A Generalized Global Rank Test for Multiple, Possibly Censored, Outcomes

Ritesh Ramchandani

Harvard School of Public Health

August 5, 2014

Motivation

- Multiple Outcomes in clinical studies when a single endpoint is not adequate.
- ALS: ALSFRS-R and mortality.
- Heart Disease: Rehospitalization and mortality.
- Stroke: Different outcome scales to assess recovery

Methods for multivariate outcomes

- Multiple comparisons
  - Examples: Bonferroni; Hochberg's step-up procedure (1988)
- Composite outcomes for time-to-event (e.g. progression free survival)
- Global Tests
  - Hotelling's (1934) T^2: Assumes normality
  - O'Brien's (1984) nonparametric rank-sum test
- We are interested in a nonparametric global test of the overall efficacy of a treatment over multiple outcomes.

Some joint tests based on ranks

- O'Brien (1984): Rank subjects with respect to each outcome; average each subject's ranks, and perform rank-sum test.
- Finkelstein-Schoenfeld Joint Rank Test (1999): Survival and longitudinal outcome. First rank pairs of subjects on survival; if tied or incomparable then rank on longitudinal outcome. Moye (1992,2011) proposed a related test.
- Wittkowski (2004): Patient A is ranked higher than B if all outcomes for patient A are better or equal to those of patient B.

Examples of scoring pairs

	A		B	
Time	ALSFRS-R	Alive?	ALSFRS-R	Alive?
0	32	Yes	32	Yes
1	35	Yes	31	Yes
2	37	Censored	25	Yes
3	NA	NA	NA	No

- A is censored at time 2, B dies at time 3. Survival outcome score is 0 since indeterminate
- At the last common time they were observed, A had a higher ALSFRS-R score than B did, so the ALSFRS-R outcome score is 1.
- 1. O'Brien: Overall Score of 0.5 (average of both outcome scores)
  2. Finkelstein-Schoenfeld: Score of 1 (survival indeterminate, so use ALSFRS-R)
  3. Wittkowski: Score of 1 (ALSFRS-R better, and survival not known to be worse)

General Procedure

- For each pair of subjects i and j in different groups, score for each outcome k:

$$r_{ijk} = \begin{cases} 1, \text{ if } i \text{ did better than } j \\ -1, \text{ if } j \text{ did better than } i \\ 0, \text{ if indeterminate} \end{cases}$$

- Assign a univariate score for the pair i and j that is a function of each the scores for the p outcomes.
- The test statistic will be based on the sum of the univariate scores over one of the groups.

General Procedure

- Suppose we have two groups, and p outcomes.
- Denote $E[r_{ijk}] = \theta_k$. Can be thought of as treatment effect for outcome k.
  - Constructed so that $\theta_k = 0$ when the treatment has no effect.
- Define a composite scoring function $\phi(\mathbf{r})$ that maps the individual outcome ranks to a univariate score. Examples:
  - O'Brien: $\phi(\mathbf{r}) = \sum_k^p r_k$
  - Joint rank test: $\phi(\mathbf{r}) = r_1 + I(r_1 = 0)r_2 + ... + I(r_1 = ... = r_{p-1} = 0)r_p$
  - Wittkowski: $\phi(\mathbf{r}) = I(max_k\{r_k\} > 0) - I(min_k\{r_k\} < 0)$
  - Combination: $\phi(\mathbf{r}) = r_1 + I(r_1 = 0)\frac{1}{p-1}\sum_{k=2}^p r_k$

Test Statistic

- Let n, m be the sample sizes in groups 1 and 2 respectively, and $N = n + m$ is the total sample size
- Test Statistic:

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

- U-statistic that estimates the parameter $\theta_\phi = E[\phi(r_1(X_1, Y_1), ..., r_p(X_p, Y_p))]$, for a given choice of ϕ
  - Global treatment effect
- Under H_0 and some regularity conditions, $\sqrt{N}U \to N(0, \sigma^2)$
- Can estimate σ^2 consistently from the data.
- Can also estimate power for different values of θ_ϕ under H_1.

Weights

- In addition, we can choose a function ϕ that assigns different weights to each outcome.
- For Example, if we have weights $(w_1, ...w_p)$:
  - Weighted O'Brien: $\phi(r_1, ..., r_p) = \sum_k^p w_k r_k$
  - Weighted Joint Rank: $\phi(r_1, ..., r_p) = w_1 r_1 + I(r_1 = 0) w_2 r_2 + ... + I(r_1 = ... = r_{p-1} = 0) w_p r_p$
- How to choose weights?
  - Importance of outcomes (utility).
  - Minimize variance
  - Optimal: Maximize power under a particular alternative hypothesis.
  - Adaptively choose optimal weights by estimating from other strata.
  - Should restrict weights to the positive quadrant, i.e. $w_k \geqslant 0$ for all k.

O'Brien as sum of correlated U-Statistics

- We can write O'Brien's global test as a sum of outcome-specific U-statistics.
- The weighted O'Brien statistic is given by $\mathbf{w}'\mathbf{U} = w_1 U_1 + w_2 U_2 + ... w_p U_p$
- Then $\sqrt{N}\mathbf{w}'\mathbf{U} \rightarrow N(0, \mathbf{w}'\Lambda\mathbf{w})$, where $\Lambda = cov(\mathbf{U})$

Optimal Weights

- Solution that maximizes the power function yields $\mathbf{w} = \Lambda^{-1}\theta$, where $\theta = (\theta_1, ..., \theta_p)'$, the marginal treatment effects for each outcome.
  - Note: this may give negative weights; we can use numerical optimization to restrict weights to be positive.
  - Minimum variance weights correspond to the case where $\theta = (1, ..., 1)$
- Choose θ for a particular alternative hypothesis, estimate Λ from the data.
- Similarly, can write Joint-Rank test as sum of correlated U-statistics, and optimal solution follows.

Simulations: Mortality and ALSFRS-R

- Simulations based on a trial of Celebrex for ALS.
- Data generated from a shared parameter model with patient-specific random effects for ALSFRS-R functional rating scores, and hazard for survival is a function of ALSFRS trajectory and treatment effect (Healy 2012).
- First column denotes treatment effect for survival and ALS, respectively: None, mild, or moderate (mod).

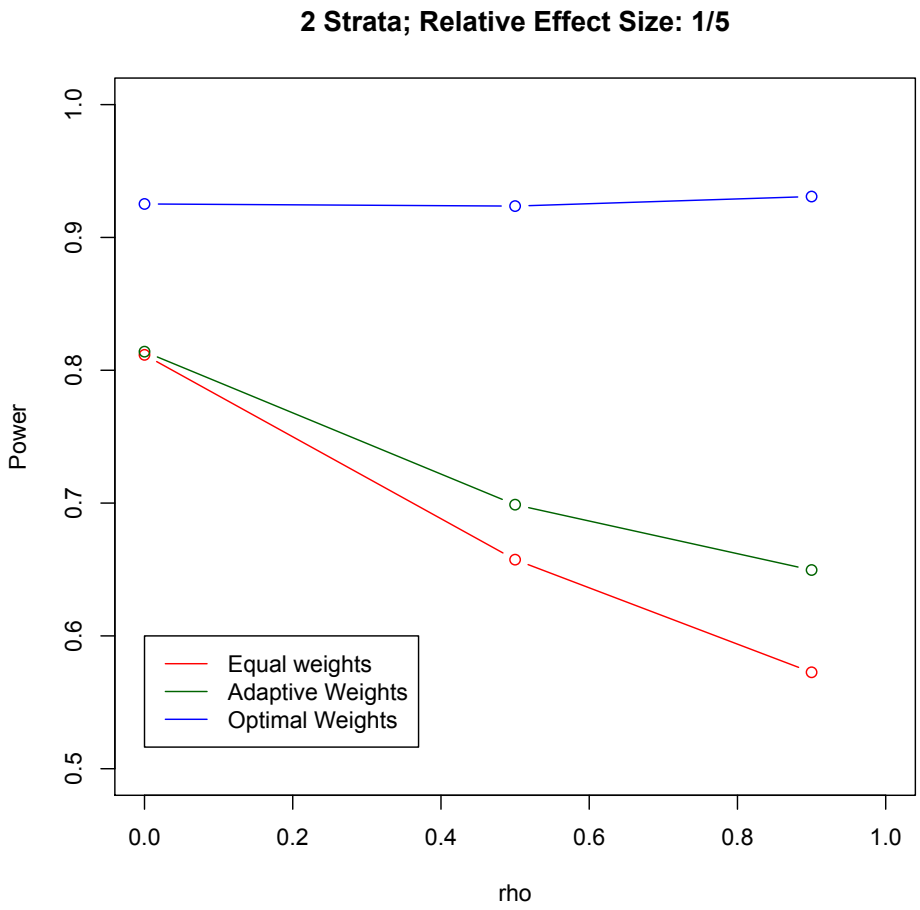
(Surv,ALS)	O'Brien	O'Brien$_w$ ($\bar{w}_{opt}$)	Joint-Rank	JR$_w$ ($\bar{w}_{opt}$)
(Mod,Mild)	.52	.59 (90,10)	.49	.59 (93,7)
(Mild,Mod)	.61	.72 (3,97)	.52	.57 (32,68)
(Mod,None)	.25	.61 (1,0)	.31	.61 (1,0)
(None,Mod)	.35	.74 (0,1)	.21	.62 (0.1,1)*

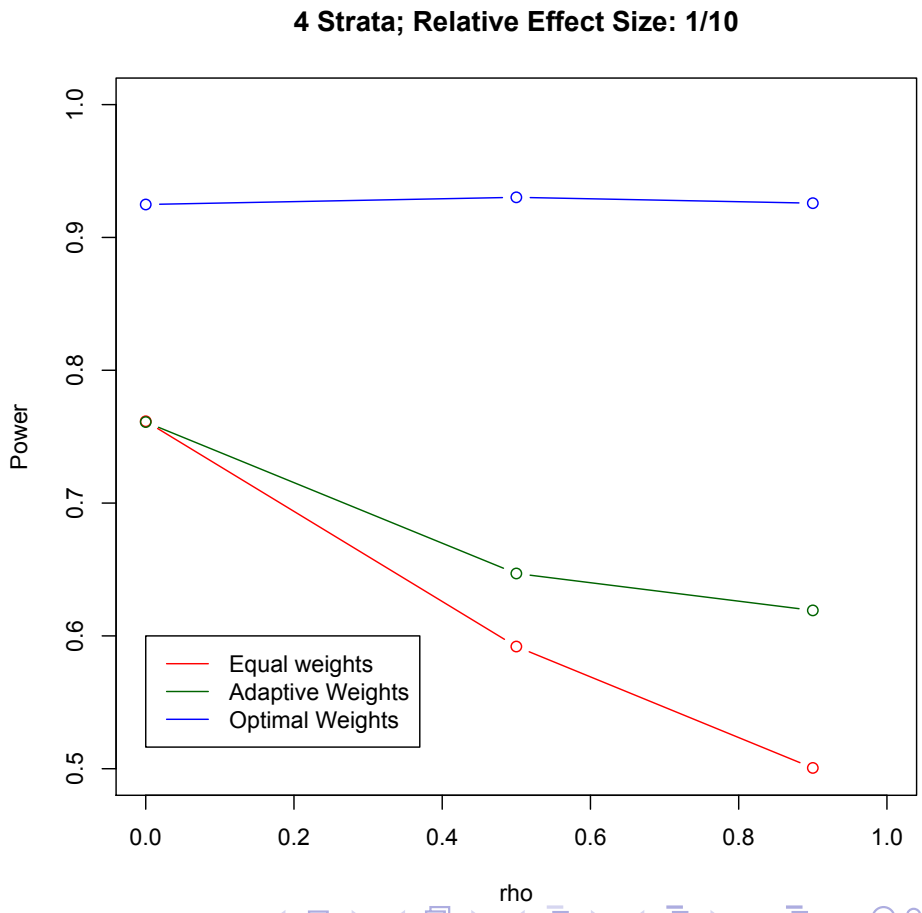
Adaptive Weights

- Optimal weight requires knowledge of the parameter values we expect to see under an alternative hypothesis.
- Can be estimated from previous studies, but in general unknown.
- With stratified data (e.g. different centers in a clinical trial), we can estimate the necessary weights in each of the 'previous' strata, and then apply the optimal weights to the 'next' stratum, and continue in this manner (Fisher 1998)
- Disadvantages:
  - Can make interpretation difficult.
  - Can yield erroneous weights, particularly when there is a treatment by strata interaction.

Adaptive Weights: Simulations

- 2 outcomes, uncensored; generated from multivariate normal with variance 1, correlation ρ.





Simulations Summary

- Type 1 error controlled at the nominal level for small samples, unequal variances, unequal sample sizes, and unequal censoring distributions.
- Optimal weight improves efficiency of existing tests, but weights in general are unknown. May be estimated from prior data.
- Adaptive weighting can improve power, but only if effect sizes vary significantly between outcomes and correlation is moderate to high.

Example Data Analysis: Ceftriaxone ALS Trial

- Stage 3 Double Blind Clinical Trial of Ceftriaxone in Subjects with ALS
- 513 Subjects monitored for rate of decline in ALSFRS-R scores over time, and survival
- 340 Subjects administered Ceftriaxone; 173 placebo
- Average follow up time of 1.6 years, maximum $\approx$ 6 years
- Conclusion: Survival and rate of decline were not signifcantly different between Ceftriaxone and placebo

Ceftriaxone Trial: Comparison of Joint Tests

	Test Statistic	p-value
Survival only	1.13	.26
ALS only	-0.39	.70
O'Brien	0.16	.87
O'Brien$_w$	0.85	.40
Joint-Rank	0.71	.48
Joint-Rank$_w$	0.31	.76

Covariance matrices $\hat{\Lambda}$:

O'Brien: $\begin{pmatrix} .37 & .06 \\ .06 & 1.40 \end{pmatrix}$

Joint-Rank: $\begin{pmatrix} .37 & .006 \\ .006 & .17 \end{pmatrix}$

Minimum variance weights:

$\mathbf{w} = (.815, .185)$

$\mathbf{w} = (.31, .69)$

Some things to consider

- Global Test weak under alternatives where treatment affects outcomes in different directions. That is okay for our purposes.
- Results of any given test should be carefully interpreted.
- A positive result does not necessarily mean treatment is best for all outcomes. Make use of descriptive statistics and plots, or combine with closed testing procedures.
- Not necessarily the right test if main interest in isolating which specific outcomes are meaningfully different. Use multiple comparisons.

Strengths and Limitations

- Flexible: Allows investigator to choose test based on the context of study, estimate of interest, and required sample size/power.
- Variance can easily be estimated from the data for a given composite function of the scores.
- Framework for constructing optimal weights under specific alternatives, for some tests.
- Covariate adjustment only available through stratification.

Acknowledgements

- Dianne Finkelstein and David Schoenfeld.
- Harvard School of Public Health and Massachusetts General Hospital, Department of Biostatistics.
- Supported by NIH grant T32NS048005.