Published in final edited form as:
Stat Med. 2012 December 10; 31(28): 3760–3772. doi:10.1002/sim.5447.

Weighted Logrank Tests for Interval Censored Data when Assessment Times Depend on Treatment

Michael P. Fay[*] and
National Institute of Allergy and Infectious Diseases, 6700-B Rockledge Dr, MSC 7630, Bethesda, MD 20892-7630, USA

Joanna H. Shih
National Cancer Institute, 6130 Executive Boulevard, Bethesda, Maryland 20892-7434, USA

Abstract

We consider weighted logrank tests for interval censored data when assessment times may depend on treatment, and for each individual we only use the two assessment times that bracket the event of interest. It is known that treating finite right endpoints as observed events can substantially inflate the type I error rate under assessment-treatment dependence (ATD), but the validity of several other implementations of weighted logrank tests (score tests, permutation tests, multiple imputation tests) has not been studied in this situation. With a bounded number of unique assessment times, the score test under the grouped continuous model retains the type I error rate asymptotically under ATD; however, although the approximate permutation test based on the permutation central limit theorem is not asymptotically valid under every ATD scenario, we show through simulation that in many ATD scenarios it retains the type I error rate better than the score test. We show a case where the approximate permutation test retains the type I error rate when the exact permutation test does not. We study and modify the multiple imputation logrank tests of Huang, Lee and Yu (2008, Statistics in Medicine, 27: 3217–3226), showing that the distribution of the rank-like scores asymptotically does not depend on the assessment times. We show through simulations that our modifications of the multiple imputation logrank tests retain the type I error rate in all cases studied, even with ATD and a small number of individuals in each treatment group. Simulations were performed using the interval R package. US Government work, in the Public Domain

Keywords

interval censoring; multiple imputation; rank test; permutation test; survival analysis; within cluster resampling

1. Introduction

With interval censored responses, the event is not observed exactly and only known to be within some interval. As an example, consider the bladder cancer trial conducted by the

[*]Correspondence to: National Institute of Allergy and Infectious Diseases, 6700-B Rockledge Dr, MSC 7630, Bethesda, MD 20892-7630, USA.

Veterans Administration Cooperative Urological Research Group [1, 2]. Patients entered the study with superficial bladder cancers, had them removed, were randomized to different treatment groups, and followed for tumor recurrence. The time of first tumor recurrence is not known exactly, but only known to be between the last recurrence-free clinic visit and the first clinic visit when a recurrence is seen. In this paper we focus on the placebo and thiotepa arms of the study, and note that there are more assessments per subject in the thiotepa arm (mean=14.5 visits) than the placebo arm (mean=9.7 visits), with the distributions of visits significantly different by permutation t-test (p=0.01). We are interested in tests that allow different distributions for assessment times in the different treatment groups, which we call assessment treatment dependence (ATD). In general, the differences in assessment between treatment groups will not be planned, and could be due to minor unplanned adverse events which cause one treatment arm to return for clinic assessments more often than another treatment arm. Freidlin, et al [3] discuss other causes of ATD in relation to clinical trials for cancer therapy with progression-free survival as the endpoint.

Although more rare, one could have differences in treatment assessment by design. Consider the MTN-018 study currently being designed by the Microbicide Trials Network sponsored by the National Institutes of Health. The MTN-018 study is a planned follow-up to MTN-003, the VOICE (Vaginal and Oral Interventions to Control the Epidemic) study, which is studying the safety and effectiveness of several experimental prophylaxes designed to prevent HIV infection. If there appears to be an effective and safe prophylaxis from the VOICE study, the MTN-018 study will randomize subjects to either monthly or quarterly assessment schedules for the safety of the prophylaxis. More frequent monitoring might reduce the severe adverse events if nascent adverse events can be identified and eliminated before becoming severe. In this case we have obvious treatment assessment differences, and the methods of this paper are applicable when the endpoint is the time to first serious adverse event.

For the interval censored responses, if the distribution of assessment times is independent of both event time and treatment assignment, then there are several available valid or approximately valid methods for performing weighted logrank tests (see e.g., [4, 5, 6, 7]). In this paper, we always assume the distribution of the assessment times is independent of event time *given* treatment, but allow ATD, i.e., that each treatment group may have a different distribution of assessment times. Under ATD, simple adjustments to the usual right censored weighted logrank test have been shown to inflate the type I error rate. For example, Law and Brookmeyer [8] shows that the midpoint imputation method (i.e., replacing the finite intervals with the midpoints then performing the usual right-censored logrank test) can substantially inflate the type I error rate under ATD. Others [3, 9] show similar type I error rate inflation with right endpoint imputation (replacing finite intervals with the right endpoint), which is still commonly used [9]. Freidlin, et al [3] suggest the simple fix of only using assessment times that are scheduled to be at the same time in both treatment groups (essentially ignoring all other assessments). This method of Freidlin, et al [3] can only be applied when identical scheduled assessments can be guaranteed in both groups, and because of this limitation that method will not be discussed further.

We classify the weighted logrank tests for interval censored responses that allow general assessment distributions into 4 broad categories: those derived using the marginal likelihood of the ranks [4], those derived using the grouped continuous model with inferences using a score test (see e.g., [5, 6, 10]), those derived as permutation tests on rank scores based on the nonparametric maximum likelihood estimate (NPMLE) of the distribution of all the data under the null [10, 11], and those derived using that same NPMLE and multiple imputation within subject (see [7]). The weighted logrank tests that use the marginal likelihood of the ranks [4] do not have a theoretical problem when there are a large number of assessment times, but are computationally difficult under that situation and will not be discussed further. Thus, the focus of this paper will be on the latter three categories of weighted logrank tests of which little is known under ATD.

Some work has been done on testing under interval censoring with ATD. There are conditions where the score test is asymptotically valid under ATD, which are discussed in Section 2. Others have studied ATD in some special cases. Sun [12] discusses an asymptotic test for the case of current status data (i.e., each subject has only one assessment time), and allows the assessment times to depend on the treatment assuming that the assessment times follow a proportional hazards model. Zhu, et al [13] discuss an asymptotic test under ATD and case II interval censoring, where there are 2 observed assessment times which are independent of the event time. Case II interval censoring is rare unless there are only 2 assessments per subject. Note that the ATD we are addressing in this paper is non-informative assessment because the event time is independent of the assessment given treatment; for informative assessment other methods are needed [14, 15, 16].

In Section 2 we give notation and review the conditions under which score tests under the grouped continuous model are asymptotically valid. In Section 3 we study permutation tests, showing why in general these tests perform much better than naive midpoint imputation even under ATD. We study an example in which the exact permutation test does not control the type I error rate if treatment is related to assessment times. In Section 4 we review the multiple imputation method of Huang, Lee, and Yu [7] under ATD. We show that asymptotically the rank-like scores from that method do not depend on the assessment times. We modify the method of Huang, Lee and Yu [7] to produce tests which retain type I error rate for all cases studied. In Section 5 we compare all methods by simulation, showing that our multiple imputation modifications retain type I error rate in all cases studied, and that the approximate permutation test is approximately valid. In Section 6 we applied the different tests to the bladder cancer data.

2. Score Tests under the Grouped Continuous Model

Previous work on assessment treatment dependence has focused on the score test of Finkelstein [5] and its generalizations [10]. Fay [10] discussed that these methods are asymptotically valid under ATD as long as the number of unique observed assessment times in the study, say $M - 1$ (so that those assessments partition the event time into M intervals), does not get large as the sample size increases. Sun and Chen [9] confirmed this validity by simulation for large samples and small M. Here we review the score method, and introduce notation and our ATD assumptions.

For the ith individual, let x_i be the unobserved event time, z_i be a covariate vector such as a treatment indicator, and $\mathbf{a}_i = [a_{i1}, \ldots, a_{i,k_i}]$ be a k_i-dimensional vector of assessment times, which partition the sample space into $k_i + 1$ intervals, $(a_{i,j-1}, a_{i,j}]$for $j = 1, \ldots, k_i + 1$ with $a_{i0} \equiv 0$ and $a_{i,k_i+1} \equiv \infty$. In this paper, we do not use the complete vector $\mathbf{a}_i$ but only use the last negative assessment and the first positive one, and we denote the resulting interval as $y_i = (\ell_i, r_i]$. Let the associated random variables be X_i, $\mathbf{A}_i$, K_i, and Y_i. We assume that under the null hypothesis K_i and $\mathbf{A}_i$ may be related to z_i, but given z_i, both K_i and $\mathbf{A}_i$ are independent of the event time, X_i.

Under the grouped continuous model (see e.g., [17]), we assume

$$\phi(X_i) = z_i'\beta + \varepsilon_i \quad (1)$$

where $\phi(\cdot)$ is an unknown non-decreasing transformation function and $\varepsilon_i \sim F$, where F is a known continuous distribution. Let $\phi(t) = F^{-1}\{H(t)\}$, where H is an arbitrary distribution function. Under the assumptions stated above we can write the grouped continuous model likelihood as

$$L(\beta, H(\cdot)) = \prod_{i=1}^{n} (F\left[F^{-1}\{H(r_i)\} - z_i\beta\right] - F\left[F^{-1}\{H(\ell_i)\} - z_i\beta\right])$$

Under the null that $\beta = 0$ the likelihood reduces to the usual likelihood for a single distribution from interval censored data, and the maximizer of H under the null is $\hat{H}$, the nonparametric maximum likelihood estimate (NPMLE) of the distribution ignoring the covariates z_i. Note that the NPMLE is really a class of distributions, but all members of this class have the same values at the observed assessment times, $t_1 < \cdots < t_{M-1}$, so following standard nomenclature we will call this class "the" NPMLE (see e.g., [18]). So $\hat{H}$ can be described by a vector of nuisance parameters.

Then the score statistic is

$$U = \left[\frac{\partial \log(L(\beta, \hat{H}))}{\partial \beta}\right]_{\beta=0} = \sum_{i=1}^{n} z_i c_i \quad (2)$$

where

$$c_i = \frac{-f\left[F^{-1}\left\{\hat{H}(r_i)\right\}\right] + f\left[F^{-1}\left\{\hat{H}(\ell_i)\right\}\right]}{\hat{H}(r_i) - \hat{H}(\ell_i)},$$

f is the density function associated with F, and since $\lim_{a\to 1} f(F^{-1}(a)) = 0$ and $\lim_{a\to 0} f(F^{-1}(a)) = 0$, we define $f[F^{-1}\{\hat{H}(\infty)\}] = 0$ and $f[F^{-1}\{\hat{H}(0)\}] = 0$.

The usual likelihood assumptions for the score test do not hold under continuous or nearly continuous assessment times when M increases with sample size, often causing the nuisance parameters that make up the NPMLE to approach the boundary of the parameter space. Fay [10] proposed an *ad hoc* modification to the score test when this boundary problem arises, but that *ad hoc* approach has not yet been studied by simulation even under assessment treatment independence.

The scores c_i act like ranking functions, especially when F is the logistic distribution, when this becomes a Wilcoxon-type test [10], and

$$c_i = 1 - \hat{H}(\ell_i) - \hat{H}(r_i). \quad (3)$$

When F is the extreme minimum value distribution this becomes a logrank-type test [5, 10], and

$$c_i = \frac{\hat{S}(\ell_i)\log(\hat{S}(\ell_i)) - \hat{S}(r_i)\log(\hat{S}(r_i))}{\hat{S}(\ell_i) - \hat{S}(r_i)}, \quad (4)$$

where $\hat{S}(t) = 1 - \hat{H}(t)$ and we define $0 \log(0) = 0$. Sun [6] gave a slightly different version of the logrank test (see [17, 19]).

3. Permutation Tests

Since the scores, c_i, are like ranking functions, it is natural to consider permutation methods for inferences (see [11, 19]). When the assessment times are independent of treatment, then standard permutation theory shows these type of weighted logrank tests are valid, and exact tests can be performed that way. The problem is that the theory for the permutation method breaks down when the assessment times are related to treatment. In the following, we provide heuristics and simulations to motivate that the permutation method often retains type I error rate reasonably well even under assessment-treatment dependence; however, the permutation method can inflate the type I error rate in extreme situations, and we explore one such situation in detail.

3.1. Rewriting Rank Scores

Now we rewrite the scores, c_i, to emphasize that they are a function of the empirical distribution. Let $\mathbf{y}_n = [y_1, \ldots, y_n]$, and let $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = \infty$ be all the unique observed assessment times plus 0 and ∞. Let $\gamma(\mathbf{y}_n) = \hat{\gamma} = \{\hat{\gamma}_1, \ldots, \hat{\gamma}_M\}$ be the induced partition of the event time space, so that $\hat{\gamma}_j = (t_{j-1}, t_j]$. We rewrite the null distribution from the cumulative distribution form, $H(t) = Pr[X \le t]$, into the set function form, $P(\gamma) = Pr[X \in \gamma]$, with the associated NPMLE denoted $\hat{H}$ or $\hat{P}$. For an interval $g = (a, b]$and a distribution P, define the function $c_F(g, P)$ as

$$c_F(g,P)=\frac{-f\left(F^{-1}\left[P\{(0,b]\}\right]\right)+f\left(F^{-1}\left[P\{(0,a]\}\right]\right))}{P(g)}$$

For notational ease we suppress the dependence on F and write $c(g, P)$. Then c_i is $c(y_i, \hat{P})$, and the function $c(\cdot, \cdot)$ is a ranking function which takes intervals and converts them to rank-like scores based on the NPMLE. Then we can rewrite c_i as

$$\begin{aligned} c(y_i,\hat{P})&=\sum_{j=1}^{M}\frac{\hat{P}(\hat{\gamma}_j\cap y_i)c(\hat{\gamma}_j,\hat{P})}{\hat{P}(y_i)}\\ &=\sum_{j=1}^{M}\hat{P}r\left[X_i\in\hat{\gamma}_j|X_i\in y_i\right]c(\hat{\gamma}_j,\hat{P}) \end{aligned} \quad (5)$$

In other words, the rank-score for the ith individual is a weighted sum of the rank-scores for each of the possible intervals from the partition of the event space by the observed assessment times, and the weighting is in proportion to the estimated distribution given the observed interval.

3.2. Ideal Scores

Now suppose that P is known. Let $\lim_{n\to\infty}\gamma(\mathbf{y}_n) = \gamma = \{\gamma_1, \ldots, \gamma_m\}$ be the partition created from the set of all possible assessment times. Then the ideal score is analogous to equation 5,

$$c(y_i,P)=\sum_{j=1}^{m}\frac{P(\gamma_j\cap y_i)c(\gamma_j,P)}{P(y_i)}. \quad (6)$$

Let the partition induced by the ith individual's set of assessment times be $\mathbf{g}_i = \{g_{i1}, \ldots, g_{i,k_i+1}\}$, where $g_{ij} = (a_{i,j-1}, a_{ij}]$, $j = 1, \ldots, k_i + 1$. Each g_{ij} is the union of a contiguous set of γ_j intervals in γ. Let the observed interval be $Y(X_i, \mathbf{g}_i) = g_{ij}$ if $X_i \in g_{ij}$. Then under the null that $X_i \sim P$, the expected value of the ideal score given $\mathbf{g}_i$ is

$$\begin{aligned} E\left[c\{Y(X_i,\mathbf{g}_i),P\}\right]&=\sum_{j=1}^{k_i+1}P(g_{ij})c(g_{ij},P)\\ &=\sum_{j=1}^{k_i+1}-f\left(F^{-1}\left[P\{(0,a_{ij}]\}\right]\right)+f\left(F^{-1}\left[P\{(0,a_{i,j-1}]\}\right]\right)\\ &=-f\left(F^{-1}\left[P\{(0,a_{i,m_i+1}]\}\right]\right)+f\left(F^{-1}\left[P\{(0,a_{i0}]\}\right]\right)\\ &=-f\left(F^{-1}[1]\right)+f\left(F^{-1}[0]\right)=0 \end{aligned} \quad (7)$$

where the last step comes from the definition based on limits. Thus, regardless of the vector of assessment times, the expected value of the ideal rank-like score is 0.

Note this expectation result does not prove that the permutation test based on the ideal rank scores will retain the type I error rate. Consider an extreme case to show this is not true.

Example 1—Consider a two treatment situation where all subjects in group 0 are assessed at exactly the same time, say a_0, and all subjects in group 1 are assessed at a different time, say a_1. Let $P\{(0, a_j]\} = q_j$ for $j = 0, 1$, and suppose we use the Wilcoxon-type test so that F is the logistic and $f\{F^{-1}(q)\} = q(1 - q)$. Then for group j, the ideal scores will be $c((0, a_j], P) = -(1 - q_j)$ with probability q_j and $c((a_j, \infty), P) = q_j$ with probability $(1 - q_j)$. The probability that all scores from group j equal q_j is $(1 - q_j)^{nj}$. For example, suppose $n_0 = n_1 = 5$ and $q_0 = .1$ and $q_1 = .2$ then we have a $(.9^5) * (.8^5) = .193$ probability that all the scores from group 0 are q_0 and all the scores in group 1 are q_1. The exact one-sided p-value in that situation would be $1/\binom{10}{5} = 0.004$. So we would reject at the 0.025 level at least 19.3% of the time, and the type I error rate is not controlled.

If we used the actual scores not the ideal scores, then this problem does not happen in this case; suppose $a_0 < a_1$ and suppose no $y_i = (0, a_0]$or $y_i = (0, a_1]$values are observed, then $\hat{P}\{(0, a_1]\} = 0$ making $c\{(a_0, \infty), \hat{P}\} = c\{(a_1, \infty), \hat{P}\} = 0$ so that all scores are 0 for both treatment groups. Simulating 1000 data sets under the above scenario, we rejected none. We study a case where the exact permutation test using the actual scores does not retain the type I error rate in the Section 3.3.

This next example shows how the expectation result of equation 7 helps clarify why even if the assessment times are related to treatment, often the permutation test approximately retains the type I error rate unlike the midpoint imputation method. Although type I error rate problems with midpoint imputation are known [8], this expectation result gives new intuition about how the midpoint imputation can be much worse than the permutation method with respect to type I error rate.

Example 2—Consider a two group example, where for all individuals in the control group are assessed only at time 2, and for all individuals in the treatment group make an additional assessment time and are assessed at time 1 and 2. Suppose the treatment does not affect event time and is distributed exponential with mean 1. We see as expected from the result above, that for those with events at $t \leq 2$, the logrank scores for the treatment group and the control group have the same expectation.

Treatment Group			Control Group		
g	**P(g)**	**c(g,P)**	**g**	**P(g)**	**c(g,P)**
(0, 1]	0.632	0.368			
(1, 2]	0.233	−0.265	(0, 2]	0.632+0.233	0.197
(2, ∞)	0.135	−1.265	(2, ∞)	0.135	−1.265

Note that

$$\frac{632 * 0.368 + 233 * (-.265)}{632 + 233} = 0.197$$

and the score for the (0, 2]interval for the controls is a weighted average of the scores of the treatment.

Now consider the scores from midpoint imputation under the scenario. We only impute the midpoint for those who are interval censored. In the following the c^* columns denote the logrank scores that would result if we observed either exact values at the midpoints or right censored values given by the g^* columns in the proportions given by the appropriate $P(g)$ column.

Treatment Group				Control Group			
g	**g***	**P(g)**	**c***	**g**	**g***	**P(g)**	**c***
(0, 1]	{0.5}	0.632	0.648				
(1, 2]	{1.5}	0.233	−0.411	(0, 2]	{1}	0.632+0.233	0.0517
(2, ∞)	(2, ∞)	0.135	−1.411	(2, ∞)	(2, ∞)	0.135	−1.411

Note that

$$0.0517 \neq \frac{632 * 0.648 + 233 * (-.411)}{632 + 233} = 0.389$$

and the (0, 2]interval for controls does not translate into a weighted average of the scores for the treatment after midpoint imputation. We see that the expected score for the control group are quite a bit less than for the treated group, so one might correctly infer that the midpoint imputation does not retain the type I error rate. Simulating under this scenario with 50 in each group and 10,000 replications, the logrank test using the midpoint imputation method rejects 63.0% of the time at the 5% level, while the GCM interval logrank permutation test (using the permutation central limit theorem) only rejects 3.6% of the time.

3.3. Actual Scores and Type I Error

Heimann and Neuhaus [20] and Heinze, Gnant, and Schemper [21] have shown by simulation that in some extreme situations with right-censored data and differing assessment distributions based on treatment (i.e., censoring related to treatment) that the exact permutation based on actual scores does not maintain the nominal type I error rate. In this section, we study a simple example in detail to get intuition about why this problem can occur for interval censored data as well.

Example 3—We consider an example similar to example 1 where each subject is assessed only once. Let p_{ij} be probability that a subject in the ith group is assessed at a_j, with $a_1 < a_2$. Let $\mathbf{p}_i = [p_{i1}, p_{i2}]$, and let $q_j = P\{(0, a_j]\}$. Consider the case with $n_0 = n_1 = 500$, $q_1 = .0001$, $q_2 = 0.001$, $\mathbf{p}_0 = [0.2, 0.8]$, and $\mathbf{p}_1 = [0.5, 0.5]$. We simulated this scenario and out of 1000 simulations, we reject the exact weighted logrank permutation test (estimated by Monte Carlo simulation of 99 replications each, using either the logrank scores or the Wilcoxon-type scores) 17.3% of the time at the two-sided 0.05 level.

To gain insight on why the type I error rate is not met in this situation, we give one of the simulated data sets that reject in Table 1. Notice that we get 0 for the expected sum of scores for each set of assessment intervals within each group (i.e., for Group i, $E(N) * c(y, P)$ for $y = (0, a_j]$ plus $E(N) * c(y, P)$ for $y = (a_j, \infty)$ is 0, for any i, j), however, the analogous sum for the observed sum of scores is not zero. We give the scores in tabular form:

	−0.0016	0	0.9984
Group=0	402	98	0
Group=1	227	272	1

We see that Group 1 has much fewer with the lowest score and the only one with the highest score. Note that the ideal scores associated with (0, a_1]and (0, a_2]are close to each other and close to 1, while the ideal scores associated with (a_1, ∞) and (a_2, ∞) are close to each other and close to 0. Of the scores close to 0, we can see by looking at the expected number in each category of response that Group 0 will virtually always have many more with the lowest score than Group 1. The key to understanding the type I error rate problem is that the scores close to 1 are very unlikely and a substantial proportion of the time all of those scores close to 1 may fall in Group 1, in which case Group 1 will have more of the highest score and is virtually certain to have much less of the lowest scores. We estimate the proportion of times all the scores close to 1 fall in Group 1 as

$$Pr\left[\text{none in Group 0}\right] \quad * \quad Pr\left[\text{at least one in Group 1}\right] \approx$$
$$\{1-(p_{01}q_1+p_{02}q_2)\}^{n_0} \quad * \quad [1-\{1-(p_{11}q_1+p_{12}q_2)\}^{n_1}]=0.16$$

This explains the type I error rate problem shown previously by simulation. Note that this situation depends on very low expected values in the high scores. If we increase both q_1 and q_2 ten-fold, then the above equation gives 0.0153, and that source of type I error is not a problem (for the 0.05 significance level at least). A simulation with 1000 replications rejects at the 0.05 level 4.2% of the time.

4. Multiple Imputation-Based Tests

Huang, Lee and Yu [7] repeatedly simulate right censored data using the NPMLE from the interval censored data, and for each replicate data set combine logrank statistics and their Martingale-based variances using ideas from within cluster resampling [22]. Huang, Lee and Yu [7] did not study ATD, but we note that their multiple imputation method is similar to a multiple imputation approach that has been used for right censored data to create logrank tests that have been shown by simulation to be valid for small sample sizes with ATD [21, 23]. For the right censoring case, Heinze, Gnant, and Schemper [21] and Wang, Lagakos, and Gray [23] estimate both the event time and the censoring distribution, then resample from both estimated distributions. For the usual application of interval censored data (including all the ones mentioned in this paper), we do not use the entire set of assessment times for each individual, but only use the assessment times that bracket the event. Using only that data, we cannot estimate the assessment time distributions for the two treatment

groups. Nevertheless, we show that when we write the imputed weighted logrank scores of Huang, Lee and Yu [7] in an ideal way assuming that the event time distribution under the null is known, then under the null hypothesis the distribution of those ideal imputation scores does not depend on the assessment times. Although the actual scores use the NPMLE of the distribution of the combined data, because of the consistency of that NPMLE even with differing assessment times in different groups, we expect that the method of Huang, Lee and Yu [7] will perform well with large samples. We justify by simulation that the Huang, Lee and Yu [7] method approximately retains the type I error rate, and importantly we offer two modifications to the Huang, Lee and Yu [7] method that each retain type I error rate in all simulated scenarios including some with very small sample sizes.

4.1. Ideal Imputation

Consider again the ideal permutation scores, given by equation 6 which assume[$\gamma_1, \ldots, \gamma_m$]and P are known. A natural imputation strategy is to replace the ideal score, $c(y_i, P)$ with an ideal imputed score, denoted $C(y_i, P)$, where we define $C(y_i, P)$ as a pseudo-random sample, where we sample $c(\gamma_j, P)$ with probability $P(y_i \cap \gamma_j)/P(y_i)$. The beauty of imputation (in the ideal case at least) is that the distribution of the imputed values does not depend on the assessment time intervals, say $\mathbf{g}_i$. Then given $\mathbf{g}_i$, the probability that an ideal imputation score will equal the ℓth of the rank scores using the partition $[\gamma_1, \ldots, \gamma_m]$, is

$$Pr[C\{Y(X_i, \mathbf{g}_i), P\} = c(\gamma_\ell, P)|\mathbf{g}_i] = \sum_{j=1}^{k_i+1} Pr(X_i \in g_{ij}) \frac{P(g_{ij} \cap \gamma_\ell)}{P(g_{ij})} = P(\gamma_\ell). \quad (8)$$

We see that the distribution of $C\{Y(X_i, \mathbf{g}_i), P\}$ does not depend on $\mathbf{g}_i$.

4.2. Definition of Tests

Huang, Lee and Yu [7] imputed failure times based on the NPMLE, recalculated logrank scores from the imputed failure times, recalculated the usual Martingale-based variance for each imputation, then combined that information from the imputations using ideas also used in within cluster resampling [22, 24]. Huang, Lee and Yu [7] showed that using within cluster resampling ideas performed better than previous multiple imputation strategies for interval censored data.

Recall $\mathbf{a}_i$ are the assessment times for the ith individual and $t_1 < \cdots \prec t_{M-1}$ are the union of the observed assessment times for all n individuals. Let i_ℓ and i_r be indeces such that $\ell_i = t_{i\ell}$ and $r_i = t_{i_r}$. Then define the imputed score analogous to the ideal imputed scores; let the imputed scores $C(y_i, \hat{P})$ be a pseudo-random sample from $c(\hat{\gamma}_{i\ell+1}, \hat{P}), \ldots, c(\hat{\gamma}_{i_r}, \hat{P})$ with associated probabilities $w_{i,\ell+1}, \ldots, w_{ir_i}$, where

$$w_{ij} = \frac{\hat{P}(\hat{\gamma}_j \cap y_i)}{\hat{P}(y_i)}$$

Let $C^{(j)}(\hat{P}) = [C^{(j)}(y_1, \hat{P}), \ldots, C^{(j)}(y_n, \hat{P})]$ be the sample of n values from the jth imputation. Then we have three ways we can perform inferences from the imputations.

1. For each imputation, we treat the scores as ordered responses and follow exactly the Huang, Lee and Yu7 [] method.

2. We use the imputed scores directly; for each imputation calculate the permutational variance of those scores, and combine the imputations following Huang, Lee and Yu [7]. Specifically, let $T^{(j)}=\sum_{i=1}^{n} z_i C^{(j)}(y_i, \hat{P})$ and let $V^{(j)}$ be the associated permutational variance. Then we treat the mean of the $T^{(j)}$ values as normal with variance,

$$\hat{V}_p=\frac{\sum_{j=1}^{J} V^{(j)}}{J}-\frac{\sum_{j=1}^{J}(T^{(j)}-\overline{T})(T^{(j)}-\overline{T})'}{J-1}$$

   where $\bar{T}$ is the mean of the test statistics on the imputations.

3. We can use a Monte Carlo permutation estimate to calculate the p-values. Let $T_{jk}=\sum_{i=1}^{n} z_i^{(jk)} C^{(j)}(y_i, \hat{P})$, where $\mathbf{z}^{(jk)}=\left[z_1^{(jk)}, \ldots, z_n^{(jk)}\right]$ is the kth permutation of the covariates associated with that imputation, and we define $\mathbf{z}^{(j0)} = \mathbf{z}$. Then when the test statistic of the permutation test rejects for large values, the exhaustive p-value is defined as $E\{I(T_{jk} \geq T_{j0})\}$, where the expectation is defined over the imputations and the permutations. We estimate the exhaustive p-value by Monte Carlo simulation as

$$p=\frac{1+\sum_{j=1}^{J}\sum_{k=1}^{K} I(T_{jk} \geq T_{j0})}{1+JK},$$

   where J and K are the number of imputations and Monte Carlo permutations respectively. The ones are added to the numerator and denominator since if each within subject resampling created a valid permutation test then defining p this way will ensure a valid p-value [25]. For discussion on ways to choose J and K, or calculate the p in a more complicated way to get better precision see [26].

4.3. Asymptotic Distribution of Imputed Scores under the Null Hypothesis

Theorem 1—Assume the assumptions stated in the second paragraph of Section 2 (i.e., for the ith subject, the distribution of the number of assessments, K_i, and the vectors assessment times $\mathbf{A}_i$ may depend on treatment z_i, but failure time, X_i, is independent of both z_i and of K_i and $\mathbf{A}_i$). Let $\hat{P}_n$ be the NPMLE based on interval responses from all n individuals, $\mathbf{y}_n$, Let $\gamma(\mathbf{y}_n) = \hat{\gamma}_n$ be the partition induced by $\mathbf{y}_n$, and $\gamma = [\gamma_1, \ldots, \gamma_m]$ be the smallest possible partition induced by an infinite sample. Suppose that for any ε there exists some $N(\varepsilon)$ such that $Pr[\hat{\gamma}_n = \gamma] > 1 - \varepsilon$ for all $n > N(\varepsilon)$. Let $\mathbf{g}_h$ be an arbitrary partition of the event space, then

$$Pr\left[C\left\{Y(X_i,\mathbf{g}_h),\hat{P}_n\right\}=c\left(\gamma_\ell,\hat{P}_n\right)|\mathbf{g}_h\right]\rightarrow P(\gamma_\ell).$$

To prove this theorem, we first note that if we ignore treatment assignment, then the K_i and $\mathbf{A}_i$ are independent of X_i and the usual assumptions for consistency of $\hat{P}_n$ hold (see e.g., [18, 27]). In other words, $\hat{P}_n(\gamma_j) \rightarrow P(\gamma_j)$ for $j = 1, \ldots, m$. Then the theorem follows using Slutzky's theorem on an analogous equation to equation 8.

Theorem 1 shows us that asymptotically the distribution of any one of the imputed scores under the null hypothesis does not depend on the covariates, z_i; however, it does not completely define the joint distribution of the imputed scores within a treatment group. A formal proof showing that a permutation test based on the imputed scores is asymptotically valid even when the inspection process depends on z_i is more difficult. To study this more, and to help elucidate finite sample properties, we perform simulations in the next section.

5. Simulations

5.1. Description

We describe the simulations using 4 scenarios which describe general assessment time models and 5 implementations which may have different samples sizes and or assessment parameters. Here are the 4 scenarios:

1. **Mixed Discrete Assessment (MDA):** There are 10 equally spaced possible assessment times at 1, 2, …, 10, with 2 mandatory assessment times (at 3 and 10) and 8 optional ones (all the others). Under the null, the mandatory assessment times occur when 45.1% and 86.5% of the events are expected to have occurred. We assume the optional assessments are independent of each other and the probability of an optional assessment in group j is ξ_j.
2. **Continuous Assessment (CA):** In this scenario the assessments take place in continuous time, so that there are an infinite number of possible assessment times. The ith subject has K_i assessments, where $K_i - 1$ is distributed Poisson with mean ξ_j if the subject is in group j. The K_i assessments are independent and uniformly distributed on 0 to 10.
3. **Decreasing Probability of Assessment (DPA):** In this scenario the assessments occur in discrete time at 1, 2, …, 10, and we consider cases where the probability of making each assessment is independent and equal to Pr[Assess at time=t] = $\exp(-t/\xi_j)$. When $\xi_j = 5$ the probability of an assessment at t ranges from 82% at t=1 to 13.5% at t=10. For $\xi_j = 50$ the probability of assessment ranges from 98% at t=1 to 81.9% at t=10.
4. **One Extreme Assessment (OEA):** In this scenario we repeat example 3, assuming that there are 500 subjects in each group, and each subject has one assessment occurring at either $a_1 = 0.0005$ or $a_2 = 0.005$. The probability that a subject in the jth group is assessed at a_i is p_{ij}, with $\mathbf{p}_j = [p_{1j}, p_{2j}]$, $\mathbf{p}_0 = [0.2, 0.8]$, and $\mathbf{p}_1 = [0.5, 0.5]$.

The distribution of the event times for group j is exponential with mean μ_j. Under the null $\mu_0 = \mu_1 = 5$, while under the alternative μ_0 and μ_1 are defined differently for each implementation in order to have interesting power results. Here is a description of the 5 implementations under the first three scenarios:

- **a.** Moderate and equal sample sizes ($n_0 = n_1 = 50$) with equal assessment models for the two groups. The assessment parameters in the first 3 scenarios are: (1) $\xi_0 = \xi_1 = 0.5$, (2) $\xi_0 = \xi_1 = 5$, (3) $\xi_0 = \xi_1 = 5$. The alternative simulations have $\mu_0 = 4$ and $\mu_1 = 6$.
- **b.** Moderate and equal sample sizes ($n_0 = n_1 = 50$) with unequal assessment models. The assessment parameters are: (1) $\xi_0 = 0.25$, $\xi_1 = 0.75$, (2) $\xi_0 = 2.5$, $\xi_1 = 7.5$, (3) $\xi_0 = 5$, $\xi_1 = 50$. The alternative simulations have $\mu_0 = 4$ and $\mu_1 = 6$.
- **c.** Small and equal sample sizes ($n_0 = n_1 = 5$) with unequal assessment models. The assessment parameters are the same as in (b). The alternative simulations have $\mu_0 = 2$ and $\mu_1 = 8$.
- **d.** Unequal sample sizes ($n_0 = 5$ and $n_1 = 50$) with unequal assessment models. The assessment parameters are the same as in (b). The alternative simulations have $\mu_0 = 3$ and $\mu_1 = 7$.
- **e.** Unequal sample sizes switched. These are the same as (d) except $n_0 = 50$ and $n_1 = 5$.

We label each situation (i.e., each scenario/implementation) by the number letter combination (e.g., 1a or 3d). For each data set we perform a two-sided logrank test using Sun's (1996) formulation, and using one of the following tests:

REI (Right Endpoint Imputation): is the usual logrank test for right censored data (using survdiff in the survival package) after assuming that all non-right censored observations had the event exactly observed at the right endpoint;

pMC: is the permutation test, but instead of completely enumerating all possible permutations we take a Monte Carlo sample. Let U_i be the score statistic from the ith replicate, and U_0 be the score statistic from the original data, then the one-sided pMC p-value is $(1 + \#(U_i \geq U_0))/(1 + R)$, where $R = 299$ is the number of Monte Carlo replications. *If* the complete enumeration permutation test is exact then (even with ties allowed) this Monte Carlo method will retain the type I error rate for any R [25];

Score: is the score test using the *ad hoc* adjustment of Fay [10] if necessary;

wsrMC: is the within subject resampling (WSR) test described in Section 4.2 number 3 using 299 imputations and 299 Monte Carlo replications;

wsrHLY: is the within subject resampling (WSR) test of Huang, Lee and Yu [7] using 299 imputations;

wsrPCLT: The within subject resampling (WSR) test described in Section 4.2 number 2 using 299 imputations.

The simulations were done on the Biowulf Linux cluster at NIH (http://biowulf.nih.gov) in R (version 2.11.1) using the interval package [17] (version 1.0–1.2), which has options for all of the above tests, except REI which used the survival package.

5.2. Results and Interpretation

We present the simulations under the null in Table 2. First notice that under implementations *a* (1a, 2a, and 3a), where there are equal assessment models for both groups, that all the methods retain the type I error rate. In the other situations where the assessment models are unequal between the two groups, the REI test usually does not retain the type I error rate. In some situations, the REI method has very high rejection rates under the null.

For the moderately sized simulations with assessment-treatment dependence (implementations *b*), we note that all of the methods except REI appear to retain the type I error rate fairly well. This is especially important for the score method under continuous assessment (2b). Previously, no simulations had been done to assess the *ad hoc* adjustment to the score test for the continuous assessment proposed by Fay [10], and the simulation 2b shows that it retains the type I error rate at least in that one situation.

For smaller sample sizes (implementations *c*, *d* and *e*), consider first the asymptotic methods (PCLT, Score, wsrHLY, and wsrPCLT). The asymptotic methods based on the PCLT (PCLT, wsrPCLT) appear to retain the type I error rate much better than the other two (Score, wsrHLY). The PCLT method is generally quite good in all situations, although it does not retain the type I error rate in 3e, whereas the wsrPCLT does slightly better in terms of retaining the type I error rate. Now consider the non-asymptotic methods (pMC and wsrMC). First, note that the pMC is not exact as it does not retain the type I error rate in all situations (see 2e, 3e, and 4). The wsrMC is still a candidate for an exact method since it does retain the type I error rate in all simulations that we considered; however, we have not proven exactness.

In Table 3 we give the simulated power under alternative hypotheses. For shorthand we will call a test “valid” for a particular situation if it did not have simulated type I error rate significantly greater than 5%. We do not consider the REI test a viable option because it had simulated error more than 10% for several situations. Recall the PCLT test had simulated size less than 5% in nearly all situations studied, but we see in Table 3 that the price for retaining that type I error rate can be a substantial drop in power (see 1c, 1d,2c,2d,3c,3d). It appears that the method of Huang, Lee and Yu [7] (i.e., wsrHLY) should be preferred over the score test since the powers are similar in Table 3, while in Table 2 when either test has simulated type I error rates larger than 5%, the score test rates are always worse. There is no uniformly best test in the sense of maximizing the power within each situation compared to the other tests which are valid. For example, in 1e the PCLT test appears to have the best power of the valid tests, while in 1d the PCLT has considerably less power than the wsrMC.

Note that the Monte Carlo tests (pMC and wsrMC) may slightly increase power by taking more Monte Carlo replications. We expect that by taking only 299 replications the power will be at least 90% of the power that could be achieved by complete enumeration (see [28], p. 155). The effect on the power of the multiple imputation methods by taking less than an

infinite number of imputations is more difficult to elucidate because the imputations are not simply a series of Bernoulli replications as are the Monte Carlo replications; nevertheless, we suspect that little power will be gained by taking more replications.

We repeated all simulations with $n_a = 5$ for any group (implementations *c*, *d*, and *e*) after replacing the associated sample sizes with $n_a = 25$. The corresponding simulated type I error rates are generally closer to the nominal levels (except the REI test), and the powers are larger as one would expect with larger sample sizes. Those results are not shown.

6. Application

We apply the different tests to the bladder cancer trial mentioned in the introduction [1]. Data for the placebo (n=47) and thiotepa (n=38) arms only are available at the Royal Statistical Society Data Sets website (http://www.blackwellpublishing.com/rss/) as a supplement to Sun and Wei [2]. We perform a logrank-type test using Sun's (1996) scores on the time to first recurrence. Using the interval R package [17], with $z_i = 0$ for placebo or $z_i = 1$ for thiotepa, we get that $U = -4.49$ implying that the recurrence times are on average later for the thiotepa group. For the multiple imputation methods we use 999 imputations and for the Monte Carlo methods we use 999 replications. From the simulation section, we expect that the PCLT, wsrPCLT and wsrMC tests to retain the type I error rate even if the assessment times are related to the treatment. The two-sided p-values are similar for all the methods: REI, p=0.337; pMC, p=0.178; PCLT, p=0.165; Score, p=0.162; wsrMC, p=0.224; wsrHLY, p=0.168; wsrPCLT, p=0.165. Similarly, we can perform a Wilcoxon-type test which weights early events more. The results are similar: REI, p=0.217; pMC, p=0.254; PCLT, p=0.220; Score, p=0.213; wsrMC, p=0.263; wsrHLY, p=0.223; wsrPCLT, p=0.222.

7. Discussion and Recommendations

In this paper we have explored weighted logrank tests for interval censored data under assessment treatment dependence. Although previously only the score test from the grouped continuous model with large sample size and small number of assessment times was known to be valid under ATD, we have shown that in other situations other methods may perform as well or better. Specifically, we have given heuristic and simulation-based justification to show that in many cases permutation-based weighted logrank tests are valid under ATD, and additionally we have given asymptotic and simulation-based justification to show that the weighted logrank test of Huang, Lee and Yu [7] is often approximately valid under ATD. Importantly, we have developed two modifications of the method of Huang, Lee and Yu [7] that retain type I error rate under simulations with small sample sizes and ATD. Furthermore, we have added more justification for the score test under continuous assessment and have shown through simulation that the *ad hoc* adjustment of Fay [10] appears to retain the type I error rate for moderate sample sizes but only under assessment-treatment independence. All the tests done in the simulation (except right endpoint imputation) have been made available as options in the R package interval (see [17] for a description of its use).

If the full vector of assessment times is available for each subject, then one could test for ATD using standard tests on the number of assessments such as the t-test or the Wilcoxon-

Mann-Whitney test. We do not recommend testing for ATD and choosing the particular logrank implementation based on the results of the test for ATD, since ATD may be present and not detectable by significance test. It is better to use methods that retain the type I error rate regardless of the presence of ATD.

Based on the work of this paper we make the following recommendations:

- Most importantly, when testing interval censored data do not use either midpoint or right endpoint imputation. These methods can severely inflate the type I error rate when the assessment times are related to treatment. Despite the fact that this has been known for a long time (see [8]), these naive methods continue to be used [9], and our simulations have reconfirmed that they should not routinely be used.
- If the researcher wants a non-random method that approximately retains the type I error rate in most situations even with small sample sizes, then the method based on the permutational central limit theorem (PCLT) is recommended. Because this method is not based on simulations (i.e., is non-random), it is faster to calculate than the simulation based methods (i.e., within subject resampling methods) and two statisticians need not use the identical software to get the same answer with the same data.
- If the primary concern of the researcher is with retaining type I error rate even with very small sample sizes, then the simulations indicate that the wsrPCLT method works well. Theoretically, we know that it is unlikey to be exact because it is based on asymptotic theory, and a conservative approach would be to use the wsrMC method. Although, both methods are technically random since they are based on simulating many imputations, by increasing the number of imputations, and for the wsrMC method increasing the Monte Carlo replications as well, those sources of variability can be made as small as needed.
- If the sample sizes are equal, the method of Huang, Lee and Yu [7] (labeled wsrHLY in simulations) may be a reasonable option since it bounded the simulated type I error in almost all cases studied and had substantially greater power than the PCLT in some situations. When the sample sizes where not equal, the wsrHLY method had simulated sizes greater than 5% but less than 7%.

Acknowledgments

The authors thank Pam Shaw for helpful comments on the paper.

References

1. Byar, D. The verterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparison of placebo, pyridoxine, and topical thiotepa. Plenum; New York: 1980. Bladder Tumors and Other Topics in Urological Oncology, chap.
2. Sun J, Wei LJ. Regression analysis of panel count data with covariate-dependent observation and censoring times. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2000; 62(2):293–302.

3. Freidlin B, Korn EL, Hunsberger S, Gray R, Saxman S, Zujewski JA. Proposal for the use of progression-free survival in unblinded randomized trials. Journal of Clinical Oncology. 2007; 25(15):2122–2126. [PubMed: 17513819]
4. Self SG, Grosman EA. Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. Biometrics. 1986; 42:521–530. [PubMed: 3105615]
5. Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics. 1986; 42:845–854. [PubMed: 3814726]
6. Sun J. A non-parametric test for interval-censored failure time data with application to aids studies. Statistics in Medicine. 1996; 15:1387–1395. [PubMed: 8841649]
7. Huang J, Lee C, Yu Q. A generalized log-rank test for interval-censored failure time data via multiple imputation. Statistics in Medicine. 2008; 27:3217–3226. [PubMed: 18254128]
8. Law C, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. Statistics in Medicine. 1992; 11:1569–1578. [PubMed: 1439361]
9. Sun X, Chen C. Comparison of finkelstein's method with the conventional approach for interval censored data analysis. Statistics in Biopharmaceutical Research. 2010; 2:97–108.
10. Fay MP. Rank invariant tests for interval censored data under the grouped continuous model. Biometrics. 1996; 52:811–822. [PubMed: 8805758]
11. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society A. 1972; 135:185–207.
12. Sun J. A nonparametric test for current status data with unequal censoring. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1999; 61(1):243–250.
13. Zhu C, Yuen K, Sun J, Zhao X. A nonparametric test for interval-censored failure time data with unequal censoring. Communications in StatisticsTheory and Methods. 2008; 37(12):1895–1904.
14. Farrington C, Gay N. Interval-censored survival data with informative examination times: parametric models and approximate inference. Statistics in medicine. 1999; 18(10):1235–1248. [PubMed: 10363342]
15. Finkelstein D, Goggins W, Schoenfeld D. Analysis of failure time data with dependent interval censoring. Biometrics. 2002; 58(2):298–304. [PubMed: 12071402]
16. Zhang Z, Sun L, Sun J, Finkelstein D. Regression analysis of failure time data with informative interval censoring. Statistics in Medicine. 2007; 26(12):2533–2546. [PubMed: 17072823]
17. Fay MP, Shaw PA. Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. Journal of Statistical Software. 2010; 36(2):1–34. URL http://www.jstatsoft.org/v36/i02/.
18. Gentleman R, Geyer C. Maximum likelihood for interval censored data: Consistency and computation. Biometrika. 1994; 81:618–623.
19. Fay MP. Comparing several score tests for interval censored data (Corr: 1999V18 p2681). Statistics in Medicine. 1999; 18:273–285. [PubMed: 10070674]
20. Heimann G, Neuhaus G. Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. Biometrics. 1998; 54:168–184. [PubMed: 9544515]
21. Heinze G, Gnant M, Schemper M. Exact log-rank tests for unequal follow-up. Biometrics. 2003; 59:1151–1157. [PubMed: 14969496]
22. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. Biometrika. 2001; 88:420–429.
23. Wang R, Lagakos S, Gray R. Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. Biostatistics. 2010; 11:676–692. [PubMed: 20439258]
24. Follmann D, Proschan M, Leifer E. Multiple outputation: Inference for complex clustered data by averaging analyses from independent data. Biometrics. 2003; 59(2):420–429. [PubMed: 12926727]
25. Fay MP, Follmann DA. Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests. The American Statistician. 2002; 56(1):63–70.
26. Follmann D, Fay M. Exact inference for complex clustered data using within-cluster resampling. Journal of Biopharmaceutical Statistics. 2010; 20(4):850–869. [PubMed: 20496210]

27. Schick A, Yu Q. Consistency of the GMLE with mixed case interval-censored data. Scandinavian Journal of Statistics. 2000; 27(1):45–55.

28. Davidson, A.; Hinkley, D. Bootstrap Methods and Their Application. Cambridge University Press; New York: 1997.

Table 1

Table of one simulated data set that rejects from example 3. N is the number observed for the given group and y_i, and $E(N)$ is the expected number. The $c(y_i, \hat{P})$ is the observed Wilcoxon-type score, where $\hat{P}$ is the NPMLE from the simulated data.

Group	y_i	N	$E(N)$	$c(y_i, \hat{P})$	$c(y_i, P)$
0	$(0, a_1]$	0	0.010	-	0.9999
	(a_1, ∞)	98	99.990	0.0000	−0.0001
	$(0, a_2]$	0	0.400	-	0.9990
	(a_2, ∞)	402	399.960	−0.0016	−0.0010
1	$(0, a_1]$	0	0.025	-	0.9999
	(a_1, ∞)	272	249.975	0.0000	−0.0001
	$(0, a_2]$	1	0.250	0.9984	0.9990
	$(a_2, \infty]$	227	249.750	−0.0016	−0.0010

Stat Med. Author manuscript; available in PMC 2014 May 21.

Table 2

Percent Rejected from Simulations Under the **Null** Hypothesis using Nominal Two-sided 5 percent level. Values preceded with * are significantly larger than the nominal 5% by exact one-sided 0.025 binomial test. We used the logrank test (Sun's [1996] implementation). In Simulation Description: SN=simulation number, Scn=scenario, $\hat{E}(M)$=average number of assessment times in each simulation, eqA= equal Assessment distributions for both treatments (yes/no). See text for complete descriptions of simulations and tests. Simulations where based on 10,000 replications and all tests were applied to the same 10,000 data sets.

Simulation Description						Percent Rejected						
SN	**Scn**	n_0	n_1	$\hat{E}$ (*M*)	**eqA**	**REI**	**pMC**	**PCLT**	**Score**	**wsrMC**	**wsrHLY**	**wsrPCLT**
1a	MDA	50	50	11.0	yes	5.2	4.5	4.9	5.4	4.5	5.2	4.8
1b	MDA	50	50	11.0	no	*21.5	4.8	4.9	5.4	4.3	5.1	4.9
1c	MDA	5	5	8.3	no	*8.8	3.1	3.4	*7.8	2.3	5.3	3.4
1d	MDA	5	50	11.0	no	4.6	4.0	3.6	*7.0	4.0	*6.5	3.9
1e	MDA	50	5	10.9	no	*18.0	5.2	4.4	*6.8	4.3	*6.3	4.0
2a	CA	50	50	156.8	yes	4.9	4.5	4.7	5.4	3.0	4.9	4.7
2b	CA	50	50	153.0	no	*10.9	4.7	5.0	*5.7	2.6	5.1	5.0
2c	CA	5	5	16.2	no	*8.0	3.2	3.8	*8.3	1.0	5.1	3.8
2d	CA	5	50	89.7	no	2.8	3.4	3.2	*6.8	1.8	*5.9	4.5
2e	CA	50	5	79.3	no	*15.6	*6.1	5.4	*7.2	3.0	*6.1	4.1
3a	DPA	50	50	11.0	yes	5.1	4.6	4.9	5.4	2.0	4.9	4.8
3b	DPA	50	50	11.0	no	*6.9	4.7	4.9	*5.7	3.1	5.1	5.0
3c	DPA	5	5	8.5	no	*6.8	3.4	4.0	*8.0	2.1	*5.6	4.0
3d	DPA	5	50	11.0	no	4.3	3.3	2.9	*6.5	2.5	*5.7	4.9
3e	DPA	50	5	10.9	no	*10.9	*6.0	*5.8	*7.2	4.0	*6.6	4.0
4	EA	500	500	3.0	no	0.2	*14.2	0.2	0.2	0.0	0.2	0.2

Table 3

Percent Rejected from Simulations Under the **Alternative** Hypothesis using Nominal Two-sided 5 percent level. Values preceded with * are significantly larger than the nominal 5% under the Null hypothesis of Table 2. We used the logrank test (Sun's [1996] implementation). In Simulation Description: SN=simulation number, Scn=scenario, $\hat{E}(M)$=average number of assessment times in each simulation, eqA= equal Assessment distributions for both treatments (yes/no). See text for complete descriptions of simulations and tests. Simulations where based on 10,000 replications and all tests were applied to the same 10,000 data sets.

Simulation Description						Percent Rejected						
SN	**Scn**	n_0	n_1	$\hat{E}(M)$	**eqA**	**REI**	**pMC**	**PCLT**	**Score**	**wsrMC**	**wsrHLY**	**wsrPCLT**
1a	MDA	50	50	11.0	yes	45.0	43.1	45.0	46.7	43.8	45.9	44.9
1b	MDA	50	50	11.0	no	*10.1	43.2	45.0	46.8	43.5	45.2	45.0
1c	MDA	5	5	7.8	no	*28.1	25.0	27.1	*48.7	24.9	41.1	27.0
1d	MDA	5	50	11.0	no	19.9	23.8	11.6	*43.5	22.3	*41.7	12.2
1e	MDA	50	5	10.7	no	*22.1	40.4	49.2	*39.1	39.4	*36.3	47.4
2a	CA	50	50	156.4	yes	39.9	40.5	42.4	44.5	34.6	42.8	42.3
2b	CA	50	50	151.9	no	*14.4	37.7	39.3	*42.3	29.9	40.3	39.4
2c	CA	5	5	15.6	no	*23.2	22.0	27.1	*43.8	13.3	34.5	27.3
2d	CA	5	50	86.6	no	19.4	18.7	11.2	*39.3	14.5	*37.1	9.8
2e	CA	50	5	79.4	no	*20.1	*39.9	48.0	*38.2	32.7	*33.9	44.9
3a	DPA	50	50	11.0	yes	36.4	36.1	38.1	39.6	25.6	38.7	38.2
3b	DPA	50	50	11.0	no	*23.7	35.5	37.1	*41.2	30.5	39.0	37.0
3c	DPA	5	5	8.2	no	*32.0	24.5	26.8	*47.1	22.9	*38.2	26.8
3d	DPA	5	50	11.0	no	27.9	20.6	11.4	*40.1	20.0	*38.5	11.6
3e	DPA	50	5	10.7	no	*27.8	*39.3	*46.9	*36.1	35.3	*32.9	43.8
4	EA	500	500	3.0	no	0.0	*5.2	0.0	0.1	0.0	0.0	0.0