A new method to identify significant endpoints in a closed test setting

MCP Conference, Vienna: July 11, 2007

Carlos Vallarino, Takeda Pharmaceuticals
Takeda Global Res & Dev, Deerfield, IL, USA
cvallarino@tgrd.com

"For my part I know nothing with any certainty, but the sight of the stars makes me dream." van Gogh

Co-authors: J. Romano, R. Bittman, M. Wolf

Outline

I. Background. FWE. Closed testing. O'Brien. Common effect direction. Concept of consonance.

II. Simple sum test is maximin. New method for testing the intersection is consonant and maximin.

III. Application to the PROactive clinical trial. WLW method. Permutation test alternative.

Multi-experiment Trial

Simultaneously test s hypotheses $H_1, \ldots, H_s$: common in clinical trials to have s measures of efficacy or "endpoints" per patient.

Go beyond the usual $P(\text{Type } I \text{ error})$ to the familywise error rate $(\text{FWE}_P) \equiv \text{Prob}_P$ (at least 1 true H_i is rejected).

Require $\text{FWE}_P \leq \alpha \quad \forall P,$

where our model $P = \{P_\theta,\ \theta \in \Omega\}$ and $H_i : P \in \omega_i \subset \Omega$.

Aim: allocate FWE_P so as to maximize "power".

The Classics. Bonferroni: Given p-value $\hat{p}_j$ for testing H_j, reject any H_i if $\hat{p}_i \leq \alpha/s$. This controls FWE_P.

Stepdown Methods: Holm. Order the p-values and reject corresponding $H_{(1)}, \ldots, H_{(j)}$ from smallest to largest possible. Sidak is another *stepdown* method.

A *stepup* method starts with the largest p-value: Hochberg, Rom.

While big improvements over Bonferroni, still can be conservative.

Stretching the FWE Budget

- *Intersection Hypothesis.* For a subset $K \subseteq \{1, \ldots, s\}$, let $H_K \equiv \bigcap_{i \in K} H_i = \bigcap_{i \in K} \omega_i$ so H_K is true iff $\theta \in \bigcap_{i \in K} \omega_i$.

Closure Method. To simultaneously test $H_1, \ldots, H_s$, Marcus, Peritz and Gabriel (1976) reduce the problem to constructing single tests that control the usual probability of Type I error.

Suppose an α-level test of H_K exists for every subset K. Then, the decision rule that rejects H_i if and only if H_K is rejected for all subsets K for which $\{i\} \subseteq K$ strongly controls the FWE.

Stretching the FWE Budget

So, in order for H_i to be deemed significant, every intersection hypothesis which includes H_i must be deemed significant. For $s = 3$, to reject H_2, must reject $H_{\{1,2,3\}}$, $H_{\{1,2\}}$, $H_{\{2,3\}}$ and H_2.

How to test the intersection?

A. The max. Incorporating the dependence structure of p-values. Westfall and Young (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.*

Further work by van der Laan et al. (2004). Romano and Wolf (2005) test any intersection hypothesis by bootstrapping the distribution of the maximum test statistic or minimum p-value.

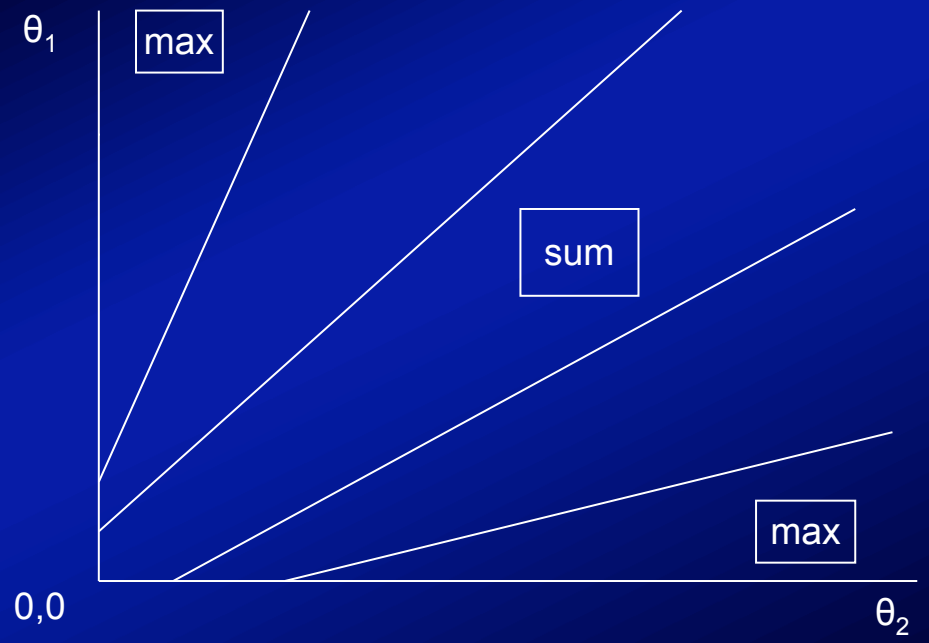
B. The sum. When Common Effect Direction can be assumed, O'Brien (1984) developed a test statistic for the intersection hypothesis more powerful than Hotelling's T^2 (in 2 dimensions).

- *Common Effect Direction.* Means $\{\theta : \theta_i \leq -\epsilon \text{ or } \geq \epsilon \ , \ i = 1, 2\}$.

Sum vs. Max. Lehmacher et al. (1991) state that Bonferroni, and tests based on the maximum test statistic, are useful in situations where one difference stands out from the rest; O'Brien, and tests based on the sum, succeed when all treatment effects are similar.

Intersection Test

Power comparison in 2 dimensions



Consonance

A method is *consonant* if the rejection of an intersection hypothesis implies the rejection of at least one subset hypothesis it contains.

- e.g. rejection of $\theta_1 = \theta_2 = 0$ entails rejection of $\theta_1 = 0$ or $\theta_2 = 0$.

Not all methods generated by the closure principle are *consonant*.

"... consonance is only a desirable property."

BLUE: Test (i) defined by circle of radius 2.448; RED: Test (ii) defined by square with halflength 2.234; $\alpha = 0.05$

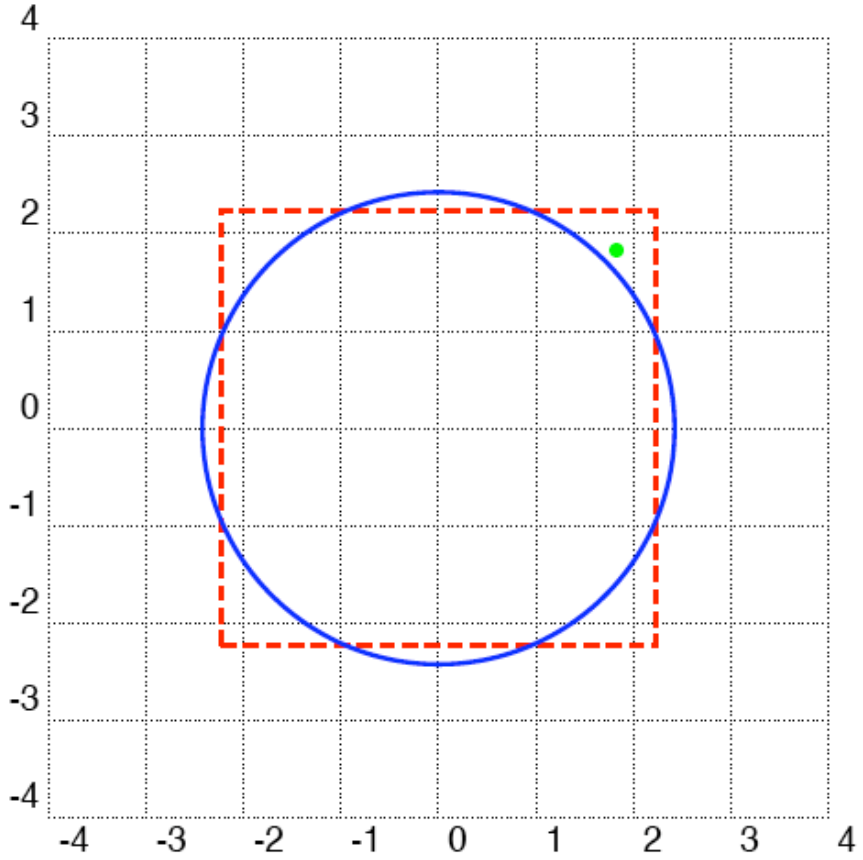


Figure 1: The pt (1.83, 1.83) leads to rejection of $H_{\{1,2\}}$ using (i) but not (ii), but applying closure, no rejections of individual H_i (1.96).

Rejection region of improved Test (i).

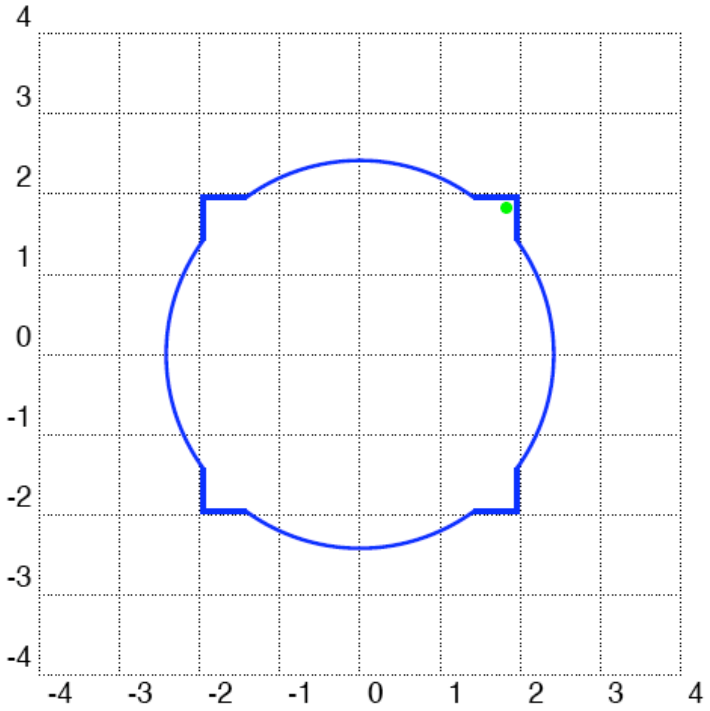


Figure 2: 'Union' of circle (radius 2.421) and square (side 2x1.96) removes dissonant points like (1.83, 1.83), which no longer rejects. Note (2.43, 0) would reject H_1. This test is consonant.

Focus on 2 dimensions. Asymptotic version: observe (X_1, X_2) bivariate normal with $X_i \sim N(\theta_i, 1)$ and known correlation ρ.

Common effect direction. Assume (θ_1, θ_2) either in quadrants I, III.

$H_i : \ \theta_i = 0$ versus $H_i' : \ \theta_i \neq 0$. Want to apply closure.

Test for H_i based on UMPU test, which rejects if $|X_i| > z_{1-\frac{\alpha}{2}}$.

Choice of intersection test? One possibility, determine the maximin test over $\{\min(|\theta_1|, |\theta_2|) \geq \epsilon\}$ or over $\{|\theta_1 + \theta_2| \geq \epsilon\}$.

Proposition: In either region where minimum power is maximized and for all $\epsilon > 0$, the maximin test rejects $H_{\{1,2\}} : \theta_1 = \theta_2 = 0$ if

$$|X_1 + X_2| > z_{1-\frac{\alpha}{2}} (2 + 2\rho)^{1/2} \ . \tag{1}$$

Notice that $|X_1 + X_2|$ in (1) can be large without either $|X_i|$ being sufficiently large to reject its individual H_i (dissonance).

Theorem: Optimal consonant, maximin, level α test is given by

$$\{(X_1, X_2) : \ |X_1 + X_2| > r(1-\alpha), \ \max(|X_i|) > z_{1-\frac{\alpha}{2}}\} \ , \qquad (2)$$

where the constant $r(1-\alpha)$ is determined so that this rejection region has probability α under $(\theta_1, \theta_2) = (0, 0)$.

Proofs: visit ssrn.com/abstract=938950 or see technical report.

The optimal rejection region of (2) resembles that of (1) but adds the necessary restrictions to make the test consonant.

Rejection regions for simple sum test and its consonant improvement when $\rho = 0$

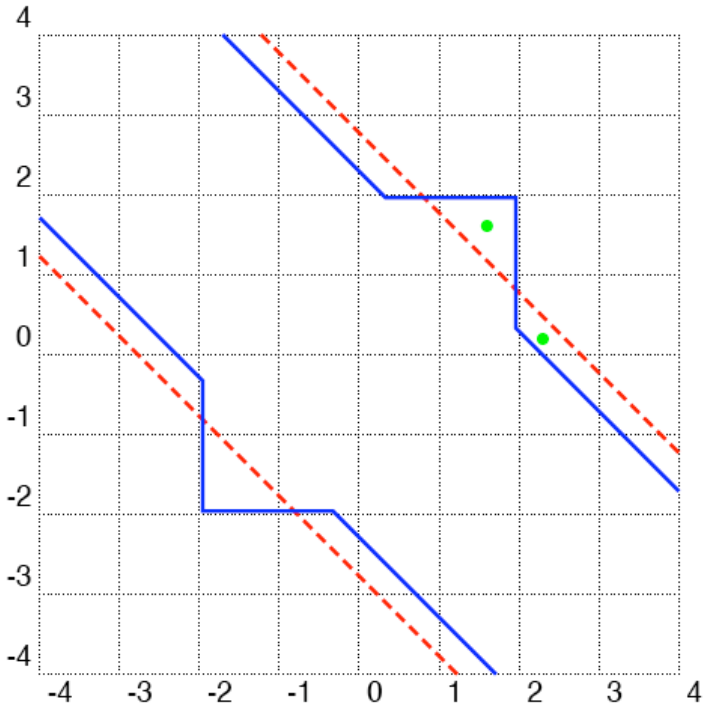


Figure 3: Simple sum test rejects outside dashed red band. Improved test rejects outside solid blue band. The pt $(1.6, 1.6)$ accepts both H_i. The pt $(2.3, 0.2)$ rejects H_1 for improved test only.

Motivating Case Study: PROactive

Randomized, double-blind clinical trial to investigate the effect of oral glucose lowering drug *pioglitazone* on macrovascular outcomes. $n = 5238$ patients with T2 diabetes and a history of heart disease.

Primary endpoint - time to first occurrence of: death, stroke, non-fatal MI (incl. silent myocardial infarction), acute coronary syndrome (ACS), leg amputation, various procedural interventions.

Secondary endpoint - time to 1st occurrence of the most serious and objective components: death, stroke, non-fatal MI (excl. silent MI).

- Two interim analyses reduced nominal FWE available to 0.044

PROactive Clinical Trial

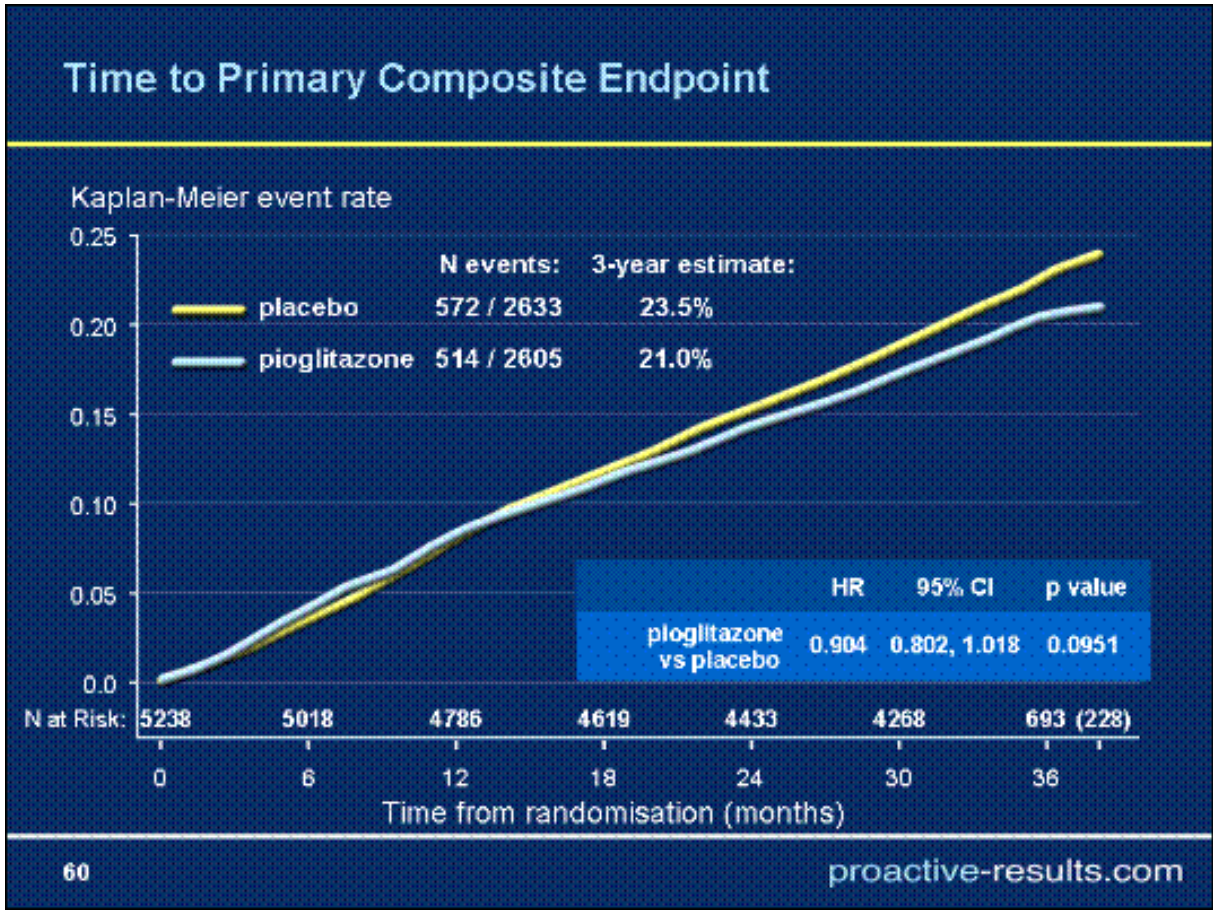


Figure 4: The log-rank test (Mantel-Haenszel test) for the Primary endpoint yields a p-value of 0.095.

Secondary Endpoint

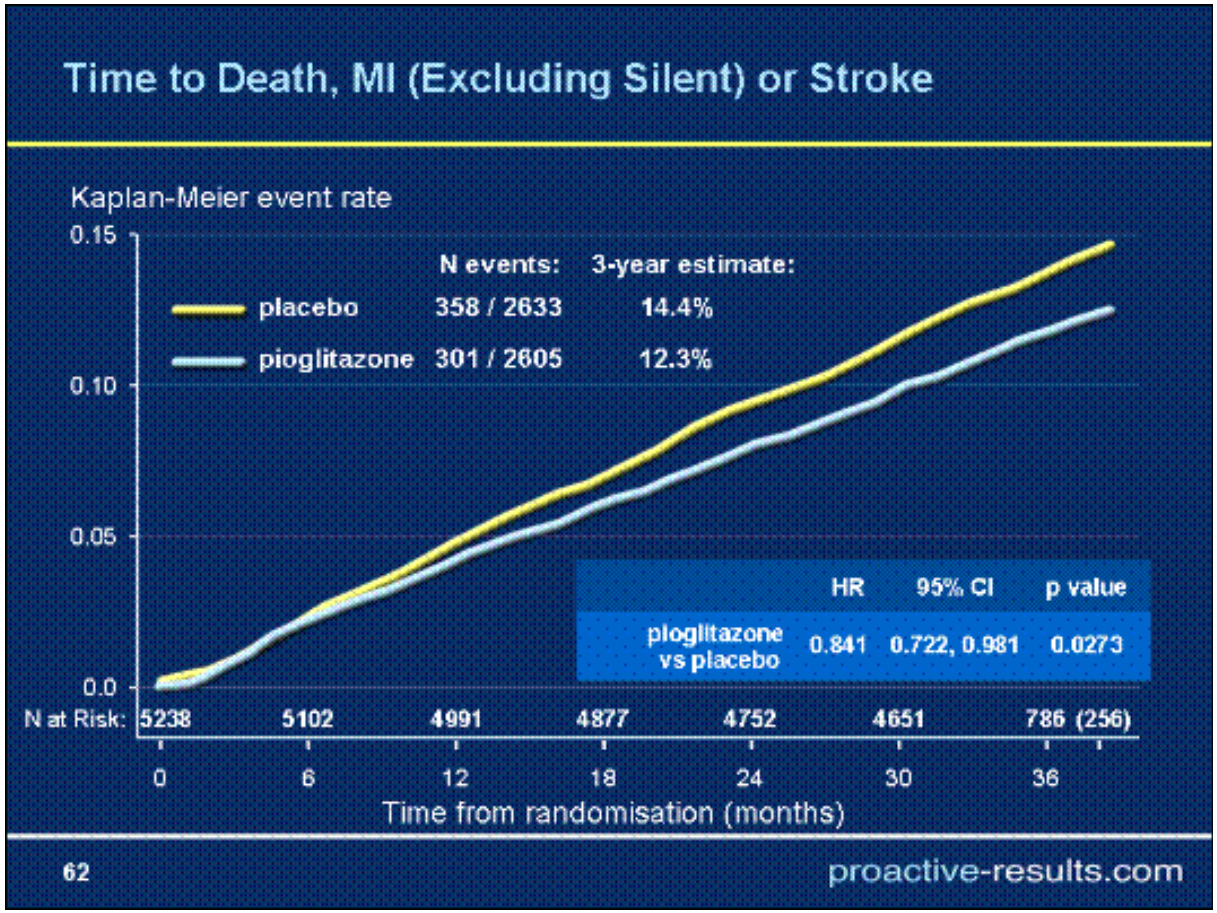


Figure 5: The log-rank test for the Secondary endpoint yields a p-value of 0.027.

Is the secondary endpoint significant? Dormandy et al. (2005) were criticized for claiming it was, because a MCP was not used.

Could a MCP have been applied? How to combine the individual log-rank tests into a test of the intersection and apply a closed test?

Try the simple sum and consonant sum tests! Must consider the endpoints co-primary. Note from the definition: high correlation expected and assumption of common effect direction justified. Individual H_i already tested with the log-rank test.

WLW method. Use relation of log-rank test to PH model to apply Wei et al. (1989) method of marginal distributions. Sum studentized parameter estimates of the 2 endpoints (asymptotically normal) and estimate $\hat{\rho} = 0.74$ from the "sandwich" estimator of the covariance matrix; see Liang and Zeger (1986).

Results. Consonant sum test yields adjusted p-value of 0.036 for $H_{\{1,2\}}$ (vs. 0.038 for the simple sum).

Closed test: $H_{\{1,2\}}$ – p-value $< 0.044 \implies$ reject.

H_2 – log-rank test p-value $0.027 < 0.044 \implies$ reject.

Second endpoint declared significant.

Permutation test. Randomly permute treatment, placebo labels. More robust. Useful with small samples. Yielded identical results.

Conclusion: New consonant sum test a more powerful tool under common effect direction.

Had PROactive evaluated its endpoints as co-primary in a closed test setting, both the simple sum test and the new consonant sum test would have identified one endpoint as statistically significant.

References

[1] Dormandy, J.A., Charbonnel, B., Eckland, D.J., et al. (2005). Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study. *Lancet* **366**, 1279–1289.

[2] Lehmacher, W., Wassmer, G., and Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling experimentwise error rate. *Biometrics* **47**, 511–521.

[3] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

[4] Marcus, R., Peritz, E., and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

[5] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

[6] Romano, J.P., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100**, 94–108.

[7] van der Laan, M.J., Dudoit, S. and Pollard, K.S. (2004). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology* **3**, No. 1, Article 14. URL `www.bepress.com/sagmb/vol3/iss1/art14`.

[8] Wei, L.J., Lin, D.Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *JASA* **84**, 1065–1073.

[9] Westfall, P. and Young, S. (1993). *Resampling-based Multiple Testing.* John Wiley, New York.