Biostatistics (2015), **16**, 4, *pp.* 785–798
doi:10.1093/biostatistics/kxv012
Advance Access publication on March 21, 2015

Semiparametric likelihood inference for left-truncated and right-censored data

CHIUNG-YU HUANG*

Sidney Kimmel Comprehensive Cancer Center and Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

cyhuang@jhu.edu

JING NING

The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

JING QIN

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

SUMMARY

This paper proposes a new estimation procedure for the survival time distribution with left-truncated and right-censored data, where the distribution of the truncation time is known up to a finite-dimensional parameter vector. The paper expands on the Vardis multiplicative censoring model (Vardi, 1989. Multiplicative censoring, renewal processes, deconvolution and decreasing density: non-parametric estimation. *Biometrika* **76**, 751–761), establishes the connection between the likelihood under a generalized multiplicative censoring model and that for left-truncated and right-censored survival time data, and derives an Expectation–Maximization algorithm for model estimation. A formal test for checking the truncation time distribution is constructed based on the semiparametric likelihood ratio test statistic. In particular, testing the stationarity assumption that the underlying truncation time is uniformly distributed is performed by embedding the null uniform truncation time distribution in a smooth alternative (Neyman, 1937. Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 150–199). Asymptotic properties of the proposed estimator are established. Simulations are performed to evaluate the finite-sample performance of the proposed methods. The methods and theories are illustrated by analyzing the Canadian Study of Health and Aging and the Channing House data, where the stationarity assumption with respect to disease incidence holds for the former but not the latter.

Keywords: Biased sampling; Cross-sectional studies; Prevalent sampling; Profile likelihood; Smooth tests of goodness of fit.

*To whom correspondence should be addressed.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

1. INTRODUCTION

Incident and prevalent cohort study designs are two primary approaches for collecting survival data in observational studies. When it is not feasible to conduct an incident cohort study because of limited resources or other constraints, a prevalent cohort study is a good alternative. Under a prevalent cohort study design, only individuals who have the disease of interest but have not yet experienced the failure event are enrolled, and the observed survival times are subject to left truncation in addition to the usual right censoring because those who have experienced the failure event before the recruitment time are not observable, and those who are recruited may not experience the failure event before the end of the study (Zelen and Feinleib, 1969; Lagakos *and others*, 1988). As a result, the observed survival times are a biased sample of the survival times that occur in the target population, as the sampling scheme favors subjects with slower disease progression. Statistical methods that fail to account for left truncation usually lead to substantial overestimation of the survival time.

Under the stationarity assumption that the incidence of disease onset is constant over time, the truncation time variable follows a uniform distribution, and, as shown in Vardi (1989), the likelihood for survival data that arise in a prevalent cohort study is proportional to the likelihood for data that are subject to multiplicative censoring. Hence, the Expectation–Maximization (EM) algorithm developed for the multiplicative censoring models, which has been shown to be fully efficient in Vardi and Zhang (1992) and Asgharian *and others* (2002), can be readily applied to analyze left-truncated and right-censored survival data when the stable disease condition holds. The stationarity assumption, however, can be easily violated in prevalent cohort studies. For example, in the event of an infectious disease outbreak, the number of people infected usually grows exponentially rather than linearly over time. Hence, the truncation time is unlikely to be uniformly distributed. The existing methods for non-parametric estimation of left-truncated and right-censored data have been based on the conditional likelihood, conditioning on the observed truncation time (Lynden-Bell, 1971; Wang, 1991; Tsai *and others*, 1987) so that information about the truncation time distribution is not required in the estimation procedure. Although knowing the censoring distribution does not provide additional information for estimating the survival time distribution using censored data, as pointed out in Wang (1991), the efficiency of the estimator can be improved substantially for truncated data if the truncation time distribution can be parameterized or fully specified. When the distribution of the truncation time is completely specified, Mandel (2007) generalized Vardi's EM algorithm to obtain the non-parametric maximum likelihood estimator for the survival time distribution. When the distribution of the truncation time variable is known up to a parameter, Shen (2007, 2009) applied the pseudo-profile likelihood procedure (Severini and Wong, 1992) to replace the nuisance parameters; that is, the distribution function of the survival times in the marginal likelihood of the truncation times with a consistent estimator that depends on the parameter of interest. As illustrated later, the resulting estimator is not the maximum likelihood estimator, and hence is not fully efficient. Moreover, the convergence of the iteration algorithm lacks rigorous justifications.

This paper is organized as follows. In Section 2, we study a generalized multiplicative censoring model and develop an EM algorithm to obtain the maximum likelihood estimator for the expanded model. In Section 3.1, we investigate a semiparametric truncation model where the distribution of truncation variable is parameterized up to an unknown parameter. We show in Section 3.2 that, with proper reparameterization, the semiparametric maximum likelihood estimator for the failure time distribution and the truncation time distribution can be easily obtained by employing the EM algorithm developed under a generalized multiplicative censoring model. Asymptotic properties of the proposed estimator are established. To assess whether the stationarity assumption of a constant incidence rate holds for the occurrence of the initiating event, in Section 3.3 we propose a semiparametric likelihood ratio test by embedding the null truncation time distribution in a smooth alternative (Neyman, 1937). In Section 4, simulation studies show that the proposed semiparametric maximum likelihood estimator and the semiparametric likelihood ratio test work

well. We also apply the proposed methodology to two data sets: one from the Canadian Study of Health and Aging and the other is the Channing House data. We show that the stationarity assumption holds for the former but not the latter. A discussion concludes in Section 5.

2. The generalized multiplicative censoring model

In this section, we propose a generalized multiplicative censoring model. Consider non-negative random variable X with density function $g(t)$. Let U be a uniform[0, 1] random variable independent of X. Let $h(t, \boldsymbol{\theta})$ be a density function, where $\boldsymbol{\theta}$ belongs to a compact set Θ in $\mathbb{R}^p$. Define the random variable $Z = H^{-1}\{U \cdot H(X, \boldsymbol{\theta}), \boldsymbol{\theta}\}$ with $H(t, \boldsymbol{\theta}) = \int_0^t h(u, \boldsymbol{\theta})\, \mathrm{d}u$ and $H^{-1}(t, \boldsymbol{\theta}) = \min\{u : H(u, \boldsymbol{\theta}) \geqslant t\}$. Here Z is referred to as subject to the generalized multiplicative censoring. Assume that g and h have support on $[0, \tau]$. In the special case where H is the distribution function of the uniform $[0, \tau]$ random variable, we have $Z = UX$, hence the model reduces to the multiplicative censoring model considered by Vardi (1989). Moreover, given $X = x$, Z has a conditional density function $h(z, \boldsymbol{\theta})/H(x, \boldsymbol{\theta})$ for $z \leqslant x \leqslant \tau$, hence the marginal density function of Z can be shown to be $\int_{u \geqslant z} \{h(z, \boldsymbol{\theta})/H(u, \boldsymbol{\theta})\}\, \mathrm{d}G(u)$, where $G(t) = \int_0^t g(u)\, \mathrm{d}u$.

Define $Y = \Delta X + (1 - \Delta)Z$, where Δ is a binary indicator independent of (X, U). Thus $Y = X$ if $\Delta = 1$, and $Y = H^{-1}\{U \cdot H(X, \boldsymbol{\theta}), \boldsymbol{\theta}\}$ if $\Delta = 0$. Let $(y_1, \delta_1), \ldots, (y_n, \delta_n)$ be n independently realizations of (Y, Δ). The full likelihood function under the generalized multiplicative censoring model is proportional to

$$\mathcal{L}_{MC}(G, \boldsymbol{\theta}) = \prod_{i=1}^{n} g(y_i)^{\delta_i} \left\{ \int_{u \geqslant y_i} H(u, \boldsymbol{\theta})^{-1}\, \mathrm{d}G(u) \right\}^{1-\delta_i} h(y_i, \boldsymbol{\theta})^{1-\delta_i}. \tag{2.1}$$

Intuitively, one can estimate G and $\boldsymbol{\theta}$ by applying the profile likelihood method.

First, for a fixed $\boldsymbol{\theta}$, we derive the EM algorithm to obtain the non-parametric maximum likelihood estimator for G. Let $\{t_1, \ldots, t_L\}$ be the ordered and distinct values of $\{y_1, \ldots, y_n\}$. Define $\xi_j = \sum_{i=1}^n \delta_i I(y_i = t_j)$ and $\eta_j = \sum_{i=1}^n (1 - \delta_i) I(y_i = t_j)$. Thus ξ_j is the number of complete observations at t_j and η_j is the number of multiplicatively censored observations at t_j. Applying a similar argument as in Vardi (1989) and Mandel (2007), we can show that, for a fixed $\boldsymbol{\theta}$, the problem of maximizing (2.1) is equivalent to maximizing

$$\prod_{j=1}^{L} p_j^{\xi_j} \left\{ \sum_{k=j}^{L} H(t_k, \boldsymbol{\theta})^{-1} p_k \right\}^{\eta_j},$$

subject to the constraints $p_j \geqslant 0$, $j = 1, \ldots, L$, and $\sum_{j=1}^L p_j = 1$, where p_j is the jump size of G at t_j. Note that the log-likelihood for the complete data $(x_1, \ldots, x_n)$ is given by $\sum_{j=1}^L \sum_{i=1}^n I(x_i = t_j) \log p_j$. It follows from the result that the conditional density function of X given $Z = z$ is

$$\frac{H(x, \boldsymbol{\theta})^{-1} g(x)}{\int_{u \geqslant z} H(u, \boldsymbol{\theta})^{-1}\, \mathrm{d}G(u)};$$

given the current estimated probabilities $\boldsymbol{p}^{old} = \{p_j^{old}, j = 1, \ldots, L\}$ and the observed value $Z_i = z_i$, the conditional expectation is given by

$$v_l = E\{I(x_i = t_l) \mid z_i, p^{old}\} = H(t_l, \boldsymbol{\theta})^{-1} p_l^{old} \left\{ \sum_{t_k \geqslant z_i} H(t_k, \boldsymbol{\theta})^{-1} p_k^{old} \right\}^{-1}.$$

Replacing the missing indicator variable by its conditional expectation in the complete likelihood and maximizing the likelihood function yields the updated estimates

$$p_l^{new} = \frac{1}{n}\left\{\xi_l + \sum_{i=1}^{n}(1-\delta_i)E\{I(x_i = t_l) \mid z_i, \boldsymbol{p}^{old}\}\right\}$$
$$= \frac{1}{n}\left[\xi_l + H(t_l,\boldsymbol{\theta})^{-1}p_l^{old}\sum_{k=1}^{l}\eta_k\left\{\sum_{j=k}^{L}H(t_j,\boldsymbol{\theta})^{-1}p_j^{old}\right\}^{-1}\right] \tag{2.2}$$

for $l = 1, \ldots, L$.

Let $\{\hat{g}_{\boldsymbol{\theta}}(t_1), \ldots, \hat{g}_{\boldsymbol{\theta}}(t_L)\}$ be the probabilities that the EM algorithm converges to. Thus, for fixed $\boldsymbol{\theta}$, the maximum likelihood estimator for G under the proposed generalized multiplicative censoring model is given by $\hat{G}_{\boldsymbol{\theta}}(t) = \sum_{l=1}^{L}\hat{g}_{\boldsymbol{\theta}}(t_l)I(t_l \leqslant t)$. Then, by the profile likelihood approach, the parameter $\boldsymbol{\theta}$ can be estimated by maximizing the full likelihood $\mathcal{L}_{MC}(\hat{G}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ under generalized multiplicative censoring.

3. Application to left-truncated and right-censored data

3.1 *Model setup*

Next, we consider the estimation problem for data under cross-sectional sampling with follow-up. Let T^0 be the time from disease incidence to the failure event of interest in a target population. Let $f(t)$, $F(t)$, and $S(t)$ denote the density function, the cumulative density function, and the survival function of T^0. Let A^0 be the time from disease incidence to the (potential) study recruitment time, where the density function of A^0 lies in a parametric family $\{h(\cdot,\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$. In a cross-sectional study, the prevalent population consists of individuals with the disease who have not experienced the failure event at the recruitment time; that is, individuals whose time to failure satisfies $T^0 \geqslant A^0$. Let (T, A) be the observed random variables in the prevalent population; then (T, A) has the same distribution function as (T^0, A^0) conditional on $T^0 \geqslant A^0$. The joint density function of (T, A) evaluated at (t, a) is

$$\frac{h(a,\boldsymbol{\theta})f(t)}{\mu(\boldsymbol{\theta}, F)}I(t \geqslant a),$$

where $\mu(\boldsymbol{\theta}, F) = \int_0^\infty H(t,\boldsymbol{\theta})\,\mathrm{d}F(t)$, with $H(t,\boldsymbol{\theta}) = \int_0^t h(u,\boldsymbol{\theta})\,\mathrm{d}u$.

In practice, the observation of survival time can be terminated before an individual experiences an event. We assume that the residual life time $T - A$ is subject to random censoring by an independent variable C. Let $Y = \min(T, A + C)$ be the observed survival time, and $\Delta = I(T \leqslant A + C)$ be the indicator of the failure event. Let the observed data $\{(y_i, a_i, \delta_i), i = 1, \ldots, n\}$ be independent and identically distributed copies of (Y, A, Δ). Our goal is to estimate the survival time distribution as well as the truncation time distribution by maximizing the full likelihood function

$$\mathcal{L}(\boldsymbol{\theta}, F) = \prod_{i=1}^{n}\left\{\frac{h(a_i,\boldsymbol{\theta})f(y_i)}{\mu(\boldsymbol{\theta}, F)}\right\}^{\delta_i}\left\{\frac{h(a_i,\boldsymbol{\theta})S(y_i)}{\mu(\boldsymbol{\theta}, F)}\right\}^{1-\delta_i}, \tag{3.1}$$

which involves both the parametric component $\boldsymbol{\theta}$ and the non-parametric component $F(\cdot)$. In the absence of censoring, the semiparametric maximum likelihood estimator is given in Wang (1989). Her method, however, cannot be applied in the presence of censoring.

3.2 *The semiparametric maximum likelihood estimator*

We now derive the semiparametric maximum likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{F})$ that maximizes the full likelihood (3.1) for left-truncated and right-censored data. Define the functions $g(t) = H(t, \boldsymbol{\theta}) f(t)/\mu(\boldsymbol{\theta}, F)$ and $G(t) = \int_0^t g(u)\,\mathrm{d}u$. Thus, g is a well-defined probability density function. The full likelihood (3.1) can be reparameterized as

$$\mathcal{L}(\boldsymbol{\theta}, G) = \mathcal{L}_1(G) \times \prod_{i=1}^{n} \{h(a_i, \boldsymbol{\theta}) H(y_i, \boldsymbol{\theta})^{-\delta_i}\}, \tag{3.2}$$

where $\mathcal{L}_1(G) = \prod_{i=1}^n g(y_i)^{\delta_i} \{\int_{u \geqslant y_i} H(u, \boldsymbol{\theta})^{-1}\,\mathrm{d}G(u)\}^{1-\delta_i}$. Thus, for fixed $\boldsymbol{\theta}$, $\mathcal{L}_1$ is proportional to the full likelihood function $\mathcal{L}_{MC}$ in (2.1) under the generalized multiplicative censoring model. Moreover, the second term on the right-hand side in (3.2) does not involve F. Thus, holding $\boldsymbol{\theta}$ fixed, the full likelihood is maximized by the maximizer of $\mathcal{L}_1$. Let $\hat{g}_{\boldsymbol{\theta}}(t_l)$ be the estimated jump size at time t_l for fixed $\boldsymbol{\theta}$ obtained by applying the EM algorithm (2.2), where $\{t_1, \ldots, t_L\}$ are the ordered and distinct values of $\{y_1, \ldots, y_n\}$. Let $\hat{G}_{\boldsymbol{\theta}}(t)$ be the corresponding estimate of the cumulative distribution function. Replacing $G(t)$ with $\hat{G}_{\boldsymbol{\theta}}(t)$ in $\mathcal{L}$, we have the profile likelihood of $\boldsymbol{\theta}$

$$\mathcal{L}_P(\boldsymbol{\theta}) = \prod_{i=1}^{n} \hat{g}_{\boldsymbol{\theta}}(y_i)^{\delta_i} \left\{ \int_{u \geqslant y_i} H(u, \boldsymbol{\theta})^{-1}\,\mathrm{d}\hat{G}_{\boldsymbol{\theta}}(u) \right\}^{1-\delta_i} \times \prod_{i=1}^{n} \{h(a_i, \boldsymbol{\theta}) H(y_i, \boldsymbol{\theta})^{-\delta_i}\}.$$

Let $\widehat{\boldsymbol{\theta}}$ be the maximizer of the profile likelihood $\mathcal{L}_P(\boldsymbol{\theta})$; then the maximum likelihood estimator $\hat{F}$ of F is given by

$$\hat{F}_n(t) = \sum_{l=1}^{L} \frac{H(t_l, \hat{\boldsymbol{\theta}})^{-1} \hat{g}_{\hat{\boldsymbol{\theta}}}(t_l)}{\sum_{k=1}^{L} H(t_k, \hat{\boldsymbol{\theta}})^{-1} \hat{g}_{\hat{\boldsymbol{\theta}}}(t_k)} I(t_l \leqslant t).$$

Note that $\hat{F}_n$ assigns positive probability mass at possibly all censored and uncensored event times $\{t_1, \ldots, t_L\}$ (Vardi, 1989; Qin *and others*, 2011) and the likelihood function is strictly concave in p_j, $j = 1, \ldots, L$. Arguing as in Vardi (1989), we can show that, given the unique supporting points $\{t_1, \ldots, t_L\}$, the maximizer of (2.1) is unique for each given $\boldsymbol{\theta}$, and that the EM algorithm (2.2) converges to the unique maximizer since the set of the probability measure is convex.

Another possible attempt to estimate $\boldsymbol{\theta}$ and F is to apply the iteration algorithm considered in Shen (2007, 2009). For any fixed θ^*, the author applied the same EM algorithm described in Section 2 to obtain an estimator, denoted by $\hat{g}_{\theta^*}(t)$, for $g(t)$. Then the estimator

$$\hat{F}_{\theta^*}(t) = \sum_{l=1}^{L} \frac{H(t_l, \boldsymbol{\theta}^*)^{-1} \hat{g}_{\boldsymbol{\theta}^*}(t_l)}{\sum_{k=1}^{L} H(t_k, \boldsymbol{\theta}^*)^{-1} \hat{g}_{\boldsymbol{\theta}^*}(t_k)} I(t_l \leqslant t)$$

can be shown to be consistent for the true F when $\boldsymbol{\theta}^*$ is the true parameter value. Replacing $S(t)$ with $S^*(t) = 1 - \hat{F}_{\theta^*}(t)$, the author proposed to maximize the marginal likelihood of A

$$\mathcal{L}_M(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{h(a_i, \boldsymbol{\theta}) S^*(a_i)}{\int_0^\infty h(u, \boldsymbol{\theta}) S^*(u)\,\mathrm{d}u}$$

with respect to θ to obtain an updated estimate of θ. The algorithm then iterates until convergence. Note that, in the absence of censoring, we have

$$\hat{F}_\theta(t) = \frac{\sum_{l=1}^{L} H(t_l, \boldsymbol{\theta})^{-1} I(t_l \leqslant t)}{\sum_{k=1}^{L} H(t_k, \boldsymbol{\theta})^{-1}}.$$

Thus, the estimator proposed by the author is equivalent to maximizing the marginal likelihood of the truncation times with $S(t)$ replaced with $1 - \hat{F}_\theta(t)$, that is, maximizing

$$\mathcal{L}_M(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ h(a_i, \boldsymbol{\theta}) \times n^{-1} \sum_{l=1}^{L} H(t_l, \boldsymbol{\theta})^{-1} I(t_l > a_i) \right\}.$$

On the other hand, the maximum likelihood estimator proposed in our paper is obtained by maximizing the full likelihood with $F(t)$ replaced with $\hat{F}_\theta(t)$, that is,

$$\mathcal{L}(\boldsymbol{\theta}, \hat{F}_\theta) = \prod_{i=1}^{n} \frac{h(a_i, \boldsymbol{\theta})}{H(y_i, \boldsymbol{\theta})}.$$

It is obvious that the estimator considered in Shen (2007, 2009) is not the maximum likelihood estimator, hence it is not expected to be fully efficient.

For the remainder of this paper, we assume that the support of $A^0 + C$ contains that of T^0, so that F is estimable on the entire support of $[0, \tau]$. This assumption is reasonable in many applications, for example, the disease incidence in a population could have occurred in the distant past. If the maximum support of $A^0 + C$, say, τ_1, is smaller than the maximum support of T^0, τ, then the proposed method estimates the conditional distribution function of $T^0 \mid T^0 \leqslant \tau_1$, that is, $F(t)/F(\tau_1)$. In contrast, if the lower limit of the support for $A^0 + C$, say, τ_2, is > 0, then the proposed method estimates the conditional distribution function of $T^0 \mid T^0 \geqslant \tau_2$, that is, $\{F(t) - F(\tau_2)\}/\{1 - F(\tau_2)\}$.

To establish the large-sample properties of the maximum likelihood estimator, we impose the following conditions:

(A1) The true parameters $(\boldsymbol{\theta}_0, F_0)$ belong to the space $\Theta \times \mathcal{F}$, where Θ is a known compact set in $\mathbb{R}^p$, and $\mathcal{F} = \{F : [0, \tau] \to [0, 1], F$ is a non-decreasing function and $\Lambda(t) = -\log\{1 - F(t)\}$ satisfying $\Lambda(0) = 0$, $\Lambda(\tau) < \infty\}$. Assume that the true cumulative hazard function $\Lambda_0(t) = -\log\{1 - F_0(t)\}$ has a derivative $\lambda_0(t)$ that is differentiable and satisfies $0 < \inf_{t\in[0,\tau]} \lambda_0(t) \leqslant \sup_{t\in[0,\tau]} \lambda_0(t) < \infty$.

(A2) The density function $h(t, \boldsymbol{\theta})$ of the truncation time is positive on $[0, \tau]$ and is differentiable with respect to $\boldsymbol{\theta}$. Moreover, both $h(t, \boldsymbol{\theta})$ and its partial derivative $h^{(1)}(t, \boldsymbol{\theta}) = \partial h(t, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ are bounded on $[0, \tau]$.

(A3) The censoring time for the residual life time is not degenerate at 0, that is, $\mathrm{pr}(C > 0) > 0$. The distribution function of C is absolutely continuous.

(A4) The information matrix $-\partial^2 E\{\mathcal{L}(\boldsymbol{\theta}, \hat{F}_{\boldsymbol{\theta}})\}/\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}$ evaluated at the true value $\boldsymbol{\theta}_0$ is positive definite.

Denote by $(\boldsymbol{\theta}_0, F_0)$ the true parameter values and by $(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)$ the maximum likelihood estimator that maximizes $\mathcal{L}(\boldsymbol{\theta}, F)$. The large-sample properties of $(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)$ are summarized in Theorem 3.1, with proofs given in the supplementary material available at *Biostatistics* online.

THEOREM 3.1 Under regularity conditions (A1)–(A4), the maximum likelihood estimators $(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)$ are consistent for the product of the Euclidean topology and the topology of uniform convergence on $[0, \tau]$,

that is,

$$|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| + \sup_{t\in[0,\tau]} |\hat{F}_n(t) - F_0(t)| \to 0$$

almost surely as $n \to \infty$. Moreover, $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0, \hat{F}_n - F_0)$ converges weakly to a tight mean zero Gaussian process $-\phi_0'\{\dot{\mathcal{U}}_0^{-1}(\mathcal{W})\}$ as $n \to \infty$, where definitions of ϕ_0', $\dot{\mathcal{U}}$, and $\mathcal{W}$ are given in the supplementary material available at *Biostatistics* online.

The strong consistency is proved by using the classical Kullback–Leibler information approach (Murphy, 1994; Parner, 1998). The asymptotic normality can be established by applying the general Z-estimator convergence theorem (van der Vaart and Wellner, 1996, Theorem 3.3.1). Although the variance–covariance matrix of $(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)$ can be obtained by the empirical plug-in version of the asymptotic variance given in Theorem 3.1, it is computationally complicated. As an alternative, we may use the bootstrap resampling method to estimate the variance of the maximum likelihood estimators. By the weak convergence stated in Theorem 3.1 and arguing as in van der Vaart and Wellner (1996, Chapter 3.6), the bootstrap method is expected to produce valid estimates of the variance of $(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)$.

3.3 *Test for stationarity of the incidence rate*

In addition to the survival time distribution, the truncation time distribution may be of independent interest. In particular, researchers may want to know whether there is a temporal change in the disease incidence. If the incidence rate increases over time, the underlying truncation time random variable is likely to be right skewed. In contrast, if the incidence of disease is constant over time, the underlying truncation time is uniformly distributed.

To test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, we consider the semiparametric likelihood ratio test statistic

$$R = -2\log\left\{\frac{\mathcal{L}(\boldsymbol{\theta}_0, \hat{F}_0)}{\mathcal{L}(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)}\right\},$$

where $\hat{F}_0$ maximizes the likelihood function $\mathcal{L}(\boldsymbol{\theta}_0, F)$ under the null. It follows from Theorem 3.1 and an extension of the general semiparametric likelihood ratio statistic theorem in Murphy and van der Vaart (1997) that R has a χ_p^2 distribution as $n \to \infty$, where p is the dimension of the vector parameter $\boldsymbol{\theta}$. This result can be used not only in hypothesis testing but also to construct approximate confidence sets for $\boldsymbol{\theta}$.

Next, we propose formal statistical tests for checking the stationarity assumption of the uniform truncation time distribution. Testing the stationarity assumption on disease incidence is an important yet understudied problem. To the best of our knowledge, only a few publications, including Asgharian *and others* (2006), Mandel and Betensky (2007), and Addona and Wolfson (2006), have focused on this problem. All three papers provided graphical examinations and formal tests of the stationarity condition based on the equality in the distribution of the observed truncation time A and that of the residual survival time $T - A$. In what follows, we propose a formal test based on the semiparametric likelihood ratio test statistic.

Following the spirit of Neyman's smooth tests of goodness of fit, we propose to embed the uniform density function to a parametric family of density functions that differ smoothly from the uniform density function. Neyman (1937) constructed a smooth alternative of order K to the null density function $h_0(t)$ given by

$$h(t, \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left\{\sum_{k=1}^{K} \theta_k e_k(t)\right\} h_0(t),$$

where $\{e_k(t), k = 1, \ldots, K\}$ is a set of orthonormal functions which satisfies

$$\int_0^\tau e_k(t)e_l(t)h_0(t)\,\mathrm{d}t = \delta_{kl} \quad \text{for } k, l = 0, 1, 2, \ldots, K,$$

with $e_0(t) = 1$ for $t \in [0, \tau]$, $\delta_{kl} = 1$ if $k = l$ and $= 0$ if $k \neq l$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, and $c(\boldsymbol{\theta})$ is a normalizing constant. When h_0 is the uniform density function, the orthonormal functions can be chosen to be the Legendre polynomials. In this way, an order K alternative is a polynomial of order K, while the null hypothesis is equivalent to a polynomial of degree 0. In other words, the smooth alternatives are given by

$$h(t, \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left\{\sum_{k=1}^{K} \theta_k t^k\right\}, \quad 0 \leqslant t \leqslant \tau, \tag{3.3}$$

with $c(\boldsymbol{\theta}) = [\int_0^\tau \exp\{\sum_{k=1}^K \theta_k t^k\}\,\mathrm{d}t]^{-1}$. Thus, testing for the uniform distribution is equivalent to testing $\boldsymbol{\theta} = \mathbf{0}$ against $\boldsymbol{\theta} \neq \mathbf{0}$ in (3.3).

To test whether the underlying truncation distribution is uniform, we consider the semiparametric likelihood ratio test

$$R = -2 \log\left\{\frac{\mathcal{L}(\mathbf{0}, \hat{F}_n^V)}{\mathcal{L}(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)}\right\},$$

where $(\hat{\boldsymbol{\theta}}_n, \hat{F}_n)$ is the semiparametric maximum likelihood estimator that maximizes $\mathcal{L}(\boldsymbol{\theta}, F)$ under Model (3.3), and $\hat{F}_n^V$ is Vardi's non-parametric maximum likelihood estimator under the null. It was recommended by Neyman (1937) and further investigated in Rayner and Rayner (2001) that a polynomial of degree 4 usually suffices to test for uniformity. In our experience, the smooth alternative with $K = 3$ yields good power and is more numerically stable than tests with higher-order polynomials when applied to left-truncated and right-censored data. In the remainder of this paper, we set $K = 3$ for the smooth alternative model (3.3).

4. Simulations and data analysis

4.1 *Monte Carlo simulations*

To evaluate the finite-sample performance of the proposed methods, we conducted a series of Monte Carlo simulations. The first set of simulations compared the performance of the proposed semiparametric maximum likelihood estimator to existing methods. We generated survival time T^0 from a truncated Weibull random variable with shape and scale parameters of (0.7, 1). The underlying truncation time A^0 was independently generated from a truncated exponential distribution with survival function $\exp(-t)/\{1 - \exp(-\tau)\}$. To form a prevalent cohort, realizations of (A^0, T^0) were generated repeatedly until there were $n = 200$ subjects satisfying the sampling constraint $A^0 \leqslant T^0$. The censoring time C for the residual survival time $T - A$ in the prevalent cohort was generated from a uniform distribution on the interval $[0, \tau_c]$, where τ_c was chosen so that the overall censoring rate was approximately 0%, 25%, and 50%. We fit each generated data set by the proposed profile likelihood method, using the *optim* function with the option "Nelder-Mead" in R.

Five different methods were applied to estimate the failure time distribution: (I) the truncation product-limit estimator; (II) Vardi's non-parametric maximum likelihood estimator under uniform truncation time distribution (Vardi, 1989); (III) the estimator obtained by the iteration algorithm considered in Shen (2007, 2009) with the assumption of a Weibull truncation time distribution; (IV) the proposed estimator with the assumption of a Weibull truncation time distribution; and (V) the proposed estimator with the assumption that the truncation time has a smooth alternative density (3.3) with $K = 3$. Table 1 summarizes the

Table 1. *Simulation results for various estimators for the failure time distribution.*

		Product-limit		NPMLE			Shen's method			Weibull			Smooth alternative		
pr($\Delta = 1$)	$F(t)$	Bias	SE	Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE	RE
1	0.2	−1	14	5	13	1.10	−3	11	1.54	−2	12	1.48	−2	11	1.63
	0.4	−1	11	11	10	1.42	−3	10	1.33	−1	10	1.38	−2	9	1.53
	0.6	0	8	14	5	2.32	−2	7	1.24	−1	7	1.33	−1	7	1.45
	0.8	0	4	11	2	5.10	−1	4	1.21	−1	4	1.27	−1	4	1.35
0.75	0.2	−1	14	4	13	1.10	−3	11	1.52	−2	12	1.46	−2	11	1.63
	0.4	−1	12	11	10	1.42	−3	10	1.33	−1	10	1.36	−2	9	1.52
	0.6	0	8	14	5	2.27	−2	7	1.23	−1	7	1.31	−1	7	1.43
	0.8	0	4	11	2	5.25	−1	4	1.19	0	4	1.24	−1	4	1.32
0.5	0.2	−1	14	4	13	1.13	−3	11	1.53	−2	12	1.47	−2	11	1.57
	0.4	−1	12	10	10	1.42	−3	10	1.34	−2	10	1.37	−2	10	1.46
	0.6	0	8	13	6	2.21	−2	7	1.24	−1	7	1.31	−1	7	1.37
	0.7	0	5	11	2	5.18	−1	4	1.18	0	4	1.21	0	4	1.25

Product-limit, the truncation product-limit estimator; NPMLE, Vardi's non-parametric maximum likelihood estimator; Shen's method; Weibull, the proposed estimator with Weibull truncation time distribution; Smooth Alternative, the proposed estimator with smooth alternative truncation time distribution; Bias, the empirical bias ($\times 10^2$); SE, the empirical standard deviation ($\times 10^2$); RE, the empirical variance of the truncation product-limit estimator divided by that of an estimator.

Monte Carlo bias, standard deviation, and relative efficiency for each estimator at $t = 0.12, 0.38, 0.88$, and 1.97 based on 1000 replications. The true cumulative distribution functions at the selected time points are 0.2, 0.4, 0.6, and 0.8, respectively. As expected, Vardi's estimator is biased, because the underlying truncation time is not uniformly distributed. Methods III, IV, and V outperform the truncation product-limit estimator by Method II in terms of Monte Carlo standard deviation. The proposed method with the assumption of a smooth alternative density (Method V) works very well and is most efficient when estimating the failure time distribution. Methods III and IV, which employ the Weibull density, also yield small biases. Interestingly, compared with Method IV, the estimator of the failure time distribution obtained using Method III can be slightly more efficient at early time points, but is equally efficient at later time points. For the parameters in the truncation distribution, Methods IV and V show clear superiority than Method III. Specifically, the biases in the estimated shape and scale parameters of the Weibull density by Method V are (0.011, 0.008), (0.010, 0.011), and (0.005, 0.032) with standard deviations (0.071, 0.136), (0.071, 0.142), and (0.072, 0.158), respectively, when the proportion of uncensored subjects is 100%, 75%, and 50%. When using Method III, the corresponding biases under three scenarios are respectively (−0.027, 0.035), (−0.027, 0.043), and (−0.034, 0.084) with standard deviations (0.090, 0.156), (0.091, 0.198), and (0.093, 0.238), which are larger than those by the proposed method (Method V).

The second set of simulations evaluated the power of the proposed semiparametric likelihood ratio test under various scenarios. We simulated survival time T^0 from a truncated exponential distribution with density function $\exp(-t)/\{1 - \exp(-10)\}$ for $t \in (0, 10]$ and $T^0/10$ from a beta distribution with parameters 0.5 and 5. The underlying truncation times were generated so that $A^0/10$ followed the uniform distribution on [0, 1] and a beta distribution with shape parameters 0.75 and 1. The censoring times were generated from uniform distributions so that the proportions of uncensored subjects were 100%, 75%, 50%. For each set of simulations, we considered different sample sizes $n = 100, 200$. The significance level of the semiparametric likelihood ratio test was set at 0.05. Table S.1 in the supplementary material available at *Biostatistics* online summarizes the estimated size and power of the proposed semiparametric likelihood ratio test for testing $H_0 : \theta_1 = \theta_2 = \theta_3 = 0$ in the smooth alternative density (3.3).

For comparison, we also applied the paired logrank test proposed by Mandel and Betensky (2007) that compares the truncation time distribution and the residual survival time distribution, and reported the size and power of the test in Table S.1 in the supplementary material available at *Biostatistics* online. When the underlying truncation time is uniformly distributed, the estimated sizes of both tests are close to the predetermined significance level (0.05). As expected, when the truncation time distribution is not uniform, the power to reject the null hypothesis increases with the sample size but decreases with the proportion of censored subjects. The proposed test is more powerful than the paired logrank test when the proportion of censored subjects is low, and is as efficient as its competitor when the censoring proportion is high.

4.2 *Analysis of Canadian study of health and aging*

In this section, we report the results of data analysis for a cohort of prevalent cases in one of the largest epidemiologic studies of dementia, the Canadian Study of Health and Aging. From February 1991 to May 1992, an extensive survey was carried out and a total of 1132 persons aged 65 and older with dementia were identified in this first phase of the study. For each study subject, a diagnosis of possible Alzheimer's disease, probable Alzheimer's disease, or vascular dementia was assigned, and the date of dementia onset was determined by interviewing care-givers. Information on mortality was collected between January 1996 and May 1997.

We considered a subset of the study data by excluding those with missing date of onset or classification of dementia subtype. Moreover, as in Wolfson *and others* (2001), those with observed survival time of 20 or more years were excluded because these subjects are considered unlikely to have Alzheimer's disease or vascular dementia. As a result, a total of 807 dementia patients were included in our analysis. Among them, 388 had a diagnosis of probable Alzheimer's disease, 249 had possible Alzheimer's disease, and 170 had vascular dementia. In the second phase of the study a total of 627 deaths were recorded, among which 302 subjects has a diagnosis of probable Alzheimer's, 189 possible Alzheimer's, and 136 vascular dementia.

We first applied the proposed semiparametric likelihood ratio test to check the stationarity assumption that the incidence of dementia is constant over time within each subgroup. The p-values were 0.14, 0.17, and 0.28 for possible Alzheimer's, probable Alzheimer's and vascular dementia, respectively. Figure 1 shows the estimated cumulative distribution functions using the truncation product-limit estimator, Vardi's non-parametric maximum likelihood estimator, and the proposed estimator with the assumption that the truncation time has a smooth alternative density (3.3) with $K = 3$. In general, the three estimated survival curves are reasonably close to each other. Interestingly, the distribution function obtained by the proposed method is closer to the product-limit estimator than Vardi's non-parametric maximum likelihood estimator. In fact, as shown in Figure 5 of Asgharian *and others* (2002) and Figure 4 of Asgharian *and others* (2006), the truncation times appear to have a cyclic effect according to year. Figure 1 also shows the relative efficiency of the other two estimators compared with the truncation product-limit estimator at selected time points using these three estimators. To estimate the standard errors of the estimated survival probabilities, we adopted a non-parametric bootstrap method by sampling 807 subjects with replacement from the data set. The resampling procedure was repeated 2000 times, and the standard error was estimated by the standard deviation of the 2000 survival probability estimates at each time point. As expected, the proposed estimator is more efficient than the truncation product-limit estimator, but is less efficient than the non-parametric maximum likelihood estimator under uniform truncation time distribution.

4.3 *Analysis of nursing home data*

We next illustrate our methods by analyzing the well-known Channing House data (Hyde, 1977), which recorded age at entry and age at death for 462 residents of Channing House, a retirement center in Palo

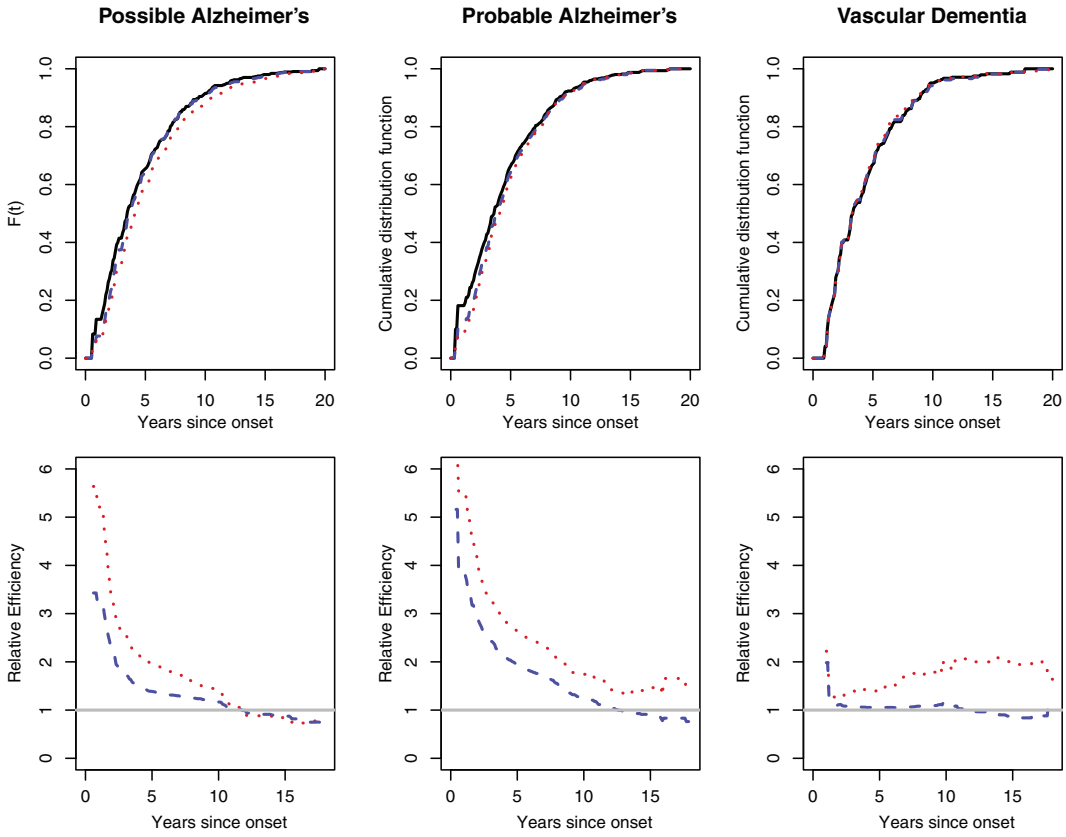

Fig. 1. Upper panel: Estimated distribution functions using the truncation product-limit estimator (solid line), Vardi's non-parametric maximum likelihood estimator (dotted line), and the proposed estimator (dashed line) for different diagnosis subtypes. Lower panel: relative efficiency compared with the truncation product-limit estimator for Vardi's non-parametric maximum likelihood estimator (dotted line) and the proposed estimator (dashed line).

Alto, California, from its opening in 1964 to the data collection date July 1, 1975. The survival time is left-truncated by the age at entry and right-censored by end of study or loss to follow-up. As in Wang (1991), we considered a subset of 438 residents (94 male and 344 female) who survived longer than 866 months. The proposed semiparametric likelihood ratio test rejected the assumption that the entry age was uniformly distributed for both male and female residents (both with a p-value < 0.001). The upper panels of Figure 2 show the estimated distribution functions of the survival time for males and for females using the truncation product-limit estimator, Vardi's non-parametric likelihood estimator and the proposed semiparametric estimator with smooth alternative truncation distribution. Moreover, the lower panels of Figure 2 show the estimated distribution functions of the truncation time for different gender groups using the non-parametric estimator proposed in Wang (1991) and the smooth alternative density with $K = 3$. Compared with males, the truncation time distribution for females further departs from the uniform distribution. The comparison of relative efficiency based on the bootstrap standard error estimates (data not shown), however, suggested that the product-limit estimator is preferred in this case, as the smooth alternative is as or less efficient than the truncation product-limit estimator.

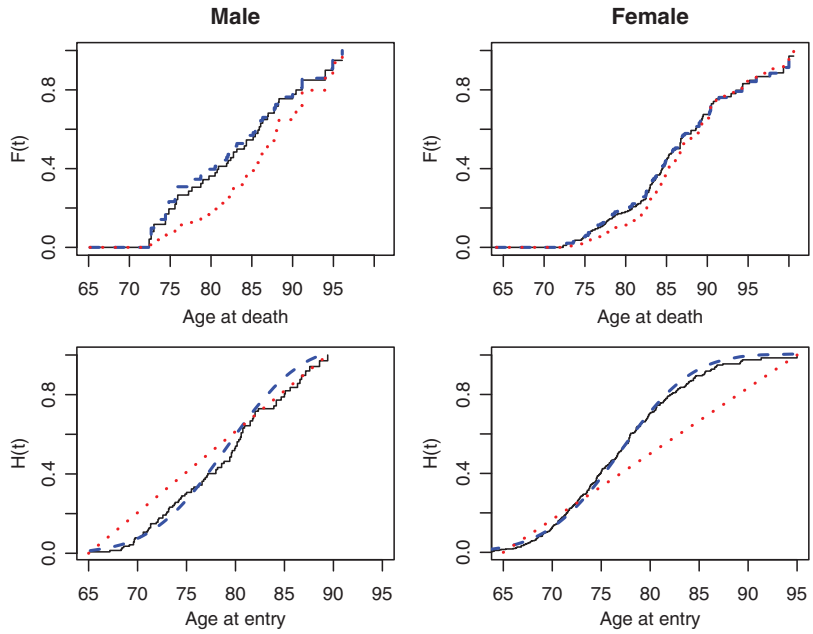

Fig. 2. Estimates of $F(t)$ and $H(t)$ for Channing House data using Wang's non-parametric estimator (solid line), Vardi's non-parametric maximum likelihood estimator (dotted line), and the proposed estimator with the smooth alternative (dashed line).

5. REMARK

The purpose of this paper is 2-fold: first, to generalize Vardi's multiplicative censoring model with a unified EM algorithm for model estimation, and second, to establish the connection between the likelihood for data subject to the generalized multiplicative censoring ($\mathcal{L}_{MC}$) and that for left-truncated and right-censored data ($\mathcal{L}$), so that the EM algorithm developed for $\mathcal{L}_{MC}$ can be applied to obtain the maximum profile likelihood estimator for $\mathcal{L}$. Although the asymptotic properties of the maximum likelihood estimator for $\mathcal{L}_{MC}$ are not discussed in this paper, they can be established by applying a similar argument as that for the maximum profile likelihood estimator for $\mathcal{L}$.

Left-truncated data can be viewed as selection-biased samples with sampling weights being proportional to the distribution functions of the truncation times. Many authors, including Vardi (1985) and Gill *and others* (1988), considered non-parametric estimation for selection bias models with known, non-negative weight functions, and Gilbert *and others* (1999) generalized Vardi's model to allow for the weighting functions to depend on an unknown parameter. However, these methods usually cannot be applied directly to the problem considered in this paper because the left-truncated survival times are further subject to right censoring. It is of interest to generalize the methods for selection bias models to deal with censoring. Future research is warranted.

6. SOFTWARE

Software in the form of R code, together with a sample input data set and complete documentation is available on request from the corresponding author.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors are grateful to Professors Masoud Asgharian, Ian McDowell, and Christina Wolfson for sharing the Canadian Study of Health and Aging data. The data reported in the example were collected as part of the CSHA. *Conflict of Interest*: None declared.

Funding

This work was supported in part by grants CA016672 and CA006973 from the National Institutes of Health. The core of the CSHA study was funded by the Seniors' Independence Research Program through the National Health Research and Development Program of Health Canada (Project no.6606-3954-MC(S)). Additional funding was provided by Pfizer Canada Incorporated through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP Project 6603-1417-302(R), Bayer Incorporated, and the British Columbia Health Research Foundation Projects 38 (93-2) and 34 (96-1). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada.

References

Addona, V. and Wolfson, D. B. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Analysis* **12**(3), 267–284.

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., Wolfson, D. B. and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in Medicine* **25**(10), 1751–1767.

Gilbert, P. B., Lele, S. R. and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86**, 27–43.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* **16**, 1069–1112.

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika* **64**, 225–230.

Lagakos, S. and Barraj, L. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **75**(3), 515–523.

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

Mandel, M. (2007). Nonparametric estimation of a distribution function under biased sampling and censoring. *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. IMS Lecture Notes Onograph Series* **54**, 224–238.

Mandel, M. and Betensky, R. A. (2007). Testing goodness of fit of a uniform truncation model. *Biometrics* **63**, 405–412.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics* **22**, 712–731.

MURPHY, S. A. (1997). Semiparametric likelihood ratio inference. *The Annals of Statistics* **25**, 1471–1509.

NEYMAN, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 150–199.

PARNER, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26**, 183–214.

QIN, J., NING, J., LIU, H. AND SHEN, Y. (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association* **106**(496), 1434–1449.

RAYNER, G. D. AND RAYNER, J. C. W. (2001). Power of the Neyman smooth tests for the uniform distribution. *Journal of Applied Mathematics and Decision Sciences* **5**(3), 181–191.

SEVERINI, T. A. AND WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of Statistics* **20**, 1768–1802.

SHEN, P.-S. (2007). A general semiparametric model for left-truncated and right-censored data. *Journal of Nonparametric Statistics* **19**, 113–129.

SHEN, P.-S. (2009). Semiparametric analysis of survival data with left truncation and right censoring. *Computational Statistics and Data Analysis* **53**, 4417–4432.

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.

VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

VARDI, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13**, 178–203.

VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76**, 751–761.

VARDI, Y. AND ZHANG, C. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *The Annals of Statistics* **20**(2), 1022–1039.

WANG, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* **84**, 742–748.

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

WOLFSON, C., WOLFSON, D. B. AND ASGHARIAN, M. (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine* **344**(15), 1111–1116.

ZELEN, M. AND FEINLEIB, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56**, 601–614.

[*Received August 28, 2014; revised February 17, 2015; accepted for publication February 25, 2015*]