Lifetime Data Anal (2017) 23:57–82
DOI 10.1007/s10985-015-9355-7

Recent progresses in outcome-dependent sampling with failure time data

Jieli Ding[1] · Tsui-Shan Lu[2] · Jianwen Cai[3] · Haibo Zhou[3]

Received: 30 July 2015 / Accepted: 22 December 2015 / Published online: 13 January 2016

Abstract An outcome-dependent sampling (ODS) design is a retrospective sampling scheme where one observes the primary exposure variables with a probability that depends on the observed value of the outcome variable. When the outcome of interest is failure time, the observed data are often censored. By allowing the selection of the supplemental samples depends on whether the event of interest happens or not and oversampling subjects from the most informative regions, ODS design for the time-to-event data can reduce the cost of the study and improve the efficiency. We review recent progresses and advances in research on ODS designs with failure time data. This includes researches on ODS related designs like case–cohort design, generalized case–cohort design, stratified case–cohort design, general failure-time ODS design, length-biased sampling design and interval sampling design.

Keywords Case–cohort design · ODS design · Failure time data

✉ Haibo Zhou
zhou@bios.unc.edu

Jieli Ding
jlding.math@whu.edu.cn

Tsui-Shan Lu
tslu@math.ntnu.edu.tw

Jianwen Cai
cai@bios.unc.edu

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

[2] Department of Mathematics, National Taiwan Normal University, Taipei 116, Taiwan

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

1 Introduction of original ODS design

Epidemiologic and biomedical observational studies that relate outcome of interest to individual exposure and other characteristics play a key role in understanding the determinants of diseases in humans. In many such studies, the major budget and cost typically arise from the assembling of primary exposure variables. Large cohort studies with simple random sampling are often too expensive to conduct for investigators with a limited budget. To reduce the cost and to achieve a prespecified power level, alternative cost-effective designs and procedures are thus desirable for studies with a limited budget.

An *outcome-dependent sampling (ODS) design* is a general term that describes a retrospective sampling scheme where one observes the primary exposure variables with a probability that depends on the observed value of the outcome variable. It is a useful, and more importantly, a cost-effective alternative to the more standard random sampling design. Under an ODS design, information on the primary exposure variables is assembled only for a sample that is selected from the underlying cohort population in a manner other than simple random sampling. The principal idea of an ODS design is to concentrate resources where there is the greatest amount of information. By allowing the selection probability of each individual in the ODS sample to depend on the outcome, the investigators attempt to enhance the efficiency and reduce the cost of a study. The well-known example is the case–control study in epidemiology, which is an ODS scheme for a binary outcome variable (Cornfield 1951).

Although in some applications the outcome event is intrinsically binary/categorical, there are many situations in which the outcome variable is actually measured continuously (e.g. failure time). One commonly used approach in epidemiologic studies is to dichotomize the continuous outcome and use the available methods for binary outcome (White 1982; Breslow and Cain 1988; Scott and Wild 1991; Weinberg and Wacholder 1993; Schill et al. 1993; Breslow and Holubkov 1997; Breslow et al. 2003). Another commonly used approach in these situations is to stratify the range of the continuous outcome variable and then sampling observations according to stratum-specific selection probabilities (Imbens and Lancaster 1996; Lawless et al. 1999).

Recent work has focused on a more general ODS design for continuous outcomes. Such an ODS design usually includes a simple random sample from the underlying population and some additional supplemental samples which are determined by the scales of outcome. The advantage of such an ODS design is that, while providing overall information about the population, it allows the investigators to target sample certain regions of the population that are believed to be more informative. There are very active researches on such sampling schemes. Weaver (2001) and Zhou et al. (2002) developed a semiparametric empirical likelihood inference procedures. Weaver and Zhou (2005) proposed a maximum estimated likelihood estimation approach. Chatterjee et al. (2003), Song et al. (2009), and Zhou et al. (2011b) developed inferential methodologies for the two-stage ODS design, which make efficient use of any additional outcome data that may be available for the entire study population. Qin and Zhou (2011) and Zhou et al. (2011a, d) studied the inference procedures for ODS design under the partial linear models. Schildcrout and Heagerty (2008), Schildcrout and Rathouz (2010), and Schildcrout et al. (2012) discussed ODS design and proposed analysis approaches for longitudinal data. Ding et al. (2012) developed a regression analysis under an ODS design for a missing data problem. A useful extension of ODS design is developed by allowing the selection probability to depend on not only the outcome but also an auxiliary variable, which is referred as outcome and auxiliary-dependent subsampling (OADS). Wang and Zhou (2006, 2010) and Zhou et al. (2011c) considered and proposed inference procedures for data from the OADS sampling scheme.

In this paper, we review recent progresses for ODS design with failure time data. In the rest of the paper, Sect. 2 reviews the research work in univariate failure time data case. Section 3 reviews the work in multivariate failure time case.

2 ODS designs with a univariate failure time

In the above literature, studies were based on that data on outcome variables are completely observed. When the outcome of interest is failure time, the observed data are often censored. In this section, we introduce some biased-sampling schemes for failure time data and review the developments of the inferential methodologies for these biased-sampling studies.

2.1 Case–cohort design and its variations

For the time-to-event data, the *case–cohort design* (Prentice 1986) is one of the most widely used biased-sampling scheme for censored failure-time data. The key idea of this study design is to obtain the measurements of primary exposure variables only on a subset of the entire cohort (subcohort) and all the subjects who experience the event of interest (cases) in the cohort. Thus, the case–cohort study designs are particularly useful for large-scale cohort studies with a low disease rate or for cohort studies with exposure expensive to measure.

The requirement of sampling all the cases in the original case–cohort design will limit the application of case–cohort study designs since some diseases might not be rare. In such cases, a *generalized case–cohort design* have been proposed where, in addition to a subcohort, the information on exposure is assembled only for a subset of the failures instead of all the failures to reduce the cost. In many applications, certain exposure variables, which are relatively easy and cheap to be measured, are observed on all of the subjects in the cohort. Such data are referred to as the first-phase covariate data. Complete measurements of primary exposure which are expensive to assemble are only collected for the subcohort and all cases. These data are referred to as the second-phase covariate data. To improve the efficiency of the original case–cohort design, a *stratified case–cohort design* suggests to select a subcohort by a stratified sampling scheme which depends on the available first-phase covariate data. Besides the generalized case–cohort design and the stratified case–cohort design, many variations of the sampling schemes based on the original case–cohort design have been developed, and we refer to such biased-sampling schemes as modified case–cohort designs.

The motivation, importance, and broad potential applications of case–cohort designs are widely discussed in the literature. Parametric models for case–cohort designs have been studied (Kalbfleisch and Lawless 1988; Nan et al. 2006). Statistical methods for fitting case–cohort data with semiparametric survival models have also been developed for the proportional hazards model (Prentice 1986; Self and Prentice 1988; Lin and Ying 1993; Barlow 1994; Chen and Lo 1999; Borgan et al. 2000; Chen 2001c; Cai and Zeng 2004, 2007; Kulich and Lin 2004; Qi et al. 2005; Breslow and Wellner 2007), the additive hazards model (Kulich and Lin 2000; Sun et al. 2004), the proportional odds model (Chen 2001a), the accelerated failure time model (Kong and Cai 2009), the semiparametric transformation models (Chen 2001b; Kong et al. 2004; Lu and Tsiatis 2006), among others. Various estimating procedures have been proposed for data from case–cohort studies. These have proceeded mainly along two lines, likelihood-based approaches and estimating-equation-based approaches.

Throughout this section, we suppose that there exists a study population of N independent individuals. Let $\widetilde{T}_i$ denote the potential failure time and C_i denote the censoring time for subject i $(i = 1, \ldots, N)$. The observed time is $T_i = \min(\widetilde{T}_i, C_i)$. Let $\Delta_i = I(\widetilde{T}_i \leq C_i)$ denote the failure indicator for subject i, $Y_i(t) = I(T_i \geq t)$ denote the at-risk process and $N_i(t) = \Delta_i I(T_i \leq t)$ denote the counting process, where $I(\cdot)$ is an indicator function. Let $Z_i(t)$ be a p-dimensional exposure variable for subject i at time t. Let β be a p-dimensional regression parameter of interest. Let τ denote the study end time.

2.1.1 *Original case–cohort designs*

In the landmark article of Prentice (1986), the case–cohort design was first formally proposed, in which a subcohort is selected randomly from the full cohort, and the complete information of exposure are only observed for the subcohort subjects and additional cases outside the subcohort. Under the proposed case–cohort design, Prentice (1986) considered a relative risk regression model:

$$\lambda\left(t|Z_i(t)\right) = \lambda_0(t) r\left(\beta' Z_i(t)\right), \quad i = 1, \ldots, N,$$

where $\lambda_0(t)$ is the baseline hazard function and $r(\cdot)$ is a known function with $r(0) = 1$, for example, $r(x) = e^x$ for the proportional hazards model. Prentice (1986) proposed a pseudo-likelihood approach for estimation of the parameter β by maximizing the following objective function:

$$L_1(\beta) = \prod_{i=1}^{N} \left[\frac{r\left(\beta' Z_i(T_i)\right)}{\sum_{l \in \widetilde{R}(T_i)} r\left(\beta' Z_l(T_i)\right)} \right]^{\Delta_i} \tag{1}$$

where $\widetilde{R}(t) = \{i : N_i(t) \neq N_i(t-)\} \cup S_0$, and S_0 denotes the index set of the subcohort. Note that the objective function (1) is a modification of the partial-likelihood function (Cox 1975) that weights the contributions of the cases and subcohort differently. Since the expression in (1) does not generally possess a partial-likelihood interpretation, it was termed as pseudo-likelihood. After the publish of Prentice (1986), the case–cohort design and related statistical methodologies have been extensively studied.

Self and Prentice (1988) further elaborated such pseudo-likelihood estimators by slightly modifying the risk set $\widetilde{R}(t)$ used in (1) to S_0. Estimators obtained from the modified pseudo-likelihood function was proved to be asymptotically equivalent to the pseudo-likelihood estimators defined in Prentice (1986). Asymptotic distribution theory for such pseudo-likelihood estimators and corresponding cumulative failure rate estimators were presented. Lin and Ying (1993) and Barlow (1994) further discussed the pseudo-likelihood methods and provided different ways to obtain easily computed variances for the estimators.

Chen and Lo (1999) improved the pseudo-likelihood estimators (Prentice 1986) by utilizing the information of all cases in constructing the risk set to derive estimating equations. Under the proportional hazards model,

$$\lambda\big(t|Z_i(t)\big) = \lambda_0(t)\exp\big\{\beta' Z_i(t)\big\}, \quad i = 1, \ldots, N, \tag{2}$$

the pseudo-likelihood (1) of Prentice (1986) yields the score function:

$$U_1(\beta) = \sum_{i=1}^{N} \int_0^{\tau} \left[Z_i(t) - \frac{\sum_{l \in \widetilde{R}(t)} Z_l(t) e^{\beta' Z_l(t)}}{\sum_{l \in \widetilde{R}(t)} e^{\beta' Z_l(t)}} \right] dN_i(t) = 0. \tag{3}$$

Chen and Lo (1999) modified the risk set $\widetilde{R}(t)$ used in the above score function by including more information of cases. Let N_1 (n_1) and N_0 (n_0) be the numbers of cases and controls in the cohort (subcohort), respectively. Let R_1 ($\widetilde{R}_1$) and R_0 ($\widetilde{R}_0$) be the index sets of all cases and all controls in the cohort (subcohort), respectively. Denote $R_1(t) = \{i : T_i \geq t, \ i \in R_1\}$ and $\widetilde{R}_0(t) = \{i : T_i \geq t, \ i \in \widetilde{R}_0\}$, i.e., the risk sets defined on R_1 and $\widetilde{R}_0$, respectively. The following estimating equation was proposed by Chen and Lo (1999),

$$U_2(\beta) = \sum_{i=1}^{N} \int_0^{\tau} \left[Z_i - \frac{\frac{\hat{p}}{N_1} \sum_{l \in R_1(t)} Z_l e^{\beta' Z_l} + \frac{1-\hat{p}}{n_0} \sum_{l \in \widetilde{R}_0(t)} Z_l e^{\beta' Z_l}}{\frac{\hat{p}}{N_1} \sum_{l \in R_1(t)} e^{\beta' Z_l} + \frac{1-\hat{p}}{n_0} \sum_{l \in \widetilde{R}_0(t)} e^{\beta' Z_l}} \right] dN_i(t) = 0, \tag{4}$$

where $\hat{p}$ is an estimator of the population case probability $p = P(\Delta = 1)$. They derived a class of estimating equations by using different estimators of p. Let $\hat{p} = n_1/n$, then (4) becomes

$$U_3(\beta) = \sum_{i=1}^{N} \int_0^{\tau} \left[Z_i - \frac{\frac{n_1}{nN_1} \sum_{l \in R_1(t)} Z_l e^{\beta' Z_l} + \frac{1}{n} \sum_{l \in \widetilde{R}_0(t)} Z_l e^{\beta' Z_l}}{\frac{n_1}{nN_1} \sum_{l \in R_1(t)} e^{\beta' Z_l} + \frac{1}{n} \sum_{l \in \widetilde{R}_0(t)} e^{\beta' Z_l}} \right] dN_i(t) = 0.$$

Substitute $\hat{p} = N_1/N$ into (4), giving

$$U_4(\beta) = \sum_{i=1}^{N} \int_0^{\tau} \left[Z_i - \frac{\frac{1}{N} \sum_{l \in R_1(t)} Z_l e^{\beta' Z_l} + \frac{N_0}{n_0 N} \sum_{l \in \widetilde{R}_0(t)} Z_l e^{\beta' Z_l}}{\frac{1}{N} \sum_{l \in R_1(t)} e^{\beta' Z_l} + \frac{N_0}{n_0 N} \sum_{l \in \widetilde{R}_0(t)} e^{\beta' Z_l}} \right] dN_i(t) = 0.$$

Including more information of cases in constructing the estimating equations, the estimators proposed by Chen and Lo (1999) improve the pseudo-likelihood estimators of Prentice (1986) by achieving better efficiency.

Kulich and Lin (2000) proposed an inverse probability weighted estimating approach for the regression parameters of the additive hazards model, which has the form

$$\lambda\big(t|Z_i(t)\big) = \lambda_0(t) + \beta' Z_i(t), \quad i = 1, \ldots, N. \tag{5}$$

Under the case–cohort design, let ξ_i be the subcohort indicator, having the value 1 if the ith subject being selected into the subcohort and 0 otherwise. Denote $\pi_i = P(\xi_i = 1)$ to be the selection probability of the ith subject. Applying the inverse probability weighted approach, Kulich and Lin (2000) defined the weights as

$$w_i = \Delta_i + \big(1 - \Delta_i\big)\xi_i/\pi_i, \quad i = 1, \ldots, N,$$

and derived the weighted estimating equation as

$$U_5(\beta) = \sum_{i=1}^{N} \int_0^{\tau} w_i \big\{Z_i(t) - \bar{Z}(t)\big\}\big\{dN_i(t) - Y_i(t)\beta' Z_i(t)dt\big\} = 0, \tag{6}$$

where $\bar{Z}(t) = \frac{\sum_{i=1}^{N} w_i Y_i(t) Z_i(t)}{\sum_{i=1}^{N} w_i Y_i(t)}$. The resulting estimator has a closed form:

$$\hat{\beta}_5 = \left[\sum_{i=1}^{N} \int_0^{\tau} w_i Y_i(t) \big\{Z_i(t) - \bar{Z}(t)\big\}^{\otimes 2} dt\right]^{-1} \left[\sum_{i=1}^{N} \int_0^{\tau} \big\{Z_i(t) - \bar{Z}(t)\big\} dN_i(t)\right],$$

where $a^{\otimes 2} = aa'$. Kulich and Lin (2000) studied how to fit the case–cohort data to the addictive hazards model, which is an important alternative to the proportional hazards model when researchers are interested in risk differences rather than risk ratios. Including the information of primary exposure from all the cases in $\bar{Z}(t)$ regardless of whether or not they belong to the subcohort, the proposed estimating procedure makes fuller use of the exposure information from both the cases and controls. Furthermore, the proposed method can also be applied to the situations that the subcohort is selected by Bernoulli sampling with arbitrary selection probabilities or possibly stratified simple random sampling.

Chen (2001a) proposed a weighted semiparametric likelihood method for case–cohort studies under the proportional odds model, in which,

$$P\big(\widetilde{T}_i > t|Z_i\big) = \frac{1}{1 + \exp\big\{\beta' Z_i\big\}\Lambda_0(t)}, \quad i = 1, \ldots, N,$$

where $\Lambda_0(t)$ is the baseline cumulative hazard function. Let S_0 be the index set of the subcohort and S_1 be the index set of the cases outside the subcohort. Denote $\widetilde{S}_0 = \{i \in S_0,\ \Delta_i = 0\}$ and $\widetilde{S}_1 = \{i \in S_0 \cup S_1,\ \Delta_i = 1\}$. Let $n_{\widetilde{S}_0}$ and $n_{\widetilde{S}_1}$ denote the sample sizes of $\widetilde{S}_0$ and $\widetilde{S}_1$, respectively. Chen (2001a) derived the estimation of regression parameter β by maximizing the objective function:

$$L_2(\beta) = \prod_{i\in\widetilde{S}_0}\left[\frac{1}{1+\exp\left\{\beta' Z_i\right\}\Lambda_0(T_i)}\right]^{\frac{N-n_{\widetilde{S}_1}}{n_{\widetilde{S}_0}}} \prod_{i\in\widetilde{S}_1}\frac{\exp\left\{\beta' Z_i\right\}d\Lambda_0(T_i)}{\left[1+\exp\{\beta' Z_i\}\Lambda_0(T_i)\right]^2}, \quad (7)$$

where $\Lambda_0(t)$ is restricted to the class of monotonically increasing functions of the form $\Lambda_0(T_i) = \sum_{j\in\widetilde{S}_1} Y_i(T_j)\lambda_j$, that is, with jumps at the observed failure times only. The proposed objective function (7) is a geometrically weighted version of the so-called complete-case likelihood function, hence it was called simply the weighted semiparametric likelihood. Chen (2001a) studies the case–cohort design under the proportional odds model, which is a potentially useful alternative in some applications when the proportional hazards model does not fit the data well. The proposed estimating procedure is applicable to the semiparametric transformation model. Particularly, the estimators of Chen and Lo (1999) can be generated by the approach of Chen (2001a) under appropriate weighting scheme under the proportional hazards model.

Kong et al. (2004) considered the following semiparametric transformation models:

$$H(\widetilde{T}_i) = -Z_i'\beta + \varepsilon_i, \quad i = 1, \ldots, N, \quad (8)$$

where H is an unspecified strictly increasing function and ε is a random error with a known distribution function F. The main idea of Kong et al. (2004) is to regard case–cohort design as a special case of general missing data problems. The exposure variables are missing by design in case–cohort studies, so the missing mechanism is clearly known. Following the inference of model (8) for complete data, Kong et al. (2004) introduced an extra parameter $\gamma = H(t_0)$, where t_0 is a prespecified constant such that $P(\min(\widetilde{T}, C) > t_0) > 0$, and then obtained the parameter vector $\theta = (\beta', \gamma)'$. Suppose a subcohort of size n is selected randomly from the cohort. Let ξ_i denote the indicator for the ith subject being selected into the subcohort. Assume $P(\xi_i = 1) = \pi = n/N$, which means each subject has the same probability of being selected into the subcohort.

Motivated by the idea of weighting the incomplete data by the inverse selection probabilities, Kong et al. (2004) defined a weight w_{ij} to reflect the contribution of a pair of subjects i and j to the estimating function as

$$w_{ij} = w_i w_j,$$

where

$$w_i = \Delta_i + (1-\Delta_i)\xi_i/\pi.$$

For estimation of the parameter vector θ, the weighted estimating equation was proposed

$$U_6(\theta) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{ij}\rho_{ij}(\theta)\dot{\eta}_{ij}(\theta) \left[\frac{\Delta_j I\big\{\min(T_i, t_0) \geq T_j\big\}}{\hat{G}_n^2(T_j)} - \eta_{ij}(\theta) \right] = 0, \tag{9}$$

where $\rho_{ij}(\theta)$ is a positive weight function, $\eta_{ij}(\theta) = \int_{-\infty}^{\gamma}\{1 - F(v + Z_i'\beta)\}dF(v + Z_j'\beta)$, $\dot{\eta}_{ij}(\theta) = \partial\eta_{ij}(\theta)/\partial\theta$, and $\hat{G}_n(\cdot)$ is the Kaplan–Meier estimator of the survival function for censoring time based on the subcohort data. Two types of weight were used in the estimating function (9), in which, the weight w_{ij} was applied to take into account the sampling design effect, and $\rho_{ij}(\theta)$ was introduced to improve the efficiency of the estimating equations. In some applications, the proportional hazards model may not fit the data well, or researchers may be interested in modelling the association from different aspects. The semiparametric transformation models, incorporating a variety of nonproportional hazards models, can be a more flexible choice in such situations. Kong et al. (2004) established statistical methods for the case–cohort data under the semiparametric transformation models.

Lu and Tsiatis (2006) developed a way of weighted estimating equations for parameters of the semiparametric transformation models in (8) under the case–cohort design. Inspired by the methods of semiparametric transformation models for the complete data, Lu and Tsiatis (2006) considered a martingale process defined as

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)d\Lambda\big\{H(s) + Z_i'\beta\big\},$$

where $\Lambda(t)$ denotes the cumulative hazard function for ε in (8). Suppose a subcohort of size n is selected randomly from the cohort. Let ξ_i denote the indicator for the ith subject being selected into the subcohort. Assume each subject has the same probability $P(\xi_i = 1) = \pi = n/N$ of being selected into the subcohort. Lu and Tsiatis (2006) adopted weights as

$$w_i = \Delta_i + (1 - \Delta_i)\xi_i/\pi, \quad i = 1, \ldots, N,$$

and proposed to use the following estimating equations

$$U_9(\beta) = \sum_{i=1}^{N} w_i \left[dN_i(t) - Y_i(t)d\Lambda\big\{H(t) + Z_i'\beta\big\}\right] = 0, \quad (t \geq 0),\ H(0) = -\infty,$$

$$U_{10}(\beta) = \sum_{i=1}^{N} \int_0^{\tau} w_i Z_i \left[dN_i(t) - Y_i(t)d\Lambda\big\{H(t) + Z_i'\beta\big\}\right] = 0,$$

to estimate functions for H and β. As we mentioned before, Kong et al. (2004) studied the case–cohort design under semiparemetric transformation models by regarding the case–cohort data as a special missing data problem. In the model of Kong et al. (2004), the censoring time was assumed to be independent of exposure variables. Lu and Tsiatis (2006) developed a different way of estimating parameters. The proposed procedure makes use of a martingale integral representation and an inverse probability weighted approach in constructing the estimating equations. The proposed method of Lu and Tsiatis (2006) allows the censoring time to depend on exposure variables. Breslow and Wellner (2007) further studied a theory of inverse probability weighted methods under the semiparametric model with two-phase stratified samples.

Qi et al. (2005) considered weighted estimators for the proportional hazards model (2) with missing exposure variables. Suppose that some elements of the exposure variable Z are missing. Define $Z_i = (Z_i^m, Z_i^c)$, where Z_i^c denotes the exposure variables for the ith subject that are always observed and Z_i^m denotes the exposure variables that are sometimes missing. Let ξ_i denote the selection indicator, which equals 1 if Z_i^m is available and 0 if Z_i^m is missing for the ith subject. The missing-data mechanism is determined by the distribution of ξ_i given (T_i, Δ_i, Z_i^c), which is Bernoulli with probability $\pi_i = P(\xi_i = 1|T_i, \Delta_i, Z_i^c)$. When the selection probability $\pi = (\pi_1, \ldots, \pi_N)'$ is known, under the proportional hazards model, Qi et al. (2005) first proposed a weighted estimating function as

$$U_7(\beta, \pi) = \sum_{i=1}^{N} w_i \int_0^\tau \left[Z_i - \frac{\sum_{l=1}^N w_l Y_l(t) Z_l e^{\beta' Z_l}}{\sum_{l=1}^N w_l Y_l(t) e^{\beta' Z_l}} \right] dN_i(t) = 0, \tag{10}$$

where

$$w_i = \xi_i / \pi_i, \quad i = 1, \ldots, N.$$

The estimator obtained by the above estimating equation may not be efficient. In an attempt to improve efficiency, an estimator of π was used in the above estimating function (10). Qi et al. (2005) applied nonparametric kernel smoothing techniques to estimate π based on observed data, including complete and incomplete observations. Let W denote the variables on which an estimator of π is allowed to depend. The rth-order kernel function K is a piecewise smooth function, which satisfies $\int K(u)du = 1$, $\int u^m K(u)du = 0$, for $m = 1, \ldots, (r-1)$, $\int u^r K(u)du \neq 0$, and $\int K(u)^2 du < \infty$. Define $K_h(\cdot) = K(\cdot/h)$, where h is the bandwidth. The following estimator

$$\hat{\pi} = \frac{\sum_{i=1}^N \xi_i K_h(w - W_i)}{\sum_{j=1}^N K_h\left(w - W_j\right)},$$

was proposed to estimate π. Replacing π in the weighted estimating function (10) with $\hat{\pi}$, Qi et al. (2005) derived a new estimating equation:

$$U_8(\beta, \pi) = \sum_{i=1}^{N} \int_0^\tau \hat{w}_i \left[Z_i - \frac{\sum_{l=1}^N \hat{w}_l Y_l(t) Z_l e^{\beta' Z_l}}{\sum_{l=1}^N \hat{w}_l Y_l(t) e^{\beta' Z_l}} \right] dN_i(t) = 0,$$

where

$$\hat{w}_i = \xi_i / \hat{\pi}_i, \quad i = 1, \ldots, N.$$

Under the proportional hazards model with missing exposure variables, Qi et al. (2005) presented both simple weighted and kernel-assisted fully augmented weighted estimators, and the latter one is more efficient than the former one. The proposed methods require neither a model for the missing-data mechanism nor specification of the conditional distribution of the missing exposure variable. The proposed methods allow the missing-data mechanism to depend on outcome variables and observed exposure variables, which makes the proposed estimating procedure applicable to various cohort sampling designs, including the case–cohort design.

Nan et al. (2006) studied how to fit case–cohort data to a linear regression model, which models the relationship of the failure time and the primary exposure variable directly as the form:

$$\widetilde{T}_i = \beta' Z_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{11}$$

where, given (Z_i, C_i), the ε_i's are independent and identically distributed with an unknown distribution. For the case–cohort study, the estimating equation was proposed

$$U_{11}(\beta) = \sum_{i=1}^{N} \int_0^{\tau} \left[Z_i - \frac{\sum_{l \in S_0} Z_l Y_l \left(u + \beta' Z_l\right)}{\sum_{l \in S_0} Y_l \left(u + \beta' Z_l\right)} \right] dN_i \left(u + \beta' Z_i\right) = 0, \tag{12}$$

where S_0 denotes the index set of the subcohort. Nan et al. (2006) developed the statistical methods for the case–cohort design under a linear regression model, which is an important alternative way of analyzing failure time data. The proposed weighted estimating equations are derived by modifying the linear ranks tests and estimating equations which arise from full-cohort data, using similar methods to those applied by Self and Prentice (1988) for the proportional hazards model.

Kong and Cai (2009) considered the accelerated failure time model:

$$\log(\widetilde{T}_i) = \beta' Z_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{13}$$

where ε_i's are independent and identically distributed random errors with an unspecified distribution. Define $e_i(\beta) = \log(T_i) - Z_i'\beta$, $N_i(\beta; t) = \Delta_i I(e_i(\beta) \leq t)$, and $Y_i(\beta; t) = I(e_i(\beta) \geq t)$, for $i = 1, \ldots, N$. For the case–cohort design, suppose a subcohort of size n is selected randomly from the cohort. Let ξ_i denote the indicator for the ith subject being selected into the subcohort. Assume each subject has the same probability $\pi = n/N$ of being selected into the subcohort. Kong and Cai (2009) adopted the weights as

$$w_i = \Delta_i + (1 - \Delta_i)\xi_i/\pi, \quad i = 1, \ldots, N,$$

and developed a rank-based estimating equation approach by solving the score function

$$U_{12}(\beta) = \sum_{i=1}^{N} \Delta_i \phi\left(\beta; e_i(\beta)\right) \left\{ Z_i - \widetilde{Z}\left(\beta; e_i\left(\beta\right)\right) \right\} = 0, \tag{14}$$

where $\widetilde{Z}(\beta;t) = \frac{\sum_{l=1}^{N} w_l Z_l Y_l(\beta;t)}{\sum_{l=1}^{N} w_l Y_l(\beta;t)}$, and ϕ is a possibly data-dependent weight function. The choices of $\phi(\beta;t) = 1$ and $\phi(\beta;t) = N^{-1}\sum_{i=1}^{N} Y_i(\beta;t)$ correspond to the log-rank and Gehan statistics, respectively. By linearly relating the natural logarithm of the failure time to the exposure, the accelerated failure time model may be attractive to model failure time data in some applications. Kong and Cai (2009) developed a rank-based estimating approach for analyzing the case–cohort data under the accelerated failure time model. Furthermore, the proposed method is also valid for the usual linear model. Compared with the estimating function (12) used by Nan et al. (2006), the proposed estimating approach includes failures outside the subcohort in constructing $\widetilde{Z}(\beta;t)$ in Eq. (14). Therefore, the estimators of Kong and Cai (2009) may be more efficient.

2.1.2 Generalized case–cohort designs

The case–cohort design is used primarily to reduce the cost involved in the assembly of the exposure information. The censoring times of the subjects who are not included in the subcohort may be much less costly to obtain. Chen (2001b) studied the case–cohort design modified by considering the information of the censoring times of subjects not included in the subcohort to parameter estimation. They considered a more general specification of semiparametric transformation regression models, which assumes

$$P\left(\widetilde{T}_i > t|Z_i\right) = \phi\big(\beta, Z_i, H(t)\big), \quad i = 1, \ldots, N, \tag{15}$$

where ϕ is assumed known. Model (15) reduces to the usual semiparametric models by choosing certain specified form of ϕ. Let φ be the derivative of $-\phi$ with respect to the third argument. Denote $G(t) = \int \phi(\beta, z, H(t))dQ(z)$, where Q is the marginal distribution of the exposure variable Z, and $\bar{G}(t) = 1 - G(t)$. Let $\nu(\beta, Q, G(t))$ denote the inverse transformation.

Chen (2001b) proposed a maximum conditional-profile-likelihood method to fit the above model (15) to data from the modified case–cohort design. Let S_0 be the index set of the subcohort, S_1 be the index set of the cases that are not selected in the subcohort, and S_2 be the index set of the remaining subjects. With the observation (T_i, Δ_i, Z_i), for $i \in S_0 \cup S_1$, and (T_i, Δ_i), for $i \in S_2$, they proposed the conditional likelihood as:

$$L_3(\beta) = \prod_{i \in S_0 \cup S_1} \left[\frac{\varphi\Big(\beta, Z_i, \nu\big(\beta, Q, G(T_i)\big)\Big)}{\int \varphi\Big(\beta, z, \nu\big(\beta, Q, G(T_i)\big)\Big) dQ(z)}\right]^{\Delta_i} \times \left[\frac{\varphi\Big(\beta, Z_i, \nu\big(\beta, Q, G(T_i)\big)\Big)}{G(T_i)}\right]^{1-\Delta_i} \times \prod_{i \in S_0 \cup S_1 \cup S_2} G(T_i)^{\Delta_i}\left\{d\bar{G}(T_i)\right\}^{1-\Delta_i}.$$

Since neither G nor Q is known, G was replaced by the Kaplan–Meier estimator and Q was replaced by the empirical estimator based on the random subcohort S_0 to obtain a conditional profile likelihood function. Then an estimator of β can be derived by maximizing the resulting profile likelihood. Chen (2001b) considered the problem of fitting a more flexible semiparametric transformation regression models to data from a modified case–cohort design, in which the efficiency gain may arise because the censoring times of all the censored subjects in the cohort are included.

Chen (2001c) defined a generalized case–cohort design, which consists of a number of sampling steps. Each step takes a random sample from a certain subset of the cohort, and the design of the sample size and subset at each step and of the total numbers of steps is independent of the observed exposure. Such generalized case–cohort design covers case–control design, nested case–control design and original case–cohort design. Under the proportional hazards model (2) for data from the proposed generalized case–cohort design, Chen (2001c) developed a weighted estimating equation approach, in which the weights were obtained from an idea of estimating each missing exposure variable by a local average. Let ξ_i denote the indicator, equaling 1 if the ith subject is sampled and 0 otherwise. Let $0 = t_0 \le t_1 \le \cdots \le t_{a_N} = \tau$ and $0 = s_0 \le s_1 \le \cdots \le s_{n_N} = \tau$ be two partitions of $[0, \tau)$. The partitions may be data dependent but should only depend on $(T_i, \Delta_i, \xi_i),\ i = 1, \ldots, N$. Let $r_N(t, d)$ be a step function defined on $[0, \tau) \times \{0, 1\}$ such that

$$r_N(t,d) = \begin{cases} \dfrac{\sum_{l=1}^{N} \Delta_l \xi_l I\left(T_l \in [t_{i-1}, t_i)\right)}{\sum_{l=1}^{N} \Delta_l I\left(T_l \in [t_{i-1}, t_i)\right)}, & \text{if } d = 1\text{and } t \in [t_{i-1}, t_i), \\ \dfrac{\sum_{l=1}^{N} (1-\Delta_l) \xi_l I(T_l \in [s_{j-1}, s_j))}{\sum_{l=1}^{N} (1-\Delta_l) I(T_l \in [s_{j-1}, s_j))}, & \text{if } d = 0 \text{ and } t \in [s_{j-1}, s_j), \end{cases}$$

for $1 \le i \le a_N$ and $1 \le j \le b_N$. Defining the weights as

$$w_i = \frac{\xi_i}{r_N\left(T_i, \Delta_i\right)}, \quad i = 1, \ldots, N,$$

Chen (2001c) proposed the weighted estimating equation as

$$U_{13}(\beta) = \sum_{i=1}^{N} \int_0^{\tau} w_i \left[h\left(Z_i(t)\right) - \frac{\sum_{l=1}^{N} w_l Y_l(t) h\left(Z_l(t)\right) e^{\beta' Z_l(t)}}{\sum_{l=1}^{N} w_l Y_l(t) e^{\beta' Z_l(t)}} \right] dN_i(t) = 0,$$

where h is an exposure-related process.

The sample reuse approach via local averaging proposed by Chen (2001c) is more efficient than the typical approach via inclusion probabilities. By choosing $h(x) = x$, Chen (2001c) improved the pseudo-likelihood estimators of Prentice (1986). Despite more complexity and difficulty, a semiparametric efficient estimator can be obtained by choosing h to be the exposure variable transformed by the inverse of an estimated linear integral operator. Samuelsen et al. (2007) further discussed Chen's approach and pointed out how it is related to stratified case–cohort analysis. They studied an extension of Chen's generalized case–cohort design to allow for surrogate-dependent sampling and showed how such data may be analyzed with the post-stratification method.

Cai and Zeng (2007) developed the case–cohort design to a generalized case–cohort design, where only a fraction of cases instead of all cases are sampled. The main difference of such generalized case–cohort design from the original case–cohort design is that not all the remaining cases are selected for assembling the exposure measurements. Cai and Zeng (2007) proposed a general log-rank test statistic, which was constructed by approximating the risk set and the event process of the complete data using the sampled data. Kang and Cai (2009) and Kang et al. (2013) further developed the statistical inference for generalized case–cohort studies with multiple disease outcomes. The methods can be easily reduced to the situation with univariate outcome. We will discuss in details later.

2.1.3 Stratified case–cohort designs

In order to improve the efficiency of the case–cohort study by making better use of the first-phase covariate data, several literature discussed stratified case–cohort designs.

Kulich and Lin (2004) developed a general class of weighted estimators under a stratified case–cohort designs. Consider a cohort of N subjects who can be divided into K mutually exclusive strata based on a discrete random variable V, which represents the first-phase covariate information. Let ξ denote the selection indicator of a subject into the subcohort. For each $k = 1, \ldots, K$, let $P(\xi = 1|V = k) = \pi_k$. Let N_k denote the number of subjects in the kth stratum. Under the stratified case–cohort design, complete observations $(T_{ki},\ \Delta_{ki},\ Z_{ki}(t),\ 0 \leq t \leq \tau,\ V_{ki},\ \xi_{ki} = 1)$ are available for all subcohort subjects, and at least $(T,\ \Delta_{ki} = 1,\ Z_{ki}(T_{ki}))$ are observed for the cases, where the subscript $\{ki\}$ denotes the ith subject in the kth stratum. Under the proportional hazards model in (2), Kulich and Lin (2004) proposed a weighted estimating approach for estimating the regression parameter β by solving the score function

$$U_{14}(\beta)=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\int_0^{\tau}\left[Z_{ki}(t)-\frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k} w_{ki}(t)Y_{ki}(t)Z_{ki}(t)e^{\beta' Z_{ki}(t)}}{\sum_{k=1}^{K}\sum_{i=1}^{N_k} w_{ki}(t)Y_{ki}(t)e^{\beta' Z_{ki}(t)}}\right]dN_{ki}(t)=0. \tag{16}$$

Various proposals for the potentially time-varying weight $w_{ki}(t)$ yield different case–cohort estimators.

Kulich and Lin (2004) further extended the above method to a doubly weighted estimation method by incorporating arbitrary stochastic processes as time-varying weights into the empirical sampling probabilities. Let $\mathbf{A}_{ki}(t)$ be a diagonal matrix with m potentially different random processes on the diagonal. Consider the following estimators of the subcohort sampling probabilities:

$$\hat{\boldsymbol{\pi}}_k(t) = \left[\sum_{i=1}^{N_k}(1-\Delta_{ki})\,\mathbf{A}_{ki}(t)\right]^{-1}\left[\sum_{i=1}^{N_k}(1-\Delta_{ki})\,\xi_{ki}\mathbf{A}_{ki}(t)\right],$$

which yields m estimators of π_k on the diagonal of $\hat{\boldsymbol{\pi}}_k(t)$. Each estimator can be interpreted as an empirical sampling proportion based on the control, with the contribution of each control weighted by a component of $\mathbf{A}_{ki}(t)$. The weight matrix was defined as

$$\mathbf{w}_{ki}(t) = \Delta_{ki}\mathbf{I}_m + (1 - \Delta_{ki})\,\xi_{ki}\hat{\boldsymbol{\pi}}_k^{-1}(t),$$

where $\mathbf{I}_m$ is an $m \times m$ identity matrix. Kulich and Lin (2004) considered the estimating equation in (16) as

$$U_{15}(\beta) = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\int_0^{\tau}\left\{Z_{ki}(t) - \bar{Z}_{DW}(t,\beta)\right\}dN_{ki}(t) = 0,$$

where

$$\bar{Z}_{DW}(t,\beta) = \left\{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\mathbf{w}_{ki}(t)Y_{ki}(t)e^{\beta' Z_{ki}(t)}\right\}^{-1} \times \left\{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\mathbf{w}_{ki}(t)Y_{ki}(t)Z_{ki}(t)e^{\beta' Z_{ki}(t)}\right\}.$$

The estimators proposed by Borgan et al. (2000) can be regarded as the special cases for the above doubly weighted estimators. To reduce the efficiency loss caused by misspecification of model, Kulich and Lin (2004) combined the doubly weighted estimator with the estimator of Borgan et al. (2000) to obtain a combined doubly weighted estimator. The proposed estimators Kulich and Lin (2004) may be more efficient than the estimators of Chen and Lo (1999), Borgan et al. (2000) and Chen (2001c) by choosing appropriate weight functions.

Nan et al. (2006) also developed weighted estimating equation methods for a stratified case–cohort designs under the linear regression model (11). If a variable Z^* that is highly correlated with the primary exposure variable Z is available for all the subjects in the cohort, selecting the subcohort using a stratified sampling scheme based on Z^* can improve efficiency. An independent Bernoulli sampling method, in which

$$\pi\left(Z_i^*\right) = P\left(i \in S_0 | Z_i^*\right), \quad i = 1, \ldots, N,$$

was considered to select the subcohort. Nan et al. (2006) proposed the following estimating equation for such stratified case–cohort design as

$$U_{16}(\beta) = \sum_{i=1}^{N}\int_0^{\tau}\left[Z_i - \frac{\sum_{l=1}^{N} w_l Z_l Y_l\left(u + \beta' Z_l\right)}{\sum_{l=1}^{N} w_l Y_l\left(u + \beta' Z_l\right)}\right]dN_i\left(u + \beta' Z_i\right) = 0, \qquad (17)$$

where

$$w_i = I\left(i \in S_0\right)/\pi\left(Z_i^*\right), \quad i = 1, \ldots, N,$$

and S_0 denotes the index set of the subcohort. Nan et al. (2006) extended their works on the original case–cohort design (see 12) to the stratified case–cohort design. By selecting the subcohort using a stratified sampling scheme, which makes better use of the first-phase covariate information, the estimator obtained from (17) can improve efficiency.

Kong and Cai (2009) also extended the proposed estimating procedure under the accelerated failure time model (13) for the original case–cohort design (see 14) to the stratified case–cohort design. For the stratified case–cohort design, the full cohort is supposed to consist of K strata of sizes $N_1, \ldots, N_K$. Let n_k denote the size of samples selected from the kth stratum into the subcohort. Let $\pi_k = n_k/N_k$ be the sampling proportion of the subcohort in the kth stratum. Kong and Cai (2009) proposed the following rank-based estimating equation

$$U_{17}(\beta) = \sum_{k=1}^{K}\sum_{i=1}^{N_k} \Delta_{ki}\phi\left(\beta; e_{ki}(\beta)\right)\left\{Z_{ki} - \widetilde{Z}_k\left(\beta; e_i(\beta)\right)\right\} = 0, \tag{18}$$

where $\widetilde{Z}_k(\beta; t) = \frac{\sum_{l=1}^{N_k} w_{kl} Z_{kl} Y_{kl}(\beta; t)}{\sum_{l=1}^{N_k} w_{kl} Y_{kl}(\beta; t)}$, and

$$w_{ki} = \Delta_{ki} + (1 - \Delta_{ki})\,\xi_{ki}/\pi_k, \quad i = 1, \ldots, N.$$

Since the stratified sampling design further improves the efficiency when the stratification variable is a good surrogate of the primary exposure, the proposed method of Kong and Cai (2009) can further enhance the efficiency. The proposed methods are also valid for the usual semiparametric linear model.

2.2 General failure-time ODS design

While case–cohort is an efficient design, especially when failure is rare, i.e., in the high censoring situations, this design may not be practically feasible to implement if the failure is not rare. In case–cohort design and generalized case–cohort design, the selection of the supplemental samples depends on whether the event of interest happens or not. As ODS design with continuous outcome, if an exposure variable is related to the outcome, then the subjects, whose observed failure time are very long or short, should be of more information about the exposure-response relationship.

To take advantage of the ODS scheme for right-censored data to yield more powerful and efficient inferences, Ding et al. (2014) proposed a *general failure-time ODS sampling design*. In such a general failure-time ODS design, a random sample (SRS) from the full cohort is selected. In addition, the range of observed failure time of all the cases is partitioned into mutually exclusive and exhaustive strata, and a supplemental sample from each stratum is selected. The measurements of primary exposure variables are only assembled for these two components. Like case–cohort designs, the general failure-time ODS design enriches the observed sample by selectively including certain failure subjects. The development of such a general ODS design for failure time data is interesting and significant in building a cost-effective sampling design in survival analysis related studies. Several authors studied the statistical inference methodologies for data from the general failure-time ODS design.

To further highlight the general ODS design with failure time data, suppose that there exists a study population of N independent individuals. Assume that the range of observed failure time of all the cases is partitioned into K mutually exclusive and exhaustive strata: $A_k = (a_{k-1}, a_k],\ k = 1, \ldots, K$, by some known constants $\{a_i, i = 1, \ldots, K\}$ which satisfy $0 = a_0 < a_1 < \cdots < a_{k-1} < a_k = \tau$. The general failure-time ODS design: First, a random sample of size n_0 from the full cohort, denoted by the SRS sample, is selected. In addition, a supplemental sample of size n_k is selected from each of the above kth stratum of cases. The samples from these two components constitute the ODS sample. Suppose that n_k is fixed by design for $k = 1, \ldots, K$. Denote $n = \sum_{k=0}^{K} n_k$ to be the total size of the ODS sample. Let V, S_0 and S_k be the index set of the total ODS sample, the SRS sample and the supplemental sample from the kth stratum, respectively. Hence, the observed data for the failure-time ODS design can be summarized as: (T_i, Δ_i, Z_i) when $i \in S_0$, and $(T_i, \Delta_i, Z_i |\ \Delta_i = 1, T_i \in A_k)$ when $i \in S_k,\ k = 1, \ldots, K$.

Ding et al. (2014) developed such a general failure-time ODS scheme and established estimation procedures under the proportional hazards model in (2). The likelihood function based on the observed data from the general failure-time ODS design is proportional to

$$\begin{aligned} L_4(\beta, Q_Z, \Lambda_0, S_C) = & \left[\prod_{i \in S_0} \left\{f_{\beta,\Lambda_0}(T_i|Z_i)\right\}^{\Delta_i} \left\{\overline{F}_{\beta,\Lambda_0}(T_i|Z_i)\right\}^{1-\Delta_i}\right] \\ & \times \left[\prod_{k=1}^{K} \prod_{i \in S_k} f_{\beta,\Lambda_0}(T_i|Z_i)\right] \times \left[\prod_{k=0}^{K} \prod_{i \in S_k} q_Z(Z_i)\right] \\ & \times \left[\prod_{k=1}^{K} \left\{\int_{\mathcal{Z}} \int_{A_k} f_{\beta,\Lambda_0}(t|Z) S_C(t) dt dQ_Z(Z)\right\}^{-n_k}\right], \end{aligned} \tag{19}$$

where $f_{\beta,\Lambda_0}(t|Z)$ and $\overline{F}_{\beta,\Lambda_0}(t|Z)$ are the conditional density function and survival function of $\widetilde{T}$ given Z with the baseline cumulative hazard function $\Lambda_0(t)$, respectively, $Q_Z(\cdot)$ and $q_Z(\cdot)$ denote the cumulative distribution and density function of Z, respectively, and $S_C(t)$ are the survival function of the censoring time C. Because the nonparametric portion (Q_Z, Λ_0, S_C) cannot be separated from the above likelihood function (19) that combines both the conditional parametric likelihood and the marginal semiparametric likelihood, Ding et al. (2014) developed an estimated maximum semiparametric empirical likelihood approach for estimation of the regression parameter.

By replacing Λ_0 with the Breslow-Aalen estimator $\hat{\Lambda}_0$ and S_C with the Nelson-Aalen estimator $\hat{S}_C$ based on the SRS data in the above joint likelihood, an estimated likelihood function was obtained as

$$l_5(\beta, Q_Z) = \sum_{i \in S_0} \log h_\beta\left(T_i, \Delta_i, Z_i\right) + \sum_{k=1}^{K} \sum_{i \in S_k} \log f_{\beta, \hat{\Lambda}_0}\left(T_i | Z_i\right)$$
$$+ \sum_{k=0}^{K} \sum_{i \in S_k} \log q_Z(Z_i) - \sum_{k=1}^{K} n_k \log \pi_k, \tag{20}$$

where $h_\beta(T_i, \Delta_i, Z_i) = \left(\frac{e^{\beta' Z_i}}{\sum_{l \in S_0} Y_l(T_i) e^{\beta' Z_l}} \right)^{\Delta_i}$, and for $k = 1, \ldots, K$,

$$\pi_k \equiv \int_{\mathcal{Z}} P_k(Z; \beta) dQ_Z(Z) \equiv \int_{\mathcal{Z}} \left(\int_{A_k} f_{\beta, \hat{\Lambda}_0}(t|Z) \hat{S}_C(t) dt \right) dQ_Z(Z),$$

which are the stratum-specific estimated probabilities of the failure time across all cases. Maximizing the estimated likelihood (20) with respect to (β, Q_Z) by a semi-parametric empirical approach without specifying Q_Z, the resulting profile likelihood function was obtained

$$l_6(\beta, \pi) = \sum_{i \in S_0} \log h_\beta\big(T_i, \Delta_i, Z_i\big) + \sum_{k=1}^{K} \sum_{i \in S_k} \log f_{\beta, \hat{\Lambda}_0}\big(T_i | Z_i\big)$$
$$- \sum_{i \in V} \log \left[n_0 \left\{ 1 + \sum_{k=1}^{K} \frac{n_k}{n_0 \pi_k} P_k\big(Z_i; \beta\big) \right\} \right] - \sum_{k=1}^{K} n_k \log \pi_k, \tag{21}$$

where $\pi' = (\pi_1, \ldots, \pi_K)$. The proposed estimator is the β that maximizes (21).

Yu et al. (2015) developed a weighted pseudo-score estimator for the regression parameters of the additive hazards model (5) for data from the general failure-time ODS design. Let ξ_i indicate, by the values 1 or 0, whether or not the ith subject is selected into SRS portion. Let η_{ik} denote whether or not the ith subject from the stratum A_k is selected into the supplemental sample. For estimating the regression parameter β, the following weighted pseudo-score equation was proposed by applying the inverse probability weighted approach,

$$U_{18}(\beta) = \sum_{i=1}^{N} \int_0^\tau w_i \left\{ Z_i(t) - \bar{Z}(t) \right\} \left\{ dN_i(t) - Y_i(t) \beta' Z_i(t) dt \right\} = 0,$$

where $\bar{Z}(t) = \frac{\sum_{i=1}^{N} w_i Y_i(t) Z_i(t)}{\sum_{i=1}^{N} w_i Y_i(t)}$, and the weights w_i were defined as

$$w_i = \left(1 - \Delta_i\right) \xi_i \left(\hat{\rho}_0 \hat{\rho}_V\right)^{-1} + \Delta_i \xi_i \left(1 - \zeta_i\right) \left(\hat{\rho}_0 \hat{\rho}_V\right)^{-1} + \Delta_i \xi_i \zeta_i$$
$$+ \Delta_i \left(1 - \xi_i\right) \sum_{k=1}^{K} \frac{\pi_k \left(1 - \hat{\rho}_0 \hat{\rho}_V\right) \zeta_{ik} \eta_{ik}}{\hat{\rho}_k \hat{\rho}_V},$$

where $\hat{\rho}_0 = n_0/n$, $\hat{\rho}_k = n_k/n$, $\hat{\rho}_V = n/N$, and $\zeta_i = \sum_{k=1}^{K} \zeta_{ik}$ and $\zeta_{ik} = I(T_i \in A_k)$.

The general failure-time ODS design proposed by Ding et al. (2014) is an improvement over the case–cohort design and the generalized case–cohort design, because the general failure-time ODS design allows the sample selection of cases to depend on the timing of disease endpoints, i.e., by oversampling subjects from the most informative regions. To reap in the benefit of such a general failure-time ODS design, Ding et al. (2014) developed a new inferential method and provided an estimated maximum semiparametric empirical likelihood estimator for the parameters of primary interest under the proportional hazards model. For the additive hazards model, which focuses on risk differences rather than risk ratios, Yu et al. (2015) studied a weighted pseudo-score estimating procedure for estimation of regression parameter. The proposed estimators have a closed form and are easy to compute. Some suggestions for using the proposed method by evaluating the relative efficiency of the proposed method against simple random sampling design and the optimal allocation of the subsamples for the proposed design were derived. The above researches suggest that the general failure-time ODS design is a biased-sampling design which can enhance study efficiency and reduce study cost. Such a general failure-time ODS design can be an important alternative to the case–cohort design and the generalized case–cohort design in survival-data related studies. Further developments on the general failure-time ODS design are desirable.

2.3 Other biased-sampling designs with failure time data

Other important biased-sampling designs with failure time data include length-biased sampling and interval sampling. When survival data arise from prevalent cases ascertained through a cross-sectional study, it is well known that the survivor function corresponding to these data is *length biased*. Length-biased sampling is frequently a convenient and economical sampling scheme for analyzing failure time data. The phenomenon of length bias has been first noticed in the context of anatomy by Wicksell (1925). Later systematically studied by Vardi (1982, 1989), Wang (1991), Correa and Wolfson (1999), Asgharian et al. (2002), and Asgharian and Wolfson (2005), among others. When analyzing prevalent cohort survival data with exposure variables, failure times are not a random sample from the study population. Thus, the corresponding exposure variables are also biased because they are associated with the long-term survivors. Related sampling issues have been discussed, e.g. Patil and Rao (1978), Patil et al. (1988), and Bergeron et al. (2008).

Wang (1996) first adopted the proportional hazards model to fit length-biased failure time data. The proposed estimation procedures used a bias-adjusted risk set sampling for the construction of the pseudo-likelihood. Ghosh (2008) proposed an estimating equation approach, which allows the length-biased data are subject to right censoring. Tsai (2009) obtained a pseudo-partial likelihood for proportional hazards models with biased-sampling data by embedding the biased-sampling data into left-truncated data. Shen et al. (2009) studied how to model exposure effects for length-biased data under transformation and accelerated failure time model. Qin and Shen (2010) proposed inverse weighted equation methods for estimating the regression parameter of the proportional hazards model. Qin et al. (2011) proposed new EM algorithms for the maximum likelihood estimators of the nonparametric and semiparametric proportional hazards models for right-censored length-biased data.

Often in practice, instead of right censored, the event time is interval censored, that is, the event time for a subject falls into some random time interval. Under the proportional hazards model, Li et al. (2008) considered case–cohort data with interval censoring, where the inspection time intervals were assumed to be fixed. Current status data are a special type of interval censored data in which the inspection time interval are random. Li and Nan (2011) considered a family of semiparametric regression models for the current status data in two-phase sampling designs, which include case–cohort designs as special cases. A weighted likelihood method was proposed by regarding two-phase sampling designs as a special missing data problem.

In many applications, interests often lie on the occurrence of two or more consecutive failure events and the relationship between event times. In such situations, data are often collected conditional on the first failure event which occurs within a specific time interval, and this fact induces bias. This type of sampling is referred to as *interval sampling*, where the first event is retrospectively identified and the subsequent failure events are observed during follow-up. Interval sampling occurs because only subjects with disease within a specific time interval can be included, and the data represent a nonrandomly sampled subset of the population.

Recent researches include that, among others, Zhu and Wang (2012) developed the statistical features and bias of observed data in relation to interval sampling. Semiparametric methods were proposed under semi-stationarity and stationarity. Zhu and Wang (2014) proposed nonparametric estimation of the association between bivariate failure times based on Kendall's tau for interval sampling data. A nonparametric estimator was derived, where the contribution of each comparable and order able pair was weighted by the inverse of the associated selection probability. Zhu and Wang (2015) obtained bias-corrected estimators of marginal survival functions and estimated association parameter of copula model by a two-stage procedure. Inference of association measure in copula model was developed, where exposure variables were incorporated into the survival distribution via the proportional hazards model.

3 ODS designs for multivariate failure time

An advantage of the case–cohort study design is that the subcohort can be used for multiple disease outcomes. Taking this advantage, in many studies, multiple case–cohort studies are conducted for different diseases using the same subcohort. A commonly used method for dealing with multiple disease outcomes is to analyze each disease separately. However, this approach does not allow comparison of the effects of risk factors for different diseases, because it does not account for the repeated use of the subcohort as well as the correlation between outcomes. Recently, several methodologies have been developed to analyze case–cohort and generalized case–cohort data with multiple disease outcomes.

Suppose that there are N independent subjects in a cohort study and there are K diseases outcomes of interest. Consider independent failure time response vectors

$\widetilde{T}_i = (\widetilde{T}_{i1}, \ldots, \widetilde{T}_{iK})'$ for $i = 1, \ldots, N$. Let C_{ik} denote the potential censoring time for outcome k of subject i. The observed time is $T_{ik} = \min(\widetilde{T}_{ik}, C_{ik})$. Let $\Delta_{ik} = I(\widetilde{T}_{ik} \leq C_{ik})$ denote the right censoring indicator for outcome k of subject i, $Y_{ik}(t) = I(T_{ik} \geq t)$ denote the at-risk process and $N_{ik}(t) = \Delta_{ik} I(T_{ik} \leq t)$ denote the counting process. Let $Z_{ik}(t)$ be a p-dimensional exposure variable corresponding to the kth disease outcome for subject i at time t. Let β be a p-dimensional parameter of interest. Let τ denote the study end time.

Kang and Cai (2009) proposed to fit data from the case–cohort design with multiple disease outcomes with a marginal intensity process model:

$$\lambda_{ik}\left(t|Z_{ik}(t)\right) = Y_{ik}(t)\lambda_{0k}(t)\exp\left\{\beta' Z_{ik}(t)\right\}, \quad i=1,\ldots,N;\ k=1,\ldots,K, \tag{22}$$

where $\lambda_{0k}(t)$ is an unspecified baseline hazard function for disease outcome k. A subject may experience all, only some, or even none of the K diseases. Model (22) can incorporate failure-type-specific effects and includes the model

$$\lambda_{ik}\left(t|Z^*_{ik}(t)\right) = Y_{ik}(t)\lambda_{0k}(t)\exp\left\{\beta_k' Z^*_{ik}(t)\right\},$$

as a special case. By defining $\beta = (\beta_1', \ldots, \beta_k', \ldots, \beta_K')'$ and $Z_{ik}(t) = (0, \ldots, 0, Z^*_{ik}(t)', 0, \ldots, 0,)'$, disease-specific effects can be obtained by model (22).

Under the case–cohort design, suppose a subcohort of size n is selected from the cohort by simple random sampling. Let ξ_i denote the indicator for the ith subject being selected into the subcohort and $P(\xi_i = 1) = \pi = n/N$ denote the selection probability of the ith subject. The observed data structure for the kth disease outcome of the ith subject is $(T_{ik}, \Delta_{ik}, \xi_i, Z_{ik}(t), 0 \leq t \leq T_{ik})$ when $\xi_i = 1$ or $\Delta_{ik} = 1$ and $(T_{ik}, \Delta_{ik}, \xi_i)$ when $\xi_i = 0$ and $\Delta_{ik} = 0$. Kang and Cai (2009) developed the following pseudo-partial-likelihood score equation for the estimation of β,

$$U_{19}(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\int_0^{\tau}\left[Z_{ik}(t) - \frac{\sum_{i=1}^{N} w_{ik}(t)Y_{ik}(t)Z_{ik}(t)e^{\beta' Z_{ik}(t)}}{\sum_{i=1}^{N} w_{ik}(t)Y_{ik}(t)e^{\beta' Z_{ik}(t)}}\right] dN_{ki}(t) = 0, \tag{23}$$

where $w_{ik}(t)$ is a time-varying weight function which has the form:

$$w_{ik}(t) = \Delta_{ik} + \left(1 - \Delta_{ik}\right)\xi_i \hat{\pi}_k^{-1}(t),$$

where

$$\hat{\pi}_k(t) = \frac{\sum_{i=1}^{N}\left(1 - \Delta_{ik}\right)\xi_i Y_{ik}(t)}{\sum_{i=1}^{N}\left(1 - \Delta_{ik}\right)Y_{ik}(t)}.$$

This weight function equals to 1 for the cases regardless of wether they belong to the subcohort or not, and $\hat{\pi}_k^{-1}(t)$ for the sampled censored subjects, where $\hat{\pi}_k(t)$ is the estimator of the true sampling probability π. $\hat{\pi}_k(t)$ denotes the number of sampled censored subjects divided by the number of censored subjects remaining in the risk set at time t, which means $\hat{\pi}_k(t)$ is constructed using only censored subjects. This type of time-varying weight function, as it was discussed under the univariate failure time context, may enhance the efficiency.

Kim et al. (2013) further improved efficiency for the case–cohort studies with multiple disease outcomes under the marginal proportional hazards regression model (22). The new weights

$$\tilde{w}_{ik}(t) = \left\{1 - \prod_{j=1}^{K}\left(1 - \Delta_{ij}\right)\right\} + \prod_{j=1}^{K}\left(1 - \Delta_{ij}\right)\xi_i \tilde{\pi}_k^{-1}(t),$$

where

$$\tilde{\pi}_k(t) = \frac{\sum_{i=1}^{N} \xi_i \left\{\prod_{j=1}^{K}(1 - \Delta_{ij})\right\} Y_{ik}(t)}{\sum_{i=1}^{N} \left\{\prod_{j=1}^{K}(1 - \Delta_{ij})\right\} Y_{ik}(t)},$$

were used to replace $w_{ik}(t)$ in the score function (23) to obtain the proposed pseudo-likelihood estimator. The weight function $\tilde{w}_{ik}(t)$ takes the failure status of the other diseases into consideration, and thus the proposed estimator will use the available exposure information for other diseases, which makes the estimators proposed by Kim et al. (2013) are more efficient than the estimators of Kang and Cai (2009).

Under the generalized case–cohort design, suppose a subcohort of size n is sampled from the cohort by simple random sampling. After the sampling of a subcohort, subsequent samplings of cases outside the subcohort follow. For the kth disease, let $n_{c,k}$ denote the number of cases that are outside the subcohort. Let η_{ik} denote the indicator for the ith subject outside the subcohort with the kth disease being selected into the sample. Denote by $q_k = P(\eta_{ik} = 1|\Delta_{ik} = 1, \xi_i = 0) = n_{c,k}/(N_k - n_k)$ the selection probability of subjects who have the kth disease but are outside the subcohort, where N_k and n_k denote the number of the kth disease cases in the cohort and in the subcohort, respectively. The observed data structure for the kth disease outcome of the ith subject is $(T_{ik},\ \Delta_{ik},\ \xi_i,\ \eta_{ik},\ Z_{ik}(t),\ 0 \le t \le T_{ik})$ when $\xi_i = 1$ or $\eta_{ik} = 1$ and $(T_{ik},\ \Delta_{ik},\ \xi_i,\ \eta_{ik})$ when $\xi_i = 0$ and $\eta_{ik} = 0$. When $q_k = 1$ for all k, it reduces to the original case–cohort design that samples all the cases outside the subcohort.

For the generalized case–cohort study with multiple disease outcomes, Kang and Cai (2009) also fitted data to the marginal proportional hazards model (22), and constructed the following weighted estimating functions for the estimation of the hazards regression parameter β:

$$U_{20}(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\int_0^{\tau} w_{ik}(t)\left[Z_{ik}(t) - \frac{\sum_{i=1}^{N} w_{ik}(t)Y_{ik}(t)Z_{ik}(t)e^{\beta' Z_{ik}(t)}}{\sum_{i=1}^{N} w_{ik}(t)Y_{ik}(t)e^{\beta' Z_{ik}(t)}}\right] dN_{ki}(t),$$

where

$$w_{ik}(t) = \Delta_{ik}\xi_i + \left(1 - \Delta_{ik}\right)\xi_i \hat{\pi}_k^{-1}(t) + \Delta_{ik}\left(1 - \xi_i\right)\eta_{ik}\hat{q}_k^{-1}(t), \qquad (24)$$

and

$$\hat{\pi}_k(t) = \frac{\sum_{i=1}^{N}\left(1 - \Delta_{ik}\right)\xi_i Y_{ik}(t)}{\sum_{i=1}^{N}\left(1 - \Delta_{ik}\right)Y_{ik}(t)}, \quad \hat{q}_k(t) = \frac{\sum_{i=1}^{N}\Delta_{ik}\left(1 - \xi_i\right)\eta_{ik}Y_{ik}(t)}{\sum_{i=1}^{N}\Delta_{ik}\left(1 - \xi_i\right)Y_{ik}(t)}. \qquad (25)$$

This idea is similar to their proposed method for the original case–cohort design (see 23). Subcohort cases are weighted by 1, and subjects censored for disease k in the subcohort are weighted by $\hat{\pi}_k^{-1}(t)$. The sampled non-subcohort cases are weighted by the inverse of their estimated sampling probabilities, $\hat{q}_k^{-1}(t)$, where $\hat{q}_k(t)$ denotes the number of sampled non-subcohort cases with the kth disease outcome divided by the number of non-subcohort cases with the kth disease outcome remaining in the risk set at time t.

Kang et al. (2013) considered fitting marginal additive hazards regression model for the generalized case–cohort designs with multiple disease outcomes. The model is

$$\lambda_{ik}\big(t|Z_{ik}(t)\big) = \lambda_{0k}(t) + \beta' Z_{ik}(t), \quad i = 1, \ldots, N; \; k = 1, \ldots, K, \tag{26}$$

where $\lambda_{0k}(t)$ is an unspecified baseline hazard function for disease outcome k. Model (26) also incorporates disease-specific effects like model (22). Kang et al. (2013) proposed the weighted estimating equation:

$$U_{21}(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\int_0^{\tau} w_{ik}(t)\left\{Z_{ki}(t) - \bar{Z}_k(t)\right\}\left\{dN_{ik}(t) - Y_{ik}(t)\beta' Z_{ik}(t)dt\right\} = 0,$$

where $\bar{Z}_k(t) = \frac{\sum_{i=1}^{N} w_{ik} Y_{ik}(t) Z_{ik}(t)}{\sum_{i=1}^{N} w_{ik} Y_{ik}(t)}$, and $w_{ik}(t)$ and $\hat{\pi}_k(t)$ have the same definitions as (24) and (25). The resulting estimator possesses a closed form:

$$\hat{\beta}_{21} = \left[\sum_{i=1}^{N}\sum_{k=1}^{K}\int_0^{\tau} w_{ik}(t)Y_{ik}(t)\big\{Z_{ik}(t) - \bar{Z}_k(t)\big\}^{\otimes 2} dt\right]^{-1} \times \left[\sum_{i=1}^{N}\sum_{k=1}^{K}\int_0^{\tau} w_{ik}(t)\big\{Z_{ik}(t) - \bar{Z}_k(t)\big\} dN_{ik}(t)\right].$$

Besides the multiple disease outcome data, other kinds of multivariate failure time data have become increasingly common in practice as a result of growing interest in studying disease incidence and clustering due to environmental factors and genetics. Lu and Shih (2006) proposed case–cohort designs adapted to clustered failure time data. The main principle of their proposed case–cohort designs is to determine the random subcohort from which the exposure data are assembled in addition to those from all cases. Lu and Shih (2006) considered several sampling schemes and developed the estimation procedures by fitting the proposed case–cohort design with clustered data under the proportional hazards model. The proposed approaches were derived by the principle similar to that of the pseudo-likelihood function of Self and Prentice (1988), but extended to accommodate the proposed subcohort selection procedures and to account for intracluster association.

The above literature on ODS designs with a univariate and multivariate failure time also studied the asymptotic properties, e.g. consistency and asymptotic normality, of the proposed estimators. Feasible estimators of variance, usually the small-sample expression of the large-sample formula, are naturally derived from the asymptotic variance of the proposed estimators. Some technical challenges arise from the biased-sampling designs because the observations are not independent. For example, asymptotic properties have been established by using techniques such as empirical likelihood method, empirical process theory, martingale convergence theory, and U-statistics theory, etc. The key for the studies of theoretical results is to address the challenges introduced by the counterpart data of the subcohort or supplemental samples in the ODS designs with survival data.

4 Discussion

Epidemiologic studies often require a long follow-up of subjects in order to observe meaningful outcome results. The cost for a large cohort study and a long period of follow-up time could be very expensive. Efficient sampling designs and statistical methods, which can reduce the study cost and improve the study power under a limited budget, are always desirable. Several cost-effective biased-sampling designs for failure time data have been developed and various estimating procedures have been proposed. This paper reviewed recent progresses in ODS designs with failure time data.

One advantage of the general failure-time ODS design is, while providing an overall information, to allow the sample selection of cases to depend on the timing of disease endpoints. The general failure-time ODS design is an improvement of the simple random sampling design, the case–cohort design and the generalized case–cohort design, especially in the situations that the disease rate is not low or investigators have not enough budget to sample all cases. Despite the progresses in the development of analyzing failure time data from a biased-sampling design, the methodologies to address data from such a general failure-time ODS design have been limited.

Extensions of the constructions of weighted estimating equations or likelihood functions would be worthwhile to consider. One extension is to adopt time-varying weights instead of weights based on simple sampling probabilities. Another extension is to include information available from the first-phase data in estimating equations or likelihood functions. For example, if the observed times are available for all the subjects in the cohort, incorporating failure times or censoring times of those who do not belong to the ODS samples in constructing estimating equations or likelihood functions could enhance the efficiency. Due to the fact that applying a stratified sampling scheme for selecting the subcohort could improve the efficiency of case–cohort designs, future developments of a stratified failure-time ODS design is justified, where the SRS portion is selected by a stratified sampling scheme.

In more and more applications, investigators tend to take interests in multivariate failure-time outcomes. Future researches include incorporating information of some always observed auxiliary variables in the weight functions of the estimating equations to improve efficiency further. For example, similar idea of Kulich and Lin (2004) could be modified to fit case–cohort data with multiple disease outcomes. Recent works of case–cohort design with multivariate failure times have focused on estimating equation approaches. Specifying the joint distribution of the correlated failure times from the same subject, nonparametric maximum likelihood estimations based on the joint likelihood function for case–cohort data will derive more efficient estimators. In order to make use of the advantage of an ODS design, which could oversample from the regions of most information, the development of a multivariate failure-time ODS design will be an interesting topic.

Acknowledgements This research is supported in part by U.S. National Institutes of Health (R01ES021900, P01CA142538 to H.Z. and J.C.), National Science Foundation of China (11101314 to J.D.).

References

Asgharian M, M'Lan CE, Wolfson DB (2002) Length-biased sampling with right censoring: an unconditional approach. J Am Stat Assoc 97:201–209

Asgharian M, Wolfson DB (2005) Asymptotic behaviour of the npmle of the survivor function when the data are length-biased and subject to right censoring. Ann Stat 33:2109–2131

Barlow W (1994) Robust variance estimation for the case-cohort design. Biometrics 50:1064–1072

Bergeron PJ, Asgharian M, Wolfson DB (2008) Covariate bias induced by length-biased sampling of failure times. J Am Stat Assoc 103:737–742

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J (2000) Exposure stratified case-cohort designs. Lifetime Data Anal 6:39–58

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

Breslow NE, Holubkov R (1997) Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J R Stat Soc B 59:447–461

Breslow NE, McNeney B, Wellner JA (2003) Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. Ann Stat 31:1110–1139

Breslow NE, Wellner JA (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. Scand J Stat 34:86–102

Cai J, Zeng D (2004) Sample size/power calculation for case-cohort studies. Biometrics 60:1015–1024

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

Chatterjee N, Chen YH, Breslow NE (2003) A pseudo-score estimator for regression problems with two-phase sampling. J Am Stat Assoc 98:158–168

Chen HY (2001a) Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. J Am Stat Assoc 96:1446–1458

Chen HY (2001b) Fitting semiparametric transformation regression models to data from a modified case-cohort design. Biometrika 88:255–268

Chen K (2001c) Generalized case-cohort sampling. J R Stat Soc B 63:791–809

Chen K, Lo S (1999) Case-cohort and case-control analysis with Coxs model. Biometrika 86:755–764

Cornfield J (1951) A method of estimating comparative rates from clinical data: applications to cancer of lung, breast, and cervix. J Natl Cancer I 11:1269–1275

Correa JA, Wolfson DB (1999) Length-bias: some characterizations and applications. J Stat Comput Sim 64:209–219

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

Ding J, Liu L, Peden DB, Kleeberger SR, Zhou H (2012) Regression analysis for a summed missing data problem under an outcome-dependent sampling scheme. Can J Stat 40:282–303

Ding J, Zhou H, Liu L, Cai J, Longnecker MP (2014) Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. Biostatistics 15:636–650

Ghosh D (2008) Proportional hazards regression for cancer studies. Biometrics 64:141–148

Imbens GW, Lancaster T (1996) Efficient estimation and stratified sampling. J Econ 74:289–318

Kalbfleisch JD, Lawless JF (1988) Likelihood analysis of multi-state models for disease incidence and mortality. Stat Med 7:147–160

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

Kang S, Cai J, Chambless L (2013) Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the Atherosclerosis Risk in Communities (ARIC) study. Biostatistics 14:28–41

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

Kong L, Cai J (2009) Case-cohort analysis with accelerated failure time model. Biometrics 65:135–142

Kong L, Cai J, Sen PK (2004) Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. Biometrika 91:305–319

Kulich M, Lin DY (2000) Additive hazards regression for case-cohort studies. Biometrika 87:73–87

Kulich M, Lin DY (2004) Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc 99:832–844

Lawless JF, Wild CJ, Kalbfleisch JD (1999) Semiparametric methods for response-selective and missing data problems in regression. J R Stat Soc B 61:413–438

Li Z, Gilbert P, Nan B (2008) Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. Biometrics 64:1247–1255

Li Z, Nan B (2011) Relative risk regression for current status data in case-cohort studies. Can J Stat 39:557–577

Lin DY, Ying Z (1993) Cox regression with incomplete covariate measurements. J Am Stat Assoc 88:1341–1349

Lu S, Shih JH (2006) Case-cohort designs and analysis for clustered failure time data. Biometrics 62:1138–1148

Lu W, Tsiatis AA (2006) Semiparametric transformation models for the case-cohort study. Biometrika 93:207–214

Nan B, Yu M, Kalbfleisch JD (2006) Censored linear regression for case-cohort studies. Biometrika 93:747–762

Patil GP, Rao CR (1978) Weighted distributions and size-biased sampling with applications to wildlife population and human families. Biometrics 34:179–189

Patil GP, Rao CR, Zelen M (1988) Weighted distributions. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences. Wiley, New York, pp 565–571

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

Qi L, Wang CY, Prentice RL (2005) Weighted estimators for proportional hazards regression with missing covariates. J Am Stat Assoc 100:1250–1263

Qin J, Ning J, Liu H, Shen Y (2011) Maximum likelihood estimations and EM algorithms with length-biased data. J Am Stat Assoc 106:1434–1449

Qin J, Shen Y (2010) Statistical methods for analyzing right-censored length-biased data under Cox model. Biometrics 66:382–392

Qin G, Zhou H (2011) Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. Biostatistics 12:506–520

Samuelsen SO, Anestad H, Skrondal A (2007) Stratified case-cohort analysis of general cohort sampling designs. Scand J Stat 34:103–119

Schildcrout JS, Heagerty PJ (2008) On outcome dependent sampling designs for longitudinal binary response data with time-varying covariates. Biostatistics 9:735–749

Schildcrout JS, Mumford SL, Chen Z, Heagerty PJ, Rathouz PJ (2012) Outcome dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. Stat Med 31:2441–2456

Schildcrout JS, Rathouz PJ (2010) Longitudinal studies of binary response data following case-control and stratified case-control sampling: design and analysis. Biometrics 66:365–373

Schill W, Jockel KH, Drescher K, Timm J (1993) Logistic analysis in case-control studies under validation sampling. Biometrika 80:339–352

Scott AJ, Wild CJ (1991) Fitting logistic regression models in stratified case-control studies. Biometrics 47:497–510

Self SG, Prentice RL (1988) Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Stat 16:64–81

Shen Y, Ning J, Qin J (2009) Analyzing length-biased data with semiparametric transformation and accelerated failure time models. J Am Stat Assoc 104:1192–1202

Song R, Zhou H, Kosorok MR (2009) On semiparametric efficient inference for two-stage outcome dependent sampling with a continuous outcome. Biometrics 96:221–228

Sun J, Sun L, Flournoy N (2004) Addictive hazards model for competing risks analysis of the case-cohort design. Commun Stat Theor M 33:351–366

Tsai WY (2009) Pseudo-partial likelihood for proportional hazards models with biased-sampling data. Biometrika 96:601–615

Vardi Y (1982) Nonparametric estimation in the presence of length bias. Ann Stat 10:616–620

Vardi Y (1989) Multiplicative censoring, renewal processes, deconvolution and decreasing density. Biometrika 76:751–761

Wang MC (1991) Nonparametric estimation from cross-sectional survival data. J Am Stat Assoc 86:130–143

Wang MC (1996) Hazards regression analysis for length-biased data. Biometrika 83:343–354

Wang X, Zhou H (2006) A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. Biometrics 62:1149–1160

Wang X, Zhou H (2010) Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. Biometrics 66:502–511

Weaver MA (2001) Semiparametric methods for continuous outcome regression models with covariate data from an outcome dependent subsample. PhD Thesis, University of North Carolina, Chapel Hill

Weaver MA, Zhou H (2005) An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. J Am Stat Assoc 100:459–469

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

White JE (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 115:119–128

Wicksell SD (1925) The corpuscle problem: a mathematical study of a biometric problem. Biometrika 17:84–99

Yu J, Liu Y, Sandler DP, Zhou H (2015) Statistical inference for the additive hazards model under outcome-dependent sampling. Can J Stat 43(3):436–453

Zhou H, Weaver MA, Qin J, Longnecker M, Wang MC (2002) A semiparametric empirical likelihood method for data from an outcome dependent sampling scheme with a continuous outcome. Biometrics 58:413–421

Zhou H, Qin G, Longnecker MP (2011a) A partial linear model in the outcome-dependent sampling setting to evaluate the effect of prenatal PCB exposure on cognitive function in children. Biometrics 67:876–885

Zhou H, Song R, Qin J (2011b) Statistical inference for a two-stage outcome dependent sampling design with a continuous outcome. Biometrics 67:194–202

Zhou H, Wu Y, Liu Y, Cai J (2011c) Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. Biostatistics 12:521–534

Zhou H, You J, Qin G, Longnecker MP (2011d) A partially linear regression model for data from an outcome-dependent sampling design. J R Stat Soc C 60:559–574

Zhu H, Wang MC (2012) Analysing bivariate survival data with interval sampling and application to cancer epidemiology. Biometrika 99:345–361

Zhu H, Wang MC (2014) Nonparametric inference on bivariate survival data with interval sampling: association estimation and testing. Biometrika 101:519–533

Zhu H, Wang MC (2015) A semi-stationary Copula model approach for bivariate survival data with interval sampling. Int J Biostat 11:151–173