen, K. (2001). Generalized case-cohort sampling. *J. Roy. Statist. Soc. Ser. B* **63**, 791-809.

Chen, K. (2004). Statistical estimation in the proportional hazards model with risk set sampling. *Ann. Statist.* **32**, 1513-1532.

Chen, K. and Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **86**, 755-764.

Chen, Y.-H. and Zucker, D. M. (2009). Case-cohort analysis with semiparametric transformation models. *J. Stat. Plan. Infer.* **139**, 3706-3717.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple diseases outcomes. *Biometrika* **96**, 887-901.

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

Kong, L. and Cai, J. (2009). Case-cohort analysis with accelerated failure time model. *Biometrics* **65**, 135-42.

Kulich, M. and Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Amer. Statist. Assoc.* **99**, 832-844.

Liu, M., Lu, W., Shore, R. E. and Jacquotte, A. Z. (2010). Cox regression model with time-varying coefficients in nested case-control studies. *Biostatistics* **11**, 693-706.

Lu, W. and Tsiatis, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika* **93**, 207-214.

Nan, B. (2004). Efficient estimation for case-cohort studies. *Canad. J. Statist.* **32**, 403-419.

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

Thomas, D. C. (1977). Appendum to "Methods of cohort analysis: appraisal by application to asbestos mining," by Liddell, F. D. K., McDonald, J. C. and Thomas, D. C. *J. R. Statist. Soc.* A **140**, 469-490.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge, U.K.

Yao, Y. (2015). Maximum likelihood method for linear transformation models with cohort sampling data. *Statist. Sinica* **25**, 1231-1248.

Zeng, D. and Lin, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J. Amer. Statist. Assoc.* **109**, 371-383.

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.

E-mail: yaoyuan@hkbu.edu.hk

Department of Statistics, School of Management, Fudan University, Shanghai, 200433, P. R. China.

E-mail: wenyu@fudan.edu.cn

Department of Mathematics, Hong Kong University of Science and Technology, Kowloon, Hong Kong.

E-mail: makchen@ust.hk

(Received August 2015; accepted January 2016)