Statistics and Its Interface Volume 13 (2020) 373–383

Semiparametric accelerated failure time modeling for multivariate failure times under multivariate outcome-dependent sampling designs

Tsui-Shan Lu, Sangwook Kang*, and Haibo Zhou

Researchers working on large cohort studies are always seeking for cost-effective designs due to a limited budget. An outcome-dependent sampling (ODS) design, a retrospective sampling scheme where one observes covariates with a probability depending on the outcome and selects supplemental samples from more informative segments, improves the study efficiency while effectively controlling for the budget. To take the advantage of the ODS scheme when multivariate failure times are main response variables, relevant study designs and inference procedures need to be studied.

In this paper, we consider a general multivariate-ODS design for multivariate failure times under the framework of a semiparametric accelerated failure time model. We develop a weighted estimating equations approach, based on the induced smoothing method, for parameter estimation. Extensive simulation studies show that our proposed design and estimator are more efficient than other competing estimators based on simple random samples. The proposed method is illustrated with a real data set from the Busselton Health Study.

Keywords and phrases: Biased sampling, Induced smoothing, Rank-based estimation, Resampling, Weighted estimating equations, Sandwich variance estimation.

1. INTRODUCTION

In some modern biomedical and epidemiological cohort studies, major sources of high costs can be attributed to a long follow-up period and acquirement of expensive covariate measurements. Investigators have always been interested in seeking alternative cost-effective study designs to reduce the cost while retaining the study power. An outcome-dependent sampling (ODS) design is a retrospective sampling scheme where one selects an overall random sample from the underlying study cohort and some additional supplemental samples through some probability sampling schemes, depending on the level of the outcome variable [7, 14, 33, 35, 39, 40, 41, 42]. The principal idea is to take a biased sampling by concentrating resources on the segments having the greatest amount of information. By allowing the probability of each individual to be selected into the sample via an ODS design to depend on the observed value of the outcome, researchers can enhance the study efficiency while reducing the overall cost of the study, especially in large-scale cohort studies. Recent studies have discussed such generalized ODS designs with continuous outcome variables.

For failure time data, the case-cohort study design [30] is the most widely used biased-sampling scheme for a rare disease situation. When diseases are not rare, the generalized case-cohort study design along with its various estimating procedures have been proposed for a single disease outcome. A stratified case-cohort design further improves the study efficiency by dividing the cohort into some mutually exclusive strata based on a discrete random variable. Statistical inferences for case-cohort data under failure time models have been developed and discussed in most literatures, such as the commonly-used proportional hazards model [3, 4, 6, 7, 9, 31, 37], the additive hazards model [26, 34], the accelerated failure time (AFT) model [23] and others.

Recently, Ding et al. [16] considered a general failure-time ODS design for right-censored data, where a simple random sample is selected from the underlying cohort and some additional supplemental samples are drawn from the strata of interest which are mutually-exclusively partitioned from the range of observed time of all the cases. They proposed maximum semiparametric empirical likelihood estimation under the framework of the Cox proportional hazards model. Yu et al. [37] further developed a weighted pseudo-score estimator for fitting additive hazards models under the ODS scheme. Work above mentioned has focused on analyzing univariate failure time data with a single disease outcome under the ODS design.

In practice, multivariate data have increasingly been encountered in many contexts. It might be of interest to consider several disease outcomes or several subtypes of a disease simultaneously. One could deal with multiple disease outcomes by analyzing each disease separately. Without considering the induced correlation among outcomes, however, this could lead to less efficient or erroneous results. For example, in the Busselton Health Study [13], it was of interest

*To whom correspondence should be addressed.

to explore the potential risk factors on the coronary heart disease and stroke events. It is desirable to model times to coronary heart disease and stroke events in a multivariate framework since the two endpoints were observed from the same subject and hence correlated. As a result, the correlated nature of failure times has to be taken into consideration when analyzing such data.

Literature for inferences on multivariate failure-time data under biased sampling schemes is mostly limited on the case-cohort study design with clustered failure times [23, 28, 38]. For multivariate data under the ODS design (multivariate-ODS), Lu, Longnecker, and Zhou [29] proposed an empirical likelihood inference procedures for based on the continuous responses within a cluster. To the best of our knowledge, multivariate-ODS designs for multivariate failure time data and relevant inference procedures have not yet been explored.

In this paper, we propose a statistical inference procedure for fitting multivariate failure time data from multivariate-ODS designs. As the underlying model for failure times, we consider a semiparametric AFT model, which directly links the failure time to covariates through a log function without specifying the error distribution. For estimation, we consider a computationally efficient induced smoothing method for rank-based estimating equations [4]. We take a marginal model approach to handle the correlated feature among failure times. We incorporate the inverse of sampling probability weights to cover the biased sampling feature of the multivariate-ODS design. The remainder of the paper is organized as follows. In Section 2, we present the notation and the data structure under the multivariate-ODS design with multivariate failure time data. The estimation procedure based on the induced smoothing for semiparametric AFT models under the multivariate-ODS scheme is developed. In Section 3, we conduct simulation studies to evaluate finite sample performances of our proposed estimators. We apply the proposed methods to analyze the Busselton Health Study data in Section 4. Discussion and final remarks are given in Section 5.

2. DESIGN AND ESTIMATING APPROACH

2.1 Multivariate AFT model

Suppose that there are N independent subjects in a cohort study with K disease outcomes of interest. Let $\boldsymbol{T}_i = (T_{i1}, \ldots, T_{iK})^\top$ be the log-transformed independent failure time response vector and $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{iK})^\top$ denote the corresponding log-transformed censoring time vector, where $i = 1, \ldots, N$ and $k = 1, \ldots, K$. The observed time is $X_{ik} = \min(T_{ik}, C_{ik})$ and the corresponding observed time vector is $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})^\top$. Let $\triangle_{ik} = I(T_{ik} \leq C_{ik})$ denote an indicator for failure, where $I(\cdot)$ is an indicator function. The corresponding failure time indicator vector is then $\boldsymbol{\Delta}_i = (\triangle_{i1}, \ldots, \triangle_{iK})^\top$. Let $Y_{ik}(t) = I(X_{ik} \geq t)$ denote the at-risk process and $N_{ik}(t) = I(X_{ik} \leq t, \triangle_{ik} = 1)$ denote the counting process for outcome k of subject i. Let $\boldsymbol{Z}_{ik}$ be a p-dimensional covariate vector corresponding to the kth disease outcome for subject i. We assume that T_{ik} and C_{ik} are independent conditional on $\boldsymbol{Z}_{ik}$. Let τ be the study end time. Then the marginal semiparametric AFT model [9, 10] is

$$T_{ik} = \boldsymbol{Z}_{ik}^\top \boldsymbol{\beta} + \epsilon_{ik}, \quad i = 1, \ldots, N; k = 1, \ldots, K, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of fixed and unknown parameters of interest and the error terms, $\epsilon_i = \{\epsilon_{i1}, \ldots, \epsilon_{iK}\}$, are independently and identically distributed for each subject i. A subject may experience all, only some, or even none of the K disease outcomes. One may also incorporate disease-specific effects in the model.

2.2 Multivariate-ODS designs

Under a multivariate-ODS design, the sampling procedure is conducted in two steps. In the first step, a simple random sample (SRS) without replacement is drawn from the underlying study cohort. In the second step, supplemental samples from certain segments of the cohort are selected. In specific, suppose that M domains in the observed times, A_m, $m = 1, \ldots, M$, are defined and subsequent samplings from A_ms are followed. Typical ODS designs are with $M = 2$ to capture the responses with large and small values [29, 37] and we assume this. Let ξ_i and η_{im} denote the indicators for the SRS of size n_0 and the supplemental sample from A_m of size n_m, respectively. Let ζ_{im} denote the indicator for $\boldsymbol{X}_i$ to be included in A_m, i.e., $I(\boldsymbol{X}_i \in A_m)$ where $\zeta_i = \sum_{m=1}^{M} \zeta_{im}$. Then, for example, A_ms can be defined as $A_1 = \{X_{i1} > a_1, \ldots, X_{iK} > a_K, \Delta_{ik} = 1 \text{ for some or all } k\}$ and $A_2 = \{X_{i1} < b_1, \ldots, X_{iK} < b_K, \Delta_{ik} = 1 \text{ for some or all } k\}$, $k = 1, \ldots, K$, respectively. Here, $\boldsymbol{a} = \{a_k,\ k = 1, \ldots, K\}$ and $\boldsymbol{b} = \{b_k,\ k = 1, \ldots, K\}$ are known constants vectors used as cutpoints satisfying $\{a_k > b_k, \forall k\}$. In this way, the domains of interest in the observed times are defined to large and small observed failure times. Supplemental samples are randomly drawn from each of these two domains with sizes n_1 and n_2, respectively. When $K = 1$, this reduces to the definition of the supplemental components in the ODS design for univariate failure time data [37]. Without loss of generality, we consider the case when $K = 2$, i.e., each subject will have at most two endpoints. The cutpoints are then set to be a_1, a_2, b_1 and b_2.

Without censoring, the supplemental components can be defined in the same way as those in the multivariate-ODS design for continuous responses [29]. Due to censoring, however, exact failure times might not be observable; that is, the clusters might include censored failure times or might not include any failure times. Thus, the supplemental components are restricted to clusters having at least one failure time. In particular, we consider two sampling designs to select the supplemental samples:

- MODS Design 1 (All Failures Design) - every observed time in a cluster is a failure and all in a cluster satisfy the criteria set by the cutpoints to be sampled in the supplemental samples: either all failure times greater than the designated cutpoints or all lower than the designated cutpoints.
- MODS Design 2 (At-least-one Failure Design) - we do not require all the observed times in a cluster to be failure times as in MODS Design 1. In other words, as long as there exists at least one failure in a cluster satisfies the criteria set by cutpoints, that cluster is eligible to be selected into the supplemental samples.

As one can expect, the sizes of supplemental components for MODS Design 1 would be smaller than those for MODS Design 2 due to the restriction that all the elements in a cluster need to be failures. For both designs, the resulting sample is composed of three components:

(i) MODS Design 1 (All Failure Design):

(a) SRS Component ($\xi_i = 1$) : $\{\boldsymbol{X}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i\}$, $i = 1, \cdots, n_0$;

(b) Supplemental component 1 ($\zeta_{i1} = 1, \eta_{i1} = 1$): $\{\boldsymbol{X}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i \mid X_{ik} > a_k \text{ and } \Delta_{ik} = 1\}$, $i = 1, \ldots, n_1$ and $k = 1, \ldots, K$;

(c) Supplemental component 2 ($\zeta_{i2} = 1, \eta_{i2} = 1$): $\{\boldsymbol{X}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i \mid X_{ik} < b_k \text{ and } \Delta_{ik} = 1\}$, $i = 1, \ldots, n_2$ and $k = 1, \ldots, K$;

(ii) MODS Design 2 (At-least-one Failure Design):

(a) SRS Component ($\xi_i = 1$): $\{\boldsymbol{X}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i\}, i = 1, \cdots, n_0$;

(b) Supplemental component 1 ($\zeta_{i1} = 1, \eta_{i1} = 1$): $\{\boldsymbol{X}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i \mid X_{ik} > a_k \text{ and for some } k, \Delta_{ik} = 1\}$, $i = 1, \ldots, n_1$ and $k = 1, \ldots, K$;

(c) Supplemental component 2 ($\zeta_{i2} = 1, \eta_{i2} = 1$): $\{\boldsymbol{X}_i, \boldsymbol{\Delta}_i, \boldsymbol{Z}_i \mid X_{ik} < b_k \text{ and for some } k, \Delta_{ik} = 1\}$, $i = 1, \ldots, n_2$ and $k = 1, \ldots, K$.

Note that covariate information is collected only for the sampled subjects. The samples from above components consist of the observed sample and the total sample size is $n = n_0 + n_1 + n_2$.

2.3 Estimation of model parameters

Let $e_{ik}(\beta) = X_{ik} - \boldsymbol{Z}_{ik}^\top \beta$ be the residual for disease outcome k in subject i. For full cohort data, rank-based estimating equations with a Gehan type weight are defined as follows [21]

(2)
$$U(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} \Delta_{ik}(\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl}) I\{e_{jl}(\beta) \geq e_{ik}(\beta)\}$$

An induced smoothed version of (2) is $\tilde{U}(\beta) = \mathrm{E}_W\{U(\beta + N^{-1/2}\Gamma^{-1/2}W)\}$ [22] where

(3)
$$\tilde{U}(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} \Delta_{ik}(\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl}) \Phi\left\{\frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ijkl}}\right\},$$

$W \sim \mathcal{N}(0, I_p)$, $r_{ijkl}^2 = N^{-1}(\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl})^\top \Gamma^{-1}(\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl})$, $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $\mathrm{E}_W\{\cdot\}$ denotes an expectation with respect to W, and Γ is typically set to I_p, a p-dimensional identity matrix.

For the data under an ODS scheme, $\boldsymbol{Z}_{ik}$s are available only for the ODS sample. Thus, (3) cannot be evaluated. To overcome this, we propose a weighted estimating equations approach where the weights are the inverse of the sampling probabilities. Since supplemental components are sampled at the subject level, our proposed weight is also constructed at the subject level. Let $p_m = n_m/N$ be the SRS portion ($m = 0$) and supplemental portions ($m = 1$ and 2) in the underlying study cohort. Sampling probabilities for the SRS and supplemental components are p_0 and $r_m = n_m/(N_m - n_{0,m}), m = 1$ and 2, respectively, where N_m and $n_{0,m}$ are the sizes of the full cohort and SRS sample in A_m. Then,

(4)
$$\begin{aligned} w_i = {} & \xi_i \prod_{k=1}^{K}(1-\Delta_{ik})\, p_0^{-1} + \xi_i \left\{1 - \prod_{k=1}^{K}(1-\Delta_{ik})\right\}(1-\zeta_i) p_0^{-1} \\ & + \xi_i \left\{1 - \prod_{k=1}^{K}(1-\Delta_{ik})\right\}\zeta_i \\ & + (1-\xi_i)\left\{1 - \prod_{k=1}^{K}(1-\Delta_{ik})\right\}\sum_{m=1}^{M}\zeta_{im}\eta_{im} r_m^{-1}. \end{aligned}$$

The weighted version of $\tilde{U}(\beta)$ incorporating w_i defined in (4) is

(5)
$$\begin{aligned} \tilde{U}_c(\beta) = {} & \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} w_i w_j \Delta_{ik} \\ & \times (\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl}) \Phi\left\{\frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ijkl}}\right\}. \end{aligned}$$

The proposed estimation procedures are in the similar spirit as those in Chiou, Kang, and Yan [9, 10] in which clustered failure times from stratified case-cohort designs were considered. The proposed weight functions and associated estimating functions, however, reflect the novel multivariate-ODS design that contains the case-cohort design as a special case - $M = 1$ and $a_k = 0$ for all $k = 1, \ldots, K$. The estimator of β, $\widehat{\beta}$ is defined as the solution to $\tilde{U}_c(\beta) = 0$. Then, using the argument in Appendix of Chiou, Kang and Yan [10] and Chapter 3 of Yu et al. [37], $\hat{\beta}$ can be shown to be consistent and asymptotically normal. The detailed derivation is provided in Appendix A.

2.4 Variance estimation

For estimation of the variance of $\widehat{\beta}$, we propose a robust sandwich estimator aided by a resampling method. This type of variance estimation has been previously employed in rank-based estimation with the induced smoothing for fitting semiparametric AFT models [9, 10]. Here we estimate the asymptotic variance-covariance function of $\widehat{\beta}$ by $A^{-1}(\widehat{\beta})V(\widehat{\beta})A^{-1}(\widehat{\beta})$. We obtain $A(\cdot)$ and $V(\cdot)$ separately. $A(\cdot)$ can be directly calculated by taking the first derivative of (5). Specifically,

$$A(\beta) = n^{-1}\frac{\partial}{\partial \beta^{\top}}\tilde{U}_c(\beta).$$

To calculate $V(\cdot)$, we use a multiplier resampling approach. We generate N independent and identically distributed multipliers us from a positive random variable having both mean and variance one. Given the realized values of us, we can calculate one bootstrap replicate $\tilde{U}_c^*(\widehat{\beta})$ where

$$\tilde{U}_c^*(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} u_i u_j w_i w_j \Delta_{ik} \times (\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl})\Phi\left\{\frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ijkl}}\right\}.$$

By independently repeating this process of generating us and evaluating $\tilde{U}_c^*(\widehat{\beta})$ B times, we obtain B bootstrap replicates of $\tilde{U}_c^*(\widehat{\beta})$s. Then, $V(\widehat{\beta})$ can be obtained by the sample variance of B $\tilde{U}_c^*(\widehat{\beta})$s. Although this is a resampling procedure requiring evaluations of $\tilde{U}_c^*(\widehat{\beta})$s many times to obtain a variance estimate, the computational load is not heavy since we only need evaluations of $\tilde{U}_c^*(\widehat{\beta})$s.

3. SIMULATION STUDIES

We assess the performance of the estimates under our proposed MODS Designs 1 and 2 for finite samples by conducting extensive simulation experiments. As mentioned earlier, we consider $K = 2$ and generate bivariate failure times, $\boldsymbol{T}_i = (T_{i1}, T_{i2})$ for each subject i in a full cohort, from

$$(6)\quad T_{ik} = \beta_0 + \beta_1 Z_{1ik} + \beta_2 Z_{2ik} + \beta_3 Z_{3ik} + \epsilon_{ik}, \quad k = 1, 2,$$

where Z_{1ik} follows Bernoulli(0.5), $Z_{2ik}, Z_{3ik} \sim \mathcal{N}(0, 1)$, and ϵ_{ik} either follows the standard normal distribution or the Gumbel distribution. We consider a common error distribution with both being the standard normal distributions or distinct error distributions with one following the standard normal distribution ($k = 1$) while the other one following the Gumbel distribution ($k = 2$). We set $\beta_0 = 1$ and $\beta_1 = \beta_2 = \beta_3 = 0.5$. The censoring times are generated from the uniform distribution $[0, c]$ with c being chosen to have the censoring rate of approximately 80%. All the cases are partitioned into three strata by the cutpoints (q_1, q_3) quantiles of failure times, for which we investigate two pairs, (10%, 90%) and (30%, 70%). We first randomly select the SRS of size n_0, set to be either 100 or 200. We then sample the supplemental components of n_1 and n_2 from those remaining in the low stratum (10% or 30%) and the high stratum (70% or 90%), respectively, with different supplemental sampling fractions. Note that the cutpoints for each setting are chosen on the basis of a large population (40,000) and fixed throughout the simulation studies.

For each configuration, we set the full cohort size to be 1,000 and obtain n_0, n_1 and n_2 following MODS Designs 1 and 2. We compare our proposed estimator ($\widehat{\beta}_{MODS}$), with two competing estimators in our simulation study: the estimator based on the SRS portion of the sample from MODS Designs ($\widehat{\beta}_{SRS0}$) and the estimator from a random sample of the same size as the sample from MODS Designs ($\widehat{\beta}_{SRS1}$). The mean of the parameter estimates (Means), the sample standard deviations (SDs), the mean of the estimated standard errors (ESEs) and the efficiencies relative to the MODS (REFF) defined as the ratio of the standard error estimates for $\widehat{\beta}_{SRS0}$ to that of $\widehat{\beta}_{MODS}$ or $\widehat{\beta}_{SRS1}$ to that of $\widehat{\beta}_{MODS}$ are obtained from 2,000 generated data sets.

The simulation results under the different SRS sizes and supplemental sampling fractions for our proposed MODS Designs 1 and 2 are summarized in Tables 1 and 2, respectively. The results in Table 1 suggest that all of the coefficient estimates are approximately unbiased under all the scenarios considered. Our proposed variance estimator seems to provide a good estimate of the true variability. We note that the proposed estimator, $\widehat{\beta}_{MODS}$, is the most efficient among all the estimators in most of the circumstances. The fact that $\widehat{\beta}_{MODS}$ is more efficient than $\widehat{\beta}_{SRS1}$ indicated that our MODS Design 1 is favored over the SRS of the same sample size, even when the cutpoints are further out from (0.3, 0.7) to (0.1, 0.9) and therefore, fewer cases are included. As expected, when the SRS size increases to 200 with other settings being held fixed, the standard errors estimates decrease. We also calculated the REFFs and observed that most of the REFFs are greater than 1, again suggesting that $\widehat{\beta}_{MODS}$ is more efficient among three estimators. Table 2 provides the simulation results when using MODS Design 2 to select the supplemental samples. Note that the sizes of the supplemental samples,(n_1, n_2), increased substantially under MODS Design 2. Overall findings are similar to those in Table 1: estimates for the regression coefficients are virtually unbiased. The standard errors estimates under the three designs considered are in good agreement with the sample standard deviations. Most importantly, MODS Design 2 produce more efficient estimators than an SRS with the same size does. Finally, among all greater-than-one REFFs, we observed more efficiency gains when $n_0 = 100$, the cutpoints = $(0.3, 0.7)$ and the supplemental proportion = 0.5. This implies that a small sample with a smaller fraction of the remaining subjects satisfying the criteria can produce good performance.

Table 1. Simulation results for MODS Design 1 (All Failures Design) with the full cohort size $N = 1000$ and censoring rate $= 80\%$, based on the model $T_{ik} = \beta_0 + \beta_1 Z_{1ik} + \beta_2 Z_{2ik} + \beta_3 Z_{3ik} + \epsilon_{ik}, \quad k = 1, 2$, where $Z_{1ik} \sim Bernoulli(0.5)$, $Z_{2ik}, Z_{3ik} \sim \mathcal{N}(0,1)$, and $\epsilon_{ik} \sim \mathcal{N}(0,1)$

Cutpoints		Mean			SD(ESE)			REFF		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
(0.3, 0.7)	- $n_0 = 100$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (9.29, 16.36)$									
	$\widehat{\beta}_{SRS0}$	0.507	0.513	0.513	0.238(0.242)	0.133(0.126)	0.132(0.127)	1.19	1.18	1.19
	$\widehat{\beta}_{SRS1}$	0.507	0.511	0.512	0.216(0.214)	0.121(0.112)	0.115(0.113)	1.08	1.08	1.05
	$\widehat{\beta}_{MODS}$	0.510	0.513	0.515	0.201(0.202)	0.113(0.106)	0.110(0.105)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.5, 0.5)$, $(n_1, n_2) = (5.95, 10.15)$									
	$\widehat{\beta}_{SRS0}$	0.528	0.523	0.520	0.238(0.240)	0.132(0.126)	0.129(0.126)	1.14	1.19	1.20
	$\widehat{\beta}_{SRS1}$	0.519	0.530	0.521	0.216(0.225)	0.128(0.119)	0.123(0.119)	1.03	1.16	1.15
	$\widehat{\beta}_{MODS}$	0.521	0.519	0.518	0.208(0.202)	0.110(0.106)	0.107(0.107)	1	1	1
	- $n_0 = 200$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (8.45, 14.49)$									
	$\widehat{\beta}_{SRS0}$	0.498	0.511	0.508	0.171(0.167)	0.092(0.090)	0.093(0.089)	1.15	1.19	1.19
	$\widehat{\beta}_{SRS1}$	0.502	0.508	0.510	0.156(0.157)	0.085(0.083)	0.084(0.083)	1.05	1.11	1.08
	$\widehat{\beta}_{MODS}$	0.504	0.509	0.510	0.149(0.144)	0.077(0.076)	0.078(0.076)	1	1	1
	- $n_0 = 200$, supplemental proportion $= (0.5, 0.5)$, $(n_1, n_2) = (5.27, 9.20)$									
	$\widehat{\beta}_{SRS0}$	0.508	0.510	0.511	0.167(0.166)	0.090(0.089)	0.089(0.089)	1.10	1.14	1.12
	$\widehat{\beta}_{SRS1}$	0.517	0.510	0.510	0.159(0.169)	0.087(0.086)	0.088(0.086)	1.05	1.10	1.11
	$\widehat{\beta}_{MODS}$	0.510	0.512	0.511	0.151(0.149)	0.080(0.078)	0.079(0.079)	1	1	1
(0.1, 0.9)	- $n_0 = 100$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (1.96, 4.14)$									
	$\widehat{\beta}_{SRS0}$	0.518	0.515	0.517	0.232(0.237)	0.127(0.126)	0.126(0.125)	1.04	1.05	1.07
	$\widehat{\beta}_{SRS1}$	0.521	0.519	0.516	0.237(0.235)	0.131(0.125)	0.127(0.124)	1.06	1.09	1.08
	$\widehat{\beta}_{MODS}$	0.518	0.514	0.517	0.223(0.225)	0.121(0.118)	0.117(0.118)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.5, 0.5)$, $(n_1, n_2) = (1.26, 2.68)$									
	$\widehat{\beta}_{SRS0}$	0.517	0.517	0.521	0.248(0.244)	0.133(0.129)	0.139(0.128)	1.05	1.06	1.09
	$\widehat{\beta}_{SRS1}$	0.507	0.511	0.512	0.235(0.239)	0.128(0.125)	0.127(0.126)	1.00	1.02	1.00
	$\widehat{\beta}_{MODS}$	0.518	0.517	0.521	0.236(0.231)	0.125(0.121)	0.127(0.119)	1	1	1
	- $n_0 = 200$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (1.73, 3.85)$									
	$\widehat{\beta}_{SRS0}$	0.511	0.514	0.512	0.169(0.168)	0.090(0.089)	0.092(0.090)	1.04	1.06	1.05
	$\widehat{\beta}_{SRS1}$	0.503	0.508	0.506	0.166(0.164)	0.089(0.087)	0.089(0.087)	1.01	1.06	1.02
	$\widehat{\beta}_{MODS}$	0.508	0.513	0.511	0.163(0.160)	0.084(0.083)	0.087(0.084)	1	1	1
	- $n_0 = 200$, supplemental proportion $= (0.5, 0.5)$, $(n_1, n_2) = (1.08, 2.38)$									
	$\widehat{\beta}_{SRS0}$	0.507	0.512	0.507	0.166(0.168)	0.091(0.089)	0.087(0.090)	1.04	1.07	1.06
	$\widehat{\beta}_{SRS1}$	0.506	0.506	0.503	0.162(0.167)	0.089(0.088)	0.092(0.088)	1.01	1.05	1.12
	$\widehat{\beta}_{MODS}$	0.507	0.512	0.507	0.160(0.161)	0.085(0.085)	0.082(0.085)	1	1	1

SD: standard deviation of the parameter estimates; ESE: the mean of the standard error of the estimator; REFF: relative efficiency over MODS; $\widehat{\beta}_{MODS}$ denotes the proposed estimator based on our proposed MODS design; $\widehat{\beta}_{SRS0}$ and $\widehat{\beta}_{SRS1}$ are the standard estimators based on the SRS sample from MODS and the SRS sample with the same size as MODS design, respectively.

Table 3 provides additional simulation results for considering different error distributions, unequal censoring rates and unbalanced supplemental sampling proportions to further examine the robust property of our proposed estimator. We investigated the performance of $\widehat{\beta}_{MODS}$ when the two failure-time distributions ϵ_{i1} and ϵ_{i2} are chosen to follow the standard normal distribution and the Gumbel distribution $(0, 1)$, respectively. In addition to equal supplemental sampling fractions, we consider unequal proportions of two supplemental samples, (0.5, 0.8) and (0.8, 0.5). We also assign different censoring rates, 85% and 75%, to $k = 1$ and $k = 2$, respectively. The results in Table 3 indicate that $\widehat{\beta}_{MODS}$ is

Table 2. Simulation results for MODS Design 2 (At-least-one Failure Design) with the full cohort size $N = 1000$ and censoring rate $= 80\%$, based on the model $T_{ik} = \beta_0 + \beta_1 Z_{1ik} + \beta_2 Z_{2ik} + \beta_3 Z_{3ik} + \epsilon_{ik}, \quad k = 1, 2$, where $Z_{1ik} \sim Bernoulli(0.5)$, $Z_{2ik}, Z_{3ik} \sim \mathcal{N}(0, 1)$, and $\epsilon_{ik} \sim \mathcal{N}(0, 1)$

Cutpoints		Mean			SD(ESE)			REFF		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
(0.3, 0.7)	- $n_0 = 100$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (70.3, 44.6)$									
	$\widehat{\beta}_{SRS0}$	0.513	0.513	0.510	0.247(0.242)	0.132(0.129)	0.129(0.128)	1.67	1.68	1.63
	$\widehat{\beta}_{SRS1}$	0.502	0.511	0.510	0.156(0.161)	0.086(0.085)	0.086(0.086)	1.06	1.10	1.08
	$\widehat{\beta}_{MODS}$	0.506	0.511	0.506	0.148(0.148)	0.078(0.077)	0.079(0.077)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.5, 0.5)$, $(n_1, n_2) = (44.2, 44.4)$									
	$\widehat{\beta}_{SRS0}$	0.515	0.517	0.530	0.238(0.235)	0.132(0.124)	0.129(0.125)	1.69	1.73	1.65
	$\widehat{\beta}_{SRS1}$	0.508	0.514	0.513	0.174(0.172)	0.097(0.093)	0.098(0.092)	1.17	1.24	1.24
	$\widehat{\beta}_{MODS}$	0.502	0.507	0.513	0.149(0.150)	0.078(0.077)	0.079(0.078)	1	1	1
(0.1, 0.9)	- $n_0 = 200$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (24.4, 22.5)$									
	$\widehat{\beta}_{SRS0}$	0.502	0.506	0.506	0.171(0.166)	0.091(0.090)	0.092(0.089)	1.23	1.23	1.23
	$\widehat{\beta}_{SRS1}$	0.507	0.506	0.500	0.153(0.150)	0.082(0.079)	0.082(0.079)	1.10	1.11	1.10
	$\widehat{\beta}_{MODS}$	0.502	0.506	0.505	0.139(0.140)	0.074(0.073)	0.075(0.072)	1	1	1
	- $n_0 = 200$, supplemental proportion $= (0.5, 0.5)$, $(n_1, n_2) = (15.3, 14.2)$									
	$\widehat{\beta}_{SRS0}$	0.518	0.508	0.512	0.168(0.170)	0.089(0.089)	0.090(0.092)	1.14	1.19	1.23
	$\widehat{\beta}_{SRS1}$	0.512	0.503	0.509	0.160(0.155)	0.084(0.082)	0.088(0.083)	1.07	1.12	1.17
	$\widehat{\beta}_{MODS}$	0.514	0.509	0.510	0.149(0.145)	0.075(0.075)	0.075(0.076)	1	1	1

SD: standard deviation of the parameter estimates; ESE: the mean of the standard error of the estimator; REFF: relative efficiency over MODS; $\widehat{\beta}_{MODS}$ denotes the proposed estimator based on our proposed MODS design; $\widehat{\beta}_{SRS0}$ and $\widehat{\beta}_{SRS1}$ are the standard estimators based on the SRS sample from MODS and the SRS sample with the same size as MODS design, respectively.

in consistent performance with those in Tables 1 and 2. This suggests that our proposed method is robust and produces reliable and efficient results when applying to the situation with distinct error distributions within a subject or the unbalanced assignment of supplemental samples and various censoring rates.

4. ANALYSIS OF THE BUSSELTON HEALTH STUDY DATA

The Busselton Health Study is a prospective cohort study conducted in Western Australia over the 17-year period, 1981-1998, with comprehensive surveys in cardiovascular risk factors and disease data along with other experimental collection. The main goal was to examine the association between serum ferritin levels and cardiovascular disease. The study cohort consists of 1,612 men and women aged 40-89 years old who participated in the 1981 Busselton Health Survey and had not previously diagnosed coronary heart disease or ischemic stroke at that time. The participants then were asked to complete a health and lifestyle questionnaire, followed by various measurements and tests. The outcomes of interest were times to first coronary heart disease event and first stroke event, following up from the 1981 survey. Those who left the study or were followed up until December 31, 1988 were considered as censored. Among the 1,612 subjects in the cohort, 285 and 159 experienced coronary heart disease and stroke, respectively. Previous studies [11, 13, 24] have focused on the analyses under case-cohort study designs originally implemented. In order to illustrate our proposed MODS Designs and estimation methods while taking the advantage of this rich data set, we considered triglycerides and systolic blood pressure (SBP) available for the full cohort as the main risks factors. We implemented the following designs as described in the simulation studies: (1) MODS Design 1 (All Failure Design): $n_0 = 200$, supplemental proportion = 80% and cutpoints = (0.3, 0.7); (2) MODS Design 2 (At-least-one Failure Design): $n_0 = 200$, supplemental proportions $= (n_1, n_2) = (80\%, 50\%)$ and cutpoints = (0.3, 0.7). Here we allow different supplemental proportions for MODS Design 2 to illustrate the robustness of our method. The results of fitting the Busselton Health Study Data using two MODS Designs based on 1,000 repetitions are listed in Table 4.

First, as we have observed in the simulation studies, the resulting supplemental sample sizes are much smaller under MODS Design 1 due to the more restrictive criteria. We also note that the effect of triglycerides level was found to be significantly negative at the $\alpha = 0.05$ level, which was agreed by the three estimators in MODS Design 2 - the corresponding 95% confidence interval did not include 0. Under MODS

Table 3. Simulation results with the full cohort size $N = 1000$ and the cutpoints = (30%, 70%), based on the model $T_{ik} = \beta_0 + \beta_1 Z_{1ik} + \beta_2 Z_{2ik} + \beta_3 Z_{3ik} + \epsilon_{ik},\ k = 1, 2$, where $Z_{1ik} \sim Bernoulli(0.5)$, $Z_{2ik}, Z_{3ik} \sim \mathcal{N}(0, 1)$, $\epsilon_{i1} \sim \mathcal{N}(0, 1)$ and $\epsilon_{i2} \sim Gumbel(0, 1)$

Censoring proportion		Mean			SD(ESE)			REFF		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
	MODS Design 1									
$(0.8, 0.8)$	- $n_0 = 100$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (6.55, 10.9)$									
	$\widehat{\beta}_{SRS0}$	0.528	0.518	0.524	0.248(0.253)	0.131(0.137)	0.132(0.135)	1.14	1.13	1.16
	$\widehat{\beta}_{SRS1}$	0.517	0.522	0.521	0.228(0.234)	0.121(0.125)	0.122(0.124)	1.05	1.04	1.07
	$\widehat{\beta}_{MODS}$	0.526	0.517	0.522	0.217(0.222)	0.114(0.121)	0.115(0.116)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.8, 0.5)$, $(n_1, n_2) = (6.67, 6.77)$									
	$\widehat{\beta}_{SRS0}$	0.510	0.526	0.523	0.250(0.253)	0.133(0.142)	0.132(0.136)	1.12	1.14	1.13
	$\widehat{\beta}_{SRS1}$	0.523	0.515	0.522	0.233(0.253)	0.123(0.128)	0.123(0.128)	1.06	1.03	1.06
	$\widehat{\beta}_{MODS}$	0.506	0.522	0.523	0.223(0.225)	0.117(0.124)	0.117(0.121)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.5, 0.8)$, $(n_1, n_2) = (4.20, 10.7)$									
	$\widehat{\beta}_{SRS0}$	0.528	0.525	0.524	0.247(0.252)	0.132(0.142)	0.133(0.134)	1.12	1.15	1.18
	$\widehat{\beta}_{SRS1}$	0.521	0.516	0.513	0.231(0.236)	0.122(0.128)	0.122(0.130)	1.05	1.04	1.14
	$\widehat{\beta}_{MODS}$	0.525	0.523	0.521	0.217(0.225)	0.115(0.123)	0.116(0.114)	1	1	1
$(0.85, 0.75)$	- $n_0 = 100$, supplemental proportion $= (0.8, 0.5)$, $(n_1, n_2) = (3.93, 7.39)$									
	$\widehat{\beta}_{SRS0}$	0.542	0.547	0.544	0.236(0.242)	0.123(0.128)	0.123(0.128)	1.08	1.12	1.13
	$\widehat{\beta}_{SRS1}$	0.543	0.544	0.545	0.226(0.226)	0.118(0.116)	0.118(0.117)	1.01	1.02	1.03
	$\widehat{\beta}_{MODS}$	0.539	0.546	0.544	0.214(0.224)	0.110(0.114)	0.110(0.113)	1	1	1
	MODS Design 2									
$(0.8, 0.8)$	- $n_0 = 100$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (56.7, 63.5)$									
	$\widehat{\beta}_{SRS0}$	0.526	0.521	0.523	0.248(0.252)	0.131(0.136)	0.131(0.136)	1.77	1.84	1.84
	$\widehat{\beta}_{SRS1}$	0.511	0.508	0.510	0.164(0.165)	0.088(0.090)	0.088(0.088)	1.15	1.22	1.19
	$\widehat{\beta}_{MODS}$	0.513	0.510	0.510	0.144(0.143)	0.074(0.074)	0.074(0.074)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.8, 0.5)$, $(n_1, n_2) = (57.2, 39.8)$									
	$\widehat{\beta}_{SRS0}$	0.521	0.528	0.517	0.247(0.253)	0.132(0.141)	0.130(0.137)	1.60	1.71	1.76
	$\widehat{\beta}_{SRS1}$	0.511	0.511	0.511	0.174(0.172)	0.093(0.094)	0.093(0.097)	1.09	1.14	1.25
	$\widehat{\beta}_{MODS}$	0.505	0.514	0.510	0.152(0.158)	0.078(0.082)	0.079(0.078)	1	1	1
	- $n_0 = 100$, supplemental proportion $= (0.5, 0.8)$, $(n_1, n_2) = (36.4, 63.2)$									
	$\widehat{\beta}_{SRS0}$	0.520	0.516	0.520	0.246(0.248)	0.131(0.140)	0.132(0.140)	1.64	1.82	1.79
	$\widehat{\beta}_{SRS1}$	0.505	0.512	0.509	0.172(0.176)	0.092(0.094)	0.092(0.094)	1.16	1.22	1.20
	$\widehat{\beta}_{MODS}$	0.508	0.510	0.511	0.146(0.152)	0.076(0.077)	0.076(0.078)	1	1	1
$(0.85, 0.75)$	- $n_0 = 100$, supplemental proportion $= (0.5, 0.5)$,$(n_1, n_2) = (48.1, 46.0)$									
	$\widehat{\beta}_{SRS0}$	0.546	0.546	0.546	0.240(0.246)	0.124(0.125)	0.124(0.129)	1.54	1.62	1.62
	$\widehat{\beta}_{SRS1}$	0.534	0.536	0.537	0.169(0.170)	0.088(0.088)	0.088(0.090)	1.06	1.13	1.13
	$\widehat{\beta}_{MODS}$	0.535	0.539	0.537	0.157(0.159)	0.079(0.077)	0.079(0.080)	1	1	1

Design 1, however, the significance of triglycerides was detected only by the proposed estimator, not in the other competing estimators. On the other hand, the SBP effect was not statistically significant under both designs and for all three methods. We also note that our proposed estimator is the most efficient one in both MODS Designs 1 and 2 and produced the narrowest confidence intervals among three. Clearly, $\widehat{\beta}_{MODS}$ yielded relatively smaller standard error es-

Table 4. Analysis of the Busselton Health Study

MODS Design		Estimates		SE (95% CI)	
		SBP	Triglycerides	SBP	Triglycerides
1	- $n_0 = 200$, supplemental proportion $= (0.8, 0.8)$, $(n_1, n_2) = (6.87, 5.19)$				
	$\widehat{\beta}_{SRS0}$	−0.087	−0.388	0.147 (−0.375, 0.202)	0.200 (−0.780, 0.004)
	$\widehat{\beta}_{SRS1}$	−0.075	−0.401	0.129 (−0.328, 0.179)	0.298 (−0.985, 0.183)
	$\widehat{\beta}_{MODS}$	−0.072	−0.396	0.124 (−0.314, 0.171)	0.179 (−0.747, −0.044)
2	- $n_0 = 200$, supplemental proportion $= (0.8, 0.5)$, $(n_1, n_2) = (83.9, 55.4)$				
	$\widehat{\beta}_{SRS0}$	−0.086	−0.393	0.147 (−0.373, 0.202)	0.194 (−0.772, −0.013)
	$\widehat{\beta}_{SRS1}$	−0.060	−0.350	0.110 (−0.277, 0.156)	0.122 (−0.589, −0.110)
	$\widehat{\beta}_{MODS}$	−0.044	−0.364	0.069 (−0.180, 0.092)	0.105 (−0.570, −0.158)

Cutpoints: (0.3, 0.7); CI: confidence interval.

timates in both risk factors and designs than $\widehat{\beta}_{SRS1}$ did. Moreover, we considered different supplemental sampling portions for n_1 and n_2 under MODS Design 2. Although the standard error estimates are substantially smaller in MODS Design 2 compared with those in MODS Design 1 due to the larger sample sizes in supplemental portions, the estimates for the effect of triglycerides do not differ much. This indicates that our proposed estimators are robust to the designs and unbalanced supplemental sampling proportions.

5. DISCUSSION

We proposed a general multivariate-ODS design for the multiple disease outcomes and clustered survival failure-time data under the framework of the semiparametric AFT model and answered a much needed call for cost-efficient studies. By taking advantages of a multivariate-ODS scheme [29] and recent advances in computing the estimators using the generalized estimating procedures with the induced smoothing approach and the sandwich variance estimator [9, 10], we established new statistical inference procedures. The main advantage of our proposed general failure-time MODS Designs is that researchers can select not only a simple random sample but also different supplemental samples with various criteria to further improve the study efficiency, especially when the disease rate is low. Moreover, the proposed MODS Designs 1 and 2 (All Failures v.s. At-least-one Failure) offer a flexibility for researchers in implementing multivariate-ODS designs in practice.

The proposed estimators are shown to be consistent and asymptotically normal. In the simulation studies, the results suggested that both MODS Designs 1 and 2 worked well. Our proposed estimator is more efficient than the estimator based only on the SRS portion of the sample from MODS Designs and the estimator based on an SRS of the same size as the sample from MODS Designs.

In general, the clusters composed only of failures would be most informative. So we recommend using such clusters as supplemental components (MODS Design 1). As the censoring rate gets higher, however, sizes of such clusters get smaller. They would get even smaller with a larger number of members in a cluster. Due to this low sample size, gains in efficiency from using the proposed MODS Design 1 could be challenging as not many clusters are likely to be selected into the supplemental samples. In these cases, we recommend using clusters containing at least one failure as supplemental components (MODS Design 2). Such clusters might not be as informative as those containing only failures but it is relatively much easier to identify such clusters. Moreover, our simulation studies demonstrated that our proposed estimator based on such MODS Designs still leads to a substantial improvement in efficiency. Study designs having a similar spirit are not uncommon. A case-control-family study [17, 19] is an example; cases and controls, known as the probands, are sampled followed by subsequent samplings of relatives for each case and control with an ascertainment of information on covariates and outcomes. This design can be viewed as one of our proposed MODS Designs in which a cluster is formed by a case or control proband and his/her relatives. Clusters containing case probands comprise supplemental components where a cluster contains at least one failure - case proband - but does not necessarily contain failures only.

Extended development of multivariate failure time data via a multivariate-ODS scheme under additive hazards model or other survival models will be possible directions for future studies. Another important topic for future studies is to develop model-checking and goodness-of-fit procedures for data from a multivariate failure-time ODS design under a particular survival model.

APPENDIX A. APPENDIX SECTION

In this section, we provide proofs of the consistency and asymptotic normality of the proposed estimator $\hat{\beta}$. We assume the following conditions:

1. The parameter space $\mathbb{B}$ containing β_0 is a compact set of $\mathbb{R}^p$.
2. $\sum_{k=1}^{K} \|Z_{ik}\| + K$ is bounded almost surely by a nonrandom constant $(i = 1, \ldots, N)$.
3. $\mathrm{Var}(\epsilon_{11}) < \infty$.
4. The matrix $\tilde{A}(\beta_0) = \lim_{N\to\infty} \partial U(\beta_0)/\partial\beta_0^\top$ is nonsingular.
5. Let $f_0(\cdot)$ denote the marginal density associated with model error term ϵ_{11}. Then, $f_0(\cdot)$ and $f_0'(\cdot)$ are bounded functions on $\mathbb{R}$ with
$$\int_{\mathbb{R}} \left\{ \frac{f_0'(t)}{f_0(t)} \right\}^2 f_0(t)\mathrm{d}t < \infty$$
6. The marginal distribution of C_{ik} is absolutely continuous and has a bounded density $g_{ik}(\cdot)$ on $\mathbb{R}$ for $i = 1, \ldots, N$ and $k = 1, \ldots, K$).
7. As $N \to \infty$, $p_m \to \tilde{p}_m (0 < \tilde{p}_m < 1)$ and $r_m = n_m/(N_m - n_{0,m}) \to \tilde{r}_m (0 < \tilde{r}_m < 1)$ for $m = 0, 1, 2$.

Conditions 1 - 6 are identical to those imposed by Chiou, Kang and Yan [10]. Condition 7 is added to ensure the desired asymptotic convergence of the ODS samples.

We first provide a proof of the consistency of $\hat{\beta}$. Let $\phi(\cdot)$ denote the density function of the standard normal random variable. The convex objective functions of $\tilde{U}(\beta)$ and $\tilde{U}_c(\beta)$, $\tilde{L}(\beta)$ and $\tilde{L}_c(\beta)$, respectively, are then

$$\tilde{L}(\beta) = \frac{1}{N(N-1)} \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} \Delta_{ik} \Bigg[\{e_{jl}(\beta) - e_{ik}(\beta)\} \times \Phi\left\{\sqrt{N}\kappa_{ijkl}(\beta)\right\} + \frac{r_{ikjl}}{\sqrt{N}} \phi\left\{\sqrt{N}\kappa_{ijkl}(\beta)\right\} \Bigg],$$

$$\tilde{L}_c(\beta) = \frac{1}{N(N-1)} \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} w_i w_j \Delta_{ik} \times \Bigg[\{e_{jl}(\beta) - e_{ik}(\beta)\} \Phi\left\{\sqrt{N}\kappa_{ijkl}(\beta)\right\} + \frac{r_{ikjl}}{\sqrt{N}} \phi\left\{\sqrt{N}\kappa_{ijkl}(\beta)\right\} \Bigg],$$

where $\kappa_{ijkl}(\beta) = \dfrac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ijkl}}$.

By applying Lemma 2 in Johnson and Strawdermann [22], $\lim_{N\to\infty} \sup_{\beta\in\mathbb{B}} |\tilde{L}(\beta) - L_0(\beta)| = 0$ where $L_0(\beta)$ is strictly convex for $\beta \in \mathbb{B}$. It can also be shown that $\lim_{N\to\infty} \sup_{\beta\in\mathbb{B}} |\tilde{L}_c(\beta) - \tilde{L}(\beta)| = 0$ by the strong law of large numbers for U-statistics [32], asymptotic convergence results on finite population sampling [18], and Lemma 1 in Kong, Cai and Sen [25]. Combining these two results and by applying the triangle inequality, $\lim_{N\to\infty} \sup_{\beta\in\mathbb{B}} |\tilde{L}_c(\beta) - L_0(\beta)| = 0$. Condition 4 ensures that $L_0(\beta)$ is strictly convex at β_0, a unique minimizer of $L_0(\beta)$. Then, the unique minimizer of $L_c(\beta)$, $\hat{\beta}$, converges to β_0 almost surely [1, Corollary II.2].

To establish the asymptotic normality of $\hat{\beta}$, we first show the asymptotic normality of $\tilde{\beta}$, solution to $U_c(\beta) = 0$ where $U_c(\beta)$ is the weighted version of $U(\beta)$ and

$$U_c(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{K} w_i w_j \Delta_{ik} (\boldsymbol{Z}_{ik} - \boldsymbol{Z}_{jl}) \times I\{e_{jl}(\beta) \geq e_{ik}(\beta)\}.$$

Then, we show the asymptotic equivalence between the distributions of $\sqrt{N}(\tilde{\beta} - \beta_0)$ and $\sqrt{N}(\hat{\beta} - \beta_0)$.

Let $M_{ik}(\beta; t) = N_{ik}(\beta; t) - \int_{-\infty}^{t} I\{e_{ik}(\beta) \geq u\} \lambda_0(u)du$ where $\lambda_0(\cdot)$ is the common hazard function for $\epsilon_{ik}s$. Define $\boldsymbol{S}^{(d)}(\beta; t) = N^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K} \boldsymbol{Z}_{ik}^{\otimes d} I\{e_{ik}(\beta) \geq t\}$ and $\boldsymbol{S}_c^{(d)}(\beta; t) = N^{-1}\sum_{i\equiv1}^{N}\sum_{k=1}^{K} w_i \boldsymbol{Z}_{ik}^{\otimes d} I\{e_{ik}(\beta) \geq t\}$ $(d = 0, 1)$. Further, define $\bar{\boldsymbol{Z}}(\beta; t) = \boldsymbol{S}^{(1)}(\beta; t)/S^{(0)}(\beta; t)$ and $\bar{\boldsymbol{Z}}_c(\beta; t) = \boldsymbol{S}_c^{(1)}(\beta; t)/S_c^{(0)}(\beta; t)$. The limiting quantities of $\boldsymbol{S}^{(d)}(\beta; t)$ and $\bar{\boldsymbol{Z}}(\beta; t)$ are $\boldsymbol{s}^{(d)}(\beta; t)$ and $\bar{\boldsymbol{z}}(\beta; t) = \boldsymbol{s}^{(1)}(\beta; t)/s^{(0)}(\beta; t)$, respectively.

By applying Lemma 1 of Yu et al. [37] and Lemma 1 of Jin, Lin and Ying [21] to the stochastic integral representation of $U_c(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \int_{-\infty}^{\infty} w_i S_c^{(0)}(\beta; t) \{\boldsymbol{Z}_{ik} - \bar{\boldsymbol{Z}}_c(\beta; t)\} dN_{ik}(\beta; t)$, it can be shown that

$$U_c(\beta_0) = \sum_{i=1}^{N}\sum_{k=1}^{K} \boldsymbol{u}_{ik}(\beta_0) + \sum_{i=1}^{N}\sum_{k=1}^{K} (w_i - 1)\boldsymbol{u}_{ik}(\beta_0) + o_p(\sqrt{N}) \tag{7}$$

where $\boldsymbol{u}_{ik}(\beta) = \int_{-\infty}^{\infty} s^{(0)}(\beta; t) \{\boldsymbol{Z}_{ik} - \bar{\boldsymbol{z}}(\beta; t)\} dM_{ik}(\beta; t)$. Since

$$w_i - 1 = \prod_{k=1}^{K}(1 - \Delta_{ik}) \left(\frac{\xi_i}{p_0} - 1 \right) + (1 - \zeta_i) \left\{ 1 - \prod_{k=1}^{K}(1 - \Delta_{ik}) \right\} \left(\frac{\xi_i}{p_0} - 1 \right) + (1 - \xi_i) \left\{ 1 - \prod_{k=1}^{K}(1 - \Delta_{ik}) \right\} \sum_{m=1}^{M} \zeta_{im} \left(\frac{\eta_{im}}{r_m} - 1 \right),$$

the second term in (7) is decomposed into

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\prod_{k=1}^{K}(1 - \Delta_{ik}) \left(\frac{\xi_i}{p_0} - 1 \right) \boldsymbol{u}_{ik}(\beta_0) + \sum_{i=1}^{N}\sum_{k=1}^{K}(1 - \zeta_i) \times \left\{ 1 - \prod_{k=1}^{K}(1 - \Delta_{ik}) \right\} \left(\frac{\xi_i}{p_0} - 1 \right) \boldsymbol{u}_{ik}(\beta_0) + \sum_{i=1}^{N}\sum_{k=1}^{K}(1 - \xi_i) \times \left\{ 1 - \prod_{k=1}^{K}(1 - \Delta_{ik}) \right\} \sum_{m=1}^{M} \zeta_{im} \left(\frac{\eta_{im}}{r_m} - 1 \right) \boldsymbol{u}_{ik}(\beta_0).$$

These three terms are asymptotically uncorrelated. Moreover, the first term in (7) and these three terms are asymptotically uncorrelated. Thus, by applying Lemma 3 in the

supplementary materials of Kang and Cai [23] and the multivariate central limit theorem, we have the desired asymptotic normality of $\sqrt{N}^{-1}U_c(\beta_0)$ whose mean is 0 and asymptotic covariance function is $\Sigma_F(\beta_0)+\Sigma_S(\beta_0)$ where $\Sigma_F(\beta_0) = \mathrm{E}\left[\sum_{k=1}^K \boldsymbol{u}_{ik}(\beta_0)\right]^{\otimes 2}$ and $\Sigma_S(\beta_0) = \frac{1-\tilde{p}_0}{\tilde{p}_0}\mathrm{Var}\left[\prod_{l=1}^K(1-\Delta_{1l})\sum_{k=1}^K \boldsymbol{u}_{1k}(\beta_0)\right] + \frac{1-\tilde{p}_0}{\tilde{p}_0}\mathrm{Var}\left[\left\{1-\prod_{l=1}^K(1-\Delta_{1l})\right\}(1-\zeta_i)\sum_{k=1}^K \boldsymbol{u}_{1k}(\beta_0)\right] + \sum_{m=1}^M(1-\tilde{p}_0)\frac{1-\tilde{r}_m}{\tilde{r}_m}\times \mathrm{Var}\left[\left\{1-\prod_{l=1}^K(1-\Delta_{1l})\right\}\zeta_1\sum_{k=1}^K \boldsymbol{u}_{1k}(\beta_0)\right]$. The consistency of $\tilde{\beta}$ to β_0 follows from the similar arguments of showing the consistency of $\hat{\beta}$. Using this with the arguments in Theorem 2 of Ying [36], it can be shown that $\sqrt{N}(\tilde{\beta}-\beta_0) = -\tilde{A}^{-1}(\beta_0)\sqrt{N}^{-1}U_c(\beta_0)+o_p(1+\sqrt{N}\|\tilde{\beta}-\beta_0\|)$. Then, by incorporating the asymptotic normality of $\sqrt{N}^{-1}U_c(\beta_0)$, $\sqrt{N}^{-1}(\tilde{\beta}-\beta_0)$ is asymptotically normally distributed with mean 0 and covariance function $\tilde{A}^{-1}(\beta_0)\{\Sigma_F(\beta_0)+\Sigma_S(\beta_0)\}\tilde{A}^{-1}(\beta_0)$,

To establish the equivalence of the distributions of $\sqrt{N}(\tilde{\beta}-\beta_0)$ and $\sqrt{N}(\hat{\beta}-\beta_0)$ asymptotically, it is sufficient to show that, as $N\to\infty$, (i) $\partial\tilde{U}_c(\beta)/\partial\beta^\top$ converges to $\tilde{A}(\beta)$ in probability uniformly in $\beta\in\mathbb{B}$, and (ii) $\sqrt{N}^{-1}\left\{\tilde{U}_c(\beta)-U_c(\beta)\right\}$ converges to $\tilde{A}(\beta)$ in probability uniformly in $\beta\in\mathbb{B}$. To show (i), we decompose $\partial\tilde{U}_c(\beta)/\partial\beta^\top-\tilde{A}(\beta)$ into $\left\{\partial\tilde{U}_c(\beta)/\partial\beta^\top-\partial\tilde{U}(\beta)/\partial\beta^\top\right\}+\left\{\partial\tilde{U}(\beta)/\partial\beta^\top-\tilde{A}(\beta)\right\}$. The second term converges to 0 in probability uniformly in $\beta\in\mathbb{B}$ by Lemma 3 in Johnson and Strawdermann [22]. The first term can also be shown to converge to 0 in probability uniformly in $\beta\in\mathbb{B}$ by applying the strong law of large numbers for U-statistics [32], Lemma 1 in Kong, Cai and Sen [25], and the asymptotic convergence results on finite sampling [18]. Combining these two and by applying the triangle inequality, we have the desired result.

For (ii),

$$\begin{aligned}
&\sqrt{N}^{-1}\left\{\tilde{U}_c(\beta)-U_c(\beta)\right\}\\
&=\frac{1}{N(N-1)}\sum_{i=1}^N\sum_{k=1}^K\sum_{j=1}^N\sum_{l=1}^K w_iw_j\Delta_{ik}(\boldsymbol{Z}_{ik}-\boldsymbol{Z}_{jl})\frac{1}{\kappa_{ijkl}(\beta)}\\
&\quad\times\sqrt{N}\kappa_{ijkl}(\beta)\left[\Phi\left\{\sqrt{N}\kappa_{ijkl}(\beta)\right\}-I(\kappa_{ijkl}(\beta)\geq 0)\right]\\
&=\frac{1}{N(N-1)}\sum_{i=1}^N\sum_{k=1}^K\sum_{j=1}^N\sum_{l=1}^K w_iw_j\Delta_{ik}(\boldsymbol{Z}_{ik}-\boldsymbol{Z}_{jl})\frac{1}{\kappa_{ijkl}(\beta)}\\
&\quad\sqrt{N}\kappa_{ijkl}(\beta)\Phi\left\{-\sqrt{N}\kappa_{ijkl}(\beta)\right\}.
\end{aligned}$$

Note that, for $u\in\mathbb{R}$, $\left|u\left\{\Phi(\sqrt{N}u)-I(u\geq 0)\right\}\right| = \mathrm{sign}(u)\left\{u\Phi(\sqrt{N}^{-1}|u|)\right\}$ where $\mathrm{sign}(u) = 2I(u\geq 0)-1$. Since $\Phi(-u)\leq(\sqrt{2\pi}u)^{-1}\exp(-u^2/2)$, $\lim_{N\to\infty}\sup_{u\in\mathbb{R}}\left|u\left\{\Phi(\sqrt{N}u)-I(u\geq 0)\right\}\right| = 0$. Then, (ii) follows from this result and by applying the strong law of large numbers for U-statistics [32], Lemma 1 in Kong, Cai and Sen [25], and the asymptotic convergence results on finite sampling [18].

ACKNOWLEDGEMENTS

We are grateful to Professor Matthew Knuiman and the Busselton Population Medical Research Foundation for permission to use the data for application. We also thank Dr. Jianwen Cai for her helpful comments. This research was partly supported by the Ministry of Science and Technology in Taiwan grant (108-2118-M-003-001-MY2) for Dr. Lu, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2017R1A2B4005818) for Dr. Kang and US National Institutes of Health grants (P01-CA142538 and P30-ES010126) for Dr. Zhou.

Received 20 May 2019

REFERENCES

[1] Andersen, P., and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, **10**, 1100–1120. MR0673646

[2] Barlow, W. (1994). Robust variance estimation for the case-cohort design. *Biometrics*, **50**, 1064–1072.

[3] Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, **34**, 86–102. MR2325244

[4] Brown, B. M. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, **92**, 149–158. MR2158616

[5] Cai, J and Zeng, D. (2004). Sample size/power calculation for case-cohort studies. *Biometrics*, **60**, 1015–1024. MR2133554

[6] Cai, J and Zeng, D. (2007). Power calculation for case-cohort studies with nonrare events. *Biometrics*, **63**, 1288–1295. MR2414609

[7] 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. MR1965682

[8] Chen, K. (2001). Generalized case-cohort sampling. *Journal of the Royal Statistical Society, Series B*, **63**, 791–809. MR1872067

[9] Chiou, S. H., Kang, S. and Yan, J. (2014). Fitting accelerated failure time models in routine survival analysis with R package aftgee. *Journal of Statistical Software*, **61**, 1–23.

[10] Chiou, S. H., Kang, S. and Yan, J. (2015). Semiparametric accelerated failure time modeling for clustered failure times from stratified sampling. *Journal of the American Statistical Association*, **110**, 621–629. MR3367252

[11] Cook, J., Lipschitz, D., Miles, L. and Finch, C. (1974). Serum ferritin as a measure of iron stores in normal subjects. *The American Journal of Clinical Nutrition*, **27**, 681–687.

[12] Cox, D. R. (1975). Partial Likelihood. *Biometrics*, **62**, 269–276. MR0400509

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

[14] Ding, J., Lu, T. S. and Zhou, H. (2013). Outcome-dependent selection models. *Encyclopedia of Environmetrics*, **v4**.

[15] Ding, J., Zhou, H., Liu, Y., Cai, J. and Longnecker, M. (2014). Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. *Biometrics*, **15**, 636–650.
[16] Ding, J., Lu, T. S., Cai, J. and H. Zhou. (2017). Recent progresses in outcome-dependent sampling with failure time data. *Lifetime Data Analysis*, **23**, 57–82. MR3601684
[17] Gorfine, M., Bordo, N. and Hsu, L. (2017). A fully non-parametric estimator of the marginal survival function based on case–control clustered age-at-onset data. *Biostatistics*, **18**, 76–90. MR3612275
[18] Hájek, J. (1960), Limiting Distributions in Simple Random Sampling from a Finite Population. *Pub. Math. Inst. Hungar. Acad. Sci.*, **5**, 361–374. MR0125612
[19] Hsu, L., Chen, L., Gorfine, M. and Malone, K. (2004). Semi-parametric estimation of marginal hazard function from case–control family studies. *Biometrics*, **60**, 936–944. MR2133546
[20] Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, **90**, 341–353. MR1986651
[21] Jin, Z., Lin, D. Y., and Ying, Z. (2006). Rank regression analysis of multivariate failure time data based on marginal linear models. *Scandinavian Journal of Statistic*, **33**, 1–23. MR2255106
[22] Johnson, L. M. and Strawderman, R. L. (2009). Induced smoothing for the semiparametric accelerated failure time model: asymptotic and extensions to clustered data. *Biometrika*, **90**, 341–327. MR2538758
[23] Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, **96**, 887–901. MR2767277
[24] Knuiman, M. W., Divitini, M. L., Olynyk, J. K., Cullen, D. J. and Bartholomew, H. C. (2003). Serum ferritin and cardiovascular disease: A 17-year-follow-up study in Busselton, Western Australia. *American Journal Statistical of Epidemiology*, **158**, 144–149.
[25] Kong, L., Cai, J., and Sen, P. K. (2006). Asymptotic results for fitting semiparametric transformation models to failure fime data from case-cohort studies. *Statistica Sinica*, **16**, 155–151. MR2256083
[26] Kulich, M. and Lin, D. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, **99**, 832–844. MR2090916
[27] Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, **88**, 1341–1349. MR1245368
[28] Lu, S. and Shih, J. H. (2006). Case-cohort designs and analysis for clustered failure time data. *Biometrics*, **62**, 1138–1148. MR2307439
[29] Lu, T. S., Longnecker, M. and Zhou, H. (2017). Statistical inferences for data from studies conducted with an aggregated multivariate outcome-dependent sample design. *Statistics in Medicine*, **36**, 985–997. MR3606658
[30] Prentice, R. L. (1986). A case-cohort design for epidemiologic studies and disease prevention trials. *Biometrika*, **73**, 1–11.
[31] Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, **16**, 64–81. MR0924857
[32] Serfling, R. (2009). *Approximation Theorems of Mathematical Statistics*, (Vol. 162). John Wiley & Sons. MR0595165
[33] Song, R., Zhou, H. and Kosorok, M. R. (2009). On semiparametric efficient inference for two-stage outcome dependent sampling with a continuous outcome. *Biometrics*, **96**, 221–228. MR2482147
[34] Sun, J., Sun, L. and Flournoy, N. (2004). Additive hazards model for competing risks analysis of the case-cohort design. *Communications in Statistics - Theory and Methods*, **33**, 351–366. MR2045320
[35] 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. MR2160550
[36] Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics*, **21**, 76–99. MR1212167
[37] Yu, J., Liu, Y., Sandler, D. P. and Zhou, H. (2015). Statistical inference for the additive hazards model under outcome-dependent sampling. *The Canadian Journal of Statistics*, **43**, 436–453. MR3388326
[38] Zhang, H., Schaubel, D. E. and Kalbfleisch, J. D. (2011). Proportional hazards regression for the analysis of clustered survival data from case-cohort studies. *Biometrics*, **67**, 18–28. MR2898813
[39] Zhou, H. and Weaver, M. (2001). Outcome-dependent selection models. *Encyclopedia of Environmetrics*, **v3**, 1499–1502.
[40] Zhou, H., Weaver, M. A., 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. MR1908182
[41] Zhou, H., Chen, W., Rissanen T., Korrick, S., Hu, H., Salonen, J., and Longnecker, M. (2007). Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, **18**, 461–468.
[42] 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. MR2898831

Tsui-Shan Lu
Department of Mathematics
National Taiwan Normal University
Taipei, Taiwan
E-mail address: tslu@ntnu.edu.tw

Sangwook Kang
Department of Applied Statistics
Yonsei University
Seoul, Korea
E-mail address: kanggi1@yonsei.ac.kr

Haibo Zhou
Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27514
E-mail address: zhou@bios.unc.edu