INFERENCE ON QUANTILE RESIDUAL LIFE FOR LENGTH-BIASED SURVIVAL DATA

Samia H. Lopa⋆

Department of Obstetrics, Gynecology, and Reproductive Sciences
University of Pittsburgh, Pittsburgh, PA, USA
Email: shw30@pitt.edu

Jong-Hyeon Jeong

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA
Email: jjeong@pitt.edu

SUMMARY

Length biased data occurs when a prevalent sampling is used to recruit subject into a study that investigates the time from an initial event to a terminal event. Such data are usually left-truncated and right-censored. While there have been accurate and efficient methods to estimate the survival function, not much work have been done regarding estimation of residual life time distribution or the summary parameters such as the median and quantiles of the residual life. In this paper we propose two ways to estimate the quantiles of the residual life time at fixed time points accounting for the length biased and censored nature of the data. We provide the asymptotic properties of these estimators and investigate them through simulation studies. Considering that the variance of these estimators require density estimation, we suggest an alternate approach to obtain the confidence intervals for the residual function. We apply these methods to a breast cancer dataset from National Surgical Adjuvant Breast and Bowel Project (NSABP).

Keywords and phrases: Left Truncation; Length-Biased Data; Quantiles; Residual Life; Right-Censoring

AMS Classification: 62N99

1 Introduction

In epidemiological studies it is often of interest to study the time between the occurrence of an initial event and a terminating event. For example, in breast cancer studies it is of interest to study time from recurrence to death. In mental health studies it is important to estimate the time between the onset of depression and the remission to understand the disease history and its evolution over time. In the ideal scenario, every subject in the population who had experienced the initial event would be followed until the occurrence of the terminating event and the difference between the times of two events would be computed, which would result in the distribution of the time to terminating event

---

⋆ Corresponding author

from the initiating event. In reality it is not possible to follow each member of the cohort for the occurrence of these paired events as it requires unlimited follow-up time. Two practical strategies are usually taken: incident cohort study and prevalent cohort study.

An incident cohort in epidemiology is formed by individuals who experience the initial events within a specified calendar time interval. These cases are then followed for a further fixed time period until failure, loss to follow-up or end-of-study. For example, a group of women aged 18 or older may be followed for breast cancer occurrence over a period of five years and then those who had breast cancer would be followed until death, recurrence, or termination of the study. A prevalent cohort is formed as a group of individuals who have experienced the initial event but have not experienced the terminating event at the time of recruitment into the study. Women diagnosed with breast cancer and alive during the recruitment period would enter the cohort for follow-up until death, recurrence, or termination of the study.
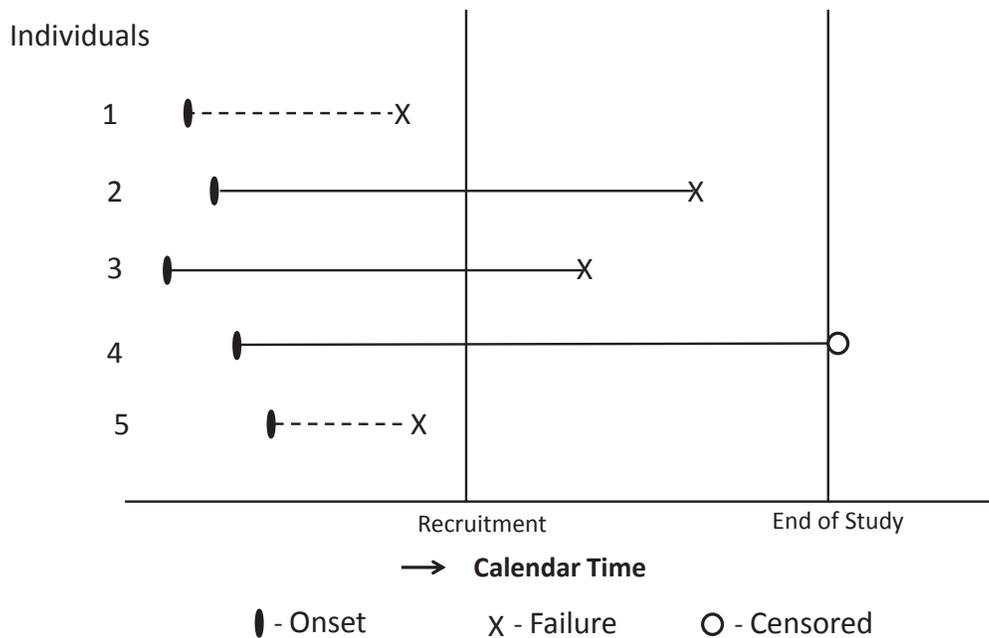

Figure 1: Prevalent cohort study with follow up.

Figure 1 shows the time course of events in a population for five individuals and how a prevalent sampling affects each of them. In a prevalent cohort study, cross-sectional sampling is used to identify the cases first and then followed until failure or censoring. Note that individuals who have already experienced both the initial and failure event will not be eligible to enter the study. Thus the prevalent cohort will be biased against individuals with shorter failure event times. Consider individuals 1 and 5 in Figure 1, both of whom experienced the initial and failure events before the study recruitment and hence will not be enrolled in the cohort. However, individuals 2 and 3 have not

experienced the failure event at the study recruitment and will be enrolled in the cohort. Similarly, patient 4 would be included as the patient had experienced the initial event before recruitment. However this patient becomes right censored at the end of the study. The exclusion of individuals 1 and 5 occurs because of the length of time of the study, and the bias towards the longer times between the two events is referred to as length bias.

As mentioned earlier, in a prevalent sampling, cases have already experienced their initial event. These initiating times are assumed to form a stationary Poisson process and is referred to as stationarity. Under stationarity the incidence rate of the disease is approximately constant over time. Stationarity assumption is not valid when there was an epidemic of disease before the study starts. But it holds in situations where the disease is stable, that is, rate of occurrence of disease (initial event) remains constant over time.

When prevalent sampling is used for recruiting a cohort, left truncation arises. In the presence of left truncation, we can only observe those individuals whose event time is longer than a given time. In addition to left truncation, survival data may also be subject to right censoring. Right censoring occurs when the event occurs after the study ends or a subject leaves the study (loss to follow-up). Thus, if the data are right censored, exact event time is not known, but it is only known to be greater than certain time (e.g. time of last contact).

Failure to account for left-truncation properly results in biased estimation (overestimation) of survival function. Wolfson et al. (2001) reported that when left-truncation was not accounted for, the survival rates for patients with dementia from the onset of the disease was overestimated such that the median survival time was 6.6 years, twice as high compared to the left-truncation adjusted median of 3.3 years. Thus a proper method of analysis is essential for making valid inference from length-biased data. In the last several decades, many authors have addressed the issues of left-truncation and right censoring in survival analysis (Wang et al., 1986; Tsai et al., 1987; Chan et al., 2012; Chan, 2013). Although the concept of truncation product limit estimator was first introduced by Lynden-Bell (1971) to account for left truncation, it has only drawn attention in the 80's. This estimator is the analogue of the product limit estimator of Kaplan and Meier (1958) for randomly censored data, and hence is referred to as truncation product limit estimator. Wang et al. (1986) discussed the asymptotic properties of this estimator. When the distribution of left truncation time is unspecified, a conditional analysis is preferred conditioning on the truncation times (Turnbull, 1976; Wang et al., 1986, 1993; Tsai et al., 1987; Lagakos et al., 1988; Wang, 1991; Wang et al., 1993; Andersen, 1993) as it provides a simple and easy-to-implement expression for the estimator. However, if the onset of a disease follows a stationary Poisson process such that the incidence rate remains constant over time (Wang, 1991), i.e., if the stationarity assumption holds, the truncation time will follow a uniform distribution and an estimator that incorporates the truncation time distribution is generally more efficient (Vardi, 1982, 1985; Gill et al., 1988; Vardi and Zhang, 1992; Asgharian et al., 2002; Asgharian and Wolfson, 2005) than truncation product limit estimator. Vardi (1982, 1985) and Gill et al. (1988) extensively discussed the issues of length-bias and proposed a nonparametric maximum likelihood estimator (NPMLE) of the underlying non-length-biased survival distribution. The unconditional estimator, though more efficient, is not easy to implement as its large-sample properties involve mathematically intractable integral equations (Huang and Qin,

2011). Wang (1987) showed that truncation product limit estimator is the nonparametric maximum likelihood estimator of the full likelihood. Wang (1991) maintains that the truncation product limit estimator can be derived by maximizing the conditional likelihood and it has no information loss when the distribution of the truncation time is not specified. Huang and Qin (2011) provided a new non-parametric estimator incorporating the distribution of truncation times under stationarity assumption. The new estimator has a closed-form expression and almost as efficient as the NPMLE. Shi et al. (2018) showed how to estimate the quantiles from length-biased and right-censored data based on the NPMLE.

Liu et al. (2015) provided NPMLE of the quantile residual lifetime with length-biased and right-censored data. Since it is based on survival distribution estimate which does not have a closed-form expression, computation of this estimator and its asymptotic properties is relatively complicated. Although Liu et al. (2015) approach can be extended to censored length-biased data, but this method does not utilize the information of the truncation variable, making it less efficient than that proposed in Wang et al. (2017).

In this paper we propose non-parametric estimators for the quantile residual life function based on the length-biased survival data. Specifically, we propose two estimators of quantile residual life function based on the survival functions estimated using truncation product limit method and Huang and Qin (2011) method, and show that the latter is more efficient than the QRL function estimated using the truncation product limit estimator. Our method is similar to those described as TJW and HQ methods in Wang et al. (2017), where the variance was estimated using a bootstrap approach. In contrast, we provide explicit variance estimator based on influence functions and implement it. In addition, we will build upon the techniques presented in Jeong et al. (2008) to construct confidence intervals for the quantiles of length-biased right-censored survival data. The paper is organized as follows. In Section 2, we describe the data set-up and notation used in the paper; Section 3 presents the proposed estimators followed by the asymptotic properties of the estimators in Section 4. In Section 5, we present a simulation study to investigate the small- and large-sample performances of our estimators. Section 6 presents the breast cancer data analysis example. We conclude with a few remarks in Section 7.

2   Data set-up and notation

Let $T_1^0, T_2^0, T_3^0, \ldots$ be i.i.d. positive random variables representing the time from disease incidence to failure event. Denote by $S(t)$ and $f(t)$ the survival and probability density function of T^0 (note that we drop the subscript i to represent a generic copy of the i.i.d variables) respectively. Let W^0 denote the calender time of disease incidence and ξ be the potential recruitment time. A prevalent population would include individuals with the disease who have not experienced the failure event at the sampling time, that is, the prevalent population would consist of failure times for which $T^0 \geq \xi - W^0 > 0$. We will use W and T to indicate the 'observed' W^0 and T^0 respectively in the prevalent population.

For stable disease we assume two conditions :

(i) the probability distribution of survival time T^0 is independent of W^0.

(ii) The occurrence rate of disease remains constant over calender time, that is, W^0 has a constant density function.

Let $A = \xi - W$ be the truncation time from disease incidence to sampling time and let $V = T - A$ be the residual survival time from sampling time. Let $f_T(t)$, $f_A(t)$ and $f_V(t)$ be the marginal density functions and $S_T(t)$, $S_A(t)$ and $S_V(t)$ be the survival functions of the corresponding the random variables T, A and V.

When conditions (i) and (ii) are satisfied, (A, T) has the joint density function:

$$f_{A,T}(a, t) = \mu^{-1} f(t) I(t > a > 0), \tag{2.1}$$

where $\mu = E(T^0)$. The marginal distribution of length-biased survival time T is then obtained as

$$f_T(t) = \int_0^t \frac{1}{\mu} f(t) da = \frac{1}{\mu} t f(t) I(t > 0). \tag{2.2}$$

The marginal distribution of truncation time A is then

$$f_A(a) = \int_a^\infty \frac{1}{\mu} f(t) dt = \frac{1}{\mu} S(a) I(a > 0). \tag{2.3}$$

Marginal distribution of $V = T - A$ is exactly same as that of A. The conditional distribution of T given A is

$$f_{(T|A=a)}(t) = \frac{f_{A,T}(t, a)}{f_A(a)} = \frac{\frac{1}{\mu} f(t)}{\frac{1}{\mu} S(a)} = \frac{f(t)}{S(a)}, \tag{2.4}$$

and the corresponding survival function is

$$S_{(T|A=a)}(t) = P(T > t | A = a) = \frac{1}{S(a)} \int_t^\infty f(t) \, dt = \frac{S(t)}{S(a)}. \tag{2.5}$$

The second assumption of occurrence rate of disease being constant over calender time is often referred to as stationary. The consequence of this is that the truncation time A given T is uniformly distributed. This follows from that fact that if a person lives up to time T, then because the initial event occurs at a constant rate, it could occur at any time before T with equal probability. The fact that the marginal distribution of V and A are identical provides us a way to test the stationary assumption. Asgharian et al. (2006) suggested to check this in the data graphically by comparing the estimated distributions (Kaplan-Meier curves) of V and A.

Our interest lies in the statistical inference for the distribution of residual lifetime at a specific timepoint s of the population represented by the survival time T^0. Specifically our objective is to estimate the $100q^{th}$ quantile residual lifetime function $Q(q, s)$ at time s, defined by,

$$P(T^0 - s > Q(q, s) | T^0 > s) = 1 - q, \tag{2.6}$$

where $0 \leq q \leq 1$. Or, equivalently, $Q(q, s)$ is the solution to the equation,

$$\frac{S(s + Q)}{S(s)} = 1 - q. \tag{2.7}$$

The $100q^{th}$ quantile residual lifetime for the prevalent population $Q^*(q, s)$ is defined as,

$$P(T - s > Q^*(q, s)|T > s) = 1 - q,$$
$$\Leftrightarrow \frac{S_T(s + Q^*)}{S_T(s)} = 1 - q,$$

where

$$S_T(s) = \int_s^\infty \frac{tf(t)}{\mu} I(t > 0)dt = \frac{1}{\mu} \int_s^\infty tf(t)dt.$$

Suppose that individual i has censoring time $C_i^* = A_i + C_i$, where C_i is the residual censoring time, the time from recruitment until the individual is censored and $T_i = A_i + V_i$, is the overall survival time, such that V_i is the time from recruitment to failure event. We can observe only $\min(C_i^*, T_i)$, We also assume that C_i is independent of (ξ_i, W_i, T_i). However, total survival time T_i and total censoring time C_i^* have A_i in common and are dependent. Thus total survival time T_i is subject to informative censoring (Vardi, 1989). Informative censoring occurs when censoring time, C_i^* gives more information on the survival time T_i, in addition to the knowledge that $T_i > C_i^*$.

Under right censoring, we observe $(W_i, A_i, \widetilde{V}_i, \Delta_i)$ where $\widetilde{V}_i = \min(V_i, C_i)$ and $\Delta_i = I(V_i \leq C_i)$, $i = 1, \ldots, n$ where n is the number of individuals in the sample. Define $Y_i = A_i + \widetilde{V}_i = \min(T_i, C_i^*)$.

3   Estimation of quantile residual life function

From Eq. (2.7) the $100q^{th}$ quantile of the residual lifetime distribution at a fixed time t_0 can be calculated as a solution to the equation

$$u(Q) = S(t_0 + Q) - (1 - q)S(t_0) = 0. \tag{3.1}$$

Note that for simplicity we have used Q for $Q(q, t_0)$. To estimate Q, one would replace the survival function $S(.)$ by its sample estimate, and solve the same equation for Q. In other words, Q would be estimated by solving the equation $\hat{u}(Q) = 0$ where,

$$\hat{u}(Q) = \hat{S}(t_0 + Q) - (1 - q)\hat{S}(t_0), \tag{3.2}$$

where $\hat{S}(t)$ is a uniformly consistent estimator of the survival function in the prevalent population (Wang, 1991).

Different choices of $\hat{S}(.)$ might lead to different estimates of Q. Here we will consider two estimates of $S(.)$. One is the traditionally used truncation product limit (TPL) estimator by Tsai et al. (1987), and the other is a recently proposed estimator by Huang and Qin (2011).

The TPL estimator of $S(.)$ from the length-biased data is calculated as

$$\hat{S}(t) = \prod_{u \in [0,t]} \{1 - d\hat{\Lambda}(u)\} \tag{3.3}$$

where $\hat{\Lambda}$ is the estimated cumulative hazard function

$$\hat{\Lambda}(t) = \int_0^t \frac{d\bar{N}(u)}{\bar{R}(u)} \tag{3.4}$$

with $\bar{N}(u) = n^{-1} \sum_{j=1}^{n} \Delta_j I(Y_j \leq u)$ and $\bar{R}(u) = n^{-1} \sum_{j=1}^{n} I(Y_j \geq u \geq A_j)$.

Truncation product-limit estimator ignores the information in the marginal distribution of A_j, which can be very inefficient in the situation where the data comes from length biased sample. Huang and Qin (2011) argued that the truncation time A has the same marginal distribution as the residual survival time V under length-biased sampling. They estimate S_A and Λ by combining the information from both A and V which leads to a more efficient estimator than the TPL estimator.

The survival function estimator proposed by Huang and Qin (2011) is given by

$$\hat{S}^*(t) = \prod_{u \in [0,t]} \{1 - d\tilde{\Lambda}(u)\} \tag{3.5}$$

with the corresponding estimated cumulative hazard function defined as

$$\tilde{\Lambda}(t) = \int_0^t \frac{d\bar{N}(u)}{\tilde{R}(u)}, \tag{3.6}$$

where

$$\tilde{R}(u) = n^{-1} \sum_{i=1}^{n} I(Y_i \geq u) - \tilde{S}_A(u),$$

$$\tilde{S}_A(u) = \prod_{t \in [0,u]} \left\{ 1 - \frac{d\tilde{B}(t)}{\tilde{K}(t)} \right\},$$

$$\tilde{B}(t) = n^{-1} \sum_{i=1}^{n} \{I(A_i \leq t) + \Delta_i I(\tilde{V}_i \leq t)\},$$

$$\tilde{K}(t) = n^{-1} \sum_{i=1}^{n} \{I(A_i \geq t) + I(\tilde{V}_i \geq t)\}.$$

Our first proposed estimator uses $\hat{S}(t)$ for $S(t)$ in Equation (3.2) and solves the equation for Q to obtain an estimator $\hat{Q}$ of Q. This means that

$$\hat{S}(t_0 + \hat{Q}) = (1 - q)\hat{S}(t_0),$$

Or, equivalently,

$$\hat{Q} = \hat{S}^{-1}((1 - q)\hat{S}(t_0)) - t_0, \tag{3.7}$$

where $\hat{S}^{-1}(\cdot)$ is the inverse of the estimated survival function defined as,

$$\hat{S}^{-1}(p) = \inf\{t : \hat{S}(t) < p\} \tag{3.8}$$

for $0 < p < 1$. Note that, this implies that $\hat{Q}$ exists only if $\hat{S}^{-1}(p)$ exists at $p = (1 - q)\hat{S}(t_0)$; or equivalently, there exists a t for which $\hat{S}(t) < (1 - q)\hat{S}(t_0)$.

In a similar fashion, we define the estimator $\hat{Q}^*$ of Q. Specifically, $\hat{Q}^*$ is the solution of the equation,

$$\hat{u}^*(Q) = 0,$$

for Q where

$$\hat{u}^*(Q) = \hat{S}^*(t_0 + Q) - (1 - q)\hat{S}^*(t_0),$$

In other words,

$$\hat{Q}^* = [\hat{S}^*((1 - q)\hat{S}^*(t_0))]^{-1} - t_0, \tag{3.9}$$

which exists only when there is at least one t for which $\hat{S}^*(t) < (1 - q)\hat{S}^*(t_0)$.

4   Asymptotic properties and Inference

It has been shown that $\hat{S}(t)$ is uniformly consistent and asymptotically normal (Wang, 1991) for $0 \leq t \leq \tau$, where τ is such that a positive proportion in the population is alive beyond τ. Moreover, this estimator is asymptotically linear such that

$$\sqrt{n}\left\{\hat{S}(t) - S(t)\right\} = n^{-1/2}\sum_{i=1}^{n}\phi_i(t) + o_p(1), \tag{4.1}$$

where

$$\phi_i(t) = \left[\int_0^t R(u)^{-2} I(Y_i \geq u \geq A_i)dF^u(u) - \frac{\Delta_i I(Y_i \leq t)}{R(Y_i)}\right] S(t),$$

and $F^u(t) = \text{pr}(\Delta = 1, Y \leq t)$ is the subdistribution function of complete observations. The variables ϕ_i, $i = 1, 2, \ldots, n$ are i.i.d. with $E(\phi_i) = 0$ and variance covariance function $\Sigma(t_1, t_2) = E\{\phi_i(t_1)\phi_i(t_2)\}$, $0 \leq t_1, t_2 \leq \tau$, and are referred to as the influence function of $\hat{S}(t)$.

Huang and Qin (2011) established that $\hat{S}^*(t)$ is uniformly consistent and asymptotically normal with

$$\sqrt{n}\left\{\hat{S}^*(t) - S(t)\right\} = n^{-1/2}\sum_{i=1}^{n}\phi_i^*(t) + o_p(1), \tag{4.2}$$

where

$$\phi_i^*(t) = \phi_i(t) + \int_0^t R(u)^{-2}\{I(A_i > u) - S_A(u) - S_A(u)\phi_i(u)\}dF^u(u),$$

$i = 1, 2, \ldots, n$, are i.i.d random variable with mean zero and the covariance function

$$\Sigma^*(t_1, t_2) = E\left[\phi_i^*(t_1)\phi_i^*(t_2)\right].$$

Now, the estimator $\hat{Q}$ is a solution to the estimating equation

$$\hat{u}(Q) = 0, \tag{4.3}$$

where $\hat{u}(.)$ is defined in (3.2), with $\hat{S}(t)$ calculated using Equation (3.3).

That is,

$$\hat{u}(Q) = \hat{S}(t_0 + Q) - (1 - q)\hat{S}(t_0). \tag{4.4}$$

Since $\hat{S}(t)$ is uniformly consistent over $t \in [0, \tau]$, $\hat{u}(Q)$ uniformly converges to

$$u(Q) = S(t_0 + Q) - (1 - q)S(t_0) \tag{4.5}$$

for $t \in [0, \tau]$. Since $u(Q_0) = 0$ at the true value $Q_0 = Q(t_0, q)$, the estimated $100q^{th}$ quantile residual life of T^0 at time t_0, $\hat{Q}$ is a consistent estimator of Q_0.

As with all quantile estimators, finding the asymptotic distribution of $\hat{Q}$ through the expansion of $\hat{u}(Q)$ involves estimation of the density function of T, which appears in the denominator of the variance expression. Since density estimators are highly variable and can often take values that are close to zero, the estimated variance of $\hat{Q}$ calculated in this manner is highly unstable. To overcome this Jeong et al. (2008) suggested directly obtaining the confidence interval for Q by inverting a standardized statistic based on $\hat{u}(Q)$. This is a procedure similar to that suggested in Brookmeyer and Crowley (1982) for obtaining confidence interval for the quantiles. Thus, one would need to find the asymptotic distribution of $\hat{u}(Q_0)$.

$$\begin{aligned}
\sqrt{n}\hat{u}(Q_0) &= \sqrt{n}\{\hat{u}(Q_0) - u(Q_0)\} \\
&= \sqrt{n}\{\hat{S}(t_0 + Q_0) - (1-q)\hat{S}(t_0) - S(t_0 + Q_0) + (1-q)S(t_0)\} \\
&= \sqrt{n}\{\hat{S}(t_0 + Q_0) - S(t_0 + Q_0)\} - \sqrt{n}(1-q)\{\hat{S}(t_0) - S(t_0)\}. \quad (4.6)
\end{aligned}$$

Using (4.1),

$$\begin{aligned}
\sqrt{n}\hat{u}(Q_0) &= n^{-\frac{1}{2}}\sum \phi_i(t_0 + Q_0) - n^{-\frac{1}{2}}(1-q)\sum \phi_i(t_0) + o_p(1) \\
&= n^{-\frac{1}{2}}\sum \phi_i^u + o_p(1), \quad (4.7)
\end{aligned}$$

where $\phi_i^u = \phi_i(t_0 + Q_0) - (1-q)\phi_i(t_0)$.

Therefore, $\sqrt{n}\hat{u}(Q_0)$ is asymptotically normally distributed with mean zero and variance

$$\begin{aligned}
V_0(Q_0) &= \text{var}(\phi_i^u) \\
&= \text{var}[\phi_i(t_0 + Q_0) - (1-q)\phi_i(t_0)] \\
&= \Sigma(t_0 + Q_0, t_0 + Q_0) + (1-q)^2\Sigma(t_0, t_0) - 2(1-q)\Sigma(t_0 + Q_0, t_0). \quad (4.8)
\end{aligned}$$

For known Q_0, this variance can be estimated by

$$\hat{V}_0(Q_0) = \hat{\Sigma}(t_0 + Q_0, t_0 + Q_0) + (1-q)^2\hat{\Sigma}(t_0, t_0) - 2(1-q)\hat{\Sigma}(t_0 + Q_0, t_0),$$

where

$$\hat{\Sigma}(t_1, t_2) = \frac{1}{n}\Sigma_{i=1}^n \hat{\phi}_i(t_1)\hat{\phi}_i(t_2)$$

with

$$\hat{\phi}_i(t) = \hat{S}(t)\left[\int_0^t \frac{I(Y_i \geq u \geq A_i)d\bar{N}(u)}{\tilde{R}^2(u)} - \frac{\Delta_i I(Y_i \leq t)}{\tilde{R}(Y_i)}\right].$$

A $100(1-\alpha)\%$ confidence interval for Q can then be constructed by inverting the wald statistic based on the asymptotic distribution of $\sqrt{n}\hat{u}(Q_0)$. More explicitly, a $100(1-\alpha)\%$ confidence interval for Q is given by,

$$\{Q : [\hat{V}_0(Q)]^{-1}n[\hat{u}(Q)]^2 < \chi_{1,1-\alpha}^2\}, \quad (4.9)$$

where $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)^{th}$ percentile of a χ^2 distribution with 1 degree of freedom.

A similar argument can be used to construct confidence intervals for Q using the other estimator $\hat{Q}^*$. Explicitly, a $100(1-\alpha)\%$ confidence interval for Q using this method is given by,

$$\{Q : [\hat{V_0}^*(Q)]^{-1} n[\hat{u}^*(Q)]^2 < \chi^2_{1,1-\alpha}\},$$

where

$$\hat{u}^*(Q) = \hat{S}^*(t_0 + Q) - (1-q)\hat{S}^*(t_0),$$
$$\hat{V_0}^*(Q) = \hat{\Sigma}^*(t_0 + Q, t_0 + Q) + (1-q)^2\hat{\Sigma}^*(t_0, t_0) - 2(1-q)\hat{\Sigma}^*(t_0 + Q, t_0),$$
$$\hat{\Sigma}^*(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_i^{\ *}(t_1)\hat{\phi}_i^{\ *}(t_2),$$

with

$$\hat{\phi}_i^{\ *}(t) = \hat{S}^*(t)\left[\int_0^t \frac{I(Y_i \geq u \geq A_i)d\bar{N}(u)}{\tilde{R}^2(u)} - \frac{\Delta_i I(Y_i \leq t)}{\tilde{R}(Y_i)}\right.$$
$$\left. + \int_0^t \left\{\frac{I(A_i > u)}{\tilde{R}^2(u)} - \tilde{S}_A(u) - \tilde{S}_A(u)\tilde{\psi}_i(u)\right\}d\bar{N}(u)\right],$$

where

$$\tilde{\psi}_i(u) = \int_0^t \frac{\{I(A_i \geq u) + I(\tilde{V}_i \geq u)\}d\tilde{B}(u)}{\tilde{K}^2(u)} - \frac{I(A_i \leq t)}{\tilde{K}(A_i)} - \frac{\Delta_i I(\tilde{V}_i \leq t)}{\tilde{K}(\tilde{V}_i)}.$$

Further details of the asymptotic properties are described in Lopa (2015).

5   Simulation Study

We conducted a simulation study to evaluate the finite-sample performance of the two estimators of the quantile residual life functions discussed in the previous sections, namely, $\hat{Q}^*$ calculated based on the Huang and Qin method using Equation (3.9), and $\hat{Q}$, calculated based on the truncation product-limit estimator using Equation (3.7). We generated 1000 Monte-Carlo samples of sizes 200 or 400 from the target population. The population parameters were chosen according to the scenarios described in Huang and Qin (2011). More explicitly, we first set the sampling time (recruiting time) ξ to be 100. The time of disease onset W^0 was generated from a uniform distribution over the interval $[0, 100]$. Survival time T^0 was independently generated from a Weibull distribution with the survival function $S(t) = \exp(-t^2/4)$. To form a prevalent cohort of the desired sample size n, the pair (W^0, T^0) were generated repeatedly until there were n pairs of observations satisfying the sampling constraint $W^0 + T^0 \geq \xi$. The residual censoring time C from enrollment to loss to follow-up was generated from a uniform distribution with a support $(1, 2)$ or $(0, 2)$ resulting respectively in 30% and 51% censoring rates.

The true $100q^{th}$ quantile residual life function at time t_0 for the Weibull distribution above is given by

$$\exp\{-(t_0 + Q(q, t_0))^2/4\} - (1-q) \times \exp(-t_0^2/4) = 0,$$

which is equivalent to the positive root of the quadratic equation

$$Q^2(q, t_0) + 2Q(q, t_0)t_0 + 4\ln(1 - q) = 0.$$

This leads us to the true $100q^{th}$ quantile residual life function at time t_0 as

$$Q(q, t_0) = \sqrt{t_0^2 - 4\ln(1 - q)} - t_0. \tag{5.1}$$

Figure 2 shows the true QRL function for various values of q, specifically, the .25-, .50-, and .75- quantile residual life function.

We show the numerical results for estimating these residual life functions at fixed time points 0.0, 0.5, and 1.0. We present the bias, average length and coverage probability of 95% confidence intervals for the two estimates, $\hat{Q}$ and $\hat{Q}^*$.
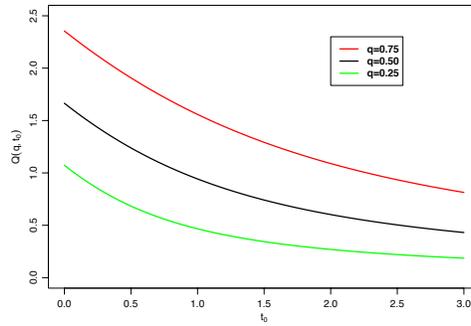

Figure 2: Graphical representation of the true quantile residual life function for various values of q.

Table 1 summarizes the mean of the 75th quantile residual life function, bias, average length of 95% confidence interval, coverage probability of 95% confidence interval for both estimators at different time points $(0.0, 0.5, 1.0)$ for two sample sizes $n = (200, 400)$ based on 1000 Monte-Carlo samples. The true 75th quantile residual life functions at the selected time points $(0.0, 0.5, 1.0)$ are respectively $(2.35, 1.91, 1.56)$. When censoring was moderate (30%), both estimators were approximately unbiased with bias being relatively smaller for the $\hat{Q}^*$ for the larger sample size of 400; the bias ranged from -0.0032 to 0.0013 for the $\hat{Q}$ and -0.0006 to 0.0064 for the $\hat{Q}^*$. In general, the coverage probabilities of 95% confidence intervals met the nominal level, specifically for the larger sample size of 400; $(91.6\% - 94.5\%)$ for $\hat{Q}$ and $(93.0\% - 95.1\%)$ for the $\hat{Q}^*$. Average length of these confidence intervals was generally wider for $\hat{Q}$. For example, with $n = 200$ the average length of the 95% confidence intervals for $Q(0.75, 0)$ was 0.47 for $\hat{Q}$ while that for the $\hat{Q}^*$ was 0.41, showing a 12.8% reduction. This reduction in the length of the confidence interval is equivalent to increase in efficiency of the corresponding $\hat{Q}^*$. When the sample size was increased from 200 to 400, the confidence intervals became narrower as shown by the decrease in the average length of the intervals for both methods, even though $\hat{Q}^*$ still resulted in narrower confidence intervals with 9.4% to 15.4% reduction in average length from that of $\hat{Q}$. This is equivalent to an efficiency gain of 21.8% to 39.7% for $\hat{Q}^*$ over the $\hat{Q}$.

Table 1: Simulation results for estimating the 75th quantile residual life function depicted in Figure 2. EST = Monte Carlo mean of the estimator, Bias = absolute bias of the estimator, AL = Monte Carlo average of the length of 95% confidence interval, CP = Coverage probability of 95% confidence interval.

% cen	t_0	n	$Q(0.75, t_0)$	$\hat{Q}$	Bias	AL	CP	$\hat{Q}^*$	Bias	AL	CP
	0.0	200	2.35	2.35	-0.0003	0.47	91.6	2.36	0.0059	0.41	93.7
30%	0.5		1.91	1.91	0.0013	0.45	91.7	1.91	0.0064	0.40	93.0
	1.0		1.56	1.56	0.0013	0.45	92.0	1.56	0.0040	0.40	93.6
	0.0	400	2.35	2.35	-0.0032	0.33	92.5	2.36	0.0004	0.29	93.6
	0.5		1.91	1.91	-0.0016	0.32	94.5	1.91	0.0006	0.28	93.7
	1.0		1.56	1.56	-0.0027	0.32	94.1	1.56	-0.0006	0.29	95.1
	0.0	200	2.35	2.36	0.0005	0.56	93.1	2.36	0.0039	0.48	93.1
51%	0.5		1.91	1.91	0.0027	0.55	92.5	1.91	0.0066	0.47	93.1
	1.0		1.56	1.56	0.0062	0.55	92.9	1.57	0.0086	0.49	95.1
	0.0	400	2.35	2.35	-0.0038	0.40	92.4	2.35	-0.0001	0.34	93.1
	0.5		1.91	1.91	-0.0018	0.39	92.2	1.91	0.0009	0.33	94.0
	1.0		1.56	1.56	-0.0022	0.39	93.3	1.56	-0.0008	0.35	93.8

Table 2: Simulation results for estimating the $.50^{th}$ quantile residual life function depicted in Figure 2. EST = Monte Carlo mean of the estimator, Bias = absolute bias of the estimator, AL = Monte Carlo average of the length of 95% confidence interval, CP = Coverage probability of 95% confidence interval.

% Cen	t_0	n	$Q(0.50, t_0)$	$\hat{Q}$	Bias	AL	CP	$\hat{Q}^*$	Bias	AL	CP
	0.0	200	1.67	1.66	-0.0004	0.45	92.3	1.67	0.0034	0.40	93.7
30%	0.5		1.24	1.24	0.0024	0.39	93.5	1.24	0.0059	0.36	94.3
	1.0		0.94	0.94	0.0016	0.35	92.2	0.94	0.0026	0.33	93.9
	0.0	400	1.67	1.66	-0.0040	0.31	92.4	1.67	0.0008	0.28	93.2
	0.5		1.24	1.24	0.0004	0.28	93.0	1.24	0.0024	0.25	92.8
	1.0		0.94	0.94	-0.0029	0.25	91.9	0.94	-0.0009	0.23	92.4
	0.0	200	1.67	1.67	0.0023	0.51	92.6	1.67	0.0049	0.46	93.1
51%	0.5		1.24	1.24	0.0042	0.47	93.9	1.25	0.0065	0.42	93.8
	1.0		0.94	0.95	0.0029	0.43	92.4	0.95	0.0053	0.39	93.9
	0.0	400	1.67	1.66	-0.0037	0.36	91.9	1.67	0.0010	0.32	93.8
	0.5		1.24	1.24	0.00003	0.33	92.0	1.24	0.0009	0.29	92.8
	1.0		0.94	0.94	-0.0011	0.30	91.9	0.94	0.0007	0.27	92.3

Table 3: Simulation results for estimating the 25th quantile residual life function depicted in Figure 2. EST = Monte Carlo mean of the estimator, Bias = absolute bias of the estimator, AL = Monte Carlo average of the length of 95% confidence interval, CP = Coverage probability of 95% confidence interval.

%Cen	t_0	n	$Q(0.25, t_0)$	$\hat{Q}$	Bias	AL	CP	$\hat{Q}^*$	Bias	AL	CP
	0.0	200	1.07	1.07	-0.0013	0.49	90.8	1.08	0.0042	0.47	93.1
30%	0.5		0.68	0.69	0.0049	0.37	91.4	0.69	0.0074	0.36	91.6
	1.0		0.47	0.47	0.0024	0.29	93.8	0.47	0.0041	0.27	93.9
	0.0	400	1.07	1.07	-0.0006	0.35	92.1	1.08	0.0026	0.34	92.8
	0.5		0.68	0.69	0.0034	0.27	91.4	0.69	0.0048	0.25	91.2
	1.0		0.47	0.47	-0.0011	0.20	93.4	0.47	-0.0001	0.19	93.6
	0.0	200	1.07	1.07	0.0016	0.53	92.2	1.08	0.0053	0.50	93.2
51%	0.5		0.68	0.69	0.0063	0.42	91.4	0.69	0.0068	0.40	91.9
	1.0		0.47	0.47	0.0039	0.34	93.6	0.47	0.0051	0.33	93.8
	0.0	400	1.07	1.07	-0.0005	0.38	93.5	1.08	0.0028	0.36	93.2
	0.5		0.68	0.69	0.0035	0.30	92.2	0.69	0.0050	0.28	92.4
	1.0		0.47	0.47	0.0007	0.24	93.6	0.47	0.0022	0.23	93.4

When the censoring rate was increased to 51%, the biases were increased for both estimators but they remained small. With the increased censoring rate, $\hat{Q}^*$ almost always provided better coverage for 95% confidence interval for both smaller (200) and larger (400) sample sizes, compared to the $\hat{Q}$.

Table 2 shows the results for estimating the median residual lifetimes at times $(0.0, 0.5, 1.0)$. The results are similar to those described for the 75th quantile estimators. Specifically both $\hat{Q}$ and $\hat{Q}^*$ are approximately unbiased and maintain the nominal coverage probability at moderate to extreme levels of censoring and sample sizes, even though $\hat{Q}^*$ remains to be more efficient (have narrower confidence intervals). Also, the coverage probabilities are almost uniformly better for $\hat{Q}^*$ compared to $\hat{Q}$. The same trend followed for the 25th percentile residual estimator [Table 3].

In summary, the two proposed estimators of quantile residual lifetime provide reasonable estimates with $\hat{Q}^*$ being more efficient than $\hat{Q}$, in general.

6   Analysis of NSABP B-20 Data

In this section we illustrate our method using a dataset from the National Surgical Adjuvant Breast and Bowel Project (NSABP) protocol B-20. Protocol B-20 was designed to determine whether addition of chemotherapy to tamoxifen would improve the beneficial effect of tamoxifen in the treatment of estrogen receptor (ER) - positive patients with axillary lymph node-negative breast cancer. Women at participating NSABP clinical centers in USA and Canada who had primary

breast cancer, histologically node-negative and ER - positive breast cancer were eligible for this trial. Patients fulfilling eligibility criteria underwent surgery (total mastectomy and lymph node dissection or lumpectomy and lymph node dissection followed by breast irradiation). Following surgery, patients providing written consent to participate in the trial were randomized to one of the three treatment groups: tamoxifen(T) alone (TAM), T plus sequential methotrexate (M) and fluorouracil (F) (MFT), T plus cyclophosphamide (C), M, and F (CMFT) stratified by age, tumor size, and tumor ER level. Between Oct 17, 1988, and March 5, 1993, a total of 2363 women were enrolled in this study (788 randomly assigned to TAM, 786 to MFT, and 789 to CMFT). Earlier findings from this clinical trial (Fisher et al., 1997) analyzing 5-year follow-up data showed that addition of chemotherapy to tamoxifen significantly improved the disease-free survival (DFS) rate by at least 4%, MFT vs. TAM (90% vs. 85%, p = .01) and CMFT vs. TAM (89% vs. 85%, p = .001). Chemotherapy plus Tamoxifen groups (MFT and CMFT) performed significantly better in DFS than tamoxifen alone. Similar results were observed for distant disease-free survival and overall survival. In a follow-up publication, Fisher et al. (2004) showed that CMFT-treated women had significantly better recurrence-free survival (89% vs. 79%, p<0.0001) and better overall survival (87% vs. 83%, p = 0.063) than women treated with tamoxifen alone over 12 years of follow-up.

This difference might prompt one to characterize the pattern of survival among patients with recurrence. Recurrence of cancer is not uncommon among breast cancer patients treated with tamoxifen. In the B-20 trial, over 18 years of follow-up, it has been observed that 17% of the women experienced a recurrence before death. Thus it might be of interest to estimate the residual survival following a recurrence in histologically node-negative and ER-positive breast cancer patients who are being treated with tamoxifen with or without chemotherapy. Thus, our initial event will be recurrence, and the primary endpoint will be death.

A total of 788 women were randomly assigned to receive Tamoxifen. Out of these 788 women, 170 women experienced a recurrence. To construct a length-biased sample of patients, we identified the date of recurrence of the first patient who experienced recurrence and set the recruitment time to be at 5 years following the first recurrence date. By the recruitment date, 120(15%) patients had died or lost-to-follow-up and hence would not be eligible to be in the sample. (Alternatively, a window can be considered for recruitment, but for simplicity, we will assume that all patients are recruited at once.) Thus, our length-biased sample consists of 50 patients who had recurrence prior to September 17, 1994 and were still being followed at the same date. We will estimate the quantile residual lifetimes of these patients.

Table 4 presents the estimated 25th percentile and median for the residual life times for patients who would experience recurrence following treatment with tamoxifen based on the NSABP B-20 data at times 0, 0.5 and 1 years after recurrence. We also provide 95% confidence intervals for both estimates. For this specific dataset the TPL estimates of quantile residuals ($\hat{Q}$) were generally larger than $\hat{Q}^*$, which employs the distribution of truncation time. For example, at the time of recurrence, the estimated median residual lifetime estimated by $\hat{Q}^*$ is approximately 2.4 years compared to 3.1 years for $\hat{Q}$. At year 1 the median residual life time (95% CI) for patients with recurrence after tamoxifen treatment is approximately 1.36 $(0.42, 3.34)$ years by $\hat{Q}^*$ compared to 2.11 $(0.49, 9.32)$ years by $\hat{Q}$. Similar results follow for the 25th percentile.

In Figure 3, we present the estimated median residual lifetimes and their pointwise confidence intervals for the NSABP B-20 patients experiencing recurrence. Consistent with the result shown in Table 4, $\hat{Q}^*$ estimates are uniformly smaller than the $\hat{Q}$ estimates. Median residual lifetime first decreases until between 1 to 1.2 years after recurrence and then increases. This could potentially be the effect of post-recurrence treatment which the patients may have taken outside the protocol. The confidence intervals are wider for $\hat{Q}$ estimators compared to $\hat{Q}^*$ estimators, as expected from the simulation results.

Table 4: Estimated quantile residual lifetimes and 95% confidence interval in Tamoxifen arm from NSABP B-20 data

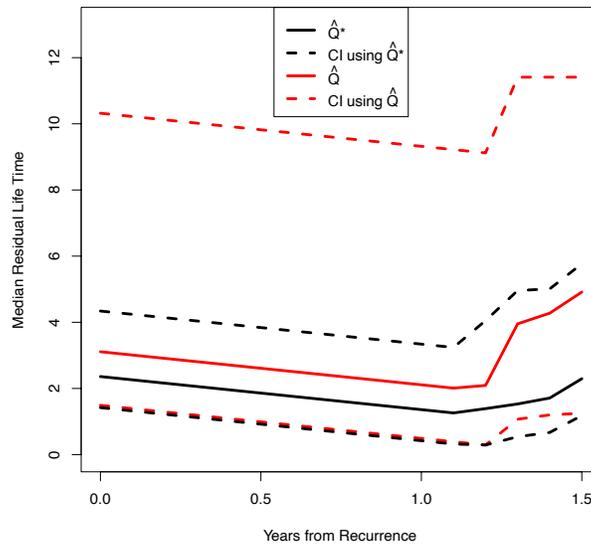| q	t_0	$\hat{Q}(q,t_0)$	95% CL	$\hat{Q}^*(q,t_0)$	95% CL
	0.0	1.48	(1.27, 3.29)	1.37	(1.26, 2.59)
.25	0.5	0.98	(0.77, 2.79)	0.87	(0.76, 2.09)
	1.0	0.48	(0.27, 2.29)	0.37	(0.26, 1.59)
	0.0	3.11	(1.49, 10.32)	2.36	(1.42, 4.34)
.50	0.5	2.61	(0.99, 9.82)	1.86	(0.92, 3.84)
	1.0	2.11	(0.49, 9.32)	1.36	(0.42, 3.34)

Figure 3: Estimated median residual lifetimes and 95% Confidence Intervals (CI) for time to death from recurrence for NSABP B-20 data.

7   Discussion

Length biased data are common in epidemiological prevalent cohort studies where time between an initial event and a terminal event is of interest. In this paper we have proposed two different approaches to estimate the length bias corrected quantile residual life function for a prevalent sample. The first estimator is based on the truncation product limit estimator (Wang et al., 1986) and the second is based on the Huang and Qin (2011) survival estimator. We have developed the asymptotic inference procedures for the two estimators and demonstrated how the confidence interval for QRL can be constructed by inverting a proposed test statistic.

Our methods are similar to those described as TJW and HQ methods in Wang et al. (2017) where variance of these methods were estimated using a bootstrap approach. In contrast, we have provided explicit variance estimators based on influence functions and implemented it in the data analysis and simulation. In addition, we have shown how to construct confidence intervals for the quantiles of length-biased right-censored survival data. Moreover, bootstrap-based inference did not provide good coverage when the censoring rates were larger [30%, maximum considered in Wang et al. (2017)]. Our influence-function-based variance estimation maintained approximately nominal coverage when censoring rate was as high as 51%.

Simulation results showed that both methods empirically lead to consistent QRL estimators, but the QRL estimators based on the Huang and Qin method is more efficient than TPL estimators as shown by the narrower confidence intervals for the former compared to the latter. Both estimators maintained nominal coverage level, though the coverage often being smaller than expected.

We have illustrated our methods with a breast cancer dataset, estimating the quantiles of residual lifetime for breast cancer patients experiencing a recurrence.

Our method has some limitations. First, the estimating equations needed to be solved numerically using grid search approach, which might take longer time when the range of the survival times is large. Second, inverting the test statistic to get the confidence interval is also done numerically, which might also require long time in some applications. Third, our study did not investigate the impact of truncation rate on the inference. Future research can investigate how search algorithms can be improved to efficiently use computing resources to calculate the estimators and confidence intervals. More research also is needed to investigate the impact of truncation rate.

8   Acknowledgment

References

Andersen, P. (1993), *Statistical Models Based on Counting Processes*, Springer Series in Statistics, Springer-Verlag GmbH.

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

Asgharian, M. and Wolfson, D. B. (2005), "Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data," *The Annals of Statistics*, 33, 2109–2131.

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

Brookmeyer, R. and Crowley, J. (1982), "A confidence interval for the median survival time," *Biometrics*, 29–41.

Chan, K. C. G. (2013), "Survival analysis without survival data: connecting length-biased and case-control data," *Biometrika*, 100, 764–770.

Chan, K. C. G., Chen, Y. Q., and Di, C.-Z. (2012), "Proportional mean residual life model for right-censored length-biased data," *Biometrika*, 99, 995–1000.

Fisher, B., Dignam, J., Emir, B., Bryant, J., DeCillis, A., Wolmark, N., Wickerham, D. L., Dimitrov, N. V., Abramson, N., Atkins, J. N., et al. (1997), "Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer," *Journal of the National Cancer Institute*, 89, 1673–1682.

Fisher, B., Jeong, J.-H., Bryant, J., Anderson, S., Dignam, J., Fisher, E. R., and Wolmark, N. (2004), "Treatment of lymph-node-negative, oestrogen-receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials," *The Lancet*, 364, 858–868.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988), "Large sample theory of empirical distributions in biased sampling models," *The Annals of Statistics*, 1069–1112.

Huang, C.-Y. and Qin, J. (2011), "Nonparametric estimation for length-biased and right-censored data," *Biometrika*, 98, 177–186.

Jeong, J.-H., Jung, S.-H., and Costantino, J. P. (2008), "Nonparametric inference on median residual life function," *Biometrics*, 64, 157–163.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, 53, 457–481.

Lagakos, S., Barraj, L., and De Gruttola, V. (1988), "Nonparametric analysis of truncated survival data, with application to AIDS," *Biometrika*, 75, 515–523.

Liu, P., Wang, Y., and Zhou, Y. (2015), "Quantile residual lifetime with right-censored and length-biased data," *Annals of the Institute of Statistical Mathematics*, 67, 999–1028.

Lopa, S. H. (2015), "Inference on quantile residual life for length-biased survival data," Ph.D. thesis, University of Pittsburgh.

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

Shi, J., Ma, H., and Zhou, Y. (2018), "The nonparametric quantile estimation for length-biased and right-censored data," *Statistics & Probability Letters*, 134, 150–158.

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

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

Vardi, Y. (1982), "Nonparametric estimation in the presence of length bias," *The Annals of Statistics*, 10, 616–620.

— (1985), "Empirical distributions in selection bias models," *The Annals of Statistics*, 178–203.

Vardi, Y. and Zhang, C.-H. (1992), "Large sample study of empirical distributions in a random-multiplicative censoring model," *The Annals of Statistics*, 1022–1039.

Wang, M. C. (1987), "Product limit estimates: A generalized maximum likelihood study," *Communications in Statistics-Theory and Methods*, 16, 3117–3132.

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

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993), "Statistical models for prevalent cohort data," *Biometrics*, 1–11.

Wang, M.-C., Jewell, N. P., and Tsai, W.-Y. (1986), "Asymptotic properties of the product limit estimate under random truncation," *the Annals of Statistics*, 1597–1605.

Wang, Y., Zhou, Z., Zhou, X.-H., and Zhou, Y. (2017), "Nonparametric and semiparametric estimation of quantile residual lifetime for length-biased and right-censored data," *Canadian Journal of Statistics*, 45, 220–250.

Wolfson, C., Wolfson, D. B., Asgharian, M., M'Lan, C. E., Østbye, T., Rockwood, K., and Hogan, D. (2001), "A reevaluation of the duration of survival after the onset of dementia," *New England Journal of Medicine*, 344, 1111–1116.