The effect of left-truncation on the coverage of confidence intervals for pregnancy outcome probabilities

Arthur Allignol[1], Franziska Königsbauer[1], and Markus Pauly[1]*

[1]Institute of Statistics, Ulm University, Germany.

August 2, 2018

Abstract

Estimation of pregnancy outcome probabilities from observational cohorts is complicated by the fact that the data are left-truncated as women usually become aware of their pregnancy several weeks after conception. Moreover, the data are subject to competing risks. Indeed a pregnancy may end in a live-birth, an induced abortion or a spontaneous abortion. Thus the use of survival techniques, in particular, the Aalen-Johansen estimator of the cumulative incidence function, is advocated to estimate pregnancy outcome probabilities. However, due to left-truncation, the number of individuals at risk might be small at the beginning of the time interval, possibly leading to unstable estimates propagating to the whole time span. It turned out in simulations that the usual asymptotic approximation for forming point-wise confidence intervals may become highly liberal with coverage probabilities as low as 60%. To this end, we investigate the use of more involved procedures for the construction of confidence intervals in simulation studies and provide insightful solutions. The results are applied to a study on the effect of coumarin derivatives' use during pregnancy on the risk of spontaneous abortion.

Keywords: Aalen-Johansen, Abortion, Bootstrap, Competing Risks, Delayed Entry.

* e-mail: markus.pauly@uni-ulm.de

1 Introduction

Prenatal development is the most vulnerable phase in human life. Drug toxicities but also insufficiently treated maternal conditions may result in life-long handicaps of the newborn. Besides birth defects, spontaneous abortions and miscarriages — usually defined in epidemiological studies as pregnancy losses up to 20 or 22 weeks of gestational age (Slama et al., 2014) — are the most frequent adverse pregnancy outcome and are estimated to occur in around 15% of all clinically recognized pregnancies (Borrell and Stergiotou, 2013). Lupattelli et al. (2014) estimate that 80% of pregnant women use at least one drug to treat an acute or chronic condition. The role of Teratology Information Services (TIS) is to advise pregnant women and policy makers about the risk of adverse drug reactions in pregnancy. This counselling aims at both reducing the rate of elective terminations based on irrationally overestimated drug risks and at proposing better and safer medical treatment in case of a maternal disease. Even if the percentage of preventable pregnancy losses might be low, better care can have a huge public health impact. As an example, if 15% of pregnancy losses were preventable, then 90,000 families per year in the United States would spare themselves such grief (Hogue, 2016). Additionally to their counselling role, TIS collect data of outstanding value. These are prospectively collected and, contrary to registry and prescription data, contain information on spontaneous abortion and precise timing of drug exposure (Grzeskowiak et al., 2012).

Pregnant women usually contact a TIS once the pregnancy is recognized and are followed until at least the planned delivery date. A consequence of this sampling is that the data are left-truncated. Indeed, the natural time-scale is time since conception (or last menstrual period) but the women enter the study at different times after conception. Thus simple binomial proportions are biased and survival techniques should be used (Meister and Schaefer, 2008; Allignol et al., 2010; Andersen et al., 2012; Slama et al., 2014). Besides, observations are subject to competing risks as pregnancy may end in a spontaneous abortion, induced abortion or a live-birth. Meister and Schaefer (2008) advocate the use of the cumulative incidence function estimated via the Aalen-Johansen estimator for quantifying, say, the probability of spontaneous abortion. However, and contrary to standard survival situations with only right-censoring, delayed study entries may lead to early random time intervals with small risk sets. Observed events during such intervals may lead to unstable estimates which propagate over the whole time span. Friedrich et al. (2017) proposed a stabilized Aalen-Johansen estimator for such situations but found out in simulations that some scenarios led to coverage probabilities as low as 60% using the usual asymptotic approximation. Alternatively, Efron (1979) suggested the bootstrap for constructing confidence intervals with improved small sample behaviour. Akritas (1986) and Lo and Singh (1986) generalized the classical bootstrap approach to simple survival models with right censoring. An alternative procedure described in Lin et al. (1993) is the Wild bootstrap(Beyersmann et al., 2013; Dobler and Pauly, 2014; Dobler et al., 2017). In this approach, random multipliers with expectation zero and variance one are used for calculating critical values for inference. In this paper, we analyse variants of both bootstrap procedures with respect to their ability to enhance the under coverage of the usual confidence intervals.

This article is structured as follows: In Section 2 we introduce the underlying competing risks model and the Aalen-Johansen estimator. Classical and modified bootstrap-based confidence intervals for the probabilities of interest are described in Section 3. Simulation results are outlined in Section 4. In Section 5 the presented bootstrap methods are applied to the real data set of a pregnancy outcome studySchaefer et al. (2006); Meister and Schaefer (2008). Finally, a discussion closes the paper in Section 6.

the results are discussed in Section 6.

2 Competing Risks Multi-State Model and Non-parametric Estimation

Consider a competing risk process $(X(t), t \geqslant 0)$ with a finite number of states $X(t) \in \{0, 1, \dots k\}$, $k \geqslant 2$. In this model besides the initial state 0, with $P(X(0) = 0) = 1$, there exists k absorbing states representing the competing risks. For ease of presentation however, we assume two competing states in the sequel. The event time T is defined as

$$T = \inf\{t > 0 \mid X(t) \neq 0\}. \tag{1}$$

The type of event is then given by $X(T) \in \{1, 2\}$ and the instantaneous risk of their occurrences are described by the cause specific hazards

$$\alpha_{0j}(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t, X(T) = j \mid T \geqslant t)}{\Delta t}, j = 1, 2, \tag{2}$$

which we assume to exist. We further assume that $\int_0^\tau \alpha_{0j}(u)du < \infty$, for a terminal time $\tau < \infty$.

Another important quantity in the competing risk setting is the cumulative incidence function (CIF), that describes the probability of a $0 \to j$ transition on the interval $[0, t]$:

$$P_{0j}(t) = P(T \leqslant t, X(T) = j) = \int_0^t P(T > u-)\alpha_{0j}(u)du, j = 1, 2, \tag{3}$$

for $t < \tau$, and where $P(T > t)$ is the survival probability.

Since pregnancy outcome study data are usually only subject to left truncation, we assume a left truncated model without right censoring as in Keiding and Gill (1990). The left-truncation times L are assumed to be independent of $(T, X(T))$.

left-truncation arises if study entry happens after time origin 0. In pregnancy studies, the time since conception (or last menstrual period) would be considered as the natural time scale, but most women become aware of their pregnancy some time after conception, women who experience an event before clinical detection never enter the study.

We now consider independent replications $(T_i, L_i, X_i(T_i)), 1 \leqslant i \leqslant n$, of the competing risks process corresponding to n individuals. For each event type $j \in \{1, 2\}$ we then define the individual event and at-risk processes as

$$N_{0j,i}(t) = \mathbf{1}(L_i < T_i \leqslant t, X_i(T_i) = j) \tag{4}$$

$$Y_i(t) = \mathbf{1}(L_i < t \leqslant T_i). \tag{5}$$

By summation of (4) and (5) over all individuals $1 \leqslant i \leqslant n$ we obtain $Y(t)$, the number of individuals at risk in the initial state just before t, and $N_{0j}(t)$ the number of observed type j events on the interval $[0, t]$. As described in Andersen et al. (1993) the counting process $N_{0j}(t)$ has intensity $Y(t)\alpha_{0j}(t)$ and

$$M_{0j}(t) = N_{0j}(t) - \int_0^t Y(u)\alpha_{0j}(u)du, j = 1, 2 \tag{6}$$

is a martingale. From this it follows that the Aalen-Johansen estimator of the cumulative incidence function

$$\hat{P}_{0j}(t) = \int_0^t \frac{\hat{P}(T > u-)dN_{0j}(u)}{Y(u)}, \quad j = 1, 2, \tag{7}$$

is asymptotically normal (Andersen et al., 1993). Here $\hat{P}(T > u)$ denotes the Kaplan-Meier estimator of the survival probability. The complementary weighted log-minus-log transformed $100(1 - \alpha)\%$-confidence intervals based on the central limit theorem were considered as in Beyersmann et al. (2012):

$$1 - (1 - \hat{P}_{0j}(t))^{exp\left(z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_{AJ}(t)}{(1-\hat{P}_{0j}(t))\cdot\log(1-\hat{P}_{0j}(t))}\right)}. \tag{8}$$

Here $\hat{\sigma}^2_{AJ}(t)$ is a consistent Greenwood-type variance estimator as defined in Equation (6) of Allignol et al. (2010) and $z_{1-\alpha}$ is the $(1 - \alpha)$-quantile of the standard normal distribution.

3 Bootstrap

It turned out that the asymptotic confidence intervals given in (8) may lead to extremely low coverage probabilities in case of left-truncation(Friedrich et al., 2017, and Section 5). Since this is clearly apparent in our pregnancy outcome study we investigated several resampling approaches and their potential to enhance the coverage.

3.1 Efron's Bootstrap

Efron's (1979, 1981) bootstrap method is based on drawing n times with replacement from the data. In our competing risks setting, data is given by the trajectories of the n independent competing risks processes $X_i, i = 1 \ldots, n$, or equivalently by the pairs $(L_i, T_i, X_i(T_i))$. Thus, denoting a bootstrap sample thereof as $(L_i^*, T_i^*, X_i^*(T_i^*)), i = 1, \ldots, n$, we can calculate bootstrap versions of the individual event and risk process given in (4) - (5), say $N^*_{0j,i}$ and Y_i^*, by replacing $(L_i, T_i, X_i(T_i))$ with their bootstrap counterpart. From this, bootstrap versions of the Kaplan-Meier, say $\hat{P}^*(T > u)$, and the Aalen-Johansen estimator

$$\hat{P}^*_{0j}(t) = \int_0^t \frac{\hat{P}^*(T > u-)dN^*_{0j}(u)}{Y^*(u)}, j = 1, 2 \tag{9}$$

are calculated. Roughly speaking, it then follows from theory on the bootstrap (Akritas, 1986; van der Vaart and Wellner, 1996; Gill and Johansen, 1990; Lo and Singh, 1986; Dobler, 2016) that the distributions of $W_n = \sqrt{n}(\hat{P}_{0j}(t) - P_{0j}(t))$ and $W_n^* = \sqrt{n}(\hat{P}^*_{0j}(t) - \hat{P}_{0j}(t))$ asymptotically coincide. From this several bootstrap confidence intervals can be constructed that are correct for large sample sizes. Here, the two most common approaches in applied statistics are either based on the percentile method or a bootstrap plug-in variance estimator. In both cases the bootstrap algorithm is repeated a large number, say B, of times to obtain bootstrapped Aalen-Johansen estimates $\hat{P}^{*,(b)}_{0j}(t), b = 1, \ldots, B$. Let $c^*(\alpha)$ be the $(1 - \alpha)$-quantile thereof, then $(c^*(\alpha/2), c^*(1 - \alpha/2))$ is the bootstrap percentile confidence interval for $P_{0j}(t)$. For the other we calculate the empirical variance of the bootstrapped Aalen-Johansen estimates as $\hat{\sigma}^{*2}_{AJ}(t) = (B - 1)^{-1} \sum_{b=1}^{B} (\hat{P}^{*,(b)}_{0j}(t) - \frac{1}{B} \sum_{c=1}^{B} \hat{P}^{*,(c)}_{0j}(t))^2$ and obtain a confidence interval based on bootstrapped variances by replacing $\hat{\sigma}_{AJ}(t)$ in (8) with $\hat{\sigma}^*_{AJ}(t)$.

Although these two bootstrap techniques are correct for $n \to \infty$ and usually improve the small sample performance of pure asymptotic methods, our simulation results from Section 4 did not show a clear advantage in case of heavy left-truncation. We therefore considered other bootstrap techniques that (at least heuristically) should mirror the Aalen-Johansen process much better. To motivate them recall that the construction of (8) is based on central limit theorems for the weighted and transformed Aalen-Johansen estimator $H_n = \sqrt{n}g(t)\{\Psi(\hat{P}_{0j}(t)) - \Psi(P_{0j}(t))\}$, where

$$g(t) = \log(1 - \hat{P}_{0j}(t)) \cdot (1 - \hat{P}_{0j}(0,t))/\hat{\sigma}_{AJ}(t) \quad \text{and} \quad \Psi(t) = \log(-\log(1-x)). \tag{10}$$

By the delta method, H_n is asymptotically equivalent to $\hat{H}_n = \sqrt{n}g(t)\Psi'(P_{0j}(t))W_n(t)$ and thus, asymptotically standard normal leading to the asymptotic confidence interval (8). Now, a bootstrap version of the latter is given by

$$\hat{H}_n^* = \sqrt{n}g^*(t)\Psi'(\hat{P}_{0j}(t))W_n^*(t), \tag{11}$$

where $g^*(t) = \log(1 - \hat{P}_{0j}^*(0,t)) \cdot (1 - \hat{P}_{0j}^*(0,t))/\hat{\sigma}_{AJ}^*(t)$. Denoting by $q_{1-\frac{\alpha}{2}}^*$ the (conditional) $(1-\frac{\alpha}{2})$-quantile of $\hat{H}_n^*$ a bootstrap version of the transformed confidence interval (8) of asymptotic level $(1-\alpha)$ is obtained by replacing $z_{1-\frac{\alpha}{2}}$ with $q_{1-\frac{\alpha}{2}}^*$ in (8).

Note, that we also considered a similar bootstrap approximation of H_n and even investigated versions of both with weight function g instead of g^* (results not shown here). All of them considerably improved the coverage probability of the previous approaches but the confidence intervals based on $\hat{H}_n^*$ performed the best and are solely reported for clarity.

3.2 Wild Bootstrap

Another common resampling strategy is given by the wild bootstrap. For right-censored competing risks data it was first proposed by Lin et al. (1993) and later generalized by other authors (Beyersmann et al., 2013; Dobler and Pauly, 2014; Dobler et al., 2017); also allowing for left-truncated observations. Due to its computational efficiency it is often the resampling method of choice for complex time to event situations. To introduce it in the present context recall (Andersen et al., 1993; Beyersmann et al., 2013), that the normalized Aalen-Johansen estimator W_n possesses the martingale representation

$$\begin{aligned} W_n(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \int_0^t \frac{S_2(u)}{\frac{1}{n}Y(u)} dM_{01;i}(u) + \int_0^t \frac{P_{01}(0,u)}{\frac{1}{n}Y(u)} dM_{02;i}(u) \right. \\ & \left. - P_{0j}(0,t) \int_0^t \frac{dM_{01;i}(u) + dM_{02;i}(u)}{\frac{1}{n}Y(u)} \right\} \end{aligned} \tag{12}$$

with $S_2(t) = 1 - P_{02}(t)$. Let $G_{0j;i}$, $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant 2$, be i.i.d. zero-mean random variables with variance one. Then a general wild bootstrap version of W_n is defined by replacing the unknown martingales $M_{0j,i}$ with randomly weighted counting processes $G_{0j,i}N_{0j,i}$:

$$\begin{aligned} W_n^\star(t) =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \int_0^t \frac{\hat{S}_2(u)G_{01;i}}{\frac{1}{n}Y(u)} dN_{01;i}(u) + \int_0^t \frac{\hat{P}_{01}(0,u)G_{02;i}}{\frac{1}{n}Y(u)} dN_{02;i}(u) \right. \\ & \left. - \hat{P}_{0j}(0,t) \int_0^t \frac{G_{01;i}dN_{01;i}(u) + G_{02;i}dN_{02;i}(u)}{\frac{1}{n}Y(u)} \right\}, \end{aligned} \tag{13}$$

where $\hat{S}_2(t) = 1 - \hat{P}_{02}(t)$. Writing $W_n^\star(t) = \sqrt{n}(\hat{P}_{0j}^\star(0,t) - \hat{P}_{0j}(0,t))$ as in the case of Efron's bootstrap, a wild bootstrap version of $\hat{P}_{0j}(0,t)$ is thus given by $\hat{P}_{0j}^\star(0,t) = n^{-1/2}W_n^\star(t) + \hat{P}_{0j}(0,t)$. Substituting Efron's bootstrap version $\hat{P}_{0j}^*(0,t)$ with $\hat{P}_{0j}^\star(0,t)$ in the confidence intervals from Section 3.1 we obtain various wild bootstrap confidence intervals. Again the best thereof is based on approximating H_n with $\hat{H}_n^\star(t) = \sqrt{n}g^\star(t)\Psi'(\hat{P}_{0j}(0,t))W_n^\star(t)$, i.e. on replacing $z_{1-\frac{\alpha}{2}}$ in (8) with the respective (conditional) $(1-\frac{\alpha}{2})$-quantile $q_{1-\frac{\alpha}{2}}^\star$ from $\hat{H}_n^\star$.

4 Simulation Study

In order to evaluate the performance of the different bootstrap methods, we conducted several simulation studies based on the pregnancy data set. As in Allignol et al. (2010) we generated competing risks data with only two competing states. We consider a linearly decreasing cause specific hazard function for the event of interest $\alpha_{01}(t) = -1.7 \cdot 10^{-4} \cdot t + 0.017$. For the competing event we used the Weibull-type cause specific hazard $\alpha_{02}(t) = 1.4/27^{1.4} \cdot t^{0.4}$. Both hazard functions are displayed in Figure 1.

The competing risks data were generated as described in Beyersmann et al. (2009), i.e,

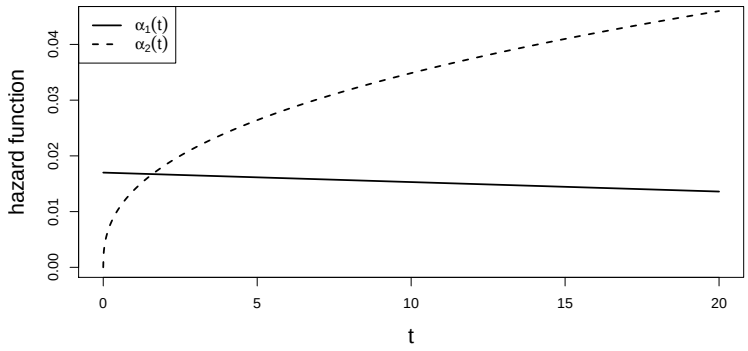

Figure 1: Hazard functions for scenarios 1 and 2.

1. The event times T are generated according to the all-cause hazard $\alpha_{0\cdot}(t) = \alpha_{01}(t) + \alpha_{02}(t)$ using the inversion method.

2. Given a survival time $T = t$, a binomial experiment is run, which decides with probability $\alpha_{01}(t)/\alpha_{0\cdot}(t)$ for an event of type 1.

3. Left-truncation times L are simulated independently of $(T, X(T))$.

In the first two scenarios, the left-truncation times follow a skewed normal distribution with density function(Azzalini, 1985)

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left[\frac{x-\xi}{\omega}\right]\right), \tag{14}$$

where (ξ, ω, α) are the location, scale and shape parameters, respectively. $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and the cumulative distribution function of the standard normal distribution, respectively. *Scenario 1* includes rather extreme scenarios, with $(\omega, \alpha) = (4.3, -8)$ and varying location parameter ($\xi \in \{8, 12, 16, 20\}$) to simulate different truncation probabilities. In Figure 2 the impact of the location parameter on the density function of the skewed normal distribution is shown. In

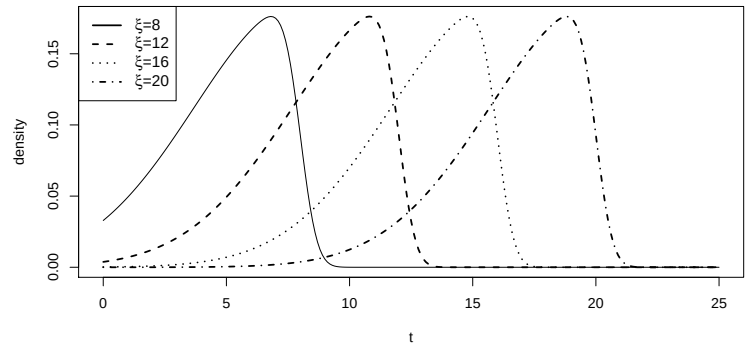

Figure 2: Density of skewed normal distribution with $(\omega, \alpha) = (4.3, -8)$ and different location parameters ξ.

particular, the likelihood that an individual starts in state 0 at time $t = 0$ is very low. *Scenario 2* is less extreme. The setting is the same as in Scenario 1, but the parameters of the truncation distribution are chosen such that approximately 10% of the individuals had a study entry at time point $t = 0$, i.e., $(\xi, \alpha) = (0, 1)$ and $\omega = \{6.9, 13.8, 23.3, 38.3\}$ (Figure 3). In *Scenario 3*, we considered a more simple set-up with exponentially distributed survival times and $\alpha_{01}(t) = 0.03$ as well as $\alpha_{02}(t) = 0.08$. The truncation times are also exponentially distributed with rate $\lambda \in \{0.6, 0.2728, 0.1547, 0.097\}$.

We run simulations for different sample sizes $m \in \{100, 200\}$. Since the results are similar we only show those for $m = 200$. Because of left-truncation, the number of individuals under observation is random with $n \leqslant m$, e.g., for $m = 200$ and $\xi = 16$ in scenario 1, there are on average $\bar{n} = 117$ individuals in the study. For each scenario, we calculated the coverage probabilities of the confidence intervals introduced in Sections 2 - 3 based on $1,000$ simulation runs each. In particular, we investigated the asymptotic method (8) based on the log-minus-log transformation (*Asympt*), the classical bootstrap methods based on the percentile method (*PercM*) and variance estimates (*VarBoot*), respectively,

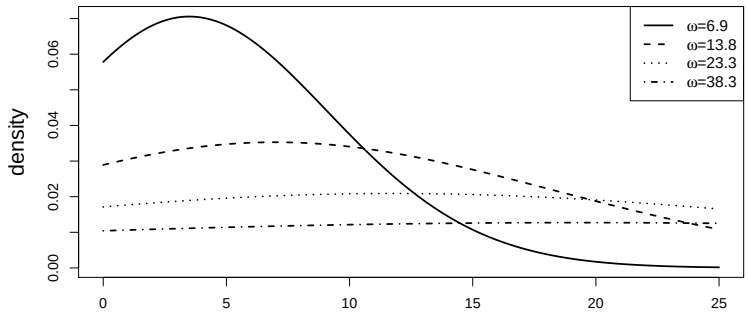

Figure 3: Density of skewed normal distribution with $(\xi, \alpha) = (0, 1)$ and different scale parameters ω.

Table 1: Coverage probabilities at time point $t = 20$ for different scenarios of underlying truncation distributions. *PercM:* Percentile method using Efron's bootstrap. *VarBoot:* Plug-in variance based on Efron's bootstrap. *TBoot:* modified Efron confidence intervals based on (11). *TWBoot:* Wild bootstrap confidence intervals.

	$m, \bar{n}$	Asympt.	PercM	VarBoot	TBoot	TWBoot
Scenario 1	$m = 200, \bar{n} = 169$	92.40	91.60	92.70	95.10	94.70
	$m = 200, \bar{n} = 143$	83.90	80.70	83.40	89.70	87.70
	$m = 200, \bar{n} = 117$	71.10	63.20	71.50	87.30	83.20
	$m = 200, \bar{n} = 94$	62.00	49.00	63.20	95.00	93.20
Scenario 2	$m = 200, \bar{n} = 169$	94.10	93.30	94.20	95.20	95.60
	$m = 200, \bar{n} = 142$	92.30	90.70	92.40	95.50	95.50
	$m = 200, \bar{n} = 117$	92.50	91.80	92.60	95.30	95.50
	$m = 200, \bar{n} = 93$	94.10	91.50	94.00	96.50	95.70
Scenario 3	$m = 200, \bar{n} = 169$	94.10	94.30	94.10	94.50	94.30
	$m = 200, \bar{n} = 143$	94.60	94.20	95.10	96.10	95.80
	$m = 200, \bar{n} = 117$	93.20	92.70	93.80	94.70	94.60
	$m = 200, \bar{n} = 94$	94.60	94.00	94.60	95.90	95.40

as well the more involved transformed bootstrap based confidence intervals based on (11) (*TBoot*) and (13) (*TWBoot*). For the latter we focus on standardized Poisson-$(Poi(1) - 1)$ wild bootstrap weights since they showed a slightly better performance than the usual standard normal multipliers (results not shown here).

Since we were interested in the confidence interval for spontaneous abortion, we focus on the coverage probability at time $t = 20$. The results are presented in Table 1. For Scenario 1 it is readily seen that the first three methods (*Asympt, PercM, VarBoot*) are highly liberal with coverage probabilities as low as 49% in the most extreme scenario. In particular, the classical bootstrap based on the percentile method (*PercM*) results in slightly lower values than the asymptotic confidence interval. This may be caused by the fact that in contrast to the asymptotic confidence intervals, the intervals were neither transformed nor weighted. The intervals based on bootstrapped variances (*VarBoot*) yield comparable coverage probabilities to that of the asymptotic confidence intervals. On the contrary, the modified bootstrap (*TBoot*) and Wild bootstrap confidence intervals (*TWBoot*) both result in higher coverage probabilities, that are close to the nominal level.

Scenario 2 — that allows for around 10% of the individuals to enter the study at time 0 — leads to higher coverage probabilities for all confidence intervals and the nominal 95% level is much better approximated here. Again the modified bootstrap based confidence intervals (*TBoot* and *TWBoot*) showed the best coverage.

In the more simple setting of Scenario 3 all procedures controlled the coverage probability adequately with slight advantages for the enhanced bootstrap methods.

5 Data Example

We reanalyse the data from Meister and Schaefer (2008). It contains information on 1186 pregnant women who — or whose physician — contacted the Teratology Information Service in Berlin, Germany. Among those women, 173 were therapeutically exposed to coumarin derivatives, an anticoagulant, while 1013 women not exposed to known teratogens served as controls. Data were left-truncated as women usually become aware of their pregnancy several weeks after conception. Moreover, they are subject to competing risks as a pregnancy may end either in a spontaneous abortion,

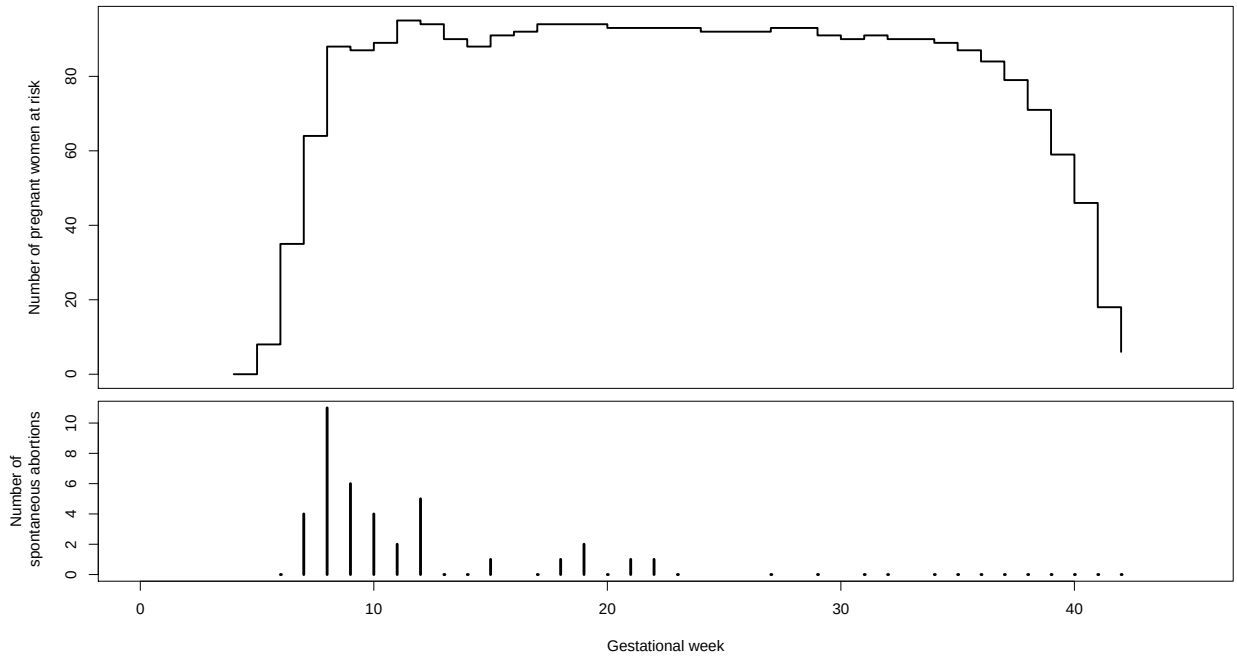

Figure 4: The upper panel displays $y(t)$, i.e., the number of pregnant women at risk over time. The bottom panel gives the number of spontaneous abortions at each gestational week.

induced abortion or a live-birth. We refer to Meister and Schaefer (2008) for more details on the data and Allignol et al. (2010) for an investigation of the independent left-truncation assumption and a comparison of different variance estimators.

The interest lies in estimating the probability of spontaneous abortion for the exposed women, and induced abortion and live-birth are subsumed into one competing endpoint. Figure 4 depicts the number of treated women at risk and the number of spontaneous abortion over the course of time. No women entered the cohort before week 4 but the first event happens at week 6 while 35 women were under observation. The risk set then grows steadily until around week 10.

The CIF of spontaneous abortion is displayed in Figure 5 along with the asymptotic 95% complementary-log-minus-log-transformed confidence intervals (grey lines). The asymptotic confidence intervals are compared to bootstrapped confidence intervals based on $10,000$ bootstrapped samples.

All variants of Efron's bootstrap investigated in Section 4 lead to confidence intervals that are close to the asymptotic ones. In contrast, confidence intervals based on the wild bootstrap are larger than the asymptotic pointwise confidence intervals. But note that the wild bootstrap confidence intervals are computed for $t > 7$ due to numerical instabilities at the earliest event time.

6 Discussion and Outlook

It is well known that ignoring left-truncation in time to event data may lead to biased estimates of endpoint probabilities(Meister and Schaefer, 2008). The effect of left-truncation on the coverage has, however, not been investigated considerably. Motivated by a recent pregnancy outcome study we investigated this effect in a competing risks set-up. Since pregnant women usually enter the study weeks after conception, left-truncation occurs naturally in these studies. It turned out that the usual $\log$-minus-$\log$ transformed confidence intervals (Lin, 1997; Allignol et al., 2010) for pregnancy outcome probabilities are very sensitive to the shape of the underlying left-truncation distribution. The same observation has also been made for naive bootstrap-based intervals obtained from the percentile method or bootstrapped variances. In particular, for rather left tailed truncation distributions the amount of truncation considerably increased their liberality leading to low coverage up to $50 - 60\%$ in the most extreme scenario. To tackle this problem, we proposed two more enhanced confidence intervals based on bootstrap and wild bootstrap techniques. Besides utilizing an adequate transformation, also bootstrapping the underlying weight functions shaped up as a key issue for gaining much better coverage probabilities. In the data example, we have illustrated their applicability.

In future research we plan to investigate the coverage of usual confidence intervals in more complex multi-state models. This is relevant even in the absence of left-truncation: For instance, in an illness-death model without recovery, initial state 0, intermediate illness state 1 and absorbing death state 2, there will be internal left-truncation due to 0

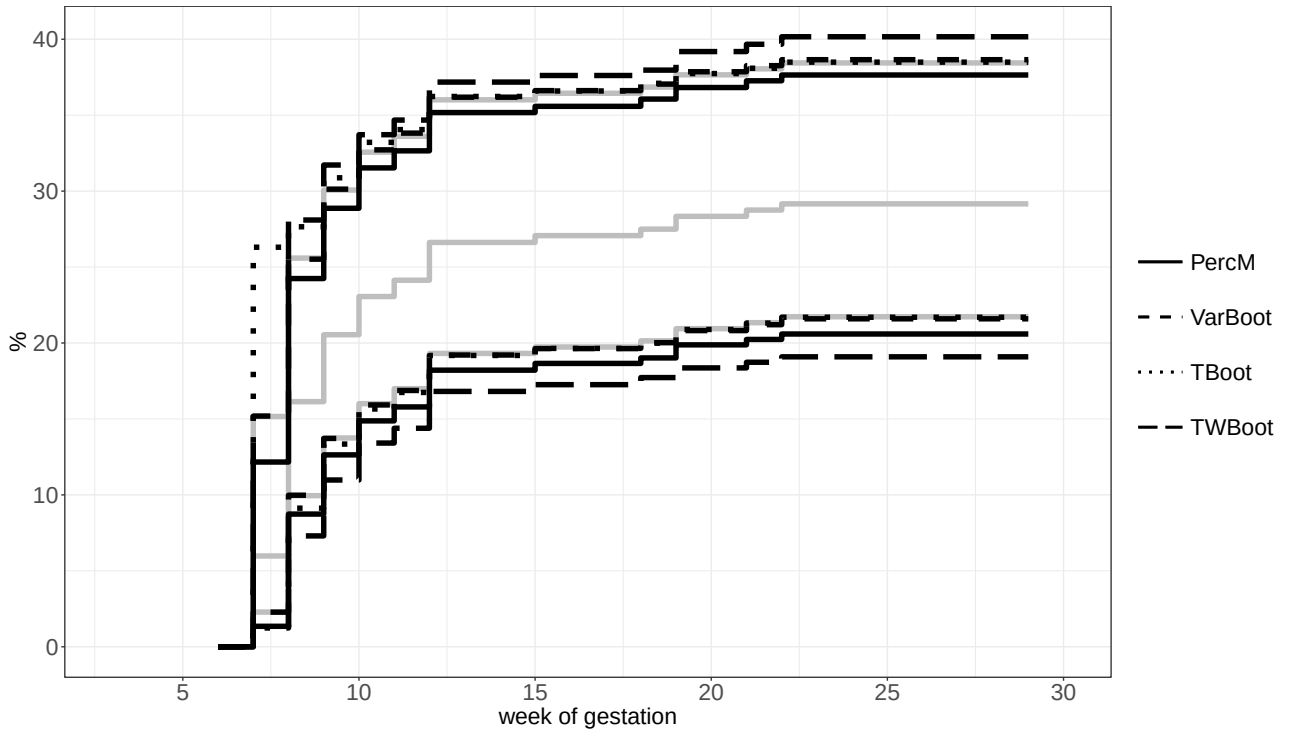

Figure 5: Cumulative incidence of spontaneous abortion for the women exposed to coumarin derivatives along with complementary log-log transformed 95% confidence intervals in grey. *PercM:* Percentile method using Efron's bootstrap. *VarBoot:* Plug-in variance based on Efron's bootstrap. *TBoot:* modified Efron confidence intervals based on (11). *TWBoot:* Wild bootstrap confidence intervals.

$\rightarrow 1$ transitions. Moreover, it is of research interest to investigate whether the more accurate confidence intervals lead to different results in the analysis of other studies with non-negligible left-truncation. For example, the analysis of pregnancy outcome studies with different strata should be revisited with the novel methods.

Acknowledgements

This work was supported by the German Research Foundation (DFG).

References

M.G. Akritas. Bootstrapping the Kaplan-Meier Estimator. *J. Amer. Statist. Assoc.*, 81:1032–1038, 1986.

A. Allignol, M. Schumacher, and J. Beyersmann. A Note on Variance Estimation of the Aalen-Johansen Estimator of the Cumulative Incidence Function in Competing Risks, with a View towards Left-Truncated Data. *Biometrical Journal*, 52:126–137, 2010.

A.-M. N. Andersen, P.K. Andersen, J. Olsen, M. Grønbæk, and K. Strandberg-Larsen. Moderate alcohol intake during pregnancy and risk of fetal death. *International Journal of Epidemiology*, 41(2):402–413, 2012.

P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer, New York, NY, 1993.

R. Arboretti, R. Fontana, F. Pesarin, and L. Salmaso. Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring. *Statistical methods in medical research, to appear*, 2018.

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher. Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28:956–971, 2009.

J. Beyersmann, A. Allignol, and M. Schumacher. *Competing Risks and Multistate Models with R*. Springer, New York, NY, 2012.

J. Beyersmann, S. Di Termini, and M. Pauly. Weak Convergence of the Wild Bootstrap for the Aalen-Johansen Estimator of the Cumlative Incidence Function of a Competing Risk. *Scandinavian Journal of Statististics*, 40:387–402, 2013.

A Borrell and I Stergiotou. Miscarriage in contemporary maternal–fetal medicine: targeting clinical dilemmas. *Ultrasound in Obstetrics & Gynecology*, 42(5):491–497, 2013.

D. Dobler. Bootstrapping the kaplan–meier estimator on the whole line. *Annals of the Institute of Statistical Mathematics*, pages 1–34, 2016.

D. Dobler and M. Pauly. Bootstrapping Aalen-Johansen processes for competing risks: Handicaps, solutions, and limitations. *Electron. J. Statist.*, 8:2779–2803, 2014.

D. Dobler, J. Beyersmann, and M. Pauly. Non-strange weird resampling for complex survival data. *Biometrika*, 104(3): 699–711, 2017.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7:1–26, 1979.

S. Friedrich, J. Beyersmann, U. Winterfeld, M. Schumacher, and A. Allignol. Nonparametric estimation of pregnancy outcome probabilities. *The Annals of Applied Statistics*, 11(2):840–867, 2017.

R.D. Gill and S. Johansen. A Survey of Product-Integration with a View Towards Application in Survival Analysis. *Ann. Statist.*, 18:1501–1555, 1990.

L.E. Grzeskowiak, A.L. Gilbert, and J.L. Morrison. Exposed or not exposed? exploring exposure classification in studies using administrative data to investigate outcomes following medication use during pregnancy. *American Journal of Epidemiology*, 68(5):459–467, 2012.

C.J. Hogue. Invited commentary: preventable pregnancy loss is a public health problem. *American Journal of Epidemiology*, 183(8):709–712, 2016.

N. Keiding and R.D. Gill. Random Truncation Models and Markov Processes. *Annals of Statististics*, 18:582–602, 1990.

D. Lin. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, 16:901–910, 1997.

D. Lin, L.J. Wei, and Z. Ying. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80:557–572, 1993.

S.-H. Lo and K. Singh. The Product-Limit Estimator and the Bootstrap: Some Asymptotic Representations. *Probability Theory and Related Fields*, 71:455–465, 1986.

A. Lupattelli, O. Spigset, M.J. Twigg, K. Zagorodnikova, A.-C. Mårdby, M.E. Moretti, M. Drozd, Alice Panchaud, Katri Hämeen-Anttila, and Andre Rieutord. Medication use in pregnancy: a cross-sectional, multinational web-based study. *BMJ open*, 4(2):e004365, 2014. doi: 10.1136/bmjopen-2013-004365.

R. Meister and C. Schaefer. Statistical methods for estimating the probability of spontaneous abortion in observational studies – Analyzing pregnancies exposed to coumarin derivatives. *Reprod. Toxicol.*, 26:31–35, 2008.

C. Schaefer, D. Hannemann, R. Meister, E. Eléfant, W. Paulus, T. Vial, M. Reuvers, E. Robert-Gnansia, J. Arnon, M. De Santis, M. Clementi, E. Rodriguez-Pinilla, A. Dolivo, and P. Merlob. Vitamin K antagonists and pregnancy outcome: A multi-centre prospective study. *Thromb Haemost*, 95:949–957, 2006.

R. Slama, F. Ballester, M. Casas, S. Cordier, M. Eggesbø, C. Iniguez, M. Nieuwenhuijsen, C. Philippat, S. Rey, and S. Vandentorren. Epidemiologic tools to study the influence of environmental factors on fecundity and pregnancy-related outcomes. *Epidemiologic Reviews*, 36(1):148–164, 2014.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, NY, 1996.