The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 29

Survival Curve Estimation with Dependent Left Truncated Data Using Cox's Model

Todd Mackenzie, *Dartmouth College*

Recommended Citation:
Mackenzie, Todd (2012) "Survival Curve Estimation with Dependent Left Truncated Data Using Cox's Model," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 29.
DOI: 10.1515/1557-4679.1312

Survival Curve Estimation with Dependent Left Truncated Data Using Cox's Model

Todd Mackenzie

Abstract

The Kaplan-Meier and closely related Lynden-Bell estimators are used to provide nonparametric estimation of the distribution of a left-truncated random variable. These estimators assume that the left-truncation variable is independent of the time-to-event. This paper proposes a semiparametric method for estimating the marginal distribution of the time-to-event that does not require independence. It models the conditional distribution of the time-to-event given the truncation variable using Cox's model for left truncated data, and uses inverse probability weighting. We report the results of simulations and illustrate the method using a survival study.

KEYWORDS: delayed entry, inverse probability weighting

1 Introduction

Truncation prevents observations for a subset of a sample space. In a left-truncated study, neither the dependent variable, Y, nor the truncating variable, V, are observable if $Y < V$. Examples of double-truncation arise in astronomy as described by Efron and Petrosian (1992, 94). A common epidemiological example of left-truncation is the prevalent cohort. In the prevalent cohort, left-truncated times-to-event arise because subjects are identified for study inclusion after the inception time (i.e., the time at which the time-to-event begins). Age at death is left-truncated if a subject does not enter the study at birth. Left-truncation is also referred to as *delayed entry*.

The Cox model for left-truncated data (Andersen et al, 1997, Keiding, 1992, Gail et al, 2009) is frequently used to model covariate effects on age at event. In this version of Cox's model, age is used as the time-scale (i.e., argument of the hazard function) and the counting process style of data specification is used, which consists of the start time (i.e., the age of left-truncation), the stop time (age at event or right censoring) and the event indicator.

If the left-truncation variable is independent of the time-to-event of interest, the distribution of the time-to-event can be estimated using the estimator of Lynden-Bell (1971). The distribution of an estimator which is both left-truncated and right-censored can be estimated using the estimator of Kaplan and Meier (1958). In this case the risk set used for calculating the Kaplan-Meier is defined differently than in the non-truncated setting: the risk set at a particular time, t, is that set of subjects for whom the time-to-event and censoring time exceed t, and whose left-truncation is less than t. The Lynden-Bell estimator is not applicable when right-censoring is present. It also differs from the Kaplan-Meier in that the former is left-continuous but the latter is right-continuous. The asymptotics of the Kaplan-Meier under independent truncation have been discussed by Woodroofe (1985), Wang, Jewell and Tsai (1986) and Tsai, Jewell and Wang (1987). A treatment of truncation in terms of Markov processes was given by Keiding and Gill (1990). Estimation of marginal survival in the setting of a truncation variable that is independent but parameterized was addressed by Wang (1989). Length-biased sampling (Wang, 1996, Ashgarian et al, 2002) can be considered a special case of left-truncation. The Kaplan-Meier is sensitive to small risk sets if there is left-truncation. For instance, if there is an event when the size of the risk set is one, the resulting Kaplan-Meier decreases to zero. Lai and Ying (1991) proposed a method for overcoming this by only taking increments in the Kaplan-Meier when the risk set exceeds $n\alpha$ for some $0 < \alpha < 1$ (e.g. $\alpha = 1/3$).

If the truncation variable and time-to-event variable are not independent of the survival time, and furthermore unparameterized, then the Kaplan-Meier and Lynden-Bell are not consistent estimators of the time-to-event distribution. Positive correlations lead to underestimation, and negative correlations cause overestimation (Keiding, 1992). Tsai (1990) proposed a generalization of the Kendall tau test statistic to test for independence of truncation times and times-to-event in left-truncated right-censored data sets. Efron and Petrosian (1992,94) discussed its application with data that are left-truncated but not right-censored. The test is developed further by Martin and Betensky (2005). A test based on the product-moment correlation has been proposed by Chen et al (1996). Alternatively, Jones and Crowley (1992) proposed the use of Cox's model for testing the association. In particular, they model the conditional distribution of time-to-event given truncation time using Cox's model for left-truncated data (Andersen et al, 1997). The conditional distribution can be used to predict time-to-event based on truncation time.

The marginal distribution of the time-to-event retains importance if there is a correlated left-truncation time. The conditional distribution of the time-to-event given the left-truncation time may be of interest in some applications, but left-truncating time may be a nuisance variable. Chaieb et al (2006) developed an estimator of the time-to-event distribution in the presence of non-independent truncation. They provide a nonparametric estimate of the time-to-event distribution using a copula to model the joint distribution of time-to-event and left-truncation time. Implementation of their method requires the user to specify a copula from an Archimedian family. Choice of this copula was discussed by Beaudoin et al (2008).

We propose a semi-parametric method for estimating the distribution of a time-to-event from left-truncated data that does not assume independence or any parametric form for the distribution of the truncation variable. This method starts by modelling the dependence of survival on the truncation time using Cox's model for left-truncated data (Andersen et al, 1997, Keiding, 1992). It proceeds by invoking inverse probability estimation (Horvitz and Thompson, 1952, Robins and Rotnitsky, 1992). The Kaplan-Meier has been shown to be an inverse probability estimator (Satten and Datta, 2001, Shen, 2003). In addition to the marginal distribution of survival times, this approach also yields an estimate of the distribution of the left-truncation times, and an estimator of the probability of truncation. We report the behavior of our estimator using simulations, and illustrate our method using left-truncated data from users of the health system of the Veterans Administration (VA) to estimate the survival of VA users after the age of 65. We conclude with a discussion addressing the strengths and limitations of this method.

2 Methods

2.1 Notation and Assumption of Non-informative Censoring

Let Y be the time-to-event of interest, with cumulative distribution function (CDF) F_Y, and let $S_Y = 1 - F_Y$. Let F_V be the CDF of the truncating variable, V. We assume that F_Y and F_V are continuous functions. The available data are $\{T_i, \Delta_i, V_i\}_{i=1}^n$, where $T_i = \min(Y_i, V_i + E_i)$, Δ_i equals 1 if $Y_i \leqslant V_i + E_i$, and 0 otherwise, and E_i is a follow-up time that commences when the person enters the cohort. We assume that E_i is independent of Y_i conditional on V_i and $Y_i \geqslant V_i$. In the case of no truncation where $\Pr[V = 0]$=1 this is equivalent to the usual assumption of non-informative censoring.

Suppose the distribution of the time-to-event, Y, given the left-truncation variable, V, follows a Cox model. Note that this assumption is untestable: we can test the assumption on the sample space $Y \geqslant V$, but we cannot know if it holds for $Y < V$. The assumption of this Cox model can be written, $\Pr[Y \geqslant y|V = v] = \exp[-g(v;\beta)\Lambda(y)]$ where $\Lambda(y)$ is the baseline cumulative hazard function and $g(v;\beta)$ is some continuous function of β for which $g(v;0) = 1$ (e.g., $g(v;\beta) = \exp[\beta v]$). In counting process notation we are assuming that $\mathrm{E}dN_i(t) = d\Lambda(t)g(v_i;\beta)R_i(t)$ where $N_i(t)$ is the counting process which equals 1 if $\Delta_i = 1$ and $T_i \leq t$, and is zero otherwise, $R_i(t)$ is the predictable process which equals 1 if $T_i \geq t \geq V_i$ and is 0 otherwise.

We shall let $Q = \Pr[Y \geqslant V]$, the probability that an observation is *not* excluded due to left-truncation.

2.2 Estimation

The marginal distribution of Y is given by

$$Pr[Y \geqslant y] = \int_0^\infty \exp[-g(v;\beta)\Lambda(y)]dF_V(v) \tag{1}$$

We shall estimate this by substituting estimators for β, $\Lambda(t)$, and $F_V(t)$. Estimators for β and Λ are available using existing theory. A consistent estimator of the parameter β can be found by solving the solution to the estimating equation $\sum_{i=1}^n \int_0^\tau \left[v_i - \sum_{j=1}^n R_j(t)v_j \exp[\beta v_j] / \sum_{j=1}^n R_j(t)\exp[\beta v_j] \right] dN_i(t)$ (Andersen et al, 1997, Keiding, 1992). This can be done readily using existing software by specifying the truncation time as both a covariate and as the entry time in the time-to-event triplet consisting of the time of event (or censoring), the event (censoring) indicator, and the entry time. In the statistical language R, this is accomplished using code such as coxph(Surv(tr_time, time, status)). In order to be less sensitive to small risk sets we use the correction of Lai and Ying (1991), which prevents the cumulative hazard from taking a jump if the size of risk set is less than $n^{1/3}$.

Gail et al (2009) suggested adjusting for truncation time in models that are left-truncated. For instance, they recommend that Cox's model for left-truncated data include truncation time (e.g., age at beginning of follow-up) as a covariate in addition to other covariates of interest. This is to account for any association of truncation time with survival time. In our paper, Cox's model for left-truncated data includes no other covariates except the truncation time. The parameter $\Lambda(t)$ of Cox's model for left-truncated data can be consistently estimated using the Breslow estimator for left-truncated data $\sum_{i=1}^{n} \int_0^{\tau} \frac{dN_i(t)}{\sum_{j=1}^{n} R_j(t) \exp[\hat{\beta} v_j]}$ (Andersen et al, 1997, Keiding, 1992).

An estimator of the distribution of V, represented by F_V, can be found using the concept of inverse probability weighting (Horvitz and Thompson, 1952, Satten and Datta, Shen, 2003). The probability of not being left-truncated if $V = v$, is $Pr[Y \geqslant V|V = v]$ which equals $\exp[-g(v;\beta)\Lambda(v)]$. The principle of inverse probability weighting is to weight an observation by the reciprocal of the probability that the observation is not missing, e.g., not truncated. For instance, the weight assigned to the observation for which $V = v$ is $1/\exp[-g(v;\beta)\Lambda(v)]$.

In order to explain this estimator we find it useful to introduce the following harmonic mean. Let $\widehat{Q}$ equal $n/\sum_{i=1}^{n} 1/\exp[-g(v_i;\widehat{\beta})\widehat{\Lambda}(v_i)]$, the harmonic mean of $\{\exp[-g(v_i;\hat{\beta})\Lambda(v_i)]\}_{i=1}^{n}$. The inverse probability weighted estimator of F_V assigns a mass of $\frac{\widehat{Q}}{n}/\exp[-g(v_i;\hat{\beta})\Lambda(v_i)]$ to the point v_i. This harmonic mean, $\widehat{Q}$, is an estimator of the probability of not being truncated. This is because $Q = Pr[Y \geqslant V]$ equals $\int_0^{\infty} Pr[Y \geqslant v]dF_V(v) = \int_0^{\infty} \exp[-g(v;\hat{\beta})\Lambda(v)]dF_V(v)$, which we estimate by the expression

$$\sum_{i=1}^{n} \left\{ \exp[-g(v_i;\widehat{\beta})\Lambda(v_i)] \frac{\widehat{Q}}{n} / \exp[-g(v_i;\widehat{\beta})\Lambda(v_i)] \right\} = \widehat{Q}. \tag{2}$$

The estimator of the truncation distribution is

$$\hat{F}_V(v) = \frac{\widehat{Q}}{n} \sum_{v_i \leq v} \exp[-g(v_i;\widehat{\beta})\{\widehat{\Lambda}(y) - \widehat{\Lambda}(v_i)\}]. \tag{3}$$

An estimate of the marginal distribution of Y is obtained by substituting estimates of β, Λ and F_V into (1) to yield

$$\widehat{S}_Y(y) = \frac{\widehat{Q}}{n}\sum_{i=1}^{n} \exp[-g(v_i;\widehat{\beta})\{\widehat{\Lambda}(y) - \widehat{\Lambda}(v_i)\}]. \tag{4}$$

This expression reduces to an expression related to the Nelson-Aalen estimator for left-truncated survival if 0 is substituted for β (work not shown). If we express the right-hand side of equation (1) using product integral notation, the resulting estimator is $\widehat{S}_Y(y) = \frac{\widehat{Q}}{n}\sum_{i=1}^{n}\prod_{v_i}^{y}[1-g(v_i;\widehat{\beta})d\widehat{\Lambda}(u)]$, which reduces to the Kaplan-Meier for $\beta = 0$.

To estimate the variance we suggest a bootstrap or jackknife approach.

2.3 Untestable Assumption

This method does make an untestable assumption: whereas it is possible to test the modelling assumption of the joint distribution of the time-to-event and the truncation time (i.e. Cox model) on the observable region, it is not possible to verify it on the truncated region. It is possible that the association determined using the observable region does not hold on the truncation region. As a consequence the estimator we propose may be biased. The extent of this bias will be driven by the discrepancy (e.g. how much the joint distribution on the truncated region differs from the model extrapolated from the observable region) and the frequency of truncation, Q. As Y and V are not observed on the region $Y < V$ we know nothing about their joint distribution on that region. It is plausible that the joint distribution actually has zero mass on the region $Y \leqslant V$. In other words, there is no actual truncation. Our method is based on the extrapolation of a Cox model on the observeable region to the truncated region and then integrating with respect to V to obtain the marginal of Y.

3 Example

In this section we estimate the overall survival curve of users of the VA health system. The survival distribution of VA users is essential to policy decisions made by the VA. We use survival data from approximately 850,000 VA patients who were randomly sampled to complete a VA survey in 1999 (MacKenzie et al, 2010). These survival data are based on records from the VA Vital Status Registry. Age at death is the time-to-event of interest, Y. The left-truncation time, V, is age at the time of the 1999 survey, as patients who died before the survey are not part of the

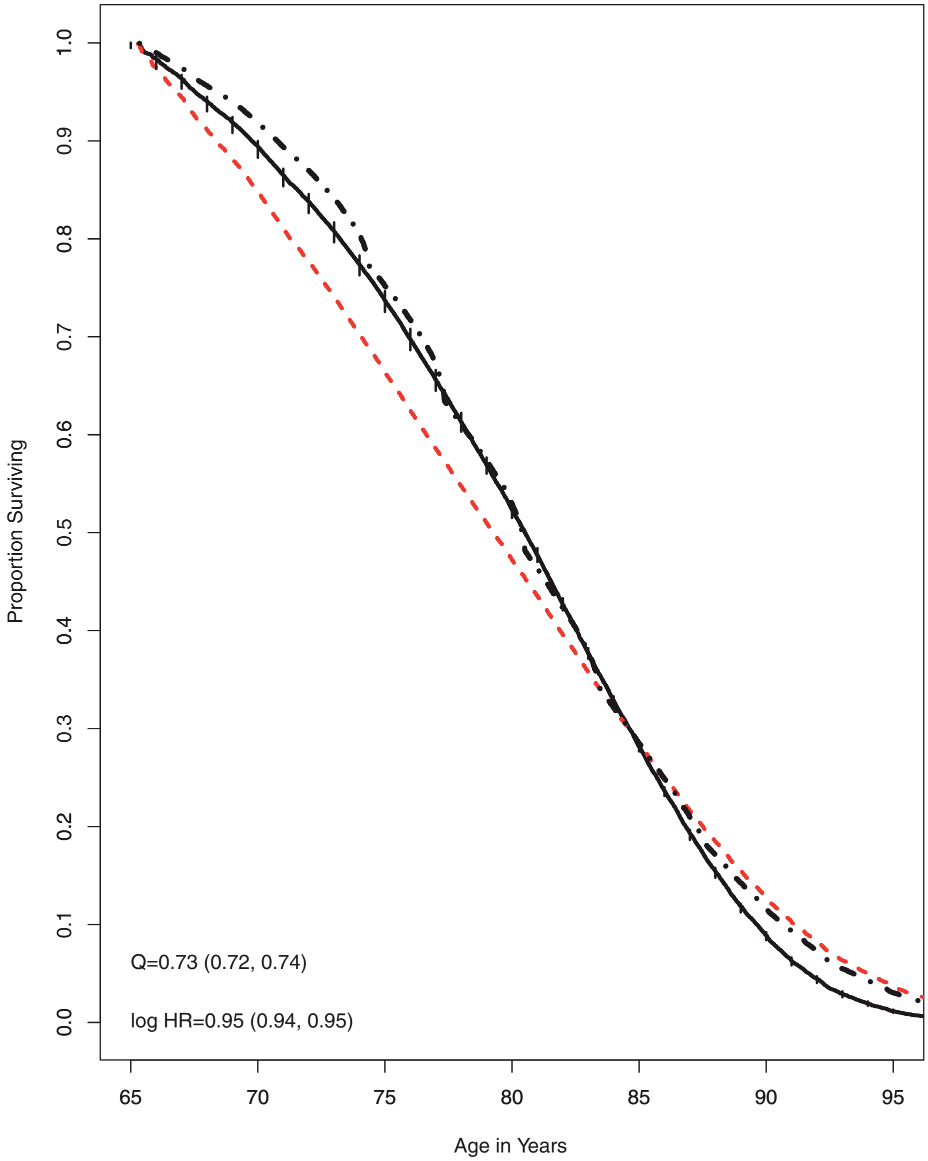

Figure 1: Survival curves for Veterans over the age of 65 based on 350 thousand respondees to a random survey. The curves have been estimated using the Kaplan-Meier for left-truncated data (dashed line) which assumes independence of survival and age at time of survey, and the method we are proposing (solid line) with 95% confidence intervals (dashed lines). This estimator assumes that the dependence of survival and age at survey follow a Cox model. The dotted line is the estimator of Chaieb et al (2006) implemented using a copula from the Frank family.

dataset. We have restricted attention to VA users who were age 65 or more in 1999. Individuals are right censored if they were alive as of 2006.

We tested the independence of the left-truncation time and age at death using Jones and Crowley's (1992) approach based on the Cox model. There was a significant association (P¡0.0001). An increase of one year in age at the time of the survey was associated with a 4.7% (95% CI: 4.1% to 5.2%) reduction in incident mortality. This means that among two VA users of the same age, the veteran who was younger when he completed the survey is more at risk. For example, a veteran of age 80 in 2010 (69 at time of survey) is more at risk at that age than was a veteran who was 80 in 2005 (74 at time of survey).

Regression splines were used to explore the functional form of this association. We determined that the log hazard is approximately linear with respect to truncation time. In addition, we found no evidence that the hazard ratio changed over time.

Figure 1 demonstrates the estimator we propose (solid line) and the Kaplan-Meier estimator (dotted line) that assumes independence of age at survey and age at death. The difference between our survival estimate and the Kaplan-Meier is 5% or more between the ages of 73 and 77. The probability of not being truncated, Q, was 72% (95% CI: 73% to 74%).

As described in section 2.3, the estimator we propose makes an untestable assumption. It assumes the conditional distribution of survival given the age at time of survey is a Cox model both on the observable region (testable) and on the truncated region (untestable). It assumes that being older at the time of survey is sign of a decreased hazard not just after the survey but before it. For instance, individuals who were destined to take the survey at age 65 (should they have lived until the period of the survey in 1999 and made a visit to the VA during that time) had a higher mortality rate before age 65 than individuals who were destined to take the survey at age 75.

4 Simulations

4.1 Methods

We evaluated the bias of the estimator we propose using a range of truncation probabilities, 0.1, 0.3, 0.5, 0.7 and 0.9, (i.e., Q =0.9,0.7,0.5,0.3,0.1), and a range of associations between the left-truncation time and the time-to-event, as measured by the hazard ratio comparing those at the 75th percentile of the truncation time to those at the 25th percentile, HR=exp[1.35 $*\beta$] $= 1/2, 2/3, 1, 3/2, 2$ (i.e., β =-0.51,

Table 1: Estimation of Q: Mean across 2500 simulations for a range of actual Q and hazard ratios, and censoring rate of 50%. Q is 1 minus the probability of truncation.

		Hazard Ratio				
Truncation %	**Q**	**1/2**	**2/3**	**1**	**1.5**	**2**
		N=200				
10	**0.9**	0.91	0.90	0.90	0.90	0.90
30	**0.7**	0.71	0.71	0.71	0.70	0.70
50	**0.5**	0.50	0.50	0.50	0.50	0.50
70	**0.3**	0.30	0.30	0.31	0.31	0.33
90	**0.1**	0.24	0.23	0.25	0.26	0.27
		N=500				
10	**0.9**	0.90	0.90	0.90	0.90	0.90
30	**0.7**	0.70	0.71	0.70	0.70	0.70
50	**0.5**	0.50	0.50	0.50	0.50	0.50
70	**0.3**	0.30	0.29	0.29	0.30	0.31
90	**0.1**	0.18	0.19	0.21	0.23	0.24

-0.30, 0.00, 0.30, 0.51). The censoring rate was fixed at 50%. We considered a sample size of 200 (100 events) and a sample size of 500 (250 events). The time-to-event followed an exponential distribution, as did the distribution of truncation times, while the censoring distribution was uniform. For each of these scenarios, 2500 replications were carried out, and estimation of the survival curve was done using the estimator we propose (4) and the Kaplan-Meier.

4.2 Results

Table 1 shows the mean of the estimator of the quantity Q, the complement of the truncation probability, for the cases of N=200 and N=500 (censoring rate fixed at 50%) as the true value of Q and the hazard ratio vary. There is very little bias when the actual Q is 0.5 or larger. For $Q = 0.3$ absolute bias is as large as 0.02. The estimator is poor when the true truncation rate is 90% ($Q = 0.1$) and when the hazard ratio is 1 or larger. It is only slightly ameliorated for a larger sample size. This poor behavior for 90% truncation rates is not due to bias of the partial likelihood estimator of the hazard ratio. Mean values of the estimator of the log hazard ratio are shown in Table 2. The partial likelihood estimator for Cox's left-

Table 2: Estimation of Cox's Hazard Ratio with Correlated Truncated Data: Table shows mean of the log hazard ratio across 2500 simulations for a range of truncation probabilities, 1-Q, and actual hazard ratios, and censoring rate of 50%.

Truncation %	**Q**	**Hazard Ratio**				
		1/2	**2/3**	**1**	**1.5**	**2**
		log Hazard Ratio				
		-0.69	**-0.40**	**0**	**0.40**	**0.69**
		N=200				
10	**0.9**	-0.70	-0.41	0.00	0.41	0.69
30	**0.7**	-0.70	-0.41	0.01	0.41	0.70
50	**0.5**	-0.70	-0.41	0.00	0.41	0.70
70	**0.3**	-0.68	-0.42	0.00	0.42	0.71
90	**0.1**	-0.71	-0.44	0.00	0.45	0.72
		N=500				
10	**0.9**	-0.70	-0.41	0.00	0.40	0.69
30	**0.7**	-0.69	-0.41	0.00	0.40	0.70
50	**0.5**	-0.70	-0.40	0.00	0.40	0.69
70	**0.3**	-0.69	-0.39	0.00	0.42	0.69
90	**0.1**	-0.72	-0.42	0.00	0.44	0.71

truncated model exhibits a minor amount of bias for N=200 and $Q = 0.1$ but is otherwise unbiased.

Figure 2 shows the simulation results for the bias of the survival curve estimator for a sample size of 200, and 50% censoring. Each panel in the figure is a plot of the bias (estimated survival curve minus actual survival curve) versus the time argument, transformed using the function that maps t to $1 - S(t)$ so it is confined to the unit interval. The solid line is the bias of the estimator we propose. The dashed line is the Kaplan-Meier for left-truncated data. An estimator is biased to the extent that it deviates from the zero line. The 25 panels are indexed vertically by hazard ratios of 1/2, 2/3, 1, 1.5 and 2, and horizontally by Q.

There is no bias in the Kaplan-Meier when the time-to-event and truncation time are independent (hazard ratio of 1). The bias of the Kaplan-Meier increases as the hazard ratio moves away from unity, and as the truncation rate increases (as Q decreases). There is little bias in the Kaplan-Meier when the truncation rate is 10% (Q=0.9).

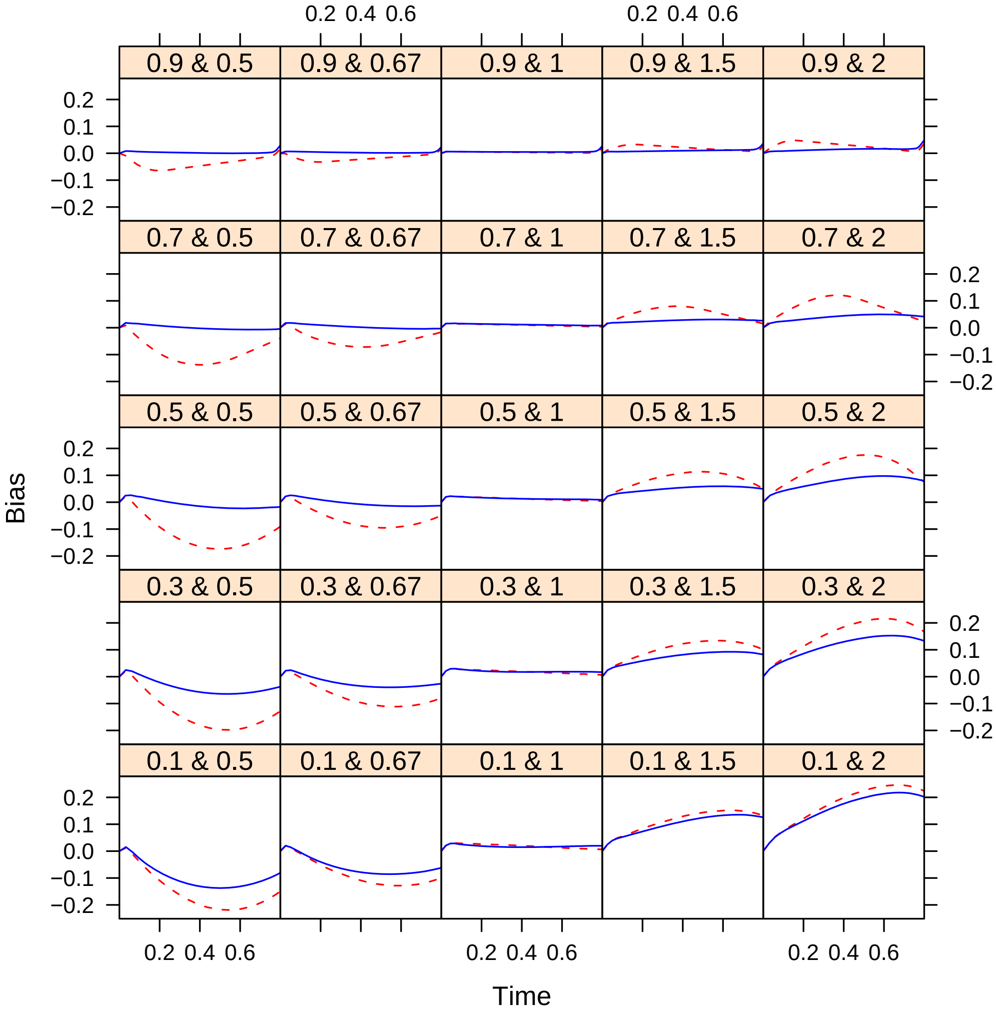

Figure 2: Results of simulations for a sample size of 200 and a range of non-truncation probabilities, Q, and a range of associations of the time-to-event and left-truncation time, HR, for a fixed right-censoring rate of 50%. There is one panel for each choice of HR from the list, 0.5, 0.67, 1, 1.5 and 2, and each choice of Q from the list 0.9, 0.7, 0.5, 0.3, 0.1. The bar above each panel shows the Q and the HR. Each panel is the bias over time as the mean value of the estimate (based on 2500 replications) minus the true value. The solid line is the estimator we propose and the dashed line is the Kaplan-Meier. The time scale has been transformed to the unit scale using the transformation, $1 - S(t)$.

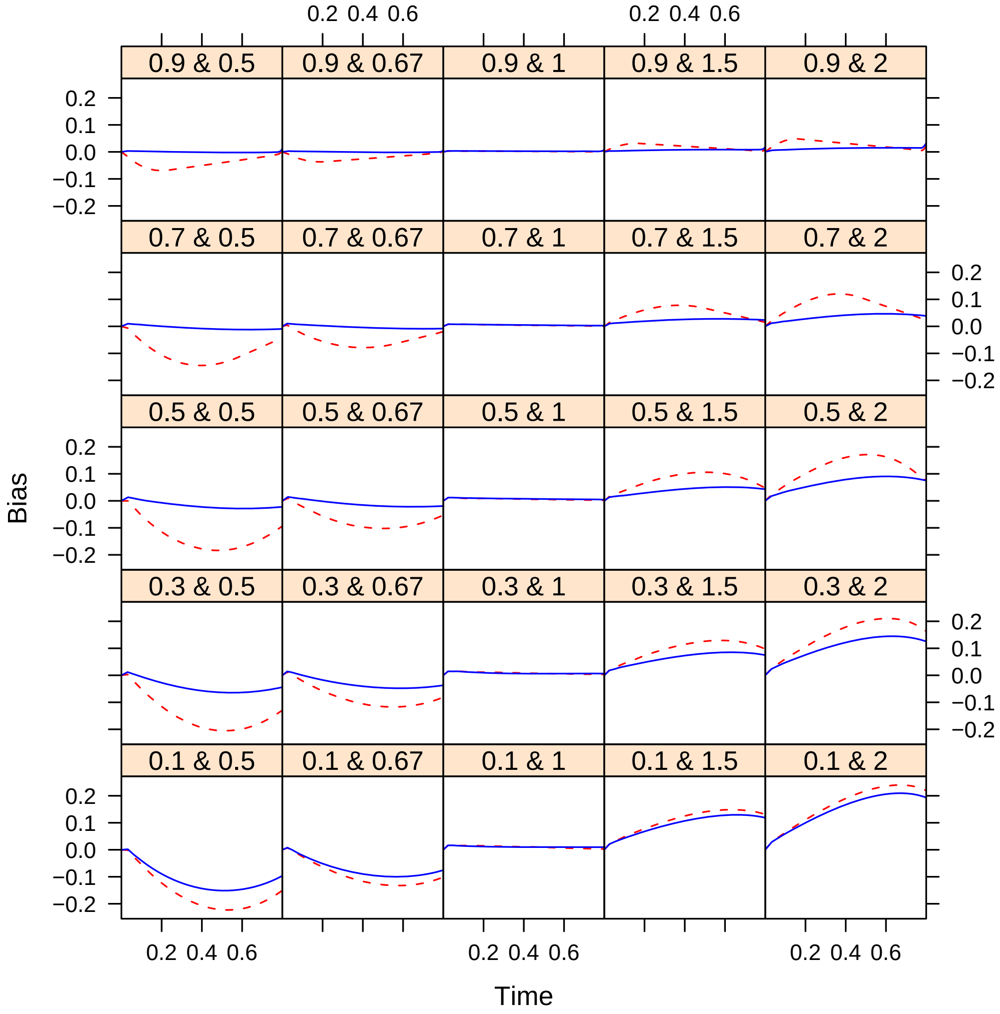

Figure 3: Results of simulations for a sample size of 500 and a range of non-truncation probabilities, Q, and a range of associations, HR, of the time-to-event and left-truncation time, for a fixed right-censoring rate of 50%. There is one panel for each choice of HR from the list, 0.5, 0.67, 1, 1.5 and 2, and each choice of Q from the list 0.9, 0.7, 0.5, 0.3, 0.1. The bar above each panel shows the Q and the HR. Each panel is a plot of the bias (mean value of the estimate based on 2500 replications minus actual value) versus time (transformed to the unit interval using the inverse cumulative distribution function). The solid line is the estimator we propose and the dashed line is the Kaplan-Meier for left-truncated data.

There is nearly zero bias of the estimator we propose if Q is 0.5 or greater and if the hazard ratio is unity or less. In contrast, the Kaplan-Meier overestimates by more than 0.2 for $Q = 0.3$ and $HR = 2$ and underestimates by more than 0.2 for $Q = 0.3$ and $HR = 1/2$. The advantage of our estimator over the Kaplan-Meier is most evident with the truncation rate between 30% and 70% and when the hazard ratio is less than unity, corresponding to a positive correlation of time-to-event and truncation time. Conversely, with $Q = 0.1$, where a full 90% of the data has been truncated, our estimator exhibits almost as much bias as the Kaplan-Meier.

Figure 3 shows simulation results for the sample size of 500 and 50% censoring. The findings are very similar to those for N=200. The estimator we propose exhibits no bias for Q of 0.3 or larger. For $Q = 0.1$ the estimator is somewhat biased but slightly less than for the case of N=200.

5 Discussion

We have proposed an estimator for the cumulative distribution function of a right-censored time-to-event that is sampled under dependent left-truncation. This method is semi-parametric. It makes no assumptions about the marginal distribution of the time-to-event, nor the marginal distribution of the truncation time. It should be considered when there is evidence of dependence between the time-to-event and the truncation time.

It assumes that the conditional distribution of the time-to-event given the truncation time follows the Cox model for left-truncated data. The advantage of this approach is the universality of Cox's model in biostatistics. It would be preferable to use a completely nonparametric approach but it is not clear that this is possible. A semi-parametric estimator has also been derived by Chaieb et al (2006). The Chaieb estimator requires specification of a copula for the joint distribution of the time-to-event and truncation time.

The appropriateness of the assumption of Cox's model can be examined and if violations of Cox's model are indicated, then a generalization of Cox's model can be used. For instance, it is not necessary to use the usual hazard ratio of $\exp(v;\beta)$. The actual functional form of the hazard ratio could be estimated using a method of nonparametric regression. The hazard ratio $g(v;\beta)$ could be generalized to a time-dependent hazard ratio, $g(v, y;\beta)$.

A clear limitation of our approach is that it involves an untestable assumption. The appropriateness of Cox's model can be evaluated on the sample space for which $Y \geqslant V$. However it cannot be evaluated on the sample space for which $Y < V$. Analogously, the approach of Chaieb et al (2006) makes the assumption that the joint distribution of Y and V can be be parameterized by a copula, which is untestable in the truncated region. Without observation of the truncated region one cannot know anything about the actual distribution of Y and V on it. It is plausible that there is zero mass on the truncated region. Our method involves the extrapolation of a model estimated using the observeable region to the truncated region. This extrapolation will lead to biased estimation of the distribution of Y to the extent that the extrapolation is poor and the probability of truncation is large. Further work could explore the development of sensitivity analyses for exploring violations of the extrapolation. Further work should also be directed at determining just how weak the untestable assumption can be made. Whether the concepts of quasi-independence for truncated data (Tsai, 1990, Martin and Betensky, 2005) can be generalized to quasi-dependence is also an area requiring further research. Untestable assumptions are not uncommon in statistics: the assumption of non-informative censoring that is required for consistent estimation of the Kaplan-Meier and most survival estimators is untestable.

The scenario of left-truncation imposes counterfactual considerations. In our example using VA data, the left-truncation time is age at the time of the survey. The truncated individuals are those who would have completed the survey had they lived that long. This is a counterfactual concept. Another example of dependent left-truncation is estimation of survival in cystic fibrosis (CF) patients using a CF registry. Individuals who die before being diagnosed with CF are left truncated. Older age at diagnosis is associated with greater longevity. Using the method of this manuscript to estimate survival for CF patients would yield an estimate of age at death among people who have been or would be diagnosed with CF if they were to live long enough. Again, this is counterfactual reasoning. It is possible to avoid the counterfactual reasoning by estimating the conditional distribution of survival given the age at diagnosis.

A limitation of marginal estimators based on truncated data, including the Kaplan-Meier and the estimator we propose, is that they may severely underestimate survival. For example, the risk set at the first event time (or any event time) may contain only one subject, in which case the Kaplan-Meier becomes zero at that point because the contribution to the product-limit is $1 - 1/1 = 0$. An approach to helping resolve this limitation has been proposed by Lai and Ying (1991). Their resolution could be adapted to our estimator.

Our Monte Carlo simulations demonstrated excellent behavior of our estimator for truncation rates at or below 50% (i.e., Q of 0.5 or more) and a hazard ratio (of the interquartile range) of unity or less. We speculate that the reason the estimator has more difficulty for hazard ratios greater than unity is due to a negative association between the time-to-event and truncation time, which is a difficult association to estimate using only data that fall on one side of the line of identity ($Y \geqslant V$). For rates of truncation of 90% (Q=0.1), the estimator we propose exhibits almost as much bias as the Kaplan-Meier for left-truncated data. Further investigation is needed to determine the cause of this poor behavior, and if possible, how to correct this estimator for bias.

Future directions include incorporation of covariates, $\overrightarrow{X}$, into the model, $\Pr[Y \geqslant V|V = v, X = x] = \exp[-g(v, x; \beta)\Lambda(y)]$. Other future directions include adapting the correlated left-truncated problem to the case of parameterized truncation distributions (Wang, 1989) and to the case of bivariate truncation as observed in applications with age-of-onset anticipation (Huang et al, 2001).

6 Appendix 1: R Code

```
dep.truncation.Survival ¡- function(time, truncation.time, status, Prob.Trunc=NULL,
Lower.Bound=NULL) {
        keep ¡- !is.na(time) & !is.na(status) & !is.na(truncation.time)
        t ¡- time[keep]
        v ¡- truncation.time[keep]
        s ¡- status[keep]
        if (is.null(Lower.Bound)) Lower.Bound ¡- 0
        ord ¡- order(v)
        v ¡- v[ord]
        t ¡- t[ord]
        s ¡- s[ord]
        o.cox ¡- coxph(Surv(v, t, s) ˜v)
        lin ¡- as.matrix(v) %*% o.cox$coef
        lin ¡- lin - mean(lin)
        o.CH ¡- Cumulative.Hazard(t, v, s, lin)
        n.t ¡- length(o.CH$time)
        tt ¡- c(0, o.CH$time)
        CH ¡- c(0, o.CH$Cum.Haz)
        if (is.null(Prob.Trunc)) {
        Prob.Trunc ¡- rep(NA, length(v))
        i.t ¡- 1+n.t
        for (i in length(v):1) {
        while(tt[i.t] ¿ v[i]) i.t ¡- i.t-1
        Prob.Trunc[i] ¡- exp(-exp(lin[i])*CH[i.t])
        }
        }
        if (!is.null(Prob.Trunc)) {
        Prob.Trunc ¡- (Prob.Trunc[keep])[ord]
        }
        # Use lower bound
        Prob.Trunc ¡- ifelse(Prob.Trunc¡Lower.Bound, Lower.Bound, Prob.Trunc)
        Q ¡- 1/mean(1/Prob.Trunc)
        S ¡- rep(NA, n.t)
        for (i.t in 1:n.t) {
        S[i.t] ¡- Q * mean(exp(-exp(lin)*o.CH$Cum.Haz[i.t]) / Prob.Trunc)
        }
        S.cond ¡- list(time=o.CH$time, surv=exp(-o.CH$Cum.Haz))
        list(survival=S, time=o.CH$time, Q=Q, log.HR=o.cox$coef, cox.iter=o.cox$iter,
S.cond=S.cond)
        }
        Cumulative.Hazard ¡- function(time, time.start=0, status, x, correction.power=1/3)
{
        if (length(time.start)==0) time.start ¡- rep(0, length(time))
        n ¡- length(time)
        ot ¡- order(time)
        ti.ot ¡- time[ot]
        ti.start.ot ¡- time.start[ot]
        st.ot ¡- status[ot]
        x.ot ¡- x[ot]
        uniq.ev.ti ¡- unique(ti.ot[st.ot==1])
        n.uniq.ev.ti ¡- length(uniq.ev.ti)
        H ¡- rep(NA, n.uniq.ev.ti)
        nr ¡- rep(NA, n)
        for (i in 1:n.uniq.ev.ti) {
        Y.i.t ¡- (ti.ot¿=uniq.ev.ti[i]) * (ti.start.ot ¡ uniq.ev.ti[i])
        H[i] ¡- ifelse(sum(Y.i.t)¡=nˆcorrection.power, 0, sum(st.ot==1 & ti.ot==uniq.ev.ti[i])
/sum(Y.i.t * exp(x.ot)))
        }
        list(time=uniq.ev.ti, Cum.Haz = cumsum(H))
        }
```

References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1997) ”Statistical Models Based on Counting Processes”, Springer: Series in Statistics.

Asgharian, M., M’Lan, C.E. and Wolfson, D. (2002) ”Length-Biased Sampling with Right Censoring: An Unconditional Approach,” *JASA*, **97(457)**:201-209.

Beaudoin, D. and Lakhal-Chaieb, L. (2008) ”Archimedean copula model selection under dependent truncation,” *Statistics in Medicine*, **27(22)**:4440-4454.

Chaieb, L.L., Rivest, L.P. and Abdous, B. (2006) ”Estimating survival under a dependent truncation,” *Biometrika*, **93**:655-669.

Chen, C.H., Tsai, W.Y. and Chao, W.H. (1996) ”The Product-moment Correlation Coefficient and Linear Regression for Truncated Data,” *Journal of the Americal Statistical Association*, **91**:1181-1186.

Efron, B. and Petrosian, V. (1992) ”A simple test of independence for truncated data with applications to redshift survey,” *Astrophysical Journal*, **399**:345-352.

Efron, B. and Petrosian, V. (1994) ”Survival analysis of the gamma-ray burst data,” *Journal of the American Statistical Association*, **89**:452-462.

Gail, M.H., Graubard, B., Williamson, D.F. and Flegal, K.M (2009) ”Comment on Choice of time scale and its effect on significance of predictors in longitudinal studies,” *Statistics in Medicine*, **28(8)**:1315-17.

Horvitz, D.G. and Thompson, D.J. (1952) ”A generalization of sampling without replacement from a finite universe”, *Journal of the American Statistical Association*, **47**:663-685.

Huang J., Vieland, V.J. and Wang, K. (2001) ”Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-onset anticipation,” *Statistica Sinica*, **11**:1047-1068.

Jones, M.P. and Crowley, J. (1992) ”Nonparametric Tests of the Markov Model for Survival Data”, *Biometrika*, **79(3)**:513-522.

Kaplan, E.L. and Meier, P. (1958) ”Nonparametric Estimation From Incomplete Observations”, *Journal of the American Statistical Association*, **53**:457-481.

Keiding, N. (1991) ”Independent Delayed Entry”, In: Klein, J.P. and Goel, P.K., eds., *Survival Analysis: State of the Art*, 309-328.

Keiding, N. and Gill, R.D. (1990) ”Random Truncation Models and Markov Processes”, *Annals of Statistics*, **18(2)**:582-602.

Klein, J.P. and Moeschberger, M.L. (1997) ”Survival Analysis: Techniques for Censored and Truncated Data”, Springer: Statistics for Biology and Health.

Lai, T.Z. and Ying, Z. (1991) ”Estimating a Distribution Function with Truncated and Censored Data”, *The Annals of Statistics*, **19**:417-442.

Lynden-Bell, D. (1971) ”A method for allowing for known observational selection in small samples applied to 3CR quasars”, *Monthly Notices of the Royal Astronomical Society*, **155**:95-118.

MacKenzie, T., Wallace, A. and Weeks, W. (2010) ”Impact of Rural Residence on Survival of Male Veteran Affairs Patients After Age 65”, *Journal of Rural Health*, **26(4)**:318-24.

Martin, E. and Betensky, R.A. (2005) ”Testing Quasi-Independence of Failure and Truncation Times via Conditional Kendall’s Tau”, *Journal of the American Statistical Association*, **100**:484-492.

Pan, W. and Chappell, R. (1999) ”A note on inconsistency of NPMLE of the distribution function from left truncated and case 1 interval censored data”, *Lifetime Data Analysis*, **5**:281-291.

Robins, J.M. and Rotnitzky, A. (1992) ”Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers,” In: Jewell, N., Dietz, K. and Farewell, V. *AIDS Epidemiology-Methodological Issues*, Boston: Birkhauser, 297-331.

Satten, G.A. and Datta, S. (2001) ”Kaplan–Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average”, *American Statistician*, **55(3)**:207-210.

Shen, P. (2003) ”The product-limit estimate as an inverse-probability-weighted average”, *Communincations in Statistics: Theory and Methods*, **32(6)**:1119-1133.

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

Tsai, W.Y. (1990) ”Testing the assumption of independence and truncation time and failure time”, *Biometrika*, **77(1)**:169-177.

Wang, M.C., Jewell, N.P. and Tsai, W.Y. (1986) ”A note on the uniform consistency of the Kaplan-Meier estimator”, *Annals of Statistics*, **14**:1597-1605.

Wang, M.C. (1989) ”A Semiparametric Model for Randomly Truncated Data”, *JASA*, **84(407)**:742-748.

Wang, M.C. (1996) ”Hazards regression analysis for length-biased data”, *Biometrika*, **83(2)**:343-354.

Woodroofe, M. (1985) ”Estimating a distribution function with truncated data”, **13**:163-177.