The Analysis of Biased Spontaneous Abortion Data from Observational Studies

Walter Faig* Ronghui Xu† Christina Chambers‡

Abstract

We investigate the analysis of spontaneous abortion (SAB) data collected via observational studies in pregnancy. Such data can be left truncated because women may enter a study any time during their pregnancy. In addition, the data has a well-defined 'cured' portion, because the majority of the pregnancies do not end in spontaneous abortion. The data are also subject to right-censoring due to loss to follow-up etc.

While left truncation has been well studied for the usual time-to-event data with or without right-censoring, it causes unique bias in the presence of a cured portion. In light of the scientific questions of interest, i.e. to identify risk factors for SAB event (yes or no), as well as to identify predictors of SAB timing among those who experience it, we consider the mixture type cure rate models for their desirable interpretations. Because the exact likelihood is difficult to maximize, we propose a weighted and an approximate complete data likelihood, which are optimized using an EM-type (ES) algorithm. Both approaches involve estimation of the left truncation distribution, which can be achieved using the cured subjects in order to simplify inference. Inference is then carried out using the semiparametric sandwich variance estimators which have closed form expressions. The approaches are examined through simulation studies, and applied to the pregnancy data from the Organization of Teratology Information Specialists (OTIS) autoimmune disease database to illustrate its ability to simultaneously answer the two scientific questions of interest which cannot otherwise be achieved with existing methodologies.

Key Words: approximate complete data likelihood; ES algorithm; inverse probability weighting; mixture cure rate models; sandwich variance estimator; weighted complete data likelihood

1. Introduction

Our work was motivated by research work carried out at 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. Research subjects are enrolled from the Teratology Information Services and through other methods of recruitment, where the mothers and their babies are followed over time. Phone interviews are conducted through the length of the pregnancy along with pregnancy diaries recorded by the mother. An outcome phone interview is conducted shortly after the pregnancy ends, and if it results in a live birth a dysmorphology exam is done within six months and with further follow-ups at one year and possibly later dates. Recently it has been of interest to assess the effects of medication exposures on spontaneous abortion (SAB) (Xu and Chambers, 2011; Chambers *et al.*, 2011). Here we examine a collection of studies on the risks and safety of autoimmune disease medications relative to adverse pregnancy outcomes, and we focus on spontaneous abortion as the outcome of interest.

By definition SAB occurs within the first 20 weeks of gestation; any pregnancy loss after that is called still birth. Ultimately we would like to know if an exposure modifies the risk of SAB for a woman, which may be increased or decreased. It is known that in the population for clinically recognized pregnancies the rate of SAB is about 12% (Wilcox

*University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093
†University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093
‡University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093

et al., 1988). On the other hand, in our database the empirical SAB rate is consistently lower than 10%. This is due to the fact that women may enter a study any time before 20 weeks' gestation. This way women who have early SAB events are less likely to be capture in our studies, and such selection bias is known as left truncation in survival analysis. Left truncation has been studied by many authors since the 1980s, and have attracted much recent attention in the context of length-biased data (Asgharian *et al.*, 2002; Qin *et al.*, 2011, among others).

The fact that the majority of the pregnant women are free of SAB is considered 'cured' in the time-to-event context. Cure rate models are well studied in the literature for right-censored data. They are used in various biomedical studies where data often include a substantial portion of 'long-term' survivors who are no longer susceptible to the event of interest (Farewell, 1982, 1986). The models effectively analyze the survival distribution of those who are susceptible along with the probability of an individual being 'cured'. In the approaches using mixture models, the logistic regression is often used to model the cured probability. For the dependency of the survival function on the covariates among the non-cured, various regression models have been considered: the Cox proportional hazards model (Kuk and Chen, 1992; Sy and Taylor, 2000), transformation models (Lu and Ying, 2004), and richly parametrized models when the shape of the hazard function is of interest (Hanson *et al.*, 2003). Cure rate models have also been developed along the lines of non-mixture models (Chen *et al.*, 1999; Zeng *et al.*, 2006).

In addition to right-censored data, cure-rate models have also been developed for interval-censored data (Kim and Jhun, 2008). To our best knowledge, however, they have not been considered for truncated data which, unlike censoring, poses a unique set of challenges. While left truncation has been well studied as mentioned above, the challenges are again unique in the presence of a cured portion. Most importantly, left-truncation leads to selection bias that needs to be explicitly counted for, and in the process of doing so computational challenges also arise, as will be seen below.

2. Model and Estimation

Let Y_i be the indicator of whether subject i experiences the event of interest, in our case SAB, $i = 1, ..., n$. It is possible that Y_i is unobserved if there is right-censoring; otherwise we observe $Y_i = 1$ if a woman has an SAB event, or 0 if a woman carries her pregnancy past 20 weeks of gestation. Note that this is different from the classic cure data in the literature, where $Y_i = 0$ is never observed. Let τ be a time (such as 20 weeks) after which an individual is no longer considered susceptible to the event; recall from the previous section that SAB is defined as loss of pregnancy prior to 20 weeks of gestation. Let T_i denote the event time if subject i experiences an SAB event. Let Q_i be the left-truncation (i.e. study entry) time and C_i be the potential right-censoring time. We assume non-informative truncation and censoring times; that is, (Q_i, C_i) with $Q_i < C_i$, and both are independent of T_i conditional on the covariates. This assumption was also used in Tsai *et al.* (1987) and Wang (1991), for example, and is consider viable in the context of our pregnancy studies. We also expand the above notation to $C_i \geq T_i \geq \tau$ if $Y_i = 0$ is observed. We define $X_i = \min(T_i, C_i, \tau)$, and $\delta_i = I(T_i \leq C_i)$; note that if $Y_i = 0$ is observed, we also have $\delta_i = 1$.

We consider the mixture cure rate model, which provides nice interpretation as explained before for our purposes of analyzing women who have and who do not have SAB events. The marginal survival function over the mixture of the two populations is given by

$$\bar{S}_i(t) = P(T_i > t) = (1 - p_i) + p_i S_i(t), \tag{1}$$

where $p_i = P(Y_i = 1)$, and $S_i(t) = P(T_i > t|Y_i = 1)$ for $t < \tau$. To model p_i and $S_i(t)$ we consider the logistic regression and the Cox proportional hazards regression, respectively. These are common regression model choices, and were used in Sy and Taylor (2000) among others. Denote Z_i the vector of covariates for the logistic regression part, and $\tilde{Z}_i$ as the vector of covariates for the Cox regression. So we have

$$p_i = \frac{\exp(\alpha' Z_i)}{1 + \exp(\alpha' Z_i)}, \tag{2}$$

where α is a vector of regression parameters. For the hazard function of T_i we have

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' \tilde{Z}_i), \tag{3}$$

where $\lambda_0(t)$ is the baseline hazard function, and β is a vector of regression parameters. Then $S_i(t) = \exp\{-\Lambda_0(t) e^{\beta' \tilde{Z}_i}\}$, with $\Lambda_0(t) = \int_0^t \lambda_0(u) du$.

Denote $\theta = (\alpha, \beta, \Lambda_0)$. Since we do not specify a parametric distribution for Q, the likelihood approach is conditional upon the observed Q_i's. Under Weibull regression models for T Struthers and Farewell (1989) considered the marginal likelihood under model (1):

$$L_m(\theta) = \prod_{i=1}^{n} \left\{ \frac{\bar{f}_i(X_i)}{\bar{S}_i(Q_i)} \right\}^{\delta_i} \left\{ \frac{\bar{S}_i(X_i)}{\bar{S}_i(Q_i)} \right\}^{1-\delta_i}, \tag{4}$$

where $\bar{f}_i(t) = -d\bar{S}_i(t)/dt = p_i f_i(t)$, and $f_i(t) = -dS_i(t)/dt$. The above likelihood does not involve the many observed $Y_i = 0$'s, which are the majority of the cases in our pregnancy data. Not making use of these $Y_i = 0$'s, as will be illustrated later, can result in substantial loss of information. Computationally the marginal likelihood is also not straightforward to maximize under the semiparametric regression models we consider here.

Instead we consider the complete data likelihood:

$$\begin{aligned} L_c(\theta; X, \delta, Y|Q) &= \prod_{i=1}^{n} L_1(\theta; Y_i|T_i > Q_i) \cdot L_2(\theta; X_i, \delta_i|Y_i = 1, T_i > Q_i)^{Y_i} \\ &= \prod_{i=1}^{n} \frac{p_i^{Y_i} S_i(Q_i)^{Y_i} (1-p_i)^{1-Y_i}}{(1-p_i) + p_i S_i(Q_i)} \cdot \left\{ \frac{f_i(X_i)}{S_i(Q_i)} \right\}^{\delta_i Y_i} \left\{ \frac{S_i(X_i)}{S_i(Q_i)} \right\}^{(1-\delta_i)Y_i} \end{aligned} \tag{5}$$

Note that the only unobserved Y_i's are those right-censored before time τ, and the observed $Y_i = 0$'s are explicitly counted for in the likelihood above. Unlike for cure rate data without left truncation, however, (5) does not split into likelihoods from the logistic and the Cox regressions separately, and this poses computational challenges. In the following we consider two approaches based on (5) to facilitate the estimation problem computationally.

A weighted likelihood

In the first approach we consider the biased sample likelihood $p_i^{Y_i}(1-p_i)^{1-Y_i}$ in place of L_1, and then weigh each observation by its estimated inverse probability $1/P(T_i > Q_i)$:

$$L_w(\theta) = \prod_{i=1}^{n} \left[p_i^{Y_i}(1-p_i)^{1-Y_i} \left\{ \frac{f_i(X_i)}{S_i(Q_i)} \right\}^{\delta_i Y_i} \left\{ \frac{S_i(X_i)}{S_i(Q_i)} \right\}^{(1-\delta_i)Y_i} \right]^{\nu_i}, \tag{6}$$

where $\nu_i = 1/\hat{P}(Q_i < T_i)$ is an estimated inverse probability of not being 'truncated out'. Inverse probability weighting (IPW) originated in Horvitz and Thompson (1952), and is often used in causal inference and missing data problems (Robins *et al.*, 1994; Hogan

and Lancaster, 2004). It is easy to see that without the weights the likelihood would be biased: assume that there is no censoring so all the Y_i's are observed, and assume that the logistic regression only has an unknown intercept α_0, it is immediately seen that the maximum likelihood estimate of $p = P(Y = 1)$ is the observed portion of $Y_i = 1$. This, as we explained earlier, is biased downward due to left-truncation. In other words, the unweighted likelihood would not account for those individuals who have early events and who are truncated out. The weighted likelihood in (6) has the computational advantage of separating the logistic and the Cox regression parameters, hence existing software for fitting the Cox model can be used in the computational algorithm. Weighted likelihood for semiparametric models was considered in Breslow and Wellner (2007) for two-phase stratified sampling.

For the weights ν_i we have considered using $P(Q_i < T_i|Q_i) = S_i(Q_i)$, but our preliminary attempt shows that it leads to numerical instability in an iterative model fitting procedure. Instead let $G(t) = P(Q < t)$ be the cumulative distribution function (CDF) of Q. For the time being we assume that G does not depend on the covariates (see more discuss on this later), and $\hat{G}$ is its estimator to be specified in Section 2.1. Then we define

$$\nu_i^{-1} =: \hat{G}(T_i) \text{ if } \delta_i = 1; \;\; \hat{p}_i\hat{G}(T_i^*) + (1 - \hat{p}_i) \text{ if } \delta_i = 0, \tag{7}$$

where $T_i^* \in (X_i, \tau)$ is an imputed value of T_i, $\hat{p}_i$ is obtained according to (2) during an iteration procedure to be described in Section 2.2, and the case for $\delta_i = 0$ is based on the fact that $P(Q_i < T_i) = P(Q_i < T_i|Y_i = 1)P(Y_i = 1) + P(Y_i = 0)$. Theoretically T_i^* should be imputed from the conditional distribution of T_i given censoring etc, but to keep the computation simple we impute it from Uniform (X_i, τ). Note that when there was no censoring, weighting by $\hat{G}(T_i)^{-1}$ was also considered in Gross (1996).

An approximate likelihood

In the second approach we consider an approximate likelihood where $S_i(Q_i) = P(T_i > Q_i|Q_i)$ in L_1 of (5) is replaced by an estimate of $G(T_i) = P(T_i > Q_i|T_i)$ if T_i is observed, and by ν_i^{-1} defined in (7) in general. The approximate likelihood is then

$$L_a(\theta) = \prod_{i=1}^{n} \frac{(p_i/\nu_i)^{Y_i}(1 - p_i)^{1-Y_i}}{(1 - p_i) + p_i/\nu_i} \cdot \left\{ \frac{f_i(X_i)}{S_i(Q_i)} \right\}^{\delta_i Y_i} \left\{ \frac{S_i(X_i)}{S_i(Q_i)} \right\}^{(1-\delta_i)Y_i}. \tag{8}$$

This approach turns out to have many common computational elements as in the first approach above.

2.1 Estimation of G

Both approaches described above requires estimation of $G(t) = P(Q < t)$. When the distribution of T does not depend on the covariates, the NPMLE of $G(t)$ was studied in Wang (1991). In the absence of censoring it reduces to a product-limit (PL) estimator for G. In addition, Turnbull (1976) studied the NPMLE for arbitrarily truncated data using a self-consistent algorithm. Notice that Q is right truncated by T for the non-cured individuals.

A more straightforward approach for our data, since the 'cured' sample is not subject to left truncation, is to estimate G using the empirical distribution function from that sample:

$$\hat{G}(t) = \frac{\sum_{i=1}^{n} I(Q_i \le t, Y_i = 0, \delta_i = 1)}{\sum_{i=1}^{n} I(Y_i = 0, \delta_i = 1)} = \frac{\sum_{i=1}^{n} I(Q_i \le t, T_i > \tau, C_i > \tau)}{\sum_{i=1}^{n} I(T_i > \tau, C_i > \tau)}. \tag{9}$$

Note that the independent censoring assumption ensures the consistency of the above estimator. This estimator is more straightforward for the overall inference, and is what we

use for the remainder of this paper. In the Appendix we also show that for the weighted approach, as long as $G(t)$ is consistently estimated, the asymptotically variance of $\hat{\theta}$ is unaffected (to the first order).

2.2 ES Algorithm

Both the weighted and the approximate likelihoods lead to a set of (complete data) estimating equations involving the parametric and the nonparametric components, in the same way the score equations from nonparametric likelihoods do. For parametric estimating equations with missing data in general, Elashoff and Ryan (2004) developed an EM-type algorithm in which the update to the parameters is found by substituting in the expected values of sufficient statistics of missing data based on current parameter estimates. They call it the ES algorithm. Rosen *et al.* (2000) also developed the same ES algorithm in the context of mixtures of marginal models. In the following we detail the algorithm for the weighted likelihood approach; it is similar for the approximate likelihood, and the corresponding details are provided in the Appendix.

From (6) the weighted complete data log-likehood is:

$$\ell_w(\theta) = \sum_{i=1}^{n} \nu_i \Big[Y_i \alpha' Z_i - \log\left\{1 + \exp(\alpha' Z_i)\right\} + Y_i \{\Lambda_0(Q_i) - \Lambda_0(X_i)\} \exp(\beta' \tilde{Z}_i) + \delta_i Y_i \{\beta' \tilde{Z}_i + \log \lambda_0(X_i)\} \Big]. \quad (10)$$

Let ℓ_i denote the contribution from the i^{th} individual to (10) without the weights ν_i. Let $0 < t_1 < ... < t_K < \tau$ be the distinct observed event times, and λ_k be the point mass of λ_0 at t_k. Let $0 < q_1 < ... < q_J < \tau$ be the distinct truncation times of the observed cured individuals, and $G_j = \hat{G}(q_j)$ as defined in (9). Taking derivatives of (10) with respect to α, β and λ_k ($k = 1, ..., K$) gives the first three of the following complete data estimating equations, and (9) gives the last estimating equation for G_j ($j = 1, ..., J-1$; note that $G_J = 1$):

$$U^{(\alpha)} = \sum_{i=1}^{n} U_i^{(\alpha)} = \sum_{i=1}^{n} \nu_i \frac{\partial \ell_i}{\partial \alpha} = \sum_{i=1}^{n} \nu_i Z_i \left\{ Y_i - \frac{\exp(\alpha' Z_i)}{1 + \exp(\alpha' Z_i)} \right\} = 0, \quad (11)$$

$$U^{(\beta)} = \sum_{i=1}^{n} U_i^{(\beta)} = \sum_{i=1}^{n} \nu_i \frac{\partial \ell_i}{\partial \beta} = \sum_{i=1}^{n} \nu_i Y_i \tilde{Z}_i [\delta_i + \{\Lambda_0(Q_i) - \Lambda_0(X_i)\} \exp(\beta' \tilde{Z}_i)] = 0, \quad (12)$$

$$U^{(\lambda_k)} = \sum_{i=1}^{n} U_i^{(\lambda_k)} = \sum_{i=1}^{n} \nu_i \frac{\partial \ell_i}{\partial \lambda_k} = \sum_{i=1}^{n} \nu_i Y_i \left\{ \frac{\delta_i I(X_i = t_k)}{\lambda_k} - \exp(\beta' \tilde{Z}_i) I(Q_i < t_k \leq X_i) \right\} = 0, \quad k = 1, ..., K, \quad (13)$$

$$U^{(G_j)} = \sum_{i=1}^{n} U_i^{(G_j)} = \sum_{i=1}^{n} \left\{ \frac{G_j}{N} - \frac{I(Q_i \leq q_j, Y_i = 0, \delta_i = 1)}{\sum_{i=1}^{n} I(Y_i = 0, \delta_i = 1)} \right\} = 0, \quad j = 1, ..., J-1. \quad (14)$$

Note that (14) does not involve any unknown parameters other than G_j or any unobserved Y_i's. We list (14) together with (11) - (13) here for the purposes of deriving the variance estimator later.

The ES algorithm proceeds iteratively for (11) - (13). We can take the initial value for α to be either zero or from a logistic regression model fit ignoring the right-censored observations, and the initial value for β and Λ_0 to be from a Cox regression model fit taking into account left truncation and treating all the cured subjects as right-censored at τ. In the

E-steps since the unobserved Y_i's enter the above complete data equations linearly, we only need to compute $\omega_i = E(Y_i|\text{observed data})$, where the expectation is computed using the current value of the parameters. Note that when Y_i is observed $\omega_i = Y_i$; otherwise

$$\begin{aligned}
\omega_i &= E(Y_i|T_i > X_i, T_i > Q_i) \\
&= P(Y_i = 1|T_i > X_i, T_i > Q_i) \\
&= \frac{P(T_i > X_i|T_i > Q_i, Y_i = 1)P(T_i > Q_i|Y_i = 1)P(Y_i = 1)}{P(T_i > X_i, T_i > Q_i|Y_i = 1)P(Y_i = 1) + P(T_i > X_i, T_i > Q_i|Y_i = 0)P(Y_i = 0)} \\
&= \frac{\exp(\alpha' Z_i)\exp\left[\{\Lambda_0(Q_i) - \Lambda_0(X_i)\}e^{\beta'\tilde{Z}_i}\right] G(T_i)}{\exp(\alpha' Z_i)\exp\left[\{\Lambda_0(Q_i) - \Lambda_0(X_i)\}e^{\beta'\tilde{Z}_i}\right] G(T_i) + 1} \\
&= \frac{\exp(\alpha' Z_i)\exp\left[\{\Lambda_0(Q_i) - \Lambda_0(X_i)\}e^{\beta'\tilde{Z}_i}\right] /\nu_i}{\exp(\alpha' Z_i)\exp\left[\{\Lambda_0(Q_i) - \Lambda_0(X_i)\}e^{\beta'\tilde{Z}_i}\right] /\nu_i + 1},
\end{aligned} \tag{15}$$

where $G(T_i)$ in the second last line is replaced by ν_i^{-1} from (7) in the above. Note that in the ES algorithm the E-step for the missing data mechanism can be specified separately from the estimating equations of the S-step.

After substituting the unobserved Y_i's in (11) - (13) by ω_i defined in (15) above, the S-step then obtains the updated parameters by solving the equations. In particular the updates for α can be obtained separately from those for β and Λ_0. From $U^{(\alpha)}$ we use Newton-Raphson methods to compute the update for α. Updates for β and Λ_0 turn out to follow the standard Cox model estimation procedure for left truncated and right-censored data with weights $\nu_i\omega_i$. That is,

$$\hat{\lambda}_k = \frac{d_k}{\sum_{i:Q_i < t_k \le X_i} \nu_i\omega_i \exp(\hat{\beta}'\tilde{Z}_i)}, \qquad k = 1, ..., K; \tag{16}$$

and $\hat{\beta}$ can be found by the corresponding weighted partial likelihood.

3. Inference

Let $\phi = (\theta, G) = (\alpha, \beta, \Lambda_0, G)$. The identifiability of the cure-rate model considered here was shown in Li *et al.* (2001). We emphasize that in our data many of the $Y_i = 0$'s are observed, which greatly improves the 'practical' identifiability of the model (Farewell, 1986).

Let $U = ({U^{(\alpha)}}^\top, {U^{(\beta)}}^\top, {U^{(\lambda)}}^\top, {U^{(G)}}^\top)^\top$, where $U^{(\lambda)}$ denotes the vector of $U^{(\lambda_k)}$ ($k = 1, ..., K$), and $U^{(G)}$ denotes the vector of $U^{(G_j)}$ ($j = 1, ..., G-1$). We use similar notations $U_i = ({U_i^{(\alpha)}}^\top, {U_i^{(\beta)}}^\top, {U_i^{(\lambda)}}^\top, {U_i^{(G)}}^\top)^\top$ for contribution from subject i. Let $\mathcal{U}_i = E\{U_i|\text{observed data}\}$, and let $\mathcal{U} = \sum_{i=1}^n \mathcal{U}_i$. At convergence of the ES algorithm described above, it can be seen that the parameter estimate $\hat{\phi}$ solves

$$\mathcal{U}(\phi) = E_\phi\{U(\phi)|\text{observed data}\} = 0. \tag{17}$$

This is in fact the first Louis (1982) formula for observed data score function when proper likelihood is used.

To estimate the variance of $\hat{\phi}$, we consider the sandwich estimator for semiparametric Z-estimators (Van der Vaart and Wellner, 1996; Kosorok, 2008). Since ν_i^{-1} from (7) involves imputation when $\delta_i = 0$, we derive the closed form expressions below when there is no right-censoring. Right-censoring for variance estimation is dealt with separately in

Section 3.1 using multiple imputation techniques. More precisely, let $h = (h_1, h_2, h_3, h_4)$ where h_1 and h_2 are vectors of the same dimension as α and β, respectively, and h_3 and h_4 are functions of bounded variation on $[0, \tau]$. Define the sandwich estimator:

$$V_n(\hat{\phi}) = \left\{ \frac{\partial \mathcal{U}}{\partial \phi}(\hat{\phi}) \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathcal{U}_i(\hat{\phi}) \mathcal{U}_i(\hat{\phi})^\top \right\} \left[\left\{ \frac{\partial \mathcal{U}}{\partial \phi}(\hat{\phi}) \right\}^{-1} \right]^\top, \tag{18}$$

where the formulas for $\partial \mathcal{U}/\partial \phi$ are given in the Appendix, and $\mathcal{U}_i$ is U_i given in (11) - (14) with Y_i replaced by ω_i. We also show in the Appendix that even though the $\mathcal{U}_i(\phi_0)$'s are correlated due to the ν_i's in them, $\sum_{i=1}^{n} \mathcal{U}_i(\hat{\phi}) \mathcal{U}_i(\hat{\phi})^\top / n$ does provide a consistent estimate of the asymptotic variance of $\sum_{i=1}^{n} \mathcal{U}_i(\phi_0)/\sqrt{n}$. If we denote h_n as the vector with elements h_1, h_2, $h_3(X_i)$ at those X_i where $\delta_i = 1$, and $h_4(Q_i)$ from those cured individuals, then $n h_n' V_n h_n$ estimates the asymptotic variance of

$$\sqrt{n}\{h_1'(\hat{\alpha} - \alpha_0) + h_2'(\hat{\beta} - \beta_0) + \int_0^\tau h_3(u) d(\hat{\Lambda}_0 - \Lambda_0)(u) + \int_0^\tau h_4(u) d(\hat{G} - G_0)(u)\}. \tag{19}$$

3.1 Multiple Imputation Variance

As mentioned previously, in order to calculate ν_i as defined in (7) when X_i is right-censored ($\delta_i = 0$), we impute a T_i^* from $U(X_i, \tau)$. To account for this additional uncertainty in the variance estimation, we can use the usual approach under multiple imputation. Multiple imputation is also used when the event times are interval censored (Pan, 2000; Chen and Sun, 2010), which also exists in SAB data in general, since a woman may not remember the exact date when it occurred; this, however, is not the focus of our paper here. The resulting variance estimator is the sum of the average variance estimates using the imputed values and an additional component due to the imputation.

For $m = 1, ..., M$ let $\hat{\phi}^{(m)}$ be the parameter estimate based on the m^{th} set of imputed data, and let $\bar{\phi}$ be the average of these M parameter estimates. Then our variance estimator with imputation is:

$$\widehat{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} V_n(\hat{\phi}^{(m)}) + \left(1 + \frac{1}{M}\right) \sum_{m=1}^{M} \frac{(\hat{\phi}^{(m)} - \bar{\phi})(\hat{\phi}^{(m)} - \bar{\phi})^\top}{M - 1}. \tag{20}$$

In our simulation studies and data analysis we use $M = 10$.

4. Spontaneous Abortion Data

The data we investigate come from the OTIS autoimmune disease database as mentioned earlier. Our sample includes pregnant women who entered a research study between 2005 and 2012. It consists of $n = 964$ women who entered the study before week 20 of their gestation. Among these 499 (52%) were pregnant women with certain autoimmune diseases who were treated with selected newer medications, 272 (28%) were with the same specific autoimmune diseases who were not treated with the selected newer medications, and the rest 193 (20%) were healthy pregnant women without autoimmune diseases who were not treated with the selected newer medications. Chambers *et al.* (2001) discussed the importance of having a diseased control group, since some of the adverse outcomes in pregnancy may be due to the disease instead of the medication. There were a total of 74 SAB events, and 21 women were lost to follow up.

4.0.1 Length-biased data and Testing the uniformity of G

Before fitting our models to the data, we note the recent methods developed for length-biased data and, if the length-biased assumption holds the conditional (on the truncation times) approach we use is not efficient (Asgharian *et al.*, 2002). This assumption means that the truncation distribution G is uniform. For left truncated and right-censored survival data, Asgharian *et al.* (2006) suggested a visual test to check whether the truncation distribution is uniform. If the truncation times are uniform over the duration of the study, the Kaplan-Meier curves for the truncation times Q and the right-censored residual times $(X - Q)$ should show no significant difference. Mandel and Betensky (2007) derived the corresponding paired log-rank test to this visual test; they also observed that for non-censored data this is equivalent to testing the distribution of the Q_i/T_i against Uniform $(0, 1)$ and the Kolmogorov-Smirnov test can be used. For our data we may apply the Kolmogorov-Smirnov test to the observed cured portion of the sample; we also apply the visual test to the non-cured portion.

We have noted before that the cured population is not subject to left truncation, hence the empirical distribution of the truncation times from the observed cured provides a consistent estimate of G, under the independent censoring assumption. The Kolmogorov-Smirnov test comparing $Q/20$ for these cured individuals with $U(0, 1)$ yields a p-value < 0.01.

4.0.2 Fitting the cure models

There are a number of risk factors for spontaneous abortion that have been identified in the literature; see for example Chambers *et al.* (2013). Strictly speaking these are known to be risk factors in the logistic part of the cure model, but here we extend them to be also considered in the Cox part of the model, i.e. for the timing of SAB events. These include maternal age, prior SAB (Y/N), prior elective abortion (TAB, Y/N), and smoking (Y/N). For these covariates we fit our regression models to the data, and the results using the weighted likelihood are given in Table 1 left columns. The results using the approximate likelihood are qualitatively similar. As before we used ten imputed values for the censored survival times to compute the variance of the estimates.

From Table 1 we see that older maternal age (> 34) significantly increase the probability of SAB in the logistic part of the model. Healthy controls have significantly lower probability of SAB compared to the autoimmune disease drug exposed women, but the probability of SAB is not significantly different between the diseased control and the exposed women. The Cox regression part of the model identified prior TAB as a significant factor for the hazard of SAB; this in the cure model context should be understood as significantly later timing of SAB during the first 20 weeks of gestation for those who had prior TAB.

It is of interest to contrast the above results with any analysis that might have been done without the methodology developed in this paper. Naively, one may fit just a logistic regression model taking yes or no SAB as outcome, and exclude the lost to follow-up subjects. As shown in the right columns of Table 1, this will still identify maternal age and disease groups as significant risk factors, with the same signs of the regression coefficients as in the cure model. But the logistic intercept will be substantially underestimated, which can lead to erroneous predicted probabilities of SAB events.

Counting for the left truncation, survival analysis methods have been advocated in the literature for the analysis of SAB data (Meister and Schaefer, 2008; Xu and Chambers, 2011). Table 1 right columns also show the results of the Cox regression model fitted to the data by treating all the cured individuals as right-censored at 20 weeks of gestation. The

Table 1: Cure rate model versus naive model fits for SAB data

	Cure model		Separate models	
	Estimate (SE)	P-value	Estimate (SE)	P-value
Logistic				
Intercept	* -1.83 (0.22)	<0.01	*-2.58 (0.22)	<0.01
Maternal Age > 34	0.77 (0.29)	<0.01	0.60 (0.25)	0.02
Prior SAB	-0.13 (0.29)	0.65	0.06 (0.27)	0.81
Prior TAB	-0.55 (0.43)	0.20	-0.21 (0.40)	0.60
Smoking	0.30 (0.35)	0.39	0.24 (0.30)	0.43
Healthy Control	-1.33 (0.46)	<0.01	-1.11 (0.45)	0.01
Diseased Control	-0.09 (0.30)	0.77	-0.14 (0.27)	0.61
Cox PH				
Maternal Age > 34	0.18 (0.44)	0.70	0.11 (0.26)	0.69
Prior SAB	0.03 (0.31)	0.93	0.26 (0.29)	0.38
Prior TAB	-0.66 (0.27)	0.02**	-0.45 (0.39)	0.25**
Smoking	-0.05 (0.45)	0.92	-0.17 (0.32)	0.60
Healthy Control	-0.54 (0.51)	0.29	-0.50 (0.50)	0.32
Diseased Control	0.11 (0.34)	0.76	0.29 (0.27)	0.29

results are such that there are no significant predictors of SAB. This, as explained before, is because we treat the majority of the women (who did not have SAB) as right-censored, leading to substantial loss of information. In addition, under the proportional hazards assumption, non-significant effect of prior TAB would translate to no significant difference in the cumulative risks of SAB, which is in fact consistent with the non-significance of prior TAB in the logistic part of the cure model. In contrast, the cure model methodology we have developed here is able to make use of the information from both the women who had SAB and those who were observed not to have SAB, as well as to separate the differential regression effects like prior TAB on both the cumulative risk of SAB as well as the timing of it among those who experience SAB.

References

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association*, **97**(457), 201–209.

Asgharian, M., Wolfson, D. B., and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in medicine*, **25**(10), 1751–1767.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, **34**(1), 86–102.

Chambers, C. D., Braddock, S. R., Briggs, G. G., Einarson, A., Johnson, Y. R., Miller, R. K., Polifka, J. E., Robinson, L. K., Stepanuk, K., and Jones, K. L. (2001). Post-marketing surveillance for human teratogenicity: a model approach. *Teratology*, **64**, 252–261.

Chambers, C. D., Johnson, D., Xu, R., and Jones, K. 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. *Poster presented at the 27th International Conference on Pharmacoepidemiology and Therapeutic Risk Management — Chicago, IL, USA*.

Chambers, C. D., Johnson, D., Xu, R., Luo, Y., Louik, C., Mitchell, A. A., Schatz, M., and Jones, K. L. (2013). Risks and safety of pandemic h1n1 influenza vaccine in pregnancy: Birth defects, spontaneous abortion, preterm delivery, and small for gestational age infants. *Vaccine*, **31**(44), 5026–5032.

Chen, L. and Sun, J. (2010). A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. *Computational Statistics & Data Analysis*, **54**(4), 1109–1116.

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**(447), 909–919.

Elashoff, M. and Ryan, L. (2004). An em algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, **13**(1), 48–65.

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**(3), 257–262.

Gross, S. T. (1996). Weighted estimation in linear regression for truncated survival data. *Scandinavian Journal of Statistics*, **23**, 179–193.

Hanson, T., Bedrick, E. J., Johnson, W. O., and Thurmond, M. C. (2003). A mixture model for bovine abortion and foetal survival. *Statistics in medicine*, **22**(10), 1725–1739.

Hogan, J. W. and Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, **13**, 17–48.

Horvitz, D. G. and Thompson, D. J. (1952). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, **169**(11), 1398–1405.

Kim, Y.-J. and Jhun, M. (2008). Cure rate model with interval censored data. *Statistics in medicine*, **27**(1), 3–14.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Kuk, A. Y. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**(3), 531–541.

Li, C.-S., Taylor, J. M., and Sy, J. P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, **54**(4), 389–395.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, **91**(2), 331–343.

Mandel, M. and Betensky, R. A. (2007). Testing goodness of fit of a uniform truncation model. *Biometrics*, **63**(2), 405–412.

Meister, R. and Schaefer, C. (2008). Statistical methods for estimating the probability of spontaneous abortion in observational studies – analyzing pregnancies exposed to coumarin derivatives. *Reproductive Toxicology*, **26**, 31–35.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, **56**(1), 199–203.

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.

Robins, J., Rotnitzky, A., and L, Z. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.

Rosen, O., Jiang, W., and Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika*, **87**, 391–404.

Struthers, C. A. and Farewell, V. T. (1989). A mixture model for time to aids data with left truncation and an uncertain origin. *Biometrika*, **76**(4), 814–817.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, **56**(1), 227–236.

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, **74**(4), 883–886.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290–295.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, **86**(413), 130–143.

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**(4), 189–194.

Xu, R. and Chambers, C. (2011). A sample size calculation for spontaneous abortion in observational studies. *Reproductive Toxicology*, **32**(4), 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.