Biostatistics (2018) **19**, 1, *pp.* 54–70
doi:10.1093/biostatistics/kxx024
Advance Access publication on May 18, 2017

Semiparametric model and inference for spontaneous abortion data with a cured proportion and biased sampling

JIN PIAO

Department of Biostatistics, The University of Texas School of Public Health, 1200 Pressler Street, Houston, TX 77030, USA

JING NING*

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Houston, TX 77030, USA

jning@mdanderson.org

CHRISTINA D. CHAMBERS

Department of Pediatrics, University of California, 9500 Gilman Drive, San Diego, La Jolla, CA 92093, USA and Department of Family Medicine and Public Health, University of California, 9500 Gilman Drive, San Diego, La Jolla, CA 92093, USA

RONGHUI XU

Department of Family Medicine and Public Health, University of California, 9500 Gilman Drive, San Diego, La Jolla, CA 92093, USA and Department of Mathematics, University of California, 9500 Gilman Drive, San Diego, La Jolla, CA 92093, USA

Summary

Evaluating and understanding the risk and safety of using medications for autoimmune disease in a woman during her pregnancy will help both clinicians and pregnant women to make better treatment decisions. However, utilizing spontaneous abortion (SAB) data collected in observational studies of pregnancy to derive valid inference poses two major challenges. First, the data from the observational cohort are not random samples of the target population due to the sampling mechanism. Pregnant women with early SAB are more likely to be excluded from the cohort, and there may be substantial differences between the observed SAB time and those in the target population. Second, the observed data are heterogeneous and contain a "cured" proportion. In this article, we consider semiparametric models to simultaneously estimate the probability of being cured and the distribution of time to SAB for the uncured subgroup. To derive the maximum likelihood estimators, we appropriately adjust the sampling bias in the likelihood function and develop an expectation-maximization algorithm to overcome the computational challenge. We apply the empirical process theory to prove the consistency and asymptotic normality of the estimators.

*To whom correspondence should be addressed.

© The Author 2017. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

We examine the finite sample performance of the proposed estimators in simulation studies and illustrate the proposed method through an application to SAB data from pregnant women.

Keywords: Biased sampling; Cure rate model; Left truncation; EM algorithm.

1. Introduction

During pregnancy, women have consistently low rates of compliance with treatment recommendations for medical conditions not related to their pregnancy. Major barriers to compliance among pregnant women have repeatedly been shown to include fear of the safety of the treatments for themselves and for their developing fetus. Some medications used to treat autoimmune disease have been associated with spontaneous abortion (SAB) during pregnancy (Visser *and others*, 2009; Skorpen *and others*, 2016). Hence, it is essential to evaluate and understand the safety and risk of treatments given to pregnant women in order to help both clinicians and pregnant women make better treatment decisions. This work was motivated by studies conducted by the Organization of Teratology Information Specialists (OTIS), which is a North American network of university or hospital-based teratology services that counsel between 70 000 and 100 000 pregnant women every year. The OTIS autoimmune disease in pregnancy database included 964 pregnant women between 2005 and 2012. During the studies, the pregnant women participated in phone interviews and recorded information in a diary throughout their pregnancy. A final outcome phone interview was conducted shortly after the pregnancy ended. While SAB as an outcome of interest is, at first sight, and perhaps ultimately, a binary endpoint, our recruitment of pregnant women leads to biased sampling. Following the research interest to assess the effects of medication exposure on SAB (Xu and Chambers, 2011; Chambers *and others*, 2011), we evaluate the relationship between the use of medications for autoimmune disease during pregnancy and the probability of experiencing SAB, as well as the time to SAB. Specifically, besides the effect of medications for autoimmune disease on the risk of experiencing SAB, we are interested in evaluating whether the use of the medications will significantly affect the distribution of time to SAB for pregnant women who experience SAB (the uncured group).

In the medical literature, SAB is defined as the natural death of an embryo or fetus before 20 weeks of gestation; any pregnancy loss after 20 weeks is called still birth (Medical Encyclopedia, National Institutes of Health website: https://www.nlm.nih.gov/medlineplus/ency/ article/001488.htm). Using this definition of SAB, the pregnant women who do not experience SAB are considered to be "cured." Hence, the population is a mixture of two subgroups: those who are non-susceptible (cured) and those who are susceptible (uncured) to SAB. Note that we are able to observe the SAB status (membership of the two subgroups) for uncensored subjects, which is different from the classical cured data. Cure rate models that consider such population heterogeneity have been well studied in the literature for time-to-event data. Most survival cure rate models have been developed on the basis of mixture models (Peng and Dear, 2000; Sy and Taylor, 2000). Various survival regression models have been considered including Cox proportional hazards models (Sy and Taylor, 2000; Kuk and Chen, 1992) and accelerated failure time models (Zhang and Peng, 2009; Li and Taylor, 2002). Also, several cure rate models have been developed along the lines of non-mixture models (Chen *and others*, 1999; Zeng *and others*, 2006).

However, the existing methods to handle survival data with a cured proportion cannot be directly applied to our motivating data because of the unique data structure of biased sampling. The data consist only of pregnant women who have not experienced the failure event, SAB, at the time of enrollment. In other words, pregnant women who have early SAB events are less likely to be included in the study and thus tend to represent left-truncated data, as indicated in Figure 1. Such a sampling bias due to left truncation is also confirmed by exploratory analysis in which the empirical SAB rate is only 7%, which is much lower than the known incidence rate (around 12%) in the general population (Wilcox *and others*, 1988). Determining the best way to adjust for sampling bias has been a longstanding statistical problem.

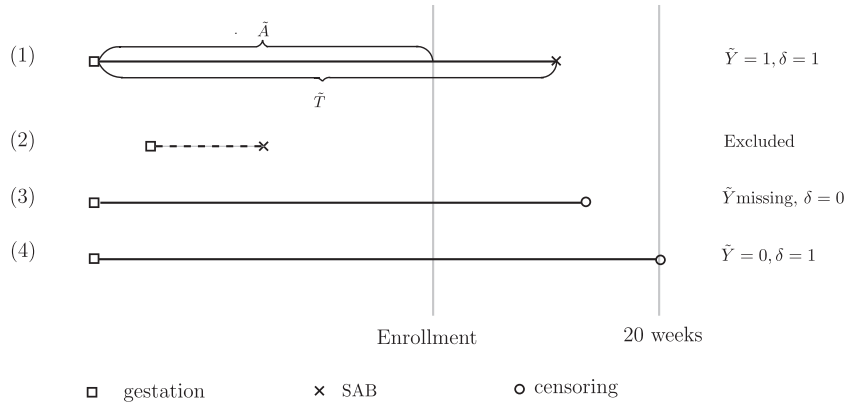

Fig. 1. Survival data from a cure model that is subject to biased sampling. Patients with IDs (1), (3), and (4) are sampled, whereas Patient (2) is excluded.

Statistical methods for analyzing survival data subject to biased sampling have been actively studied by Wang *and others* (1986), Shen *and others* (2009), Tsai (2009), Qin *and others* (2011), Kim *and others* (2013), Ning *and others* (2014) and more. However, most of the methods used in the aforementioned publications have two limitations. First, they focus on a special type of left-truncated data in which the incidence of the initial event (e.g., pregnancy) is constant over time. In our motivating study, the p-value from the test of the stationarity assumption is 0.0001 (Addona and Wolfson, 2006), indicating that such a stationarity assumption is not satisfied. Second, to the best of our knowledge, the existing methods for adjusting sampling bias have not considered the data that include a cured proportion. Our goal is to provide models and estimating procedures to simultaneously account for sampling bias and population heterogeneity.

The remainder of the article is organized as follows. In Section 2, we introduce the notations and mixture models in which the SAB status is modeled by logistic regression and the time to SAB for the uncured subgroup is characterized by a semiparametric proportional hazards model. In Section 3, we derive the full likelihood function with appropriate adjustment for biased sampling and the cured proportion, and then derive an expectation-maximization (EM) algorithm to solve the computational challenge. In Section 4, we establish the asymptotic properties of the proposed estimators. In Section 5, we report the results of simulation studies to assess the finite sample performance of the proposed method. We apply our method to the SAB data in Section 6 and provide concluding remarks in Section 7. We provide details for the proofs of the asymptotic properties in the supplementary materials available at *Biostatistics* online.

2. NOTATIONS AND MODEL

By the definition of SAB, the loss of a pregnancy prior to 20 weeks of gestation, some pregnant women appear to be free of the risk of SAB, which we consider to be the cured population. Considering that the observed data are subject to biased sampling, we introduce notations first for the target population and then for the observed biased population. Let $\widetilde{Y}$ be the status indicating whether a subject experiences the SAB event ($\widetilde{Y} = 1$) or does not experience the SAB event ($\widetilde{Y} = 0$). For subjects with $\widetilde{Y} = 1$ (uncured population), let $\widetilde{T}$ be the unbiased duration from the first day of the last menstrual period to the SAB, with a density function $f(t|\boldsymbol{Z})$ and survival function $S(t|\boldsymbol{Z})$, where $\boldsymbol{Z}$ is a $q \times 1$ vector of the covariates. Define $\tilde{A}$ to be the duration from the beginning of the pregnancy to study entry. Under biased sampling, only pregnant women who did not experience SAB prior to their enrollment were enrolled, and those

who experienced SAB were excluded from the study. In other words, we have a sampling constraint of $\widetilde{T} > \widetilde{A}$. Let (T, A) be the observed biased counterparts of unbiased $(\widetilde{T}, \widetilde{A})$. Define the censoring time from the study enrollment and censoring indicator to be C and $\delta = I(T < A + C)$, respectively. With potential right censoring, the observed time is denoted as $X = \min(T, A + C, \tau)$, where τ is the time after which an individual is no longer considered to be susceptible to the event (i.e., $\tau = 20$ in the SAB data). Note that the indicator Y is not available for subjects with $\delta = 0$. For the cured population ($\widetilde{Y} = 0$), we define $\tau \leq T \leq C$ for notational consistency. Throughout this article, we assume that (i) $\widetilde{A}$ and $\widetilde{T}$ are conditionally independent given covariates $\boldsymbol{Z}$ and $\widetilde{Y} = 1$, and (ii) C is conditionally independent of (A, T) given covariates $\boldsymbol{Z}$. Figure 1 illustrates the sampling mechanism of the data that has a cured proportion and is subject to biased sampling.

We impose a logistic regression for the risk of SAB ($\widetilde{Y}$) and a proportional hazards model for the time to SAB ($\widetilde{T}$) for subjects with $\widetilde{Y} = 1$:

$$P(\widetilde{Y} = 1|\boldsymbol{Z}_1) = \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_1)}{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_1)}, \tag{2.1}$$

$$\lambda_Z(t) = \lambda(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z}), \tag{2.2}$$

where $\boldsymbol{Z}_1 = (1, \boldsymbol{Z}^T)^T$ and $\lambda(t)$ is an unspecified baseline hazard function. For simplicity of notation, we use the same covariates in both models; however, it is easy to accommodate different sets of covariates for the two models. As discussed by Sy and Taylor (2000) and Taylor (1995), one essential assumption for model identifiability given survival data with a cured proportion is the zero-tail constraint, which refers to the conditional survival function as zero for a value of time greater than the longest time to the event of interest. In our SAB data, the constraint assumption is naturally satisfied by the definition of SAB, the natural death of an embryo or fetus before 20 weeks of gestation.

3. Likelihood and estimation procedure

Recall that the unbiased time-to-SAB data are not directly observed. Instead, the biased samples and their corresponding covariates are observed. We first consider the length-biased data, and then extend the likelihood and estimating procedure to the general left-truncated data. Length-biased data are a special case of left-truncated data in which the truncation times are uniformly distributed on a defined interval $(0, \tau)$.

Given covariates $\boldsymbol{Z}$, the probability of a subject being selected from the target population is $\pi(\boldsymbol{Z}) = P(\widetilde{T} > \widetilde{A}|\boldsymbol{Z})$, which equals $E(\widetilde{T}|\boldsymbol{Z})/\tau$ for the length-biased data. Given that the population is a mixture of cured and uncured components, the marginal survival function of the observed time T is

$$S_o(t|\boldsymbol{Z}) = \frac{P(\widetilde{Y} = 1|\boldsymbol{Z})S(t|\boldsymbol{Z})}{P(\widetilde{T} > \widetilde{A}|\boldsymbol{Z})} + P(\widetilde{Y} = 0|\boldsymbol{Z}),$$

where $S(t|\boldsymbol{Z}) = \exp\left\{-\Lambda(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z})\right\}$.

Consider a study with a sample of n subjects, with observed data of $\{\mathcal{O}_i = (X_i, A_i, \delta_i, Y_i\delta_i, \boldsymbol{Z}_i), i = 1, \ldots n\}$. Let $0 < t_1 < t_2 < \cdots < t_K$ denote the ordered distinct observed time points including censored and uncensored time, which is different from the hazard estimator in the traditional survival analysis (Qin *and others*, 2011). The true baseline hazard function $\lambda(.)$ is unspecified under the Cox model and is an infinite-dimensional parameter. In our estimation procedure, following the nonparametric maximum likelihood principle (Vardi, 1989; Qin *and others*, 2011), we assume that the estimated $\lambda(.)$ has positive masses only at distinct observed time points $(t_1, \ldots t_K)$, where the value of K depends on the observed data

and can reach infinity as $n \to \infty$. Given the conditional independence assumptions, the full likelihood function of the observed data conditional on the covariates is proportional to

$$L_n(\boldsymbol{\psi}) = \prod_{i=1}^{n} \left\{ \frac{P(\widetilde{Y}_i = 1|\boldsymbol{Z}_i, \boldsymbol{\alpha}) f(X_i|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})}{P(\widetilde{T}_i > \widetilde{A}_i|\boldsymbol{Z}_i, \boldsymbol{\psi})} \right\}^{Y_i \delta_i} P(\widetilde{Y}_i = 0|\boldsymbol{Z}_i, \boldsymbol{\alpha})^{\delta_i(1-Y_i)} S_o(X_i|\boldsymbol{Z}_i, \boldsymbol{\psi})^{1-\delta_i}, \tag{3.1}$$

where the parameter vector of interest is denoted as $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ and the density function $f(t|\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\lambda}))$ is denoted as $f(t|\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\lambda})) = \lambda(t) \exp(\boldsymbol{\beta}'\boldsymbol{Z}) \exp\left\{-\Lambda(t) \exp(\boldsymbol{\beta}'\boldsymbol{Z})\right\}$. Note that the sampling probability $P(\widetilde{T}_i > \widetilde{A}_i|\boldsymbol{Z}_i, \boldsymbol{\psi})$ involves the parameter $\boldsymbol{\lambda}$;C hence, directly maximizing the observed likelihood or the profile likelihood method is computationally prohibitive due to a lack of an analytical expression for the optimal value of parameter $\boldsymbol{\lambda}$. To overcome this computational challenge, we derive an EM algorithm that naturally incorporates the bias sampling mechanism into a missing data framework.

3.1. *EM algorithm under length-biased sampling*

There are two missing components in the observed data. First, the SAB status is not observable for any subject with censored survival time ($\delta = 0$). Conditional on the observed data, we can derive the expectation of Y_i as

$$E(Y_i|\mathcal{O}_i, \boldsymbol{\psi}) = P(Y_i = 1|\mathcal{O}_i, \boldsymbol{\psi}) = \frac{P(\widetilde{Y}_i = 1|\boldsymbol{Z}_i, \boldsymbol{\alpha}) S(X_i|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})}{P(\widetilde{Y}_i = 0|\boldsymbol{Z}_i, \boldsymbol{\alpha}) + P(\widetilde{Y}_i = 1|\boldsymbol{Z}_i, \boldsymbol{\alpha}) S(X_i|\boldsymbol{Z}_i \boldsymbol{\beta}, ,\lambda)}. \tag{3.2}$$

Next, we treat the truncated observations as missing data. For any subject i in the observed data, the data generating mechanism can be considered as sampling the unbiased time $(\widetilde{T}, \widetilde{A})$ for a random m_i times until $\widetilde{T} > \widetilde{A}$. This random integer m_i then follows a geometric distribution with parameter $\pi(\boldsymbol{Z}_i) = P(\widetilde{T} > \widetilde{A}|\boldsymbol{Z}_i)$. We denote the truncated (unobserved) data corresponding to subject i by $\mathcal{O}_i^* = \{(T_{il}^*, A_{il}^*, Y_{il}^* = 1, \boldsymbol{Z}_i),\ T_{il}^* < A_i^*, l = 1, \ldots, m_i\}$. Then the complete data for the *ith* subject include the observed data O_i, and unobserved data $\{(1 - \delta_i) Y_i, O_i^*)\}$. Accordingly, the log-likelihood function of the complete data is

$$\begin{aligned} l_c(\boldsymbol{\psi}) = & \sum_{i=1}^{n} Y_i \sum_{j=1}^{K} \sum_{l=1}^{m_i} I(T_{il}^* = t_j) \left[\boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + \log f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})\right] \\ & + \sum_{i=1}^{n} Y_i \sum_{j=1}^{K} I(X_i = t_j) \left[\boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + \delta_i \log f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})\right. \\ & \left. + (1 - \delta_i) \log S(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})\right] - \sum_{i=1}^{n} (1 - Y_i) \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\}, \end{aligned} \tag{3.3}$$

where $f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \lambda_j \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i) \exp\left\{-\sum_{l=1}^{j} \lambda_l \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\right\}$ and $S(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \exp\left\{-\sum_{l=1}^{j} \lambda_l \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\right\}$. We first select initial values $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\lambda}^{(0)})$, and let $\boldsymbol{\psi}^{(k)}$ denote the estimates of the parameters in the *kth* iteration. Following the principle of the EM algorithm, in the E-step of the $(k+1)th$ iteration, we calculate the conditional expectation of the log-likelihood function of the complete data based on the observed data and the estimated parameters from the last iteration,

$$l_E(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \omega(Y_i) \sum_{j=1}^{K} w_{ij} \left[\boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})\right] \tag{3.4}$$

$$+\sum_{i=1}^{n}\omega(Y_i)\sum_{j=1}^{K}I(X_i=t_j)\big[\boldsymbol{\alpha}'\boldsymbol{Z}_i-\log\left\{1+\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\}+\delta_i\log f(t_j|\boldsymbol{Z}_i,\boldsymbol{\beta},\boldsymbol{\lambda})$$

$$+(1-\delta_i)\log S(t_j|\boldsymbol{Z}_i,\boldsymbol{\beta},\boldsymbol{\lambda})\big]-\sum_{i=1}^{n}\{1-\omega(Y_i)\}\log\left\{1+\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\},$$

where $\omega(Y_i)=\delta_i Y_i+(1-\delta_i)E(Y_i|\mathcal{O}_i,\boldsymbol{\psi}^{(k)})$, the expected number of truncated latent subjects who would experience the failure event at t_j is

$$w_{ij}=E\left[\sum_{l=1}^{m_i}I(T_{il}^*=t_j)\Big|\mathcal{O},\boldsymbol{\psi}^{(k)}\right]=E(m_i|\mathcal{O},\boldsymbol{\psi}^{(k)})E\Big[I(T_{il}^*=t_j)|\mathcal{O}_i,\boldsymbol{\psi}^{(k)}\Big]=\frac{1-t_j/t_K}{\pi(\boldsymbol{Z}_i)}f(t_j|\boldsymbol{Z}_i,\boldsymbol{\beta}^{(k)},\boldsymbol{\lambda}^{(k)}),$$

and $\pi(\boldsymbol{Z}_i)=P(\widetilde{T}_i>\widetilde{A}_i|\boldsymbol{Z}_i,\boldsymbol{\psi}^{(k)})=\sum_{j=1}^{k}t_j f(t_j|\boldsymbol{Z}_i,\boldsymbol{\beta}^{(k)},\boldsymbol{\lambda}^{(k)})/t_K$.

In the M-step, we maximize the expected complete log-likelihood function (3.4) to update the parameter estimates. The estimates can be updated by solving the corresponding score equation, defined as the first derivative of the expected complete log-likelihood. The score equation of $\boldsymbol{\alpha}$ is

$$\sum_{i=1}^{n}\left[\omega(Y_i)\sum_{j=1}^{K}\left\{w_{ij}+I(X_i=t_j)\right\}\frac{\boldsymbol{Z}_{i1}}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_{i1})}-\{1-\omega(Y_i)\}\frac{\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_{i1})\boldsymbol{Z}_{i1}}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_{i1})}\right].\tag{3.5}$$

By solving the score equation of $\boldsymbol{\lambda}$, the maximizer for the baseline hazard can be written as a function of $\boldsymbol{\beta}$,

$$\lambda_j=\frac{\sum_{i=1}^{n}\omega(Y_i)\left\{w_{ij}+I(X_i=t_j)\delta_i\right\}}{\sum_{i=1}^{n}\sum_{k=j}^{K}\omega(Y_i)\left\{w_{ik}+I(X_i=t_k)\right\}\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)}.\tag{3.6}$$

After plugging equation (3.6) into the score equation of $\boldsymbol{\beta}$, we have the following estimating equation set

$$\sum_{i=1}^{n}\left\{\omega(Y_i)\boldsymbol{Z}_i\sum_{j=1}^{K}\Big[w_{ij}+I(X_i=t_j)\delta_i\right.\tag{3.7}$$

$$\left.-\left\{w_{ij}+I(X_i=t_j)\right\}\sum_{l=1}^{j}\frac{\sum_{i'=1}^{n}\omega(Y_{i'})\left\{w_{i'l}+I(X_{i'}=t_l)\delta_{i'}\right\}}{\sum_{i'=1}^{n}\sum_{h=l}^{K}\omega(Y_{i'})\left\{w_{i'h}+I(X_{i'}=t_h)\right\}\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{i'})}\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\Big]\right\}.$$

Hence, the updated estimate of $\boldsymbol{\psi}$ can be obtained by cycles. Specifically, given $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\lambda}^{(k+1)}$ can be obtained by equation (3.6); given $\boldsymbol{\lambda}^{(k+1)}$ and $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\alpha}^{(k+1)}$ can be calculated by equation (3.5); and given $\boldsymbol{\lambda}^{(k+1)}$ and $\boldsymbol{\alpha}^{(k+1)}$, $\boldsymbol{\beta}^{(k+1)}$ can be derived by equation (3.7).

We iterate between the E- and M-steps until the difference between the likelihoods and estimates at two successive iterations is less than a prespecified value. The proposed EM algorithm has several desirable features. First, the conditional expectations in the E-step only involve at most 1D integration. Second, in the M-step, the high-dimensional parameters $\lambda_k, k=1,\cdots,K$ are calculated explicitly (3.6), while the low-dimensional parameters can be updated through the novel use of existing software. Specifically, to solve equation (3.5) for updating parameter $\boldsymbol{\alpha}$, we can use the existing logistic regression program by creating a new data set. We first generate a data set for the unobserved and truncated subjects in which the binary

outcomes are all set to be 1. The covariates are repeated K times with $\boldsymbol{Z}_{nK} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n, \ldots, \boldsymbol{Z}_n)$. We next combine the generated data for the truncated subjects with the observed data. By using the combined data set, $\boldsymbol{\alpha}$ can be estimated by the function glm() with the "weights" option in R,

$$\text{glm}(Y^T \sim \boldsymbol{Z}, \text{weights, family} = \text{“binomial"}),$$

where $Y^T = (E(Y_1|\mathcal{O}_1, \boldsymbol{\psi}^{(k)}), \ldots, E(Y_n|\mathcal{O}_n, \boldsymbol{\psi}^{(k)}), \mathbf{1}_{nK}), \boldsymbol{Z}^T = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{Z}_{nK})$ and the weights equals $\{1, \ldots, 1, w_{11}E(Y_1|\mathcal{O}_1, \boldsymbol{\psi}^{(k)}), w_{12}E(Y_1|\mathcal{O}_1, \boldsymbol{\psi}^{(k)}), \ldots, w_{1K}E(Y_1|\mathcal{O}_1, \boldsymbol{\psi}^{(k)}), \ldots, w_{n1}E(Y_n|\mathcal{O}_n, \boldsymbol{\psi}^{(k)}), w_{n2}E(Y_n|\mathcal{O}_n, \boldsymbol{\psi}^{(k)}), \ldots, w_{nK}E(Y_n|\mathcal{O}_n, \boldsymbol{\psi}^{(k)})\}^T$.

Similarly, equation (3.7) can be solved by the existing program for right-censored data under the Cox model. First, we generate a data set for the unobserved and truncated subjects in which the failure times are constructed by repeating the observed unique survival time n times, i.e., $T_{nK} = (t_1, \ldots, t_K, \ldots, t_1, \ldots, t_K)$. The corresponding death indicator is a vector of 1, denoted as $\Delta_{nK} = (1, \ldots, 1)$. The covariates are matched with the failure times, with $\boldsymbol{Z}_{nK} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n, \ldots, \boldsymbol{Z}_n)$. After combining the generated data with the observed data, we can estimate $\boldsymbol{\beta}$ by the function coxph() with the "weight" option in R,

$$\text{coxph}(\text{Surv}(T, \Delta) \sim \boldsymbol{Z}, \text{weight}),$$

where $T^T = (x_1, \ldots, x_n, T_{nK}), \Delta^T = (\delta_1, \ldots, \delta_n, \Delta_{nK}), \boldsymbol{Z}^T = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{Z}_{nK})$ and the weight equals $\{(I(Y_1 \neq 0), \ldots, I(Y_n \neq 0), w_{11}, w_{12}, \ldots, w_{1K}, \ldots, w_{n1}, w_{n2}, \ldots, w_{nK}\}^T$. Note that the first n elements have a weight of $\{(I(Y_1 \neq 0), \ldots, I(Y_n \neq 0)\}$ since we need to exclude the cured population and have a weight of 1 for the others.

3.2. *Extension to general left-truncated data*

The stationarity assumption is required for applying the model and methods described in Section 3.1; however, that assumption can be easily violated in application. For example, in the event of an infectious disease outbreak, the number of people infected usually grows exponentially rather than linearly over time. Hence, the truncation times are unlikely to be uniformly distributed. In this section, we consider a flexible class of semiparametric models and the associated full maximum likelihood estimation for general left-truncated data. For the purpose of model identifiability, we assume a parametric model for the distribution of the truncation variable, with cumulative density function $H(\cdot|\boldsymbol{\theta})$ and density function $h(\cdot|\boldsymbol{\theta})$. The joint model of the truncation time and the time to the event of interest is not identifiable if both distributions have nonparametric components (Wang, 1989). Here, we choose the semiparametric model for the time to the event of interest (e.g., time to SAB event) and the parametric model for the truncation time. Under these assumptions, the full likelihood function of the observed data is proportional to

$$L_n(\boldsymbol{\xi}) = \prod_{i=1}^{n} \left\{ \frac{P(\widetilde{Y}_i = 1|\boldsymbol{Z}_i) f(X_i|\boldsymbol{Z}_i) h(A_i|\boldsymbol{\theta})}{P(\widetilde{T}_i > \widetilde{A}_i|\boldsymbol{Z}_i)} \right\}^{Y_i\delta_i} P(\widetilde{Y}_i = 0|\boldsymbol{Z}_i)^{\delta_i(1-Y_i)} \left\{ \frac{P(\widetilde{Y}_i = 1|\boldsymbol{Z}_i) S(X_i|\boldsymbol{Z}_i) h(A_i|\boldsymbol{\theta})}{P(\widetilde{T}_i > \widetilde{A}_i|\boldsymbol{Z}_i)} + P(\widetilde{Y}_i = 0|\boldsymbol{Z}_i) \right\}^{(1-\delta_i)}, \tag{3.8}$$

where $\boldsymbol{\xi} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}\}$.

As mentioned previously, directly maximizing the likelihood function is computationally prohibitive due to the lack of an analytical expression for the optimal value of parameter $\boldsymbol{\lambda}$. In the following equations,

we extend the EM algorithm introduced in Section 3.1 to maximize (3.8). Similarly, we treat the SAB status for censored subjects and those truncated due to the sampling mechanism as missing data and denote $\{(X_i, A_i, Y_i, \delta_i, \boldsymbol{Z}_i), \mathcal{O}_i^*, i = 1, \ldots, n\}$ as the "complete data." Accordingly, the log-likelihood of the complete data is

$$
\begin{aligned}
l_c(\boldsymbol{\xi}) = \sum_{i=1}^{n} Y_i \Big[& \log h(A_i|\boldsymbol{\theta}) + \sum_{l=1}^{m_i} \log h(A_{il}^*|\boldsymbol{\theta}) \\
& + \sum_{j=1}^{K} \sum_{l=1}^{m_i} I(T_{il}^* = t_j) \left\{ \boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + \log f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \right\} \\
& + \sum_{j=1}^{K} I(X_i = t_j) \left\{ \boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + \delta_i \log f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) + (1 - \delta_i) S(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \right\} \Big] \\
& - \sum_{i=1}^{n} (1 - Y_i) \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\}.
\end{aligned} \tag{3.9}
$$

We denote the parameter estimator from the *kth* iteration by $\boldsymbol{\xi}^{(k)} = \left\{\boldsymbol{\theta}^{(k)}, \boldsymbol{\psi}^{(k)}\right\} = \left\{\boldsymbol{\theta}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right\}$. Then given $\boldsymbol{\xi}^{(k)}$ and the observed data, we have $E(m_i|\mathcal{O}_i, \boldsymbol{\xi}^{(k)}) = \{1 - \pi(\boldsymbol{Z}_i)\}/\pi(\boldsymbol{Z}_i)$ and

$$
\pi(\boldsymbol{Z}_i) = P(\widetilde{T}_i > \widetilde{A}_i|\boldsymbol{Z}_i, \boldsymbol{\xi}^{(k)}) = \int_0^{\tau} f(u|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}) H(u|\boldsymbol{\theta}^{(k)}) \mathrm{d}u.
$$

The expected number of truncated latent subjects who would have the event at time t_j is

$$
\begin{aligned}
w_{ij} &= E\Big[\sum_{l=1}^{m_i} I(T_{il}^* = t_j)|\mathcal{O}_i, \boldsymbol{\xi}^{(k)} \Big] = E(m_i|\mathcal{O}_i, \boldsymbol{\xi}^{(k)}) E\Big[I(T_{il}^* = t_j)|\mathcal{O}_i, \boldsymbol{\xi}^{(k)} \Big] \\
&= \frac{f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)})\{1 - H(t_j|\boldsymbol{\theta}^{(k)})\}}{\int_0^{\tau} f(u|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}) H(u|\boldsymbol{\theta}^{(k)}) \mathrm{d}u}.
\end{aligned}
$$

The expectation of $\log h(A_{il}^*|\boldsymbol{\theta})$ given the observed data $\mathcal{O}_i$ under the constraint $T_{il}^* < A_{il}^*$ is

$$
E\left[\log h(A_{il}^*|\boldsymbol{\theta})|\mathcal{O}_i, \boldsymbol{\xi}^{(k)}\right] = \frac{\int F(u|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}) h(u|\boldsymbol{\theta}^{(k)}) \log h(u|\boldsymbol{\theta}) \mathrm{d}u}{\int F(u|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}) h(u|\boldsymbol{\theta}^{(k)}) du},
$$

where $F(u|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}) = \int_0^u f(v|\boldsymbol{Z}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}) \mathrm{d}v$. It follows that the expected log-likelihood function for the complete data conditional on the observed data and $\boldsymbol{\xi}^{(k)}$ is

$$
\begin{aligned}
l_E(\boldsymbol{\xi}|\boldsymbol{\xi}^{(k)}) = \sum_{i=1}^{n} \omega(Y_i) \Big[& \log h(A_i|\boldsymbol{\theta}) + E(m_i|\mathcal{O}_i, \boldsymbol{\xi}^{(k)}) E\left\{\log h(A_{il}^*|\boldsymbol{\theta}), \mathcal{O}_i, \boldsymbol{\xi}^{(k)}\right\} \\
& + \sum_{j=1}^{K} w_{ij} \Big\{ \boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + \log f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \Big\} \\
& + \sum_{j=1}^{K} I(X_i = t_j) \Big\{ \boldsymbol{\alpha}'\boldsymbol{Z}_i - \log\left\{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\} + \delta_i \log f(t_j|\boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})
\end{aligned} \tag{3.10}
$$

$$+ (1-\delta_i)\log S(t_j|\boldsymbol{Z}_i,\boldsymbol{\beta},\boldsymbol{\lambda})\Big\}\Big]$$
$$-\sum_{i=1}^{n}\{1-\omega(Y_i)\}\log\left\{1+\exp(\boldsymbol{\alpha}'\boldsymbol{Z}_i)\right\},$$

where $\omega(Y_i) = \delta_i Y_i + (1-\delta_i)E(Y_i|\mathcal{O}_i,\boldsymbol{\xi}^{(k)})$. The M-step maximizes (3.10) to update the parameter estimates. Specifically, the updates can be obtained through cycles of $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The parameters $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ can be estimated in a manner similar to that described in Section 3.1. The estimate of $\boldsymbol{\theta}$ can be derived by solving the following score equation

$$\sum_{i=1}^{n}\omega(Y_i)\left\{\frac{\dot{h}(A_i|\boldsymbol{\theta})}{h(A_i|\boldsymbol{\theta})}+E(m_i|\mathcal{O}_i)\frac{\int F(u|\boldsymbol{Z}_i,\boldsymbol{\theta}^{(k)},\boldsymbol{\lambda}^{(k)})h(u|\boldsymbol{\theta}^{(k)})\frac{\dot{h}(u|\boldsymbol{\theta})}{h(u|\boldsymbol{\theta})}\mathrm{d}u}{\int F(u|\boldsymbol{Z}_i,\boldsymbol{\theta}^{(k)},\boldsymbol{\lambda}^{(k)})h(u|\boldsymbol{\theta}^{(k)})du}\right\}, \tag{3.11}$$

where $\dot{h}(u|\boldsymbol{\theta}) = \partial h(u|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. For the implementation of the M-step, we can easily use the existing program for logistic regression and traditional right-censored data under the Cox model as described in Section 3.1.

4. ASYMPTOTIC PROPERTIES

We establish the asymptotic properties of the estimators, denoted as $\widehat{\boldsymbol{\xi}}_n = (\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Lambda}}_n)$, under general left-truncation sampling. Here, the subscript n indicates the sample size. The true values of the parameters are denoted as $\boldsymbol{\xi}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$. By the counting process formulation, the observed log-likelihood function can be rewritten as

$$l_n(\boldsymbol{\xi}) = \sum_{i=1}^{n}\left\{Y_i\int_0^{\tau}\Big[\boldsymbol{\alpha}^T\boldsymbol{Z}_{i1}+\log d\Lambda(u)+\boldsymbol{\beta}^T\boldsymbol{Z}_i-\int_0^{u}\exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)\mathrm{d}\Lambda(v)+\log h(A_i|\boldsymbol{\theta})$$
$$-\log\int_0^{\tau}S(v|\boldsymbol{Z}_i)h(v|\boldsymbol{\theta})\mathrm{d}v\Big]\mathrm{d}N_i(u)$$
ãŁŁ $$-\int_0^{\tau}\log\left[1+\frac{\exp\left\{\boldsymbol{\alpha}^T\boldsymbol{Z}_{i1}-\int_0^{u}\exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)d\Lambda(v)\right\}}{\int_0^{\tau}S(v|\boldsymbol{Z}_i)h(v|\boldsymbol{\theta})dv}h(A_i|\theta)\right]\mathrm{d}N_i(u)$$
$$+\log\left[\frac{1}{1+\exp(\boldsymbol{\alpha}^T\boldsymbol{Z}_{i1})}+\frac{\exp\left\{\boldsymbol{\alpha}^T\boldsymbol{Z}_{i1}-\int_0^{\tau}M_i(v)\exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)d\Lambda(v)\right\}}{\{1+\exp(\boldsymbol{\alpha}^T\boldsymbol{Z}_{i1})\}\int_0^{\tau}S(v|\boldsymbol{Z}_i)h(v|\boldsymbol{\theta})dv}h(A_i|\boldsymbol{\theta})\right]\Big\},$$

where $N_i(u) = I(A_i < X_i \le u)\delta_i$, $M_i(u) = I(X_i \ge u)I(X_i > A_i)$ and τ is the upper bound for the support of $\widetilde{T}$. Under the regularity conditions provided in the supplementary materials available at *Biostatistics* online, we establish strong consistency by the classical Kullback–Leibler information approach, and prove the weak convergence of the estimators by the Z-theorem for infinite-dimensional estimating equations (Van Der Vaart and Wellner, 1996).

Theorem 1: Under the regularity conditions listed in the supplementary materials available at Biostatistics online, the estimators $\widehat{\boldsymbol{\xi}}_n$ are consistent: $(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$ converge almost surely to $(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, $\sup_{0<t\le\tau}|\widehat{\Lambda}_n(t)-\Lambda_0(t)|$ converges almost surely to 0 as $n\to\infty$.

As $\widehat{\boldsymbol{\xi}}_n$ maximizes the likelihood function, the empirical Kullback–Leibler information $l_n(\widehat{\boldsymbol{\xi}}_n) - l_n(\boldsymbol{\xi}_0) \ge 0$ must always be negative. If $\widehat{\boldsymbol{\xi}}_n$ converges, say, to $\boldsymbol{\xi}^*$, then following the uniform law of large numbers, we can show that $l_n(\widehat{\boldsymbol{\xi}}_n) - l_n(\boldsymbol{\xi}_0)$ must converge to the negative Kullback–Leibler distance between $P_{\boldsymbol{\xi}^*}$

and $P_{\boldsymbol{\xi}_0}$, where $P_{\boldsymbol{\xi}}$ is the probability measure under the parameter $\boldsymbol{\xi}$. As the Kullback–Leibler information is always non-negative, it implies that $P_{\boldsymbol{\xi}^*} = P_{\boldsymbol{\xi}_0}$ almost surely. Under the regularity conditions provided in the supplementary materials available at *Biostatistics* online, model $P_{\boldsymbol{\xi}}$ is identifiable, implying that $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0$.

Theorem 2: Under the regularity conditions listed in the supplementary materials available at *Biostatistics* online, $\sqrt{n}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0)$ converges weakly to a tight, mean zero Gaussian process $-\dot{U}_0^{-1}(\mathcal{W})$, where U_0 is defined as the expectation of score function U_n under true parameter values ξ_0.

By the von Mises method for semiparametric maximum likelihood estimators (Gill *and others*, 1989), the score functions are derived by taking the derivative of $l_n(\boldsymbol{\xi})$ with respect to $\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}$, and a submodel $d\Lambda_\eta(\cdot) = \{1 + \eta\phi(\cdot)\}\, d\Lambda(\cdot)$. Here, $\phi(\cdot)$ is a bounded and integrable function, and η is a positive constant. We denote the infinite-dimensional score functions by $U_n(\boldsymbol{\psi}) \equiv \{U_{1n}(\boldsymbol{\xi}), U_{2n}(\boldsymbol{\xi}), U_{3n}(\boldsymbol{\xi}), U_{4n}(t, \boldsymbol{\xi})\}$, and its expectation under true values $\xi_0 = (\theta_0, \alpha_0, \beta_0, \Lambda_0)$ by

$$\begin{aligned} U_0(\cdot, \boldsymbol{\xi}) &\equiv \{U_{10}(\boldsymbol{\xi}), U_{20}(\boldsymbol{\xi}), U_{30}(\boldsymbol{\xi}), U_{40}(\cdot, \boldsymbol{\xi})\} \\ &= \Big\{E_0\{U_{1n}(\boldsymbol{\xi})\}, E_0\{U_{2n}(\boldsymbol{\xi})\}, E_0\{U_{3n}(\boldsymbol{\xi})\}, E_0\{U_{4n}(\cdot, \boldsymbol{\xi})\}\Big\}. \end{aligned}$$

Both the score function U_n and its expectation U_0 are defined on the parameter set $\mathcal{A} \times \mathcal{B}$, where set $\mathcal{A}$ is assumed to be compact in $\mathcal{R}^{q+2p+1}$, and the set $\mathcal{B}$ consists of nondecreasing functions in the space of functions with bounded variation. The true value $\boldsymbol{\xi}_0$ satisfies the population score function $U_0(t, \xi_0) = 0$. The estimating functions evaluated at the true value $\boldsymbol{\xi}_0$ can be written as an empirical process $\sqrt{n}U_n(\boldsymbol{\xi}_0) = \sqrt{n}\{U_n(t, \boldsymbol{\xi}_0) - U_0(t, \boldsymbol{\xi}_0)\}$. By the uniform central limit theorem, it can be shown that $\sqrt{n}U_n(\boldsymbol{\xi}_0)$ converges weakly to $\mathcal{W}$. $\mathcal{W}$ is defined as $\mathcal{W} \equiv (\mathcal{W}_1, \mathcal{W}_2)$, where $\mathcal{W}_1$ is a Gaussian random vector with covariance matrix $\Sigma_{11} = E_0\left\{U_{123n}(\xi_0)^{\otimes 2}\right\}$ and $U_{123n}(\xi_0) = \{U_{1n}(\xi_0), U_{2n}(\xi_0), U_{3n}(\xi_0)\}$, and $\mathcal{W}_2$ is a tight Gaussian process with covariance matrix $\Sigma_{22}(s, t) = E_0\{U_{4n}(s, \xi_0)U_{4n}(t, \xi_0)\}$. Denote the Fréchet derivative of $U_0(\boldsymbol{\xi})$ evaluated at $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ by $\dot{U}_0$. In the supplementary materials available at *Biostatistics* online, we outline the proof for the three main conditions for using the Z-theorem: Fréchet differentiability and invertibility, weak convergence of $\sqrt{n}U_n(\boldsymbol{\xi}_0)$ and a stochastic approximation condition of the estimating equations. Note that we show the proof under the general left-truncated sampling, which includes length-biased sampling as a special case.

4.1. *Variance estimation*

We use an EM-aided computational differentiation approach with the profile likelihood to estimate the variances of the finite dimensional estimators $\widehat{\boldsymbol{\xi}}_n$ (Chen and Little, 1999; Murphy and Van Der Vaart, 2000). By the perturbation around the obtained estimators, the information matrix can be estimated as shown below:

(1) Perturb the lth entry of $\widehat{\boldsymbol{\eta}} = (\widehat{\theta}, \widehat{\alpha}_0, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ by a small value $\epsilon = 1/n$ in the neighborhood in one direction or both directions, denoted as $\widehat{\boldsymbol{\eta}}^+ = \widehat{\boldsymbol{\eta}} + (0, \ldots, \epsilon, \ldots, 0)$ and $\widehat{\boldsymbol{\eta}}^- = \widehat{\boldsymbol{\eta}} - (0, \ldots, \epsilon, \ldots, 0)$, respectively.

(2) Use the EM algorithm to obtain $\widehat{\boldsymbol{\lambda}}_{\widehat{\boldsymbol{\eta}}^+}$ and $\widehat{\boldsymbol{\lambda}}_{\widehat{\boldsymbol{\eta}}^-}$ given $\widehat{\boldsymbol{\eta}}^+$ and $\widehat{\boldsymbol{\eta}}^-$, respectively.

(3) Approximate the lth row of the information matrix of $\widehat{\boldsymbol{\eta}}$ by

$$\frac{1}{2\epsilon}\left\{\frac{\partial l_E(\boldsymbol{\eta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}^-, \boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}_{\widehat{\boldsymbol{\eta}}^-}} - \frac{\partial l_E(\boldsymbol{\eta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\eta}}\Big|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}^+, \boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}_{\widehat{\boldsymbol{\eta}}^+}}\right\}.$$

5. SIMULATION STUDY

We conducted simulations studies to evaluate the finite sample performance of the proposed method. The SAB status indicator $\widetilde{Y}$ was generated from a logistic regression model with two covariates (Z_1, Z_2), where Z_1 followed a Bernoulli distribution with probability 0.5, and Z_2 followed a uniform distribution, $uniform(-0.5, 0.5)$. We set $\boldsymbol{\alpha} = (1.2, 1, 1)$, such that the uncured proportion was around 75%. For the uncured subjects (i.e., $\widetilde{Y} = 1$), we generated unbiased survival times $\widetilde{T}$ from a Cox proportional hazards model with covariates (Z_1, Z_2) and $\boldsymbol{\beta} = (-0.5, 1)$. For model identifiability, the baseline hazard function was chosen such that all events occurred before $\tau = 20$. Specifically, we used $\Lambda_0^{-1}(t) = 20\{1 - \exp(-t)\}$. The truncation time $\widetilde{A}_i$ was generated from a uniform distribution $U(0, \tau)$ for the length-biased data and from a truncated Weibull distribution with the density function $h(t|\boldsymbol{\theta})$ for the general left-truncated data, where $h(t|\boldsymbol{\theta}) = g(t|\boldsymbol{\theta})/G(20|\boldsymbol{\theta})$ with

$$g(t|\boldsymbol{\theta}) = \frac{\theta_1}{\theta_2}\left(\frac{t}{\theta_2}\right)^{\theta_1 - 1} \exp\left\{-\left(\frac{t}{\theta_2}\right)^{\theta_1}\right\},$$

$G(t|\boldsymbol{\theta})$ is the cumulative density function and $\boldsymbol{\theta} = (\theta_1, \theta_2) = (1.0, 2.8)$. Following the sampling mechanism, we only kept subjects with $\widetilde{T} > \widetilde{A}$ in the observed data sets. The residue censoring time was generated from a uniform distribution with varying boundaries to have different censoring rates. For a subject with censored observations, the value of the SAB status Y was set to be missing. We set $n = 300$, or 600 and used 1000 replicates for each sample size.

We first assessed the validity of our proposed estimation and inference procedures in finite samples. Tables 1 and 2 summarize the average estimates, empirical standard errors and average EM-aided standard errors for the length-biased data and general truncated data, respectively. As shown in the tables, all point estimates had negligible bias for both length-biased data and general left-truncated data. The standard errors estimated by the EM-aided procedure approximated the empirical standard errors well. Generally, the empirical biases did not change much when varying the percentage of censoring, while the standard errors consistently increased with an increasing percentage of censoring. Also, as the sample size increased from 300 to 600, the standard errors of all estimates decreased.

For comparison, we also performed a naive analysis by ignoring the unique data structure. Specifically, we first fitted a logistic regression model by excluding subjects with unknown values of Y_i, and then performed Cox proportional hazards modeling for left-truncated data by using subjects with $Y_i = 1$. As shown in the right-sided columns of Tables 1 and 2, this naive method resulted in biased estimates for all parameters in both the logistic regression model and Cox proportional hazards model, since the missing mechanism was not random in our setting.

6. DATA APPLICATION

To evaluate the entire effects of treatments for autoimmune disease on the risk of experiencing SAB and time to SAB among pregnant women, we analyzed the data from the OTIS autoimmune disease in pregnancy database that we introduced in Section 1. The data set included a total of 930 pregnant women with complete records who entered the studies before week 20 of their gestation during the years between 2005 and 2012. Among these pregnant women, 483 (51.9%) had autoimmune diseases and were treated with the medications under investigation (this group comprises the exposure group); 264 (28.4%) also had autoimmune diseases but were not treated with the medications under investigation (this group comprises the diseased control group); 183 (19.70%) were healthy pregnant women without any autoimmune diseases who were also not treated with the medications under study (this group comprises the healthy control group). There were 66 SAB events and 2 censoring events observed during the study.

Table 1. *Summary of simulation studies with length-biased data. EST*: *empirical mean*; *SD*: *empirical standard deviation*; *ESE*: *average of asymptotic standard error estimates*

				Proposed method			**Naive method**	
N	CENSOR	PARA	TRUE	EST	SD	ESE	EST	SD
300	0%	α_0	1.2	1.16	0.14	0.12	0.49	0.13
		α_1	1.0	1.05	0.27	0.23	1.27	0.26
		α_2	1.0	0.94	0.45	0.39	0.49	0.43
		β_1	-0.5	-0.50	0.14	0.14	-0.50	0.16
		β_2	1.0	1.01	0.27	0.25	1.01	0.29
	10%	α_0	1.2	1.16	0.15	0.12	0.69	0.15
		α_1	1.0	1.06	0.28	0.25	1.22	0.29
		α_2	1.0	0.94	0.47	0.41	0.59	0.48
		β_1	-0.5	-0.50	0.16	0.15	-0.43	0.16
		β_2	1.0	1.01	0.29	0.27	0.89	0.29
	30%	α_0	1.2	1.18	0.18	0.14	0.79	0.18
		α_1	1.0	1.07	0.32	0.28	1.20	0.34
		α_2	1.0	0.95	0.54	0.47	0.68	0.57
		β_1	-0.5	-0.49	0.17	0.17	-0.40	0.19
		β_2	1.0	1.01	0.32	0.30	0.83	0.34
600	0%	α_0	1.2	1.16	0.10	0.08	0.49	0.09
		α_1	1.0	1.03	0.19	0.16	1.25	0.18
		α_2	1.0	0.94	0.32	0.27	0.49	0.31
		β_1	-0.5	-0.50	0.10	0.10	-0.50	0.12
		β_2	1.0	1.00	0.17	0.17	1.00	0.19
	10%	α_0	1.2	1.17	0.11	0.09	0.69	0.10
		α_1	1.0	1.04	0.20	0.17	1.21	0.20
		α_2	1.0	0.95	0.33	0.29	0.58	0.34
		β_1	-0.5	-0.50	0.11	0.11	-0.43	0.12
		β_2	1.0	1.00	0.19	0.19	0.88	0.20
	30%	α_0	1.2	1.17	0.12	0.10	0.78	0.12
		α_1	1.0	1.04	0.22	0.20	1.18	0.23
		α_2	1.0	0.95	0.37	0.33	0.66	0.39
		β_1	-0.5	-0.49	0.12	0.12	-0.39	0.13
		β_2	1.0	1.00	0.21	0.21	0.82	0.23

Biased sampling occurred because the women who had experienced SAB early in the course of their pregnancy had been excluded from the study. As a result, the observed time to the SAB event tended to be longer than that in the general population, as illustrated in Figure 1.

We extracted the baseline covariates for the subjects, including maternal age ($\geq$35 or $<$35), smoking status, alcohol status, prior SAB status, and prior therapeutic abortion status, from the database. We performed univariate analysis to select the covariate set to use in the joint model of the risk of experiencing SAB and the time to SAB. The purpose of the univariate screening is to identify the potential confounders

Table 2. *Summary of simulation studies with general left-truncated data. EST*: *empirical mean*; *SD*: *empirical standard deviation*; *ESE*: *average of asymptotic standard error estimates*

				Proposed method			**Naive method**	
N	CENSOR	PARA	TRUE	EST	SD	ESE	EST	SD
300	0%	α_0	1.2	1.20	0.14	0.15	1.05	0.14
		α_1	1.0	1.03	0.29	0.27	1.09	0.29
		α_2	1.0	0.98	0.46	0.49	0.85	0.47
		β_1	-0.5	-0.50	0.15	0.14	-0.50	0.15
		β_2	1.0	1.01	0.25	0.25	1.00	0.26
		θ_1	1.0	1.01	0.06	0.06	1.02	0.05
		θ_2	2.8	2.78	0.24	0.24	2.46	0.15
	10%	α_0	1.2	1.22	0.17	0.18	1.88	0.21
		α_1	1.0	1.05	0.34	0.31	1.02	0.42
		α_2	1.0	0.98	0.52	0.56	1.02	0.68
		β_1	-0.5	-0.50	0.17	0.17	-0.41	0.15
		β_2	1.0	1.00	0.29	0.29	0.82	0.26
		θ_1	1.0	1.01	0.06	0.06	1.02	0.05
		θ_2	2.8	2.78	0.24	0.24	2.46	0.15
	30%	α_0	1.2	1.27	0.24	0.26	2.50	0.52
		α_1	1.0	1.10	0.49	0.46	1.02	0.99
		α_2	1.0	0.97	0.75	0.76	1.22	1.09
		β_1	-0.5	-0.50	0.21	0.21	-0.34	0.17
		β_2	1.0	1.01	0.35	0.36	0.69	0.28
		θ_1	1.0	1.00	0.06	0.06	1.02	0.05
		θ_2	2.8	2.79	0.23	0.24	2.46	0.15
600	0%	α_0	1.2	1.20	0.10	0.11	1.05	0.10
		α_1	1.0	1.01	0.21	0.18	1.07	0.20
		α_2	1.0	0.98	0.33	0.35	0.85	0.33
		β_1	-0.5	-0.50	0.10	0.10	-0.50	0.10
		β_2	1.0	1.01	0.17	0.17	1.01	0.17
		θ_1	1.0	1.00	0.04	0.04	1.01	0.03
		θ_2	2.8	2.79	0.17	0.17	2.47	0.10
	10%	α_0	1.2	1.21	0.12	0.13	1.86	0.15
		α_1	1.0	1.02	0.24	0.21	1.00	0.29
		α_2	1.0	0.97	0.37	0.41	0.99	0.49
		β_1	-0.5	-0.50	0.12	0.12	-0.40	0.10
		β_2	1.0	1.01	0.20	0.20	0.83	0.17
		θ_1	1.0	1.00	0.04	0.04	1.01	0.03
		θ_2	2.8	2.78	0.16	0.18	2.47	0.10
	30%	α_0	1.2	1.23	0.17	0.17	2.41	0.23
		α_1	1.0	1.04	0.33	0.28	0.95	0.44
		α_2	1.0	0.98	0.50	0.54	1.19	0.72
		β_1	-0.5	-0.50	0.15	0.15	-0.34	0.11
		β_2	1.0	1.01	0.25	0.25	0.70	0.20
		θ_1	1.0	1.00	0.04	0.04	1.01	0.03
		θ_2	2.8	2.79	0.16	0.18	2.47	0.10

when evaluating the risk of using autoimmune disease medications in pregnant women, which is a common practice. The univariate analysis was performed to account for the sampling bias by using the proposed method. Specifically, for each covariate, except for the treatment indicators, we have jointly modeled the cure probability and survival distribution, and used the proposed method for model fitting. Covariates with p-values smaller than 0.2 in either the logistic regression or Cox regression model were included in the final multivariate model. Our final models included maternal age (≥ 35 or <35) and treatment group indicators (exposure group, healthy control or disease control groups).

We first examined the stationarity assumption using the observed time-to-SAB data. The formal test of stationarity assumption given by Addona and Wolfson (2006) yielded a two-sided p-value of 0.0001, which indicated that the stationarity assumption did not hold and the observed time-to-SAB data were not length-biased data. We then used the Weibull distribution to model the truncation time. Both the estimated values of shape and scale parameters were very large compared with their standard errors, also suggesting the stationarity assumption did not hold in the study. Table 3 lists the estimated coefficients along with standard errors and p-values from the proposed method and the naive analyses. The model fitting of the logistic regression by the proposed method indicated that the healthy controls had significantly lower risk (p-value <0.01) of experiencing SAB compared with the other two groups after controlling for the age effect. Interestingly, our comparison of the exposure group and the disease control group suggested that the use of the medications under investigation for pregnant women with autoimmune diseases did not change their risk of experiencing SAB. Consistent with previous reports (Andersen *and others*, 2000), we found that older maternal age (≥ 35) significantly increases the risk of SAB (p-value<0.01). The Cox regression part of the joint model suggested that autoimmune disease status, use of the newer medications and maternal age did not significantly affect the distribution of time to SAB for the uncured group, although the healthy control group tended to have a lower hazard function indicating later timing of SAB events, compared with that of the other two groups after controlling for the maternal age (≥ 35 or < 35). The naive analysis that ignored the data structure had similar results for the parameters in the survival model, but had misleading results for the risk model. Specifically, the naive analysis greatly underestimated the overall risk of experiencing SAB, which is similar to our previous simulation findings when the censoring rate is low. Note that the conclusion is conditional on $\tilde{T} \geq 5$ weeks, due to a lack of instantaneous detection of pregnancy in the early stage.

7. Discussion

We have proposed new EM algorithms for biased sampling survival data with a cured proportion to obtain full likelihood maximum estimators. We first considered length-biased data and then generalized the estimation and inference procedure to general left-truncated data. As pointed out by Wang (1989), the joint model is not identifiable if distributions of the truncation time and event time of interest both have nonparametric components. Here, we choose the semiparametric model for the event time of interest (e.g., time to SAB event) and adopt a parametric model for the truncation time. One way to relax the parametric assumptions is to use a flexible parametric model, such as a truncated generalized Gamma distribution with three parameters (Stacy, 1962) for the truncation time. Specifically, the density function of a truncated generalized Gamma distribution is $h(t|\boldsymbol{\theta}) = g(t|\boldsymbol{\theta})/G(20|\boldsymbol{\theta})$ with $g(t|\boldsymbol{\theta}) = \theta_1 t^{\theta_1\theta_3-1}\mathrm{e}^{-(t/\theta_2)^{\theta_1}}/\{\Gamma(\theta_3)\theta_2^{\theta_1\theta_3}\}$, and $G(t|\boldsymbol{\theta})$ is the cumulative density function, where $(\theta_1, \theta_2, \theta_3) > 0$. The generalized Gamma distribution degenerates to the Weibull distribution if $\theta_3 = 1$, and degenerates to the Gamma distribution if $\theta_1 = 1$. In the supplementary materials available at *Biostatistics* online, we have conducted additional simulation studies to investigate the effects of model misspecification of the truncation time on the estimation of the parameters of interest, i.e., regression coefficients under the logistic regression model and Cox model. In summary, the estimators of interest have robust performance with violations of the parametric model assumptions on the truncation time. As discussed in Section 2, the zero tail constraint for survival data with

Table 3. *Estimated coefficients with standard errors* (*SE*) *and p-values for SAB data*

		Proposed method			Naive method		
		Coefficient	SE	P-value	Coefficient	SE	P-value
Logistic model							
Treatment							
	Exposed						
	Healthy control	−1.04	0.34	< 0.01	−0.91	0.45	0.04
	Disease control	0.05	0.21	0.82	0.03	0.28	0.93
Age							
	<35						
	≥35	0.67	0.19	< 0.01	0.55	0.26	0.04
Intercept		−2.01	0.14	< 0.01	−2.64	0.20	< 0.01
Cox proportional hazard model							
Treatment							
	Exposed						
	Healthy control	−0.31	0.43	0.47	−0.33	0.49	0.51
	Disease control	0.06	0.23	0.81	0.26	0.28	0.36
Age							
	<35						
	≥35	0.27	0.22	0.20	0.10	0.26	0.68
Shape parameter		3.13	0.30		2.77	0.07	
Scale parameter		10.59	0.60		12.66	0.16	

a cure portion is naturally satisfied. Different from the usual cure rate data where the long-term survivors are always right-censored, in our pregnancy studies we observe majority of the "cured" women. This greatly improves the practical identifiability of the cured portion (Farewell, 1986; Lu and Ying, 2004), as well as substantially increase the amount of information available for estimating the model parameters.

Even though the proposed point and variance estimation involves iterations, the computation is fast and efficient. The conditional expectations in the E-step of both the point and variance estimations involve at most one-dimensional integration and can be easily estimated. In the M-step, the non-specified baseline hazard function can be calculated explicitly, while the low-dimensional parameters can be updated quickly using available statistical software. For example, in a 100-run simulation for the general left-truncated data using a 3.30GHz desktop CPU under the scenario with 600 samples and 10% censoring rate, the CPU time was 3.16 hours and 0.34 hours for the point estimation and variance estimation, respectively. The average number of iterations to achieve convergence was 14, with convergence criterion defined as $\max |\eta^{(k+1)} - \eta^{(k)}| < 10^{-3}$. For the SAB data, the CPU time for fitting the final model was 0.18 hours, including the point and variance estimation.

Although this work focused on the logistic regression model for the cured proportion and the proportional hazards model for the time to the event of interest, the proposed estimation and inference method can be extended to other types of models such as the probit model for the cured proportion and the proportional odds model for the event time. In applications, one challenge when applying the proposed method is model checking. Due to the biased sampling issue, the distribution of the observed data is not representative of that of the target population. Accordingly, standard diagnostic tools, such as model checking tests of proportionality for traditional survival data, cannot be directly applied here. Developing rigorous statistical tools for model checking is beyond the scope of this article, and is a worthy objective for future research.

8. Software

Software in the form of R code and documentation is online at https://github.com/JPiao7u089/Cured-Proportion-and-Biased-Sampling.git.

Supplementary material

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

Acknowledgments

We gratefully acknowledge support from the National Cancer Institute at the National Institutes of Health (CA016672 and CA193878). *Conflict of Interest*: None declared.

References

Addona, V. and Wolfson, D. B. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime data analysis* **12**, 267–284.

Andersen, A. M. N., Wohlfahrt, J., Christens, P., Olsen, J. and Melbye, M. (2000). Maternal age and fetal loss: population based register linkage study. *BMJ* **320**, 1708–1712.

Chambers, C. D., Johnson, D. L., Xu, R., Taylor, S., Rosillon, D., Wolleswinkel, J. H. and Baril, L. (2011). Challenges and design of a prospective, observational cohort study to assess the risk of spontaneous abortion following administration of human papillomavirus (hpv) bivalent (types 16 and 18) recombinant vaccine. In: Brian L. Strom (editor), *Pharmacoepidemiology and Drug Safety*, Volume 20. Malden, MA USA: Wiley. pp. S358–S358.

Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **94**, 896–908.

Chen, M. H., Ibrahim, J. G. and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**, 909–919.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics* **14**, 257–262.

Gill, R. D, Wellner, J. A and Præstgaard, J. (1989). Non-and semi-parametric maximum likelihood estimators and the von Mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics* **16**, 97–128.

Kim, J. P., Lu, W., Sit, T. and Ying, Z. (2013). A unified approach to semiparametric transformation models under general biased sampling schemes. *Journal of the American Statistical Association* **108**, 217–227.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531–541.

Li, C. S. and Taylor, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine* **21**, 3235–3247.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika* **91**, 331–343.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, 449–465.

Ning, J., Qin, J. and Shen, Y. (2014). Score estimating equations from embedded likelihood functions under accelerated failure time model. *Journal of the American Statistical Association* **109**, 1625–1635.

Peng, Y. and Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56**, 237–243.

QIN, J., NING, J., LIU, H. AND SHEN, Y. (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association* **106**, 1434–1449.

SHEN, Y., NING, J. AND QIN, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association* **104**, 1192–1202.

SKORPEN, C. G., HOELTZENBEIN, M., TINCANI, A., FISCHER-BETZ, R., ELEFANT, E., CHAMBERS, C., DA SILVA, J., NELSON-PIERCY, C., CETIN, I., COSTEDOAT-CHALUMEAU, N. *and others*. (2016). The eular points to consider for use of antirheumatic drugs before pregnancy, and during pregnancy and lactation. *Annals of the Rheumatic Diseases* **75**, 795–810.

STACY, E. W. (1962). A generalization of the gamma distribution. *The Annals of mathematical statistics* **33**, 1187–1192.

SY, J. P. AND TAYLOR, J. M. G. (2000). Estimation in a cox proportional hazards cure model. *Biometrics* **56**, 227–236.

TAYLOR, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* **51**, 899–907.

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

VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

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

VISSER, K., KATCHAMART, W., LOZA, E., MARTINEZ-LOPEZ, J.A., SALLIOT, C., TRUDEAU, J., BOMBARDIER, C., CARMONA, L., VAN DER HEIJDE, D., BIJLSMA, J.W.J. *and others*. (2009). Multinational evidence-based recommendations for the use of methotrexate in rheumatic disorders with a focus on rheumatoid arthritis: integrating systematic literature research and expert opinion of a broad international panel of rheumatologists in the 3e initiative. *Annals of the Rheumatic Diseases* **68**, 1086–1093.

WANG, M. C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* **84**, 742–748.

WANG, M. C., JEWELL, N. P. AND TSAI, W. Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics* **14**, 1597–1605.

WILCOX, A. J., WEINBERG, C. R., O'CONNOR, J. F., BAIRD, D. D., SCHLATTERER, J. P., CANFIELD, R. E., ARMSTRONG, E. G. AND NISULA, B. C. (1988). Incidence of early loss of pregnancy. *New England Journal of Medicine* **319**, 189–194.

XU, R. AND CHAMBERS, C. (2011). A sample size calculation for spontaneous abortion in observational studies. *Reproductive Toxicology* **32**, 490–493.

ZENG, D., YIN, G. AND IBRAHIM, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association* **101**, 670–684.

ZHANG, J. AND PENG, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis* **15**, 455–467.

[*Received August 26, 2016; revised March 21, 2017; accepted for publication April 12, 2017*]