Comparison of Post Hoc Multiple Pairwise Testing Procedures as Applied to Small *k*-Group Logrank Tests

Moonseong Heo[1,*] and Andrew C. Leon[2,3]

[1]*Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA*

[2]*Department of Psychiatry;* [3]*Department of Public Health, Weill Medical College of Cornell University, New York, NY, USA*

Abstract: The logrank test is widely used to compare groups on distribution of survival time in the presence of censoring. There is no convention for post hoc pairwise comparisons after a significant omnibus *k*-group logrank test. This simulation study compares four post hoc pairwise testing procedures: Bonferroni, Dunn-Šidák, Hochberg, and unadjusted post hoc logrank test procedure. Evaluation criteria include, familywise type I error rate, correct decision rate, number of correctly rejected pairs, and false discovery rate. We demonstrated that when conditioned upon rejection of the omnibus test, multiplicity adjustments may be unnecessary and can be overly conservative when *k* is at most 4, or number of comparisons is no greater than 6. This is supported by the results that the performance of the unadjusted post hoc logrank test procedure is preferred over the others on all criteria except for the false discovery rate. The Hochberg procedure appears to be superior among the adjustments examined. Data from a clinical trial for suicide prevention illustrate these approaches where number of comparison groups is often limited.

Keywords: Logrank test, multiplicity adjustment, post hoc tests, survival analysis.

INTRODUCTION

Survival analysis is a class of statistical procedures that can assess the efficacy of multiple (*k*) treatments by comparing distributions of time to event, or survival time *T*, in the presence of censoring in clinical trials where the number of comparison groups is often limited. For example, trials with *k*=3 will compare an investigational treatment with both the current standard treatment and placebo. Psychiatric trials in particular include studies of time to remission from major depression (e.g., [1]), time to recurrence of major depression (e.g., [2]) and time to discontinuation from study medication [3]. In the illustration that follows, remission of depressive symptoms is the target "terminal event" in survival analysis terminology.

The differences in distributions of survival times *T*, a main study hypothesis in many clinical trials, can be tested typically applying logrank χ^2 tests [4, 5]. Specifically, an *omnibus* logrank χ^2 test with $k - 1$ degrees of freedom (df) with a significance level α_G = .05 can be applied to test a global null hypothesis

$$H_0: S_1(t) = S_2(t) = ... = S_k(t) \quad (1)$$

of equality in distributions of the survival time among *k* (≥ 2) groups, where $S_i(t)=P(T_i > t)$ is the survival function of the survival times in group *i*. When *k*>2 and the global hypothesis is rejected by the omnibus test, *post hoc* multiple comparison procedures are needed to identify the specific pairs of groups that significantly differ in treatment efficacy. (By post hoc we mean "after rejection of the global test.") Although such effort was made recently in Lieberman *et al.* [3] with the Hochberg adjustment procedure [6], no convention exists that guides the choice of approaches in applied settings.

There had been a few studies on multiplicity adjustment in survival analysis. For example, Logan *et al.* [7] compared performances of Bonferroni-type procedures, simulated martingale approaches and closed test procedures [8] for pairwise comparisons of survival distributions between groups. Chen [9] proposed a multiple comparison testing procedure based on Slepian's inequality [10] and compared with both generalized Steel's testing procedure [11] and the closed testing procedure [8] in an experimental setting with many treatment groups and one control group. These three procedures were compared through application of three two-sample logrank statistics with different weighting types—namely, Gehan's statistic [12], the logrank statistic and the Peto-Prentice statistic [13]—in multiple pairwise comparisons of survival distributions between groups treated with a treatment and a control. Nevertheless, these studies did not compare multiple adjustment procedures conditional on rejection of a global test, and more importantly the comparison criteria were based only on type I error rate and statistical power.

In some clinical trials involving survival times and beyond, however, testing the global null hypothesis (1)

*Address correspondence to this author at the Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA; Tel: (718) 920-6247; Fax: (718) 515-6039;
E-mail: moonseong.heo@einstein.yu.edu

E-ISSN: 1929-6029/13

often serves as a primary aim and followed are secondary analyses of post-hoc comparisons to identify specific pairs of groups that are significantly different in survival distributions. In this context, the primary objective of this paper is to compare four types of *general* post hoc testing procedures *conditional* on rejecting the global null hypothesis when applied to logrank tests used in *k*-group RCTs: Bonferroni, Dunn-Šidák [14], Hochberg [6], and unadjusted post hoc logrank test procedure. That is to evaluate *general* multiplicity adjustments when *specifically* aimed for post hoc logrank tests. When the global null hypothesis H_0 (1) is rejected by the omnibus test, these general multiplicity adjustment procedures can be applied to multiple logrank 1 df χ^2 tests for all $c = k(k\text{-}1)/2$ post hoc pairwise comparisons of testing H_{0ij}: $S_i(t) = S_j(t)$, $i < j \le k$. All of those procedures except for the unadjusted post hoc logrank test procedure controls for a familywise type I error (FWE), α_{FWE}, by adjusting the thresholds for significance level for each comparison. Formally, a family F of null hypotheses is a union $F = F_0 \cup F_1$, where $F_0 = \{H_{0ij} |\ i < j \le k$ such that $S_i(t) = S_j(t)$ for *some* pairs of *i* and *j*$\}$ consisting of true null hypotheses and $F_1 = \{H_{0ij} |\ i < j \le k$ such that $S_i(t) \ne S_j(t)$ for *some* pairs of *i* and *j* $\}$ consisting of false null hypothesis are mutually exclusive, i.e., $F_0 \cap F_1 = \varnothing$. The familywise type I error α_{FWE} quantifies the rate of falsely rejecting any hypothesis H_{0ij} that belong to a *special* family F that consists of *only* true null hypotheses $\{H_{0ij} |\ i < j \le k$ such that $S_i(t) = S_j(t)$ for *all* pairs of *i* and *j*$\}$. In other words, the familywise type I error quantifies the rate of falsely rejecting any null hypothesis in F when $F=F_0$. It follows that α_{FWE} = P(Reject any $H_{0ij} |\ H_{0ij} \in F = F_0$).

The primary focus of our evaluations is, however, to identify a post hoc procedure that correctly rejects only false null hypotheses that belongs to the family $F = F_0 \cup F_1$ in general. We denote this rate as correct *decision* rate, or CDR, that is, the rate of rejecting null hypotheses if and only if they belong to $F_1 \subset F$, i.e.,

$$\text{CDR} = P[(\text{Reject all } H_{0ij} |\ H_{0ij} \in F_1 \subset F) \cap (\text{Do not reject any } H_{0ij} |\ H_{0ij} \in F_0 \subset F)]. \quad (2)$$

Apparently CDR will be 0 when $F=F_0$, i.e., $F_1=\varnothing$. Other performance evaluation criteria include:

$$\text{Familywise type I error rate} = P(\text{Reject any } H_{0ij} |\ H_{0ij} \in F_0 = F); \quad (3)$$

$$\text{Number of } \textit{correctly} \text{ rejected pairs} = \#\{\text{Rejected } H_{0ij} |\ H_{0ij} \in F_1 \subset F\}, \quad (4)$$

where #{ } is the number of elements in the set { }; and

$$\text{Empirical FDR} = \#\{\text{Rejected } H_{0ij} |\ H_{0ij} \in F_0 \subset F\} \,/\, \#\{\text{Rejected } H_{0ij} |\ H_{0ij} \in F\}. \quad (5)$$

This false discovery rate (FDR) [15] is the ratio of the number of falsely rejected null hypotheses to that of all rejected null hypotheses, and is referred here to as "empirical" FDR because these evaluations are made based on Monte Carlo Simulations. When the global null hypothesis (1) is true and if any H_{0ij} is rejected based on a post hoc procedure, the empirical FDR will be 1 because $F_0=F$. In contrast, when $S_i(t) \ne S_j(t)$ for all pairs of *i* and *j* ($i < j \le k$), i.e., when $H_{0ij} \in F_1 = F$, the empirical FDR will be 0 because $F_0=\varnothing$. We believe that a comprehensive evaluation of these criteria has not been conducted for post-hoc logrank tests.

POST HOC MULTIPLE PAIRWISE COMPARISON PROCEDURES

The following procedures, adjusted or not, are "protected" in the sense that they are conditional upon the rejection of the global hypothesis, which is implicit in the terminology of "post-hoc" as noted above. Here we consider $\alpha_{FWE} = .05$.

Bonferroni Adjustment

The Bonferroni-adjusted significance level α_B partitions α_{FWE} evenly among *c* tests, resulting in $\alpha_B = \alpha_{FWE}/c$. This adjustment is rather conservative in a sense that familywise error based on α_B is less than the pre-specified α_{FWE} even if independence among pairs is assumed. In this case, it follows that $1-(1-\alpha_B)^C = 1-(1-\alpha_{FWE}/c)^C < \alpha_{FWE} = .05$ for all $c > 1$ with $\lim_{C\to\infty}[1-(1-\alpha_{FWE}/c)^C] = 1-\exp(-\alpha_{FWE}) > .04877$, a lower bound. For instance, with $c = 3$ pairwise comparisons among $k = 3$ groups with $\alpha_{FWE} = .05$, $\alpha_B = .05/3 = .01667$ and subsequently $1-(1-.01667)^3 = .04918$. When this adjustment is implemented, the pairs of groups with *p*-values (from 1 df logrank χ^2 test) less than α_B will be declared to differ significantly in efficacy.

Dunn-Šidák Adjustment

The Dunn-Šidák adjustment [14] was proposed to return the familywise error as the pre-specified α_{FWE} yielding an adjusted significance level for each comparison as $\alpha_{DS} = 1-(1-\alpha_{FWE})^{1/C}$. Familywise error based on this adjustment α_{DS} is precisely equal to the pre-specified α_{FWE} if independence among pairs is assumed [9]; i.e., $1-(1-\alpha_{DS})^C = 1-((1-\alpha_{FWE})^{1/C})^C = \alpha_{FWE} = .05$ for all *c*. When this adjustment is implemented, the pairs of groups with *p*-values less than α_{DS} will be declared to differ significantly in

efficacy. Thus the Dunn-Šidák adjustment should have somewhat greater FWE than the Bonferroni adjustment because $\alpha_{DS} > \alpha_B$ for all k or c. For instance, FWE of the Dunn-Šidák adjustment is greater by 0.00083 and 0.00103 for $c = 3$ and $c = 6$, respectively, than that of the Bonferroni adjustment. Nevertheless, in practice the interpretation of results from the two approaches will seldom differ.

Hochberg's Step-Up Procedure

Individual pairwise comparisons are tested in *descending* order of the c number of p-values into $p_{(c)} > p_{(c-1)} > \ldots > p_{(1)}$. Each successively smaller p-value has a more rigorous α-threshold: $\alpha_{FWE} > \alpha_{FWE}/2 > \ldots > \alpha_{FWE}/(c-1) > \alpha_{FWE}/c$. For instance for c=3 and α_{FWE} =.05, the successive α-levels are .05, .025, and .01667. The sequential Hochberg tests do not reject null pairwise hypotheses until the first significant comparison with p-value $p_{(i)}$ such that $p_{(i)} < \alpha_{FWE}/i$, and then declares that all subsequent comparisons statistically significant without further examination of subsequent thresholds. In theory, familywise error of the sequentially rejective Hochberg's procedure should be greater than or equal to those based on the Bonferroni adjustment because the last threshold α_{FWE}/c of the Hochberg procedure is the Bonferroni-adjusted significance level. In other words, rejection with the Hochberg procedure is a necessary condition for the Bonferroni adjustment in rejecting at least one pair.

Unadjusted Post Hoc Logrank Test Procedure

As described in Keppel [17], the essential idea of the least significant difference testing procedure originally proposed by Fisher [18] involves two steps: 1) Test a global null hypothesis; 2) If the global null hypothesis is rejected, then proceed with pairwise comparisons using test statistics should be based on a pooled standard error yet with no alpha-threshold adjustment. In the context of ANOVA, this procedure is known to have weak control of FWE [19] despite the use of pooled standard error and increased degrees of freedom. We apply this concept to the present problem with survival distributions and call it "unadjusted post hoc logrank procedure." Specifically, this procedure does not adjust the pre-specified familywise significance level α_{FWE} for any of the post hoc pairwise comparisons when the global null hypothesis (1) is rejected by the omnibus logrank test. That is, the significance level $\alpha_{unadjusted}$ for each pair of groups is fixed as α_{FWE}, i.e., the function connecting $\alpha_{unadjusted}$ and α_{FWE} is the identity function yielding $\alpha_{unadjusted} = \alpha_{FWE}$. It follows that each pair of groups with a p-value < $\alpha_{unadjusted}$ will be declared to be significant. Furthermore, unlike post-hoc t-tests following ANOVA, the number of degrees of freedom for the pairwise logrank statistic are always one regardless of the number of observations. Nevertheless, we did not use the estimated variance of the score statistic of the omnibus test, but instead conducted simple and straightforward pairwise comparisons. We simply applied the standard two-group logrank tests for those post-hoc comparisons.

APPLICATION

The post hoc multiple comparison procedures are illustrated in an examination of the intervention effects in the Prevention of Suicide in Primary Care Elderly: Collaborative Trial (PROSPECT) study [20]. This study recruited subjects from May 1999 through August 2001 and followed them for up to 24 months. The primary report from the PROSPECT study evaluated the 24-month course of participants who were enrolled with a diagnosis of depression. The assessments were made at baseline and months 4, 8, 12, 18, and 24. Here, in this application of multiplicity adjustments for logrank test, we included subjects who met the following criteria: 1) diagnosed with major depressive disorder (MDD); 2) baseline 24-item Hamilton Depression Rating Scale (HDRS) [21] greater than 17; 3) available for baseline Clinical Anxiety Scale (CAS) [22]; 4) available for month 4 evaluations.

The analyses included 188 subjects (103 in the intervention arm and 85 in the control arm). The primary outcome was time to remission of depressive symptoms until month 18. The remission was defined as the 24-item HDRS < 10. Of particular interest was whether the intervention was more effective for subjects with anxiety measured with CAS compared to the conventional usual care control [1]. To illustrate the post hoc procedures that are compared here, subjects were classified into the following four groups: intervention with higher anxiety (IHA, N= 56), intervention with lower anxiety (ILA, N=47), control with higher anxiety (CHA, N= 46), and control with lower anxiety (CLA, N=39), where the higher vs. lower anxiety criterion was based on a whole sample median split of the baseline Clinical Anxiety Scale ratings (CAS>=6 vs. CAS<6). We acknowledge that alternatively, this could have been tested with a Cox proportional hazards model with an intervention by CAS interaction term.

The omnibus logrank test rejected the homogeneity of survival time distributions among the four groups with χ^2 = 11.615, df = 3, and p=0.0088. Figure **1** suggests that the intervention effect could depend upon severity of clinical anxiety symptoms. Table **1** shows subsequent pairwise comparison results. The unadjusted post hoc logrank test procedure indicated that the ILA subjects achieved a significantly higher remission rate than the other three groups of CHA, IHA, and CLA. Therefore, in this study, the intervention is particularly effective for subjects with low anxiety. In contrast, the results of the other three procedures indicate that ILA group achieved a significantly higher remission rate than the CHA group only. The interpretation of these results differs based on the post hoc testing procedure that is chosen. Thus a simulation study is now presented that compares the performance of those procedures.

SIMULATION DESIGN, PROCEDURES, AND EVALUATION CRITERIA

Simulation Design

A Monte Carlo simulation compares the performance of the approaches that have been described. The study design parameters that are typically seen in randomized clinical trials in psychopharmacology guided the choice of the following simulation specifications:

1. Unit of time T = day.
2. Length of trial in days L = 28 days (4 weeks).
3. Number of groups k = 3 for a trial with 3 arms of Placebo Control, Active Control Drug (A), and Investigational Drug (B); and k = 4 for a trial with 4 arms of Placebo Control, Drug A, Drug B, and Drug A plus B. (The corresponding numbers of post hoc pairwise comparisons for k=3 and k=4 are c=3 and 6, respectively.)
4. Number of subjects per group: N/group = 50, 100, and 200. Therefore, the total sample size will be kN subjects.
5. Remission rates π_R during a trial:
   a. Placebo Control: π_R = 5% or 20%;
   b. Active Control Drug (A): π_R = 30% (small), 40% (medium), or 50% (large);
   c. Investigational Drug (B): π_R = 30% (small), 40% (medium), or 50% (large).
6. The lost follow-up rate π_{LF} during the trial is fixed at 30% for each group.

Generation of Observed Survival Time

The distribution of a “true” survival time (time to remission) T is assumed to be exponential with a rate

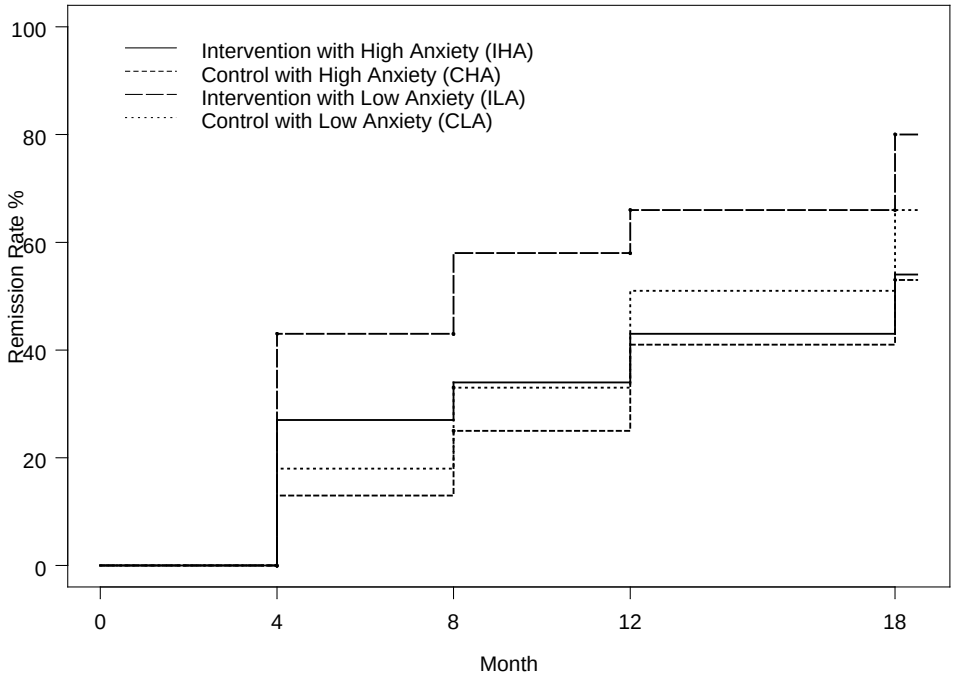

Figure 1: Cumulative remission rates (24-Item HDRS<10) in MDD subjects in the PROSPECT study with higher and lower Clinical Anxiety Scale (CAS) scores than the sample median.

Table 1: Post Hoc Pairwise Comparisons with the PROSPECT Study Subjects

Logrank			Post hoc multiple comparison procedures			
Pairs	**$\chi^2(1)$**	**p**	**Unadjusted**	**Bonferroni**	**Dunn-Šidák**	**Hochberg**
ILA vs. CHA	9.687	.0019	*	*	*	*
ILA vs. IHA	6.410	.0113	*	ns	ns	ns
ILA vs. CLA	3.948	.0469	*	ns	ns	ns
CLA vs. CHA	1.116	.2908	ns	ns	ns	ns
IHA vs. CHA	0.384	.5356	ns	ns	ns	ns
IHA vs. CLA	0.184	.6673	ns	ns	ns	ns

IHA: intervention with high anxiety; ILA: intervention with low anxiety; CHA: control with high anxiety; CLA: control with low anxiety.

parameter λ. The global null hypothesis (1) can then be expressed as follows:

$$H_0: \lambda_1 = \lambda_2 = ... = \lambda_k. \quad (6)$$

We also assume that a censoring time C is also exponentially distributed with a rate parameter ζ. We further assume that T and C are independent to each other although the practicality of this assumption is not always reasonable. Then the observed survival time $Y = \min(T,C)$ is also exponentially distributed with a rate parameter ξ, which is $\lambda + \zeta$. It follows that $P_\xi(Y<L) = 1 - \exp(-\xi L)=\pi_R+\pi_{LF}$ for $\xi = -\log(\pi_R+\pi_{LF})/L$. Furthermore, by definition, $\pi_R = P(T<C \ \& \ Y<L)$ and $\pi_{LF} = P(C<T \ \& \ Y<L)$. Therefore, $P(T<C|Y<L) = \pi_R/(\pi_R+\pi_{LF})$ and $P(C<T|Y<L) = \pi_{LF}/(\pi_R+\pi_{LF})$. It follows that once Y is drawn from $\exp(\xi)$ and if $Y<L$, then the distribution of the event indicator $D = 1(Y=T|Y<L)$ before the end of the study is Bernoulli with the "success" probability of $\pi_R/(\pi_R+\pi_{LF})$, where 1( ) is an indicator function that returns 1 if the condition in ( ) is satisfied or 0 otherwise. Therefore, we drew first randomly Y, and if $Y < L$ then we determined D based on a random draw from the Bernoulli distribution. For subjects with $Y > L$, we fixed $D = 0$. Collectively, the observed survival time will be $Y= \min(T, C)$ with event indicator D defined as 1 if "remitted" and 0 otherwise. Therefore, the subjects with time Y with $D = 0$ will be considered as *censored* or lost at time Y. The subjects with time Y with $D = 1$ will be considered as *remitted* at time Y or T. We generated $N_{sim} = 10{,}000$ data sets for each simulation specification described above.

Post Hoc Testing Procedures

We first applied the omnibus logrank χ^2 test with k-1 df to simulated data to test the global null hypothesis H_0 (6), the equality of the rate parameters among the k groups. Only if the omnibus test rejected H_0, did we proceed with a logrank χ^2 test statistic with 1 df for each of the c pairs of groups with a null hypothesis H_{0ij}: $\lambda_i = \lambda_j$ where $i < j \le k$, and reserved p-values. Each of the post hoc procedures was applied to evaluate the p-values for each pairwise comparison.

Evaluation Criteria

The following four criteria were evaluated based on simulation results. In each case, the denominator of the respective proportions or means was the total number of simulated data sets $N_{sim} = 10{,}000$. It was not based on number of rejected omnibus tests.

1. Type I error: This was computed as a proportion of simulations in which at least one pairwise null hypothesis was rejected following a significant omnibus test when, in fact, the null hypothesis (6) of equal efficacy is true, or $H_{0ij} \in F_0=F$ for all $i < j \le k$ as defined in equation (3).

2. Number of *correctly* rejected pairs: For each simulated data set, numbers of correctly rejected pairs (4) following a significant omnibus test were counted. These counts were averaged over the number of simulations. When these results are presented, the number of true false null hypotheses will be shown for reference.

3. Correct decision rate (CDR): The CDR was computed as a proportion of simulations in which a post hoc procedure rejects all of the false null hypotheses, but rejects only the false null hypotheses as defined in equation (2) following a significant omnibus test result.

4. Empirical false discovery rates (FDR): For each simulated data set, empirical FDR was computed as number of falsely rejected pairs divided by

number of all rejected pairs as defined in equation (5) following a significant omnibus test result.

Of note, the total number of rejected null hypothesis, that is #{Rejected H_{0ij}| $H_{0ij} \in F$}, can easily be obtained from the number of *correctly* rejected pairs and the empirical FDR as:

$$\#\{\text{Rejected } H_{0ij} | H_{0ij} \in F\} = \#\{\text{Rejected } H_{0ij} | H_{0ij} \in F_1 \subset F\}/(1 - \text{FDR}). \quad (7)$$

SIMULATION STUDY RESULTS

Type I Error

The empirically estimated type I error rates (Table **2**) represent familywise error conditional upon rejection of a global null hypothesis. By design, the type I error of each post hoc procedure to reject at least one pairwise hypothesis never exceeded that of the omnibus test. This, of course, is a function of the condition in the simulation study that required rejection of the omnibus test prior to pairwise testing. For the

Table 2: Type I Error for Post Hoc Pairwise Comparisons Based on N_{sim}=10,000 Simulated Data Sets.

k	N/group	π_R	Omnibus	Post hoc multiple comparison procedures			
				Unadjusted	Bonferroni	Dunn-Šidák	Hochberg
3	50	(.05, .05, .05)	0.042	0.038	0.017	0.017	0.017
		(.20, .20, .20)	0.047	0.047	0.036	0.037	0.037
		(.30, .30, .30)	0.052	0.052	0.043	0.043	0.044
		(.40, .40, .40)	0.050	0.050	0.040	0.041	0.040
		(.50, .50, .50)	0.054	0.054	0.043	0.044	0.044
	100	(.05, .05, .05)	0.047	0.047	0.033	0.033	0.033
		(.20, .20, .20)	0.050	0.050	0.041	0.041	0.041
		(.30, .30, .30)	0.050	0.050	0.042	0.043	0.043
		(.40, .40, .40)	0.052	0.052	0.044	0.044	0.045
		(.50, .50, .50)	0.051	0.051	0.042	0.043	0.042
	200	(.05, .05, .05)	0.051	0.051	0.038	0.039	0.039
		(.20, .20, .20)	0.055	0.055	0.046	0.046	0.047
		(.30, .30, .30)	0.052	0.052	0.044	0.045	0.045
		(.40, .40, .40)	0.048	0.048	0.040	0.041	0.041
		(.50, .50, .50)	0.052	0.052	0.042	0.042	0.043
4	50	(.05,.05,.05,.05)	0.043	0.042	0.010	0.010	0.010
		(.20,.20,.20,.20)	0.055	0.055	0.035	0.036	0.036
		(.30,.30,.30,.30)	0.052	0.052	0.035	0.036	0.035
		(.40,.40,.40,.40)	0.056	0.056	0.039	0.039	0.039
		(.50,.50,.50,.50)	0.058	0.058	0.040	0.040	0.040
	100	(.05,.05,.05,.05)	0.044	0.044	0.024	0.024	0.024
		(.20,.20,.20,.20)	0.056	0.056	0.039	0.039	0.039
		(.30,.30,.30,.30)	0.053	0.053	0.038	0.039	0.038
		(.40,.40,.40,.40)	0.049	0.049	0.037	0.037	0.037
		(.50,.50,.50,.50)	0.052	0.052	0.040	0.040	0.040
	200	(.05,.05,.05,.05)	0.048	0.048	0.032	0.033	0.032
		(.20,.20,.20,.20)	0.048	0.048	0.034	0.035	0.035
		(.30,.30,.30,.30)	0.052	0.052	0.038	0.038	0.038
		(.40,.40,.40,.40)	0.052	0.052	0.039	0.039	0.039
		(.50,.50,.50,.50)	0.051	0.051	0.038	0.039	0.039

most part, the unadjusted post hoc logrank test procedure without multiplicity adjustment has the identical type I error as the omnibus test as is expected based on theory [8]. The type I error rates for the other post hoc procedures are lower than that of the omnibus test. By definition, Bonferroni procedure has smaller type I error rates than both Dunn-Šidák adjustment and Hochberg procedure as demonstrated in Table **2**. Overall, for a given set of simulation specifications, type I error for each of the three procedures with multiplicity adjustments is in general greater for larger N/group, for larger remission rates, and for smaller *k*.

Number of Correctly Rejected Pairs

Table **3** presents the number of *correctly* rejected pairs for N/group =200. The order of the number of correctly rejected pairs is as follows: unadjusted post

Table 3: Number of Correctly Rejected Pairs for Post Hoc Pairwise Comparisons Based on N_{sim}=10,000 Simulated Data Sets (N/group = 200)

k	π_R	Unequal	Post hoc multiple comparison procedures			
		pairs	Unadjusted	Bonferroni	Dunn-Šidák	Hochberg
3	(.05, .2, .3)	3	2.708	2.540	2.542	2.708
	(.05, .3, .3)	2	2.000	2.000	2.000	2.000
	(.05, .2, .4)	3	2.998	2.988	2.989	2.998
	(.05, .3, .4)	3	2.669	2.501	2.504	2.669
	(.05, .4, .4)	2	2.000	2.000	2.000	2.000
	(.05, .2, .5)	3	2.999	2.994	2.994	2.999
	(.05, .3, .5)	3	2.999	2.996	2.996	2.999
	(.05, .4, .5)	3	2.721	2.556	2.559	2.721
	(.05, .5, .5)	2	2.000	2.000	2.000	2.000
	(.2, .2, .3)	2	1.285	1.059	1.063	1.135
	(.2, .3, .3)	2	1.263	1.063	1.066	1.139
	(.2, .2, .4)	2	1.996	1.988	1.988	1.993
	(.2, .3, .4)	3	2.369	2.016	2.021	2.315
	(.2, .4, .4)	2	1.995	1.987	1.987	1.991
	(.2, .2, .5)	2	2.000	2.000	2.000	2.000
	(.2, .3, .5)	3	2.700	2.538	2.541	2.700
	(.2, .4, .5)	3	2.719	2.552	2.553	2.719
	(.2, .5, .5)	2	2.000	2.000	2.000	2.000
4	(.05, .3, .3, .3)	3	3.000	3.000	3.000	3.000
	(.05, .3, .3, .4)	5	4.349	3.819	3.825	4.082
	(.05, .3, .3, .5)	5	4.998	4.979	4.980	4.995
	(.05, .3, .4, .4)	5	4.334	3.809	3.815	4.068
	(.05, .3, .4, .5)	6	5.390	4.859	4.864	5.345
	(.05, .4, .4, .4)	3	3.000	3.000	3.000	3.000
	(.05, .4, .4, .5)	5	4.451	3.930	3.937	4.199
	(.05, .4, .5, .5)	5	4.437	3.920	3.925	4.191
	(.05, .5, .5, .5)	3	3.000	3.000	3.000	3.000
	(.2, .3, .3, .3)	3	1.810	1.290	1.298	1.377
	(.2, .3, .3, .4)	5	3.735	2.668	2.679	3.029
	(.2, .3, .3, .5)	5	4.411	3.856	3.862	4.142
	(.2, .3, .4, .4)	5	4.041	3.219	3.228	3.558
	(.2, .3, .4, .5)	6	5.091	4.282	4.291	4.922
	(.2, .4, .4, .4)	3	2.994	2.963	2.964	2.975
	(.2, .4, .4, .5)	5	4.428	3.890	3.897	4.168
	(.2, .4, .5, .5)	5	4.448	3.911	3.917	4.200
	(.2, .5, .5,. 5)	3	3.000	3.000	3.000	3.000

hoc logrank test procedure > Hochberg >> Dunn-Šidák > Bonferroni, where the symbol ">>" reads "much greater." This ordering was consistent regardless of configurations of remission rates π_R. The overall pattern reflected in Table **3** was similar for N/group = 50 and 100 (not presented), albeit with a smaller number of rejected pairs with a smaller N/group. In summary, the unadjusted post hoc logrank test procedure correctly rejected greater number of false null hypotheses than the other procedures. In this regard, the Hochberg procedure rejected more than both Dunn-Šidák and Bonferroni procedures, which are close each other.

Correct Decision Rates

Table **4** presents the CDR for N/group = 200. The pattern of the results is similar to that above in the

Table 4: Correct Decision Rate (CDR) for Post Hoc Pairwise Comparisons Based on N_{sim}=10,000 Simulated Data Sets (N/group = 200)

k	π_R	Post hoc multiple comparison procedures			
		Unadjusted	Bonferroni	Dunn-Šidák	Hochberg
3	(.05, .2, .3)	0.708	0.540	0.542	0.708
	(.05, .3, .3)	1.000	1.000	1.000	1.000
	(.05, .2, .4)	0.998	0.988	0.989	0.998
	(.05, .3, .4)	0.669	0.501	0.504	0.669
	(.05, .4, .4)	1.000	1.000	1.000	1.000
	(.05, .2, .5)	0.999	0.994	0.994	0.999
	(.05, .3, .5)	0.999	0.996	0.996	0.999
	(.05, .4, .5)	0.721	0.556	0.559	0.721
	(.05, .5, .5)	1.000	1.000	1.000	1.000
	(.2, .2, .3)	0.546	0.366	0.368	0.436
	(.2, .3, .3)	0.542	0.366	0.369	0.437
	(.2, .2, .4)	0.997	0.989	0.989	0.994
	(.2, .3, .4)	0.412	0.179	0.182	0.412
	(.2, .4, .4)	0.996	0.988	0.988	0.992
	(.2, .2, .5)	1.000	1.000	1.000	1.000
	(.2, .3, .5)	0.700	0.539	0.541	0.700
	(.2, .4, .5)	0.719	0.552	0.554	0.719
	(.2, .5, .5)	1.000	1.000	1.000	1.000
4	(.05, .3, .3, .3)	1.000	1.000	1.000	1.000
	(.05, .3, .3, .4)	0.528	0.247	0.250	0.406
	(.05, .3, .3, .5)	0.998	0.980	0.981	0.996
	(.05, .3, .4, .4)	0.515	0.242	0.244	0.397
	(.05, .3, .4, .5)	0.429	0.111	0.113	0.429
	(.05, .4, .4, .4)	1.000	1.000	1.000	1.000
	(.05, .4, .4, .5)	0.588	0.297	0.300	0.465
	(.05, .4, .5, .5)	0.580	0.292	0.295	0.462
	(.05, .5, .5, .5)	1.000	1.000	1.000	1.000
	(.2, .3, .3, .3)	0.451	0.189	0.190	0.239
	(.2, .3, .3, .4)	0.262	0.044	0.045	0.146
	(.2, .3, .3, .5)	0.560	0.265	0.269	0.439
	(.2, .3, .4, .4)	0.289	0.039	0.040	0.150
	(.2, .3, .4, .5)	0.248	0.018	0.019	0.248
	(.2, .4, .4, .4)	0.995	0.967	0.968	0.978
	(.2, .4, .4, .5)	0.572	0.282	0.285	0.449
	(.2, .4, .5, .5)	0.584	0.283	0.286	0.462
	(.2, .5, .5,. 5)	1.000	1.000	1.000	1.000

Table 5: False Discovery Rate (FDR) for Post Hoc Pairwise Comparisons Based on N_{sim}=10,000 Simulated Data Sets (k=3)

N/group	π_R	Post hoc multiple comparison procedures			
		Unadjusted	Bonferroni	Dunn-Šidák	Hochberg
50	(.05, .3, .3)	0.025	0.009	0.009	0.023
	(.05, .4, .4)	0.024	0.009	0.009	0.024
	(.05, .5, .5)	0.024	0.008	0.008	0.024
	(.2, .2, .3)	0.080	0.053	0.054	0.072
	(.2, .3, .3)	0.108	0.062	0.062	0.073
	(.2, .2, .4)	0.036	0.015	0.016	0.028
	(.2, .4, .4)	0.040	0.016	0.017	0.028
	(.2, .2, .5)	0.025	0.009	0.009	0.023
	(.2, .5, .5)	0.024	0.008	0.008	0.023
100	(.05, .3, .3)	0.024	0.009	0.009	0.024
	(.05, .4, .4)	0.025	0.009	0.009	0.025
	(.05, .5, .5)	0.024	0.008	0.008	0.024
	(.2, .2, .3)	0.052	0.028	0.028	0.040
	(.2, .3, .3)	0.062	0.027	0.027	0.038
	(.2, .2, .4)	0.028	0.009	0.009	0.024
	(.2, .4, .4)	0.026	0.009	0.009	0.023
	(.2, .2, .5)	0.026	0.009	0.010	0.026
	(.2, .5, .5)	0.024	0.009	0.009	0.024
200	(.05, .3, .3)	0.026	0.009	0.009	0.026
	(.05, .4, .4)	0.022	0.008	0.008	0.022
	(.05, .5, .5)	0.024	0.008	0.008	0.024
	(.2, .2, .3)	0.038	0.016	0.016	0.029
	(.2, .3, .3)	0.040	0.015	0.015	0.029
	(.2, .2, .4)	0.023	0.008	0.008	0.023
	(.2, .4, .4)	0.024	0.009	0.009	0.023
	(.2, .2, .5)	0.024	0.008	0.008	0.024
	(.2, .5, .5)	0.025	0.008	0.009	0.025

number of *correctly* rejected pairs. Again, the order of CDR is the same regardless of configurations of remission rates as π_R: unadjusted post hoc logrank test procedure > Hochberg >> Dunn-Šidák > Bonferroni. Likewise, the overall pattern reflected in Table **3** was similar for N/group = 50 and 100 (not presented), albeit with lower CDRs. In general, CDR was higher when k=3 than when k=4 perhaps in part because the former case has smaller number of pairwise comparisons. However, the CDR depends on differences in remission rates between i-th and j-th groups. For instance, regardless of k for the same remission rates of the other active or investigative drugs, all procedures have (much) higher CDR for remission rate of placebo = 0.05 than for remission rate of placebo = 0.20, where the former case has pairs with greater differences in remission rates for the same number of unequal pairs.

Empirical False Discovery Rates

Tables **5** and **6** present the empirically estimated FDR for k = 3 and 4, respectively. For a given configuration of remission rates, the ordering of FDR among the procedures was consistent with those of CDR and the number of correctly rejected pairs, but was not necessarily consistent with that of type I error rate (Table **2**). For instance, the Hochberg procedure

Table 6: False Discovery Rate (FDR) for Post Hoc Pairwise Comparisons Based on N_{sim}=10,000 Simulated Data Sets (k=4)

N/group	π_R	Post hoc multiple comparison procedures			
		Unadjusted	Bonferroni	Dunn-Šidák	Hochberg
50	(.05, .3, .3, .3)	0.049	0.010	0.010	0.017
	(.05, .3, .3, .4)	0.015	0.003	0.004	0.007
	(.05, .3, .3, .5)	0.011	0.002	0.002	0.006
	(.05, .3, .4, .4)	0.015	0.003	0.003	0.007
	(.05, .4, .4, .4)	0.049	0.009	0.009	0.018
	(.05, .4, .4, .5)	0.015	0.003	0.003	0.007
	(.05, .4, .5, .5)	0.014	0.003	0.003	0.007
	(.05, .5, .5, .5)	0.047	0.008	0.009	0.018
	(.2, .3, .3, .3)	0.204	0.091	0.091	0.100
	(.2, .3, .3, .4)	0.034	0.013	0.013	0.016
	(.2, .3, .3, .5)	0.018	0.005	0.005	0.008
	(.2, .3, .4, .4)	0.027	0.007	0.008	0.010
	(.2, .4, .4, .4)	0.079	0.019	0.020	0.026
	(.2, .4, .4, .5)	0.018	0.004	0.004	0.007
	(.2, .4, .5, .5)	0.016	0.004	0.004	0.007
	(.2, .5, .5,. 5)	0.051	0.010	0.010	0.018
100	(.05, .3, .3, .3)	0.047	0.007	0.008	0.017
	(.05, .3, .3, .4)	0.014	0.003	0.003	0.007
	(.05, .3, .3, .5)	0.010	0.002	0.002	0.009
	(.05, .3, .4, .4)	0.014	0.003	0.003	0.008
	(.05, .4, .4, .4)	0.051	0.009	0.009	0.021
	(.05, .4, .4, .5)	0.013	0.002	0.002	0.007
	(.05, .4, .5, .5)	0.013	0.002	0.002	0.007
	(.05, .5, .5, .5)	0.048	0.008	0.009	0.018
	(.2, .3, .3, .3)	0.129	0.047	0.048	0.056
	(.2, .3, .3, .4)	0.020	0.005	0.005	0.008
	(.2, .3, .3, .5)	0.013	0.003	0.003	0.006
	(.2, .3, .4, .4)	0.016	0.003	0.003	0.006
	(.2, .4, .4, .4)	0.052	0.010	0.011	0.018
	(.2, .4, .4, .5)	0.013	0.003	0.003	0.006
	(.2, .4, .5, .5)	0.013	0.003	0.003	0.007
	(.2, .5, .5, .5)	0.047	0.009	0.009	0.019
200	(.05, .3, .3, .3)	0.048	0.008	0.009	0.018
	(.05, .3, .3, .4)	0.011	0.002	0.002	0.008
	(.05, .3, .3, .5)	0.010	0.002	0.002	0.010
	(.05, .3, .4, .4)	0.011	0.002	0.002	0.007
	(.05, .4, .4, .4)	0.049	0.008	0.008	0.019
	(.05, .4, .4, .5)	0.011	0.003	0.003	0.008
	(.05, .4, .5, .5)	0.012	0.002	0.002	0.008
	(.05, .5, .5, .5)	0.047	0.008	0.008	0.018
	(.2, .3, .3, .3)	0.072	0.017	0.017	0.024
	(.2, .3, .3, .4)	0.014	0.003	0.003	0.007
	(.2, .3, .3, .5)	0.011	0.002	0.002	0.007
	(.2, .3, .4, .4)	0.012	0.003	0.003	0.007
	(.2, .4, .4, .4)	0.049	0.009	0.009	0.018
	(.2, .4, .4, .5)	0.011	0.002	0.002	0.008
	(.2, .4, .5, .5)	0.011	0.002	0.002	0.007
	(.2, .5, .5, .5)	0.047	0.008	0.008	0.018

has greater FDRs than Dunn-Šidák procedure but they have lower type I error rates, especially when k=4.

DISCUSSION

A summary of the findings from this simulation study is as follows. The three post hoc procedures for multiplicity adjustments are relatively conservative even when conditional upon rejection of the global null hypothesis. However, this does not apply to the unadjusted post hoc logrank test procedure. The sequentially rejective Hochberg procedure generally has greater CDR than the other two adjustment procedures (Bonferroni and Dunn-Šidák), but less than the unadjusted post hoc logrank test procedure. Similarly, the Hochbergh procedure always correctly rejected more pairs than the other adjustment procedures, but less than the unadjusted post hoc logrank test procedure. When not protected, even if the number of comparisons are as small as 4, the family wise type I error rate is as high as 0.18 if the size of test for each pair is 0.05. Therefore, the inflation of type I error may not be moderate. However, when protected or when conditional upon rejection of the omnibus testing, the performance of the unadjusted procedure is good compared with the other adjusted procedures. Nevertheless, application of these findings should be limited to the use of post-hoc pairwise logrank tests within the ranges of k and c considered here; that is $3 \leq k \leq 4$ and $3 \leq c \leq 6$.

The three procedures with multiplicity adjustments have FDR less than 0.05 except for a few cases particularly with small sample size N/group = 50. However, the unadjusted post hoc logrank test procedure has FDR > 0.05 when both 1) number of null pairs is large relative to the total number of pairwise comparisons and 2) remission rates of non-null pairs are close each other. In these circumstances, the numbers of *all* rejected pairs (7) tended to be greater than those of true unequal pairs for the unadjusted post hoc logrank test procedure and for the other four post hoc comparison procedures as well. Nevertheless, the numbers of *all* rejected pairs (7) are not presented since they can be obtained from FDR (5) and the number of *correctly* rejected pairs (4).

In theory, the results might very well apply to any post hoc pairwise comparisons following rejection of other omnibus tests. For instance, an omnibus 3 df Pearson χ^2 can be applied to test homogeneity of proportions of a four category outcome between two groups in a 4 by 2 contingency table. Upon the rejection by the omnibus test, a data analyst can go forward with four 1 df Pearson χ^2 tests for four 2 by 2 contingency tables to identify specific outcome categories that differ in proportions between the two groups. In addition, the results could also apply to the case where the clinical trial protocol specifies *a priori* limited number of contrasts of interest but not all possible combinations of contrasts. In short, once the number of potential post hoc comparisons (i.e., the size of a family F) is given, the multiplicity adjustments can accordingly be adjusted except for the unadjusted post hoc procedure.

The conservative nature (beyond conditioning upon the rejection of the omnibus test) of the three procedures with multiplicity adjustments may stem from the correlations among test statistics for pairwise comparisons. That is, the correlations among outcomes may not be zero when pairwise comparisons involve the same group. For instance, when k = 3, all of the c = 3 pairwise comparisons may be correlated each other unlike post hoc comparisons following ANOVA with normal distributions. (In this latter case, the three pairwise group mean comparisons (say, μ_1 vs. μ_2, μ_1 vs. μ_2, and μ_2 vs. μ_3) are uncorrelated each other if $\text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(X_3)$, where X_i and μ_i represent a normal outcome variable and its mean for the i-th group.) When such correlations exist, the four multiple adjustment procedures are even more conservative compared to uncorrelated pairwise comparisons, yet that distinction is most apparent when correlation of outcomes >.5 [23]. Therefore, an investigation of procedures that explicitly account for the correlation among pairwise comparisons would be of value. For instance, adoption of the James approach [24] for post hoc multiple comparisons will likely be less conservative, as shown with correlated binary end points [23]. This might identify a procedure that would increase CDR without sacrificing FDR.

It is noteworthy, however, that the type I error (3) and FDR are not necessarily consistent across the procedures, particularly between the Hochberg procedure and the Bonferroni or Dunn-Šidák procedure. That is, one procedure can have comparable type I error but lower FDR. This inconsistency may be due to the fact that the Hochberg rejective procedure is based on varying threshold for ordered p-values whereas the Bonferroni or Dunn-Šidák procedure is based on a fixed threshold for all p-values. That is, the Hochberg procedure rejects pairwise hypotheses even in cases when their p-values exceed the fixed thresholds of Bonferroni or Dunn-

Šidák. Nonetheless, the number of *all* or *correctly* rejected pairs (4) and FDR are consistent across the procedures. FDR is strongly correlated with number of "null" pairs within each procedure.

In clinical trials, it is rare to compare more than four groups in any biomedical research field. For this reason, we considered only k = 3 or 4 in this paper. However, we caution that extrapolation of the present study findings to *unconditional* comparisons is unwarranted. Often in applied settings, a global null hypothesis is not of interest. Instead, there is a family of numerous null hypotheses such as those seen in neuroimaging studies (e.g., [25]) and in microarray analyses in genetic bioinformatics (e.g., [26]). Thus, in such unconditional comparisons, the unadjusted post hoc procedure should be avoided due to inflated type I error (e.g., [23]) and excessively inflated FDR. Nevertheless, we believe that a decision as to whether to apply conditional or unconditional procedures should depend on the context of studies or investigations more than on statistical considerations. For examples, if a trial intends to test effects of specific pairs of drugs targeted *a priori* in a multiple arm setting, then unconditional adjustments should be applied and power computations should be based on individual pairwise testing with adjusted significance level.

There are limitations to our simulation findings. For instance, it is unknown how relatively large number of equal pairs compared to the number of k (>2) groups would be associated with large FDR (e.g. >.05) for the unadjusted post hoc logrank test procedure. Therefore, when many sample parameter estimates among groups of interest are (or expected to be) relatively similar, the use of the unadjusted post hoc logrank test procedure may be discouraged. In which case, the Hochberg procedure may be preferred, which has low FDR and the highest CDR among the others. Finally, the study findings are based on empirical simulations with limited scenario of situations rather than on theoretical derivations.

In conclusion, we demonstrated that if conditioned on the rejection of the global hypothesis, the unadjusted post hoc logrank test procedure performs acceptably with regard to type I error, number of correctly rejected pairs, and correct decision rate, but less well on false discovery rate. Therefore, when conditioned upon the rejection of a global hypothesis any adjustment of the significance levels may not be necessary, and can be overly conservative. Nevertheless, if conditional adjustments are necessary or warranted, the Hochberg procedure may be preferred.

ACKNOWLEDGEMENT

This paper is dedicated in memory of the late Dr. Andrew Leon who unexpectedly passed away during the preparation of this manuscript. His death is a significant loss particularly to the field of statistics in psychiatry. Both authors were grateful to the PROSPECT study group for providing access to their data (R01MH59366, R01 MH59380, R01 MH59381). This research was supported in part by NIMH grants R01MH060447 (PI: Dr. Leon) and P30MH068638.

REFERENCES

[1] Alexopoulos GS, Katz IR, Bruce ML, Heo M, Ten Have TR, Raue PJ, *et al*. and the PROSPECT Group. Remission in depressed geriatric primary care patients: a report from the PROSPECT study. Am J Psychiatry 2005; 62: 718-24. http://dx.doi.org/10.1176/appi.ajp.162.4.718

[2] Reynolds CF 3rd, Frank E, Perel JM, Imber SD, Cornes C, Miller MD, *et al*. Nortriptyline and interpersonal psychotherapy as maintenance therapies for recurrent major depression: a randomized controlled trial in patients older than 59 years. J Am Med Assoc 1999; 281: 39-45. http://dx.doi.org/10.1001/jama.281.1.39

[3] Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, *et al.* for the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Investigators. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. New Engl J Med 2005; 353: 1209-23. http://dx.doi.org/10.1056/NEJMoa051688

[4] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966; 50: 163-70.

[5] Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). J Royal Statist Soc A 1972; 135: 185-206. http://dx.doi.org/10.2307/2344317

[6] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75: 800-803. http://dx.doi.org/10.1093/biomet/75.4.800

[7] Logan BR, Wang H, Zhang M-J. Pairwise multiple comparison adjustment in survival analysis. Statist Med 2005; 24: 2509-23. http://dx.doi.org/10.1002/sim.2125

[8] Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special referrnce to ordered analysis of variance. Biometrika 1976; 63: 655-60. http://dx.doi.org/10.1093/biomet/63.3.655

[9] Chen Y-I. Multiple comparisons in carcinogenesis study with right-censored survival data. Statist Med 2000; 19: 353-67. http://dx.doi.org/10.1002/(SICI)1097-0258(20000215)19:3<353::AID-SIM333>3.0.CO;2-B

[10] Slepian D. The one-sided barrier problem for Gaussian noise. Bell Syst Tech J 1962; 41: 463-501.

[11] Steel RGD. A multiple comparison rank sum test: treatments versus control. Biometrics 1959; 15; 560-72. http://dx.doi.org/10.2307/2527654

[12] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 1965; 52: 203-23.

[13] Prentice RL. Linear rank tests with right censored data. Biometrika 1978; 65: 165-79. http://dx.doi.org/10.1093/biomet/65.1.167

[14] Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. J Am Statist Assoc 1967; 62: 626-33.

[15] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statist Soc B 1995; 57: 289-300.

[16] Ury HK. A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes. Technometrics 1976; 18: 89-97. http://dx.doi.org/10.2307/1267921

[17] Keppel G. Design & Analysis: A Resercher's Handbook, Englewood Cliffs, NJ: Prentice Hall 1982; pp. 157-159.

[18] Fisher RA. The Design of Experiments, Oliver & Boyd: Edinburgh 1935.

[19] Cook RJ, Dunnett CW. Multiple comparisons, in Encyclopedia of Biostatistics P Armitage and T Colton (eds.) Chichester, UK: John Wiley and Sons 1998; p. 2739.

[20] Bruce ML, Ten Have TR, Reynolds CF 3rd, Katz II, Schulberg HC, Mulsant BH, *et al*. Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. J Am Med Assoc 2004; 291: 1081-91. http://dx.doi.org/10.1001/jama.291.9.1081

[21] Hamilton M. A rating scale for depression, J Neurol Neurosurg Psychiatry 1960; 23: 56-62. http://dx.doi.org/10.1136/jnnp.23.1.56

[22] Snaith RP, Baugh SJ, Clayden AD, Husain A, Sipple MA. The Clinical Anxiety Scale: An instrument derived from the Hamilton Anxiety Scale. Br J Psychiatry 1982; 141: 518-23. http://dx.doi.org/10.1192/bjp.141.5.518

[23] Leon AC, Heo M. A comparison of multiplicity adjustment strategies for correlated binary endpoints with application to a study of homicide victims. J Biopharmaceut Statist 2005; 15: 839-55. http://dx.doi.org/10.1081/BIP-200067922

[24] James S. The approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. Statist Med 1991; 10: 1123-35. http://dx.doi.org/10.1002/sim.4780100712

[25] Nichols TE, Hayasaka S. Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. Statist Methods Med Res 2003; 12: 419-46. http://dx.doi.org/10.1191/0962280203sm341ra

[26] Allison DB, Gadbury G, Heo M, Fernandez J, Prolla TA, Lee CK, Weindruch R. Statistical methods for the analysis of microarray gene expression data. Comput Statist Data Analysis 2002; 39: 1-20. http://dx.doi.org/10.1016/S0167-9473(01)00046-9

Received on 14-12-2012 Accepted on 19-04-2013 Published on 30-04-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.02.04