Biometrika (2019), **106**, 3, *pp.* 501–518
Printed in Great Britain
doi: 10.1093/biomet/asz016
Advance Access publication 9 June 2019

Nonparametric generalized fiducial inference for survival functions under censoring

By Y. CUI

Department of Statistics, The Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, U.S.A.

cuiy@wharton.upenn.edu

and J. HANNIG

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, 318 Hanes Hall, Chapel Hill, North Carolina 27599, U.S.A.

jan.hannig@unc.edu

Summary

Since the introduction of fiducial inference by Fisher in the 1930s, its application has been largely confined to relatively simple, parametric problems. In this paper, we present what might be the first time fiducial inference is systematically applied to estimation of a nonparametric survival function under right censoring. We find that the resulting fiducial distribution gives rise to surprisingly good statistical procedures applicable to both one-sample and two-sample problems. In particular, we use the fiducial distribution of a survival function to construct pointwise and curvewise confidence intervals for the survival function, and propose tests based on the curvewise confidence interval. We establish a functional Bernstein–von Mises theorem, and perform thorough simulation studies in scenarios with different levels of censoring. The proposed fiducial-based confidence intervals maintain coverage in situations where asymptotic methods often have substantial coverage problems. Furthermore, the average length of the proposed confidence intervals is often shorter than the length of confidence intervals for competing methods that maintain coverage. Finally, the proposed fiducial test is more powerful than various types of log-rank tests and sup log-rank tests in some scenarios. We illustrate the proposed fiducial test by comparing chemotherapy against chemotherapy combined with radiotherapy, using data from the treatment of locally unresectable gastric cancer.

Some key words: Coverage; Generalized fiducial inference; Nonparametric model; Right-censored data; Testing.

1. Introduction

Fiducial inference can be traced back to a series of articles by (R. A. Fisher, 1925, 1930, 1933, 1935), who introduced the concept as a potential replacement for the Bayesian posterior distribution. A systematic development of the idea has been hampered by ambiguity, as Brillinger (1962) describes: 'The reason for this lack of agreement and the resulting controversy is possibly due to the fact that the fiducial method has been put forward as a general logical principle, but yet has been illustrated mainly by means of particular examples rather than broad requirements.' Indeed, we contend that until recently fiducial inference has been applied to a relatively small class of parametric problems only.

Since the mid 2000s, there has been a renewed interest in modifications of fiducial inference. Hannig (2009, 2013) brings forward a mathematical definition of what he calls the generalized fiducial distribution. Having a formal definition allowed fiducial inference to be applied to a wide variety of statistical settings (Wang & Iyer, 2005, 2006a,b; Hannig et al., 2007; Hannig & Lee, 2009; Cisewski & Hannig, 2012; Wandler & Hannig, 2012; Wang et al., 2012; Lai et al., 2015; Liu & Hannig, 2017; Hannig et al., 2018).

Other related approaches include the Dempster–Shafer theory (Dempster, 2008; Edlefsen et al., 2009), inferential models (Martin & Liu, 2015), and confidence distributions (Xie & Singh, 2013; Schweder & Hjort, 2016; Hjort & Schweder, 2018). Objective Bayesian inference, which aims at finding non-subjective model-based priors, can also be seen as addressing the same basic question. Examples of recent breakthroughs related to reference prior and model selection are Bayarri et al. (2012) and Berger et al. (2009, 2012); see the review article Hannig et al. (2016) for more references.

In this paper, we apply the fiducial approach in the context of survival analysis. To our knowledge, this is the first time fiducial inference has been systematically applied to an infinite-dimensional statistical problem. However, for the use of confidence distributions to address some basic nonparametric problems see Chapter 11 of Schweder & Hjort (2016). In this article, we propose a computationally efficient algorithm to sample from the generalized fiducial distribution, and use the samples from the generalized fiducial distribution to construct statistical procedures. The median of the generalized fiducial distribution could be considered a substitute for the Kaplan–Meier estimator (Kaplan & Meier, 1958), which is a classical estimator in survival analysis. Appropriate quantiles of the generalized fiducial distribution evaluated at a given time provide pointwise confidence intervals for the survival function. Similarly, the confidence intervals for quantiles of survival functions can be obtained by inverting the generalized fiducial distribution.

The proposed pointwise confidence intervals maintain coverage in situations where classical confidence intervals often have coverage problems (Fay et al., 2013). Fay et al. (2013) and Fay & Brittain (2016) construct solutions to avoid these coverage problems. The conservative version of the proposed pointwise fiducial confidence interval is equivalent to the beta product confidence procedure of Fay et al. (2013). The other fiducial confidence interval proposed in this paper is based on log-linear interpolation and has the shortest length among all existing methods which maintain coverage. We also construct curvewise confidence intervals for survival functions. Based on the curvewise confidence intervals, we propose a two-sample test for testing whether two survival functions are equal.

We establish an asymptotic theory which verifies the frequentist validity of the proposed fiducial approach. In particular, we prove a functional Bernstein–von Mises theorem for the generalized fiducial distribution in Skorokhod's $D[0, t]$ space. Because randomness in generalized fiducial distributions comes from two distinct sources, the proof of this result is different from the usual proof of asymptotic normality for the Kaplan–Meier estimator. As a consequence of the functional Bernstein–von Mises theorem, the proposed pointwise and curvewise confidence intervals provide asymptotically correct coverage, and the proposed survival function estimator is asymptotically equivalent to the Kaplan–Meier estimator.

The proposed fiducial approach provides competitive, and in some cases superior, performance to many methods in the literature and also appears to be an alternative to the log-rank test and sup log-rank test, which are valid for goodness-of-fit testing within the class of stochastically ordered alternatives as shown in the 1980 PhD thesis by R. D. Gill, Mathematical Centre, Amsterdam.

2. Methodology

2.1. *Fiducial approach explained*

We explain the definition of a generalized fiducial distribution and start by expressing the relationship between the data Z and the parameter θ using

$$Z = G(W, \theta), \tag{1}$$

where $G(\cdot, \cdot)$ is a deterministic function termed the data-generating function and W is a collection of random variables whose joint distribution is independent of θ and completely known.

In the case of no censoring, $Z = Y$ and $W = U$, where $Y = (Y_1, \ldots, Y_n)$ are observed data and $U = (U_1, \ldots, U_n)$ are independent and identically distributed $\mathrm{Un}(0, 1)$. The inverse cumulative distribution function method for generating random variables provides a common data-generating equation for a nonparametric independent and identically distributed model:

$$Y_i = G(U_i, F) = F^{-1}(U_i) \quad (i = 1, \ldots, n), \tag{2}$$

where $F^{-1}(u) = \inf\{y \in \mathbb{R} : F(y) \geqslant u\}$ is the usual inverse of the distribution function $F(y)$ (Casella & Berger, 2002). The distribution function F itself is the parameter θ in this infinite-dimensional model. The actual observed data are generated using the true distribution function F_0.

Roughly speaking, a generalized fiducial distribution is obtained by inverting a well-chosen data-generating equation, and Hannig et al. (2016) propose a very general definition of a generalized fiducial distribution. However, in order to simplify the presentation, we will use an earlier, less general version found in Hannig (2009). The two definitions are equivalent for the models considered here.

We start by denoting the inverse image of the data-generating equation (1) by

$$Q(y, u) = \{\theta \ : \ y = G(u, \theta)\}.$$

For the special case (2) the inverse image is

$$Q(y, u) = \bigcap_{i=1}^{n} \{F : F(y_i) \geqslant u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0\}. \tag{3}$$

If y is the set of observed data and u_0 the value of the random vector U that was used to generate it, then we are guaranteed that the true parameter value θ_0 belongs to $Q(y, u_0)$. However, we only know a distribution of U and not the actual value u_0. Notice that $y = G(u_0, \theta_0)$ and therefore only values of u for which $Q(y, u) \neq \emptyset$ should be considered. Let U^* be another random variable independent of and having the same distribution as U. Since the conditional distribution of $U^* \mid \{Q(y, U^*) \neq \emptyset\}$ can be viewed as summarizing our knowledge about u_0, the conditional distribution of

$$Q(y, U^*) \mid \{Q(y, U^*) \neq \emptyset\} \tag{4}$$

can be viewed as summarizing our knowledge about θ_0.

The set $Q(y, u)$ can contain more than one element. We deal with this by selecting a representative from the closure of $Q(y, u)$. The distribution of a representative selected from (4) is a generalized fiducial distribution. Based on the theoretical results presented, the nonuniqueness

caused by this somewhat arbitrary choice disappears asymptotically. A possible conservative alternative to selecting a single representative from $Q(y,u)$ would use the theory of belief functions (Dempster, 2008; Shafer, 1976).

To describe the generalized fiducial distribution in the particular case of (2) we define, for all $s \geqslant 0$, $F^{\mathrm{L}}_{(y,u)}(s) = \inf\{F(s) : F \in Q(y,u)\}$ and $F^{\mathrm{U}}_{(y,u)}(s) = \sup\{F(s) : F \in Q(y,u)\}$. The closure of the inverse image (3) is the set of all distribution functions F that stay between $F^{\mathrm{L}}_{(y,u)}$ and $F^{\mathrm{U}}_{(y,u)}$. Also notice that $Q(y,u)$ is not empty if and only if the order of u matches the order of y componentwise, with the understanding that in the case of ties in y, the u_i corresponding to the ties could be in any order.

By exchangeability, the conditional distribution $U^* \mid \{Q(y,U^*) \neq \emptyset\}$ is the same as the distribution of $U^*_{[y]}$, where $U^*_{[y]}$ is independent and identically distributed Un(0,1), reordered to match the order of y componentwise. Next define random distribution functions by inserting the random vector $U^*_{[y]}$ for u: $F^{\mathrm{L}}(s) = F^{\mathrm{L}}_{(y,U^*_{[y]})}(s)$ and $F^{\mathrm{U}}(s) = F^{\mathrm{U}}_{(y,U^*_{[y]})}(s)$. For simplicity of notation we omit the subindex $(y, U^*_{[y]})$. The random distribution functions F^{L} and F^{U} provide stochastic lower and upper bounds on the generalized fiducial distribution.

We consider two main options in using the generalized fiducial distribution for inference. The first option is to construct conservative confidence sets. For example, when designing pointwise confidence intervals for the survival function at time s, we use quantiles of the random survival functions $S^{\mathrm{L}}(s) = 1 - F^{\mathrm{U}}(s)$ for lower bounds and quantiles of $S^{U}(s) = 1 - F^{\mathrm{L}}(s)$ for upper bounds. We will call $S^{\mathrm{L}}(s)$ and $S^{U}(s)$ the lower and upper fiducial survival functions, respectively.

The second option is to select a suitable representative of $Q(y, U^*_{[y]})$. When there are no ties present in the data, we propose to fit a continuous distribution function by using linear interpolation for the survival function on the log scale, i.e., the distribution function $F^{\mathrm{I}}_{(y,u)}(s) = 1 - \exp\{L(s)\}$, where $L(s)$ is the linear interpolation between $(0,0), (y_{(1)}, \log u_{(1)}), \ldots, (y_{(n)}, \log u_{(n)})$, and on the interval $(y_{(n)}, \infty)$ we extrapolate by extending the line between $(y_{(n-1)}, \log u_{(n-1)})$ and $(y_{(n)}, \log u_{(n)})$. We will call this the log-linear interpolation and call the corresponding random survival function $S^{I}(s) = 1 - F^{\mathrm{I}}_{(y,U^*_{[y]})}(s)$ the log-linear interpolation fiducial survival function.

In the rest of this paper we will denote Monte Carlo realizations of the lower, the upper, and the log-linear interpolation fiducial survival functions by S^L_i, S^U_i, and S^I_i $(i = 1, \ldots, m)$, respectively.

To demonstrate the fiducial distribution of this section, we draw 300 observations from Wei(20, 10). Based on these data, we plot a sample of the fiducial survival functions $S^I_i (i = 1, \ldots, 1000)$ and the empirical survival function in Fig. 1(a).

2.2. *Fiducial approach in the survival setting*

We now derive the generalized fiducial distribution using a particular choice of data-generating equation generalizing (2) in a natural way for right-censored data. We treat the situation where failure and censoring times are independent and unknown.

Let the failure times X_i $(i = 1, \ldots, n)$ follow the true distribution function F_0 and let the censoring times C_i $(i = 1, \ldots, n)$ have the distribution function R_0. We observe partially censored data $\{y_i, \delta_i\}$ $(i = 1, \ldots n)$, where $y_i = x_i \wedge c_i$ is the minimum of x_i and c_i, and $\delta_i = I\{x_i \leqslant c_i\}$ denotes the censoring indicator.

We consider the following data-generating equation:

$$Y_i = F^{-1}(U_i) \wedge R^{-1}(V_i), \quad \Delta_i = I\{F^{-1}(U_i) \leqslant R^{-1}(V_i)\} \quad (i = 1, \ldots n), \tag{5}$$

where U_i and V_i are independent and identically distributed Un$(0,1)$ and the actual observed data were generated using $F = F_0$ and $R = R_0$. Notice that Z in (1) is (Y, Δ) in (5) and that W in (1) is (U, V) in (5).

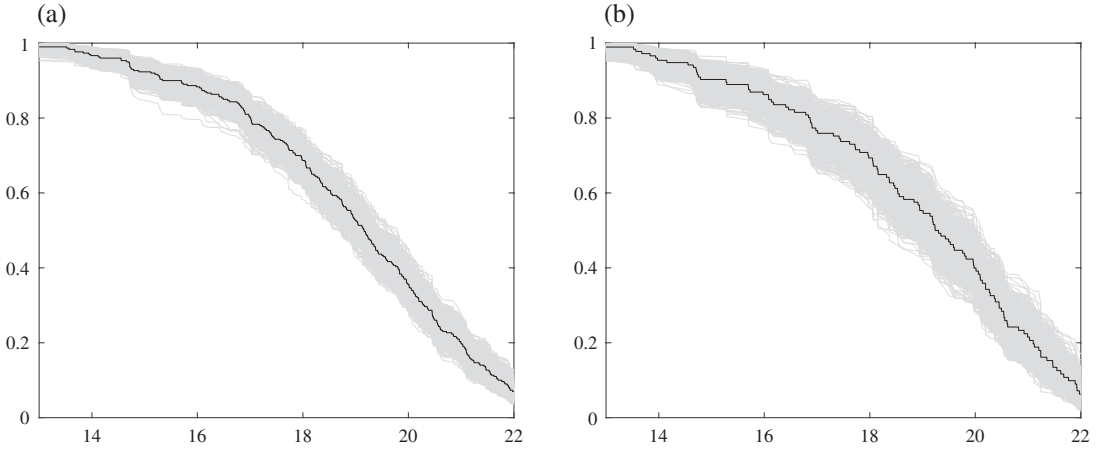

Fig. 1. A plot of Monte Carlo realizations S_i^I $(i = 1, \ldots, 1000)$ sampled from the generalized fiducial distribution based on either a sample of 300 uncensored Wei(20, 10) observations, shown as grey curves in (a), or the same 300 Wei(20, 10) observations censored by Ex(20), shown as grey curves in (b). The black curve in (a) is the empirical survival function and in (b) it is the Kaplan–Meier estimator. As expected, we observe higher uncertainty in the fiducial sample under censoring.

For a failure event $\delta_i = 1$, we have full information about the failure time x_i, i.e., $x_i = y_i$, and partial information about the censoring time c_i, i.e., $c_i \geqslant y_i$. In this case, just as in the previous section, $F^{-1}(u_i) = y_i$ if and only if $F(y_i) \geqslant u_i$ and $F(y_i - \epsilon) < u_i$ for any $\epsilon > 0$.

For a censored event $\delta_i = 0$, we know only partial information about x_i, i.e., $x_i > y_i$, and full information on c_i, i.e., $c_i = y_i$. Similarly, $F^{-1}(u_i) > y_i$ if and only if $F(y_i) < u_i$; $R^{-1}(v_i) = y_i$ if and only if $R(y_i) \geqslant v_i$ and $R(y_i - \epsilon) < v_i$ for any $\epsilon > 0$.

To obtain the inverse map, we start by inverting a single observation. If $\delta_i = 1$, the inverse map for this datum is

$$Q_1^{F,R}(y_i, u_i, v_i) = \{F : F(y_i) \geqslant u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0\} \times \{R : R^{-1}(v_i) \geqslant y_i\}.$$

If $\delta_i = 0$, the inverse map is

$$Q_0^{F,R}(y_i, u_i, v_i) = \{F : F(y_i) < u_i\} \times \{R : R(y_i) \geqslant v_i, R(y_i - \epsilon) < v_i \text{ for any } \epsilon > 0\}.$$

Combining these we obtain the complete inverse map

$$Q^{F,R}(y, \delta, u, v) = \bigcap_i Q_{\delta_i}^{F,R}(y_i, u_i, v_i) = Q^F(y, \delta, u) \times Q^R(y, \delta, v), \tag{6}$$

where

$$Q^F(y, \delta, u) = \left\{ F : \begin{cases} F(y_i) \geqslant u_i, F(y_i - \epsilon) < u_i \text{ for any } \epsilon > 0 & \text{for all } i \text{ such that } \delta_i = 1, \\ F(y_j) < u_j & \text{for all } j \text{ such that } \delta_j = 0 \end{cases} \right\} \tag{7}$$

and $Q^R(y, \delta, v)$ is analogous. The inverse of $Q^{F,R}$ in (6) is in the form of a Cartesian product. This is a direct consequence of our choice of data-generating equation, and it greatly simplifies the calculation of the marginal fiducial distribution for failure times.

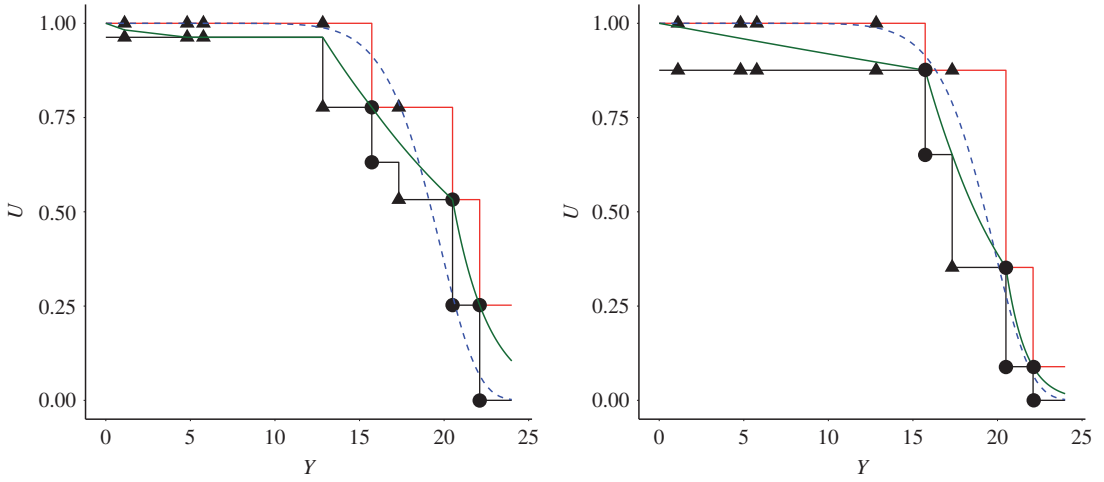

Fig. 2. Two realizations of fiducial curves for a sample of size 8 from Wei(20, 10) censored by Ex(20). Here fiducial curves refer to Monte Carlo samples S_i^L, S_i^U, and S_i^I $(i = 1, 2)$ from the generalized fiducial distribution. The red and black curves are corresponding realizations of the upper and fiducial survival functions. The green curve is the log-linear interpolation. The circles represent failure observations. The triangles represent censored observations. The dashed blue curve is the true survival function of Wei(20, 10). Since the fiducial distribution reflects uncertainty, we do not expect every fiducial curve to be close to the true survival function.

Figure 2 demonstrates the survival function representation of $Q^F(y, \delta, u)$, as defined in (7), for one dataset with $n = 8$ observations of X following Wei(20, 10) censored by Z following Ex(20). Each of the panels corresponds to a different value of u, where each u is a realization of U^*. Any survival function lying between the upper red and the lower black fiducial survival functions corresponds to an element of the closure of $Q^F(y, \delta, u)$. In particular, we plot in green the log-linear interpolation going through the failure observations as described in § 2.1 with a modification to ensure it stays in $Q^F(y, \delta, u)$. The details of the modification are given in Step 5.3 of Algorithm 1.

Algorithm 1.

Step 1. Generate n independent Un(0, 1) data items and sort them. Denote the sorted vector as $preU = (u_1, \ldots, u_n)$.

Step 2. Sort the pairs $\{y_i, \delta_i\}$ by the value of y_i. Relabel the sorted data as $(y_1, \ldots, y_n)$ and $(\delta_1, \ldots, \delta_n)$.

Step 3. Initialize $LowerFid = (0)_{n+1}$, $UpperFid = (1)_{n+1}$, where the subindex refers to the vector length.

Step 4. For $i = 1$ to n:

Let $UpperFid(i) = preU(1)$, where $preU(1)$ is the first and also the smallest element left in $preU$ and $UpperFid(i)$ is the ith element of the vector $UpperFid$.

If $\delta = 1$, set $LowerFid(i + 1) = preU(1)$, and delete $preU(1)$.

If $\delta = 0$, randomly pick one u from $preU$, set $LowerFid(i + 1) = LowerFid(i)$, and delete the selected u from $preU$.

In either case, deleting an entry decreases the dimension of $preU$ by 1.

Step 5. Output three survival functions that are needed for the conservative and log-linear interpolation methods.

5.1. Lower fiducial bound: using *LowerFid* as a fiducial curve.

5.2. Upper fiducial bound: using *UpperFid* as a fiducial curve.

5.3. Log-linear interpolation: Fit a continuous fiducial distribution by linear interpolation based on failure observations as described in § 2.1. Then correct the linear interpolation at the censoring observations so that the upper fiducial bound on the continuous distribution function, or the lower fiducial bound for the survival function, is satisfied. Let y_{n-k} $(k = 0, 1, \ldots, n-1)$ denote the last failure observation. Fit a single line after the last uncensored observation and take the maximum of $s_0, s_1, \ldots, s_k$ as the slope, where s_1 is the slope between $(y_{n-k}, \log u_{n-k})$ and $(y_{n-k+1}, \log u_{n-k+1})$, $\ldots$, s_k is the slope between $(y_{n-k}, \log u_{n-k})$ and $(y_n, \log u_n)$, and s_0 is the slope between $(\tilde{y}, \log \tilde{u})$ and $(y_{n-k}, \log u_{n-k})$, with $\tilde{y}$ being the second last uncensored observation. If there is only one failure time, $\tilde{y}$ and $\log \tilde{u}$ are 0.

Step 6. From steps 1–5 we get one curve of the fiducial distribution. Repeat steps 1–5 to get one fiducial sample with m curves.

When defining the generalized fiducial distribution, let (U^*, V^*) be independent of and have the same distribution as (U, V). Because the inverse (6) separates into a Cartesian product, and by the fact that U^* and V^* are independent, the marginal fiducial distribution for the failure distribution function F is

$$Q^F(y, \delta, U^*) \mid \{Q^F(y, \delta, U^*) \neq \emptyset\}. \tag{8}$$

As proved in the Supplementary Material, the conditional distribution in (8) can be sampled efficiently using Algorithm 1. Other choices of data-generating equations might lead to different fiducial distributions, but in the rest of this paper we only study the generalized fiducial distribution sampled from Algorithm 1.

Following § 2.1, let u be such that $Q^F(y, \delta, u) \neq \emptyset$ and define

$$F^{\mathrm{L}}_{(y,\delta,u)}(s) = \inf\{F(s) : F \in Q^F(y, \delta, u)\}, \quad F^{\mathrm{U}}_{(y,\delta,u)}(s) = \sup\{F(s) : F \in Q^F(y, \delta, u)\}. \tag{9}$$

Next we define a random distribution function $F^{\mathrm{L}}(s) = F^{\mathrm{L}}_{(y,\delta,U^*_Q)}(s)$, where $U^*_Q = (U^*_1, \ldots, U^*_n)$ is distributed as independent $\mathrm{Un}(0, 1)$ conditioned on the event $\{Q^F(y, \delta, U^*_Q) \neq \emptyset\}$. We will refer to the random survival function $S^U(s) = 1 - F^{\mathrm{L}}(s)$ as the upper fiducial survival function and its distribution as the upper fiducial distribution. The lower $S^L(s)$ and log-linear interpolated $S^I(s)$ fiducial survival functions are defined analogously. The closure of the inverse image (8) is the set of all distribution functions F that stay between F^{L} and F^{U}.

To illustrate the fiducial distribution in the right-censoring case, let the failure time X follow $\mathrm{Wei}(20, 10)$ and censoring time Z follow $\mathrm{Ex}(20)$ with sample size 300. The censoring percentage is about 60%. We plot a sample of the fiducial survival function $S^I_i(i = 1, \ldots, 1000)$ and the Kaplan–Meier estimator in Fig. 1(b). As expected, we see a wider spread of fiducial curves in the censoring case, indicating higher uncertainty.

Remark 1. The same marginal generalized fiducial distribution sampled from Algorithm 1 can also be derived for some explicit models in which failure and censoring times are dependent, as shown in the 2018 PhD dissertation by Y. Cui, University of North Carolina at Chapel Hill.

2.3. *Inference based on fiducial distribution*

We now describe how to use generalized fiducial distributions for inference, specifically, point estimation, pointwise confidence intervals for survival functions and quantiles, curvewise

confidence intervals, and testing. The actual numerical implementation will be based on a sample of the fiducial survival functions S_i^L, S_i^U, and S_i^I $(i = 1, \ldots, m)$, i.e., the lower bound, the upper bound, and the log-linear interpolation respectively, obtained from Algorithm 1 for generating Monte Carlo samples from the generalized fiducial distribution. We will call these samples of fiducial survival functions fiducial samples.

As shown in the Supplementary Material, the Kaplan–Meier estimator falls into the interval given by the expectation of the lower and upper fiducial bounds at any failure time t. However, instead of using the Kaplan–Meier estimator we propose to use the pointwise median of the log-linear interpolation fiducial distribution as a point estimator of the survival function. It follows from § 3 that the proposed estimator is asymptotically equivalent to the Kaplan–Meier estimator. Numerically, we estimate the median of the generalized fiducial distribution at time x by computing a pointwise median of the fiducial sample $S_i^I(x)$ $(i = 1, \ldots, m)$.

As explained at the end of § 2.1 we use two types of pointwise confidence intervals, conservative and log-linear interpolation, using quantiles of appropriate parts of the fiducial samples. For example, a 95% confidence log-linear interpolation confidence interval for $S(x)$ is formed by using the empirical 0.025 and 0.975 quantiles of $S_i^I(x)$. Similarly, a 95% conservative confidence interval is formed by taking the empirical 0.025 quantile of $S_i^L(x)$ as a lower limit and the empirical 0.975 quantile of $S_i^U(x)$ as an upper limit. Simulation results in § 4.1 show that the proposed confidence intervals match or outperform their main competitors in terms of coverage and length. In order to save space, in the rest of this section we present procedures based on the log-linear interpolation sample only. A conservative version can be obtained analogously.

In survival analysis, we are also interested in confidence intervals for quantile q of the survival function, where $0 < q < 1$. We obtain such a confidence interval by inverting the procedure of computing the pointwise confidence interval. Specifically, a 95% confidence interval is obtained by taking empirical 0.025 and 0.975 quantiles of the inverse of fiducial sample S_i^I evaluated at q.

Next, we discuss the use of the generalized fiducial distribution to obtain simultaneous curvewise confidence bands. In particular, for a $1 - \alpha$ curvewise confidence set we propose using a band $\{S : \|S - M\| \leqslant c\}$ of fiducial probability $1 - \alpha$, where M denotes the pointwise median of the generalized fiducial distribution and $\|\cdot\|$ is the L_∞-norm, i.e., $\|S - M\| = \max_x |S(x) - M(x)|$. Numerically we implement this by using a fiducial sample. Let

$$l_j = \|S_j^I - \hat{M}\| = \max_x |S_j^I(x) - \hat{M}(x)| \quad (j = 1, \ldots, m),$$

where $\hat{M}$ is the estimated pointwise median of the generalized fiducial distribution. Then we form the 95% curvewise confidence band $\{S : \|S - \hat{M}\| \leqslant \hat{c}\}$, where $\hat{c}$ is the 0.95 quantile of l_j.

Remark 2. The L_∞-norm determines the shape of the confidence band. Other choices, such as L_2-norm, are possible as long as the resulting confidence bands satisfy Corollary 1 in § 3.

The curvewise confidence set could be inverted for testing. The resulting test is different from the log-rank test (Mantel, 1966) and its modifications. Based on our definition of the $1-\alpha$ fiducial band, the fiducial p-value for the test

$$H_0 : S(t) = S_0(t) \text{ for all } t \quad \text{versus } H_1 : S(t) \neq S_0(t) \text{ for some } t$$

is $\text{pr}^*_{y,\delta}(\|S^I - M\| \geqslant \|S_0 - M\|)$, where $\text{pr}^*_{y,\delta}$ stands for a fiducial probability computed for observed data (y, δ), S^I stands for a random survival function following the log-linear interpolation generalized fiducial distribution, and as before M is the pointwise median of the fiducial distribution.

We estimate this p-value from a fiducial sample by finding the largest α for which the $1-\alpha$ curvewise confidence set contains S_0. In particular, let

$$l_0 = \max_x |S_0(x) - \hat{M}(x)|, \quad l_j = \max_x |S_j^I(x) - \hat{M}(x)| \quad (j = 1, \ldots, m). \tag{10}$$

Numerically, we approximate the p-value by the proportion of the fiducial sample satisfying $l_j \geqslant l_0$.

Finally, let us consider two-sample testing. For each sample, we have observed values y^i and censoring indicators δ^i $(i = 1, 2)$. The two independent log-linear interpolation generalized fiducial distributions are denoted by $S^I_{(y^i,\delta^i)}$ $(i = 1, 2)$. When testing $H_0 : S^1 - S^2 = \Theta_0$ we define a fiducial p-value as the fiducial probability $\mathrm{pr}^*_{y,\delta}(\|S^I_{(y^1,\delta^1)} - S^I_{(y^2,\delta^2)} - M_D\| \geqslant \|\Theta_0 - M_D\|)$, where M_D is the median of the difference of the two generalized fiducial distributions. Numerically, we evaluate the p-value in the same fashion as in (10). We will compare the performance of the proposed fiducial test with the log-rank test and sup log-rank test with different weights for the two-sample setting by simulation in § 4.2.

3. Theoretical results

In this section we study theoretical properties of the generalized fiducial distribution with respect to a specific data-generating equation that leads to the distribution of failure times generated by Algorithm 1. Recall that the generalized fiducial distribution is a probability distribution $\mathrm{pr}^*_{y,\delta}$ that is defined for every fixed dataset (y, δ). It can be made into a random measure $\mathrm{pr}^*_{Y,\Delta}$ in the same way as one defines the usual conditional distribution, i.e., by substituting random variables (Y, Δ) for the observed dataset (y, δ). In this section, we establish a Bernstein–von Mises theorem for this random measure assuming the observed data contain no ties with probability 1.

Praestgaard & Wellner (1993) prove a Bernstein–von Mises theorem for the exchangeably weighted bootstrap, of which the Bayesian bootstrap (Rubin, 1981) is an example. However, the result of Praestgaard & Wellner (1993) is not applicable in the survival setting due to the fact that the jump sizes of F^{L} or F^{U} are not exchangeable. Here, we study the theoretical properties of the generalized fiducial distribution in the survival setting. For simplicity, we state the results in this section using upper fiducial survival functions S^U, i.e., the lower fiducial bound of the cumulative distribution functions F^{L}. In the Supplementary Material we prove that the same results hold for S^L and S^I.

First we introduce some notation: X_i is the failure time, C_i is the censoring time, Y_i is the observed minimum of failure and censoring times, and $\Delta_i = I\{X_i \leqslant C_i\}$ is the censoring indicator. We define the counting process

$$N_i(t) = I\{Y_i \leqslant t\}\Delta_i, \quad \bar{N}(t) = \sum_{i=1}^n N_i(t)$$

and the at-risk process

$$K_i(t) = I\{Y_i \geqslant t\}, \quad \bar{K}(t) = \sum_{i=1}^n K_i(t).$$

The following lemma provides a useful alternative expression for the upper fiducial survival function, as defined immediately below (9). All proofs in this section are deferred to the

Supplementary Material. When the data (y, δ) are assumed fixed and known, the ordered failure times s_i and the functions $\bar{N}(t)$ and $\bar{K}(t)$ are nonrandom.

LEMMA 1. *For any fixed dataset without ties, the upper fiducial survival function is*

$$S^U(t) = \prod_{s_i \leqslant t} \{1 - B_i\}, \tag{11}$$

where s_i are ordered failure times and B_i are independent random variables following $\mathrm{Be}\{1, \bar{K}(s_i)\}$. *Furthermore, its expectation is*

$$\hat{S}(t) = E\{S^U(t)\} = \prod_{s_i \leqslant t} \left\{1 - \frac{1}{1 + \bar{K}(s_i)}\right\}. \tag{12}$$

As our first major result, we prove a concentration inequality for $S^U(t)$.

THEOREM 1. *The following bound holds for any fixed dataset with no ties, any $0 < t < \infty$ such that $\bar{K}(t) \geqslant 1$, and any $\epsilon > 0$:*

$$\begin{aligned} &\mathrm{pr}^*_{y,\delta}\left\{\sup_{s \leqslant t} |S^U(s) - \hat{S}(s)| \geqslant 3\epsilon^2/n^{1/2} + \bar{N}(t)/\bar{K}(t)^{-2}\right\} \\ &\quad \leqslant \bar{N}(t)\left[(1-\epsilon/n^{3/4})^{\bar{K}(t)} + 0.4^{\bar{K}(t)} + n/\left\{\epsilon^2 \bar{K}(t)\right\}^2\right]. \end{aligned} \tag{13}$$

Next we list all conditions needed for Theorem 2. The following two assumptions are also needed for theoretical study of the Kaplan–Meier estimator (Fleming & Harrington, 2011).

Assumption 1. There exists a function π such that as $n \to \infty$,

$$\sup_{0 \leqslant t < \infty} \left|\bar{K}(t)/n - \pi(t)\right| \to 0 \text{ almost surely.}$$

This assumption is very mild. For example if Y_i are independent and identically distributed, it is implied by the Glivenko–Cantelli theorem; see the discussion following Assumption 6.2.1 in Fleming & Harrington (2011) for more details.

Assumption 2. The distribution function of failure times F_0 is absolutely continuous.

Remark 3. Theorem 1 and Assumption 1 imply that the fiducial distribution is uniformly consistent on finite time intervals. In particular, provided that we have a sequence of data such that $\bar{K}(t)/n \to \pi(t) > 0$, the right-hand side of (13) is $O(n^{-1})$ whenever $\epsilon^2 = n^{-1/2}$.

Let $\tilde{S}(t) = \prod_{s \leqslant t}\{1 - \Delta\bar{N}(s)/\bar{K}(s)\}$ be the Kaplan–Meier estimator. It is well known (see for example Theorem 6.3.1 of Fleming & Harrington, 2011) that for any t satisfying $\pi(t) > 0$,

$$\sqrt{n}\{\tilde{F}(\cdot) - F_0(\cdot)\} \to \{1 - F_0(\cdot)\}W\{\gamma(\cdot)\} \quad \text{in distribution on } D[0, t], \tag{14}$$

where $\tilde{F}(t) = 1 - \tilde{S}(t)$, $\gamma(t) = \int_0^t \pi^{-1}(s)\,\mathrm{d}\Lambda(s)$, W is Brownian motion, and Λ is the cumulative hazard function. Similarly, (12) provides us with a modification of the Kaplan–Meier estimator

that also satisfies (14). We will use this modification throughout this section and in all the proofs that can be found in the Supplementary Material. Next we state two additional assumptions specific to the Bernstein–von Mises theorem for generalized fiducial distributions.

Assumption 3. Let $\int_0^t f_n(s)/\bar{K}(s)\,\mathrm{d}\bar{N}(s) \to \int_0^t \lambda(s)/\pi(s)\,\mathrm{d}s$ almost surely for any $t \in \mathcal{I} = \{t : \pi(t) > 0\}$ and any sequence of functions $f_n \to \pi^{-1}$ uniformly, where π is defined in Assumption 1.

Assumption 3 is reasonable since the probability of failure and censoring both happening in $[t, t+\Delta t)$ is of a higher order, $O\{(\Delta t)^2\}$.

Assumption 4. Let $\sup_{0\leqslant s\leqslant t} |\tilde{F}(s) - F_0(s)| \to 0$ almost surely for any $t \in \mathcal{I} = \{t : \pi(t) > 0\}$, where $\tilde{F} = 1 - \tilde{S}$, with $\tilde{S}$ the Kaplan–Meier estimator.

Remark 4. The strong consistency result of Assumption 4 has been proved for the model described in § 2.2 by Gill (1994) and Stute & Wang (1993). Moreover, Assumption 4 is only needed for establishing a strong version of Theorem 2, i.e., convergence in distribution almost surely. If the Kaplan–Meier estimator only converges in probability, then the convergence mode in Theorem 2 is in distribution in probability.

The following theorem establishes a Bernstein–von Mises theorem for the fiducial distribution. In particular, we will show that the fiducial distribution of $n^{1/2}\{F^{\mathrm{L}}(\cdot) - \hat{F}(\cdot)\}$, where $\hat{F}(\cdot) = 1 - \hat{S}(\cdot)$ and $F^{\mathrm{L}}(\cdot) = 1 - S^{U}(\cdot)$, converges in distribution on $D[0,t]$ almost surely to the same Gaussian process as in (14). To understand the somewhat unusual mode of convergence used here, notice that there are two sources of randomness present. One is the randomness of the fiducial distribution defined for each fixed dataset. The other is the usual randomness of the data. The mode of convergence here is in distribution almost surely, i.e., the centred and scaled fiducial distribution viewed as a random probability measure on $D[0,t]$ converges almost surely to the Gaussian process described in the right-hand side of (14) using the weak topology on the space of probability measures.

THEOREM 2. *Based on Assumptions 1–4, for any* $t \in \mathcal{I} = \{t : \pi(t) > 0\}$,

$$n^{1/2}\{F^{\mathrm{L}}(\cdot) - \hat{F}(\cdot)\} \to \{1 - F_0(\cdot)\}W\{\gamma(\cdot)\}$$

in distribution on $D[0,t]$ *almost surely, where* $\gamma(t) = \int_0^t \pi^{-1}(s)\,\mathrm{d}\Lambda(s)$.

Theorem 2 implies that the pointwise fiducial confidence intervals are equivalent to the asymptotic confidence intervals based on the Kaplan–Meier estimator. This fact can be also seen from Theorem 2 of Fay et al. (2013). The following corollary shows that Theorem 2 also implies that all the pointwise and curvewise confidence intervals described in § 2.3 have asymptotically correct coverage. Consequently, the tests described in § 2.3 also have asymptotically correct Type I error.

COROLLARY 1. *Let* $\Psi\{\phi(\cdot)\}$ *be a map* $D[0,t] \to \mathbb{R}$ *with the properties that there exists a function* ψ *such that*

$$\Psi\{\phi(\cdot)\} = \Psi\{-\phi(\cdot)\}, \quad \Psi\{a\phi(\cdot)\} = \psi(a)\Psi\{\phi(\cdot)\} \tag{15}$$

for all $\phi \in D[0,t]$ *and* $a > 0$, *and that the distribution of the random variable* $\Psi[\{1 - F_0(\cdot)\}W\{\gamma(\cdot)\}]$ *is continuous and the* $(1-\alpha)$*th quantile of this distribution is unique.*

Then, under the assumptions in Theorem 2, any set $C_{n,\alpha} = \{F : \Psi\{F(\cdot) - \hat{F}(\cdot)\} \leqslant \epsilon_{n,\alpha}\}$ *with* $\mathrm{pr}^*_{y,\delta}(C_{n,\alpha}) = 1 - \alpha$ *is a* $1 - \alpha$ *asymptotic confidence set for* F_0.

4. Simulation study

4.1. *Coverage of pointwise confidence intervals and mean square error of point estimators*

We present comparisons of frequentist properties of the proposed fiducial confidence intervals with a number of competing methods. We will consider two basic groups of settings, one with heavy censoring from Fay et al. (2013) and another with a moderate level of censoring from Barber & Jennison (1999). In both cases the proposed generalized fiducial distribution intervals perform comparably to or better than the other methods.

First we reproduce the settings in Fay et al. (2013) that have a very high level of censoring. Fay et al. (2013) compared their proposed beta product confidence procedure methods with a number of asymptotic methods. These include: the method of Greenwood by logarithmic transformation; the confidence interval on the Kaplan–Meier estimator using Greenwood's variance by logarithmic transformation (Therneau, 2015); modified Greenwood by logarithmic transformation, which modifies the estimator of variance for the lower limit by multiplying Greenwood's variance estimator by $K(y_i)/K(t)$ at t, where y_i is the largest observed survival less than or equal to t (Therneau, 2015); Borkowf's method by logarithmic transformation, which gives wider intervals with more censoring and assumes normality on $\log\{\tilde{S}(t)\}$, where $\tilde{S}(t)$ is the Kaplan–Meier estimator (Borkowf, 2005); shrinkage Borkowf by logarithmic transformation, which uses a shrinkage estimator of the Kaplan–Meier estimator with a hybrid variance estimator (Borkowf, 2005), the Strawderman–Wells method, which uses the Edgeworth expansion for the distribution of the studentized Nelson–Aalen estimator (Strawderman et al., 1997; Strawderman & Wells, 1997); the Thomas–Grunkemeier method, a likelihood ratio method which depends on a constrained product-limit estimator of the survival function (Thomas & Grunkemeier, 1975); constrained beta, which refers the distribution of $\tilde{S}(t)$ to a beta distribution subject to some constraints (Barber & Jennison, 1999); nonparametric bootstrap (Efron, 1981; Akritas, 1986); and constrained bootstrap, an improved bootstrap approximation subject to some constraints (Barber & Jennison, 1999).

Simulation studies reported in Fay et al. (2013) show that the above asymptotic methods have a coverage problem, i.e., the error rate of 95% confidence intervals of all these methods is larger than 5% in their high-censoring scenarios. Therefore in this section we focus on comparing the fiducial methods with our main competing methods, which are the beta product confidence procedure (Fay et al., 2013), mid-p beta product confidence procedure (Fay & Brittain, 2016); see also Chapter 11 of (Schweder & Hjort, 2016), and the binomial procedure (Clopper & Pearson, 1934), which maintain the coverage. Additionally, we include the Thomas & Grunkemeier (1975) confidence interval, which can be viewed as the empirical likelihood applied to the survival distribution for right-censored data; see pages 144–5 of Owen (2001). For each of the methods, we report the error rates of coverage and the average width of 95% confidence intervals.

In particular, we consider the following two scenarios in Fay et al. (2013). In the first scenario, the failure time is $\mathrm{Ex}(10)$ and the censoring time is $\mathrm{Un}(0, 5)$. The censoring percentage is approximately 80%. We simulate 100 000 independent datasets of size 30 and apply our methods with fiducial sample size 1000. In the second scenario, we reproduce the setting using a mixture of exponentials to mimic the pilot study of treatment in severe systemic sclerosis (Nash et al., 2007). In particular, the failure time is a mixture of $\mathrm{Ex}(0.227)$ with probability 0.187 and $\mathrm{Ex}(22.44)$ with probability 0.813, and the censoring time is $\mathrm{Un}(2, 8)$. The censoring percentage is about 65%. We simulate 100 000 independent datasets of size 34 and apply our methods with fiducial sample size 1000.

Table 1. *Error rates in percentages and average width of* 95% *confidence intervals for Scenario 1. Failure time is* Ex(10). *Censoring time is* Un(0, 5). *The censoring percentage is about* 80%

	$t=1$			$t=2$			$t=3$			$t=4$		
	L	U	W	L	U	W	L	U	W	L	U	W
FD-I	1.9	2.7	0.21	1.5	2.8	0.29	1.4	3.0	0.37	1.8	3.1	0.45
FD-C	0.0	1.4	0.26	0.3	1.6	0.36	0.1	1.5	0.46	0.0	1.4	0.63
BPCP-MM	0.0	1.3	0.26	0.3	1.4	0.35	0.1	1.3	0.46	0.0	1.0	0.62
BPCP-MC	0.0	1.3	0.25	0.4	1.5	0.35	0.1	1.5	0.46	0.0	1.4	0.63
BPCP-MP	0.0	2.2	0.23	0.8	2.3	0.32	0.4	2.2	0.41	0.0	2.0	0.57
BN	0.0	1.4	0.26	0.7	1.3	0.38	0.6	1.3	0.51	0.1	0.9	0.70
TG	6.7	2.1	0.20	3.8	2.3	0.29	4.0	2.4	0.37	5.5	2.4	0.43

FD-I, the proposed fiducial confidence interval using log-linear interpolation; FD-C, the proposed fiducial conservative confidence interval; BPCP-MM, beta product confidence procedure using method of moments; BPCP-MC, beta product confidence procedure using Monte Carlo; BPCP-MP, mid-p beta product confidence procedure; BN, Clopper–Pearson binomial confidence interval; TG, Thomas–Grunkemeier confidence interval; L, error rate that the true parameter is less than the lower confidence limit; U, error rate that the true parameter is greater than the upper confidence limit; two-sided error rate is obtained by adding the values in columns L and U; values less than 2.5% in individual columns, or 5% in aggregate, indicate good performance; W, average width of the confidence interval.

Table 2. *Error rates in percentages and average width of* 95% *confidence intervals for Scenario 2. Failure time is a mixture of* Ex(0.227) *and* Ex(22.44) *with probability* 0.187 *and* 0.813*, respectively. Censoring time is* Un(0, 5). *The censoring percentage is about* 65%

	$t=3$			$t=4$			$t=5$			$t=6$		
	L	U	W	L	U	W	L	U	W	L	U	W
FD-I	2.2	2.7	0.29	1.9	2.9	0.31	1.7	3.0	0.33	1.5	3.2	0.36
FD-C	1.2	1.7	0.33	0.7	1.8	0.36	0.4	1.8	0.40	0.1	1.7	0.46
BPCP-MM	1.3	1.7	0.33	0.7	1.7	0.35	0.4	1.6	0.39	0.1	1.4	0.46
BPCP-MC	1.2	1.8	0.32	0.7	2.0	0.35	0.4	1.9	0.39	0.1	1.9	0.46
BPCP-MP	1.8	2.1	0.30	1.6	2.4	0.32	0.9	2.5	0.36	0.4	2.3	0.41
BN	1.4	1.5	0.35	1.5	1.6	0.40	1.5	1.7	0.46	1.0	1.5	0.56
TG	1.8	2.1	0.29	2.4	2.4	0.31	2.8	2.5	0.33	3.6	2.5	0.35

See Table 1 for description of abbreviations.

The simulation results are presented in Tables 1 and 2. We see that our confidence intervals using log-linear interpolation maintain the aggregate coverage, are much shorter than the other conservative methods, but may be slightly biased to the left. Not surprisingly, the performance of the proposed conservative confidence interval is similar to that of the beta product confidence procedure method. The Thomas–Grunkemeier confidence intervals have coverage problems in these high-censoring scenarios and are only slightly shorter than the fiducial confidence intervals using log-linear interpolation.

Our second simulation study setting comes from Barber & Jennison (1999), where the censoring rate is relatively low. In the third scenario, the failure time follows Ex(10), and the censoring time is Ex(25). The sample size $n = 50$ and the censoring percentage is about 30%. In the fourth scenario, the failure time follows Ex(10), and the censoring time is Ex(50). The sample size $n = 100$ and the censoring percentage is about 15%. The empirical error rates and average width of confidence intervals from 5000 simulations are presented in Tables 3 and 4. We see that the proposed fiducial confidence intervals using log-linear interpolation do as well as the Thomas–Grunkemeier confidence intervals in terms of both coverage and average length in these settings.

Table 3. *Error rates in percentages and average width of* 95% *confidence intervals for Scenario 3. Failure time is* Ex(10). *Censoring time is* Ex(25). *The censoring percentage is about* 30%

	$S(t) = 0.8$			$S(t) = 0.6$			$S(t) = 0.4$			$S(t) = 0.2$		
	L	U	W	L	U	W	L	U	W	L	U	W
FD-I	2.4	2.3	0.22	2.6	2.4	0.27	1.8	2.8	0.29	2.0	3.1	0.25
FD-C	1.4	1.8	0.24	1.5	1.8	0.30	1.1	1.9	0.32	0.9	1.4	0.29
BPCP-MM	1.6	1.8	0.24	1.6	1.6	0.30	1.3	1.8	0.32	1.1	1.3	0.29
BPCP-MC	1.7	1.9	0.24	1.8	1.7	0.30	1.3	1.6	0.32	1.2	1.3	0.29
BPCP-MP	2.4	2.3	0.22	2.4	2.2	0.28	1.9	2.5	0.30	1.9	2.3	0.27
BN	1.7	1.9	0.24	1.7	1.5	0.31	1.5	1.8	0.34	1.2	1.3	0.32
TG	2.6	2.3	0.22	2.8	2.3	0.28	2.2	2.9	0.29	3.0	0.0	0.26

See Table 1 for description of abbreviations.

Table 4. *Error rates in percentages and average width of* 95% *confidence intervals for Scenario 4. Failure time is* Ex(10). *Censoring time is* Ex(50). *The censoring percentage is about* 15%

	$S(t) = 0.8$			$S(t) = 0.6$			$S(t) = 0.4$			$S(t) = 0.2$		
	L	U	W	L	U	W	L	U	W	L	U	W
FD-I	3.0	2.4	0.16	2.2	2.2	0.19	2.5	2.4	0.20	2.1	2.3	0.17
FD-C	2.0	2.0	0.17	1.7	1.6	0.20	1.9	1.8	0.21	1.6	1.4	0.18
BPCP-MM	2.0	1.8	0.17	1.6	1.6	0.20	1.9	1.7	0.21	1.6	1.4	0.18
BPCP-MC	2.1	1.8	0.17	2.0	1.6	0.20	1.9	1.7	0.21	1.5	1.5	0.18
BPCP-MP	3.0	2.4	0.16	2.1	1.9	0.20	2.5	2.2	0.20	2.0	2.1	0.17
BN	2.1	1.9	0.17	1.6	1.5	0.21	2.3	1.9	0.22	1.6	1.6	0.19
TG	3.3	2.4	0.16	2.1	1.9	0.19	2.5	2.3	0.20	2.0	2.3	0.17

See Table 1 for description of abbreviations.

Table 5. *Pointwise root mean square error of survival function estimators multiplied by* 100. *Failure time is* Ex(1). *Censoring time is* Un(0, 5). *The censoring percentage is about* 20%

	$S(t) = 0.99$	$S(t) = 0.9$	$S(t) = 0.75$	$S(t) = 0.5$	$S(t) = 0.25$	$S(t) = 0.1$	$S(t) = 0.01$
FD-I	1.7	5.6	8.4	10.0	9.1	6.6	3.5
BPCP-MM	2.1	5.9	8.7	10.3	9.4	6.6	3.9
BPCP-MP	2.2	6.0	8.7	10.3	9.5	7.6	1.6
KML	2.0	6.0	8.8	10.5	9.7	7.9	1.7
KMM	2.0	6.0	8.8	10.5	9.7	7.6	2.8
KMH	2.0	6.0	8.8	10.5	9.7	7.5	5.4

FD-I, method using the pointwise median of S^I as a point estimator of the survival function; BPCP-MM and BPCP-MP, associated median unbiased estimators defined in Fay et al. (2013); KML, KMH, and KMM, three variants of the Kaplan–Meier estimator (Fay et al., 2013); if the largest observation is censored, KML is defined as 0, KMH is defined as the Kaplan–Meier at the last value, and KMM $= 0.5 \times$ KML $+ 0.5 \times$ KMH for arguments beyond the largest observation.

We also perform a simulation for the root mean square error of survival functions, adopting a setting in Fay et al. (2013). Here, the failure time is Ex(1) and the censoring time is Un(0, 5). The censoring percentage is about 20%. We simulate 100 000 independent datasets of size 25 and apply our fiducial methods with fiducial sample size 10 000. Since the Kaplan–Meier estimator is not defined after the largest observation if it is censored, we follow Fay et al. (2013) and define it in three ways. We evaluate root mean square error at t, where $S(t) = 0.99, 0.9, 0.75, 0.5, 0.25, 0.1, 0.01$. We report the results in Table 5. We see that the proposed fiducial approach has the smallest root mean square error for $S(t) = 0.99, 0.9, 0.75, 0.5, 0.25, 0.1$.

Table 6. *Estimated power/ Type I error, in percentage, of level*-0.05 *tests*

Scenario	FD	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
1	87.0	94.0	87.8	91.4	89.8	89.4	88.8	89.8	81.4	87.0	85.0	84.4	85.8
2	99.4	90.2	29.0	60.8	52.2	51.4	100	84.2	48.2	61.8	58.2	57.4	100
3	49.4	6.6	16.0	8.0	10.6	10.8	37.6	30.2	56.2	46.6	51.4	51.4	30.6
4	89.2	30.6	86.8	70.2	76.4	77.8	7.0	88.0	96.8	94.8	95.8	95.8	36.2
5	4.8	6.4	6.2	6.4	6.2	6.2	6.6	5.4	4.4	5.0	5.0	5.0	4.6

FD, the proposed fiducial test; LR, the original log-rank test with weight 1 (Mantel, 1966); GW, Gehan–Breslow generalized Wilcoxon test, i.e., log-rank test weighted by the number at risk overall (Gehan, 1965); TW, log-rank test weighted by the square root of the number at risk overall (Tarone & Ware, 1977); PP, log-rank test with Peto and Peto's modified survival estimate (Peto & Peto, 1972); MPP, log-rank test with modified Peto–Peto survival estimate (Andersen & Gill, 1982); FH, Fleming–Harrington weighted log-rank test (Harrington & Fleming, 1982); the last six tests are sup versions of LR, GW, TW, PP, MPP, and FH, respectively.

4.2. *Comparison of the proposed fiducial test and log-rank tests for two-sample testing*

We compare the performance of the proposed fiducial approach with different types of tests for testing the equality of two survival functions (Dardis, 2016). A common approach to testing the difference of two survival curves is the log-rank test. There are several modifications of the log-rank tests and sup log-rank tests that consist of reweighting (Gehan, 1965; Mantel, 1966; Peto & Peto, 1972; Tarone & Ware, 1977; Andersen & Gill, 1982; Harrington & Fleming, 1982; Fleming et al., 1987; Eng & Kosorok, 2005).

Five simulation settings from Li et al. (2015) are considered here. Survival configurations, i.e., the survival functions of failure times, are plotted in Figure 1 in Li et al. (2015). The censoring times are generated from uniform distributions $\mathrm{Un}(0, a)$ and $\mathrm{Un}(0, b)$, where the values of a and b are chosen so that the censoring rates are approximately 40%. For each scenario, we simulated 500 independent datasets of size 50 and applied the proposed fiducial test with fiducial sample size 1000 as well as the 12 existing methods mentioned above. The power or Type I error of the tests at the $\alpha = 0.05$ level, i.e., the percentage of p-values less than 0.05, for all of the scenarios described below are shown in Table 6. The reported powers of the log-rank test, Gehan–Breslow generalized Wilcoxon test, and Tarone–Ware test are consistent with Tables 1–5 in Li et al. (2015).

In the first scenario, two survival curves have proportional hazard functions. For the first group, the failure time follows $\mathrm{Ex}(2)$, and the censoring time is $\mathrm{Un}(0, 6)$. For the second group, let $\mathrm{Ex}(5)$ be the distribution of the failure time and $\mathrm{Un}(0, 11)$ the distribution of the censoring time. We see that the log-rank test has the highest power, and the proposed fiducial test is comparable to other types of tests in this case.

The second scenario is a setting with an early crossing of the survival curves. For the first group, the failure time follows $\mathrm{Wei}(2.5, 30)$, and the censoring time is $\mathrm{Un}(0, 65)$. For the second group, the failure time has the hazard rate $\lambda = 0.125I\{t \leqslant 1\} + 0.01I\{t \geqslant 1\}$, and the censoring time follows $\mathrm{Un}(0, 160)$. The proposed fiducial test performs better than other types of log-rank tests except for the Fleming–Harrington weighted log-rank test and the sup Fleming–Harrington weighted log-rank test. Fleming–Harrington weighted log-rank tests; may use better weights than other log-rank tests: however, the proposed fiducial test does not need to specify any weight.

The third scenario is a setting with a middle crossing of the survival curves. For the first group, the failure time follows $\mathrm{Ex}(12)$, and the censoring time is $\mathrm{Un}(0, 28)$. For the second group, the failure time has the hazard rate $\lambda = 1/4I\{t \leqslant 2\} + 1/35I\{t \geqslant 2\}$, and the censoring time follows $\mathrm{Un}(0, 33)$. We can see that overall, the proposed fiducial test and sup log-rank tests perform better than other types of log-rank tests.

The fourth scenario is a setting with a late crossing of the survival curves. For the first group, the failure time follows $\mathrm{Wei}(1.5, 5)$, and the censoring time follows $\mathrm{Un}(0, 11)$. For the second group, the failure time has the hazard rate $\lambda = 0.5I\{t \leqslant 1.5\} + 0.1I\{t \geqslant 1.5\}$, and the censoring

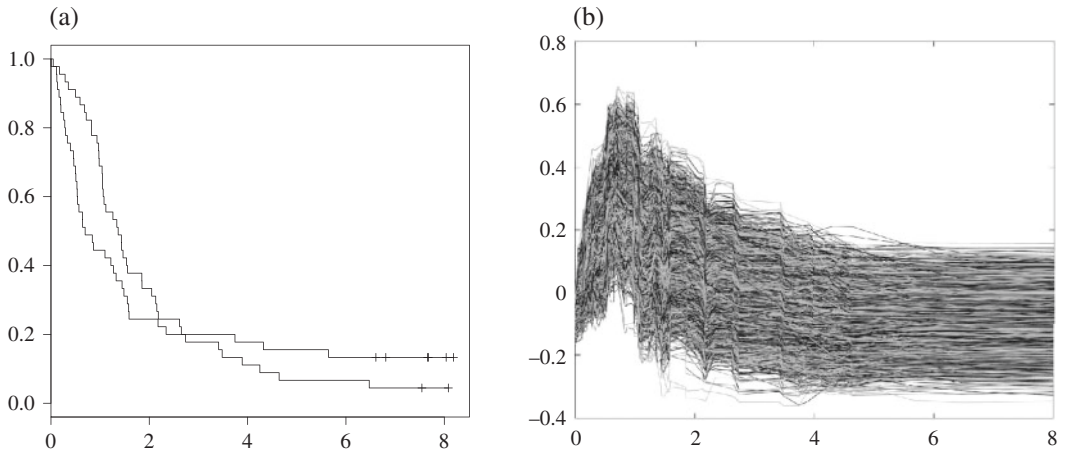

Fig. 3. (a) Kaplan–Meier estimators for two treatment groups. (b) Difference of two-sample fiducial distributions using log-linear interpolation.

Table 7. *p-values in percentages for testing the difference between chemotherapy and chemotherapy combined with radiotherapy*

	FD	LR	GW	TW	PP	MPP	FH	SLR	SGW	STW	SPP	SMPP	SFH
p-value	0.2	63.5	4.6	16.8	4.6	4.3	90.6	5.6	0.6	1.5	0.6	0.6	22.8

See Table 6 for description of abbreviations.

time follows $\mathrm{Un}(0, 10)$. Again, the proposed fiducial test and sup log-rank tests perform better than other types of log-rank tests.

To investigate Type I errors, in the fifth scenario two samples are independently generated from an exponential distribution with a hazard rate of 0.25. The censoring time is $\mathrm{Un}(0, 9)$ for both groups. We observe that the *p*-values of all methods follow a uniform distribution under H_0. The percentages of *p*-values less than 0.05 for all methods are about 0.05, indicating good Type I error performance.

The overall conclusion is that the proposed fiducial test has good power against all of the alternative hypotheses considered in Li et al. (2015). The Supplementary Material contains simulation results for additional scenarios, also showing the good power of the fiducial test.

5. GASTRIC TUMOUR STUDY

In this section, we analyse the following dataset presented in Klein & Moeschberger (2005). A clinical trial of chemotherapy against chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer was conducted by the Gastrointestinal Tumor Study Group (Schein, 1982). In this trial, 45 patients were randomized to each of the two groups and followed for several years. The censoring percentage is 13.3% for the combined therapy group, and 4.4% for the chemotherapy group. We are interested in testing whether the two treatment groups have the same survival functions.

We draw the Kaplan–Meier curves for these two datasets in Fig. 3(a). We notice that the two hazards appear to be crossing, which could pose a problem for some log-rank tests. Table 7 reports *p*-values obtained using the same 13 tests described in § 4.2.

To explain why the fiducial approach gives a small *p*-value on this dataset, we plot the sample of the difference of two fiducial distributions in Fig. 3(b). If these two datasets are from the same distribution, 0 should be well within the sample curves. However, from the picture, we can see

that the majority of curves are very far away from 0 on the interval [0.5, 1]. This indicates that the group with combined therapy has significantly worse early survival outcomes.

Acknowledgement

We thank Michael P. Fay, Michael R. Kosorok and Jonathan Williams for helpful conversations and suggestions. We thank the editor, associate editor, and reviewers for many useful comments which led to an improved manuscript. Cui and Hannig were partly supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes all proofs, and additional simulation scenarios.

References

Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *J. Am. Statist. Assoc.* **81**, 1032–8.
Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100–20.
Barber, S. & Jennison, C. (1999). Symmetric tests and confidence intervals for survival probabilities and quantiles of censored survival data. *Biometrics* **55**, 430–6.
Bayarri, M. J., Berger, J. O., Forte, A. & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40**, 1550–77.
Berger, J. O., Bernardo, J. M. & Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–38.
Berger, J. O., Bernardo, J. M. & Sun, D. (2012). Objective priors for discrete parameter spaces. *J. Am. Statist. Assoc.* **107**, 636–48.
Borkowf, C. B. (2005). A simple hybrid variance estimator for the Kaplan–Meier survival function. *Statist. Med.* **24**, 827–51.
Brillinger, D. R. (1962). Examples bearing on the definition of fiducial probability with a bibliography. *Ann. Math. Statist.* **33**, 1349–55.
Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Pacific Grove, California: Wadsworth and Brooks/Cole Advanced Books and Software, 2nd ed.
Cisewski, J. & Hannig, J. (2012). Generalized fiducial inference for normal linear mixed models. *Ann. Statist.* **40**, 2102–27.
Clopper, C. J. & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–13.
Dardis, C. (2016). *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.4.
Dempster, A. P. (2008). The Dempster-Shafer calculus for statisticians. *Int. J. Approx. Reason.* **48**, 365–77.
Edlefsen, P. T., Liu, C. & Dempster, A. P. (2009). Estimating limits from Poisson counting data using Dempster–Shafer analysis. *Ann. Appl. Statist.* **3**, 764–90.
Efron, B. (1981). Censored data and the bootstrap. *J. Am. Statist. Assoc.* **76**, 312–9.
Eng, K. H. & Kosorok, M. R. (2005). A sample size formula for the supremum log-rank statistic. *Biometrics* **61**, 86–91.
Fay, M. P. & Brittain, E. H. (2016). Finite sample pointwise confidence intervals for a survival distribution with right-censored data. *Statist. Med.* **35**, 2726–40.
Fay, M. P., Brittain, E. H. & Proschan, M. A. (2013). Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics* **14**, 723–36.
Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–25.
Fisher, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.* **xxvi**, 528–35.
Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proc. R. Soc. London* A **139**, 343–8.
Fisher, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics* **VI**, 91–8.
Fleming, T. R. & Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. Hoboken, New Jersey: John Wiley & Sons.
Fleming, T. R., Harrington, D. P. & O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *J. Am. Statist. Assoc.* **82**, 312–20.
Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–23.
Gill, R. D. (1994) Glivenko–Cantelli for Kaplan–Meier. *Math. Meth. Statist*. **3**, 76–87.
Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19**, 491–544.

HANNIG, J. (2013). Generalized fiducial inference via discretization. *Statist. Sinica* **23**, 489–514.
HANNIG, J., FENG, Q., IYER, H. K., WANG, J. C.-M. & LIU, X. (2018). Fusion learning for inter-laboratory comparisons. *J. Statist. Plan. Infer.* **195**, 64–79.
HANNIG, J., IYER, H., LAI, R. C. & LEE, T. C. (2016). Generalized fiducial inference: A review and new results. *J. Am. Statist. Assoc.* **111**, 1346–61.
HANNIG, J., IYER, H. K. & WANG, J. C.-M. (2007). Fiducial approach to uncertainty assessment: Accounting for error due to instrument resolution. *Metrologia* **44**, 476–83.
HANNIG, J. & LEE, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika* **96**, 847–60.
HARRINGTON, D. P. & FLEMING, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–66.
HJORT, N. L. & SCHWEDER, T. (2018). Confidence distributions and related themes. *J. Statist. Plan. Infer.* **195**, 1–13.
KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.
KLEIN, J. P. & MOESCHBERGER, M. L. (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer Science & Business Media.
LAI, R. C. S., HANNIG, J. & LEE, T. C. M. (2015). Generalized fiducial inference for ultra-high dimensional regression. *J. Am. Statist. Assoc.* **110**, 760–72.
LI, H., HAN, D., H. Y., CHEN, H. & CHEN, Z. (2015). Statistical inference methods for two crossing survival curves: A comparison of methods. *PLoS One* **10**, e0116774.
LIU, Y. & HANNIG, J. (2017). Generalized fiducial inference for logistic graded response models. *Psychometrika* **82**, 1097–125.
MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo. Rep. Part 1* **50**, 163–70.
MARTIN, R. & LIU, C. (2015). *Inferential Models: Reasoning with Uncertainty*. Boca Raton, Florida: Chapman & Hall/CRC.
NASH, R. A., MCSWEENEY, P. A., CROFFORD, L. J., ABIDI, M., CHEN, C.-S., GODWIN, J. D., GOOLEY, T. A., HOLMBERG, L., HENSTORF, G., LEMAISTRE, C. F. et al. (2007). High-dose immunosuppressive therapy and autologous hematopoietic cell transplantation for severe systemic sclerosis: Long-term follow-up of the US multicenter pilot study. *Blood* **110**, 1388–96.
OWEN, A. (2001). *Empirical Likelihood*. New York: Chapman and Hall.
PETO, R. & PETO, J. (1972). Asymptotically efficient rank invariant test procedures. *J. R. Statist. Soc.* A **135**, 185–207.
PRAESTGAARD, J. & WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Prob.* **21**, 2053–86.
RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130–4.
SCHEIN, P. S. (1982). A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. *Cancer* **49**, 1771–7.
SCHWEDER, T. & HJORT, N. L. (2016). *Confidence, Likelihood, Probability*. Cambridge: Cambridge University Press.
SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.
STRAWDERMAN, R. L., PARZEN, M. I. & WELLS, M. T. (1997). Accurate confidence limits for quantiles under random censoring. *Biometrics* **53**, 1399–415.
STRAWDERMAN, R. L. & WELLS, M. T. (1997). Accurate bootstrap confidence limits for the cumulative hazard and survivor functions under random censoring. *J. Am. Statist. Assoc.* **92**, 1356–74.
STUTE, W. & WANG, J.-L. (1993). The strong law under random censorship. *Ann. Statist*. **21**, 1591–607.
TARONE, R. E. & WARE, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156–60.
THERNEAU, T. M. (2015). *A Package for Survival Analysis in S*. Version 2.38.
THOMAS, D. R. & GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Am. Statist. Assoc.* **70**, 865–71.
WANDLER, D. V. & HANNIG, J. (2012). Generalized fiducial confidence intervals for extremes. *Extremes* **15**, 67–87.
WANG, J. C.-M., HANNIG, J. & IYER, H. K. (2012). Pivotal methods in the propagation of distributions. *Metrologia* **49**, 382–9.
WANG, J. C.-M. & IYER, H. K. (2005). Propagation of uncertainties in measurements using generalized inference. *Metrologia* **42**, 145–53.
WANG, J. C.-M. & IYER, H. K. (2006a). A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. *Measurement* **39**, 856–63.
WANG, J. C.-M. & IYER, H. K. (2006b). Uncertainty analysis for vector measurands using fiducial inference. *Metrologia* **43**, 486–94.
XIE, M. & SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Statist. Rev.* **81**, 3–39.

[*Received on* 25 *Ocober* 2017. *Editorial decision on* 13 *November* 2018]