HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2016 September 07.

Published in final edited form as:

Biometrics. 2016 September ; 72(3): 926–935. doi:10.1111/biom.12475.

Global Rank Tests for Multiple, Possibly Censored, Outcomes

R. Ramchandani[1,*], **D.A. Schoenfeld**[1,2], and **D.M. Finkelstein**[1,2]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave Boston, MA 02115, U.S.A

[2]Department of Biostatistics, Massachusetts General Hospital, 50 Staniford St. Boston, MA 02114, U.S.A

Summary

Clinical trials often collect multiple outcomes on each patient, as the treatment may be expected to affect the patient on many dimensions. For example, a treatment for a neurological disease such as ALS is intended to impact several dimensions of neurological function as well as survival. The assessment of treatment on the basis of multiple outcomes is challenging, both in terms of selecting a test and interpreting the results. Several global tests have been proposed, and we provide a general approach to selecting and executing a global test. The tests require minimal parametric assumptions, are flexible about weighting of the various outcomes, and are appropriate even when some or all of the outcomes are censored. The test we propose is based on a simple scoring mechanism applied to each pair of subjects for each endpoint. The pairwise scores are then reduced to a summary score, and a rank-sum test is applied to the summary scores. This can be seen as a generalization of previously proposed nonparametric global tests (e.g. O'Brien 1984). We discuss the choice of optimal weighting schemes based on power and relative importance of the outcomes. As the optimal weights are generally unknown in practice, we also propose an adaptive weighting scheme and evaluate its performance in simulations. We apply the methods to analyze the impact of a treatment on neurological function and death in an ALS trial.

Keywords

ALS; Global test; Multiple endpoints; Nonparametric; Rank-sum; U-statistic

1. Introduction

Many clinical trials are conducted to compare treatments with respect to a single primary measure, such as time to death. A single outcome, however, does not always adequately capture the entire effect of a therapy, which can impact patients in many dimensions. For example, new treatments for amyotrophic lateral sclerosis (ALS) target both mortality and different aspects of neurological function, which are measured using the ALS Functional Rating Scale (ALSFRS-R) (Cedarbaum et al., 1999). In such cases, it is useful to test the

*ritesh@mail.harvard.edu

Supplementary Materials: Web Appendices referenced in Sections 2.2, 3.1, and 4.2 are available with this paper at the Biometrics website on Wiley Online Library.

efficacy of a treatment with respect to all relevant outcomes simultaneously. The design, analysis, and interpretation of studies in the presence of multiple outcomes like these can be difficult, especially when some of the outcomes are subject to censoring. We propose flexible nonparametric global tests to summarize a treatment effect across multiple endpoints.

Several methods for combining multiple endpoints have previously been proposed. Pocock, Geller, and Tsiatis (1987) provide a global test statistic that can be used to combine any set of asymptotically normal test statistics. Many authors have also proposed nonparametric tests based only on composite ranks of a set of outcomes. O'Brien's (1984) nonparametric rank-sum method sums the ranks for each outcome, and makes inference on the combined ranks. Wei and Johnson (1985) combined Wilcoxon statistics for incomplete repeated measurement data using U-statistics. Finkelstein and Schoenfeld's joint rank test (1999) is a method that compares each pair of subjects with respect to mortality and a secondary endpoint jointly, an extension of similar joint tests proposed by Moyé et al. (1992; 2011). Wittkowski (2004) proposed a test for multivariate ordinal data using U-statistics based on a product ordering of outcomes, an idea also explored by Rosenbaum in depth (1991 (1994). Häberle, Pfahlberg, and Geffeler (2009) defined the ranking methods of many of the above referenced tests in terms of different types of partial orders.

These combined tests have increasingly attracted clinical interest for complex diseases where treatment is expected to affect multiple dimensions. Felker and Maisel (2010) suggested using global rank approaches for trials of acute heart failure, with death, dyspnea improvement, and other biomarkers as outcomes. Sun et al. (2012) assessed the performance of various global approaches using simulations based on phase II trials for acute heart failure. Healy and Schoenfeld (2012) also examined through simulation how a global test performs relative to other methods of analyzing a longitudinal and survival outcome jointly. Berry et al. (2013) proposed using a global test for ALS trials, and retrospectively applied the Finkelstein-Schoenfeld test to a phase II trial for ALS.

We propose a generalization of the aforementioned global nonparametric rank tests using U-statistics. The class of tests can be applied to settings that involve continuous, ordinal, and censored endpoints. While some of the tests that we consider in this paper have been proposed and examined in the literature, we will generalize the expression for rank-based tests of combined endpoints. The advantage of a broader generalization of these tests is that the properties of any particular test can be readily developed using the infrastructure we provide. This allows investigators the flexibility to choose an existing test (e.g. O'Brien), a weighted or modified version of an existing test, or even create a new test that may be more suitable to a different notion of treatment efficacy within a particular study. Additionally, we determine the optimal outcome weights for certain tests, and propose a novel adaptive weighting method that can be used to improve power over the ordinary global tests.

In section 2, we will describe the general test statistic and its properties under the null hypothesis. Section 3 will focus on the choice of optimal outcome weights for specific tests, including a description of the adaptive procedure for estimating weights. We will present simulation results in section 4, and an example analysis of an ALS clinical trial in section 5.

We will close by discussing the merits and drawbacks of such combined tests, and the implications in interpreting results.

2. Methods

Suppose we have two groups of patients on different treatments, and we are interested in testing a hypothesis about the efficacy of one treatment versus the other when there are multiple outcomes that have been recorded for each patient. First, we will score all pairs of patients between groups with respect to each outcome, with a score between -1 and 1. For example, if we are comparing patients *i* and *j* on survival and a quantitative outcome (e.g. ALSFRS-R score), for the pair (*i,j*) we may assign a score of -1 for survival if subject *i* failed before patient *j* (1 if *j* failed before *i*). For ALSFRS-R, we would assign a score of 1 if *i* had a higher score than patient *j* at their last common follow-up time (-1 if *i* had a lower score). Generally, for each outcome, indexed by *k*, we have a function r_k that takes data from both subjects and assigns a score of -1, 0, or 1. This function should indicate which patient did better with respect to the k^{th} outcome, with a value of 1 indicating a better outcome for subject *i* over *j*, -1 a worse outcome, and 0 the same. We will call this a pairwise rank. Note that in this example we compare *i* and *j* on ALSFRS-R at their last common follow-up time, but in reality we may want to use a different measure that accounts for some pre-treatment baseline measurement of ALSFRS-R, such as percent change or slope. The main idea is that due to censoring and death, we can only validly compare patients up to their last common follow-up time; the measure that we use should make sense within the context of the illness.

In general, let $\mathbf{x}_{ik}$, $\mathbf{y}_{jk}$ represent observed data on subjects *i* and *j* for outcome *k*, where $\mathbf{x}_{ik}$, $\mathbf{y}_{jk}$ can possibly be vectors, *i* indexes subjects on treatment ($i = 1, \ldots, n$), *j* indexes control subjects ($j = 1, \ldots, m$), and *k* indexes the outcomes ($k = 1, \ldots, p$). We assume that the complete vetor of outcome random variables $\mathbf{X}_i$, and $\mathbf{Y}_j$ are i.i.d. with respective distribution functions $F_X(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ and $F_Y(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p)$.

Suppose, for example, that x_{ik}, y_{jk} are scalar observed outcomes where a larger value is favorable; then we would write the ranking function for that outcome as $r_k(x_{ik}, y_{jk}) = I(x_{ik} > y_{jk}) - I(x_{ik} < y_{jk})$. In the case of a failure time, we will use the Gehan scoring function (1965) to score pairs. For example, let X'_{ik} and Y'_{jk} denote the follow-up time random variables for subjects *i* and *j* on outcome *k* (i.e. $X'_{ik}=\min(X_{ik}, C_i)$, where X_{ik}, C_i are the failure and censoring time random variables for subject *i*; $Y'_{jk}=\min(Y_{jk}, C_j)$ analogously), and let δ_{ik}, δ_{jk} be the indicator variables that a failure was observed. Then we have $r_k((x'_{ik}, \delta_{ik}), (y'_{jk}, \delta_{jk}))=I(x'_{ik} \geq y'_{jk})\delta_{jk} - I(x'_{ik} \leq y'_{jk})\delta_{ik}$. This will be equal to 1 if subject *i* is known to have survived longer than subject *j*, -1 if *i* is known to fail before *j*, and 0 if tied or it is indeterminate who survived longer. We will denote $E[r_k(\mathbf{x}, \mathbf{y})] = \theta_k$. This θ_k can be thought of as a marginal treatment effect for outcome *k*, where a positive value favors the treated group. Note that in the expression $r_k(\mathbf{x}, \mathbf{y})$, *x* and *y* may be vectors of data, as in the Gehan scoring function.

Now define $\mathbf{r}_{ij} = (r_1(\mathbf{x}_{i1}, \mathbf{y}_{j1}), r_2(\mathbf{x}_{i2}, \mathbf{y}_{j2}), \ldots, r_p(\mathbf{x}_{ip}, \mathbf{y}_{jp}))$. This is the vector of the scores comparing subject *i* to subject *j* on each of the *p* outcomes. The vector $\mathbf{r}_{ij} = (-1, 1, 0)$, for

example, would indicate subject i did worse than j on the first outcome, better on the second outcome, and the same or indeterminate on the third.

Once we have the vector $\mathbf{r}_{ij}$ for each pair i and j between different groups, we map it to a one-dimensional score, and then construct a test statistic based on the univariate scores for each pair of subjects. That is, we will have a function $\phi(r_1, .., r_p)$ that maps the vector of pairwise outcome scores to a single summary score. The univariate score resulting from $\phi(\mathbf{r}_{ij})$ is interpreted as a summary measure of the differences in outcomes between subjects i and j. A positive score favors subject i, a negative score subject j, and 0 favors neither.

The test statistic is given by the sum of the composite pairwise scores between the two groups:

$$U = \frac{1}{nm}\sum_{i}^{n}\sum_{j}^{m}\phi(r_{ij}) \quad (1)$$

This is simply a two-sample U-statistic that estimates the parameter $\theta_\phi = E[\phi(r_1(X_1, Y_1), \ldots, r_p(X_p, Y_p))]$. Borrowing terminology from Huang (Huang, Woolson, and O'Brien, 2008), we can think of θ_ϕ as a *global treatment effect*. It is the expectation of the composite of outcome-specific pairwise ranks, where each pairwise rank is a scaled probability between -1 and 1 that i did better than j on that outcome. Thus, it can be interpreted as something like a scaled probability of doing "better" on treatment, "better" being defined by how we summarize pairs with the function ϕ. Note that in this paper we construct the statistic so that $\theta_\phi = 0$ under the null hypothesis H_0.

2.1 Some Examples for ϕ

Below we will give examples for composite functions ϕ for some tests previously proposed in the literature. For ease of notation, we will denote the outcome-specific rank scores $r_k(\mathbf{x}_{ik}, \mathbf{y}_{ik}) = r_k$.

1. O'Brien (1984). O'Brien's proposed nonparametric procedure for comparing multiple outcomes was based on an overall rank for each subject that is obtained by summing their outcome-specific ranks, and using a rank-sum or ANOVA test based on the overall ranks. A function ϕ that would yield a test similar to O'Brien's is $\phi(r_1, \ldots, r_p) = r_1 + r_2 + \cdots + r_p$. More generally, we could weight the outcomes differently, and have $\phi(r_1, \ldots r_p) = w_1 r_1 + w_2 r_2 + \cdots + w_p r_p$, with $w_k \geq 0$ for all k.
2. Finkelstein-Schoenfeld Test (FS) (1999). This test compares a mortality outcome and a longitudinal outcome in a hierarchy, where subjects are first compared pairwise on survival, and then on the longitudinal marker if it is indeterminate who survived longer. Here r_1 is the Gehan scoring function, and r_2 ranks pairs of subjects on their longitudinal outcome at their last common follow-up time. In our framework, the function ϕ is given by $\phi(r_1, r_2) = r_1 + I(r_1 = 0)r_2$. For p outcomes arranged in a hierarchy

(Buyse, 2010), we would have $\phi(r_1, r_2, \ldots, r_p) = r_1 + I(r_1 = 0)r_2 + \ldots + I(r_1 = \cdots = r_{p-1} = 0)r_p$. We could also assign a different weight to each outcome with $\phi(r_1, r_2, \ldots, r_p) = w_1 r_1 + I(r_1 = 0)w_2 r_2 + \ldots + I(r_1 = \cdots = r_{p-1} = 0)w_p r_p$, with $w_k \geq 0$ for all k. With censored data, when there is only administrative censoring at the end of the study period, but no dropout during the study period, this is equivalent to using "worst-rank" scores (Wittes, Lakatos, and Probstfield, 1989).

3. Wittkowski (2004). Wittkowski's proposal compares subjects pairwise with respect to several ordinal measures. When all of the outcomes for subject i are at least as favorable as that of the subject j, and at least one of subject i's outcomes is more favorable, a score of 1 is assigned for the pair (-1 if subject j does better). If some outcomes are better and some are worse in the pairwise comparison, the score is 0. For ϕ, we can write $\phi(r_1, \ldots, r_p) = I(max_k\{r_k : k = 1, \ldots, p\} > 0) - I(min_k\{r_k : k = 1, \ldots, p\} < 0)$. This could be modified to score a 1 if subject 1 has more favorable outcomes than subject 2: $(\phi(r_1, \ldots, r_p) = I(\Sigma_k r_k > 0) - I(\Sigma_k r_k < 0)$. This can further be modified with weights: $(\phi(r_1, \ldots, r_m) = I(\Sigma_k w_k r_k > 0) - I(\Sigma_k w_k r_k < 0)$, with $wk > 0$ for all k.

4. Combination of different tests: To illustrate the flexibility of the test, we can also use a combination of other tests. For example, a ϕ function that combines elements of the O'Brien and FS tests could be $\phi(r_1, \ldots, r_p) = r_1 + I(r_1 = 0)\frac{1}{p-1}\sum_{k=2}^{p} r_k$. This function gives a composite score based on the the first outcome, but if the first outcome is tied, the composite score is an average of the scores for all other outcomes.

We will mainly focus on the O'Brien and FS tests in this paper, but the large-sample properties of the test hold for any appropriate function ϕ.

2.2 The Null Hypothesis and Restrictions on ϕ

The null hypothesis with which we are working is that the global treatment effect $\theta_\phi = 0$, and whenever this is the case, the test statistic will have mean 0 and should reject the null at the nominal α level. For each test described above, $\theta_\phi = 0$ holds under the strongest null hypothesis that the joint distributions in each group are the same, but it can also hold under weaker conditions. For example, with uncensored data using O'Brien's test, $\theta_\phi = 0$ when $\sum_k^p P(X_k > Y_k) - P(X_k < Y_k) = 0$. This is essentially equivalent to the null hypothesis for the modification of O'Brien's test proposed by Huang et al. (2005).

The following conditions on ϕ will always ensure a valid test under the strong null that the joint distributions of the outcomes are equal between both groups.

1. $\phi(\mathbf{0}) = 0$.
2. ϕ is an odd function, i.e. $\phi(\mathbf{r}_{ij}) = -\phi(\mathbf{r}_{ji})$. Then $\phi(\mathbf{r}_{ij}) + \phi(\mathbf{r}_{ji}) = 0$
3. $E[\phi^2(r_1(X_1, Y_1), \ldots, r_p(X_p, Y_p))] < \infty$

The first two conditions ensure that the composite scores will only differ by sign if we flip the arguments of the $r_k(\cdot, \cdot)$. By symmetry, $\theta_\phi = E[\phi(\mathbf{r}_{ij})] = 0$ under the strong null, and the test statistic will have mean 0 when this is the case. Let $N = n + m$ be the total sample size. Under H_0, when the third condition holds and $\frac{n}{N} \to \lambda$ as $N \to \infty$, it follows that $\sqrt{N}U \to N(0, \sigma^2)$, where

$$\sigma^2 = \frac{1}{\lambda} E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})] + \frac{1}{1-\lambda} E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{i'j})] \quad (2)$$

This follows from standard asymptotic theory on U-statistics (Van der Vaart, 2000). The asymptotic variance is not distribution free under H_0, as it will generally depend on the correlation between the scores among different outcomes, but can be consistently estimated from the data with:

$$\hat{\sigma}^2 = \frac{N}{(nm)^2} \left[\sum_i^n \sum_j^m \sum_{j' \neq j}^m \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'}) + \sum_i^n \sum_{i' \neq i}^n \sum_j^m \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{i'j}) \right]. \quad (3)$$

(see Web Appendix A for details).

If we have stratified data, a stratified test statistic is given by $T = \frac{\sum_{s=1}^S \sqrt{N_s} U_s}{\sqrt{\sum_{s=1}^S \hat{\sigma}_s^2}}$, where S is the total number of strata, and for the s^{th} stratum N_s is the total sample size, U_s is calculated as in (1), and $\hat{\sigma}_s^2$ is estimated as in (3). T has an asymptotic standard normal distribution, but note that the asymptotic distribution is based on the asymptotic normality of the within-strata U-statistics, which may not hold if some of the strata have very small sample sizes per treatment group.

2.3 Power and Sample Size Considerations

For a given function ϕ, probability of type 1 and type 2 errors α and β respectively, and global treatment effect $\theta_\phi > 0$ under the alternative hypothesis H_1, the power of the test can be approximated by $1 - \beta \approx 1 - \Phi(z_{1-\alpha/2} - \frac{\sqrt{N}\theta_\phi}{\sigma})$, where Φ is the standard normal cumulative distribution function, $z_{1-\alpha/2}$ is the minimum upper tail value for which we would reject H_0, and σ is the standard deviation of the U-statistic as given in (2). Then for a given power $1 - \beta$, an estimated total sample size is given by $N = \left[\frac{\sigma(z_{1-\alpha/2} - z_\beta)}{\theta_\phi} \right]^2$. It follows that $n = \lambda N$ and $m = (1 - \lambda)N$. Note that to find candidate values for θ_ϕ and σ, we would need to make some distributional assumptions on the data, and obtain the parameters analytically or by simulation. As Huang, Woolson, and O'Brien note (2008), this has no bearing on the test statistic itself, for which we do not make any parametric assumptions.

In the next section, we will show that we can write the O'Brien and Finkelstein-Schoenfeld tests as a sum of outcome-specific U-statistics, $U_1, \ldots, U_p$. Then we can construct a weighted global test of the form $\mathbf{w}'\mathbf{U}$ where $\mathbf{w}$ is a vector of weights. For these weighted tests, we can rewrite the power function in terms of the weighted component U-statistics. Let $\mathbf{U} = (U_1, \ldots, U_p)'$ be the vector of outcome-specific U-statistics, $\boldsymbol{\Lambda} = cov(\mathbf{U})$, $\boldsymbol{\theta}_\phi = (\theta_{\phi 1}, \ldots, \theta_{\phi p})' = E(\mathbf{U})$ under H_1, and $\mathbf{w} = (w_1, \ldots, w_p)'$ be a fixed weighting vector. Without loss of generality, assume $\boldsymbol{\theta}_\phi \geq 0$ in all components. We assume this because we are only interested in alternatives where the treatment is favorable on at least some outcomes, and not unfavorable on any outcomes (equivalently, we can assume $\boldsymbol{\theta}_\phi \leq 0$). Then the power of the test is given by $1 - \beta \approx 1 - \Phi(z_{1-\alpha/2} - \frac{\sqrt{N}\mathbf{w}'\boldsymbol{\theta}_\phi}{\sqrt{\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w}}})$. For optimal weights, it follows that maximizing power corresponds to maximizing $\mathbf{w}'\boldsymbol{\theta}_\phi(\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w})^{-1/2}$ with respect to $\mathbf{w}$. Note that if we assumed $\boldsymbol{\theta}_\phi \leq 0$, maximizing power corresponds to minimizing this quantity. The total sample size for given β is then $N = \mathbf{w}'\boldsymbol{\Lambda}\mathbf{w}\left[\frac{z_{1-\alpha/2} - z_\beta}{\mathbf{w}'\boldsymbol{\theta}_\phi}\right]^2$.

As a guide to choosing a particular test, one can compute the estimated power for different tests under a range of distributional assumptions and alternative hypotheses.

3. Weights

Incorporating outcome weights allows the relative importance of the outcomes to be reflected in the test. For example, in some cases the treatment may be most targeted to improving mortality, while in other cases death may be a competing risk. Weights would allow us to easily cast our statistic in terms of these different settings.

One method for choosing weights would be to base it on the importance of outcomes. These utility weights are completely determined by the investigator prior to the study. For example, in a study of ALS and survival, the rank on survival may get a larger weight than the rank on ALSFRS-R score because survival is more important. One problem with utility weights is that utility of certain outcomes may be different for different subjects, and can be arbitrarily chosen based on investigator belief. On the other hand, this may be attractive when there is a clear subset of outcomes that should dominate the statistic.

An alternative method would be to construct optimal weights by maximizing the power of of our test statistic under a particular alternative hypothesis. We can do this for both the O'Brien and the Finkelstein-Schoenfeld tests, which we describe below.

3.1 O'Brien

For O'Brien's test, note that $\boldsymbol{\phi}$ is a linear function of the individual outcome scores, so we can write the test a sum of U-statistics for each outcome, as described by Li et al. (2009). First, let $U_k = \frac{1}{nm}\sum_i^n\sum_j^m r_k(\mathbf{X}_{ik}, \mathbf{Y}_{jk})$, the U-statistic for the k^{th} outcome. The weighted O'Brien statistic is then given by $\mathbf{w}'\mathbf{U}$ where $\mathbf{w}$ is a weighting vector. Since $\sqrt{N}U_k \rightarrow N(0, \sigma_k^2)$, it follows that $\sqrt{N}\Sigma_k U_k \rightarrow N(0, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = cov(\mathbf{U})$. Then $\sqrt{N}\mathbf{w}'\mathbf{U} \rightarrow N(0, \mathbf{w}'\boldsymbol{\Lambda}\mathbf{w})$. As

noted earlier, maximizing power is equivalent to maximizing $|\mathbf{w}'\boldsymbol{\theta}|(\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w})^{-1/2}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' = (E[U_1], \ldots, E[U_p])'$. The solution to this equation is $\mathbf{w} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}$ (see Web Appendix B). We would need to choose $\boldsymbol{\theta}$ a priori under a specific alternative hypothesis we have in mind. We will assume that $\theta_k > 0$ (or $\theta_k < 0$) for all k, since these are the alternative hypotheses in which we are interested. For any distribution functions we assume on the data, we can always approximate the desired $\boldsymbol{\theta}$ by simulation, and in many cases we can solve for it analytically. For the purpose of selecting weights a priori, the covariance matrix $\boldsymbol{\Lambda}$ should also be obtained using a combination of historical data and hypothesized treatment effects. If there were no historical data available, we would have to make some distributional assumptions on the data, and then arrive at the covariance matrix analytically or through simulation. By simulation, we can calculate each of the outcome-specific U-statistics several times under specific distributional assumptions, and compute the empirical covariance matrix of the U-statistic vectors as $\boldsymbol{\Lambda}$. This can easily be done for a variety of assumptions to get a better idea of reasonable candidates for $\boldsymbol{\Lambda}$.

For computation of the test statistic, the covariance matrix $\boldsymbol{\Lambda}$ has entries $\sigma_{k,l} = cov(U_k, U_l)$, which can be estimated with:

$$\hat{\sigma}_{k,l} = \frac{N}{(nm)^2}\left[\sum_{i}^{n}\sum_{i\neq i'}^{n}\sum_{j}^{m} r_k(\mathbf{X}_{ik}, \mathbf{Y}_{jk}) r_l(\mathbf{X}_{i'l}, \mathbf{Y}_{jl}) + \sum_{i}^{n}\sum_{j}^{m}\sum_{j'\neq j}^{m} r_k(\mathbf{X}_{ik}, \mathbf{Y}_{jk}) r_l(\mathbf{X}_{il}, \mathbf{Y}_{j'l})\right].$$

Note that this variance estimate, based on the current trial, should only be used for computation of the test statistic, not for obtaining optimal weights.

3.2 Finkelstein-Schoenfeld (FS)

To find optimal weights for the FS test, we will again write the test a sum of dependent U-statistics. Suppose that the first, and most important outcome is a failure time. Let X_{i1}, Y_{j1} denote the follow-up times on this outcome for subjects i (group 1) and j (group 2). Let δ_{i1}, δ_{j1} be the indicator that a failure was observed for i and j respectively. Let $r_{ij1} = I(X_{i1} > Y_{j1})\delta_{j1} - I(X_{i1} < Y_{j1})\delta_{i1}$ be the pairwise Gehan rank for the first outcome, and in general let $r_{ijk} = I(X_{ik} > Y_{ik}) - I(X_{ik} < Y_{ik})$ be the pairwise rank for subject i vs. subject j on outcome k. Note that these ranks can also be Gehan ranks on failure and censoring times with their own δ values, but we suppress the notation for generality. Also, the non-survival outcome(s) will not be able to be measured on a subject after he or she fails or is censored, so subjects can be compared on the other outcomes based on their last common follow-up time. Now, define $e_{ij1} = 1$ and $e_{ijk} = I(r_{ij1} = 0, r_{ij2} = 0, \ldots, r_{ij,k-1} = 0)$ for $k \geq 2$. Then the test statistic is given by $\sum_k^p U_k$, where $U_k = \frac{1}{nm}\sum_i^n \sum_j^m e_{ijk} r_{ijk}$

As before $\sqrt{N}\mathbf{w}'\mathbf{U} \rightarrow N(0, \mathbf{w}'\boldsymbol{\Lambda}\mathbf{w})$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' = (E[U_1], \ldots, E[U_p])'$. The optimal weight is given by $\mathbf{w} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}$. As described in the previous section, $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ should be determined a priori for the purpose of selecting weights. For computing the test statistic, the estimate $\hat{\boldsymbol{\Lambda}}$ for $\boldsymbol{\Lambda}$ has entries

$$\hat{\sigma}_{k,l}=\frac{N}{(nm)^2}\left[\sum_{i}^{n}\sum_{i\neq i'}^{n}\sum_{j}^{m}e_{ijk}r_{ijk}e_{i'jl}r_{i'jl}+\sum_{i}^{n}\sum_{j}^{m}\sum_{j'\neq j}^{m}e_{ijk}r_{ijk}e_{ij'l}r_{ij'l}\right].$$

3.3 Optimal Weighting and Constrained Optimization

The optimal solution for O'Brien's test and the FS test can yield undesirable weights from a clinical standpoint, particularly the case where some weights are positive and others negative. At the most basic level, we want all of our weights to be positive, but we may also want to restrict certain outcomes to have some fixed, minimum, or maximum weight. This can be achieved fairly easily using a constrained optimization. Ultimately, to maximize power, we want to maximize the quantity $\delta = |\mathbf{w}'\boldsymbol{\theta}_\phi(\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w})^{-1/2}$ with respect to the vector $\mathbf{w}$. This quantity can be maximized using the *optim* function in *R* (R Core Team, 2014), and box-constraints can be put on each element of the vector $\mathbf{w}$ when using the "L-BFGS-B" method of optimization (Byrd et al., 1995). The box-constraints are simply lower and upper bounds specified for each element of the vector (which can be ∞ or $-\infty$ for upper and lower bounds, respectively). If we want any outcomes to have some fixed weight, we can do that as well. Suppose we want to fix the first outcome weight to be 1; we could still optimize the same quantity δ, but in the function simply set $w_1 = 1$, and maximize the quantity with respect to the vector of weights (w_2, …, w_p). The investigator should use their discretion to determine whether selected weights are sensible given the nature of the illness, outcomes, and treatment under study.

3.4 Adaptive Weighting

The biggest issue with attempting to use optimal weights as described above is that we need to have an idea of the parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ under the alternative hypothesis for the weights to be useful in improving power. This may be viable if we have previous studies for which we can estimate those parameters, but in general they are unknown. An adaptive weighting method can be used to avoid guessing weights prior to the study when we have multiple strata. Natural strata are frequently present in medical studies, e.g. different enrollment periods and/or centers in clinical trials. In such settings, we propose using data from "previous" strata to estimate weights for "upcoming" strata. Fisher (1998) describes the general idea, and shows that adapting weights in this manner maintains the significance level of the trial. An adaptive weighting scheme can be constructed as follows.

1. Suppose we have p outcomes and S strata. Order the strata 1, …, S. This could be a natural ordering based on the design of the study (e.g. enrollment period), or a random ordering. Let U_{sk} denote the k^{th} component U-statistic for the s^{th} stratum.

2. In the first stratum, calculate the outcome specific test statistics U_{1k}, $k = 1$, …, p as described in section 3.1 or 3.2 for the appropriate test. U_{1k} is then an estimate of $\theta_k = E[U_{sk}]$ for the subsequent strata, and $\mathbf{U}_1 = (U_{11}, \ldots, U_{1p})'$ is an estimate of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$. Estimate the covariance matrix for the first stratum, $\hat{\boldsymbol{\Lambda}}_1$.

3. Estimate the optimal weights for the second stratum with $\mathbf{w}_2=\hat{\mathbf{\Lambda}}_1^{-1}\mathbf{U}_1$ (or, if this yields negative weights, numerically optimize with constraints). Scale the weights $\mathbf{w}_2$ such that its components sum to 1, i.e. $\sum_{k=1}^{p} w_{2k}=1$. Then the numerator of the statistic for the second stratum is $\mathbf{w}_2'\mathbf{U}_2$, and the variance is $\sigma_2^2=\mathbf{w}_2'\mathbf{\Lambda}_2\mathbf{w}_2$.

4. For all of the subsequent strata, we may use a weighted average of the requisite parameters from the previous strata, weighted by the number of pairwise comparisons ineach of the strata. That is, our estimate for $\boldsymbol{\theta}$ and $\mathbf{\Lambda}$ for the s^{th} stratum are given by $\hat{\boldsymbol{\theta}}_s=\frac{1}{\sum_{j=1}^{s-1} n_j m_j}\sum_{j=1}^{s-1} n_j m_j \mathbf{U_j}$, and $\hat{\sum}_s=\frac{1}{\sum_{j=1}^{s-1} n_j m_j}\sum_{j=1}^{s-1} n_j m_j \hat{\mathbf{\Lambda}}_j$, respectively. The optimal weight for the s^{th} stratum is then $\mathbf{w}_s=\hat{\sum}_s^{-1}\hat{\boldsymbol{\theta}}_s$. Note that $\hat{\mathbf{\Sigma}}$ is used in place of $\hat{\mathbf{\Lambda}}$ here to avoid confusion between the within stratum estimates of the covariance ($\hat{\mathbf{\Lambda}}$) and the average of those covariances across strata ($\hat{\mathbf{\Sigma}}$).

5. Combine the stratum-specific test statistics using a stratified statistic, as described insection 2.2.

This is a general outline, but there can be many variations on the above procedure. For example, the stratified statistic given in section 2.2 weights each of the strata equally in the overall test statistic, so a further modification can be to give different strata different weights, perhaps to upweight the strata that uses more previous information.

Alternatively, one can use Bayesian methods by setting a prior on the weights, and updating the weights with additional data. Minas et al. (2012) use a type of Bayesian method to estimate weights in the case of multivariate normal data, basing the priors on previous studies, and computing the posterior with a subset of pilot data taken from the main study data. Something similar to the above procedure can potentially fit within a group-sequential design framework as well.

The weights used for the first stratum can all be equal, or they can be estimated from historical data or simulation based on a hypothesized treatment difference between groups. In addition, the ordering of the strata should be pre-specified, as the value of the test statistic will depend on the order. A natural ordering could be based on the sample size of each stratum, or could be chronological if the strata are distinguished by enrollment period.

The main advantage of this procedure is that we are letting the data self-select the weights based on what outcomes the treatment is affecting most. A disadvantage is that we are using different outcome weights for different strata, so interpretation of the pooled stratified test becomes muddled. In addition, if we get the wrong weights we can lose power. This is more likely to happen when equal weights are already near optimal, causing us to estimate sub-optimal weights due to the variability in estimation. With censored data, there is greater

variability in weight estimation, as the optimal weights will also depend on the censoring distributions. Furthermore, the above procedure assumes the same treatment effect across strata, and thus may give sub-optimal weights when this is not the case.

4. Simulations

We assessed the performance of the O'Brien and FS tests, and their adaptively weighted counterparts, under two different scenarios. In the first scenario, we generate uncensored data on 4 outcomes, and compare the type 1 error of O'Brien's originally proposed nonparametric test (denoted T_O), our proposed version of O'Brien's test with equal weights (T_O^U), and our proposed Adaptive O'Brien test ($T_O^{U,\mathrm{Ad}}$). We also compare the power of these tests with the optimally weighted O'Brien test ($T_O^{U,\mathrm{Opt}}$). In the second scenario, we generate data based on an ALS simulation study by Healy and Schoenfeld (2012). For this scenario, we compare the type 1 errors of the proposed O'Brien, Adaptive O'Brien, FS (T_{FS}^U), and Adaptive FS ($T_{\mathrm{FS}}^{U,\mathrm{Ad}}$) tests. Additionally, we compare the power of these tests and the optimally weighted O'Brien($T_O^{U,\mathrm{Opt}}$) and $\mathrm{FS}(T_{\mathrm{FS}}^{U,\mathrm{Opt}})$ tests. To determine the appropriate covariance matrix $\mathbf{\Lambda}$ for optimal weight estimation, we did an independent simulation under the assumed distributions, and empirically estimated $\mathbf{\Lambda}$ from the U-Statistic vectors. In each setting, we generated 2 or 4 strata with no treatment by strata interaction, and used the stratified test statistic given in section 2.2. Note that for O'Brien's original test, there is no stratified statistic, so the T_O statistic is based on the full sample irrespective of the strata. For each setting, 5000 iterations were performed.

4.1 Scenario 1: Four outcomes, uncensored

To test the performance of O'Brien's test under the null hypothesis, we generated data from a multivariate normal distribution with four outcomes and zero mean for all outcomes, under both equal and unequal variances between the groups. In the equal variances setting, all outcomes had variance 1, and all correlations between outcomes were set to ρ, with the value of ρ for each setting given in Table 1. For unequal variances, the covariance matrix for group 1 was equal to 1 on the diagonals, and all off-diagonal entries were 0, indicating no correlation between outcomes. The covariance matrix for group 2 was set to (1, 4, 9, 25) on the diagonal, and all off-diagonal entries were set to 1. In Table 1, we see that when the multivariate distributions for both groups are equal, i.e. when the within group variances are equal, that T_O, T_O^U, and $T_O^{U,\mathrm{Ad}}$ all control the type 1 error at the nominal 0.05 level, including under unequal sample sizes. Under unequal variances, however, the type I error for T_O is inflated, while the type I errors for the proposed T_O^U and $T_O^{U,\mathrm{Ad}}$ statistics are still controlled at the nominal level. This was the same conclusion drawn by Huang et al. (2005) for O'Brien's original test.

Under the alternative hypothesis, we similarly generated multivariate normal data, using the same covariance matrix as the “equal variances” scenario under the null hypothesis above for both groups. The mean for each outcome was zero in group 2, and in group 1 the means were (.053, .142, .286, .507), chosen so that θ = (.03, .08, .16, .28). The results are given in

Table 2. With no correlation between the outcomes, the adaptive test $T_O^{U,\mathrm{Ad}}$ Performs similarly to the unweighted tests T_O and T_O^U, while the optimally weighted test $T_O^{U,\mathrm{Opt}}$ has significantly higher power. As the correlation (ρ) between outcomes increases, we see that the adaptive test begins to perform better than the unweighted tests, and the optimally weighted tests have an even greater power increase over the unweighted tests. This is because when the outcomes are correlated, it becomes more optimal to lower the weight on the outcomes with a smaller effect size to diminish the additional variance obtained from adding the correlated ranks between the outcomes. As the optimal weights become further away from equal, the adaptive test gains significantly more power than its unweighted counterparts. This also illustrates that whenever two outcomes are strongly correlated, we may be better off dropping one of those outcomes entirely from the statistic.

4.2 Scenario 2: Survival and Neurological Function

In this scenario, we generate data based on a clinical trial where patients are monitored for two outcomes: survival, and ALSFRS-R scores. The ALSFRS-R is a functional rating scale by which physicians evaluate the degree of neurological function in ALS patients. For every subject, we generated ALSFRS-R data for 25 time points, (0, 1, …, 24), where each time can be thought of as a month. We also generated survival times, subject to equal and unequal censoring distributions between groups in different scenarios. For the equal censoring case, we used administrative censoring in both groups at time 24. Under unequal censoring, one group had only administrative censoring at time 24, while the other group was subject to administrative censoring at time 24 or random censoring before time 24, generated from a uniform distribution.

The simulation is nearly identical to a simulation study by Healy and Schoenfeld (2012) for ALS, so we refer to their paper for details, and include a description of the model in Web Appendix C. They generated the data from a shared parameter model, where survival was correlated with ALSFRS-R trajectory through patient-specific random effects. The parameters for their model were derived from estimation of the model for data from an ALS clinical trial (Cudkowicz et al., 2006), and they varied the treatment effects for ALSFRS and survival across simulations.

In Table 3, we present results for our version of the O'Brien and FS tests, and their adaptive counterparts, under no treatment effect on ALSFRS or survival. Each test controls the type I error at the nominal level for equal and unequal censoring distributions, including under unequal sample sizes. As O'Brien's originally proposed test was not constructed for censored data, we did not assess its performance in this scenario.

In Table 4, we present power under the alternative hypothesis for the O'Brien and FS tests, and their adaptive and optimally weighted counterparts. Data was generated under different combinations of effect sizes for mortality and ALS (none, mild, moderate, strong) under the shared parameter model. In general, the T_{FS}^U test performs slightly better than T_O^U when there is a stronger treatment effect on mortality, while T_O^U performs better with a stronger effect on ALS. Additionally, the adaptive tests $T_O^{U,\mathrm{Ad}}$ and $T_{\mathrm{FS}}^{U,\mathrm{Ad}}$ perform better than the unweighted

tests when the treatment effect sizes are very different between mortality and ALS, where the optimal weights are far from equal. However, in cases where equal weights are already close to optimal (e.g. moderate effect sizes on both outcomes), the adaptive tests will do worse than the unweighted tests. In this scenario, the adaptive tests seem to be most useful as a hedge in case one of the outcomes is null or close to null, in order to minimize dilution of the statistic from combining noise from a null outcome with signal from another outcome.

5. Example: ALS Trial

We illustrate the proposed O'Brien and FS tests on data from a clinical trial of Ceftriaxone in patients with ALS (Berry et al., 2013). The 513 subjects in the trial were monitored for two endpoints: survival, and rate of decline in neurological function as measured by their ALSFRS-R scores. The scale ranges from 0-48, with a higher score indicating better function. ALSFRS-R was measured periodically in patients until death, drop-out, or the end of the study. 340 subjects were administered Ceftriaxone, and 173 placebo, with an average follow-up time of 1.6 years. We compared treatments using the stratified test statistic, with the stratum variable being site of onset (“limb-onset” or “bulbar-onset”). There were 119 subjects with bulbar-onset and 394 with limb-onset disease. We used Gehan ranks for the survival outcome, and for the ALS outcome, we compared patients pairwise on the mean of their ALSFRS-R scores up to their last common follow-up time. The component U-statistics (normalized by $\sqrt{N}$) for O'Brien's test were (1.37,0.08) in the bulbar-onset stratum and (0.18, −0.56) in the limb-onset stratum, where the first component refers to survival and the second ALSFRS-R; for the FS tests these were (1.37, −.04) and (0.18, −0.36). The estimated covariance matrices in each stratum for O'Brien's test were $\hat{\Lambda}_1 = \begin{pmatrix} .42 & .007 \\ .007 & 1.43 \end{pmatrix}$ and $\hat{\Lambda}_2 = \begin{pmatrix} .43 & .007 \\ .007 & 1.39 \end{pmatrix}$. For the FS test we had, $\hat{\Lambda}_1 = \begin{pmatrix} .42 & -.02 \\ -.02 & .11 \end{pmatrix}$ and $\hat{\Lambda}_2 = \begin{pmatrix} .43 & .003 \\ .003 & .174 \end{pmatrix}$. The normalized test statistics were 0.56 for the O'Brien test (p-value = .577), and 1.09 for the FS test (p-value = 0.275). Notice here that the FS test, which puts more emphasis on survival, is more robust to the weak treament effect in the opposite direction on the ALS outcome.

We also computed the test statistic using the adaptive method described in section 3.4. We first computed the statistics above, then estimated optimal weights for the “limb-onset stratum” using data from the “bulbar-onset” stratum. The optimal weights (restricted to be non-negative) for both tests were (1,0), i.e. with only weight on the survival outcome. The normalized adaptive test statistics were 0.96 (p-value = .340) for the O'Brien test, and 1.14 (p-value = .256) for the FS test. Observe that because the ALSFRS-R outcome is given zero weight in the second stratum, the adaptive statistics, especially O'Brien's, are less diluted by that outcome. This could be problematic, however, if treatment is actually better in one outcome and worse in another, because we would not want to erroneously conclude a positive global treatment effect in that case. This example illustrates well how the decomposition of each statistic and its variance into a weighted sum of it's components gives

us a sense of which outcomes are contributing the most and the least to the test statistic, and in which direction.

6. Discussion

We have generalized previously proposed nonparametric tests that use different methods to rank multivariate outcomes. For both uncensored and censored data, the generalization creates a class of valid tests under the null hypothesis that the two groups have the same joint distribution of outcomes, though for some tests a weaker null will suffice. For uncensored data, the proposed O'Brien test and it's weighted counterparts are valid under the Behrens-Fisher hypothesis as described by Huang (2005). With censored outcomes, the tests are valid under unequal censoring distributions between groups. The tests are also valid under unequal sample sizes.

This unified framework allows the investigator the flexibility to choose a test that fits the purposes of their study without making distributional assumptions on the data. The generalization allows for an easily estimable variance for each method, and the ability to compare the global treatment effect size and variance among different methods, which have implications in the power and sample size of the test. We have also provided a method for determining optimal weights for O'Brien's test and the FS test under a specified alternative hypothesis. Since in practice we do not know the necessary parameters to obtain the optimal weights, we have proposed an adaptive weighting method that incorporates data-driven weights. Simulations indicate that the type I error holds in the adaptive case, and that power can improve significantly in settings with differing treatment effect sizes or moderate correlation between outcomes.

When the outcome weights for these tests are based on achieving maximal power using a priori assumed treatment effects, or selected adaptively, the tests may be more difficult to interpret. With any outcome-weighting vector (even equal weights), we are projecting the vector of treatment effects onto a single dimension. By restricting all of the weights to be positive, the summary statistic represents treatment efficacy on that one dimensional space. The same is true for equal weights, except in that case each outcome contributes a similar amount to the overall treatment effect. With adaptive weights, there is the added dimension of different weighting in different strata, but the statistic still constitutes a measure of efficacy as long as all weights are positive. And from a strictly statistical standpoint, the numerator of the statistic measures the *global treatment effect* described in section 2, the expected value of the composite of pairwise ranks.

Of course, with this kind of dimension reduction, we can miss important red flags if we are not careful. For example, suppose treatment is actually harmful on survival, but that we put very little weight on survival based on an a priori hypothesis. If the treatment is strongly beneficial in the other outcomes, we could reject the null in favor of treatment despite its negative effect on survival. This type of issue can be avoided. As described earlier, investigators could restrict a subset of outcomes to have some minimal weight or fixed weight if they do not want them to be overridden in the global test. In general, because of the

loss of information that occurs with the dimension reduction, care should be taken to ensure that the most important outcomes are amply contributing to the test statistic.

Further, we recommend not completely divorcing the global tests from considering each outcome individually, whether through providing summary statistics or additional statistical testing. The multiple U-statistic framework of the O'Brien and FS tests is useful for this, as they reveal the magnitude and direction of treatment effects for each outcome that contributes to the test statistic.

Investigators may be interested in some guidance concerning which tests may be most appropriate to use for their setting. O'Brien's test will be better powered when treatment favors most or all outcomes with similar effects, or when treatment is more favorable on the uncensored outcomes, as it uses all available pairwise comparisons on each outcome. The FS test is most applicable when there is a clear hierarchy of outcomes, and better powered when the treatment is most favorable towards the top of that hierarchy. Additionally, the FS test may be better powered when outcomes are very correlated, as the test removes much of the additional variance due to that correlation.

It is important to understand what these U-statistics are measuring. The *global treatment effect* θ_ϕ that these statistics estimate are sometimes complex functions of the marginal or joint distributions of the data, including censoring distributions. The choice of ϕ should be carefully considered, and should be a reflection of what constitutes efficacy of the treatment within the context of the study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank committee member Rebecca Betensky for her input and discussion on this project. Thanks to the referees and the Associate Editor for their thoughtful suggestions, which have helped improve this paper. This work was supported by NIH grant T32NS048005.

References

Berry JD, Miller R, Moore DH, Cudkowicz ME, Van Den Berg LH, Kerr DA, et al. The combined assessment of function and survival (cafs): A new endpoint for als clinical trials. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2013; 14:162–168. [PubMed: 23323713]

Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Statistics in medicine. 2010; 29:3245–3257. [PubMed: 21170918]

Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 1995; 16:1190–1208.

Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, Nakanishi A, et al. study group, B. A., complete listing of the BDNF Study Group, A. The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. Journal of the neurological sciences. 1999; 169:13–21. [PubMed: 10540002]

Cudkowicz ME, Shefner JM, Schoenfeld DA, Zhang H, Andreasson KI, Rothstein JD, et al. Trial of celecoxib in amyotrophic lateral sclerosis. Annals of neurology. 2006; 60:22–31. [PubMed: 16802291]

Felker GM, Maisel AS. A global rank end point for clinical trials in acute heart failure. Circulation: Heart Failure. 2010; 3:643–646. [PubMed: 20841546]

Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Statistics in medicine. 1999; 18:1341–1354. [PubMed: 10399200]

Fisher LD. Self-designing clinical trials. Statistics in medicine. 1998; 17:1551–1562. [PubMed: 9699229]

Gehan EA. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika. 1965; 52:203–223. [PubMed: 14341275]

Häberle L, Pfahlberg A, Gefeller O. Assessment of multiple ordinal endpoints. Biometrical Journal. 2009; 51:217–226. [PubMed: 19197963]

Healy BC, Schoenfeld D. Comparison of analysis approaches for phase iii clinical trials in amyotrophic lateral sclerosis. Muscle & nerve. 2012; 46:506–511. [PubMed: 22987690]

Huang P, Tilley BC, Woolson RF, Lipsitz S. Adjusting o'brien's test to control type i error for the generalized nonparametric behrens–fisher problem. Biometrics. 2005; 61:532–539. [PubMed: 16011701]

Huang P, Woolson RF, O'Brien PC. A rank-based sample size method for multiple outcomes in clinical trials. Statistics in medicine. 2008; 27:3084–3104. [PubMed: 18189338]

Li Q, Liu A, Yu K, Yu KF. A weighted rank-sum procedure for comparing samples with multiple endpoints. Statistics and its interface. 2009; 2:197. [PubMed: 19823699]

Minas G, Rigat F, Nichols TE, Aston JA, Stallard N. A hybrid procedure for detecting global treatment effects in multivariate clinical trials: theory and applications to fmri studies. Statistics in medicine. 2012; 31:253–268. [PubMed: 22170084]

Moyé LA, Davis BR, Hawkins CM. Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. Statistics in medicine. 1992; 11:1705–1717. [PubMed: 1485054]

Moyé LA, Lai D, Jing K, Baraniuk MS, Kwak M, Penn MS, et al. Combining censored and uncensored data in a u-statistic: Design and sample size implications for cell therapy research. The international journal of biostatistics. 2011; 7:1–29.

O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984:1079–1087. [PubMed: 6534410]

Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics. 1987:487–498. [PubMed: 3663814]

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2014.

Rosenbaum PR. Some poset statistics. The Annals of Statistics. 1991:1091–1097.

Rosenbaum PR. Coherence in observational studies. Biometrics. 1994:368–374. [PubMed: 8068837]

Sun H, Davison BA, Cotter G, Pencina MJ, Koch GG. Evaluating treatment efficacy by multiple endpoints in phase ii acute heart failure clinical trials: Analyzing data using a global method. Circulation: Heart Failure. 2012:742–749. [PubMed: 23065036]

Van der Vaart, AW. Asymptotic statistics. Vol. 3. Cambridge university press; 2000.

Wei L, Johnson WE. Combining dependent tests with incomplete repeated measurements. Biometrika. 1985; 72:359–364.

Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. Statistics in medicine. 1989; 8:415–425. [PubMed: 2727465]

Wittkowski KM, Lee E, Nussbaum R, Chamian FN, Krueger JG. Combining several ordinal measures in clinical studies. Statistics in medicine. 2004; 23:1579–1592. [PubMed: 15122738]

Table 1

Type I Error (%), Scenario 1: Uncensored data, 4 outcomes. n, m = sample size per strata in each group, respectively. T_O = O'Brien original test; T_O^U=Proposed O'Brien test; $T_O^{U,\mathrm{Ad}}$=Proposed Adaptive O'Brien test. ρ = correlation between outcomes; for unequal variances below, group 1 covariance $\Sigma_1 = diag(1, 1, 1, 1)$; for group 2, Σ_2 has elements (1,9,16,25) on the diagonal, and $\Sigma_{2ij} = 1$ for $i \neq j$.

Variances	ρ	No. Strata	n	m	T_O	T_O^U	$T_O^{U,\mathrm{Ad}}$
Equal	0	2	15	15	4.7	4.2	4.3
			30	30	5.4	5.0	5.8
			100	100	4.9	4.8	5.0
			80	40	5.0	4.8	5.1
		4	15	15	5.5	5.7	6.0
			30	30	5.1	5.1	5.4
			100	100	5.0	5.0	5.1
			80	40	4.3	4.2	5.1
	0.5	2	15	15	5.6	5.0	4.9
			30	30	5.2	4.7	4.9
			100	100	5.4	5.3	5.4
			80	40	5.1	4.9	5.0
		4	15	15	5.6	5.5	5.5
			30	30	4.9	5.0	5.3
			100	100	5.3	5.3	5.6
			80	40	5.7	5.7	5.4
Unequal	See Caption	2	15	15	6.5	4.6	4.7
			30	30	6.5	4.8	4.9
			100	100	6.6	4.9	5.3
			80	40	6.0	4.7	5.3
		4	15	15	7.2	5.5	5.9
			30	30	6.7	5.0	5.2
			100	100	6.6	4.8	5.2
			80	40	5.9	4.8	4.9

Table 2

Power (%), Scenario 1: Uncensored data, 4 outcomes, θ_ϕ= (.03,.08,.16,.28). *n*, *m* = sample size per strata in each group, respectively. T_O = O'Brien original test; T_O^U=Proposed O'Brien test; $T_O^{U,\mathrm{Ad}}$=Proposed Adaptive O'Brien test; $T_O^{U,\mathrm{Opt}}$=Optimal O'Brien Test. w_{opt} = Optimal weight vector. ρ = correlation between outcomes.

ρ	No. Strata	*n*	*m*	T_O	T_O^U	$T_O^{U,\mathrm{Ad}}$	$T_O^{U,\mathrm{Opt}}$	w_{opt}
0	2	20	20	55.6	54.1	52.6	71.6	(.053, .136, .281, .530)
		40	40	84.5	84.2	84.8	95.6	
	4	10	10	55.5	56.7	53.2	74.3	
		20	20	85.5	85.7	83.9	95.7	
0.2	2	30	30	54.5	53.7	59.4	80.1	(0, .024, .276, .700)
		60	60	84.1	83.8	90.4	98.2	
	4	15	15	53.7	54.4	61.3	82.7	
		30	30	84.2	84.4	90.7	98.3	
0.5	2	30	30	38.7	37.7	52.6	77.0	(0, 0, .094, .906)
		60	60	66.4	66.1	84.2	96.8	
	4	15	15	39.4	39.7	56.5	77.9	
		30	30	66.6	66.5	86.5	96.9	
0.8	2	30	30	30.8	30.2	50.1	75.8	(0, 0, 0, 1)
		60	60	53.2	52.8	79.6	97.4	
	4	15	15	30.7	30.9	57.0	78.2	
		30	30	52.6	52.9	84.2	97.0	

Table 3

Type 1 Error (%), Scenario 2: Survival and ALSFRS; T_O^U=Propsed O'Brien test, $T_O^{U,\mathrm{Ad}}$=Proposed Adaptive O'Brien test, T_{FS}^U=Proposed Finkelstein − Schoenfeld(FS)test, $T_{\mathrm{FS}}^{U,\mathrm{Ad}}$=Proposed Adaptive FS test.

Censoring	No. Strata	*n*	*m*	T_O^U	$T_O^{U,\mathrm{Ad}}$	T_{FS}^U	$T_{\mathrm{FS}}^{U,\mathrm{Ad}}$
Equal (52 %)	2	15	15	4.8	5.4	4.8	4.7
		30	30	4.7	5.1	4.7	4.8
		100	100	5.0	5.3	4.7	5.5
		80	40	4.9	4.5	5.1	4.3
	4	15	15	4.8	4.8	5.1	5.2
		30	30	5.1	5.7	5.5	5.7
		100	100	4.7	4.8	4.9	5.2
		80	40	5.0	4.7	5.2	5.3
Unequal (52 %, 80%)	2	15	15	5.0	5.6	4.8	5.4
		30	30	4.6	4.9	4.6	5.3
		100	100	5.5	5.2	5.4	5.6
		80	40	4.7	4.7	4.5	4.6
	4	15	15	4.9	5.1	5.2	5.9
		30	30	5.0	5.1	4.8	5.4
		100	100	4.9	4.8	5.0	5.3
		80	40	5.0	5.3	5.2	5.4

Table 4

Power(%), Scenario 2: Survival and ALSFRS; T_O^U=Proposed O'Brien test, $T_O^{U,\mathrm{Ad}}$=Proposed Adaptive O'Brien test, $T_O^{U,\mathrm{Opt}}$=Optimally Weighted O'Brien Test, T_{FS}^U=Proposed Finkelstein- Schoenfeld(FS)test, $T_{\mathrm{FS}}^{U,\mathrm{Ad}}$=Proposed Adaptive FS test, $T_{\mathrm{FS}}^{U,\mathrm{Opt}}$=Optimally Weighted FS Test.

Effect Size (Survival,ALS)	No. Strata	*n*	*m*	T_O^U	$T_O^{U,\mathrm{Ad}}$	$T_O^{U,\mathrm{Opt}}$	T_{FS}^U	$T_{\mathrm{FS}}^{U,\mathrm{Ad}}$	$T_{\mathrm{FS}}^{U,\mathrm{Opt}}$
Moderate,Mild	2	50	50	45.7	48.6	59.6	48.1	49.2	59.6
		100	100	73.9	78.9	87.2	77.8	79.4	87.4
	4	25	25	47.0	49.3	58.9	49.2	48.5	58.9
		50	50	74.6	78.8	87.0	77.6	78.2	87.0
Strong,Mild	2	30	30	45.4	55.5	70.8	47.7	56.0	70.8
		60	60	75.2	85.7	95.3	78.1	86.7	95.3
	4	15	15	47.1	59.0	73.4	50.8	58.8	73.4
		30	30	76.1	87.2	95.1	78.4	87.0	95.1
Mild,Moderate	2	50	50	57.3	57.9	65.8	52.9	51.5	59.1
		100	100	85.0	86.0	91.5	80.8	78.3	86.0
	4	25	25	55.9	56.7	65.0	52.1	48.1	58.8
		50	50	86.6	86.7	92.0	82.9	77.9	87.5
Mild,Strong	2	25	25	53.7	58.5	70.2	50.0	51.1	62.4
		50	50	83.3	88.3	94.5	78.7	82.1	90.3
	4	15	15	62.9	67.1	79.2	57.8	58.3	71.3
		25	25	83.6	89.3	94.3	79.8	80.8	89.8
Moderate,None	2	50	50	22.0	38.7	63.7	31.3	44.4	63.7
		100	100	38.9	67.1	89.5	55.3	74.0	89.5
	4	25	25	23.7	44.0	61.9	33.3	46.9	61.9
		50	50	39.8	72.1	89.7	56.4	76.7	89.7
None,Moderate	2	50	50	30.0	45.5	65.1	20.6	35.6	62.9
		100	100	53.9	76.8	92.1	37.2	65.6	90.9
	4	25	25	29.8	49.5	66.5	20.9	40.1	64.4
		50	50	54.6	79.0	91.9	37.2	67.6	90.8
Moderate,Moderate	2	30	30	50.7	47.0	51.0	49.6	44.9	49.5

Biometrics. Author manuscript; available in PMC 2016 September 07.

Effect Size (Survival,ALS)	No. Strata	n	m	T_O^U	$T_O^{U,\mathrm{Ad}}$	$T_O^{U,\mathrm{Opt}}$	T_{FS}^U	$T_{\mathrm{FS}}^{U,\mathrm{Ad}}$	$T_{\mathrm{FS}}^{U,\mathrm{Opt}}$
		60	60	80.0	76.9	80.1	78.6	S74.4	78.6
	4	15	15	50.3	45.2	50.3	48.9	41.9	48.9
		30	30	80.2	75.0	80.2	78.5	69.7	78.3