HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2018 January 27.

Published in final edited form as:

Stat Methods Med Res. 2018 February ; 27(2): 384–397. doi:10.1177/0962280216630342.

A robust semi-parametric warping estimator of the survivor function with an application to two-group comparisons

Alan D Hutson

Department of Biostatistics, University at Buffalo, USA

Abstract

In this note, we develop a new and novel semi-parametric estimator of the survival curve that is comparable to the product-limit estimator under very relaxed assumptions. The estimator is based on a beta parametrization that *warps* the empirical distribution of the observed censored and uncensored data. The parameters are obtained using a pseudo-maximum likelihood approach adjusting the survival curve accounting for the censored observations. In the univariate setting, the new estimator tends to better extend the range of the survival estimation given a high degree of censoring. However, the key feature of this paper is that we develop a new two-group semi-parametric exact permutation test for comparing survival curves that is generally superior to the classic log-rank and Wilcoxon tests and provides the best global power across a variety of alternatives. The new test is readily extended to the *k* group setting.

Keywords

Product-limit estimator; log-rank test; Wilcoxon-rank sum test; censored data

1 Introduction

The product-limit estimator[1] of the survival function is one of the most popular and well-studied methods for estimating the survival distribution. The original paper[1] is one of the most cited statistical papers of all time.[2] Its deficiencies are also well-documented, particular as it pertains to a large proportion of extreme observations being right censored.[3,4] Even so, the product limit estimator has great utility as a descriptor of data with support on the positive real line given a right censoring mechanism and reduces nicely to the classic empirical distribution function estimator given no censoring.

The convergence properties of the product-limit estimator are technically difficult to derive, but have been thoroughly investigated, e.g. see Chen and Lo[5] who proves the product-limit estimator converges in probability to the population survival function under certain conditions within the range of the observed data. In this note we will introduce a new and

Corresponding author: Alan D Hutson, Department of Biostatistics, University at Buffalo, 706 Kimball Tower, 3435 Main St., Buffalo, NY 14214-3000, USA. ahutson@buffalo.edu.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

novel semi-parametric "warping" estimator of survival function, which mimics very closely the behavior of the product-limit estimator and serves as a basis for a new method for estimating the survival function and generating a new two-group test for comparing two survival curves.

As background in terms of outlining the derivation of the product limit estimator of the survival function we start with standard notation. Towards this end, let $X_1, X_2, \ldots, X_n$ denote i.i.d. failure times and let $C_1, C_2, \ldots, C_n$ denote the corresponding i.i.d. noninformative right censoring times, $i = 1, 2, \ldots, n$. Given right censoring we only observe $g \leq n$ of the X's. Now let $0 < x_{(1)} < x_{(2)} < \cdots x_{(g)}$ be the distinct (no ties) ordered observed failure times. The classic maximum likelihood based derivation of the product-limit estimator starts by assuming the underlying distribution is discrete with probabilities $\pi_j = P(X = x_{(j)})$, $j = 1, 2, \ldots, g$ for $g \leq n$. Given a discrete hazard of $h_j = P(X = x_{(j)} | X \geq x_{(j)})$ for $0 \leq h_j \leq 1$ we have that $\pi_1 = h_1$, $\pi_2 = (1 - h_1)h_2, \cdots, \pi_g = (1 - h_1)(1 - h_2)\cdots(1 - h_{g-1})h_g$. Then the estimator of $S(x) = P(X > x) = \prod_{x_{(j)} \leq x}(1 - h_j)$ in the discrete case is given as

$$\widehat{S}(x) = \prod_{x_{(j)} \leq x}(1 - \widehat{h}_j) \quad (1)$$

where the estimates of the discrete hazard parameters follow from the likelihood[6]

$$log\, L = \sum_{j=1}^{g} d_j log\, h_j + (r_j - d_j) log(1 - h_j) \quad (2)$$

where d_j denotes the number of events and r_j denotes the number at risk at time $x_{(j)}$, $j = 1, 2, \cdots, g$. The maximization of (2) with respect to the parameters h_j, $j = 1, 2, \cdots, g$ yields the estimates $\widehat{h}_j = d_j / r_j$, where r_j denotes the number of subjects at risk at time $x_{(j)}$ and d_j denotes the number of subjects who fail at time $x_{(j)}$ For a technical treatment with respect to the behavior of the product-limit estimator and how it translates to continuous case see Chen and Lo.[5] It is well-known, but not immediately obvious, that the product-limit estimator reduces to the classic empirical estimator of the survival function $\widehat{S}(x) = 1 - \sum_{i=1}^{n} I(x_{(i)} \leq x)/n$ when there are no censored observations, where $I(\cdot)$ denotes the indicator function. We will be comparing our new estimator to that of (2) throughout this note.

Two common methods for comparing survival curves that go hand-in-hand with the product-limit estimator are the ubiquitous log-rank test[7] and the generalization of the Wilcoxon rank-sum test.[8] The log-rank test is fully efficient against alternatives in which the hazard rates are proportional across time. The generalized Wilcoxon test weighs comparisons of the survival curve at an earlier point in time. In this note we develop a new two-group rank based comparison based on our semi-parametric warped estimator of the quantile function and show that in many common scenarios it outperforms the classic tests in terms of relative

efficiency and hence could reduce the cost of carrying forth large-scale phase III trials, which incorporate a time-to-event endpoint.

In section 2, we develop the new warped univariate survival function estimator. In section 3, we provide a small simulation study comparing the product-limit estimator to the new estimator. In section 4, we provide a basic data example using a well-known dataset. In section 5, we develop an exact permutation test and its corresponding approximation based on pseudo-likelihood methods for comparing two survival curves. In section 6, we provide a two-group example. Finally, in section 7, we provide a simulation study for the new test and compare it to the classic log-rank and Wilcoxon tests most commonly found in statistical software packages.

2 Survival function estimator

Prior to developing our method we first define the common survival analysis notation. As with the product-limit estimator let $X_1, X_2, \cdots, X_n$ denote the i.i.d. absolutely continuous failure times and let $C_1, C_2, \cdots, C_n$ denote the corresponding i.i.d. absolutely continuous noninformative right censoring times such that we observe $T_i = \min(X_i, C_i)$ and censoring indicator $\delta_i = I_{(X_i < C_i)}$, $i = 1, 2, \cdots, n$. Within the context of the methods of this section we assume X and C have support over the entire positive real line. It is well known that the survival function for T is equal to the product of the survival functions for X and C, and given as

$$S_T(t) = S_X(t) S_C(t) \quad (3)$$

where the corresponding distribution functions are denoted as $F_T(t) = 1 - S_T(t)$, $F_X(t) = 1 - S_X(t)$ and $F_C(t) = 1 - S_C(t)$, respectively. Hence in one sense the distance between $S_T(t)$ and $S_X(t)$, and $F_T(t)$ and $F_X(t)$, is dictated by censoring survivor function $S_C(t)$ (or F_C), i.e. as the $P(S_X(t) < S_C(t))$ increases the closer $S_T(t)$ and $S_X(t)$, and $F_T(t)$ and $F_X(t)$, are in terms of absolute distances over the range of t. As is well known, only the T_i's are completely observed in the standard survival analysis framework as compared to the X_i's unless there are no censored observations (or vice versa with the C_i's), which is the basis of our approach, i.e. we can use the observed T_i's to estimate $S_X(t)$ directly and accurately, and is there any benefit for taking such an approach given the known estimation methods already in existence such as the product-limit estimator. The foundation of our approach is that theoretically, for a known distribution function $F_T(t)$, we should be able to *warp* $F_T(t)$ such that

$$F_X(t) \approx g(F_T(t)) \quad (4)$$

for some function g under some basic assumptions.

In terms of the choice for g there are several candidates that maintain the mapping of the (0,1) to (0,1) probability space in a monotone continuous one-to-one fashion. For our

purpose we chose the beta distribution function $g(x)=B_{p,q}(x)=\int_0^x t^{p-1}(1-t)^{q-1}/\beta(p,q)\mathrm{d}t$, where $\beta_{p,q}(x)=\int_0^1 t^{p-1}(1-t)^{q-1}\mathrm{d}t$. In this context, for known $F_T(t)$, we can define a class of parametric distribution and survival functions as follows

$$F_X(t)=B_{p,q}[F_T(t)] \quad (5)$$

$$S_X(t)=1-B_{p,q}[F_T(t)] \quad (6)$$

with corresponding density and quantile functions given as

$$f_X(t)=b_{p,q}[F_T(t)]f_T(t) \quad (7)$$

$$P_X(u)=Q_T[B_{p,q}^{-1}(u)] \quad (8)$$

where $b_{p,q}(\cdot)$ denotes a beta density, $B_{p,q}^{-1}(u)$ denotes the beta distribution quantile function, and $Q_T(u)=F_T^{-1}(u)$, $0 < u < 1$. The parameters of interest that are to be estimated are p and q. The case of no censored observations corresponds to $p = 1$ and $q = 1$.

As an example used to illustrate our foundational concept, we simulated pairs of X's and C's from sample sizes $n = 10, 100, 1000, 10{,}000$ each from independent standard exponential distribution functions, where $S_X(t) = S_C(t) = \exp(-t)$, which yields a known $S_T(t) = \exp(-2t)$ as the underlying truth. Suppose we modeled the data assuming in a standard fashion $S_X(t) = \exp(-\theta t)$ using standard maximum likelihood techniques versus warping an assumed known distribution function for T given as $F_T(t) = 1 - \exp(-2t)$ with density $f_T(t) = 2\exp(-2t)$ such that the warping distribution and density functions for X using the definitions above are given as $F_X(t) = B_{p,q}(1 - \exp(-2t))$ and $f_X(t) = 2b_{p,q}(1 - \exp(-2t))\exp(-2t)$, respectively. We then can estimate p and q using maximum likelihood using the likelihood function

$$L=\prod_{i=1}^{n} f_X(t_i)^{\delta_i}(1-F_X(t_i))^{1-\delta_i} \quad (9)$$

$$=\prod_{i=1}^{n}\{b_{p,q}[F_T(t_i)]f_T(t_i)\}^{\delta_i}(1-B_{p,q}[F_T(t_i)])^{1-\delta_i} \quad (10)$$

For this scenario, with 1000 Monte Carlo simulations per sample size, we calculated the average maximum absolute difference and simulation standard error between

$\widehat{S}_X(t)=exp(-\widehat{\theta}_t)$ and $\widehat{S}_X(t)=1-B_{\widehat{p},\widehat{q}}(1-exp(-2t))$ for t = 0 to 2 by 0.01. The results are presented in Table 1. We can see heuristically that the two distributions are converging over the range of t we selected as the sample size increases. This is a typical result for this model across various parametric families. Philosophically, one does not know which model may be driving the underlying stochastic process. It is certainly reasonable to consider the functional parametric form of the density $g_X(t) = 2b_{p,q}(1-\exp(-2t))\exp(-2t)$ as a reasonable choice among many choices of parametric models. However, this point is not the overarching goal of this note, where in this model can be modified to develop a strong semi-parametric approximation approach for general and flexible modeling of a survival process as given below.

Semi-parametric warped survival model

Instead of assuming a known continuous parametric form for $F_T(t)$ as in equations (5) to (8) we now estimate $F_T(t)$ using a rescaled version of the empirical distribution function estimator given as $\widehat{F}_T(t)=\sum_{i=1}^{n} I_{(t\leq t_i)}/(n+1)$, which maintains the same large sample properties as the classic empirical distribution function estimator. We also utilize $\widehat{f}_T(t)=1/n$ in a standard fashion. Rescaling $\widehat{F}_T(t)$ avoids the difficulties of applying it as the argument within the beta density and distribution functions as part of our pseudo-likelihood estimation procedure, described below, relative to obtaining estimates for p and q. In the trivial case of no censoring this approach results in distribution and survival function estimators with a $1/(n + 1)$ step between ordered observations, corresponding to the expected value between successive ordered uniform (0,1) observations.

Now we have our semi-parametric forms for the estimators of the distribution, survival and density functions given as

$$\widehat{F}_X(t)=B_{\widehat{p},\widehat{q}}\left(\widehat{F}_T(t)\right) \quad (11)$$

$$\widehat{S}_X(t)=1-B_{\widehat{p},\widehat{q}}\left(\widehat{F}_T(t)\right) \quad (12)$$

$$\widehat{f}_X(t)=b_{\widehat{p},\widehat{q}}\left(\widehat{F}_T(t)\right)/n \quad (13)$$

respectively, where as mentioned earlier $B_{p,q}(\cdot)$ denotes a beta distribution function and $b_{p,q}(\cdot)$ denotes a beta density. We will illustrate that this model behaves quite similarly to the product-limit estimator when the censoring distribution has positive support. Some modifications are necessary for shifted censoring distributions and are given later in this section. The next step is to outline how to obtain estimates for p and q.

Pseudo-likelihood estimator for the parameters *p* and *q*

In essence, we have a model for the rankits of X through T accounting for censoring versus using the raw data, e.g. suppose T = (3, 6, 4, 2*), where * denotes a censored observation, then the corresponding transformed data is given by the uniform rankits through $\widehat{F}_T(t)$ as 2/5, 4/5, 3/5, 1*/5. Basically, we are fitting a beta density on the rankits accounting for censored observations and that is why we term this model a semi-parametric type model. The pseudo-likelihood function for our semi-parametric rankit-based approach therefore has the form

$$L=\prod_{i=1}^{n} b_{p,q}\left[\widehat{F}_T(t_i)\right]^{\delta_i}\left(1-B_{p,q}\left[\widehat{F}_T(t_i)\right]\right)^{1-\delta_i} \quad (14)$$

since $\widehat{f}_T(t_i)=1/n$ is constant over all *i*. The maximization of L at (14) or log L relative to finding $\widehat{p}$ and $\widehat{q}$ follows as per standard maximum likelihood methods. Standard numerical software routines such as SAS PROC NLMIXED (SAS Institute, Cary, NC) may be used for the purpose of finding the values for $\widehat{p}$ and $\widehat{q}$. In general, we recommend a grid search of a range of starting values corresponding to *p* and *q*. The properties of the pseudo-likelihood estimators are given as follows.

Theorem 1—As $n\rightarrow\infty$, $\sqrt{n}(\widehat{p}-p,\widehat{q}-q)$ has a centered bivariate normal distribution with variance–covariance matrix $B^{-1}\Sigma B^{-1}$, where B is the standard maximum likelihood based information matrix associated with the standard beta, $b_{p,q}$, density and Σ is the variance–covariance matrix of a two-dimensional random vector whose components are given by

$$\partial\delta_1 log\left\{b_{p,q}[F_T(T_1)]\right\}+(1-\delta_1)log\left\{(1-B_{p,q}[F_T(T_1)])\right\}/\partial p+W_p(T_1) \quad (15)$$

$$\partial\delta_1 log\left\{b_{p,q}[F_T(T_1)]\right\}+(1-\delta_1)log\left\{(1-B_{p,q}[F_T(T_1)])\right\}/\partial q+W_q(T_1) \quad (16)$$

where $\delta_1=I_{(X_1<C_1)}$ and

$$W_p(T_1)=\int I_{(F(X_1)<u)}\frac{\partial^2}{\partial p\partial u}log(b_{p,q}(u))\mathrm{d}B_{p,q}(u) \quad (17)$$

$$W_q(T_1)=\int I_{(F(X_1)<u)}\frac{\partial^2}{\partial q\partial u}log(b_{p,q}(u))\mathrm{d}B_{p,q}(u) \quad (18)$$

Proof—The technical details have been worked out in an elegant fashion for the case of a semi-parametric copula model with marginal distribution functions estimated by the empirical distribution function estimator.[9] The result in Theorem 1 follows directly from the theoretical developments used in the copula approach in section 4 of the copula paper[9] by simply replacing the multivariate copula function with the univariate beta density, which is essentially a special case of the higher dimension copula model. Estimates of the variance-covariance matrix $B^{-1}\Sigma B^{-1}$ are not as straightforward to obtain and we recommend bootstrap resampling for this purpose.

Note that pseudo-likelihood approach is not that dissimilar to the likelihood approach to estimation for the product-limit estimator when considering the alternative approach to that given at (2). If we denote the ordered observed T_i's as $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ and let $\pi_i = P(X \leq t_{(i)} \,|\, \delta_i = 1) - P(X \leq t_{(i)} \,|\, \delta_i = 1)$. The likelihood for the product-limit estimator is rank based in the T's and can be alternative expressed as

$$L = \prod_{i=1}^{n-1} \pi_i^{\delta_i} \left(1 - \sum_{j=1}^{i} \pi_j\right)^{1-\delta_i} \times \left(1 - \sum_{j=1}^{n-1} \pi_j\right) \quad (19)$$

In fact the product-limit estimator and the semi-parametric warped estimator behave very similarly, which should not be that surprising given the functional forms of the likelihoods above. However, the semi-parametric warped based estimator will have some advantages over the product-limit estimator in terms of inference and in the face of heavy censoring proportions.

Example—As an illustration of the new semi-parametric warp estimation method we simulated $n = 1000$ pairs of failure and censoring times with $X \sim \exp(1)$ and $C \sim \exp(1/2)$. From equation (14) we obtained $\hat{p} = 0.96$ and $\hat{q} = 0.34$. We plotted the survival curve estimators for the semi-parametric warped estimator overlayed with the product-limit estimator and the true underlying survival function at $S(t_{(i)}) = exp(-t_{(i)})$, $i = 1, 2, \cdots, 1000$, where $t_{(i)}$ denotes the ith ordered observation. The plot depicted in Figure 1 provides a glimpse into the large sample behavior of the estimator being similar to the product-limit estimator. Careful inspection of the tail shows that the last value for $\widehat{S}(t)$ is at $t = 1.74$, while the semi-parametric warp estimator has a value for $\widehat{S}(t)$ over the range of the t's at $t = 2.3$. This gives a glimpse in terms of the potential advantage of the new estimator in the face of heavy censoring in terms of extending the estimator a bit further of the range of T. The reason for this behavior is that the new estimator is based on a transformation of a proper empirical distribution function $\widehat{F}_T$. Hence, there will always be corresponding values for $\widehat{S}_X(t)$ even if a proportion of the last observations are censored as compared to the product limit estimator. The larger the number of censored observations at the tail of the distribution the larger the range of estimates for $\widehat{S}_X(t)$ are available for the semi-parametric warped estimator as compared to the product limit estimator. As a simple illustration, let our contrived data set consist of observed times 1+,2,3,4+,5+ with "+" denoting a right-censored

observation. Then the product limit estimator has estimates of $\widehat{S}_X(1)=1$, $\widehat{S}_X(2)=.75$, $\widehat{S}_X(3)=.5$, $\widehat{S}_X(4)=.$: (or .5 depending on the definition) and $\widehat{S}_X(5)=.$: (or 0 depending upon the definition). The semi-parametric warped estimator has estimates $\widehat{S}_X(1)=.96$, $\widehat{S}_X(2)=.88$, $\widehat{S}_X(3)=.75$, $\widehat{S}_X(4)=.58$ and $\widehat{S}_X(5)=.35$.

Modification for shifted censoring distribution

A modification of the distribution function, survival function and density function at equations (5) to (8) is necessary when C has known support (d, ∞]. This may occur for example in a clinical trial with a survival endpoint with a minimum defined follow-up time, e.g. 5 years, such that $d = 5$. Modification for shifted censoring distributions for the distribution function, survival function and density function for a known d take the forms

$$F_X(t)=\begin{cases} F_T(t), & \text{if } t \le d \\ \beta B_{p,q}(F_T(t)), & \text{if } t>d \end{cases} \quad (20)$$

$$S_X(t)=\begin{cases} S_T(t), & \text{if } t \le d \\ 1-B_{p,q}(F_T(t)), & \text{if } t>d \end{cases} \quad (21)$$

$$f_X(t)=\begin{cases} f_T(t), & \text{if } t \le d \\ \beta b_{p,q}(F_T(t))f_T(t), & \text{if } t>d \end{cases} \quad (22)$$

where

$$\beta=\frac{F_T(d)}{B_{p,q}(F_T(d))}. \quad (23)$$

The semi-parametric estimator follows as before by first replacing $F_T(t)$ with $\widehat{F}(t)=\sum_{i=1}^{n} I_{(t\le t_i)}/(n+1)$, $\widehat{f}_T(t)=1/n$ and applying the same likelihood function as given at (14) and noting that the warping only occurs after d. This is similar to fitting a truncated distribution to the observations greater than d.

Example—As an illustration of the new semi-parametric warp estimation method relative to a shifted distribution we simulated $n = 1000$ pairs of failure and censoring times with $X \sim$ exp(1) and $C \sim U(1, 2)$, such that $d = 1$. From equation (14) and using the forms of the density and survivor functions at (22) and (21) we obtained $\widehat{p}=0.48$ and $\widehat{q}=0.78$ to account for the warped part of the distribution. We plotted the survival curve estimators for the semi-parametric warped estimator overlayed with the product-limit estimator. The plot depicted in Figure 2 provides a glimpse into the large sample behavior of the estimator being similar to

the product-limit estimator. As in the nonshift example earlier the warped estimator provides a more extended estimator relative to the range of T. The two distributions are virtually equivalent for $T < d$ with step sizes of $1/n$ versus $1/(n + 1)$, respectively.

3 Univariate simulation study

In this section, we illustrate how well our semi-parametric warped survival model fits to a general set of distributions when compared to the product-limit estimator versus simulating directly from a known true model defined at (5). The point being is to demonstrate the new semi-parametric warped survival estimator is well-suited for general estimation purposes.

We chose to simulate data of sample sizes $n = 50, 200, 500$ and computed the mean squared error (MSE) for the lower, middle, and upper quartiles, $Q(1/4)$, $Q(1/2)$, and $Q(3/4)$, respectively from the Weibull(α, λ) family of distributions with distribution function given as $F(t)=1-e^{-(x/\lambda)^{\alpha}}$. The censoring distribution was give as a standard exponential distribution. The results based on 1000 Monte Carlo replications are given in Table 2. In addition, the proportion of times an estimator was available is provided, i.e. the value for a given quartile may be missing if the last observation or observations are censored. The MSE values are based solely on observed values. The quantile estimator corresponding to the warped survival model is given as

$$\widehat{Q}_X(u)=\widehat{F}_T^{-1}(B^{-1}_{\widehat{p},\widehat{q}}(u)) \quad (24)$$

where $\widehat{F}_T^{-1}(\cdot)$ denotes the empirical quantile function for T, $B^{-1}_{p,q}(\cdot)$ is the beta quantile function and $\widehat{p}$ and $\widehat{q}$ come from the psuedo-likelihood estimation at (14).

As seen from Table 2 the warped estimator provides an estimate of the quartiles a greater proportion of the time and consistently has a better MSE than that of the product-limit estimator when the estimator is available the majority of simulations. The exception for having worse performance relative to the MSE is when the product-limit estimator does not produce a quartile estimator in substantial proportion of cases. If we were to use the maximum observation[5] as per large sample theory the MSE values would be similar to that of the warped estimator. As can be seen, this phenomenon occurs even for large samples in terms of not producing a traditional estimator for the upper quartile for the product-limit estimator whereas the warped estimator provided an upper quartile estimate 100% of the time for moderate to large samples with reasonable MSE values.

4 Univariate data example

In order to illustrate our warped survival estimator versus the product-limit estimator we utilized a textbook example[10] of $n = 43$ observations from a surgically placed catheter group and fit time-to-infection. The data are as follows:

> *Infection times*: 1.5, 3.5, 4.5, 4.5, 5.5, 8.5, 8.5, 9.5, 10.5, 11.5, 15.5, 16.5, 18.5, 23.5, 26.5

Censored observations: 2.5, 2.5, 3.5, 3.5, 3.5, 4.5, 5.5, 6.5, 6.5, 7.5, 7.5, 7.5, 7.5, 8.5, 9.5, 10.5, 11.5, 12.5, 12.5, 13.5, 14.5, 14.5, 21.5, 21.5, 22.5, 22.5, 25.5, 27.5

The plot of the two curves is overlayed in Figure 3. The warped curve is more refined as it has more "steps". The quartile estimates from the warped survival curve based on (24) along with the corresponding 95% bootstrap percentile confidence intervals based on 10,000 resamples are given as $\widehat{Q}(1/4)=9.5(7.5,\ 14.5)$, $\widehat{Q}(1/2)=21.5(13.5,\ 25.5)$, and $\widehat{Q}(3/4)=25.5(21.5,.)$. Similarly, for the product-limit estimator based on standard methods we have the point estimators and the corresponding 95% confidence intervals given as $\widehat{Q}(1/4)=10.5(5.5,\ 16.5)$, $\widehat{Q}(1/2)=18.5(11.5,.)$, and $\widehat{Q}(3/4)=26.5(23.5,.)$. Practically speaking tied observations are not an issue since the empirical estimator for the observed values can accommodate ties readily.

5 Two sample test for survival curve differences

In this section, we develop a new and novel two-group pseudo-likelihood ratio test for testing $H_0 : S_X(t) = S_Y(t)$ versus $H_1 : S_X(t) \neq S_Y(t)$ for populations corresponding to the random variables X and Y. The new test easily extends to the k group setting and utilizes the semi-parametric warped estimator developed above at its core. The test follows along the lines of the Wilcoxon rank-sum test, for which the test statistic is a function of the pooled ranks between X and Y. This test has exact type I error control and can be applied for general two and k group comparisons.

Towards this end let $X_1, X_2, \cdots, X_{n_x}$ denote i.i.d. absolutely continuous failure times and let $C_1, C_2, \cdots, C_{n_x}$ denote the corresponding i.i.d. absolutely continuous non-informative right censoring times such that we observe $R_i = \min(X_i, C_i)$ and $\delta_i^x = I_{(X_i<C_i)}$, $i = 1, 2, \cdots, n_x$ from group 1 and let $Y_1, Y_2, \cdots, Y_{n_y}$ denote the i.i.d. absolutely continuous failure times and let $D_1, D_2, \cdots, D_{n_y}$ denote the corresponding i.i.d. absolutely continuous non-informative right censoring times such that we observe $T_j = \min(Y_j, D_j)$ and $\delta_j^y = I_{(Y_j<D_j)}$, $j = 1, 2, \cdots, n_y$. Let the total sample size be denoted as $n = n_x + n_y$. The key assumption for this test is that the censoring distributions for group 1 and group 2 are equivalent, i.e. $F_C = F_D$, such as what might be seen in a randomized clinical trial setting.

In the semi-parametric warped setting, our test for the equivalence of survival curves and/or distribution functions may now be written compactly as

$$H_0: B_{p_x,q_x}[F_R(t)] = B_{p_y,q_y}[F_T(t)] \quad (25)$$

$$H_1: B_{p_x,q_x}[F_R(t)] \neq B_{p_y,q_y}[F_T(t)]$$

where $F_R(t)$ and $F_T(t)$ are the distribution functions for the observed values.

Construction of the pseudo-likelihood ratio test

In order to test the hypothesis at (26) we propose a pseudo-likelihood ratio test using the following steps:

1. Estimate the distribution function under H_0 for the combined values for R_i, $i = 1, 2, \cdots, n_X$ and T_j, $j = 1, 2, \cdots, n_Y$ as

$$\widehat{F}_{\text{pooled}}(t) = \frac{\sum_{i=1}^{n} I_{(t \le z_i)}}{n+1}, \quad (26)$$

where the $n \times 1$ vector $\mathbf{z} = (\mathbf{r}|\mathbf{t})$ is the concatenation of the two sets of observed value vectors from group 1 and group 2, respectively. This format is similar to using the pooled ranks in the Wilcoxon rank-sum test, where the pooled ranks are given as $(n+1)\widehat{F}_{\text{pooled}}(t)$.

2. Denote the observed values of $\widehat{F}_{\text{pooled}}(t)$ as a function of $\mathbf{z}$ corresponding to group 1 as $\widehat{F}^1_{\text{pooled}}(t)$ and group 2 as $\widehat{F}^2_{\text{pooled}}(t)$, i.e. the rankits broken out by group.

3. Denote the pseudo-likelihood under H_0 as

$$L_0 = \prod_{i=1}^{n_x} b_{p_z,q_z}\left[\widehat{F}^1_{\text{pooled}}(z_i)\right]^{\delta_i^x}\left(1 - B_{p_z,q_z}\left[\widehat{F}^1_{\text{pooled}}(z_i)\right]\right)^{1-\delta_i^x} \times \prod_{i=n_x+1}^{n} b_{p_z,q_z}\left[\widehat{F}^2_{\text{pooled}}(z_i)\right]^{\delta_j^y}\left(1 - B_{p_z,q_z}\left[\widehat{F}^2_{\text{pooled}}(z_i)\right]\right)^{1-\delta_j^y} \quad (27)$$

where $p_z = p_x = p_y$ and $q_z = q_x = q_y$ under H_0 at (26).

4. Denote the pseudo-likelihood under H_1 as

$$L_1 = \prod_{i=1}^{n_x} b_{p_x,q_x}\left[\widehat{F}^1_{\text{pooled}}(z_i)\right]^{\delta_i^x}\left(1 - B_{p_x,q_x}\left[\widehat{F}^1_{\text{pooled}}(z_i)\right]\right)^{1-\delta_i^x} \times \prod_{i=n_x+1}^{n} b_{p_y,q_y}\left[\widehat{F}^2_{\text{pooled}}(z_i)\right]^{\delta_j^y}\left(1 - B_{p_y,q_y}\left[\widehat{F}^2_{\text{pooled}}(z_i)\right]\right)^{1-\delta_j^y} \quad (28)$$

5. Denote the observed value of the pseudo-likelihood ratio statistic as $\Delta = -2(\log(L_0) - \log(L_1))$.

6. Permute the groups B times and for each permutation calculate B permuted pseudo-likelihood ratio statistics denoted as $\Delta_i^* = -2(log(L_0) - log(L_1))$, $i = 1,2,\ldots,B$.

7. The Monte Carlo approximate permutation p-value is given as $\sum_{i=1}^{B} I_{(\Delta_i^* > \Delta)}/B$. The exact p-value can be calculated across all permutations as necessary. However, for large B the approximation will be sufficiently accurate.

Comment 1—It is critical to note the key point of this work in that our test is exact in terms of controlling the type I error given the permutation framework even if the underlying model does not perfectly fit the data. Hence, this test can be considered as a general semi-parametric test for comparing survival curves.

Comment 2—The approach above is readily extended to the k group setting by simply adding additional terms to L_0 and L_1 at (27) and (28), respectively, corresponding to group 3 and upwards.

In general, the large sample distributions for pseudo-likelihood ratio tests are complex[11] and are only distributed asymptotically chi-squared under certain conditions. Generally, the distribution consists of a weighted average of chi-square distributions. However, heuristically we have discovered that $\Delta = -2(\log(L_0) - \log(L_1))$ follows well a chi-squared distribution with 2 degrees of freedom similar to standard large sample theory and provides almost identical p-values to that of the permutation test. An extensive simulation study comparing the new test (and its chi-square approximation) versus the classic log-rank test and Wilcoxon rank-sum test follows our illustrative example in the next section.

6 Two-group comparison

In this section, we provide a simple two-group survival curve comparison using the surgically place catheter data (group 1) from section 4 with percutaneous placed catheter data[10] (group 2) given as:

Infection times: 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 2.5, 2.5, 3.5, 6.5, 15.5

Censored observations: 0.5, 0.5, 0.5, 0.5, 0.5,0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 1.5, 1.5, 2.5, 2.5, 2.5, 2.5, 3.5, 3.5, 3.5, 3.5, 4.5, 4.5, 4.5, 5.5, 5.5, 5.5, 5.5, 5.5, 6.5, 7.5, 7.5, 7.5, 8.5, 8.5, 8.5, 9.5, 9.5, 10.5, 10.5, 10.5, 11.5, 11.5, 12.5, 12.5, 12.5, 12.5, 14.5, 14.5, 16.5, 16.5, 18.5, 19.5, 19.5, 19.5, 20.5, 22.5, 24.5, 25.5, 26.5, 26.5, 28.5

The product-limit estimates for the survival curves are given by Figure 4, where the percutaneous group is shifted to the left. The emphasis going forward is that our new test will provide an alternative and oftentimes much more powerful test than the standard log-rank or Wilcoxon tests found in most statistical software packages. For the permutation two-group warped survival test we use $B = 1000$ permutations. The exact permutation p-value was 0.005. The approximate pseudo-likelihood based p-value based on a chi-square distribution with 2 degrees of freedom was 0.0004. The classic log-rank and Wilcoxon p-values were $p = 0.4013$ and $p = 0.0097$, respectively. In our simulation study, in the next section we will show that the new warped survival test oftentimes is much more powerful than both the log-rank and Wilcoxon tests and that the chi-square approximation is virtually identical to the permutation test. The permutation p-value is exact under the exchangeability

assumption even if the warped survival model only approximates the true underlying population structure.

The example SAS code corresponding to this section may be found at https://www.researchgate.net/publication/289540278_SAS_CODE_FOR_A_Robust_Semi-Parametric_Warping_Estimator_of_the_Survivor_Function_With_an_Application_to_Two_Group_Comparisons

7 Simulation comparison of the new test with the log-rank and Wilcoxon tests

We compared the type I error and power between the exact permutation (EP) Method, approximate likelihood ratio (ALR), the classic log-rank (LR) test, and the classic Wilcoxon (WN) test for testing $H_0 : S_X(t) = S_Y(t)$ versus $H_1 : S_X(t) = S_Y(t)$ for Weibull (W), Log-Normal (LN), and Log-Logistic (LL) distributions. The results are found in Tables 3 to 5. For the simulations we used 500 Monte Carlo resamples, $n_x = n_y = 35$ and a non-informative censoring distribution $C \sim \exp(1)$ for both groups. For the Monte Carlo approximation to the permutation test we used 100 Monte Carlo permutations. The upper left column for each simulation table presents the type I error control for each method. As expected all methods control the type I error appropriately. In terms of statistical power the new permutation method provides the best global power across methods. There are a few instances where the log-rank test or the Wilcoxon test is superior to each other; however, in each instance the new permutation test is comparable and generally provides the best global coverage given an unknown underlying distribution. For example in Table 3 when group 2 is W(3,1), the power for the permutation test is 0.924 as compared to 0.118 for the log-rank test and 0.650 for the Wilcoxon test. However, when the group 2 distribution is W(2,1/2) the power for the permutation test is 0.856 as compared to 0.636 for the log-rank test and 0.178 for the Wilcoxon test. In this instance, in these examples the power magnitude for the log-rank test and the Wilcoxon test is reversed, yet the permutation test maintains better power for each alternative. In some instances, the new test is dramatically more sensitive than the log-rank and/or the Wilcoxon test towards testing the hypothesis of interest, e.g. in Table 4 with the LN(0,1/2) alternative for group 2 the power for the permutation test is 0.504 as compared to 0.064 for the log-rank test and 0.234 for the Wilcoxon test. These results consistently hold across a variety of distributions and sample sizes and are not presented. Finally, it should be noted that the ALR test approximates well the permutation test and could be used given large sample sizes when the permutation test may be computationally infeasible.

8 Summary remarks

In this note, we consider the classic mathematical statistics relationship $S_T = S_X(t)S_C(t)$, where $T = \min(X, C)$ as described in the introduction. The key idea is that $F_X(t) \approx g(F_T(t))$ where g is what we termed a *warping function*. The choice of g is arbitrary as long as $g(F_T(t))$ is a proper distribution function. Since for a given sample size n we can arrive at a proper estimator for $F_T(t)$ given as $\widehat{F}_T(t) = \sum_{i=1}^{n} I_{(t \le t_i)}/(n+1)$ this implies that for the appropriate g, $g(\widehat{F}_T(t))$ is also a proper distribution function estimator. For our purposes we

defined g to be a beta distribution function, which in turn allows for $g(\widehat{F}_T(t))$ to take on several shapes similar to what has been evolving in the parametric literature using beta transformation families.[12] The method is quite robust in that in the most basic sense we are working with a transformation of rankits and not the raw data values. The theoretical framework for this estimator follows nicely from previous work about pseudo-likelihood methods.

The utility of this semi-parametric approach is not necessarily found in the estimation of S_X but is more directed towards the two and k group comparisons of survival curves. We provide an exact α level test, that as might be expected has substantial power gains over the more traditional tests when the proportional hazards assumptions do not hold relative to the log-rank test and the weighted proportional hazards assumptions do not hold relative to the Wilcoxon test. For large samples we provide an asymptotic chi-square approximation, which also appears to provide very accurate type I error control in the medium sample size setting. In general, the new exact permutation test is relatively efficient when the various proportional hazards assumptions do hold. It was a rare event in our simulation study where the log-rank or Wilcoxon test was more efficient than the new semi-parametric approach. Hence, in terms of a global test we believe this new test has a great deal of utility, particularly when the proportional hazards assumption may be assumed to be unreasonable, e.g. a delayed treatment effect of one group compared to another, which may lead to a change-point type relationship relative to the respective hazard function ratio. Similarly, under classic permutation test exchangeability assumptions the test is always exact in terms of type I error control.

Acknowledgments

We wish to acknowledge the referees. Their comments greatly enhanced the presentation of our proposed approach.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001412.

References

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958; 52:457–481.
2. Ryan TP, Woodall WH. The most-cited statistical papers. J Appl Stat. 2005; 32:461–474.
3. Miller RG Jr. What price Kaplan-Meier. Biometrics. 1983; 39:1077–1081. [PubMed: 6671119]
4. Oakes D. A note on the Kaplan-Meier estimator. Am Statist. 1983; 47:39–40.
5. Chen K, Lo S-H. On the rate of uniform convergence of the product-limit estimator: Strong and weak laws. Ann Stat. 1997; 25:1050–1087.
6. Cox, DR., Oakes, D. Analysis of survival data. New York: Chapman & Hall/CRC; 1984.
7. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natnl Cancer Inst. 1959; 22:719–448.
8. Prentice RL. Linear rank tests with right censored data. Biometrika. 1978; 34:167–179.
9. Genest C, Ghoudi K, Rivest LP. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika. 1995; 82:543–552.

10. Klein, JP., Moeschberger, ML. Survival analysis: Techniques for censored and truncated data. New York: Springer; 2003.
11. Liang K-Y, Self SG. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. J Roy Stat Soc Ser B. 1996; 58:785–796.
12. Jones MC. Families of distributions arising from distributions of order statistics. Test. 2004; 13:1–43.

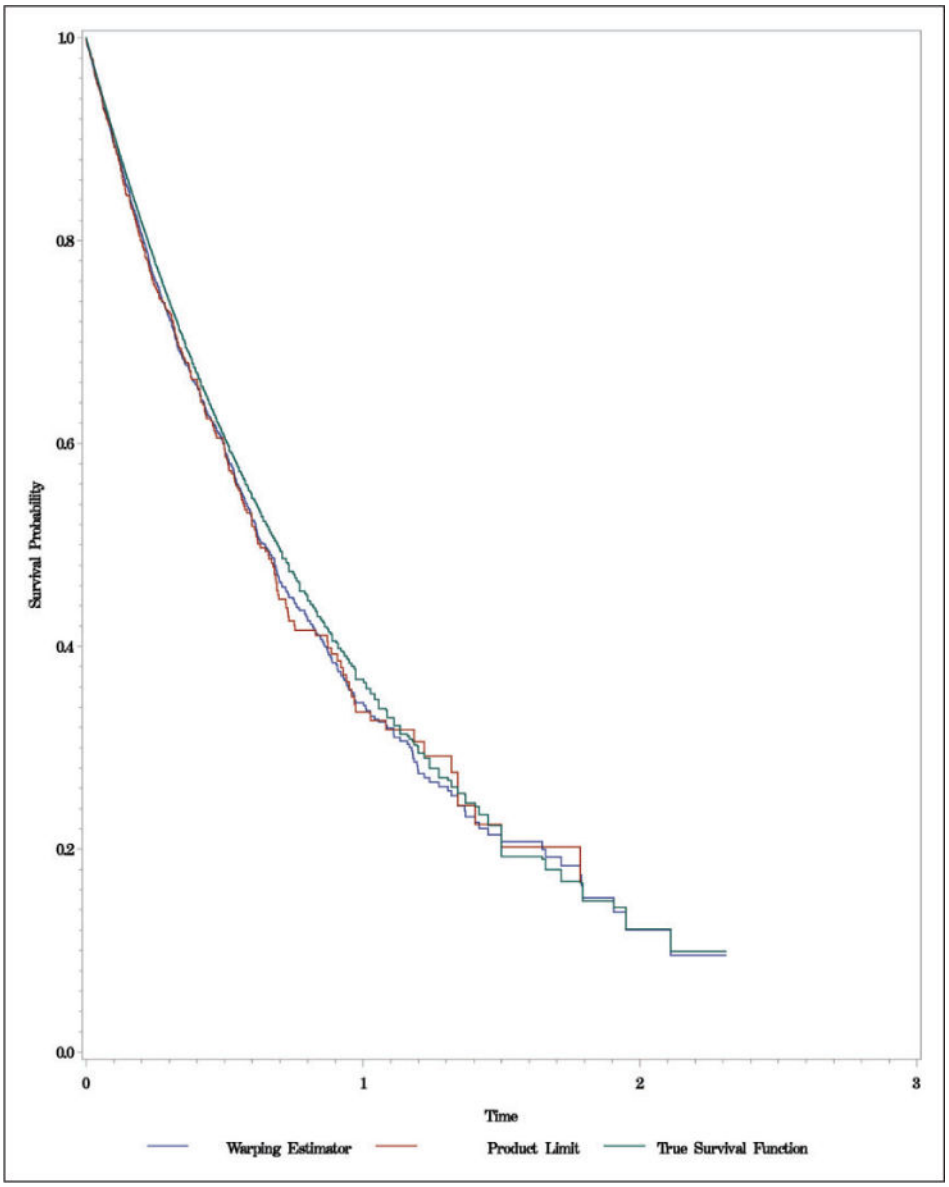

Figure 1.

Estimate of $\widehat{S}(x)$ from simulated data comparing the semi-parametric warped model versus the product-limit estimator versus true underlying survival distribution from a sample size of $n = 1000$ with $X \sim \exp(1)$ and $C \sim \exp(1/2)$.

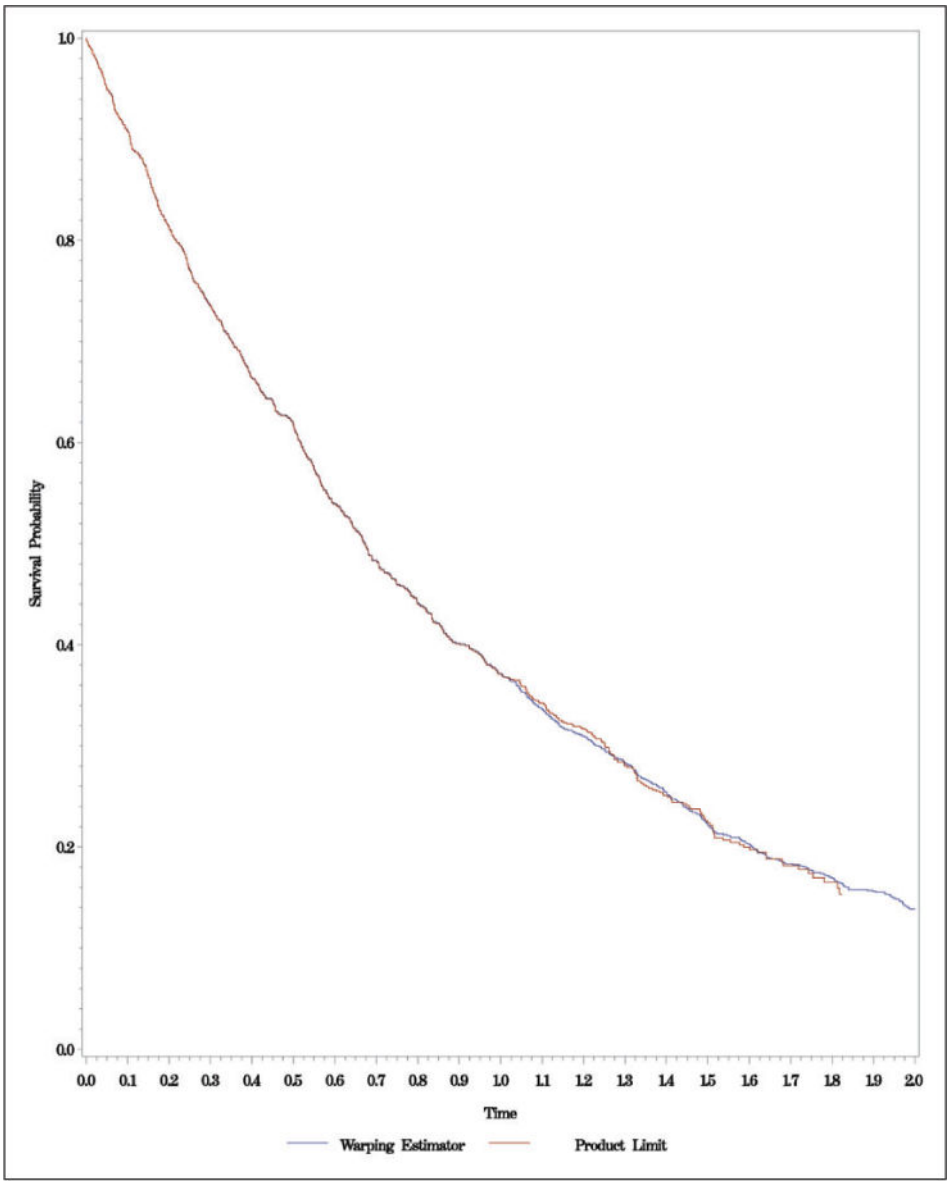

Figure 2.

Estimate of $\widehat{S}(x)$ from simulated data comparing the semi-parametric warped model versus the product-limit estimator from a sample size of $n = 1000$ with $X \sim \exp(1)$ and $C \sim U(1,2)$.

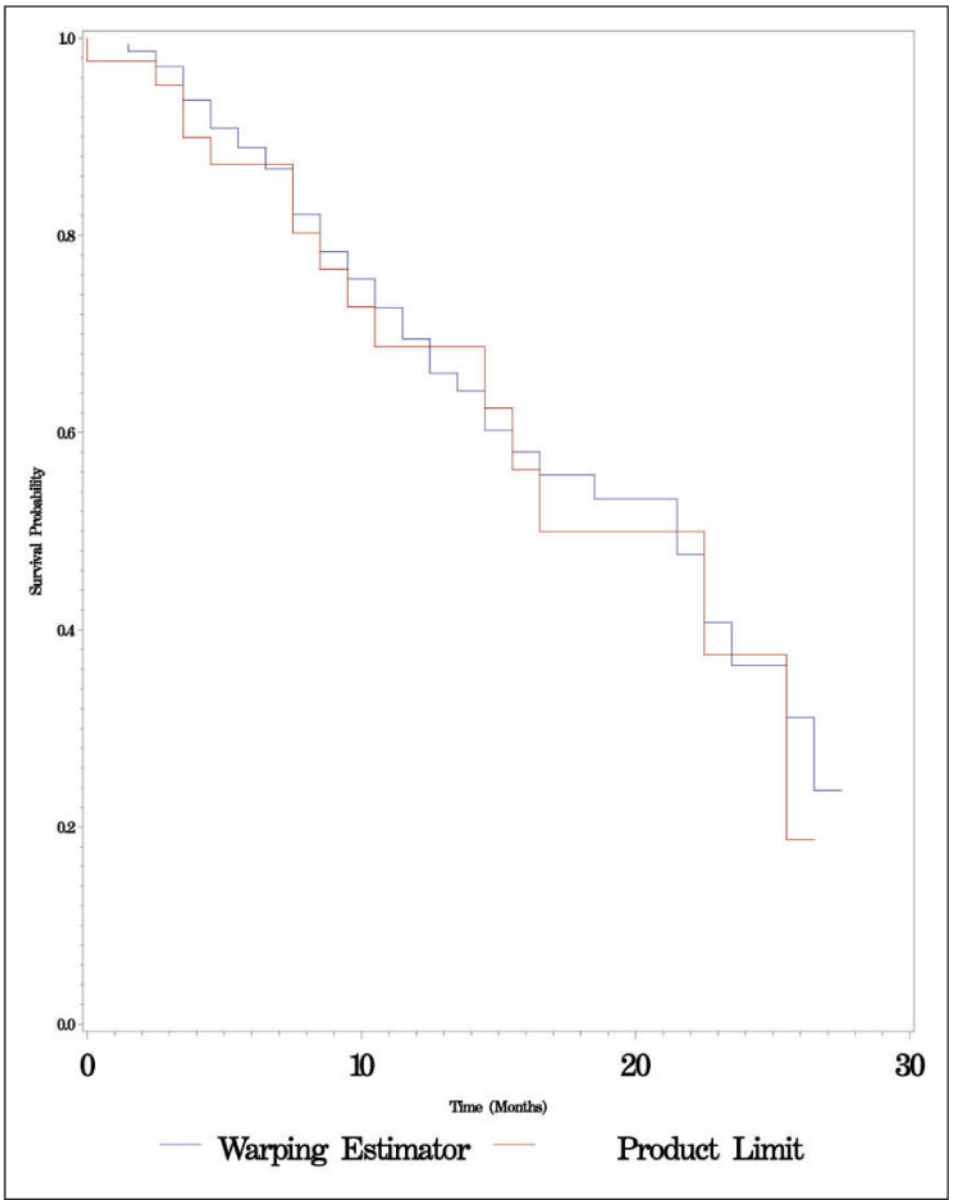

Figure 3.
Time to infection estimates for surgically placed catheter group based on the semi-parametric warped model versus the product-limit estimator.

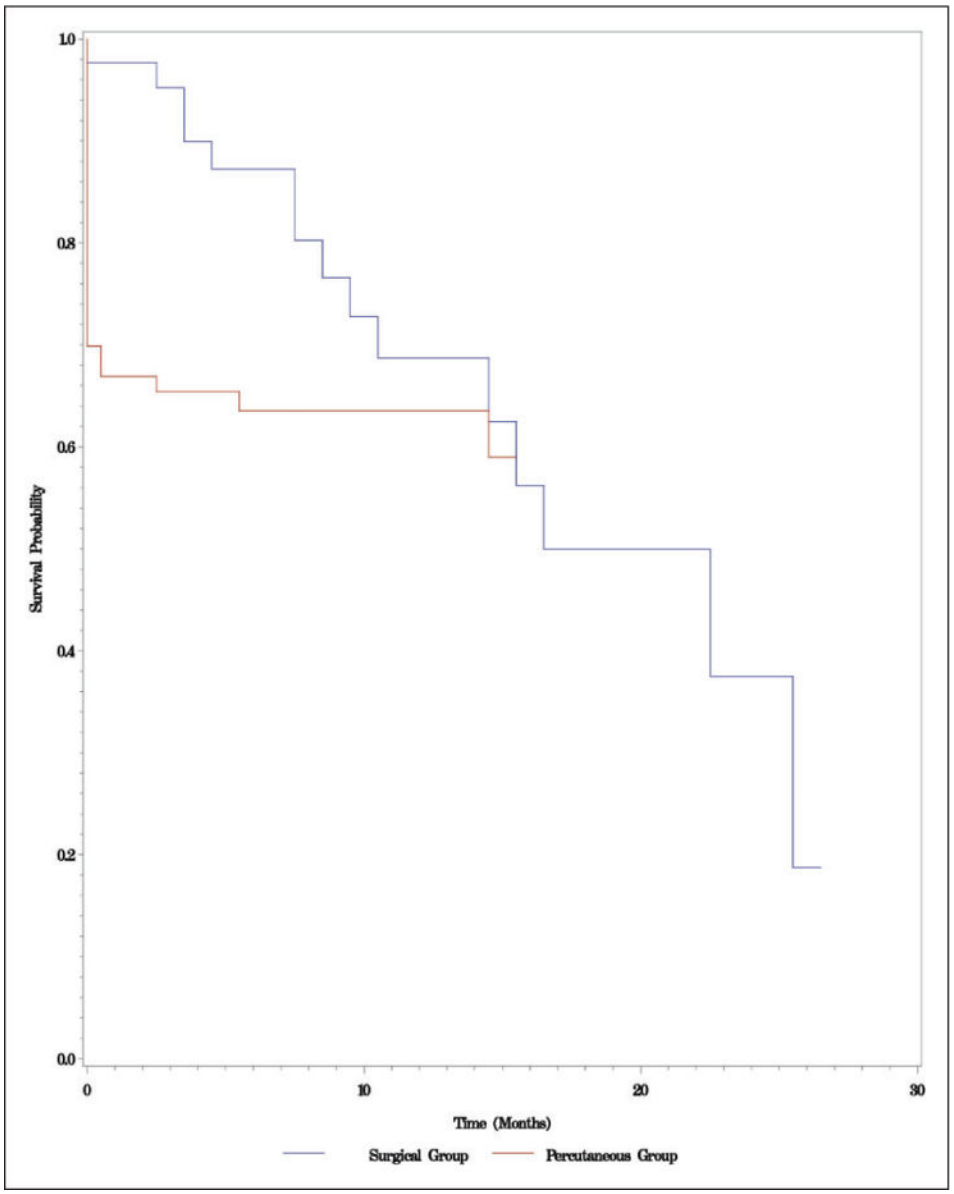

Figure 4.
Product-limit estimator for the time to infection estimates for surgically placed catheter group versus the percutaneous catheter group.

Table 1

Simulation results: Average absolute maximum difference between the two parametric survival curve estimators for $t = 0$ to 2 by 0.01.

n	Mean ± Std Error
10	0.086 ± 0.079
100	0.020 ± 0.016
1000	0.006 ± 0.004
10,000	0.001 ± 0.001

Table 2

MSE for Weibull quartile estimation with $C \sim \exp(1)$ with percent of observed quartiles. A comparison of the product-limit (PL) estimator to the warped survival (WS) estimator.

Estimator	*n*	*Q*(1/4)	%	*Q*(1/2)	%	*Q*(3/4)	%
Weibull(1,1)							
WS	50	0.006	100	0.023	100	0.118	100
PL	50	0.008	100	0.041	100	0.226	93.3
WS	200	0.002	100	0.006	100	0.031	100
PL	200	0.002	100	0.009	100	0.047	99.9
WS	500	0.001	100	0.002	100	0.012	100
PL	500	0.001	100	0.003	100	0.016	100
Weibull(3,3)							
WS	50	0.130	99.0	0.264	99.4	0.724	99.0
PL	50	0.222	95.2	0.253	80.9	0.336	61.2
WS	200	0.037	100	0.067	100	0.133	100
PL	200	0.055	100	0.106	99.3	0.173	89.1
WS	500	0.016	99.9	0.026	100	0.062	100
PL	500	0.018	99.9	0.033	100	0.095	98.6
Weibull(3/4,3/4)							
WS	50	0.003	100	0.016	100	0.107	100
PL	50	0.004	100	0.025	100	0.202	94.9
WS	200	0.001	100	0.004	100	0.031	100
PL	200	0.001	100	0.005	100	0.050	99.9
WS	500	¡0.001	100	0.002	100	0.012	100
PL	500	¡0.001	100	0.002	100	0.015	100
Weibull(3/4,3)							
WS	50	0.059	100	0.102	100	4.933	100
PL	50	0.118	99.8	0.292	81.0	3.879	31.5
WS	200	0.012	100	0.043	100	1.924	100
PL	200	0.017	100	0.087	98.0	1.629	43.7
WS	500	0.005	100	0.043	100	1.094	100

Estimator	*n*	*Q*(1/4)	%	*Q*(1/2)	%	*Q*(3/4)	%
PL	500	0.006	100	0.087	100	1.14	58.1

Table 3

Comparison of type I error and power between the exact permutation (EP) Method, approximate likelihood ratio (ALR), the classic log-rank (LR) test, and the classic Wilcoxon (WN) test for testing $H_0 : S_X(t) = S_Y(t)$ versus $H_1 : S_X(t) \neq S_Y(t)$ with group 1 distributed as W(1,1), $n_x = n_y = 35$ and censoring distribution $C \sim \exp(1)$ for both groups.

	Group 2				
Estimator	**W(1,1)**	**W(1/2,1)**	**W(1/2,2)**	**W(1,1/2)**	**W(1,3)**
EP	0.046	0.522	0.460	0.452	0.688
ALR	0.044	0.484	0.430	0.426	0.684
LR	0.036	0.126	0.070	0.556	0.768
WN	0.048	0.394	0.098	0.432	0.648

Estimator	**W(3,1)**	**W(2,1/2)**	**W(3,3)**	**W(1/3,1/3)**	**W(2,2)**
EP	0.924	0.856	0.999	0.980	0.918
ALR	0.944	0.874	0.999	0.978	0.920
LR	0.118	0.636	0.999	0.750	0.876
WN	0.650	0.178	0.999	0.976	0.944

Table 4

Comparison of type I error and power between the exact permutation (EP) Method, approximate likelihood ratio (ALR), the classic log-rank (LR) test, and the classic Wilcoxon (WN) test for testing $H_0 : S_X(t) = S_Y(t)$ versus $H_1 : S_X(t) \neq S_Y(t)$ with group 1 distributed as LN(0,1), $n_x = n_y = 35$ and censoring distribution $C \sim \exp(1)$ for both groups.

	Group 2				
Estimator	**LN(0,1)**	**LN(−1/2,1/2)**	**LN(−1/2,1)**	**LN(−1/2,2)**	**LN(0,1/2)**
EP	0.048	0.718	0.224	0.758	0.504
ALR	0.040	0.716	0.214	0.756	0.520
LR	0.056	0.548	0.298	0.310	0.064
WN	0.042	0.186	0.302	0.688	0.234

Estimator	**LN(0,2)**	**LN(1/2,1/2)**	**LN(1/2,1)**	**LN(1/2,2)**	**LN(-3,3)**
EP	0.494	0.862	0.294	0.390	0.999
ALR	0.496	0.868	0.280	0.392	0.999
LR	0.090	0.574	0.334	0.074	0.998
WN	0.282	0.886	0.348	0.066	0.998

Stat Methods Med Res. Author manuscript; available in PMC 2018 January 27.

Table 5

Comparison of type I error and power between the exact permutation (EP) Method, approximate likelihood ratio (ALR), the classic log-rank (LR) test, and the classic Wilcoxon (WN) test for testing $H_0 : S_X(t) = S_Y(t)$ versus $H_1 : S_X(t) \neq S_Y(t)$ with group 1 distributed as LL(0,1), $n_x = n_y = 35$ and censoring distribution $C \sim \exp(1)$ for both groups.

	Group 2				
Estimator	**LL(0,1)**	**LL(−1/2,1/2)**	**LL(−1/2,1)**	**LL(−1/2,2)**	**LL(0,1/2)**
EP	0.044	0.520	0.224	0.500	0.438
ALR	0.036	0.528	0.214	0.438	0.452
LR	0.050	0.186	0.298	0.180	0.084
WN	0.030	0.080	0.302	0.440	0.252

Estimator	**LL(0,2)**	**LL(1/2,1/2)**	**LL(1/2,1)**	**LL(1/2,2)**	**LL(−3,3)**
EP	0.328	0.652	0.116	0.262	0.998
ALR	0.282	0.654	0.112	0.232	0.996
LR	0.092	0.428	0.152	0.044	0.904
WN	0.204	0.740	0.158	0.090	0.992