Published in final edited form as:

Stat Med. 2016 July 20; 35(16): 2726–2740. doi:10.1002/sim.6905.

Finite Sample Pointwise Confidence Intervals for a Survival Distribution with Right-Censored Data

Michael P. Fay* and Erica H. Brittain

National Institute of Allergy and Infectious Diseases, 5601 Fishers Lane MSC 9820, Bethesda MD 20892, USA

Abstract

We review and develop pointwise confidence intervals for a survival distribution with right censored data for small samples, assuming only independence of censoring and survival. When there is no censoring, at each fixed time point the problem reduces to making inferences about a binomial parameter. In this case, the recently developed beta product confidence procedure (BPCP) gives the standard exact central binomial confidence intervals of Clopper and Pearson. Additionally, the BPCP has been shown to be exact (gives guaranteed coverage at the nominal level) for progressive type II censoring, and has been shown by simulation to be exact for general independent right censoring. In this paper, we modify the BPCP, to create a “mid-p” version, that reduces to the mid-p confidence interval for a binomial parameter when there is no censoring. We perform extensive simulations on both the standard and mid-p BPCP using a method of moments implementation that enforces monotonicity over time. All simulated scenarios suggest that the standard BPCP is exact. The mid-p BPCP, like other mid-p confidence intervals, has simulated coverage closer to the nominal level but may not be exact for all survival times, especially in very low censoring scenarios. In contrast, the two asymptotically-based approximations have lower than nominal coverage in many scenarios. This poor coverage is due to the extreme inflation of the lower error rates, although the upper limits are very conservative. Both the standard and the mid-p BPCP methods are available in our `bpcp` R package.

Keywords

beta product confidence procedure; exact test; Kaplan-Meier estimator; mid p-value; survival analysis

1. Introduction

In our recent paper written with Michael Proschan, we developed the beta product confidence procedure (BPCP) that gives pointwise confidence intervals for a single survival distribution for right censored data [1]. The BPCP is intended to be exact in the sense that it guarantees coverage at the nominal level for all sample sizes regardless of the survival or censoring distribution as long as the censoring is independent of the survival. Under

*Correspondence to: Michael P. Fay, National Institute of Allergy and Infectious Diseases, 5601 Fishers Lane MSC 9820, Bethesda MD 20892, USA. mfay@niaid.nih.gov.

progressive type II censoring, when a proportion of the remaining individuals are censored immediately following each failure time, the BPCP has been shown to be exact. In the special case when there is no censoring, the BPCP for the survival distribution at a fixed point in time is equivalent to the standard exact binomial confidence intervals [1, 2]. Simulations have shown that the BPCP appears to be exact under general independent censoring as well [1]. In this paper, we give further simulations that support this claim.

When designing confidence intervals to be exact for discrete data in general, in order to guarantee coverage over the whole parameter space those confidence intervals are typically conservative for almost all of the parameter space. The BPCP shares this property. Because of this property, for discrete data there is typically not one universally agreed upon choice for the confidence interval. Some statisticians advocate that for discrete data a confidence interval should “on average” cover the true parameter, meaning that for some parameters it will be slightly conservative and for other parameters it will be slightly anti-conservative, but averaged over all possible values of the parameter the coverage is close to the nominal level (see [3, 4]). Other statisticians think that guaranteed coverage is more essential to a confidence interval and covering the true parameter with the correct level only “on average” is not adequate (see [5, 6]). So statisticians in the former camp would think that being anti-conservative at some parameter values is an acceptable trade off in order to get smaller confidence intervals, as long as the coverage is close to nominal on average. Of course, some statisticians might be more flexible, and use one or the other type of confidence interval based on the situation. The BPCP is designed to guarantee coverage, and in this paper, we extend the BPCP to a procedure that is designed to have approximately the nominal level of coverage “on average” over all the parameter space. We call this extension the mid-p BPCP, since it produces the mid-p confidence interval for a binomial parameter when there is no censoring. The mid-p BPCP has shorter width confidence intervals than the BPCP. This paper is not taking the position that the achieving nominal coverage on average is better than guaranteed coverage, but if the former criterion is one's goal, then the mid-p BPCP is a good option.

A mid-p confidence interval (CI) is formed by inverting the series of tests that make inferences using the mid p-value [7]. For a simple null hypothesis, the traditional p-value is defined as the probability of observing equal or more extreme data under the null hypothesis, while the mid p-value is the probability of observing more extreme data plus one half the probability of observing equally extreme data [7, 8]. Without censoring, mid-p confidence intervals have coverage that is sometimes larger and sometimes smaller than the nominal level depending on the true parameter, but the coverage is roughly equal to the nominal level on average over all true parameter values. With censoring, the mid-p version of the BPCP can have conservative coverage if there is heavy censoring.

Here is an outline of the paper. In Section 2, we review exact and mid-p inferences for binomial data, and discuss properties of confidence intervals for discrete data. In Section 3 we review the BPCP for inferences on the survival distribution for a single sample of right censored data. We then show how the original BPCP, which we call the standard BPCP, may be adjusted to give a mid-p version of the BPCP. In Section 4 we simulate the coverage of the both the standard BPCP and the mid-p BPCP in 11 situations, showing how the standard

BPCP maintains coverage while some existing methods can give very poor coverage. We show that the mid-p BPCP has coverage closer to the nominal level. In Section 5 we go through two applications and show how to use the `bpcp` R package to get the CIs calculated in this paper. We end with a discussion section.

2. Exact and mid-p Inferences for Binomial Data

Because the no censoring case reduces to the binomial case, we first review non-asymptotic inferences for the binomial case. Let $X_1, \dots, X_n$ be random variables representing the failure times of the n individuals in our sample. Our interest is $Pr[X_i > t] = S(t) \equiv \theta$ for a specific time of interest, t . Let y be the number of individuals that survive past t , then the associated random variable, Y , has $Y \sim \text{Binomial}(n, \theta)$. The standard $100(1 - \alpha)\%$ exact binomial CI [2] is an exact central CI because it guarantees that the lower error rate (the probability that the true parameter is less than the lower confidence limit) and the upper error rate (the probability that the true parameter is greater than the upper limit) are both less than or equal to $\alpha/2$. We can write these intervals in relationship to the one-sided p-values. Let $p_L(y, \theta_0) = Pr[Y \leq y; \theta_0]$ be the one-sided p-value for testing $H_0: \theta = \theta_0$ versus $H_1: \theta > \theta_0$, and similarly let $p_U(y, \theta_0) = Pr[Y \geq y; \theta_0]$ be the one for testing $H_0: \theta = \theta_0$ versus $H_1: \theta < \theta_0$. Then the $100(1 - \alpha)\%$ exact central interval is defined by the two one-sided $100(1 - \alpha/2)\%$ limits:

$$\begin{aligned} L(y, 1 - \frac{\alpha}{2}) &= \min\{\theta: p_L(y, \theta) \geq \frac{\alpha}{2}\} = \begin{cases} 0 & \text{if } y=0 \\ \theta: p_L(y, \theta) = \frac{\alpha}{2} & \text{if } y>0, \end{cases} \\ &\text{and} \\ U(y, 1 - \frac{\alpha}{2}) &= \max\{\theta: p_U(y, \theta) \geq \frac{\alpha}{2}\} = \begin{cases} 1 & \text{if } y=n \\ \theta: p_U(y, \theta) = \frac{\alpha}{2} & \text{if } y<n. \end{cases} \quad (1) \end{aligned}$$

Thus, if $L(y, 1 - \alpha/2) > \theta_0$ then we can reject the null hypothesis $H_0: \theta = \theta_0$ at a level $\alpha/2$. Similarly, $U(y, 1 - \alpha/2)$ can be used to test the other one-sided hypothesis.

The one-sided mid p-values are

$$p_{Lmid}(y, \theta_0) = Pr[Y > y; \theta_0] + \frac{1}{2}Pr[Y = y; \theta_0] \text{ and } p_{Umid}(y, \theta_0) = Pr[Y < y; \theta_0] + \frac{1}{2}Pr[Y = y; \theta_0].$$

The associated mid-p confidence intervals are calculated by substituting p_{Lmid} for p_L and p_{Umid} for p_U into the equations of expression 1.

In Figure 1 we plot the error rates for every value of θ_0 for both the exact central CIs and the associated mid-p CIs for $n = 30$. We reverse the horizontal axis so that time flows from left to right. First consider the exact central CIs. Notice the saw-toothed patterns in both the one-sided error rates. Only periodically does the error rate get very close to the 0.025 level. So in order to guarantee one-sided coverage at the $1 - \alpha/2$ level for all values of θ , the two-sided $100(1 - \alpha)\%$ CI is conservative for almost all values of θ . If we give up centrality and only require $100(1 - \alpha)\%$ two-sided coverage, we can reduce the confidence interval widths (see Section 3.4).

Now consider the mid-p CIs, the right panels in Figure 1. The pattern is similar to the exact central CIs, except the one-sided error rates are not bound to be less than or equal to 0.025, but fluctuate around 0.025. Because of the zero probability of errors at the extremes, the average one-sided error rates using a uniform prior distribution are not exactly equal to 0.025 but are slightly less than 0.025. Habiger, McCann, and Tebbs [9] studied confidence intervals designed to have the mean coverage equal to the nominal value, but we do not explore that here.

The choice between an exact central CI and a mid-p CI depends on the application and statistical philosophy. The exact central CI is appropriate when one wants to guarantee coverage. This may be important for example in a regulatory setting. Guaranteeing coverage may also be desired as a matter of presentation, since at any time point we can say the survival is significantly larger than the exact central lower limit at the $\alpha/2$ level. The mid-p CI is used when narrower confidence intervals are more important than strict coverage at the nominal level. For times when there are a large number of subjects at risk, there is very little difference between the two methods.

3. Beta Product Confidence Procedure and its Mid-p Version

After reviewing the beta product confidence procedure (BPCP), we introduce a new mid-p version of the BPCP. We assume a continuous survival distribution. We first describe both versions of the BPCP under continuous assessment, but then allow discrete assessment times, which can allow for tied failure times.

3.1. Continuous Assessment

3.1.1. Progressive Type II Censoring—In this section assume the failure times $X_1, \dots, X_n$ are independent and continuous, and they may be unobserved. As in Section 2, let $\theta = S(t) = P\{X_i > t\}$. Using standard Martingale notation, let $Y(t)$ be the number of individuals alive and at risk of failure just prior to time t . Then Y of the previous section is $Y(t+) = \lim_{\epsilon \rightarrow 0} Y(t + \epsilon)$, the number that have survived past t . Let $T_1 < T_2 < \dots < T_k$ be the observed ordered failure times. Consider first the special case of independent progressive Type II censoring, where all censoring occurs on a randomly selected proportion of the survivors immediately following each failure time. Fay *et al.* [1] used the probability integral transformation and uniform order statistics to show that the distribution of $S(T_1)$ is beta with parameters $Y(T_1) = n$ and 1. Here we use a different convention, we define the distribution of $S(T_1)$ as $Beta(n, 1)$, where $S(t) = P\{X_i > t\}$. The new convention avoids issues with the confidence intervals exactly at failure times not enclosing the associated Kaplan-Meier survival estimate. For continuous data, the probability of failure times falling exactly on t are zero, so this is just a definitional convention change, and will not change coverage probabilities. For details see [10]. We write a beta random variable with parameters a and b as $B(a, b)$. Let $q(a, W)$ be the a th quantile of the random variable W . Then an exact $100(1 - \alpha)\%$ confidence interval for $S(t)$ at $t = T_1$ is

$$(q\{\alpha/2, B(n, 1)\}, q\{1 - \alpha/2, B(n, 1)\}).$$

The distribution of $\bar{S}(\bar{T}_j)$ has the same distribution as the product of j independent random variables,

$$\bar{S}(\bar{T}_j) = \frac{\bar{S}(\bar{T}_1)}{\bar{S}(\bar{T}_0)} \frac{\bar{S}(\bar{T}_2)}{\bar{S}(\bar{T}_1)} \cdots \frac{\bar{S}(\bar{T}_j)}{\bar{S}(\bar{T}_{j-1})},$$

where $\bar{S}(\bar{T}_0) = \bar{S}(0) = 1$. The k th term, $\bar{S}(\bar{T}_j)/\bar{S}(\bar{T}_{j-1})$, can be shown to be $B(Y(\bar{T}_j), 1)$ using similar methods as for the derivation of the distribution of $\bar{S}(\bar{T}_1)$. So the distribution of $\bar{S}(\bar{T}_j)$ for $j = 1, 2, \dots, k$ is the same as the distribution of:

$$W_j = \prod_{i=1}^j B\{Y(\bar{T}_i), 1\}.$$

We define the distributions at the extremes $T_0 = 0$ and $T_{k+1} = \infty$, letting W_0 and W_{k+1} be point masses at 1 and 0, respectively. Although we know the distribution of $\bar{S}(\bar{T}_j)$ at the failure times, we want to get confidence intervals for $\bar{S}(t)$, the usual survival distribution evaluated at any fixed t . The solution of the BPCP is to act conservatively. Since the survival distribution is decreasing, if $T_j < t < T_{j+1}$ we look backward and use $W^-(t) = W_j$ to get a conservative upper limit for $\bar{S}(t)$. Similarly, if $T_j < t < T_{j+1}$ we look forward and use $W^+(t) \equiv W_j B\{Y(t+), 1\}$ to get a conservative lower limit for $\bar{S}(t)$, where we use $t+$ since $\bar{S}(t) = \bar{S}(t+)$. With this notation, we define the BPCP $100(1 - \alpha)\%$ CI for $\bar{S}(t)$ as:

$$(q\{\alpha/2, W^+(t)\}, q\{1 - \alpha/2, W^-(t)\}). \quad (2)$$

With no censoring, the BPCP confidence intervals between failure times are the usual exact binomial confidence intervals [1, 2].

3.1.2. General Independent Censoring—Now consider the more general independent censoring situation. We continue to define the BPCP by equation 2. Now the $B\{Y(t+), 1\}$ expression in $W^+(t)$ of equation 2 is motivated by the conditional distribution, $\bar{S}(t | t > T_j)$, being approximated by the first order statistic from a uniform with $Y(t)$ elements (similar to the derivation above), so that $\bar{S}(t | t > T_j)$ uses $Y(t+)$. For general censoring, when there is censoring between the j th and $(j+1)$ th failures, then $Y(t)$ will be a decreasing step function with steps at the censoring times. Thus, in contrast to conventional procedures, the lower confidence limit decreases as fewer individuals are followed, while the upper limit (like the Kaplan-Meier estimator) does not change unless a failure occurs. Table 1 of Section 5 gives the details for a specific application. This changing of the lower limit at censoring times is an important quality of the BPCP and is important for bounding the lower error rate (i.e., guaranteeing that proportion of times the lower limit is greater than the true $\bar{S}(t)$ is less than half the nominal level). Previous confidence intervals have been shown not to bound the lower error rate [1].

3.1.3. Mid-p Version of Beta Product Confidence Procedure—For the mid-p BPCP our intent is to be closer to nominal coverage as opposed to guaranteeing coverage. For $T_j < T < T_{j+1}$, the true distribution for $S(T)$ is approximately between (in a stochastic ordering sense) the distributions associated with $W^-(T)$ and $W^+(T)$. So a natural simple strategy for getting a confidence interval for $S(t)$ when $T_j < t < T_{j+1}$ is to use a distribution associated with an equally likely random draw of either $W^-(t)$ or $W^+(t)$. In other words, we chose a distribution that is “halfway” between the known distributions. Specifically, we define the $100(1 - \alpha)\%$ mid-p BPCP for $S(t)$ as

$$(q\{\alpha/2, W^*(t)\}, q\{1 - \alpha/2, W^*(t)\}), \quad (3)$$

where $W^*(t) = UW^-(t) + (1 - U)W^+(t)$, and U is Bernoulli with mean 0.5. When $t < T_1$ then $W^-(t)$ is a point mass at 1, and since $1 - \alpha/2 > 0.5$, $q\{1 - \alpha/2, W^*(t)\} = 1$. Similarly, when there are no more individuals at risk at t (i.e., $t > T_k$ and $Y(t) = 0$), then $W^+(t)$ is a point mass at 0 and $q\{\alpha/2, W^*(t)\} = 0$. We call this the mid-p BPCP, and we show in the appendix that when there is no censoring the resulting CI is the mid-p confidence interval.

3.2. Discrete Assessments and Tied Failure Times

In order for the theory of the BPCP to work we need to assume continuous failure times, but real data often have tied failure times. To allow for these ties, we assume continuous failure times but additionally assume that we only make assessments at a finite set of times. In other words, the failure times are grouped into intervals bracketed by the assessments. For example, if deaths are observed only to the nearest day, then we could keep the continuity assumption but act as if the assessments of the continuous responses only happen at midnight for each day (i.e., we only know what day an individual died or is censored, not the exact time). Suppose we observe data at time zero and m other assessment times, $0 = g_0, g_1, g_2, \dots, g_m$, which divides positive time into $m + 1$ intervals, where the i th interval is $(g_{i-1}, g_i]$, and $g_{m+1} \equiv \infty$. Let d_i and n_i be the number of deaths within the i th interval and the number at risk at the beginning of the i th interval, respectively. We make the conventional censoring assumption that if deaths and censoring happen in the same interval, that all the censoring happens after all the deaths. Then with these assumptions, we can unambiguously define $W^-(g_i)$ for all i using the following algorithm [10]:

1. Let $W^-(g_0)$ be a point mass at 1.
2. Proceed iteratively, so that at g_i
 - if $d_i = 0$ then $W^-(g_i) = W^-(g_{i-1})$ and
 - if $d_i > 0$ then $W^-(g_i) = W^-(g_{i-1})B(n_i - d_i + 1, d_i)$.

The beta random variable factor comes from the fact that conditional on n_i subjects at risk, the conditional survival distribution of the d_i th failure of the n_i at risk (given no censoring occurs between g_{i-1} and the d_i th failure) is $B(n_i - d_i + 1, d_i)$. Further, $W^+(g_i) = W^-(g_i)B(n_{i+1}, 1)$ for all i , where we define $B(0, 1)$ as a point mass at 0.

The BPCP for grouped data is defined conservatively, so that the $100(1 - \alpha)\%$ BPCP CI for $S(t)$ when $g_{i-1} < t < g_i$ is defined as

$$(q\{\alpha/2, W^+(g_i)\}, q\{1 - \alpha/2, W^-(g_{i-1})\}). \quad (4)$$

We define the mid-p BPCP for grouped data analogously to how it was defined for the non-grouped data. We create a random variable that randomly chooses between the random variable used for the lower BPCP limit and the one used for the upper BPCP limit. Thus, the $100(1 - \alpha)\%$ BPCP mid-p CI for $S(t)$ when $g_{i-1} < t < g_i$ is

$$(q\{\alpha/2, W_g^*(t)\}, q\{1 - \alpha/2, W_g^*(t)\})$$

where $W_g^*(t) = UW^-(g_{i-1}) + (1 - U)W^+(g_i)$, where U is Bernoulli with mean 0.5. Note that as the assessment gets closer and closer to continuous (i.e., $g_{i-1} \rightarrow g_i$), then this definition approaches the one for the mid-p BPCP with continuous assessments.

3.3. Estimating Quantiles of the Random Variables

To estimate $q(a, W)$ we can use Monte Carlo sampling and take the a th empirical quantile. This can be made as precise as needed by increasing the number of Monte Carlo samples at the cost of increasing computer time. A disadvantage is that it is simulation-based, so that different researchers using different software will not always get the same confidence intervals for the same data. Alternatively, for reasonable precision, a method of moments estimator is typically faster and is reproducible but is slightly more approximate.

Following [1], in [1] we proposed a method of moments estimator, where we estimated the quantiles of beta product random variables assuming the beta product is approximately beta with the same first two moments. We showed through simulation that the method of moments estimator worked quite well.

Following [1], let $B(a, 0)$ be a point mass at 1, and $B(0, b)$ be a point mass at 0. Let $\mathbf{a} = [a_1, \dots, a_h]$ and $\mathbf{b} = [b_1, \dots, b_h]$ be vectors with all values > 0 and no j such that $a_j = b_j = 0$. Define the beta product random variable, $BP(\mathbf{a}, \mathbf{b})$, such that

$$BP(\mathbf{a}, \mathbf{b}) \sim \prod_{i=1}^h B(a_i, b_i),$$

where $\sim$ denotes “is distributed as” and the terms in the product are all independent. Then let

$$u = E(BP(\mathbf{a}, \mathbf{b})) = \prod_{i=1}^h \frac{a_i}{a_i + b_i}, \text{ and}$$

$$t = E(BP^2(\mathbf{a}, \mathbf{b})) = \prod_{i=1}^h \frac{a_i(a_i + 1)}{(a_i + b_i)(a_i + b_i + 1)},$$

using method of moments for a beta distribution gives

$$a = \frac{u(u-t)}{t-u^2}$$

and

$$b = \frac{(1-u)(u-t)}{t-u^2}. \quad (5)$$

Now consider the mid-p version. For this we need to calculate quantiles of $W^* = UW^- + (1 - U)W^+$, where W^- and W^+ are beta product random variables and U is Bernoulli with mean 0.5. When W^- is a point mass at 1, then $q(1 - \alpha/2, W^*) = 1$, while if W^+ is a point mass at 0 then $q(\alpha/2, W^*) = 0$. To estimate the quantiles when neither of W^- or W^+ is a point mass is more complicated. First, we use the method of moments to approximate W^- and W^+ as beta random variables. Since the distributions are continuous, there exists some $a_1 + a_2 = a$ such that

$$Q = q(a_1, W^-) = q(a_2, W^+)$$

In other words, $Pr[W^- \leq Q] = a_1$ and $Pr[W^+ \leq Q] = a_2$. So that $Pr[W^* \leq Q] = 0.5a_1 + 0.5a_2 = a/2$. So Q is the $(a/2)$ th quantile for W^* . We use numeric methods (e.g., the `uniroot` function in R) to find Q .

Although theoretically the BPCP (either standard or mid-p version) is non-increasing in time for both the lower and upper limits, the method of moments estimator as described above (and used in [1]) can sometimes lead to increasing limits. This non-monotonicity primarily occurs in the upper mid-p limit. Therefore, in this paper we propose an improvement to the method for both BPCP versions. We simply impose monotonicity after calculating the method of moments estimators as above. This is the default in the current version of the `bpcp` R package (version 1.3.0), and is the method used in all the applications and simulations in this paper. For more details on when non-monotonicity may occur with the original method of moments, see [12].

3.4. Non-central Confidence Intervals

The $100(1 - \alpha)\%$ BPCP CI is designed to guarantee that both the lower and the upper error rates are less than or equal to $\alpha/2$ for all survival functions S and all censoring distributions, say G . Similarly, the $100(1 - \alpha)\%$ mid-p BPCP CI is “centered” so that we expect that on average the lower and upper error rates are about $\alpha/2$ for all S and G . Other extensions of the BPCP and its mid-p version are possible if we remove the centrality goals, so that our new goals are total error rates of less than or equal α (for BPCP CI) or on average about α (for

mid-p BPCP CI). For example, we can expand the class of $100(1 - \alpha)\%$ BPCP CIs to include

$$(q\{\alpha_1, W^+(t)\}, q\{1 - \alpha_2, W^-(t)\})$$

where $\alpha_1 + \alpha_2 = \alpha$. We can similarly expand the mid-p BPCP CIs. Unfortunately, it is not an easy problem to determine the α_1 and α_2 pair such that the confidence interval has good properties for all survival distributions and censoring distributions. To see the difficulty, we return to the simple binomial problem where some optimality results have been solved.

Blaker [13] (p. 792) provides a nice summary of three exact confidence procedures for the binomial problem that bound the two-sided error rate only. Three properties related to these procedures are: (1) smallest length confidence sets, (2) nestedness (i.e., if $1 - \alpha_1 < 1 - \alpha_2$ then the $100(1 - \alpha_1)\%$ confidence set will always be a subset of the $100(1 - \alpha_2)\%$ confidence set), and (3) confidence sets are intervals. It is not possible to achieve all three goals simultaneously. Sterne [14] achieves properties (1) and (2), Casella [15] (expanding on [16]) achieves properties (1) and (3). Blaker's CIs [13] are guaranteed to be subsets of the exact central CIs and have properties (2) and (3) but do not have property (1). Thus, two-sided tests that invert Blaker's CIs are more powerful than the two-sided tests associated with the exact central CIs. All three improvements give up the exact centrality property (the guaranteed bounding of both one-sided error rates at $\alpha/2$).

These non-central exact binomial methods are not easily generalized to the confidence interval on $\theta = S(t)$. First, the sample space for the binomial problem has $n + 1$ elements, while the sample space for the survival problem has $2^{n+1} - 1$ elements in the continuous case. To see this, let $T_i^* = \min(X_i, C_i)$ where C_i is the i th censoring value, and $\delta_i = I(X_i \leq C_i)$. Then all the information needed to calculate either version of the BPCP confidence intervals at t , are the number of T_i^* values less than t , say N_t , their ordering, and the associated δ_i values for those N_t ordered values. Then the total sample space given $N_t = n_t$ is 2^{n_t} , and the total (unconditional) sample space is $\sum_{i=0}^n 2^i = 2^{n+1} - 1$. Second, even if we enumerate the entire sample space, the parameter space is infinite dimensional. In order to get the distribution of the responses, we need the survival distribution, S , and the censoring distribution, G (at least up until t). Because of these difficulties, we will not attempt to study non-central confidence intervals in this paper.

4. Simulations

In [1] and its supplement, we compared the BPCP and 10 other confidence procedures using simulations in two scenarios at 4 time points each. In that paper, all 10 comparators had inflated type I error rates for some time points, while the BPCP bounded the type I error rate in all simulated cases. In this paper, we extend the scope of the simulations dramatically, testing 11 scenarios, and covering the entire range of time points in each. To keep the results manageable, we compared the BPCP and mid-p BPCP to only two of the comparators, chosen because they had comparable coverage to most and were fast to calculate. The first

comparator we call *Modified Lower*, and it is the normal approximation on the log transformed Kaplan-Meier using the Greenwood variance (see [17], p. 16-17 for discussion of different transformations), where the variance for the lower limit is multiplied by factor greater than or equal to 1. The Modified Lower method is calculated using the `survfit` function of the `survival` R package with the `conf.lower="modified"` option [18]. The second comparator method we call *Borkowf-shrink*. This method uses a shrinkage estimator of the Kaplan-Meier, takes a log transformation, and uses the hybrid variance estimator [19]. Both comparators generally had much better simulated coverage than the simple log transformed Greenwood method in [1].

We first simulated the error rates associated with the same set of parameters used in Table 1 of [1], where $n = 30$, the failure is exponential with mean of 10, and the censoring is uniform censoring 0 to 5, with 10,000 replications. We computed error rates associated for each value of $S(t)$ from 1 to 0.607 decreasing by 0.001; note that $S(5) = 0.607$, so that this is the minimum relevant survival proportion.

The results in the first two columns of Figure 2 are analogous to Figure 1, except now with censoring. In Table 1 of [1], we only presented simulated error rates at 4 time points: $S(1) = 0.9$, $S(2) = 0.82$, $S(3) = 0.74$ and $S(4) = 0.67$. We plot the results from [1] as gray points, and we see that they match our current simulation to within rounding and simulation error. Figure 2 illustrates the type I error rates across possible testing times (corresponding to the relevant range of survival times). The Modified Lower procedure produces a confidence interval of (1, 1) when there are no events; this yielded extremely high type I error rates at the beginning of the path. We also considered an alternate version, where these cases were considered non-rejections (not shown), and the results were strikingly similar to Borkowf-shrink. In summary, in this particular exponential setting, the Modified Lower and Borkowf-shrink procedures are generally highly inflated in the lower error rates and highly conservative in upper error rates across testing times. The BPCP is highly conservative for lower error rates and somewhat conservative for upper error rates, typically ranging between 0.01 and 0.015. For this one case, the mid-p BPCP is less conservative than the BPCP for both lower and upper error rates. The mid-p BPCP upper values are generally consistent with 0.025; however, there are small regions of early test times where the simulated error rate is statistically significantly above 0.025 (using an exact one-sided level 0.025 binomial test on 10,000 replications, any simulated error at or above 0.0282 is significantly larger than 0.025), but the inflation is modest, especially when compared with the inflation of the lower rates of the Modified Lower and Borkowf-shrink methods.

In this non-parametric setting, the time scale is irrelevant, all that matters is the relationship of the survival distribution to the censoring distribution. We explore a range of survival to censoring relationships by simulating 11 additional scenarios. Consider the first column in Figure 3, which portrays these 11 relationships. The y-axis is the censoring distribution, $G(t) = P[C_i \leq t]$, where C_i is the censoring time for the i th individual, and the x-axis is the corresponding $S(t)$ (plotted from 1 to 0, which is equivalent to the cumulative distribution $F(t) = 1 - S(t)$ plotted from 0 to 1). So, for example, in the first row, essentially everyone has an event before anyone is censored, whereas in the fifth row, events and censoring are equally likely throughout time. The first nine patterns trend from essentially no censoring to

the other end of the spectrum of essentially no events. The final two scenarios have more of a mixture pattern, where at some periods there are many events happening relative to censoring, and at other periods there is much censoring relative to events. To create these patterns, we simulated data with uniform censoring from zero to one, and failure times from one of 11 different beta distributions. We plot the $S(t)$ by $G(t)$ for the 11 scenarios in the first column of Figures 3 and 4, and The beta parameters are (in order from top to bottom of the first column of Figure 3 or Figure 4): $\alpha = 1, \beta = 100000$; $\alpha = 1, \beta = 50$; $\alpha = 1, \beta = 7$; $\alpha = 1, \beta = 2$; $\alpha = 1, \beta = 1$; $\alpha = 2, \beta = 1$; $\alpha = 7, \beta = 1$; $\alpha = 50, \beta = 1$; $\alpha = 100000, \beta = 1$; $\alpha = 100, \beta = 100$; $\alpha = 0.1, \beta = 0.1$. Again, we conducted 10,000 replications for each path, with error rates computed across the full range of $S(t)$ in increments of $1/300$. That is, we estimated the error rates with 11 parameters sets and 299 time points each, for a total of 3289 simulations, each with 10,000 replications, covering a wide range of survival scenarios.

Figures 3 and 4 are the main results of the paper, depicting the lower error rates and upper error rates respectively. We summarize the results below for $n = 30$:

BPCP: This method appears to guarantee coverage; the estimated error rates were not statistically significantly higher than 0.025 in any of the 3289 simulations. The error rates are quite close to 0.025 when there is little censoring, and become increasingly conservative with more censoring. This conservatism is more excessive for the lower error rates than the upper.

Mid-P BPCP: The results mirror those of BPCP closely; however the rates are consistently closer to the nominal 0.025. The error rates are sometimes inflated where there has been relatively little censoring at the test time, but the inflation is modest and vacillating.

Modified Lower and Borkowf Shrink: Both of these methods have highly inflated lower error rates, and conversely are very conservative with respect to upper error rates.

The simulation for the 11 scenarios were repeated for sample sizes of $n = 10$, $n = 100$ and $n = 1000$. The patterns are similar except the simulated error rates appear closer to the nominal rates as the sample size increases. The complete graphical results are given in the supplementary material.

Because all four methods are asymptotically accurate, these simulations provide important comparisons for how fast the type I error rates approach the nominal. Here we summarize all the simulation results ($n = 10, 30, 100$, and 1000). Only BPCP essentially guarantees coverage for all situations evaluated. (There were 2 out of 3289 with $n = 10$ and 1 out of 3289 with $n = 1000$ with error rates significantly above 0.025 at the two-sided 0.05 level, but if all rates equalled the nominal 0.025 we would expect more than that by chance.) The BPCP is conservative, but often the conservatism is more modest than in the parameters presented in the original simulations of [1]. The mid-p version of the BPCP gives closer to nominal levels than the standard BPCP, at the expense of a few areas of modest inflation with low censoring and small sample size. In fact, the mid-p BPCP simulated error rates are often close to the nominal 0.025 for n as large as 30, especially for the upper error rates. Both the Modified Lower and the Borkowf-shrink can have considerable type I error rate inflation for the lower error rate with smaller sample sizes across different scenarios. They

even have substantial inflation in some extreme censoring situations when $n = 1000$. For the upper error rates, both the Modified Lower and the Borkowf-shrink are more conservative than the already generally conservative BPCP methods, with the differences in rates between all the methods shrinking with increasing sample size. Although we have focused on comparing the one-sided error rates to the nominal 0.025, note that since the lower error rates for the Modified Lower and the Borkowf-shrink can be substantially larger than 0.05 the overall coverage is necessarily less than the nominal 95% in those situations. All-in-all, even with sample sizes as large as $n = 1000$ the BPCP methods, especially the mid-p version, have a clear advantage over the other methods in terms of type I error rates.

5. Application and Software

The `bpcp` function in the `bpcp` R package (in version 1.3.0 or later) can calculate both the standard BPCP and the mid-p BPCP and is available on CRAN (<http://cran.us.r-project.org/>). In Figure 5 we plot the Kaplan-Meier survival curve, the BPCP, and the mid-p BPCP for time to event from two studies. Both versions of the BPCP use the method of moments implementation with enforced monotonicity.

The left panel of Figure 5 is from a small study ($n=34$) of severe systemic sclerosis with 12 deaths observed during the study [20]. Systemic sclerosis is an autoimmune disease. For this pilot study all subjects received a new high-dose immunosuppressive therapy, and the inclusion criteria was designed to only include severe patients that had a mortality risk of approximately 50% at 5 years with conventional treatment.

In Table 1, we illustrate the BPCP procedure process for the first 17 time intervals of the sclerosis study. For example, for the interval that starts at the time of the first event, $W^*(t) = B(33, 1)B(34, 1)$; using equation 5, this product corresponds to $B(33, 2)$. The 0.025th quantile from a beta distribution with parameters 33 and 2 is 0.847, which is the lower confidence interval for the BPCP. The estimated survival at 5 years is 63.6% with 95% confidence intervals of (41.1%, 80.9%) by BPCP and (43.2%, 79.6%) by mid-p BPCP. Generally, competing methods have lower and upper limits that are too high (see simulation section of this paper, and the simulations and Figure 1 of [1]). For example, the 95% limits at 5 years are (47.9%, 84.5%) for the Modified Lower method (normal approximation based on log transformed Kaplan-Meier survival with the Greenwood formula for the variance, modified for the lower limit [18]) and (46.3%, 86.4%) for the Borkowf-shrink method (Borkowf estimator based on the log transformation using a Kaplan-Meier estimator shrunk towards 0.5 [19]).

We can use the confidence intervals on the survival function to get confidence intervals on quantiles using the quantile function which accepts results from the `bpcp` function [1]. For example, the BPCP gives a 95% confidence interval for the time t when $S(t) = 0.80$ as (0.16, 6.35) and the mid-p BPCP gives (0.16, 4.93). For the median time of death the upper limit is infinity for both methods, giving 95% confidence intervals of (4.14, ∞) for BPCP and (4.22, ∞) for the mid-p BPCP.

Although the methods of this paper focus on small studies, to see how they apply to a large study, we consider a large study ($n=4028$) of the National Wilms Tumor Study Group [21], using data in the `survival` R package [22]. Wilms tumor is a childhood kidney disease. We use this data only to illustrate unadjusted Kaplan-Meier curves and confidence intervals on time to relapse for a large sample, and ignore many of the important issues of study design and predictive covariates needed for a proper analysis of these data (see [21]). The 95% confidence intervals and Kaplan-Meier estimates are given on the right panel of Figure 5. As we expect, the difference between the mid-p and standard BPCP is imperceptible at the beginning when there are large numbers at risk. Due to extensive censoring the lower limit for the Wilms Tumor data decreases drastically at the end, since the at-risk sample size decreases at the end (at 10 years there are 936 subjects at risk, but by 15 years there are only 128 subjects still at risk). The Greenwood lower limit does not have that decrease at the end (not shown here, but see Figure 3 of [1]).

6. Discussion

In this paper, we have revisited the beta product confidence procedure [1] for a survival distribution, which was designed to guarantee coverage regardless of sample size or censoring distribution, as long as the censoring is independent of survival. We have performed extensive simulations that all show that the BPCP guarantees coverage. Additionally, we have developed a mid-p version of the BPCP. The mid-p BPCP reduces to a mid-p confidence interval for the binomial when there is no censoring. As expected, simulations indicate the mid-p BPCP has coverage closer to the nominal level on average than the BPCP, but it does not guarantee coverage in all situations.

In a previous paper [1], we had shown through simulation that the best previously developed methods do not always provide guaranteed coverage. In this paper, we consider a wide range of survival settings by conducting simulations at 299 time points across 11 censoring (G) by failure (F) paths. There are infinite numbers of paths, and we evaluated only 11 of them. Nevertheless, there appears to be some clear patterns across the paths we studied. In general, the simulations show that the two BPCPs have better coverage properties than two competing normal theory ones. The two normal theory methods tend to have higher limits than needed for both the upper and lower case (see Figures 2, 3, and 4). Further, we have shown that both BPCPs can create confidence intervals for the median and other quantiles.

In summary, we recommend the standard BPCP when guaranteed coverage is desired, and the mid-p BPCP when coverage that is close to the nominal level is desired.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank two anonymous reviewers for their close reading of the paper which stimulated several important improvements to the original draft, including the graphical representation of the simulations and the new convention for defining the BPCP confidence interval exactly at failure times. We thank Terry Therneau for suggesting that a

less conservative version of the BPCP would be useful. The simulations used the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (<http://biowulf.nih.gov>).

A. Equivalence of Mid-p BPCP with No Censoring to mid-p Confidence Interval for Binomial

We want to show that the mid-p BPCP CI at $T_0, T_1, \dots, T_n$ is equal to the mid-p CI for Y survivals out of n , for $Y = n, n-1, \dots, 0$. The mid-p BPCP $100(1-\alpha)\%$ CI at T_j is $(q(\alpha/2, W^*(T_j)), q(1-\alpha/2, W^*(T_j)))$, where $W^*(T_j) = UW_j + (1-U)W_{j+1}$ and as before $W_0 = 1$ and $W_{n+1} = 0$, and

$$W_j = \prod_{i=1}^j B\{Y(T_i), 1\} = \prod_{h=n}^{n-j+1} B\{h, 1\} = B(n-j+1, j),$$

where the last step uses properties of products of beta distributions (see e.g., [23], p. 158 or [1], equation 2.1). In other words, as long as $W^*(T_j)$ is continuous (i.e., not T_0 and T_n), we can define the lower (θ_L) and upper (θ_U) limits of the $100(1-\alpha)$ mid-p BPCP CI for $\mathcal{S}(T_j)$ as the solutions to

$$\Pr[W^*(T_j) \leq \theta_L] = \frac{\alpha}{2}, \quad (6)$$

$$\Pr[W^*(T_j) \leq \theta_U] = 1 - \frac{\alpha}{2}. \quad (7)$$

Then,

$$\Pr[W^*(T_j) \leq \theta] = \Pr[UB(n-j+1, j) + (1-U)B(n-j, j+1) \leq \theta] = \frac{1}{2} \Pr[B(n-j+1, j) \leq \theta] = \frac{1}{2} \Pr[B(n-j, j+1) \leq \theta].$$

(8)

Using the relationship between the beta and binomial distributions (see e.g., [23], p. 82), we have

$$\Pr[B(n-j, j+1) \leq \theta] = \Pr[Y \geq n-j; n, \theta],$$

where $Y \sim \text{Binomial}(n, \theta)$. Let $Y = n-j$ be the observed number of survivors at T_j and equation 8 becomes,

$$\begin{aligned}
& \Pr[W^*(t) \leq \theta] \\
&= \frac{1}{2} \Pr[Y \geq y \\
&\quad + 1; n, \theta] \\
&\quad + \frac{1}{2} \Pr[Y \geq y; n, \theta] \\
&= \Pr[Y > y; n, \theta] \\
&\quad + \frac{1}{2} \Pr[Y \\
&\quad = y; n, \theta] \\
&= p_{Lmid}(y, \theta) \\
&= 1 - \Pr[Y < y; n, \theta] \\
&\quad - \frac{1}{2} \Pr[Y \\
&\quad = y; n, \theta] \\
&= 1 - p_{Umid}(y, \theta)
\end{aligned}$$

Solving $p_{Lmid}(y, \theta) = \alpha/2$ for θ gives θ_L of equation 6, and solving $p_{Umid}(y, \theta) = \alpha/2$ gives θ_U of equation 7.

For T_0 then $W^-(T_0)$ is a point mass at 1 and $W^+(T_0) = B(n, 1)$, and $Y = n$. Now $W^*(T_0)$ is not fully continuous so we need to be careful about the definition of the quantile function. For general W ,

$$q(a, W) = \inf\{w: \Pr[W \leq w] \geq a\}.$$

The lower limit is

$$\begin{aligned}
& q\{\alpha/2, W^*(T_0)\} \\
&= \inf\{\theta: \Pr[W^*(T_0) \leq \theta] \geq \alpha/2\} \\
&= \inf\left\{\theta: \frac{1}{2}\Pr[1 \leq \theta] \right. \\
&\quad \left. + \frac{1}{2}\Pr[B(n, 1) \leq \theta] \geq \alpha/2\right\} = \inf\left\{\theta: \frac{1}{2}\Pr[B(n, 1) \leq \theta] \geq \alpha/2\right\} \\
&= \left\{\theta: \frac{1}{2}\Pr[B(n, 1) \leq \theta] \right. \\
&\quad \left. = \alpha/2\right\} \\
&= \left\{\theta: \frac{1}{2}\Pr[Y \right. \\
&\quad \left. = n; n, \theta] \right. \\
&\quad \left. = \alpha/2\right\} \\
&= \{\theta: p_{L\text{mid}}(y, \theta) = \alpha/2\}.
\end{aligned}$$

The upper limit is

$$q\{1 - \alpha/2, W^*(T_0)\} = \inf\left\{\theta: \frac{1}{2}\Pr[1 \leq \theta] + \frac{1}{2}\Pr[B(n, 1) \leq \theta] \geq 1 - \alpha/2\right\} = 1.$$

Finally, when T_n then $Y = 0$, $W^+(T_n)$ is a point mass at 0, and $W^-(T_n) = B(1, n)$. So the lower limit is

$$q\{\alpha/2, W^*(T_n)\} = \inf\left\{\theta: \frac{1}{2}\Pr[0 \leq \theta] + \frac{1}{2}\Pr[B(1, n) \leq \theta] \geq \alpha/2\right\} = 0.$$

The upper limit is

$$\begin{aligned}
& q\{1 - \alpha/2, W^*(T_n)\} \\
&= \inf \left\{ \theta: \frac{1}{2} \Pr[\theta \leq \theta] \right. \\
&\quad \left. + \frac{1}{2} \Pr[B(1, n) \leq \theta] \geq 1 - \alpha/2 \right\} \\
&= \left\{ \theta: \frac{1}{2} \right. \\
&\quad \left. + \frac{1}{2} \Pr[B(1, n) \leq \theta] = 1 - \alpha/2 \right\} = \left\{ \theta: \frac{1}{2} + \frac{1}{2} (1 - \Pr[B(1, n) > \theta]) = 1 - \alpha/2 \right\} \\
&= \left\{ \theta: \frac{1}{2} \Pr[B(1, n) > \theta] \right. \\
&\quad \left. = \alpha/2 \right\} \\
&= \left\{ \theta: \frac{1}{2} \Pr[Y \right. \\
&\quad \left. = 0; n, \theta] \right. \\
&\quad \left. = \alpha/2 \right\} \\
&= \{\theta: p_{U_{\text{mid}}}(y, \theta) = \alpha/2\}.
\end{aligned}$$

References

1. Fay MP, Brittain EH, Proschan MA. Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics*. 2013; 14(4):723–736. [PubMed: 23632624]
2. Clopper C, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934; 26:404–413.
3. Agresti A, Coull BA. Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*. 1998; 52(2):119–126.
4. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion (with discussion). *Statistical science*. 2001; 16(2):101–133.
5. Casella G. [interval estimation for a binomial proportion]: Comment. *Statistical Science*. 2001; 16(2):120–122.
6. Corcoran C, Mehta C. [interval estimation for a binomial proportion]: Comment. *Statistical Science*. 2001; 16(2):122–124.
7. Hirji, K. *Exact Analysis of Discrete Data*. Chapman and Hall/CRC; New York: 2006.
8. Lancaster H. Significance tests in discrete distributions. *Journal of the American Statistical Association*. 1961; 56(294):223–234.
9. Habiger JD, McCann MH, Tebbs JM. On optimal confidence sets for parameters in discrete distributions. *Statistics & Probability Letters*. 2013; 83(1):297–303.
10. Fay, MP.; Brittain, EH. Beta product confidence procedure confidence intervals and discrete data. 2016. <http://cran.r-project.org/web/packages/bpcp/pdf> document in bpcp R package, Version 1.3.0 or later

11. Fan D. The distribution of the product of independent beta variables. *Communications in Statistics-Theory and Methods*. 1991; 20(12):4043–4052.
12. Fay, MP.; Brittain, EH. Notes on method of moments and monte carlo methods for calculating beta product confidence procedure and its mid-p version. 2015. <http://cran.r-project.org/web/packages/bpcp/pdf> document in bpcp R package, Version 1.2.5 or later
13. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*. 2000; 28(4):783–798.
14. Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika*. 1954:275–278.
15. Casella G. Refining binomial confidence intervals. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*. 1986:113–129.
16. Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association*. 1983; 78(381):108–116.
17. Therneau, T.; Grambsch, P. *Modeling Survival Data: Extending the Cox Model*. Springer; New York: 2000.
18. Therneau T. A Package for Survival Analysis in S. 2012 R package version 2.36-12.
19. Borkowf CB. A simple hybrid variance estimator for the Kaplan-Meier survival function. *Statistics in Medicine*. 2005; 24(6):827–851. [PubMed: 15558837]
20. Nash R, McSweeney P, Crofford L, Abidi M, Chen C, Godwin J, Gooley T, Holmberg L, Henstorf G, LeMaistre C, et al. High-dose immunosuppressive therapy and autologous hematopoietic cell transplantation for severe systemic sclerosis: long-term follow-up of the us multicenter pilot study. *Blood*. 2007; 110(4):1388. [PubMed: 17452515]
21. Breslow N, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 1999; 48(4):457–468.
22. Therneau T. A Package for Survival Analysis in S. 2015 R package version 2.38.
23. Casella, G.; Berger, RL. *Statistical Inference*. second. Duxbury Press; 2002.

Figure 1.

Error rates for 95% confidence intervals for the binomial case with $n = 30$ and different values of θ_0 . Left plots are for the exact central CIs [2], and right plots are for the mid-p CIs [7]. We reverse the horizontal axis so that time flows from left to right.

Figure 2.

Simulated error rates by survival function for 95% confidence intervals using exponential failure with mean 10 and uniform censoring from 0 to 5 (note $S(5) = 0.607$) with $n = 30$ and 10,000 replications. The columns are the 4 methods: BPCP [1], mid-p BPCP, `conf.lower="modified"` for `survfit` [18], Borkowf's log transformation with the shrinkage of the Kaplan-Meier estimator [19]. Gray dots are the associated simulation results from Table 1 of [1].

Figure 3.

Simulated one-sided type I error rates (lower limit greater than the true value) for 4 confidence procedures with nominal one-sided error rate 0.025, with $n = 30$. There are 11 scenarios (described in text), with each scenario represented a row of panels. The leftmost panel on each row describes the scenario by depicting the 299 points that were simulated on a $S(t)$ by $G(t)$ plot, with $S(t_i) = (300 - i)/300$ for $i = 1, \dots, 299$. Simulations based on 10,000 replications. The other 4 panels on each row plot the simulated type I error rate by $S(t_i)$. The nominal rate is 0.025. For clarity the area under the error lines are shaded, so consequently fully shaded panels denote all error rates greater than 0.10, and fully white panels denote all error rates of 0. The vertical axis for each row are the same, but to avoid clutter we only give axis labels on even numbered rows.

Figure 4.

Simulated one-sided type I error rates (upper limit greater than the true value) for 4 confidence procedures with nominal one-sided error rate 0.025, with $n = 30$. There are 11 scenarios (described in text), with each scenario represented a row of panels. The leftmost panel on each row describes the scenario by depicting the 299 points that were simulated on a $F(t)$ by $G(t)$ plot, with $S(t_i) = (300 - i)/300$ for $i = 1, \dots, 299$. Simulations based on 10,000 replications. The other 4 panels on each row plot the simulated type I error rate by $S(t_i)$. The nominal rate is 0.025. For clarity the area under the error lines are shaded so consequently fully white panels denote all error rates of 0. In the BPCP column, there are no simulated type I error rates significantly (at the one-sided 0.025 level) greater than 0.025.

Figure 5. Kaplan-Meier estimator of survival (thick black), 95% BPCP (thick gray), and 95% mid-p BPCP (thin black) for two data sets. Left panel is data from a severe systemic sclerosis study [20] and has 34 patients with 12 observed deaths. Right panel shows data from the National Wilms Tumor Study Group [21] and has 4028 subjects with 571 observed relapsed, so that 85.8% of them are censored.

Table 1

Kaplan-Meier survival estimates, 95% BPCP and mid-p BPCP (both method of moments implementations) from the first 17 time intervals (excluding single point intervals) from the 34 subjects from [20]. Time intervals are in years, with **boldface** denoting death and *italics* denoting censoring. There are no ties. $B(a:b, 1) = \prod_{n=a}^b B(n, 1)$.

time interval	$W^-(t)$	$W^+(t)$	Kaplan-Meier	95% BPCP	95% mid-p BPCP
[0, 0.04)	1	$B(34, 1)$	1.00	(0.897, 1.000)	(0.916, 1.000)
[0.04 , 0.06)	$B(34, 1)$	$B(33 : 34, 1)$	0.97	(0.847, 0.999)	(0.863, 0.999)
[0.06 , 0.16)	$B(33 : 34, 1)$	$B(32 : 34, 1)$	0.94	(0.803, 0.993)	(0.819, 0.990)
[0.16 , 0.18)	$B(32 : 34, 1)$	$B(31 : 34, 1)$	0.91	(0.763, 0.981)	(0.778, 0.977)
[0.18 , 0.22)	$B(31 : 34, 1)$	$B(30 : 34, 1)$	0.88	(0.725, 0.967)	(0.740, 0.961)
[0.22 , 0.34)	$B(30 : 34, 1)$	$B(29 : 34, 1)$	0.85	(0.689, 0.950)	(0.704, 0.944)
[0.34 , 0.52)	$B(29 : 34, 1)$	$B(28 : 34, 1)$	0.82	(0.655, 0.932)	(0.669, 0.925)
[0.52 , 1.05)	$B(28 : 34, 1)$	$B(27 : 34, 1)$	0.79	(0.621, 0.913)	(0.635, 0.905)
[1.05 , <i>2.01</i>)	$B(27 : 34, 1)$	$B(26, 1)B(27 : 34, 1)$	0.76	(0.588, 0.893)	(0.602, 0.884)
[<i>2.01</i> , <i>2.44</i>)	$B(27 : 34, 1)$	$B(25, 1)B(27 : 34, 1)$	0.76	(0.587, 0.893)	(0.600, 0.884)
[<i>2.44</i> , <i>3.01</i>)	$B(27 : 34, 1)$	$B(24, 1)B(27 : 34, 1)$	0.76	(0.585, 0.893)	(0.599, 0.884)
[<i>3.01</i> , <i>3.58</i>)	$B(27 : 34, 1)$	$B(23, 1)B(27 : 34, 1)$	0.76	(0.583, 0.893)	(0.598, 0.884)
[<i>3.58</i> , 3.68)	$B(27 : 34, 1)$	$B(22, 1)B(27 : 34, 1)$	0.76	(0.580, 0.893)	(0.596, 0.884)
[3.68 , <i>3.71</i>)	$B(22, 1)B(27 : 34, 1)$	$B(21, 1)B(22, 1)B(27 : 34, 1)$	0.73	(0.541, 0.869)	(0.557, 0.859)
[<i>3.71</i> , <i>3.74</i>)	$B(22, 1)B(27 : 34, 1)$	$B(20, 1)B(22, 1)B(27 : 34, 1)$	0.73	(0.539, 0.869)	(0.555, 0.859)
[<i>3.74</i> , <i>4.08</i>)	$B(22, 1)B(27 : 34, 1)$	$B(19, 1)B(22, 1)B(27 : 34, 1)$	0.73	(0.536, 0.869)	(0.553, 0.859)
[4.08 , 4.14)	$B(22, 1)B(27 : 34, 1)$	$B(18, 1)B(22, 1)B(27 : 34, 1)$	0.73	(0.533, 0.869)	(0.551, 0.859)