Published in final edited form as:

Stat Med. 2013 June 15; 32(13): 2278–2291. doi:10.1002/sim.5637.

Empirical likelihood-based confidence intervals for length-biased data

J. Ning^a, J. Qin^b, M. Asgharian^c, and Y. Shen^a

^aDepartment of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

^bBiostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892, USA

^cDepartment of Mathematics and Statistics, McGill University, Montreal, QC, H3A 2K6, Canada

Abstract

Logistic or other constraints often preclude the possibility of conducting incident cohort studies. A feasible alternative in such cases is to conduct a cross-sectional prevalent cohort study for which we recruit prevalent cases, i.e. subjects who have already experienced the initiating event, say the onset of a disease. When the interest lies in estimating the lifespan between the initiating event and a terminating event, say death for instance, such subjects may be followed prospectively until the terminating event or loss to follow-up, whichever happens first. It is well known that prevalent cases have, on average, longer lifespans. As such they do not constitute a representative random sample from the target population; they comprise a biased sample. If the initiating events are generated from a stationary Poisson process, the so-called stationarity assumption, this bias is called length bias. The current literature on length-biased sampling lacks a simple method for estimating the margin of errors of commonly used summary statistics. We fill this gap using the empirical likelihood-based confidence intervals by adapting this method to right-censored length-biased survival data. Both large and small sample behaviors of these confidence intervals are studied. We illustrate our method using a set of data on survival with dementia, collected as part of the Canadian Study of Health and Aging.

Keywords

confidence interval; length-biased data; empirical likelihood ratio test; mean; median; quantile; survival function

1. Introduction

Length bias often arises in prevalent cohort studies when evaluating the long-term survival outcome. A prevalent sampling design which draw samples from subjects with a condition or disease (e.g diagnosed with dementia) at the time of enrollment is generally more

efficient and practical than an incident sampling design. The cohort is formed by individuals who have experienced the initial event (e.g., the onset of the disease) but have not experienced the failure event (e.g., death) at the time of recruitment. The individuals are then followed over time to monitor endpoints such as disease progression or death. Under this sampling design, individuals with shorter survival times are selectively excluded, while those with longer survival are more likely included in the dataset. In the presence of biased sampling, the standard methods for analyzing traditional survival data which do not properly account for sampling bias can lead to biased inference. Length-biased data are special case of left-truncated data under the stationarity assumption, which assumes that the initiation time (onset of the disease) is generated from a stationary Poisson process [1, 2, 3]. Although the statistical issues and model complications related to length-biased data have been recognized in the statistics and epidemiology literatures for decades [4], systematic solutions for statistical inferences lag behind methodology development for traditional survival data.

The Canadian Study of Health and Aging (CSHA) serves as a typical example of such studies. CSHA randomly contacted 14,026 adults of ages 65 years or older throughout Canada and invited them to participate in a health survey. A total of 10,263 adults agreed to participate in the study. Among them, 1132 people were identified as having dementia at the screening examination. These individuals with dementia were then followed until their death or dates of last follow-up in 1996. One of the study objectives for CSHA was to explore the survival duration of patients with dementia among adults age 65 and older [5]. The estimation of the survival duration from a prevalent cohort study can be biased when adjustments are not made for the effect of the sampling mechanism, which prefers subjects with longer survival times. Indeed, subjects who die rapidly after disease onset are less likely to be included in the study (left truncation). As demonstrated by Wolfson et al. [5], the bias-adjusted estimate of the median was about one-half of the median derived from the Kaplan-Meier estimate for conventional survival data.

Nonparametric estimation of the survival distribution when data are subject to length-biased sampling has been studied extensively over the last few decades by Vardi [6], Wang [7], Asgharian et al [8], Asgharian and Wolfson [9], Ning et al. [10], Gill and Keiding [11] and Cook and Bergeron [12] among others. Even though the large sample properties have been derived for the estimated survival distribution from the full likelihood approach (e.g., [8, 9]), their asymptotic variance has to be obtained by solving an integral equation. This complex variance-covariance form does not facilitate a variance estimator with practical feasibility for constructing the confidence intervals of such estimators. A similar complication is expected for the estimation of the confidence intervals for important summary statistics such as the mean, median and other quantiles of the target population.

Estimating the confidence intervals for the summary statistics is a fundamental part of any statistical analysis. For uncensored data, Madansky [13], Thomas and Grunkemeier [14], Owen [15] and Qin [16, 17] introduced the nonparametric empirical likelihood ratio function to construct confidence intervals for differentiable functions of statistics without estimating the variance of the statistics. Matthews [18] extended the method to estimate the confidence intervals for the cumulative hazard function and for Kaplan-Meier estimates for traditional time-to-event data. Other related work using empirical likelihood-based confidence intervals

includes that of Zhou [19], Zhou and Jeong [20] for traditional right-censored data. We will generalize the empirical likelihood ratio method to right-censored length-biased data in order to construct nonparametric confidence intervals for important summary statistics and the survival function without estimating the variances of the corresponding estimators.

For traditional right-censored data, the Kaplan-Meier estimator is proposed to estimate survival function $S(t)$, which can be derived from the nonparametric maximum likelihood under an independent right-censoring mechanism. Commonly used summary statistics such as the mean and median of the data can be expressed as a functional form of $S(\cdot)$. Given the Kaplan-Meier estimator, however, it is hard to get a reliable estimate of the mean for traditional right-censored data if the follow-up time is not long enough. However, the mean survival time can be an important statistic for risk assessment in insurance industry. Although length-biased failure time data share some common features with traditional right-censored data, they are different from traditional failure time data and have a unique structure. The mean of length-biased right-censored data, for instance, can be consistently estimated given Vardi's estimator derived from the nonparametric maximum likelihood, as long as there is a positive follow-up time, regardless of the follow-up length. Moreover, comparing the mean difference in a two-sample setting is an important statistical problem, but can be difficult for right-censored length-biased data. Using a nonparametric empirical likelihood approach, we can readily propose a confidence interval for the estimated mean and the difference in the means in the two-sample setting.

The aforementioned applications and theoretical challenges in analyzing length-biased data call for practical statistical inference tools. The objective of this paper is to construct such inferential tools for nonparametric confidence intervals of commonly used summary functions based on an empirical likelihood ratio approach for length-biased right-censored data.

2. Empirical likelihood for length-biased data

In this section, we introduce our notations, Vardi's multiplicative censorship model [6], and present some preliminary derivations of empirical likelihood (EL) methods for length-biased right-censored data. To illustrate the EL method for length-biased right-censored data, we consider the setting of the CSHA cross-sectional follow-up study through which prevalent cohort survival data were collected on adults with dementia.

2.1. Vardi's multiplicative censorship model

Vardi [6] revealed that the nonparametric maximum likelihood estimate (NPMLE) of the survival distribution under multiplicative censoring is equivalent to the NPMLE of the survival distribution of the observed length-biased data. Under the multiplicative censoring mechanism, assume that $y_1, \dots, y_k$ are independently identically distributed random variables from a cumulative distribution function (CDF) G ; and the second sample is subject to a uniform random multiplicative censoring, $z_1, \dots, z_m$. In another word, $z_j = u_j v_j$ is the product of a random variable v_j which follows the same CDF of G and u_j follows an independent uniform (0,1) distribution, where $j = 1, \dots, m$. Given the above specifications,

one can express probability density function of $Z = t$ in terms of the probability distribution of G as follows,

$$f_z(t) = \int_{z \geq t} \frac{1}{z} dG(z), \quad t > 0.$$

The full likelihood function for the estimation of G based on the observed data $y_1, \dots, y_k$ and $z_1, \dots, z_m$ can be written as

$$L(G) = \prod_{i=1}^k dG(y_i) \prod_{j=1}^m \left\{ \int_{z \geq t} \frac{1}{z} dG(z) \right\}. \quad (1)$$

Vardi proposed the EM algorithm to estimate NPMLE of G by considering $(y_1, \dots, y_k, v_1, \dots, v_m)$ as the ‘complete data’ and of $(y_1, \dots, y_k, z_1, \dots, z_m)$ as the ‘incomplete data’. In the following section, we will see that the full likelihood function of the observed length-biased data leads to the same expression as in (1), although the motivating problems are different.

2.2. Likelihood for length-biased data

Let $x_1, \dots, x_n$ denote the observed times subject to right censoring under length-biased sampling measured from the disease onset to failure, and $\delta_1, \dots, \delta_n$ denote the corresponding censoring indicators. The censoring time measured from the disease onset is denoted by C , and $\delta = I(X < C)$. Let g and G be the respective density and CDF of the observed (biased) X . We retain f and F for the corresponding functions of the unbiased $\tilde{X}$ from the target population. We assume that τ is a finite upper bound of the support for the unbiased survival time.

Under length-biased sampling, the unbiased distribution function F of $\tilde{X}$ is related to the biased distribution G of X as

$$f(x)/\mu = x^{-1}g(x), \quad S(x)/\mu = \int_x^\tau u^{-1}dG(u), \quad x > 0, \quad (2)$$

where $S(x) = 1 - F(x)$ and μ is the mean of $\tilde{X}$ and can be expressed by $\mu = [\int_0^\tau u^{-1}dG(u)]^{-1}$. Adapting notations to be parallel to those of Vardi [6], the likelihood function for the observed right-censored length-biased data is proportional to

$$\prod_{i=1}^n \frac{f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}}{\mu},$$

which can be proportionally reformulated by G using (2) as

$$L(G) = \prod_{i=1}^n \{dG(x_i)\}^{\delta_i} \left\{ \int_{u \geq x_i} \frac{1}{u} dG(u) \right\}^{1-\delta_i}. \quad (3)$$

It is worth noting that the likelihood functions of (1) and (3) are equivalent, even though the data are observed under different mechanisms.

Let $0 < t_1 < t_2 \dots < t_h$ denote the distinct values of $x_1, \dots, x_n$ in increasing order. In general $h \leq n$. If there are ties, $h < n$, $h = n$ otherwise. Let ξ_j and ζ_j be respectively the multiplicity of uncensored and censored times at t_j :

$$\xi_j = \sum_{i=1}^n \delta_i I(x_i = t_j), \quad \zeta_j = \sum_{i=1}^n (1 - \delta_i) I(x_i = t_j).$$

The NPMLE of $G(\cdot)$ has a positive probability mass on both the censored and failure times. Let p_j denote the probability mass on all distinct observations $t_j, j = 1, \dots, h$ for $G(\cdot)$. The empirical log likelihood for the observed length-biased data is derived as

$$\ell(\mathbf{p}) = \sum_{j=1}^h \left[\xi_j \log(p_j) + \zeta_j \log \left\{ \sum_{k=j}^h t_k^{-1} p_k \right\} \right], \quad (4)$$

where $\mathbf{p} = (p_1, \dots, p_h)$, and is subject to the constraints

$$p_j \geq 0 \quad (j=1, 2, \dots, h), \quad \sum_{j=1}^h p_j = 1. \quad (5)$$

The log likelihood, equation (4), under the above constraints can be expressed as

$$\ell_0(\mathbf{p}) = \sum_{j=1}^h \left[\xi_j \log(p_j) + \zeta_j \log \left\{ \sum_{k=j}^h t_k^{-1} p_k \right\} \right] - \lambda_1 \left(\sum_{i=1}^h p_i - 1 \right). \quad (6)$$

The maximization of the log likelihood, equation (6), with respect to $\mathbf{p}$ leads to Vardi's NPMLE for the distribution function $G(\cdot)$.

3. Empirical likelihood ratio tests for length-biased data

We first present necessary modifications of the EL-based confidence intervals and adaptation of the EL-based expectation-maximization (EM) algorithm for the estimation of confidence intervals for length-biased right-censored data. In the subsections that follow, we describe in detail how to estimate the confidence interval for the mean, median and pointwise survival function, respectively.

3.1. Empirical likelihood ratio test approach

The basic principle for constructing the confidence interval of a general differentiable statistical functionals for right-censored data is to obtain the empirical likelihood ratio statistics under the specified constraints using the Lagrange multiplier method [18, 19]. Similar to the definition in Owen (1988), we require the functionals be Fréchet differentiable. Although the existence of a Fréchet derivative is a strong condition, the simple functional forms considered here satisfy this condition. Most of the summary functions of a random variable can be expressed by its corresponding cumulative distribution function, thus a general constraint can be specified as follows, using notations similar to those of Owen [15] and Zhou [19]:

$$\int_0^\tau q(t) dF(t) = \gamma, \text{ or } \int_0^\tau [q(t) - \gamma] dF(t) = 0, \quad (7)$$

where $q(t)$ is a left continuous function of the bounded variation that can be specified for various summary statistics.

For length-biased right-censored data, the empirical likelihood approach is based on the estimation of the biased distribution function $G(t)$ instead of the unbiased distribution $F(t)$. The constraint for length-biased data can be equivalently formulated as

$$\int_0^\tau q_\gamma(t) dG(t) = 0, \quad (8)$$

where $q_\gamma(t) = [q(t) - \gamma]/t$.

To construct the confidence interval for γ , one needs to maximize the above log-likelihood function, equation (6), with respect to $\mathbf{p}$ under an extra constraint H_0 ,

$$H_0: \sum_{i=1}^h q_\gamma(t_i) p_i = 0.$$

The corresponding empirical log likelihood can be expressed as

$$\ell_1(\mathbf{p}) = \sum_{j=1}^h \left[\xi_j \log(p_j) + \zeta_j \log \left\{ \sum_{k=j}^h t_k^{-1} p_k \right\} \right] - \lambda_1 \left(\sum_{i=1}^h p_i - 1 \right) - \lambda \sum_{i=1}^h q_\gamma(t_i) p_i. \quad (9)$$

To facilitate the EM algorithm to calculate the empirical likelihoods, we use a weighted probability product to express the expected likelihood with complete data, similar to the work of Owen [21], $\prod_{i=1}^n p_i^{w_i}$, for which the nonnegative weights $w_1, \dots, w_n$, are self-explained in the subsequent section. The expected log-likelihood function for $\ell(\mathbf{p})$ is

$$\ell_0^E(\mathbf{p}) = \sum_{i=1}^h w_i \log(p_i) - \lambda_1 \left(\sum_{i=1}^h p_i - 1 \right). \quad (10)$$

The maximization of the log empirical likelihood (9) with the extra constraint H_0 can be derived by maximizing the following expected log-likelihood function:

$$\ell_1^E(\mathbf{p}) = \sum_{i=1}^h w_i \log p_i - \lambda_1 \left(\sum_{i=1}^h p_i - 1 \right) - \lambda \sum_{i=1}^h q_\gamma(t_i) p_i.$$

Maximizing the above objective with respect to $\mathbf{p}$, λ_1 , and λ , we obtain

$$p_i = \frac{w_i}{\sum_j w_j + \lambda q_\gamma(x_i)},$$

where λ is the solution of the equation

$$\sum_{i=1}^h \frac{w_i q_\gamma(x_i)}{\sum_j w_j + \lambda q_\gamma(x_i)} = 0.$$

The above expression can be used to propose a constrained EM algorithm to calculate the confidence interval of the quantity γ . Let $\hat{\mathbf{p}}^0$ and $\hat{\mathbf{p}}^1$ denote the NPMLE of $\mathbf{p}$ by maximizing $\ell(\mathbf{p})$ and $\ell_1(\mathbf{p})$, respectively. Under some regularity conditions [15, 22], one can show that the likelihood ratio statistic

$$R(\gamma) = 2(\ell_0(\hat{\mathbf{p}}^0) - \ell_1(\hat{\mathbf{p}}^1(\gamma))) \\ = 2 \left[\max_{\mathbf{p}} \left\{ \ell(\mathbf{p}), \text{ subject to } \sum_{i=1}^n p_i = 1, p_i \geq 0 \right\} - \max_{\mathbf{p}} \left\{ \ell(\mathbf{p}), \text{ subject to } \sum_{i=1}^h p_i = 1, p_i \geq 0, \sum_{i=1}^h p_i q_\gamma(t_i) = 0 \right\} \right] \quad (11)$$

converges to a χ^2 distribution with one degree of freedom under H_0 . It is clear that λ depends on γ through $\hat{\mathbf{p}}^1(\gamma)$. By evaluating $\{\gamma | R(\gamma) \leq Q_{1-\alpha}\}$ where $Q_{1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with one degree of freedom, we can obtain a $1-\alpha$ confidence interval for γ via an asymptotic chi-squared distribution.

The NPMLE under constraint (8), $\hat{\mathbf{p}}^1$, has the same properties as the unconstrained NPMLE $\hat{\mathbf{p}}^0$ listed in [6].

Property 1. The likelihood function (3) under constraint (8) has a unique maximizer, say $\hat{\mathbf{p}}^1$.

Property 2. The likelihood (3) under constraint (8) increases with each iteration of the EM algorithm described above.

Property 3. The algorithm converges to $\hat{p}^1$, the unique maximizer of the likelihood (3) under constraint (8).

Property 4. Let $\hat{G}_1(t) = \sum_i \hat{p}_i^1 I(t_i \leq t)$ be the nonparametric maximum likelihood estimate of G . Then $\hat{G}_1(t)$ is a consistent estimate of $G(t)$ for $t > 0$ under constraint H_0 .

Properties 1–3 follow immediately from Vardi's argument upon noticing that the feasible set of the maximization problem with constraint (8) is still a convex set. Property 4 also follows from a similar argument to that of Vardi (1989, p 756) upon replacing equation (2.12) of Vardi by

$$\lambda q_\gamma(t) + 1 = \frac{k(t)}{t} \int_0^t \{1 - F_C(y)\} dy + \frac{1}{t} \int_0^t F_C(y) \frac{\int_y^\tau \frac{G(dz)}{z}}{\int_y^\tau \frac{G(dz)}{zk(z)}} dy$$

where $k(t) = G(dt)/\hat{G}_1(dt)$ and F_C is the CDF of censoring time C measured from study enrollment, and noticing that $\lambda \rightarrow 0$ as $n \rightarrow \infty$, using an argument similar to that of Owen (1990, pages 100–101).

3.2. Computational algorithm of empirical likelihood ratio test

Both the nonparametric maximum likelihood estimator of $G(\cdot)$ with and without the extra constraint (8) can be obtained by using the EM algorithm for length-biased data. The EM algorithm for the NPMLE of probabilities without constraint H_0 is given by Vardi [6]. Here we only detail the EM algorithm with the extra constraint under H_0 .

At the first step, we may select any $p_j^{(1)}$ satisfying $\sum_j p_j^{(1)} = 1$ and $p_j^{(1)} > 0$ on $t_j, j = 1, \dots, h$.

E-step—Given the current estimate of $G(\cdot)$, denoted by $\hat{G}^{(1)}(\cdot)$, that has positive probability mass $\hat{p}_j^{(1)}$ at t_j , we compute the weight w_j as

$$\begin{aligned} w_j^E &= E_{\hat{G}^{(1)}} \left\{ \sum_{i=1}^n \delta_i I(x_i = t_j) + \sum_{i=1}^n (1 - \delta_i) I(x_i = t_j) \mid (x_1, \dots, x_n) \right\} \\ &= \xi_j + t_j^{-1} \hat{p}_j^{(1)} \sum_{l=1}^j \zeta_l \left(\sum_{i=l}^h t_i^{-1} \hat{p}_i^{(1)} \right)^{-1}. \end{aligned}$$

M-step—With the weights obtained in the E-step, we maximize the expected log likelihood conditional on the observed data with Lagrange terms:

$$\ell_1^E = \sum_{j=1}^h w_j^E \log p_j - \lambda_1 \left(\sum_{j=1}^h p_j - 1 \right) - \lambda \sum_{j=1}^h q_\gamma(t_j) p_j. \quad (12)$$

By maximizing equation (12), at the k th step ($k > 1$) we replace $\hat{p}_j^{(k-1)}$ with

$$\hat{p}_j^{(k)} = \frac{w_j^E}{\sum_l w_l^E + \lambda q_\gamma(t_j)},$$

where λ is the solution of the equation

$$\sum_{j=1}^h \frac{w_j^E q_\gamma(t_j)}{\sum_l w_l^E + \lambda q_\gamma(t_j)} = 0.$$

With a given convergence criterion, we can solve p_j iteratively. Let $\hat{p}_j^0$ and $\hat{p}_j^1$ denote the NPMLEs of p_j , $j = 1, \dots, h$ for the log likelihood with and without the extra constraint (ℓ and ℓ_1), respectively. The empirical likelihood ratio statistic, equation (11), can be computed by plugging in two NPMLEs.

When inverting the empirical likelihood ratio test statistic to obtain the corresponding confidence intervals, we use the nonparametric maximum likelihood estimator $\hat{\gamma}$ as the center of the confidence interval. We then search an interval centered around $\hat{\gamma}$ for all possible values of γ to find the lower and upper bounds of the confidence interval. Specifically, by searching the grid in that range, the values of γ that do not lead to the rejection of the chi-squared test at significance level α are included in the confidence interval, and the minimum and the maximum of γ , which lead to the rejection of the likelihood ratio test, are the lower and upper bounds of the confidence interval of $\hat{\gamma}$. Given the unified presentation of the EM algorithm with a general function of interest $q_\gamma(t)$ in this section, we will next describe the estimation of the confidence interval for commonly used summary statistics.

3.3. Confidence interval of the mean

It is of practical importance to estimate the mean of the target population (i.e., $\tilde{X}$) and its confidence interval for right-censored length-biased data. To estimate the confidence interval for the population mean of $\tilde{X}$, we will test the null hypothesis of

$$H_0: \int_0^\tau x dF(x) = \mu, \text{ or } \int_0^\tau q_\mu(x) dG(x) = 0,$$

where $q_\mu(x) = (x - \mu)/x$. By imposing the above constraint to $G(x)$, we have the following equivalent empirical likelihood expression

$$\sum_{j=1}^h p_j q_\mu(t_j) = 0. \quad (13)$$

Using the EM algorithm described in Section 3.2 and replacing $q_{\gamma}(t)$ by $q_{\mu}(t)$, we can maximize the log likelihood with the above constraint, and denote the corresponding NPMLE as $\hat{\mu}^1$, which is a function of μ . We show in the Appendix that as $n \rightarrow \infty$

$$R(\mu) = 2\{\ell_0(\hat{\mu}^0) - \ell_1(\hat{\mu}^1(\mu))\} \xrightarrow{d} \chi^2(1),$$

where “ $\xrightarrow{d}$ ” means convergence in the distribution. Thus the empirical likelihood ratio confidence interval for μ can be obtained by searching all values of μ in which the likelihood ratio statistic $R(\mu)$ falls within the acceptance range of the chi-squared test at the $(1 - \alpha)$ significance level, i.e., $\{\mu \mid R(\mu) \leq Q_{1-\alpha}\}$.

When it is of interest to compare the means of two groups, we can generalize the setting to test the equality of $\mu_1 = \mu_2$ for a two-sample problem by deriving the confidence interval for $\mu_2 - \mu_1$ [16]. Let the null hypothesis be $\Delta = 0$, where $\Delta = \mu_2 - \mu_1$. Denote $t_{11}, \dots, t_{1h_1}$ and $t_{21}, \dots, t_{2h_2}$ as the observed distinct failure and censoring times in the two groups, with the respective underlying biased distribution functions of G_1 and G_2 . Let

$$q_{\mu_1}(x) = \frac{x - \mu_1}{x}, \quad q_{\mu_2}(x) = \frac{x - \mu_2}{x} = \frac{x - \mu_1 - \Delta}{x}.$$

We need to maximize the joint likelihood function as

$$\ell_0(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i=1}^{h_1} w_{1i} \log p_{1i} + \sum_{j=1}^{h_2} w_{2j} \log p_{2j} - \lambda_1 \left(\sum_{i=1}^{h_1} p_{1i} - 1 \right) - \lambda_2 \left(\sum_{j=1}^{h_2} p_{2j} - 1 \right),$$

to solve the MLE of $\mathbf{p}_1 = (p_{11}, \dots, p_{1h_1})$ and $\mathbf{p}_2 = (p_{21}, \dots, p_{2h_2})$ at t_{ki} , $k = 1, 2$ and $i = h_1$ or h_2 for $\ell(\mathbf{p}_1, \mathbf{p}_2)$, which are Vardi's estimates. Similarly, we maximize $\ell(\mathbf{p}_1, \mathbf{p}_2)$ with the additional constraints of

$$\sum_{i=1}^{h_1} p_{1i} q_{\mu_1}(t_{1i}) = 0, \quad \sum_{j=1}^{h_2} p_{2j} q_{\mu_2}(t_{2j}) = 0.$$

Note that the NPMLE of $(\mathbf{p}_1, \mathbf{p}_2)$ solved from ℓ does not depend on μ_1 when μ_1 is also considered to be an unknown parameter. For a given value of Δ , we solve the MLE $(\mathbf{p}_1, \mathbf{p}_2)$ and μ_1 from ℓ_1 . The empirical likelihood ratio statistic for Δ is defined through

$$R(\Delta) = 2\{\ell_0(\hat{\mu}_1^0, \hat{\mu}_2^0) - \ell_1(\hat{\mu}_1^1(\hat{\mu}_1), \hat{\mu}_2^1(\hat{\mu}_1, \Delta))\} = 2\{\max_{\mathbf{p}_1, \mathbf{p}_2} \ell_0(\mathbf{p}_1, \mathbf{p}_2) - \max_{\mathbf{p}_1, \mathbf{p}_2, \mu_1} \ell_1(\mathbf{p}_1, \mathbf{p}_2)\} \xrightarrow{d} \chi^2(1).$$

3.4. Confidence interval of quantiles (median)

To estimate the confidence interval for the population median of $\tilde{X}$, we will test the null hypothesis of

$$H_0: \int_{\theta}^{\tau} dF(x) = 0.5.$$

We impose the extra constraint, which can be equivalently expressed in terms of the empirical likelihood form as

$$\sum_{j=1}^h p_j q_{\theta}(t_j) = 0, \quad \text{where} \quad q_{\theta}(t) = \{I(t > \theta) - 0.5\} / t. \quad (14)$$

Using the EM algorithm described in Section 3.2, we can maximize the log likelihood with the above constraint, which leads to the log-likelihood function $\ell(\hat{\mathbf{p}}^1(\theta))$. We then estimate the $1 - \alpha$ confidence interval for θ by the $1 - \alpha$ quantile of the chi-squared distribution with 1 degree of freedom as follows,

$$R(\theta) = 2\{\ell_0(\hat{\mathbf{p}}^0) - \ell_1(\hat{\mathbf{p}}^1(\theta))\} \xrightarrow{d} \chi^2(1).$$

The confidence interval of the median can be constructed in a similar manner. The method can be easily extended to estimate the $1 - \alpha$ confidence interval for any quantile of $F(\cdot)$ by replacing 0.5 with other probability values in (14).

3.5. Confidence interval of the survival function

The estimation of the point-wise confidence interval of the survival function $S(t)$ of the target population can be derived from the nonparametric likelihood ratio statistic. Note that the survival function of $\tilde{X}$ can be expressed by the biased distribution function $G(t)$ as follows,

$$S(t) = \int_0^{\tau} I(x \geq t) dF(x) = \left\{ \int_0^{\tau} I(x \geq t) x^{-1} dG(x) \right\} \left\{ \int_0^{\tau} u^{-1} dG(u) \right\}^{-1}.$$

The extra constraint in constructing the likelihood ratio test for a fixed t_0 is

$$\int_0^{\tau} \{I(x \geq t_0) - S(t_0)\} dF(x) = 0,$$

which is equivalent to

$$\sum_{j=1}^h p_j q_{S(t_0)}(t_j) = 0, \quad \text{where} \quad q_{S(t_0)}(x) = \{I(x \geq t_0) - S(t_0)\} / x. \quad (15)$$

The corresponding empirical likelihood ratio statistic is

$$R(S(t_0)) = 2\{\ell_0(\hat{\mathbf{p}}^0) - \ell_1(\hat{\mathbf{p}}^1(t_0))\},$$

where $\ell(\hat{\mathbf{p}}^0)$ denotes the maximum log likelihood attained at $\hat{\mathbf{p}}^0$ (i.e., Vardi's estimate), and $\ell(\hat{\mathbf{p}}^1(t_0))$ denotes the maximum log likelihood attained under constraints (5) and (15).

Similarly, we can prove that $R(S(t_0)) \xrightarrow{d} \chi^2(1)$ under constraint (15). Simultaneous confidence bands for the survival function on all the observed time points $t_j, j = 1, \dots, h$ may be constructed in a manner similar to that for traditional right-censored survival data [23]. However, the large sample properties for the estimation of simultaneous confidence bands is much more complex and beyond the scope of this paper.

4. Simulation

We conducted simulation studies to check the performance of the likelihood ratio test confidence intervals under various sample sizes with length-biased data. In each simulation, we generated a sample of i.i.d copies of $(\tilde{X}, \tilde{A})$, where the logarithm of unbiased failure time $\tilde{X}$ was generated identically either from a uniform (0, 1) distribution or from a Normal(1/2, 1/12) distribution, and the truncation time $\tilde{A}$ was generated from a uniform distribution with an upper bound bigger than the upper bound of $\tilde{X}$ to ensure the stationarity assumption. We kept only the pairs satisfying $\tilde{A}_i < \tilde{X}_i$. The residual censoring time measured from the examination time was independently generated from a uniform distribution on $(0, U_c)$, where U_c is chosen to yield a desired censoring percentage: 15% and 30%. The cohort size was chosen to be 100, 200 and 400, representing small, moderate and large sample sizes, respectively. For each scenario, we analyzed the specified cohort of samples 3000 times. The coverage probability was calculated as the proportion of the time that the 95% confidence intervals contain the true value of the parameter out of 3000 repetitions.

Table 1 summarizes the empirical 95% coverage probabilities of the confidence intervals for survival means and medians by using the empirical likelihood ratio test. We also compared the performance of the empirical likelihood ratio test method with the percentile bootstrap confidence interval. The coverage probability of the percentile bootstraps confidence intervals are based on 200 bootstrap resamples. The two methods have similar performance in terms of coverage probabilities, but the bootstrap confidence interval method is much more computationally intensive and time-consuming. For example, with a cohort size of 400 the empirical likelihood ratio test method is 30 times faster than the bootstrap method for calculating one confidence interval of the population mean. The simulation results for the mean differences are listed in Table 2. One can observe that the empirical coverage probabilities are in a reasonable range. As expected, the likelihood ratio test confidence

intervals become narrower as sample size increases, and wider as the censoring percentage increases.

Figure 1 plots the average point-wise 95% confidence intervals for the survival function by using the empirical likelihood ratio test. Again the corresponding empirical 95% coverage probabilities (not shown due to space limitations) are close to the nominal value of 0.95, but tend to be conservative when t is at two tails.

5. A real data example

In this section, we illustrate our proposed method to construct confidence intervals using data from the Canadian Study of Health and Aging (CSHA). The CSHA is a nation-wide observational study of dementia and other health-related problems among elderly citizens in Canada [24, 5]. Among study participants, 1132 individuals having dementia were identified and classified into subcategories of dementia in the first phase of the CSHA. Their dates of censoring or death were collected prospectively during the second or third phase of the CSHA. It is of interest to estimate and compare the mean and median survival times by each diagnosis type. As demonstrated by Wolfson et al. [5], the observed survival time from the onset of dementia is length-biased. This finding was tested by Asgharian et al. and Addona and Wolfson [2, 25].

Our analysis is limited to 818 subjects with a diagnosis of probable Alzheimer's disease, possible Alzheimer's disease, or vascular dementia. The estimated survival mean and median along with the 95 % confidence intervals for each subtype of dementia are listed in Table 3. The results indicate a trend for a shorter survival time for vascular dementia; hence we apply the proposed method to obtain the 95 % confidence intervals for the survival differences between the different subtypes of dementia. We found the following mean differences in survival among the three cohorts: 0.976 years with 95% confidence interval (0.087, 1.765) between probable Alzheimer's disease and vascular dementia; 0.529 years with 95% confidence interval (-0.360, 1.318) between possible Alzheimer's disease and vascular dementia; and 0.447 years with 95% confidence interval (-0.443, 1.336) between possible and probable Alzheimer's disease.

6. Closing remarks

For right-censored length-biased data, the likelihood ratio-based approach provides a practically feasible method for constructing confidence intervals for the survival function and other commonly used summary statistics that can be expressed as functionals of the survival function. The approach is based on the unconditional nonparametric likelihood of Vardi [6], which should be asymptotically efficient. The EM algorithm for the constrained NPMLE was generalized from Vardi's EM algorithm without the constraint. The proposed method performs well in terms of coverage probability; the boundaries of the confidence interval are naturally contained in $[0,1]$ for the survival function. Because of the complex formulae for the variance-covariance estimators of Vardi's NPMLE for length-biased data [9], the empirical likelihood ratio approach for confidence interval estimation is more appealing without requiring calculation of the variance for the summary statistic of interest. In addition, the computation is much less intensive than that of the bootstrap method (e.g.,

30 times faster than the bootstrap method). More importantly, as pointed out by Hall and La Scala [26], the empirical likelihood method has precision, in addition to computation, which is a general advantage over the bootstrap method.

For general left-truncated survival data, many authors [27, 28, 29, 30] had studied the properties of nonparametric distribution estimation for left-truncated data. As noted by Wang [31], the efficiency of these estimators is less than ideal because the estimators are derived from the likelihoods conditioning on the truncation times A , although the sampling bias is properly adjusted in these methods. Specifically, by decomposing the full likelihood to the product of the conditional likelihood of X given truncation time A and the marginal likelihood of A , one has

$$L = \prod_{i=1}^n \frac{f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}}{S(a_i)} \prod_{i=1}^n \frac{S(a_i)}{\mu}.$$

In contrast to the conditional likelihood approach for left-truncated data, our estimation of confidence intervals is based on the full likelihood (expression (3)) using all available data. Asgharian et al. (2002)[8] showed the efficiency gain can be up to four times more when estimating $S(t)$.

For traditional right-censored data, it is well known that the population mean cannot be consistently estimated if the follow-up is not long enough in the incident cohorts. Specifically, if the support of the censoring variable is shorter than that of the failure times. To avoid this complication, the median survival time instead of the mean survival time has been widely used as a summary statistic for traditional right-censored survival data. Unlike traditional survival data, length-biased survival data have a unique feature, such that the population mean and the underlying distribution can be consistently estimated as long as there is a positive follow-up. Intuitively, subjects who are recruited to the study experienced disease onset prior to the recruitment time, which may be considered as “free” follow-up time (i.e., the backward recurrence time) as an extra piece of information to assure the estimability of the mean. Another difference between the Kaplan-Meier estimate for traditional survival data and Vardi’s NPMLE estimate for length-biased data is that Vardi’s estimate has a positive mass at both the right-censored and failure time points; whereas the KM estimate has jumps on the failure times only.

Acknowledgments

The data reported in this article were collected as part of the Canadian Study of Health and Aging. The core study was funded by the Seniors Independence Research Program, through the National Health Research and Development Program (NHRDP) 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 no. 6603-1417-302(R)), Bayer Incorporated, and the British Columbia Health Research Foundation (Projects no. 38 (93-2) and no. 34 (96-1)). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada. This work was supported in part by grants from the National Institute of Health (CA079466 and CA016672) and Natural Sciences and Engineering Research Council of Canada.

References

1. Asgharian M. Biased sampling with right censoring: a note on sun, cui, tiwari(2002). *The Canadian Journal of Statistics*. 2003; 31:349–350.
2. Asgharian M, Wolfson DB, Zhang X. Checking stationarity of the incidence rate using prevalent cohort survival data. *Statist Med*. 2006; 25:1751–1767.
3. Gill RD, Keiding N. Product-limit estimators of the gap time distribution of a renewal process under different sampling patterns. *Lifetime Data Analysis*. 2010; 16:571–579. [PubMed: 20309631]
4. Gordis, L. *Epidemiology*. Philadelphia, PA: W. B. Saunders Company; 2000.
5. Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, Hogan DB. the Clinical Progression of Dementia Study Group. A reevaluation of the duration of survival after the onset of dementia. *New Engl J Med*. 2001; 344:1111–1116. [PubMed: 11297701]
6. Vardi Y. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*. 1989; 76:751–761.
7. Wang MC. Nonparametric estimation from cross-sectional survival data. *J Am Statist Assoc*. 1991; 86:130–143.
8. Asgharian M, M'Lan CE, Wolfson DB. Length-biased sampling with right censoring: An unconditional approach. *J Am Statist Assoc*. 2002; 97:201–209.
9. Asgharian M, Wolfson DB. Asymptotic behavior of the unconditional npmle of the length-biased survivor function from right censored prevalent cohort data. *Ann Statist*. 2005; 33:2109–2131.
10. Ning J, Qin J, Shen Y. Non-parametric tests for right-censored data with biased sampling. *J R Statist Soc B*. 2010; 72:609–630.
11. Gill RD, Keiding N. Product-limit estimators of the gap time distribution of a renewal process under different sampling patterns. *Lifetime Data Analysis*. 2010; 16:571–579. [PubMed: 20309631]
12. Cook RJ, Bergeron PJ. Information in the sample covariate distribution in prevalent cohorts. *Statist Med*. 2011; 30:1397–1409.
13. Madansky A. Approximate confidence limits for reliability of series and parallel systems. *Technometrics*. 1965; 7:495–503.
14. Thomas DR, Grunkemeier GL. Confidence interval estimation of survival probabilities for censored data. *J Amer Statist Assoc*. 1975; 70:865–871.
15. Owen A. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*. 1988; 75:237–249.
16. Qin J. Empirical likelihood in biased sample problems. *Ann of Statist*. 1993; 21:1182–1196.
17. Qin J. Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann Statist*. 1999; 27:1368–1384.
18. Matthews DE. Likelihood-based confidence intervals for functions of many parameters. *Biometrika*. 1988; 75:139–144.
19. Zhou M, Li G. Empirical likelihood analysis of the buckley-james estimator. *J Multi Ana*. 2008; 99:649–664.
20. Zhou M, Jeong JH. Empirical likelihood ratio test for median and mean residual lifetime. *Statistics in Medicine*. 2011; 30:152–159. [PubMed: 21204121]
21. Owen A. Empirical likelihood ratio confidence regions. *Ann Statist*. 1990; 18:90–120.
22. Murphy S, van der Vaart A. Semiparametric likelihood ratio inference. *Ann Statist*. 1997; 25:1471–1509.
23. Hollander N, McKeague I, Yang J. Likelihood ratio-based confidence bands for survival functions. *J Amer Statist Assoc*. 1997; 92 :215–226.
24. Lindsay J, Sykes E, McDowell I, Verreault R, Laurin D. More than the epidemiology of alzheimer's disease: contributions of the canadian study of health and aging. *Can J Psychiatry*. 2004; 49(2):83–91. [PubMed: 15065741]
25. Addona V, Wolfson DB. A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Anal*. 2006; 12:267–274. [PubMed: 16917734]

26. Hall P, Lascala B. Methodology and algorithms of empirical likelihood. *Int Statist Rew.* 1990; 58:109–127.
27. Woodroffe M. Estimating a distribution function with truncated data. *Ann Statist.* 1985; 13:163–177.
28. Wang MC, Jewell NP, Tsai WY. Asymptotic properties of the product limit estimate under random truncation. *Ann Statist.* 1986; 14:1597–1605.
29. Lagakos SW, Barraj LM, De Gruttola V. Nonparametric analysis of truncated survival data, with applications to aids. *Biometrika.* 1988; 75:515–523.
30. Kalbfleisch JD, Lawless JF. Inference based on retrospective ascertainment: analysis of the data on transfusion-related aids. *Journal of the American Statistical Association.* 1989; 84:360C372.
31. Wang, MC. *Encyclopedia of Biostatistics.* 2005. Length Bias.
32. Pan X, Zhou M. Using one-parameter sub-family of distributions in empirical likelihood ratio with censored data. *Journal of Statistical Planning and Inference.* 1999; 75:379–392.

7. Appendix

This section contains an outline of the proofs of the results stated in previous sections.

Asymptotic distribution of the Empirical likelihood

Assume $h(t)$ is another function that is left continuous and $h(t)q_{\gamma}(t) \geq 0$, and then define a class of functions

$$\mathcal{H}_{q_{\gamma}}^G = \{h(\cdot) | h(\cdot) \text{ is left continuous, } \int h^2 dG < \infty, h(t)q_{\gamma}(t) \geq 0\}.$$

We furthermore define a one-parameter distribution function family as follows [32],

$$\mathcal{A}_h^G = \left\{ F_{\lambda}(t) | F_{\lambda}(t) = \sum_{j:t_j \leq t} p_{j,\lambda}, \quad p_{j,\lambda} = \frac{\hat{p}_j^0}{1 + \lambda h(t_j)} \left(\sum_{i=1}^h \frac{\hat{p}_i^0}{1 + \lambda h(t_i)} \right)^{-1} \right\}.$$

We can thus form the log of the empirical likelihood ratio for this family by

$$\mathcal{R}_h = \{ \ell(F_{\lambda}) - \ell(\hat{\mathbf{p}}^0) | F_{\lambda} \in \mathcal{A}_h^G \}.$$

Note that the constraint equation $\sum_{i=1}^h q_{\gamma}(t_j) p_{j,\lambda} = 0$ has one unique solution, denoted as λ_0 .

Applying Taylor's expansion for $\sum_{i=1}^h q_{\gamma}(t_j) p_{j,\lambda_0}$, We have

$$\sum_{j=1}^h \frac{q_{\gamma}(t_j) \hat{p}_j^0}{1 + \lambda_0 h(t_j)} = \sum_{j=1}^h q_{\gamma}(t_j) \hat{p}_j^0 - \lambda_0 \sum_{j=1}^h q_{\gamma}(t_j) h(t_j) \hat{p}_j^0 + o(\lambda_0). \quad (16)$$

From equation (16) we have

$$\lambda_0 = \frac{\sum_{j=1}^h q_\gamma(t_j) \hat{p}_j^0}{\sum_{j=1}^h q_\gamma(t_j) h(t_j) \hat{p}_j^0} + o_p(n^{-1/2}).$$

By using the large sample properties of Vardi's estimator [9], it can be shown, as $n \rightarrow \infty$,

$$\sqrt{n} \left(\sum_{j=1}^h q_\gamma(t_j) \hat{p}_j^0 \right) \xrightarrow{d} N(0, \sigma_{q_\gamma}^2)$$

and

$$\sum_{j=1}^h q_\gamma(t_j) h(t_j) \hat{p}_j^0 \xrightarrow{P} \int_0^\tau q_\gamma(t) h(t) dG(t).$$

By Slutsky's theorem, $n\lambda_0^2 \rightarrow c_{h1}\chi^2(1)$ in distribution, where

$c_{h1} = \sigma_{q_\gamma}^2 / (\int_0^\tau q_\gamma(t) h(t) dG(t))^2$. Define the log empirical likelihood with respect to $p_{j,\lambda}$ as

$$\ell(\lambda) = \log \prod_{j=1}^h p_{j,\lambda}^{\xi_j} \left(\sum_{k=j}^h t_k^{-1} p_{k,\lambda} \right)^{\zeta_j}.$$

From this definition $\ell(0)$ is the log empirical likelihood without extra constraints. We apply Taylor's expansion for $\ell(\lambda)$ around λ_0 ,

$$\ell(\lambda_0) - \ell(0) = \lambda_0 \ell'(0) + \frac{\lambda_0^2}{2} \ell''(0) + o_p(1),$$

where $\ell(\lambda)$ and its first and second derivatives of $\ell(\lambda)$ are,

$$\begin{aligned} \ell(\lambda) &= \sum_{j=1}^h \xi_j \log \hat{p}_j^0 - \sum_{j=1}^h \xi_j \log(1 + \lambda h(t_j)) - n \log \left\{ \sum_{i=1}^h \frac{\hat{p}_i^0}{1 + \lambda h(t_i)} \right\} + \sum_{j=1}^h \zeta_j \log \left\{ \sum_{k=j}^h t_k^{-1} \hat{p}_k^0 / (1 + \lambda h(t_k)) \right\} \\ \ell'(\lambda) &= - \sum_{j=1}^h \frac{\xi_j h(t_j)}{1 + \lambda h(t_j)} + \frac{n \sum_{k=1}^h \hat{p}_k^0 h(t_k) / (1 + \lambda h(t_k))^2}{\sum_{k=1}^h \hat{p}_k^0 / (1 + \lambda h(t_k))} - \sum_{j=1}^h \zeta_j \frac{\sum_{k=j}^h t_k^{-1} \hat{p}_k^0 h(t_k) / (1 + \lambda h(t_k))^2}{\sum_{k=j}^h t_k^{-1} \hat{p}_k^0 / (1 + \lambda h(t_k))} \end{aligned} \quad (17)$$

$$\begin{aligned} \ell''(\lambda) = & \sum_{j=1}^h \frac{\xi_j h^2(t_j)}{\{1+\lambda h(t_j)\}^2} - \sum_{j=1}^h \zeta_j \frac{\left\{ \sum_{k=j}^h t_k^{-1} \hat{p}_k^0 h(t_k) / (1+\lambda h(t_k))^2 \right\}^2}{\left\{ \sum_{k=j}^h t_k^{-1} \hat{p}_k^0 / (1+\lambda h(t_k)) \right\}^2} + n \frac{\left\{ \sum_{k=1}^h \hat{p}_k^0 h(t_k) / (1+\lambda h(t_k))^2 \right\}^2}{\left\{ \sum_{k=1}^h \hat{p}_k^0 / (1+\lambda h(t_k)) \right\}^2} \\ & + 2 \sum_{j=1}^h \zeta_j \frac{\sum_{k=j}^h t_k^{-1} \hat{p}_k^0 h^2(t_k) / (1+\lambda h(t_k))^3}{\sum_{k=j}^h t_k^{-1} \hat{p}_k^0 / (1+\lambda h(t_k))} - 2n \frac{\sum_{k=1}^h \hat{p}_k^0 h^2(t_k) / (1+\lambda h(t_k))^3}{\sum_{k=1}^h \hat{p}_k^0 / (1+\lambda h(t_k))}. \end{aligned} \quad (18)$$

Equations (17) and (18) show that $\ell^0(0) = 0$ and the second derivative evaluated at 0 is

$$\ell''(0) = \sum_{j=1}^h \xi_j h^2(t_j) - \sum_{j=1}^h \zeta_j \frac{\left\{ \sum_{k=j}^h t_k^{-1} \hat{p}_k^0 h(t_k) \right\}^2}{\left\{ \sum_{k=j}^h t_k^{-1} \hat{p}_k^0 \right\}^2} + 2 \sum_{j=1}^h \zeta_j \frac{\sum_{k=j}^h t_k^{-1} \hat{p}_k^0 h^2(t_k)}{\sum_{k=j}^h t_k^{-1} \hat{p}_k^0} + n \left(\sum_{j=1}^h \hat{p}_j^0 h(t_j) \right)^2 - 2n \sum_{j=1}^h \hat{p}_j^0 h^2(t_j). \quad (19)$$

Hence the large sample properties of Vardi's estimator [9] imply $-\frac{1}{n} \ell''(0) \rightarrow c_{h2}$ in probability, where

$$c_{h2} = \int_0^\tau h^2(t) / t \int_0^t (1 - F_{C_0}(s)) ds dG(t) + \int_0^\tau \frac{\left\{ \int_x^\tau 1/s h(s) dG(s) \right\}^2}{\int_x^\tau 1/s dG(s)} dF_{C_0}(x) - \left(\int_0^\tau h(t) dG(t) \right)^2$$

By Slutsky theorem, $-2\mathcal{R}_h \xrightarrow{d} c_{h1} c_{h2} \chi^2(1)$. Then, by arguments similar to those of Pan and Zhou[32] and Zhou and Li [19], it can be shown that the infimum of the constant $c_{h1} c_{h2}$ over h is one, $\inf_h c_{h1} c_{h2} = 1$. Then the asymptotic distribution of $R(\gamma)$ follows from the same argument as that shown in Pan and Zhou [32].

Figure 1.
Plot of estimated survival function with 95% Confidence intervals

Table 1
Empirical coverage probabilities of confidence intervals for survival means and medians

Distribution	Cohort Size	Cen%	μ	$\hat{\mu}$	95% CP_E	95% CP_B	θ	$\hat{\theta}$	95% CP_E	95% CP_B
Log-uniform	100	15	2.833	2.833	0.981	0.937	2.718	2.723	0.936	0.940
		30		2.834	0.982	0.940		2.722	0.912	0.940
	200	15		2.834	0.923	0.940		2.721	0.973	0.974
		30		2.834	0.925	0.943		2.719	0.983	0.929
Log-normal	400	15		2.833	0.957	0.944		2.721	0.975	0.940
		30		2.833	0.946	0.934		2.719	0.982	0.921
	100	15	2.834	2.836	0.979	0.932	2.718	2.721	0.944	0.940
		30		2.836	0.978	0.927		2.721	0.941	0.936
	200	15		2.837	0.964	0.936		2.724	0.958	0.942
		30		2.835	0.936	0.947		2.719	0.940	0.938
	400	15		2.834	0.947	0.935		2.719	0.964	0.938
		30		2.834	0.944	0.933		2.719	0.958	0.935

95% CP_E : Coverage probability of the 95% confidence interval by empirical likelihood method; 95% CP_B : Coverage probability of the 95% confidence interval by bootstrap method.

Table 2
Empirical coverage probabilities of confidence intervals for survival mean differences by empirical likelihood method

Distribution	Cohort Size	Cen%	$\hat{\alpha}$	95%CP	Length of CI
Log-uniform	100	15	1.838	0.913	0.597
		30	1.840	0.899	0.715
	200	15	1.836	0.938	0.556
		30	1.836	0.910	0.583
	400	15	1.840	0.979	0.537
		30	1.840	0.968	0.574
Log-normal	100	15	1.840	0.914	0.574
		30	1.832	0.901	0.719
	200	15	1.838	0.937	0.547
		30	1.838	0.918	0.592
	400	15	1.839	0.974	0.527
		30	1.838	0.963	0.549

Table 3

Empirical likelihood ratio test confidence intervals of survival means (years) and medians (years) for dementia data

Subtype	Mean with 95%CI	Median with 95%CI
Vascular Dementia	4.337(3.748, 4.927)	3.279(2.290, 4.268)
Probable Alzheimer's Disease	4.867(4.178, 5.356)	4.074(3.285, 4.663)
Possible Alzheimer's Disease	5.313(4.624, 5.903)	4.202(3.513, 4.892)