LENGTH-BIASED SEMI-COMPETING RISKS MODELS FOR CROSS-SECTIONAL DATA: AN APPLICATION TO CURRENT DURATION OF PREGNANCY ATTEMPT DATA

BY ALEXANDER C. McLAIN1, SIYUAN GUO1,∗, MARIE THOMA2, AND JIAJIA ZHANG1,†

1Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208 USA mclaina@mailbox.sc.edu; ∗siyuang@email.sc.edu; †jzhang@mailbox.sc.edu
2Department of Family Health Services, University of Maryland, College Park, MD, 20742 USA mthoma@umd.edu

Cross-sectional length-biased data arise from questions on the at-risk time for an event of interest from those who are at-risk but have yet to experience the event. For example, in the National Survey on Family Growth (NSFG), women who were currently attempting to become pregnant were asked how long they had been attempting pregnancy. Cross-sectional survival analysis methods use the observed at-risk times to make inference on the distribution of the unobserved time-to-failure. However, methodological gaps in these methods remain such as how to handle semi-competing risks. For example, if the women attempting pregnancy had undergone fertility treatment during their current pregnancy attempt. In this paper, we develop statistical methods that extend cross-sectional survival analysis methods to incorporate semi-competing risks. They can be used to estimate the distribution of the length of natural pregnancy attempts (i.e., without fertility treatment) while correctly accounting for women that sought fertility treatment prior to being sampled using cross-sectional data. We demonstrate our approach based on simulated data and an analysis of data from the NSFG. The proposed method results in separate survival curves for: time-to-natural-pregnancy, time-to-fertility treatment, and time-to-pregnancy after fertility treatment.

1. Introduction. Cross-sectional studies of time-to-event outcomes occur when data are observed on the right-censored time for an event of interest, from those currently at-risk. Such length-biased cross-sectional survival data are seen in studies of time-to-pregnancy (TTP) (Weinberg and Gladen, 1986; Keiding et al., 2002; McLain et al., 2014), residential migration (Allison, 1985; Yamaguchi, 2003), length of hospitalization time after surgery (Mandel and Rinott, 2014; Mandel, 2015) and survival of dementia patients (Bergeron, Asgharian and Wolfson, 2008; Carone, Asgharian and Jewell, 2014). The benefits of cross-sectional time-to-event designs include that it may be easier to sample people while they are currently at-risk (e.g., with hospitalization patients), are more cost and time effective than prospective studies and they will include subjects that never experienced the event (missing from some retrospective TTP studies, Scheike and Keiding, 2006). A difficulty with data arising from such studies is that the responses are length biased (subjects with longer event times are more likely to be at-risk) and entirely right censored since all subjects have not experienced the event of interest.

Cox (1969) originally developed the survival analysis theory for how to provide inference on the unobserved event of interest (see van Es, Klaassen and Oudshoorn, 2000, for a formal derivation). Parametric survival models have been considered by Allison (1985), Weinberg and Gladen (1986), Keiding et al. (2002) and Keiding et al. (2012), among others. It is possible that a time-dependent intermittent event occurs during some subjects’ at-risk periods.

MSC2020 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: cross-sectional data, infertility, length-biased survival, semi-competing risks.
That is, between the start of the at-risk period and the sampling time the subject experienced an intermittent event that changed the failure rate. The semi-competing risks framework (also referred to as the illness-death model) can handle this type of data. Mandel and Fluss (2009) and Mandel (2010) considered a competing risks illness–death model under cross-sectional sampling to estimate the probability of illness and the rates of illness and death among hospitalized patients. Their methods, however, required follow-up on the total event time which will not be available in some cross-sectional studies. Keiding (1991) considered a situation very similar to that given here (see Section 6 therein) though he does not give a full treatment of the problem.

In this paper, we develop statistical methods that can be used to analyze cross-sectional length-biased survival data that are subjected to semi-competing risks. Our methods only require data available at the cross-sectional sampling time. We use a piecewise constant hazard model to estimate the underlying distribution of the event of interest and the semi-competing risk. We use a latent variable treatment of the semi-competing risks where the inference is based on the latent outcomes, which some have critiqued as lacking real world interpretability (Andersen and Keiding, 2012). An alternative approach is to model the cumulative incidence function or subdistribution, based on the joint probability of the outcome and event type (Fine and Gray, 1999; Fine, Jiang and Chappell, 2001). The motivation for using the latent variable framework is that the theoretical outcome of interest from this model coincides with TTP if fertility treatment were not an option. This outcome is representative of fecundity, the biologic ability to achieve pregnancy. Alternatively, the subdistribution of TTP and not entering fertility treatment has less biologic relevance. We do note that modeling the cumulative incidence function is more preferable for the other outcomes. For example, Duron et al. (2013) estimate the cumulative incidence function medical consultation for fecundity problems. TTP if fertility treatment were not an option is a counterfactual outcome, which calls for causal inference survival analysis methods. These methods have strong assumptions that may be hard to justify with the available data (discussed more in Section 7). As a result, in this paper we focus on estimating the distribution of TTP while correcting for the informative censoring caused by infertility treatment through a semi-competing risks model.

The outline of the paper is as follows. In Section 2, we introduce the motivating data set and the application of interest. In Section 3, we discuss the standard theory for cross-sectional length biased data, give the assumptions of the procedure, and derive the densities and likelihood when semi-competing risks are present. In Section 4, we present an approach to incorporate dependence between the latent variables. In Section 5, we present results from simulation studies. In Section 6, we apply our methods to data from the National Survey on Family Growth (NSFG) to estimate the distribution of TTP while correctly accounting for women that have undergone fertility treatment (the semi-competing risk) and compare our results to competing methods. We conclude with some discussion.

2. Motivating Data. Cross-sectional data on the length of pregnancy attempt are used in reproductive epidemiology to estimate the survival curve of TTP, a technique that has been referred to as the current duration approach (Keiding et al., 2002). The data arise when women are asked, “Are you currently trying to get pregnant” and if the woman said ‘yes,’ “How long have you been currently attempting pregnancy?” The responses to the latter question are referred to as the woman’s current duration of pregnancy attempt. In the NSFG, women are also asked to list the dates and types of fertility treatment, which complicates the analysis if fertility treatment occurred during the current attempt. That is, the moment a woman began fertility treatment, her ‘natural’ pregnancy attempt stopped and her pregnancy attempt on fertility treatment began.
The conceptual data model for a successful pregnancy attempt is presented in Figure 1 where $T$ represents the time-to-natural-pregnancy, which we refer to TTPN, $Y$ is the time-to-fertility treatment (TTFT) and $Z$ the time-to-pregnancy after fertility treatment (TTPFT) where $Z = 0$ if $T < Y$. This is viewed as a semi-competing risks problem or through the illness-death model, since fertility treatment always occurs before pregnancy, but does not always occur. Here, TTP is used to denote time-to-pregnancy which could be aided by fertility treatments. We note that pregnancy attempts can be ‘unsuccessful’ and end for other reasons (e.g., giving up). That is, either a natural pregnancy attempt or a pregnancy attempt after fertility treatment can end without the women getting pregnant. The current duration approach provides inference on the end of attempts and cannot differentiate between ‘successful’ and ‘unsuccessful’ pregnancy attempts (Keiding et al., 2002, 2012).

As opposed to the standard semi-competing risks scenario, in a cross-sectional sample we observe the current at-risk period and the time (if any) that the intermittent event occurred. For example, let $\sigma$ denote the time the woman’s pregnancy attempt begins and $\tau$ the sampling time. Without loss of generality, we assume that $\tau$ is equal across subjects. In cross-sectional data, we only observe what occurred before $\tau$ for subjects at risk at $\tau$.

The observed data are displayed in Figure 1 for five hypothetical women as a Lexis diagram. Events that occur in the gray area of the figure are unobserved with cross-sectional data. For example, with women 1 and 3 we observed the current length of their natural pregnancy attempts. For woman 2, we observe the TTFT ($Y$) and the current duration of their fertility treatment. At $\tau$ it is not known that woman 1 will get pregnant naturally and woman 3 will enter fertility treatment. No data from women 4 or 5 are observed since they are not at-risk for pregnancy when surveyed. This is a unique feature that sets it apart from standard length-biased survival data where subjects are followed from $\tau$ to the event of interest or censoring.

3. Methods. This section begins with the data notation and an overview of the derivation of the likelihood function for a general dependence structure (full details are given in Sec-
3.1. Notation, assumptions and likelihood. The observed data consist of independent realizations of $D_i = \{X_i, Z^*_i, \delta_i\}$ for $i = 1, 2, \ldots, n$, where $X_i$ is the observed time at risk (i.e., $\tau_i - \sigma_i$ in Figure 1), $Z^*_i$ is the time since the intermittent event where $Z^*_i = 0$ if the intermittent event has not occurred, and $\delta_i = I(Z^*_i > 0)$ for subject $i$. For ease of notation, let $Y$ denote the time to the intermittent event. Note that $Z^*$ is the observed time since the intermittent event (i.e., the at-risk time since the intermittent event), while $Z$ is the (unobserved) full time from the intermittent event until failure. If $\delta_i = 0$, then $X_i$ represents the current duration of $T_i \wedge Y_i$ where $T \wedge Y = \min(T, Y)$. If $\delta_i = 1$, we observe $Y_i$ and $Z^*_i$. We use $f$, $F$, $\bar{F}$ to denote a generic probability measure, distribution function and survival function, respectively, and subscript the appropriate random variable. Our derivation of the likelihood extends results given by Dauxois, Guilloux and Kirmani (2014) on length-biased competing risks data and arguments given by Vardi (1989). Most of the technical details of the derivation are given in the Supplemental Materials, here we outline the main concepts.

Our formation of the likelihood uses some important regularity conditions. Throughout we assume that (a) the beginning of the at-risk periods ($\sigma_i$) are governed by a homogeneous Poisson process with rate $\psi$, and (b) conditional on $\sigma_i$ the distributions of $T$, $Y$ and $Z$ are independent of calendar time (we discuss these assumptions further in Section 6 and Section A of the Supplemental Materials).

For $\delta = 0$ the derivation of the likelihood follows standard current duration results. Given $\delta = 0$, the results of Cox (1969) can be used to derive the form of the distribution of $X|\delta = 0$. Further, in the appendix we derive $\Pr(T|\delta = 0)$. In total, for those with $\delta = 0$ the likelihood contribution is

\begin{equation}
\label{eq:delta0}
\bar{f}_{X,\delta}(x,0) = \frac{\bar{F}_{T \wedge Y}(x)}{E(T \wedge Y + Z)}.
\end{equation}

The forms for $\bar{F}_{T \wedge Y}$ and $E(T \wedge Y + Z)$ are discussed below.

To derive the likelihood contribution for $\delta = 1$, it is useful to first consider the case where $Z$ is fully observed (i.e., if followup past $\tau$ were available) then use results on multiplicative censoring (Vardi, 1989) to obtain the density for the observed cross-sectional data at $\tau$. In the Supplemental Material, we show that the likelihood of $(Y, Z, \delta = 1)$ is

\begin{equation}
\label{eq:delta1}
f_{Y,Z,\delta}(y, z, 1) = \int y \, z f_{T, Y, Z}(u, y, z) du \frac{u}{E(T \wedge Y + Z)}.
\end{equation}

Since $Z$ occurs last in the temporal ordering of the variables it is most natural to consider the distribution of $Z|X, Y$, and represent the integrand in (3.2) as $z f_{T, X, Z}(u, y, z) = z f_{Z|T, Y}(z) f_{T, Y}(u, y)$ where, for brevity, we use $f_{Z|T, Y}(z) = f_{Z|T=t, Y=y}(z)$ throughout. Given the steady state assumptions, $\delta = 1$ and $Y$, we were equally likely to observe the current duration of $Z$ at any time point (van Es, Klaassen and Oudshoorn, 2000). As a result, the observed cross-sectional variable $Z^* \stackrel{d}{=} U \tilde{Z}$ where $U \sim \text{Uniform}(0, 1)$, $\tilde{Z} \stackrel{d}{=} Z|Y, \delta = 1$ and $\stackrel{d}{=} \text{denotes equal in distribution}$. Following these results we obtain the likelihood contribution of $(Y, Z^*, \delta = 1)$ as

\begin{equation}
\label{eq:delta1}
f_{Y, Z^*, \delta}(y, z, 1) = \int y \, z \bar{F}_{Z|T, Y}(z) f_{T, Y}(u, y) du \frac{u}{E(T \wedge Y + Z)}
\end{equation}

(see the Supplemental Materials for a formal derivation).
Combining the above results, and using the form for \( E(T \wedge Y + Z) \) given in the appendix, the full likelihood is given by

\[
L(D_i|\theta) \propto \frac{\bar{F}_{T \wedge Y}(X_i)^{(1-\delta_i)} \left\{ \int_{Y_i}^{\infty} \bar{F}_{Z|T,Y}(Z_i^*) f_{T,Y}(u,Y_i) du \right\}^\delta_i}{\int \bar{F}_{T \wedge Y}(u) du + \int \bar{F}_{T,Y}(u) du dv}.
\]

with \( \theta = \{ F_T, F_Y, F_Z, R \} \subset \Theta \) denoting the parameters where \( R \) denotes the dependence parameters (discussed further in Section 4). Performing maximum likelihood analysis of the complete data \( D = \{ D_1, D_2, \ldots, D_n \} \) amounts to maximizing \( l(\theta|D) = \sum_{i=1}^{n} \log \{ L(D_i|\theta) \} \) with respect to \( \theta \) such that \( \theta \subset \Theta \).

3.2. Likelihood under various dependence assumptions. It is common that semi-competing risks will have dependence through direct relationships or unobserved covariates. Below we discuss how the form of the likelihood in (3.4) can be simplified under some independence assumptions. Under complete independence, i.e., \( T \perp Y, T \perp Z \) and \( Y \perp Z \) where \( \perp \) denotes independence, we have

\[
L_C(D_i|\theta) \propto \frac{\bar{F}_{T}(X_i) \bar{F}_{Y}(X_i)^{(1-\delta_i)} \left\{ \bar{F}_{T}(Y_i) f_{Y}(Y_i) \bar{F}_{Z}(Z_i^*) \right\}^\delta_i}{\int \bar{F}_{T}(u) \bar{F}_{Y}(u) du + E(Z) \int \bar{F}_{T}(u) f_{Y}(u) du}.
\]

A more realistic assumption is to assume that \( Z \perp (T,Y) \), but still model the dependence between \( T \) and \( Y \). In this case, the likelihood is

\[
L_Z(D_i|\theta) \propto \frac{\bar{F}_{T \wedge Y}(X_i)^{(1-\delta_i)} \left\{ \bar{F}_{T|Y}(Y_i) f_{Y}(Y_i) \bar{F}_{Z}(Z_i^*) \right\}^\delta_i}{\int \bar{F}_{T \wedge Y}(u) du + E(Z) \int \bar{F}_{T|Y}(u) f_{Y}(u) du}.
\]

This assumption leads to a relatively computationally friendly likelihood (similar to complete independence), but \( Z \perp (T,Y) \) could be a strong assumption in practice.

Lastly, we considered two conditional independence assumptions. First, suppose \( T \perp Z|Y \), that is \( T \) and \( Z \) are conditionally independent given \( Y \). In this case, the likelihood is

\[
L_{TZ|Y}(D_i|\theta) \propto \frac{\bar{F}_{T \wedge Y}(X_i)^{(1-\delta_i)} \left\{ \bar{F}_{T|Y}(Y_i) f_{Y}(Y_i) \bar{F}_{Z}(Z_i^*) \right\}^\delta_i}{\int \bar{F}_{T \wedge Y}(u) du + \int E(Z|Y) \bar{F}_{T|Y}(u) f_{Y}(u) du}.
\]

Second, consider \( Y \perp Z|T \), which results in

\[
L_{YZ|T}(D_i|\theta) \propto \frac{\bar{F}_{T \wedge Y}(X_i)^{(1-\delta_i)} \left\{ \int_{Y_i}^{\infty} \bar{F}_{Z|T}(Z_i^*) f_{T,Y}(u,Y_i) du \right\}^\delta_i}{\int \bar{F}_{T \wedge Y}(u) du + \int E(Z|T) \bar{F}_{Y|T}(v) f_T(v) dv}.
\]

Assuming forms for all conditional distributions and expectations are available, it is evident that assuming \( T \perp Z|Y \) adds a relatively small amount of computational intensity over the \( Z \perp (T,Y) \) or complete independence assumptions and has a more flexible dependence structure. The \( Y \perp Z|T \) assumption adds computational complexity due to the additional integral in the numerator. Both conditional independence assumptions have considerable computational advantage over the general dependence structure. Particularly, the denominator of (3.4) is computationally demanding due to the double integral of \( E(Z|T,Y) \) (which must be evaluated numerically). As a result, we focus on the \( Y \perp Z|T \) assumption which (as discussed in Section 6) is well-motivated by our motivating example.
4. Dependence and Distributional Specifications. Building on the results from the previous section, we wish to model the dependence of \((T, Y, Z)\) such that forms for conditional distributions are available and flexible to various dependence assumptions. Further, in our motivating example, the hypothesized direction of the dependence between \(T\) and \(Y\) is negative, in that those with longer TTNP would have shorter TTFT. As a result, we consider a dependence structure that is amenable to positive or negative dependence. The Gaussian copula is particularly appealing since it incorporates negative dependence, has straightforward representations of conditional distributions and can incorporate conditional independence. The three-dimensional Gaussian copula is stated as

\[
C_R(u_1, u_2, u_3) = \Phi_R\{\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)\},
\]

where \(\Phi^{-1}\) is the inverse distribution function of a standard normal and \(\Phi_R\) represents the distribution function of a multivariate Gaussian distribution with mean \(0\) and correlation matrix \(R\) with elements \(r_{ij}\). The joint distribution function is given by \(\Pr(T \leq t, Y \leq y, Z \leq z) = C_R\{F_T(t), F_Y(y), F_Z(z)\}\).

The conditional independence assumptions discussed in Section 3.2 can be made by setting specific portions of \(R^{-1}\) to zero. To form the conditional distributions let \(W_T = \Phi^{-1}\{F_T(T)\}\), similarly for \(W_Y\) and \(W_Z\). The distribution function of \(W_Z|(W_T, W_Y)\) is given by

\[
\Pr(W_Z \leq w_z|W_{TY} = w_{TY}) = \Phi\left(\frac{w_{j+1} - r_Z\sigma_{Z|TY}^{-1}w_{TY}}{\sigma_Z}\right)
\]

where \(W_{TY} = (W_T, W_Y)\), \(R_{TY}^{-1}\) is the inverse correlation matrix of \(W_{TY}\), \(r_Z = (r_{TZ}, r_{YZ})\) are the correlations between \((W_T, W_Y)\) and \(W_Z\), and \(\sigma_{Z|TY}^2 = 1 - r_{Z|TY}^2 r_{TY}^{-1}r_Z\). The density and distribution functions of all the required conditional distributions can be formed following (4.1). For both simulations and real data analysis we assumed the direction of the dependence was known and appropriately modified the parameter space of the correlation parameters.

To model the marginal distributions of \(T\), \(Y\), and \(Z\) we seek a form that is flexible and amenable to the properties of the data in our motivating example. McLain et al. (2014) noted that data from the NSFG is subject to digit preference at 6, 12, 18, etc., months. They proposed a piecewise constant hazard model with knot selection designed to minimize the effect of digit preference. The cumulative hazard function for the piecewise constant hazard model states that \(H(t|\alpha, t) = \sum\{\{t_j < t\} - t_{j-1}\}\alpha_j\{\{t_j \leq t\} - t_{j-1}\}\), where \(t = (t_0, t_1, t_2, \ldots, t_L)\) denote the knot locations with \(0 \equiv t_0 < t_1 < \ldots < t_L \equiv \infty\), \(\alpha = (\alpha_1, \ldots, \alpha_L)\) and \(\alpha_j\) is the hazard rate over \([t_{j-1}, t_j)\). The survival functions for \(T\), \(Y\), and \(Z\) are given by \(F(\cdot|\alpha_T, t_T)\), \(F(\cdot|\alpha_Y, t_Y)\), and \(F(\cdot|\alpha_Z, t_Z)\) where \(\bar{F}(t|\alpha, t) = \exp\{-H(t|\alpha, t)\}\).

The main computational challenges of the likelihood in (3.7) (which we mainly focus on) is computing the integrals in the numerator and in the later part of the denominator. Neither can be simplified outside of a full Gaussian assumption. A Taylor series approximation (up to fourth order) was attempted to simplify the \(E(Z|T)\) expectation but was found to be unsatisfactory. As a result, we considered two estimation strategies: i) a mixed quadrature approach with Legendre quadrature for the integral in the numerator and a mix of Legendre quadrature and adaptive quadrature in the denominator, and ii) full adaptive quadrature, which is much more computationally demanding. The accuracy of the likelihood approximation methods was compared to the \(Z \perp (T, Y)\) likelihood in (3.5) (where calculations are straightforward) for a variety of situations under the \(Z \perp (T, Y)\) assumption. This comparison showed that full adaptive quadrature was more accurate than mixed, particularly with skewed distributions. In our simulations, we use the mixed quadrature approach since it had satisfactory
results and markedly shorter computation time. In the data analysis, we calculated the likelihood using the slower, more accurate full adaptive quadrature. All code was written in the R environment (R Core Team, 2020) and is available in the Supplemental Material and online at https://github.com/alexmclain/CDSRC.

Identifiability is an important issue in competing risk models. There is a well-known identifiability issue with dependent competing risks (Tsiatis, 1975), where multiple marginal distributions and dependence specifications can result in equivalent models. As a result, the estimated marginal distributions are reliant on a correct specification of the dependence structure. As discussed in Section A of the Supplementary Materials, the two main assumptions discussed in the beginning of Section 3.1 are needed to identify the parameters. Briefly, assumption (a) implies that the probability a subject was included in the sample is proportional to their total survival time, which relates the desired variable $T$ to a length biased version of $T$ (this assumption can be relaxed if data on the start of events are available). Further, assumption (b) is used to control for the censoring and implies multiplicative censoring by a uniform random variable. These assumptions are vital to relating the observed data to the random variables of interest.

5. Simulation Studies. To test the properties of the proposed method, numerous simulation studies were run. The simulation studies were run by simulating data for $T$, $Y$ and $Z$ using distributions designed to closely resemble their underlying biological processes in the motivating example. The current duration data were generated using methods described in McLain et al. (2014). The data were simulated using via two distributional assumptions. The first used the assumed piecewise constant hazard model discussed in Section 4, where $t_T = \{2, 5, 9, 14\}$, $t_Y = \{6, 9, 12\}$, and $t_Z = \{2, 10\}$ with the following rates $\log(\alpha_T) = (-2.5, -2.375, -2.25, -2.125, -2)$, $\log(\alpha_Y) = (-3.5, -3.0, -2.0, -1.5)$ and $\log(\alpha_Z) = (-1, -2, -3)$. The second setting used a Weibull($k, \lambda$) distribution with mean $\lambda \Gamma(1 + 1/k)$ where $(k, \lambda)$ were set to $(3/4, 6)$, $(4, 10)$ and $(3/4, 2)$ for $T$, $Y$ and $Z$, respectively. The data were generated using a Gaussian Copula with the following pairs of $(\rho_{TY}, \rho_{TZ})$: $(0.2, 0.2), (-0.4, 0.4), (-0.6, 0.6), (-0.2, 0.6), (-0.2, 0.4)$ and $(-0.4, 0.6)$. Here, $\rho_{YZ} = \rho_{TY} \rho_{TZ}$ under the $Y \perp Z | T$ assumption. Additional simulations which considered a $t$ Copula and other dependence scenarios are presented in the Section C of the Supplementary Material (conclusions were similar to below). All simulations used $n = 1,000$ for 500 iterations.

The knot locations were set to the percentiles of the observed $X_i$, $Y_i$ and $Z_i$ data. For the piecewise constant model the number of knots was equal to the true number of knots, while for the Weibull setting each distribution used 8 knots. Starting values of the parameters were obtained using a standard current duration approach applied to the $X_i$ and $Z_i$ data for $\alpha_T$ and $\alpha_Z$, respectively, while $\alpha_Y$ was set such that it resulted in the percentiles observed from $Y_i$. We compare our model a ‘standard’ piecewise constant current duration model which analyzes the data without correction (knots and starting values chosen the same way). That is, the observed $X$ data are analyzed using McLain et al. (2014) ignoring the $Z$ information. We also tested other approaches where subjects that experience the intermediate event were adjusted using a covariate or dropped. For the data simulated here, these resulted in estimates with markedly worse properties. We discuss these standard approaches more thoroughly in Section A of the Supplemental Material.

In Tables 1 and 2, we present the bias of the average estimated survival function along with Monte Carlo standard error (SE) for select values of $t$. For the piecewise constant data, the proposed method has relatively small bias when estimating the survival function of $T$, the main parameter of interest, regardless of the dependence parameters. This suggests that the proposed method can estimate the distribution of $T$ with low bias and is robust to misspecification in the knot locations. The standard approach resulted in uniformly more bias.
Simulation results of the proposed procedure along with a standard approach for data simulated with a piecewise constant hazard and \( n = 1,000 \). Displayed is the average bias (BIAS) and Monte Carlo standard error (SE) of \( \hat{F}_T(t) \) (estimated for both approaches), along with \( \hat{F}_Y(t) \), \( \hat{F}_Z(t) \) and \( \rho_{XY}, \rho_{XZ} \) (estimated with the proposed method only).

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \rho_{TY} = -0.2 )</th>
<th>( \rho_{TY} = -0.4 )</th>
<th>( \rho_{TY} = -0.6 )</th>
<th>( \rho_{TY} = -0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho_{TZ} = 0.2 )</td>
<td>( \rho_{TZ} = 0.4 )</td>
<td>( \rho_{TZ} = 0.6 )</td>
<td>( \rho_{TZ} = 0.6 )</td>
</tr>
<tr>
<td>True</td>
<td>BIAS(SE)</td>
<td>BIAS(SE)</td>
<td>BIAS(SE)</td>
<td>BIAS(SE)</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>3</td>
<td>0.773</td>
<td>-0.036 (0.081)</td>
<td>-0.028 (0.084)</td>
<td>-0.039 (0.079)</td>
</tr>
<tr>
<td>6</td>
<td>0.578</td>
<td>-0.053 (0.059)</td>
<td>-0.054 (0.065)</td>
<td>-0.054 (0.058)</td>
</tr>
<tr>
<td>12</td>
<td>0.294</td>
<td>-0.102 (0.021)</td>
<td>-0.076 (0.025)</td>
<td>-0.083 (0.022)</td>
</tr>
<tr>
<td>24</td>
<td>0.060</td>
<td>0.013 (0.008)</td>
<td>0.037 (0.011)</td>
<td>0.028 (0.009)</td>
</tr>
</tbody>
</table>

Standard Method

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \hat{F}_T(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.773</td>
</tr>
<tr>
<td>6</td>
<td>0.578</td>
</tr>
<tr>
<td>12</td>
<td>0.294</td>
</tr>
<tr>
<td>24</td>
<td>0.060</td>
</tr>
</tbody>
</table>

Proposed Method

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \hat{F}_Y(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.913</td>
</tr>
<tr>
<td>6</td>
<td>0.834</td>
</tr>
<tr>
<td>12</td>
<td>0.479</td>
</tr>
<tr>
<td>24</td>
<td>0.033</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \hat{F}_Z(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.419</td>
</tr>
<tr>
<td>6</td>
<td>0.279</td>
</tr>
<tr>
<td>12</td>
<td>0.147</td>
</tr>
<tr>
<td>24</td>
<td>0.081</td>
</tr>
</tbody>
</table>

\( \rho_{TY} \) 0.008 (0.200) 0.056 (0.159) 0.029 (0.171) 0.046 (0.138) 0.009 (0.218) 0.096 (0.241) -0.023 (0.283) -0.030 (0.301)
Simulation results of the proposed method and the standard approach for data simulated under a Weibull hazard. Displayed is the average bias (BIAS) and Monte Carlo standard error (SE) of all estimated distributions, $\rho_{TY}$ and $\rho_{TZ}$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>True</th>
<th>$F_T(t)$</th>
<th>BIAS(SE)</th>
<th>$F_T(t)$</th>
<th>BIAS(SE)</th>
<th>$F_Y(t)$</th>
<th>BIAS(SE)</th>
<th>$F_Z(t)$</th>
<th>BIAS(SE)</th>
<th>$\rho_{TY}$</th>
<th>$\rho_{TZ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.552</td>
<td>0.046 (0.065)</td>
<td>0.039 (0.066)</td>
<td>0.047 (0.068)</td>
<td>0.045 (0.065)</td>
<td>0.992</td>
<td>0.029 (0.011)</td>
<td>-0.024 (0.01)</td>
<td>-0.016 (0.007)</td>
<td>-0.033 (0.012)</td>
<td>0.062 (0.019)</td>
</tr>
<tr>
<td>6</td>
<td>0.368</td>
<td>0.004 (0.039)</td>
<td>0.009 (0.042)</td>
<td>0.013 (0.042)</td>
<td>0.018 (0.042)</td>
<td>0.878</td>
<td>-0.012 (0.035)</td>
<td>-0.018 (0.035)</td>
<td>-0.016 (0.031)</td>
<td>-0.023 (0.038)</td>
<td>-0.113 (0.055)</td>
</tr>
<tr>
<td>12</td>
<td>0.186</td>
<td>-0.094 (0.01)</td>
<td>-0.087 (0.01)</td>
<td>-0.082 (0.01)</td>
<td>-0.076 (0.011)</td>
<td>0.126</td>
<td>-0.017 (0.042)</td>
<td>-0.026 (0.039)</td>
<td>-0.010 (0.051)</td>
<td>-0.032 (0.036)</td>
<td>-0.001 (0.004)</td>
</tr>
<tr>
<td>24</td>
<td>0.059</td>
<td>-0.055 (0.001)</td>
<td>-0.053 (0.002)</td>
<td>-0.053 (0.002)</td>
<td>-0.052 (0.002)</td>
<td>0.000</td>
<td>0.000 (0.001)</td>
<td>0.000 (0.001)</td>
<td>0.002 (0.008)</td>
<td>0.000 (0.001)</td>
<td>0.002 (0.001)</td>
</tr>
</tbody>
</table>

$\rho_{TY}$ and $\rho_{TZ}$ were largely the same as for the Gaussian Copula, but with biases that were slightly magnified.

In our motivating example, the distributions of $Y$ and $Z$ along with the correlations are nuisance parameters. If these parameters were more of interest, it appears that more computationally intensive methods than the numerical approximations discussed in Section 4 should be considered.

6. Data Analysis. The NSFG is a publicly available nationally representative survey conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. The NSFG gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and reproductive health on reproductive-aged men and women in the United States (Chandra et al., 2005). For this analysis, we pooled all five cycles of the NSFG that contain sufficient information with dates ranging from 2002–2017. In total, 1244 women reported currently attempting pregnancy and gave their current duration of pregnancy attempt. Of these women, 382 (30.7%) reported receiving medical help to get pregnant within their current pregnancy attempt. The type of medical help was categorized as: ‘advice only,’ ‘infertility testing,’ ‘drugs to improve ovulation,’ ‘surgery to correct blocked tubes,’ ‘artificial insemination,’ or ‘other types of medical help.’ In this analysis, our definition of fertility
treatment excludes ‘advice only’ and ‘infertility testing,’ which left 270 women (21%) who reported the use of fertility treatments in their current attempt.

Assumption (a) discussed in Section 3.1, is satisfied when there is no trend in the start of pregnancy attempts. Slama et al. (2012) completed sensitivity analysis that showed robustness to this assumption despite the likely seasonality in pregnancy attempts. Assumption (b) requires that the distributions be independent over the sampling frame. This may be doubtful due to changes in fertility treatment practices over the period of study (impacting Y and Z) or possible changes in fecundity (Smarr et al., 2017). In the Supplemental Materials, we present analyses on the types of fertility treatment by NSFG cycle and an analysis of the observed TTFT data by NSFG data collection cycle. Both of these analyses failed to see change by NFSG cycle, which is in agreement with assumption (b). Here, we use this important assumption as a first approximation. Tests for verifying this assumption in practice and development of robust methods is an area of current research. Further, our results focus on a model using the $Y \perp Z|T$ assumption which implies that TTFT and TTPFT are independent given TTNP. While there are no studies of these latent processes to rely on, this assumption is motivated by TTPFT depending on a couple’s biological ability to reproduce, which is best measured by their TTNP. Further, the couples TTFT would be based on a physician’s prediction (based on age, lab tests, etc.) of the couples biological ability to reproduce. Thus, $T$ is a main predictor for both $Y$ and $Z$. There may, however, be unobserved confounders (e.g., socio-economic status) that create some residual dependence between $Y$ and $Z$ given $T$. As a result, we ran a sensitivity analyses with a general dependence structure and found the point estimates of the survival of $T$ to be little changed.

The knots for the piecewise model were chosen using the guidelines proposed in McLain et al. (2014). Briefly, to control for the digit preference in the data the knots are chosen such that they are outside the points of digit preference and all intervals include values with hypothesized positive and negative biases. The points of digit preference $\{6, 12, 18, 24, 36, \ldots\}$ correspond to year and half-year values. Within these possibilities the estimates of the survival function of $T$ were similar. Digit preference was not an issue for $Y$ or $Z$ since they are calculated from dates and not self-reported. For these variables, the knots were selected using their empirical distribution functions where probable distributional changepoints were identified. For our analysis, the knots for $T$, $Y$ and $Z$ were chosen to be $(2, 7, 11, 16)$, $(1, 7)$ and $(2, 4, 37)$, respectively. All confidence intervals were calculated based on 500 bootstrap replications.

We compare the proposed approach with the ‘standard’ current duration survival model used in Section 5, which ignores the infertility treatment information (discussed more in Section A of the supplemental materials). By not incorporating the fertility treatment information in the model, the standard approach estimates the distribution of time-to-pregnancy that could be fertility treatment aided, which we refer to as TTP.

In Figure 2a, we present estimates and 95% pointwise confidence intervals of the survival functions of TTP and TTNP. Additional results and figures for TTFT, TTPFT, the minimum of TTNP and TTFT $(T \land Y)$ and estimates of model parameters are given in the Supplemental Materials. The Figure shows the estimated survival function of TTNP with the proposed method given the current fertility treatment policy and effectiveness, along with our modeling assumptions on the dependent latent variables. It appears to be stochastically greater than the estimated survival function of TTP with the standard approach, which is expected since pregnancies with fertility treatment are included in the TTP curve. The estimated difference between the TTNP and TTP survival functions at 12 and 24 months are 0.06 (95% CI -0.005–0.12) and 0.09 (95% CI 0.03–0.16), respectively. Further, the differences in the TTNP and TTP distributions became more pronounced in the later months. For example, the cumulative probability at 2 years given survival past 12 months, i.e., $P(T \leq 24|T \geq 12)$ was
0.47 (95% CI 0.34–0.59) and 0.66 (95% CI 0.61–0.71) for TTNP and TTP, respectively; similarly \( \Pr(T \leq 36 | T \geq 12) \) was estimated to be 0.48 (95% CI 0.33–0.62) and 0.74 (95% CI 0.71–0.78) for TTNP and TTP, respectively. The flat portions of the survival curves are likely due to the digit preference in the data discussed in more detail in Section A of the Supplemental Materials. The differences in the proposed and standard methods appear to be due to the effectiveness of fertility treatment. The proportion of women not pregnant 3 months after fertility treatment began was 0.34 (95% CI 0.02–0.65), however, after 3 months of fertility treatment its impact appeared to diminish (see Figure 3 in the Supplemental Material).

Figure 2b shows the estimated survival curve of TTNP by parity. This figure shows that nulliparous women (those that have not given birth to a fetus with a gestational age of 24 weeks or more) appear to get pregnant more slowly than parous women, though there is increased uncertainty in the estimates versus the curve with all women.

7. Discussion. In this article, we proposed new methodology to estimate the survival function from cross-sectional length biased data with semi-competing risks. We derived the likelihood of the latent survival distributions using theory on semi-competing risks in length biased survival analysis and cross-sectional length-biased data. The semi-competing risks framework allows for the estimation of the main variable of interest while correcting for the impact of the intermittent event. In our motivating example, the latent variable treatment is advantageous since the theoretic outcome of interest is the length of pregnancy attempts if fertility treatments do not exist. This outcome is desirable, since it most closely represents the biologic capacity to reproduce and has real-world applications. For TTFT and TTPFT, the interpretation provided by cumulative incidence rates (see Duron et al., 2013) is likely preferable to that provided by the latent variable treatment. For example, the cumulative incidence rate estimates the proportion of couples that undergo fertility treatment by time \( t \), while the proposed method can estimate the proportion of couples that undergo fertility treatment by time \( t \) when none can get pregnant.

Our analysis produced findings not available with previous methods. This includes differences in survival probabilities at 12 or 24 months and the rate of events after 1 year, which is
a common time when couples will start seeking treatment. For example, the proposed method estimates that 48% (95% CI 33%–62%) of women still attempting at 1 year will get pregnant before 3 years if fertility treatments were not an option, while the standard approach estimates that 74% (95% CI 71%–78%) of women still attempting at 1 year will get pregnant before 3 years. This results in a substantial difference in the proportion of late pregnancies and quantifies how important medical fertility treatment is for couples still attempting at 1 year. Further, we found a moderate to strong negative dependence between TTNP and TTFT, indicating those with longer (shorter) TTNP have a shorter (longer) TTFT. This is likely due to missing explanatory variables related to fertility and fertility treatment. For example, many conditions such as endometriosis, pelvic inflammatory disease, age over 35, or previous treatment of cancer can lengthen TTP, and couples with such conditions are likely to seek fertility treatment earlier. The dependence between TTNP and TTFT was small, suggesting that this could be ignored for this data.

The methods developed here update the current duration approach to account for important aspects of some data applications; however, there are still issues that require further consideration. The estimates of the proposed methodology require parametric assumptions for each of the marginal distributions and on the dependence of the latent variables, which cannot be verified. As a result, they could be sensitive to the current fertility treatment policy and effectiveness of fertility treatment. Further, while proposed method accounts for the association between TTNP and TTFT biases could arise if the probability of treatment depends on the unobserved TTNP outcome. Exploration of these issues is best done under the causal inference framework, which could be used to estimate the causal survival function of TTNP (i.e., the counterfactual outcome where the time-varying treatment never occurs) and the causal treatment effect. These quantities have yet to be estimated and are of interest to the fields of fecundity and infertility treatment effectiveness. Various aspects of this problem have been studied for time-independent treatment’s and length-biased sampling (Cheng and Wang, 2012; Ertefaie, Asgharian and Stephens, 2014; Cheng and Wang, 2015) or with time-dependent treatments under observational study (Hernán et al., 2005; Li, Schaubel and He, 2014; He et al., 2020; Simoneau et al., 2020). A main challenge here is that the current data are cross-sectional and fertility treatment is a time-dependent exposure; extending existing methods to account these aspects is a challenge. For example, sequential ignorability (Robins, 2000), which extends the no unmeasured confounder assumption to the setting where treatment is time-dependent, may be hard to justify with the NSFG data. The validity of adapting prognostic risk matching methods for observational data to the cross-sectional setting needs future investigation. We will continue exploring potential methods for accounting for a time-dependent treatment in the cross-sectional study in the NSFG data in future work. As discussed in Section 2, the current duration approach cannot differentiate between pregnancy attempts that end in pregnancy or ‘unsuccessful’ attempts those that end for other reasons (e.g., giving up). With the proposed approach unsuccessful pregnancy attempts will only impact our parameter of interest (the survival function of TTNP) if the couple ends their attempt without trying fertility treatment. For standard methods all ‘unsuccessful’ pregnancy attempts will impact the survival function of TTP (Keiding et al., 2002). Nevertheless, this issue could result in TTNP survival probabilities with negative bias when the rate attempts of ending without a successful pregnancy or fertility treatment is non-zero. For current duration results to be fully understood more research on the rate and reasons for people stopping their pregnancy attempts is needed.

Our proposed method is the first attempt to extend the current duration approach to the semi-competing risk framework. The NSFG provides a rich data source with information on the timing of events, which is not typically available in most data sources on fertility and infertility patterns. In addition, NSFG is collected on an ongoing basis to examine biologic
CROSS-SECTIONAL LENGTH-BIASED SEMI-COMPETING RISKS DATA

13

fecundity patterns with future rounds of data. Considering further extensions such as including covariate data within the semi-competing risks framework are needed as large racial, ethnic, and geographic disparities persist in risk factors for infertility and for individuals who seek access to diagnostic and treatment services (Centers for Disease Control and Prevention, 2014; Farley Ordovensky Staniec and Webb, 2007). As demonstrated in the data analyses and simulations, appropriately accounting for semi-competing risks decreases the bias of survival function estimates and provides further information on the rate of fertility treatments and their success.

Acknowledgements. The authors would like to thank the Editor, Associate Editor and two reviewers for helpful comments that improved the manuscript. This research was supported through the Eunice Kennedy Shriver National Institute of Child Health and Human Development grant number 1R03HD097287-01.

SUPPLEMENTARY MATERIAL

Supplement Material to “Length-biased semi-competing risks models for cross-sectional data: an application to current duration of pregnancy attempt data” (doi: TBC; .pdf). This supplement provides additional discussion on the statistical assumptions, detailed arguments to form the likelihood, additional results for the simulation studies, and additional results and parameter estimates from the data analysis.

Code: Functions and sample data (doi: TBC; .zip). In this material we provide R code (R Core Team, 2020) with functions to implement the proposed methods. This information is also available online at https://github.com/alexmclain/CDSCR.

REFERENCES


