Joint modeling of survival and backwards recurrence outcomes: an analysis of factors associated with fertility treatment

Siyuan Guo¹, Jiajia Zhang¹, and Alexander C. McLain^{1,*}

¹Department of Epidemiology and Biostatistics, University of South Carolina

*Corresponding author

Abstract

The increase in methods focused on various types of survival outcomes has allowed practitioners to analyze data that are difficult or expensive to prospectively observe. Still, there are populations that are challenging to study. For example, obtaining a representative sample of couples attempting to become pregnant is difficult due to the dynamic nature of the population. This has led to an increase in the use of cross-sectional designs yielding backwards recurrent survival outcomes. In this paper, we consider the analysis of a survival outcome where subjects are observed if they are at-risk for a separate dependent survival outcome. The motivation for this problem is to determine which factors are associated with time-to-fertility-treatment (TTFT) among women currently attempting pregnancy in a cross-sectional sample. We propose appending a marginal accelerated failure time (AFT) model on TTFT with a conditional model on time-to-pregnancy (TTP) given TTFT to account for their dependence and avoid biases. We address challenges that arise due to the censoring of TTFT and the resulting increased computational complexity. The performance is validated via comprehensive simulation studies. We apply our approach to data from the National Survey of Family Growth to estimate the association insurance type has on TTFT, and estimate the impact of fertility treatment on TTP.

1 Introduction

Cross-sectional studies have the benefits of being low cost, time effective, and are particularly useful when studying cohorts which are difficult to sample such as dynamic populations (Vandenbroucke and Pearce, 2012). Examples of dynamic populations include couples who are attempting pregnancy (Scheike and Keiding, 2006; McLain et al., 2014), hospitalized patients after surgery (Mandel and Rinott, 2014; Mandel, 2015), and dementia patients (Bergeron et al., 2008; Carone et al., 2014). The defining feature of dynamic populations is that the length of time in the population, which we denote by T , is random. Commonly, it is of interest to study other outcomes within the dynamic population, which we denote by S . For example, we may be interested in the time-to-fertility-treatment (TTFT) among couples who are attempting pregnancy, biomarkers among hospital patients, or the time to needing in-home care among dementia patients. When T and S are dependent, this dependence needs to be accounted for to avoid biases.

There have been many different approaches proposed to analyze bivariate survival data. These include shared frailty models (Oakes, 1989; Liu et al., 2004; Hanagal et al., 2017), copula models (Genest and Rivest, 1993; Shih and Louis, 1995; Zheng and Klein, 1995; Oakes and Ritz, 2000), or semi-competing risk approaches (Fine et al., 2001; Mandel and Fluss, 2009; Mandel, 2010; McLain et al., 2021). In this paper we take an approach that is similar to the selection model framework commonly used in longitudinal analyses where the data are missing not at random (Rubin, 1976; Little and Rubin, 2019). Within this framework, the selection model consists of a marginal model for the outcome of interest and a missingness model conditional on the outcome of interest. The benefit of the selection model framework is that the model for the outcome of interest is marginal, neither the missingness pattern or a shared random effect needs to be specified for interpretation of coefficients.

Our motivating example is to study risk factors for prolonged TTFT among couples who are attempting pregnancy. Infertility is recognized as a disease of the reproductive system by the World Health Organization along with the American Society for Reproductive Medicine and is a disability under the Americans with Disabilities Act (Zegers-Hochschild et al., 2009;

Practice Committee of the American Society for Reproductive Medicine, 2013). The right to procreate is considered a fundamental human right, yet 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 Ordoventsky Staniec and Webb, 2007). Couples attempting pregnancy is a challenging population to study since it varies over time as couples get pregnant and new couples begin their pregnancy attempts. Prospective cohort studies are difficult as they have to identify and follow couples planning to start a pregnancy attempt in the future (Scheike and Keiding, 2006; Keiding et al., 2012), and retrospective studies will sample only from people who got pregnant and thus their results suffer from lack of generalizability. Alternatively, cross-sectional studies sample a large group of individuals at high-risk for attempting pregnancy and result in a sample from the population of interest.

The benefit of using a cross-sectional study is that it results in a sample from a hard to reach group. The challenge is that the outcomes are length-biased and entirely right censored. Backwards recurrent survival methods, proposed by Cox (1969), relate the full time-to-event to the cross-sectional current duration, and been further explained in van Es et al. (2000). Keiding et al. (2002) used both parametric and nonparametric methods to estimate the survival function from current duration data. McLain et al. (2014) adjusted for time-independent covariates in the model and potential digit-preference in the outcomes. McLain et al. (2021) proposed a semi-competing risk model with a Gaussian copula of TTP and TTFT. However, their focus was on the survival function of TTP, they did not estimate the impact of the fertility treatment on the TTP or asses risk factors for prolonged TTFT.

In this paper, we propose a discrete backward recurrent survival model that incorporates time-varying covariates with time-varying coefficients. In the grouped/discrete survival framework, the time-varying coefficient model has been considered (Kauermann et al., 2005; Wong et al., 2011). Specifically, Kauermann et al. (2005) proposed a discrete survival model with time-varying coefficients and Wong et al. (2011) extended the grouped proportional hazards model with time-varying coefficient. However, there is still a gap in incorporating a time-varying covariate with a time-varying coefficient. Specifically, the length of time the time-

varying covariate impacts survival may be unknown. For example, in our motivating data it is not known how long fertility treatments will impact the probability of pregnancy. As a result, we use a parsimonious form for the time-varying coefficient where the length of the impact is an estimated parameter. Further, we append this model with a marginal Accelerated Failure Time (AFT) model on TTFT (Buckley and James, 1979; Wei, 1992). This is akin to the selection model framework: a marginal model outcome of interest (here TTFT) is included with a conditional model for the missingness given the outcome of interest (here the above model for TTP given TTFT). Our main goal is to estimate risk factors for prolonged TTFT with a secondary goal of estimating the impact of the fertility treatment on TTP.

The paper is organized with the following outline. In Sections 2, we will introduce the notation, the general form of model, specify our model for the time-dependent coefficient and covariate. In Section 3, we will describe the estimation method and present results on how the computations can be simplified. We validated our model and compare it to competing approaches using comprehensive simulation studies Section 4. In Section 5, we demonstrate the proposed model on data from multiple cycles of the National Survey on Family Growth (NSFG) to determine risk factors for prolonged TTFT and the association between TTP and fertility treatment. In Section 6, we conclude with some discussion and propose future research directions.

2 Methods

In this Section, we develop the modeling approach that we'll use to estimate the impact of fertility treatment on the length of pregnancy attempts. In Section 2.1, we first discuss the main concepts needed for modeling current duration data and then introduce the discrete backwards recurrent Cox model which will serve as our baseline model. In Section 2.2, we incorporate time-varying covariates with time-varying effects. In Section 2.3, we give full specification of the time-varying coefficient and baseline hazard which we use in our experiments.

2.1 Notation and preliminaries

As described in the introduction, the conceptual data generation is that subjects begin their at-risk period at some calendar time σ and are sampled at time τ . For those at-risk at τ , we observed the current duration at risk denoted by $X = \tau - \sigma$. The goal of the estimation procedure is to provide distributional and inferential estimates for the total time at-risk denoted by T . For our motivating example, the goal is to estimate the distribution of TTP which is slightly different than T . Here, T corresponds to the total length of pregnancy attempt as pregnancy attempts can end for reasons other than the couple getting pregnant (e.g., giving up). It is important to note that all distributional properties are made on this T variable.

The difficulties in using the current duration study design are that X is length-biased (those with longer at-risk times are more likely to be at-risk) and entirely right-censored. The relationship between X and T is made by using two critical assumptions: (i) we assume that the probability a subject was included in the sample is proportional to their total at-risk time, and (ii) σ occurs according to a homogeneous Poisson process. Under these assumptions, the mass function of X , denoted as g , has form

$$g(x) = \frac{\bar{F}(x)}{\mu_T}, \quad (2.1)$$

(Cox, 1969), where $\bar{F}$ is the survival function of T , i.e. $\bar{F}(t) = 1 - P(T \leq t)$, and $\mu_T = E(T)$ which we assume to be finite. Note that (2.1) implies $g(0) = 1/\mu_T$ and $g(x) = g(0)\bar{F}(x)$. In our motivation data, the observe data are discrete so we focus throughout on this situation.

To incorporate the impact of baseline covariates $\mathbf{Z} = (Z_1, Z_2, \dots, Z_P)$ on T we assume a discrete proportional hazards model where the hazard probability of T is $P(T = j | T \geq j) = 1 - \exp[-\alpha_j \exp\{\beta^T \mathbf{Z}\}]$, where $\alpha_j \geq 0$ and $\alpha_0 \equiv 0$. As a result, the survivor function of T takes the form

$$\bar{F}(t | \mathbf{Z}) = \exp \left[- \sum_{j=0}^t \alpha_j \exp\{\beta^T \mathbf{Z}\} \right]. \quad (2.2)$$

Using equation (2.1), the mass function of X is given by

$$g(t|\mathbf{Z}) = g(0|\mathbf{Z}) \exp \left[- \sum_{j=0}^t \alpha_j \exp\{\beta^T \mathbf{Z}\} \right] \quad (2.3)$$

where $g(0|\mathbf{Z}) = (\sum_{t=0}^{\infty} \bar{F}(t|\mathbf{Z}))^{-1}$. Notice $g(x|\mathbf{Z})$ is non-increasing in x with maximum value at $g(0|\mathbf{Z})$ and $g^{-1}(0|\mathbf{Z}) = E(T|\mathbf{Z})$ is the mean of T given the covariate vector $\mathbf{Z}$. For discrete T , β_j can be interpreted as the approximate logarithms of subject-specific risk or odds ratios associated with a one unit increase in Z_j (Ecochard and Clayton, 2000).

2.2 Incorporating time-varying variables and coefficients

Now we consider a random time-varying treatment which may be observed at any time point in the current at-risk period. To this end, let S denote the first treatment time which may be censored by X . The observed data is $S^* = \min(X, S)$ and $\delta = I(S \leq X)$ where $I(\cdot)$ is the indicator function. We make the usual independent censoring assumption, which for our motivating example amounts to assuming that the survey time (τ) is independent of the treatment time (S). Further, let $\mathcal{Y}_\delta$ denote the observed treatment history given δ . Here, $\mathcal{Y}_1 = \{\mathbf{Y}_{S^*}, S^*, \delta = 1\}$ where $\mathbf{Y}_S = \{Y_j; Y_j = I(j \geq S), j = 0, 1, \dots\}$ and $Y_j = 1$ if the treatment occurred at or before time point j while $\mathcal{Y}_0 = \{(Y_0, Y_1, \dots, Y_{S^*}), S^*, \delta = 0\}$.

We estimate the impact of the treatment on T while accounting for the following factors: (i) the length of the impact is unknown and to be estimated from the model, and (ii) the treatment has no impact before it occurs. That is, we allow the impact to vary over time and extend (2.3) to include this situation. More specifically, given the treatment history $\mathcal{Y}_\delta$, let $\beta_j(S)$ denote the effect of the time-dependent treatment at time j when the treatment started at time S . The survival function T given S is modeled by

$$\bar{F}(t|\mathbf{Z}, \mathbf{Y}_S, S) = \exp \left[- \sum_{j=0}^t \alpha_j \exp\{\beta^T \mathbf{Z} + Y_j \beta_j(S)\} \right]. \quad (2.4)$$

The corresponding mass function of X given the treatment history $\mathcal{Y}_\delta$ is

$$g(t|\mathbf{Z}, \mathcal{Y}_\delta) = g(0|\mathbf{Z}, \mathcal{Y}_\delta) \exp \left[- \sum_{j=0}^t \alpha_j \exp \{ \beta_j^T \mathbf{Z} + Y_j \beta_j(S^*) \} \right]. \quad (2.5)$$

Note that we condition (2.4) on $(\mathbf{Y}_S, S)$ (i.e., the full treatment history) for notational convenience only, as it's computation only requires the history up to time t . A more accurate representation would condition on the history up to time t , e.g., $\mathcal{Y}_\delta(t)$, but this notation is more cumbersome for our purposes here. For (2.5), we condition on the observed treatment history where, if $\delta_i = 0$ (2.5) can only be used for $t = 1, \dots, X_i$ as the treatment information is unknown for $t > X_i$ (evaluating (2.5) for $t > X_i$ is not required in the likelihood calculation). We will use the notation $g(0|\mathbf{Z}, \mathcal{Y}_1) \equiv g\{0|\mathbf{Z}, (\mathbf{Y}_S, S)\}$ as needed to denote (2.5) for a particular value of S .

As in (2.3), $g(0|\mathbf{Z}, \mathcal{Y}_\delta)$ is an expectation which requires the full treatment history and its calculation depends on δ . For $\delta = 1$, the full treatment history is known by the monotonicity of the y_j 's and we have $g(0|\mathbf{Z}, \mathcal{Y}_1)^{-1} = \sum_{j=0}^{\infty} \bar{F}(j|\mathbf{Z}, \mathbf{Y}_S, S)$. For $\delta = 0$, (2.4) cannot be evaluated for $t > X$ and thus

$$\begin{aligned} g(0|\mathbf{Z}, \mathcal{Y}_0) &= E_S [g\{0|\mathbf{Z}, (\mathbf{y}_S, S)\} | \mathbf{Z}, S > S^*] \\ &= \sum_{s=S^*+1}^{\infty} g\{0|\mathbf{Z}, (\mathbf{y}_s, s)\} f_S(s|S^*), \end{aligned} \quad (2.6)$$

where $f_S(\cdot|S^*)$ denotes the probability mass function of S given $S > S^*$. In the following Section, we discuss the specification of the distribution of S and adding covariates to increase the precision in the expectation in (2.6).

2.3 Model Specification

The above form for the model is in terms of the time-varying coefficient $\beta_j(\cdot)$, the distribution of S , and the baseline hazard probability α . For our motivating example, we quantify the impact of fertility treatment by assuming a parsimonious form for $\beta_j(\cdot)$. We assume that fertility treatment has a constant impact on T for an estimated period of time, then has a

partial impact, and then has no impact. That is, we assume the time-dependent function has the form

$$\beta_j(S; \boldsymbol{\theta}) = \begin{cases} \theta_1 & S \leq j \leq S + \lfloor \theta_2 \rfloor \\ \theta_1(\theta_2 - \lfloor \theta_2 \rfloor) & j = S + \lfloor \theta_2 \rfloor + 1 \\ 0 & \text{otherwise,} \end{cases} \quad (2.7)$$

where $\lfloor \theta_2 \rfloor$ denotes the floor of the $\theta_2 \geq 0$ and $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$. Here, $\lfloor \theta_2 \rfloor + 2$ represents the treatment impact length, where the first $\lfloor \theta_2 \rfloor + 1$ time units have the full impact (θ_1) followed by one time unit with a partial impact ($\theta_1(\theta_2 - \lfloor \theta_2 \rfloor)$). For $\theta_2 = 0$, the treatment effect only occurs for one time unit. While the above form is parsimonious, it does have considerable benefits and was chosen based on our motivating data. For example, it can be optimized as a continuous function of the impact θ_1 and the length of the impact θ_2 , making it more computationally friendly than profile analytic approaches. Further, as will be discussed in Section 3, the infinite summation in (2.6) can be simplified using the properties of geometric series. Estimating the length of the treatment impact is beneficial for our motivating data as the length of time that fertility treatment has an impact is unknown.

To evaluate the likelihood the expectation in (2.6) needs to be evaluated which requires a form for the distribution of S . The expectation incorporates the uncertainty S for those where S is not observed. It is possible, however, to decrease the uncertainty in the distribution of S using available data on covariates that may be related to S . By incorporating covariate data into the distribution of S the integral in (2.6) can become more concentrated and result in less variability in the parameter estimates. By incorporating the distribution of S into the likelihood, the impact of the covariate data will be determined by all subjects. Let $\tilde{S}$ denote the continuous time to treatment variable of which we can observe $S = \lfloor \tilde{S} \rfloor$. We model the distribution of $\tilde{S}$ with an Accelerated Failure Time (AFT) model of time-independent covariates $\mathbf{W} = (W_1, W_2, \dots, W_q)$, i.e.

$$\log(\tilde{S}) = \gamma_0 + \sum_{j=1}^q \gamma_j W_j + \sigma \omega \quad (2.8)$$

where $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, \dots, \gamma_q\}$ and ω is the error term with a pre-specified distribution. By

assuming different assumptions on the distribution of ω , such as standard normal, extreme value, log-gamma, the distribution of Y will can be changed to log-normal, Weibull and Gamma distributions, respectively. The probability mass function of S is then $f_S(s|\mathbf{W}) = \bar{F}_{\bar{S}}(s - 1|\mathbf{W}) - \bar{F}_{\bar{S}}(s|\mathbf{W})$ where $\bar{F}_{\bar{S}}(\cdot|\mathbf{W})$ is the survival function corresponding to (2.8).

There are multiple models for α , the baseline hazard probability, using either continuous or discrete hazard functions. In this paper, due to the existence of digit preference that respondents preferred to report their current duration of pregnancy attempt at months of 6, 12, 18, etc., we propose to use the piecewise constant model introduced in McLain et al. (2014). The model divides the timeline into disjoint partitions with knots $\mathbf{k} = (k_1, k_2, \dots, k_m)$ and coefficients $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_{m+1}^*)$ where α_j in (2.4) is equal α_l^* for all $j \in (k_{l-1}, k_l]$ where $k_0 = 0$ and $k_{m+1} = \infty$. Suppose k is a preferred digit, we will avoid knots of $(k-2, k-1, k, k+1)$ and with this setting, the estimation bias brought from digit preference can be weaken. Further, the knots should be chosen such that observed S occur in multiple intervals. Without digit preference, knots can be selected using percentiles of the observations, equal spacing between boundaries, or other approaches. AIC can be used to compare the performance of the estimation with different numbers and positions of knots.

3 Estimation

Using the notation of the previous section, let $\mathcal{D}_i = (X_i, \mathbf{Z}_i, \mathbf{W}_i, \delta_i, \mathcal{Y}_{i\delta_i})$ denote the observed data for subject i for $i = 1, \dots, n$. The parameters to be estimated are $\Theta = (\alpha, \beta, \theta, \gamma, \sigma)$, where $\alpha = \{\alpha_1, \dots, \alpha_{m+1}\}$ for a given knot set $\mathbf{k}$. The likelihood corresponding to the observed data is given by $L(\Theta|\mathbf{D}) = \prod_{i=1}^n L(\Theta|\mathcal{D}_i)$ where $\mathbf{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$ and

$$L(\Theta|\mathcal{D}_i) = \{g(X_i|\mathbf{Z}_i, \mathcal{Y}_{i1})f_S(S_i^*|\mathbf{W}_i)\}^{\delta_i} \{\bar{F}(X_i|\mathbf{Z}_i, \mathcal{Y}_{i0})\bar{F}_{\bar{S}}(S_i^*|\mathbf{W}_i)g(0|\mathbf{Z}, \mathcal{Y}_{i0})\}^{1-\delta_i} \quad (3.9)$$

which requires the evaluation of infinite summations. For $\delta = 1$, the infinite summation can be simplified by noting that under the piecewise constant model the terms in the summation of (2.4) are constant for $j > J_i$ where $J_i = \max(S_i^* + \lfloor \theta_2 \rfloor + 1, k_m)$. In this case, the geometric

series can be used to show that

$$g(0|\mathbf{Z}_i, \mathcal{Y}_{i1})^{-1} = \sum_{t=0}^{J_i} \bar{F}(t|\mathbf{Z}_i, \mathcal{Y}_{i1}) + \bar{F}(J_i|\mathbf{Z}_i, \mathcal{Y}_{i1}) \frac{\exp\{-\alpha_{m+1} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)\}}{1 - \exp\{-\alpha_{m+1} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)\}}, \quad (3.10)$$

where $g(0|\mathbf{Z}_i, \mathcal{Y}_{i1}) = g\{0|\mathbf{Z}, (\mathbf{Y}_{iS_i}, S_i)\}$. Calculating $g(0|\mathbf{Z}, \mathcal{Y}_0) = E_S[g\{0|\mathbf{Z}, (\mathbf{y}_S, S)\} | S > S^*]$ via (2.6) requires evaluating $g\{0|\mathbf{Z}, (\mathbf{Y}_{is}, s)\}$ for all $s > S^*$. This is a computationally demanding task as J changes with the value of s . A tedious but straightforward calculation can be used to show that

$$g\{0|\mathbf{Z}_i, (\mathbf{y}_{s+k}, s+k)\} = g\{0|\mathbf{Z}_i, (\mathbf{y}_s, s)\} + D_i(s)\{1 - \exp(-k\alpha_{m+1} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i))\} \quad (3.11)$$

for $s > k_m$ where $D_i(s)$ is a function of $\mathbf{Z}_i$, α_{m+1} , $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (see the supplementary material for details on this derivation). As a result, $g\{0|\mathbf{Z}, (\mathbf{y}_s, s)\}$ can be calculated via (3.10) for $s = S^* + 1, \dots, k_m + 1$, then (3.11) can be used for the remaining terms greatly simplifying the required calculations for (3.9). See the supplemental material for more details on how the estimation approach is implemented.

Theoretically, length of the treatment affect θ_2 can be any non-negative number, but in practice setting an upper bound for θ_2 is advantageous. For example, we can only realistically expect to be able to identify θ_2 up to a certain point (e.g., the maximum observed survival time). In our simulations and data analyses, use a data-driven upper bound threshold $\theta_2^* = \max(x_i - s_i)$, where if $\theta_2 > \theta_2^*$ (2.4) and (2.5) are simplified by setting $\beta_j(S; \boldsymbol{\theta}) = \theta_1 I(j \geq S)$.

4 Simulation Study

To test the performance of the proposed model and estimation procedure, numerous simulation studies were performed. The details and motivation for how the current duration X is obtained are given in McLain et al. (2014). Briefly, for each subject and iteration of the simulation, values of T were generated until their cumulative sum is greater than a large number τ (e.g., $\tau = 10^5$) and X is the difference between τ and the largest cumulative sum smaller than τ and S was observed if $S \leq X$. Recall $\tilde{S}$ denotes the unobserved continuous time to treatment

Figure 1: For the Weibull S with 40% observed setting (a) average estimated baseline survival function of T (solid line) with the true curves (dotted line), and (b) average (black solid) and median (gray solid) estimated $\beta_j(\cdot; \theta)$ with the true value (dotted line). Both figures contain the upper and lower 5% quantiles (dashed lines).

variable, and $S = \lfloor \tilde{S} \rfloor$ is observed data used in the model fitting. The distribution of $\tilde{S}$ was the AFT model in (2.8), where $\gamma_0^* = e^{\gamma_0}$, σ , and the distribution of ω were varied. Specifically, we used a log-normal ω with $\gamma_0^* = 8$, $\sigma = 0.5$ and $\gamma_0^* = 3.5$, $\sigma = 0.5$, along with a Weibull ω with $\gamma_0^* = 10$, $\sigma = 0.25$ and $\gamma_0^* = 3.5$, $\sigma = 0.25$. The settings with larger and smaller γ_0^* resulted in approximately 20% and 40% of S being observed, respectively (i.e., $E(\delta) \approx 0.2$ or 0.4). The AFT model of S and the survival function of T have shared two covariates, i.e., $W_1 = Z_1 \sim \text{Bernoulli}(0.5)$ and $W_2 = Z_2 \sim N(0, 1)$. The regression coefficients are given in Table 1. The impact of S on T , $\beta_j(S; \theta)$, was given by (2.7) where $\theta_1 = 1$ and $\theta_2 = 4$. Given S , T was generated using (2.4). Sample size $n = 1500$ and 1000 iterations were used on all the simulations. To fit the proposed model, 5 and 8 knots for the log-normal and Weibull scenarios were used, respectively. The knots were chosen using the percentiles of the $X|\delta = 1$. We compare our results to a standard AFT model on S , where S was right censored at X if $S > X$.

Table 1 presents the mean, median, and standard deviation of regression parameters for the proposed and standard AFT models. For the proposed model, the parameters for the

Table 1: Simulation results for the proposed and standard AFT models. The distribution of $\tilde{S}$ is generated from a log-normal for (A) and (B) and a Weibull for (C) and (D) where the percentage of observed S is approximately 20% for (A) and (C), and 40% for (B) and (D).

	(A)				(B)			
	TRUE	MEAN	MEDIAN	SD	TRUE	MEAN	MEDIAN	SD
Proposed model								
γ_0^*	8.0	8.002	7.98	0.402	3.5	3.503	3.499	0.111
γ_1	0.5	0.497	0.498	0.051	0.5	0.500	0.499	0.038
γ_2	1.0	0.998	0.998	0.027	1.0	1.001	1.00	0.022
σ	0.5	0.498	0.497	0.020	0.5	0.500	0.499	0.015
β_1	-1.0	-0.977	-0.975	0.078	-1.0	-0.988	-0.986	0.070
β_2	-0.5	-0.473	-0.473	0.050	-0.5	-0.484	-0.483	0.040
Standard AFT model								
γ_0^*	8.0	9.62	9.61	0.430	3.5	4.28	4.27	0.118
γ_1	0.5	0.398	0.398	0.049	0.5	0.413	0.413	0.034
γ_2	1.0	0.938	0.938	0.024	1.0	0.870	0.870	0.018
σ	0.5	0.480	0.480	0.019	0.5	0.452	0.452	0.013
	(C)				(D)			
Proposed model								
γ_0^*	10.0	9.98	9.98	0.312	3.5	3.50	3.50	0.063
γ_1	0.5	0.501	0.502	0.032	0.5	0.500	0.501	0.023
γ_2	1.0	1.00	1.00	0.020	1.0	0.999	0.999	0.014
σ	0.25	0.249	0.249	0.011	0.25	0.249	0.248	0.009
β_1	-1.0	-1.009	-1.006	0.081	-1.0	-1.009	-1.008	0.073
β_2	-0.25	-0.246	-0.245	0.042	-0.25	-0.246	-0.248	0.038
Standard AFT model								
γ_0^*	10.0	11.1	11.1	0.358	3.5	4.36	4.36	0.130
γ_1	0.5	0.427	0.428	0.037	0.5	0.385	0.385	0.038
γ_2	1.0	0.914	0.916	0.027	1.0	0.787	0.789	0.026
σ	0.25	1.27	1.27	0.016	0.25	0.277	0.275	0.022

distribution of $\tilde{S}$ (i.e., $\gamma_0^*, \gamma_1, \gamma_2, \sigma$) all have low bias. For the standard AFT model, there is markedly high bias in the estimates of γ_0^* or σ in each setting. The percent absolute bias of the regression parameters (γ_1, γ_2) range from 0.0%-0.6% with the proposed approach and 6.2%-23% with the standard approach. The estimates of the regression parameters of T (β_1, β_2) were relatively unbiased.

Figure 1 shows the estimated and true baseline survival curves of T and $\beta_j(\cdot; \theta)$ with point-wise 90% inner quantiles for the Weibull with 20% of S observed setting (see Section B of the Supplementary Material for other settings which were similar). Figure 1a shows the survival

of T was estimated with little bias. The average and median values of $\beta_j(\cdot; \theta)$ have low bias for $j \leq 4$. For the remaining points the median estimate has less bias than the mean. The median estimate from $j = 5$ to 6 has a large jump, while the average estimate decreases more slowly resulting in negative bias for $j = 5$ and decreasing positive bias for $j > 5$. To have average estimates that more closely resemble the true value it appears that a larger sample size is required.

5 Real data analysis

We applied the proposed model to the NSFG data which is a public survey data conducted by CDC. We integrated data from seven cycles between 2002 and 2019, which contains 1442 women that reported a current duration of pregnancy attempt. Among these women, 399 (27.6%) had undergone some type of fertility treatment. Here, we aim to determine the impact of insurance type, categorized as private (which includes military or state-sponsored health plans), medicaid/medicare, or none on TTFT. We also adjust for age (categorized < 35 , $35 - 40$, and > 40), and parity (yes/no). The TTP model adjusts for fertility treatment, age and parity. A log-normal assumption of S was chosen based on model fit criteria.

Digit preference was observed in the data such that respondents were more likely to report their current duration at the time of 6, 12, 18, 24, 36 months, etc, to minimize their impact on the estimated survival curves we chose $\mathbf{k} = (3, 8, 9, 14, 27)$. Further, we set $\theta_2^{ast} = 12$ as the threshold for the impact of fertility treatment (i.e., we treat $\theta_2 > 12$ as a continual impact). To estimate the standard error and confidence intervals (CIs) 500 bootstrap iterations were completed. We compare our results to a standard AFT model on S , where if $S > X$ the data were right censored at X .

Table 2 provides estimates and 95% CIs for the variables in the proposed and standard TTFT model. The proposed model found that the impact of parity, age 35-40, age over 40, medicaid, or no insurance lead to longer TTFT than the reference group (nulliparous, age less than 35 with private insurance). Specifically, having medicaid was associated with a time ratio (TR) of 2.39 (95% CI 1.70, 3.53) versus having private insurance while being over 40 was

Table 2: NSFG data analysis results for TTFT with the proposed and standard AFT model. Displayed is the estimated coefficients and percentile based 95% bootstrap confidence intervals.

	Proposed model			Standard model		
	EST	SE	95%CI	EST	SE	95%CI
TTFT results						
INTERCEPT	3.05	0.10	(2.85, 3.25)	3.30	0.09	(3.12, 3.48)
PARITY	0.23	0.11	(0.004, 0.43)	0.38	0.11	(0.15, 0.60)
AGE						
< 35			(REFERENT)			(REFERENT)
35 – 40	0.33	0.16	(0.001, 0.65)	0.24	0.14	(-0.04, 0.52)
> 40	0.84	0.17	(0.55, 1.18)	0.68	0.17	(0.35, 1.02)
INSURANCE						
PRIVATE			(REFERENT)			(REFERENT)
NONE	0.72	0.18	(0.37, 1.09)	0.86	0.18	(0.51, 1.22)
MEDICAID	0.87	0.19	(0.53, 1.26)	1.06	0.20	(0.67, 1.45)
σ	1.44	0.04	(1.36, 1.52)	1.51	0.05	(1.40, 1.61)

associated with a TR of 2.31 (95% CI 1.73, 3.25) versus being under 35. The difference in the estimated probability mass functions between the women with private insurance and medicaid is displayed in Figure 2a for nulliparous women under 35 years old, where for women with private insurance the median occurs at 19 months while for women with medicaid the median occurs at 44 months, over 2 years later.

Generally the proposed and standard models agree on which predictor variables impact TTFT, but there are some differences in the point estimates. The impact of parity and insurance is attenuated in the proposed model versus the standard approach, while the impact of age is larger with the proposed model. For example, the TR for parity is 1.25 and 1.46 for the proposed and standard models, respectively. The estimated survival functions of TTFT for the proposed and standard model can be seen in Figure 2b for parous women under 35 years of age. For this group, the estimated median TTFT was 23 and 34 months for the proposed and standard models, respectively. We tested adding race/ethnicity to the presented model and found that it did not have an impact (data not shown).

For the TTP outcome, we found that being age 35-40 ($\hat{\beta} = -0.39$ 95% CI $-0.52, -0.27$) or over 40 ($\hat{\beta} = -0.79$ 95% CI $-0.96, -0.61$) were associated with longer TTP versus under 35. Further, parous women had higher odds of getting pregnant ($\hat{\beta} = 0.31$ 95% CI $0.22, 0.42$)

Figure 2: Estimated probability mass functions for TTFT (a) with 95% pointwise confidence intervals (thin dotted lines) for nulliparous women under 35 years of age with private (solid line) and medicaid (dashed line) insurance, and (b) for parous women under 35 years of age with the proposed model (solid line) and the standard AFT model (dashed line).

than nulliparous women. In Figure 3a, we display the estimated survival functions with 95% pointwise confidence intervals for parous women under 35 years old and nulliparous women over 40 years old (the groups with the highest and lowest odds of getting pregnant). Not including the impact of fertility treatment, the estimated median TTP for parous women under 35 years old was 10 months while for nulliparous women over 40 years old it was 28 months.

In Figure 3b, we show the estimated impact of fertility treatment by month with 95% pointwise confidence intervals. The estimated impact of infertility treatment is calculated as the average value of $\hat{\beta}_j(\cdot; \hat{\theta})$ over the 500 bootstrap iterations for each j . The impact of fertility treatment at months 1 and 2 is clear, where the estimated coefficient is $\hat{\beta}_1(\cdot; \hat{\theta}) = 1.32$ (95% CI 0.62, 2.07) for month 1 with a similar value for month 2, however, it appears to quickly diminish. For example, at month 3 the estimated coefficient is $\hat{\beta}_3(\cdot; \hat{\theta}) = 0.878$ (95% CI 0.00, 1.77) and at month 4 it is $\hat{\beta}_4(\cdot; \hat{\theta}) = 0.59$ (95% CI 0.00, 1.56). The lower bound of the CI at month 3 and after are zero. In fact, the median bootstrap $\hat{\beta}_j(\cdot; \hat{\theta})$ was zero for months 5 and after.

Figure 3: (a) Estimated survival functions for TTP with 95% pointwise confidence intervals for parous women under 35 years old (solid line) and nulliparous women over 40 years old (dashed line). (b) The estimated impact of fertility treatment $\beta_j(\cdot; \theta)$ by month with 95% pointwise confidence intervals.

6 Discussion

Outcomes from dynamic populations are known to be difficult to study, arising from the random departures and entries into the set of interest. In this paper we proposed methods to analyze survival outcomes from subjects in a dynamic population, where the time in the dynamic population and the outcome of interest are dependent. We consider cross-sectional data, without follow-up, which are a cost-effective and fast method of sampling from such populations. The observed time in the population, is an example of a backwards recurrent survival outcome which require particular assumptions and methods for estimation. In this paper, we extended previous backwards recurrent methods to incorporate a time-varying covariate with a time-varying coefficient. This model was estimated jointly with an AFT model on the outcome of interest.

In the motivation data, the interest was to identify and quantify the affect of risk factors for prolonged TTFT among women attempting pregnancy and estimate the impact of fertility treatment on TTP. Risk factors associated with use of fertility treatments has been studied

with similar data (see Chandra and Stephen, 2010; Greil et al., 2011; Chandra et al., 2014), however, all previous methods suffer from two major weaknesses. First, these studies use imprecise definitions of the population of interest. Chandra and Stephen (2010) and Greil et al. (2011) only consider women with fertility problems, which comes with the uncertainty of how to define “fertility problems” and how to handle couples that got pregnant quickly with fertility treatment. Further, Chandra et al. (2014) was a retrospective study of all women of reproductive age, which includes people in the that could have been in the “attempting couples” risk set decades before. Second, these studies all rely on a binary 0/1 indicator of fertility treatment. As a result, they cannot determine and quantify differences TTFT. Duron et al. (2013) used competing risks survival analysis methods to estimate the cumulative incidence of medical consultation for fecundity problems using prevalent cohort data, i.e., the subdistribution of fertility treatment and not getting pregnant. This method is different than the proposed method in the study design (prevalent cohort designs have follow-up while cross-sectional does not) and the use of competing risks methods.

In our analysis, we found that older age, being parous, and having non-private insurance lead to longer TTFT. Particularly, our model found that being over 40 years-of-age had a large impact TTFT with an TR of 2.31 (95% CI 1.73, 3.25) versus being less than 35 years-of-age. Further, having medicaid was associated with a TR of 2.39 (95% CI 1.70, 3.53) versus having private insurance. These indicate that the median TTFT for those over 40 or having medicaid was over 2.3 times larger than the referent groups. For nulliparous women under 35 years-of-age, the median TTFT occurs at 19 months with private insurance and at 44 months, over 2 years later, for women with medicaid. These types of findings were not available from previous studies that only considered a yes/no indicator of fertility treatment. After adjusting for these factors, we found that race/ethnicity did not have a significant impact on TTFT (this was also found by Chandra and Stephen, 2010). Our model also found a significant effect of fertility treatment on TTP for the first two months after treatment. However, the impact of treatment quickly diminished after month two and over half of the bootstrap estimates of the treatment impact length were less than 5 months.

While cross-sectional data are a key element to studying hard to reach populations, they

do come with the drawback that the outcomes commonly come from self-recall. Such data are known to have measurement error, which has vast statistical literature. However, measurement error in the outcome is something that has garnered much less attention. Recently, Oh et al. (2021) proposed a regression calibration for this setting (see also Oh et al., 2018). In the survival setting, measurement error could increase the mean which would have an impact on the current setting since the mean is required in even the most simple models. The development of methods to correct for measurement error in current duration data is a fruitful area of future research.

Acknowledgements

This research was supported through the Eunice Kennedy Shriver National Institute of Child Health and Human Development grant number 1R03HD097287-01.

Bibliography

- Bergeron, P.-J., M. Asgharian, and D. B. Wolfson (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association* 103(482), 737–742.
- Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika* 66(3), pp. 429–436.
- Carone, M., M. Asgharian, and N. P. Jewell (2014). Estimating the lifetime risk of dementia in the Canadian elderly population using cross-sectional cohort survival data. *Journal of the American Statistical Association* 109(505), 24–35.
- Centers for Disease Control and Prevention (2014, June). National public health action plan for the detection, prevention, and management of infertility, atlanta, georgia: Centers for disease control and prevention. <http://www.cdc.gov/reproductivehealth/infertility/publichealth.htm>.

- Chandra, A., C. E. Copen, and E. H. Stephen (2014). *Infertility service use in the United States: data from the National Survey of Family Growth, 1982-2010*. Number 73 in National health statistics reports. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
- Chandra, A. and E. H. Stephen (2010). Infertility service use among us women: 1995 and 2002. *Fertility and sterility* 93(3), 725–736.
- Cox, D. R. (1969). Some sampling problems in technology. In N. L. Johnson and H. Smith (Eds.), *New Developments in Survey Sampling*, pp. 506–527. Wiley, New York.
- Duron, S., R. Slama, B. Ducot, A. Bohet, D. N. Sørensen, N. Keiding, C. Moreau, and J. Bouyer (2013). Cumulative incidence rate of medical consultation for fecundity problems—analysis of a prevalent cohort using competing risks. *Human Reproduction* 28(10), 2872–2879.
- Ecochard, R. and D. G. Clayton (2000). Multivariate parametric random effect regression models for fecundability studies. *Biometrics* 56, 1023–1029.
- Farley Ordozensky Staniec, J. and N. J. Webb (2007). Utilization of infertility services: how much does money matter? *Health services research* 42(3p1), 971–989.
- Fine, J. P., H. Jiang, and R. Chappell (2001). On semi-competing risks data. *Biometrika* 88(4), 907–919.
- Genest, C. and L.-P. Rivest (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association* 88(423), 1034–1043.
- Greil, A. L., J. McQuillan, K. M. Shreffler, K. M. Johnson, and K. S. Slauson-Blevins (2011). Race-ethnicity and medical services for infertility: Stratified reproduction in a population-based sample of us women. *Journal of health and social behavior* 52(4), 493–509.
- Hanagal, D. D., A. Pandey, and A. Ganguly (2017). Correlated gamma frailty models for bivariate survival data. *Communications in Statistics-Simulation and Computation* 46(5), 3627–3644.

- Kauermann, G., G. Tutz, and J. Brüderl (2005). The survival of newly founded firms: A case-study into varying-coefficient models. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 168(1), 145–158.
- Keiding, N., O. K. Højbjerg Hansen, D. N. Sørensen, and R. Slama (2012). The current duration approach to estimating time to pregnancy. *Scandinavian Journal of Statistics* 39, 185–204.
- Keiding, N., K. Kvist, H. Hartvig, M. Tvede, and S. Juul (2002). Estimating time to pregnancy from current durations in a cross-sectional sample. *Biostatistics* 3, 565–578.
- Little, R. J. and D. B. Rubin (2019). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.
- Liu, L., R. A. Wolfe, and X. Huang (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* 60(3), 747–756.
- Mandel, M. (2010). The competing risks illness–death model under cross-sectional sampling. *Biostatistics* 11, 290–303.
- Mandel, M. (2015). Analyzing multiple cross-sectional samples with application to hospitalization time after surgeries. *Statistics in medicine* 34(26), 3415–3423.
- Mandel, M. and R. Fluss (2009). Nonparametric estimation of the probability of illness in the illness-death model under cross-sectional sampling. *Biometrika* 96(4), 861–872.
- Mandel, M. and Y. Rinott (2014). Estimation from cross-sectional samples under bias and dependence. *Biometrika* 101(3), 719–725.
- McLain, A. C., S. Guo, M. Thoma, and J. Zhang (2021). Length-biased semicompeting risks models for cross-sectional data: An application to current duration of pregnancy attempt data. *The Annals of Applied Statistics* 15(2), 1054–1067.
- McLain, A. C., R. Sundaram, M. Thoma, and G. M. Buck Louis (2014). Semiparametric modeling of grouped current duration data with preferential reporting. *Statistics in medicine* 33(23), 3961–3972.

- Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84(406), 487–493.
- Oakes, D. and J. Ritz (2000). Regression in a bivariate copula model. *Biometrika* 87(2), 345–352.
- Oh, E. J., B. E. Shepherd, T. Lumley, and P. A. Shaw (2018). Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with simex. *Statistics in medicine* 37(8), 1276–1289.
- Oh, E. J., B. E. Shepherd, T. Lumley, and P. A. Shaw (2021). Raking and regression calibration: Methods to address bias from correlated covariate and time-to-event error. *Statistics in Medicine* 40(3), 631–649.
- Practice Committee of the American Society for Reproductive Medicine (2013). Definitions of infertility and recurrent pregnancy loss: a committee opinion. *Fertility and Sterility* 99(1), 63.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.
- Scheike, T. H. and N. Keiding (2006). Design and analysis of time-to-pregnancy. *Stat. Methods Med. Res.* 15, 127–140.
- Shih, J. H. and T. A. Louis (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 51(4), 1384–1399.
- van Es, B., C. A. Klaassen, and K. Oudshoorn (2000). Survival analysis under cross-sectional sampling: length bias and multiplicative censoring. *Journal of Statistical Planning and Inference* 91(2), 295 – 312.
- Vandenbroucke, J. P. and N. Pearce (2012). Incidence rates in dynamic populations. *International journal of epidemiology* 41(5), 1472–1479.
- Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statist. Med.* 11(14-15), 1871–1879.

- Wong, M. C. M., K. F. Lam, and E. C. M. Lo (2011, 02). Analysis of multilevel grouped survival data with time-varying regression coefficients. *Statistics in Medicine* 30, 250–259.
- Zegers-Hochschild, F., G. Adamson, J. de Mouzon, O. Ishihara, R. Mansour, K. Nygren, E. Sullivan, and S. Vanderpoel (2009). International committee for monitoring assisted reproductive technology (icmart) and the world health organization (WHO) revised glossary of ART terminology, 2009. *Fertility and Sterility* 92(5), 1520–1524.
- Zheng, M. and J. P. Klein (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 82(1), 127–138.

Supplementary Materials for “Joint modeling of survival and backwards recurrence outcomes: an analysis of factors associated with fertility treatment”

Siyuan Guo¹, Jiajia Zhang¹, and Alexander C. McLain^{1,*}

¹Department of Epidemiology and Biostatistics, University of South Carolina

*Corresponding author

Section A. Likelihood function derivation details.

In this Section, we’ll discuss useful forms which can be used to estimate the likelihood. Specifically, forms to estimate the likelihood for individuals with $\delta_i = 1$ which is by far the more computationally demanding portion of implementing the estimation procedure. Recall, for individuals with $\delta_i = 1$ the likelihood contribution is $\bar{F}(X_i|\mathbf{Z}_i, \mathcal{Y}_{i0})\bar{F}_{\tilde{S}}(S_i^*|\mathbf{W}_i)g(0|\mathbf{Z}, \mathcal{Y}_{i0})$, where

$$\begin{aligned}\bar{F}_S(S_i^*|\mathbf{W}_i)g(0|\mathbf{Z}, \mathcal{Y}_{i0}) &= \bar{F}_S(S_i^*|\mathbf{W}_i) \sum_{s=S^*+1}^{\infty} g\{0|\mathbf{Z}, (\mathbf{y}_s, s)\}f_S(s|S^*, \mathbf{W}_i) \\ &= \sum_{s=S^*+1}^{\infty} g\{0|\mathbf{Z}, (\mathbf{y}_s, s)\}f_S(s|\mathbf{W}_i)\end{aligned}\tag{0.1}$$

since $f_S(s|S^*, \mathbf{W}_i) = f_S(s)/\bar{F}_S(S_i^*|\mathbf{W}_i)$. As a result, the goal is to find computationally feasible methods to calculate $g\{0|\mathbf{Z}, (\mathbf{y}_s, s)\}$ and to approximate the tail behavior of the infinite summation in (0.1).

The form for $g\{0|\mathbf{Z}, (\mathbf{y}_{s+k}, s+k)\}$ given in Section 3 is used when S is not observed and the expectation of $g\{0|\mathbf{Z}, (\mathbf{Y}_S, S)\}$ over S . Here, $g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}$ for all $s \leq k_m$ must be evaluated individually. Then the form for $g\{0|\mathbf{Z}, (\mathbf{y}_{s+k}, s+k)\}$ is used for $s > k_m$ and $k > 0$. As a

result, below we assume throughout that $s > k_m$. First, we define

$$\begin{aligned} A_1 &= \sum_{j=0}^{k_m} \alpha_j \exp(\beta^T \mathbf{Z}) \\ A_2 &= \alpha_{m+1}^* \exp(\beta^T \mathbf{Z}) \\ A_3 &= \alpha_{m+1}^* \exp(\beta^T \mathbf{Z} + \theta_1) \\ A_4 &= \alpha_{m+1}^* \exp(\beta^T \mathbf{Z} + \theta_1(\theta_2 - \lfloor \theta_2 \rfloor)). \end{aligned}$$

Note that when $k_m \leq t < s$, $\bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) = \bar{F}(t|\mathbf{Z}, \mathbf{Y}_{s+1}, s+1) = e^{-A_1 - A_2(t - k_m)}$. Further, if $s \leq t \leq s + \lfloor \theta_2 \rfloor$

$$\begin{aligned} \bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) &= \bar{F}(s-1|\mathbf{Z}, \mathbf{Y}_s, s) e^{-A_3(t-s+1)} \\ \bar{F}(t|\mathbf{Z}, \mathbf{Y}_{s+1}, s+1) &= \bar{F}(s-1|\mathbf{Z}, \mathbf{Y}_s, s) e^{-A_2 - A_3(t-s)}, \end{aligned}$$

if $t = s + \lfloor \theta_2 \rfloor + 1$,

$$\begin{aligned} \bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) &= \bar{F}(s-1|\mathbf{Z}, \mathbf{Y}_s, s) e^{-A_3(\lfloor \theta_2 \rfloor + 1) - A_4} \\ \bar{F}(t|\mathbf{Z}, \mathbf{Y}_{s+1}, s+1) &= \bar{F}(s-1|\mathbf{Z}, \mathbf{Y}_s, s) e^{-A_2 - A_3(\lfloor \theta_2 \rfloor + 1)} \end{aligned}$$

and if $t > s + \lfloor \theta_2 \rfloor + 1$

$$\begin{aligned} \bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) &= \bar{F}(s-1|\mathbf{Z}, \mathbf{Y}_s, s) e^{-A_3(\lfloor \theta_2 \rfloor + 1) - A_4 - A_2\{t - (s + \lfloor \theta_2 \rfloor + 1)\}} \\ \bar{F}(t|\mathbf{Z}, \mathbf{Y}_{s+1}, s+1) &= \bar{F}(s-1|\mathbf{Z}, \mathbf{Y}_s, s) e^{-A_2 - A_3(\lfloor \theta_2 \rfloor + 1) - A_4 - A_2\{t - (s + \lfloor \theta_2 \rfloor + 2)\}}. \end{aligned}$$

As a result, letting $B_1 = e^{-A_2 + A_3}$ and $B_2 = e^{-A_2 + A_4}$, it is easy to derive

$$r(t|\mathbf{Z}, s+1) = \frac{\bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s+1)}{\bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s)} = \begin{cases} 1 & t < s \text{ or } t \geq s + \lfloor \theta_2 \rfloor + 2 \\ B_1 & s \leq t \leq s + \lfloor \theta_2 \rfloor \\ B_2 & t = s + \lfloor \theta_2 \rfloor + 1 \end{cases}. \quad (0.2)$$

Consequently, for $s > k_m$

$$\begin{aligned}
 g\{0|\mathbf{Z}, (\mathbf{y}_{s+1}, s+1)\}^{-1} &= \sum_{t=0}^{\infty} \bar{F}(t|\mathbf{Z}, Y_s, s) \times r(t|\mathbf{Z}, s+1) \\
 &= \sum_{t=0}^{s-1} \bar{F}(t|\mathbf{Z}, Y_s, s) + B_1 \sum_{t=s}^{s+\lfloor \theta_2 \rfloor} \bar{F}(t|\mathbf{Z}, Y_s, s) \\
 &\quad + B_2 \bar{F}(s + \lfloor \theta_2 \rfloor + 1|\mathbf{Z}, \mathbf{Y}_s, s) + \sum_{t=s+\lfloor \theta_2 \rfloor+2}^{\infty} \bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) \\
 &= g\{0|\mathbf{Z}, (\mathbf{y}_s, s)\}^{-1} - (1 - B_1) \sum_{t=s}^{s+\lfloor \theta_2 \rfloor} \bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) \\
 &\quad - (1 - B_2) \bar{F}(s + \lfloor \theta_2 \rfloor + 1|\mathbf{Z}, \mathbf{Y}_s, s).
 \end{aligned}$$

The forms for $\bar{F}$ given above can be used to show that

$$\begin{aligned}
 \sum_{t=s}^{s+\lfloor \theta_2 \rfloor} \bar{F}(t|\mathbf{Z}, \mathbf{Y}_s, s) &= e^{-sA_2} e^{-A_1+A_2(1+k_m)} \sum_{t=s}^{s+\lfloor \theta_2 \rfloor} e^{-A_3(t-s+1)} \\
 &= e^{-sA_2} e^{-A_1+A_2(1+k_m)} \sum_{t=0}^{\lfloor \theta_2 \rfloor} e^{-A_3(t+1)} \\
 &= e^{-sA_2} e^{-A_1+A_2(1+k_m)-A_3} \left(\frac{1 - e^{-[\theta_2+1]A_3}}{1 - e^{-A_3}} \right)
 \end{aligned}$$

and $\bar{F}(s + \lfloor \theta_2 \rfloor + 1|\mathbf{Z}, \mathbf{Y}_s, s) = e^{-sA_2} e^{-A_1+A_2(1+k_m)-A_3} e^{-A_3\lfloor \theta_2 \rfloor - A_4}$. As a result,

$$g\{0|\mathbf{Z}, (\mathbf{Y}_{s+1}, s+1)\}^{-1} = g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}^{-1} - D e^{-sA_2} \quad (0.3)$$

where

$$D = e^{-A_1+A_2(1+k_m)-A_3} \left[(1 - B_1) \frac{1 - e^{-A_3\lfloor \theta_2+1 \rfloor}}{1 - e^{-A_3}} + (1 - B_2) e^{-A_3\lfloor \theta_2 \rfloor - A_4} \right]$$

is constant in s . Iterating (0.3) can be used to show that

$$\begin{aligned} g\{0|\mathbf{Z}, (\mathbf{Y}_{s+1}, s+k)\}^{-1} &= g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}^{-1} - D \sum_{l=0}^{k-1} e^{-(s+l)A_2} \\ &= g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}^{-1} - D e^{-sA_2} \frac{1 - e^{-kA_2}}{1 - e^{-A_2}} \\ &= g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}^{-1} - D(s)(1 - e^{-kA_2}) \end{aligned}$$

or

$$g\{0|\mathbf{Z}, (\mathbf{Y}_{s+1}, s+k)\} = \frac{g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}}{1 - g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\}D(s)(1 - e^{-kA_2})}$$

where $D(s) = D \frac{e^{-sA_2}}{1 - e^{-A_2}}$, which matches the form stated in Section 3. Further, this can be used to get

$$\begin{aligned} \sum_{s=k_m+1}^{\infty} g\{0|\mathbf{Z}, (\mathbf{Y}_s, s)\} f_s(s|\mathbf{W}) &= g\{0|\mathbf{Z}, (\mathbf{Y}_{k_m+1}, k_m+1)\} \times \\ &\quad \sum_{s=k_m+1}^{\infty} f_S(s|\mathbf{W}) \times \\ &\quad [1 - g\{0|\mathbf{Z}, (\mathbf{Y}_{k_m+1}, k_m+1)\}D(k_m+1)(1 - e^{-sA_2})]^{-1}. \end{aligned}$$

The latter summation does not have a closed form, but can be approximated by truncating it at a large number τ . For $S^* > k_m$, this can be used exchanging $k_m + 1$ with $S^* + 1$. For $S^* \leq k_m$, the summation can be broken into $[S^* + 1, k_m]$ and $[k_m + 1, \infty)$ using the above formulation. Finally, we note that when $\theta_2 > \theta_2^*$,

$$D = e^{-A_1 + A_2(1+k_m) - A_3} \left(\frac{1 - B_1}{1 - e^{-A_3}} \right).$$

Section B. More simulation results

In this section, we will present the remaining simulation results.

Figure 1: (Left) average estimated baseline survival function of T (solid line) with the true curves (dotted line), and (right) average (black solid) and median (gray solid) estimated $\beta_j(\cdot; \theta)$ with the true value (dotted line) where all figures contain the upper and lower 5% quantiles (dashed lines). S is distributed as (top) log-normal with 20% observed, (middle) Weibull with 20% observed and (bottom) Weibull with 40% observed.