Published in final edited form as:
Drug Inf J. 2010 July 1; 44(4): 405–420.

SCORE Study Report 8: Closed Tests for All Pair-Wise Comparisons of Means

Neal Oden, PhD[1], **Paul C. VanVeldhuisen, PhD**[1], **Ingrid U. Scott, MD, MPH**[2], **Michael S. Ip, MD**[3], and **the SCORE Study Investigator Group**

[1] The EMMES Corporation, Rockville, MD

[2] Departments of Ophthalmology and Public Health Sciences, Penn State College of Medicine, Hershey, PA

[3] Fundus Photograph Reading Center, University of Wisconsin, Madison, WI

Abstract

We compare five closed tests for strong control of family-wide type I error (FWE) while making all pair-wise comparisons of means in clinical trials with multiple arms such as the SCORE Study. We simulated outcomes of the SCORE Study under its design hypotheses, and used p-values from chi-squared tests to compare performance of a "pairwise" closed test described below to Bonferroni and Hochberg adjusted p-values. "Pairwise" closed testing was more powerful than Hochberg's method by several definitions of multiple-test power. Simulations over a wider parameter space, and considering other closed methods, confirmed this superiority for p-values based on normal, logistic, and Poisson distributions. The power benefit of "pair-wise" closed testing begins to disappear with 5 or more arms, and with unbalanced designs. For trials with 4 or fewer arms and balanced designs, investigators should consider using "pair-wise" closed testing in preference to Shaffer's, Hommel's, and Hochberg's approaches when making all pairwise comparisons of means. If not all p-values from the closed family are available, Shaffer's method is a good choice.

Keywords

Closed testing; Hochberg's method; Hommel's method; Shaffer's Method; Bonferroni; family-wide error; multiple testing; strong FWE control

Introduction

The Standard Care versus Corticosteroid for Retinal Vein Occlusion (SCORE) Study comprised two independent randomized controlled clinical trials designed to compare the safety and efficacy of standard care (SC) versus intravitreal injection(s) of triamcinolone acetonide (hereafter referred to as intravitreal triamcinolone) for treating vision loss associated with macular edema in the study eye of participants with retinal vein occlusion. One trial involved patients with central retinal vein occlusion (CRVO), and the other involved patients with branch retinal vein occlusion (BRVO).

From the standpoint of statistical analysis, each trial was treated separately. Each trial contained three arms with equal sample sizes: SC, 1 mg intravitreal triamcinolone, and 4 mg intravitreal triamcinolone. The primary efficacy outcome measure was the proportion of study eyes

Corresponding author: Neal Oden, PhD, The EMMES Corporation, 401 N. Washington St., Suite 700, Rockville, MD 20850, Telephone: 301-251-1161 ext 107; Fax: 301-251-1355; noden@emmes.com.

improving from baseline to the 12-month follow-up visit by at least 15 letters as determined by Electronic Early Treatment Diabetic Retinopathy Study (E-ETDRS) visual acuity testing. There was only one study eye per patient. All three pairwise comparisons between arms were of interest, with the primary analysis method being logistic regression, adjusting for baseline visual acuity, clinical site, and, in participants with BRVO, presence of baseline dense macular hemorrhage.

When, as in the SCORE Study, several null hypotheses are tested simultaneously, the probability of rejecting at least one true null is commonly referred to as the family-wide type I error rate (FWE). In the original SCORE Study statistical analysis plan (SAP), FWE was to be controlled by means of Hochberg's sequentially rejective method [1] as applied to the p-values of the three pair-wise logistic contrasts.

Numerous methods can be applied to the problem of making all pair-wise comparisons of means while still controlling FWE. Examples include the techniques of Bonferroni, Sidak [2], and Fisher [3], sequentially-rejective approaches [1,4–7], branch-and-bound algorithms [8], resampling methods [9], and ANOVA-related techniques [10–12]. Among important ways in which these approaches differ are power, intensity of computational effort, and extent to which correlations between estimates are considered.

One might expect methods that take into account correlation between pairs of estimates to be more powerful than methods that ignore it. Indeed, Grechanovsky and Hochberg [13] discuss conditions under which closed procedures have greater power than other multiple comparison methods. But the amount by which power improves, and the range of situations in which power is better, are not generally known for specific methods. This paper compares the power of Shaffer's [14] Hommel's [15], and Hochberg's [1] methods to that of a "pair-wise" closed test [16] (described below) in the SCORE Study, and generalizes the results to a wider context. We shall furnish simulation evidence suggesting that, in the context of all pair-wise comparisons between 3 or 4 balanced arms, if the underlying distribution of the data is normal, binomial, or Poisson, and the method of analysis is based on either score or likelihood ratio statistics, "pair-wise" closed testing usually offers a modest increase in power over the other methods for controlling FWE, at a cost of somewhat greater computational effort.

The SCORE Study primary analysis is complicated by the need to consider covariates and interim monitoring. To simplify, we discuss multiple testing without considering these other factors. That is, we focus on controlling FWE at a single look while performing many pair-wise comparisons.

Five methods to control type I error

Over the years, a variety of methods have been proposed to accomplish FWE control. We discuss five methods in this paper: the very popular Bonferroni and Hochberg [1] techniques, methods due to Shaffer [14] and Hommel [15], and a closed test we define below.

Bonferroni

Reject any H_i for which $p_i \leq \alpha/n$. Equivalently, reject any H_i for which $p_i^* \leq \alpha$, where $p_i^* = \min[n\, p_i, 1]$. We refer to the $\{p_i\}$ as "unadjusted" p-values, and the $\{p_i^*\}$, which are derived from $\{p_i\}$ and meant to be compared to α, as "adjusted" p-values.

Hochberg

Arrange the p-values in descending order $p_{[1]} > p_{[2]} > \ldots > p_{[n]}$ with associated hypotheses $H_{[1]}, \ldots, H_{[n]}$. Then, accept or reject hypotheses iteratively for i = 1, .., n as follows:

a. If $p_{[i]} < \alpha/i$, reject $H_{[i]}, \ldots, H_{[n]}$ and stop.

b. Otherwise, accept $H_{[i]}$ and let $i \rightarrow i+1$

Equivalently: $p_{[1]}^* = p_{[1]}$, and $p_{[i]}^* = \min(p_{[i-1]}^*, i\, p_{[i]})$, for $i = 2, \ldots, n$.

Because only the smallest p-value is compared to α/n, while all the other p-values are compared to quantities greater than α/n, Hochberg's method has a greater chance than the Bonferroni method of rejecting $H_{[1]}, \ldots, H_{[n-1]}$, and therefore greater power.

Hommel

Compute $j = \max\{i \in \{1, \ldots, n\}: p_{(n-1+k)} > k\alpha/i \text{ for } k = 1, \ldots, i\}$. If the maximum does not exist, reject all H_i, $i = 1, \ldots, n$. Otherwise, reject all H_i with $p_i \leq \alpha/j$. Wright [17] provides an algorithm for calculating adjusted p-values. Hommel's method is slightly more powerful than Hochberg's method.

Shaffer

Some inequalities may constrain others. For example, if $a > b$ and $b > c$, we must have $a > c$. Shaffer [14] took advantage of this to improve Bonferroni p-values applied to (among other things) making all pairs of comparisons between g groups. In her method, the unadjusted p-values for all $n=g(g-1)/2$ pair-wise tests are arranged in ascending order $p_{[1]} \leq p_{[2]} \leq \ldots$, with associated hypotheses of pair-wise equality $H_{[1]}, H_{[2]}, \ldots$ Then, $H_{[1]}$ is rejected if $p_{[1]} \leq \alpha/m_{[1]}$. If $H_{[1]}$ is rejected (but not otherwise), $H_{[2]}$ is rejected if $p_{[2]} \leq \alpha/m_{[2]}$, etc. Testing ends with rejection of $H_{[1]}, \ldots, H_{[i-1]}$ at the smallest i such that $p_{[i]} > \alpha/m_{[i]}$. As in the ordinary Bonferroni test, $m_{[1]} = g(g-1)/2$. For subsequent tests, $m_{[i]}$ is equal to the largest number of pair-wise hypotheses that can be true simultaneously, given rejection of the previous hypotheses $H_{[1]}, \ldots, H_{[i-1]}$. For example, consider the problem of three groups: A, B, and C, with hypotheses of equality H(AB), H(AC), and H(BC). Initially all three hypotheses are tenable, so $m_{[1]} = 3*2/2 = 3$. Suppose H(AB) is the first hypothesis rejected. Now either H(AC) or H(BC) is still individually tenable, but not both simultaneously, so that $m_{[2]} = 1$. Having rejected, say, H(AB) and H(AC), only H(BC) remains tenable, so $m_{[3]} = 1$. While Shaffer constants for three groups are the same irrespective of the order in which hypotheses are rejected, this is not true for 4 or more groups, so that Shaffer constants may sometimes be difficult to determine. We refer to this method as Shaffer's S2 method [8,18]. Holland and Copenhaver [19] present a table of maximum Shaffer constants for groups of size 3–10. These are conservative, but do not depend on the order of rejection. We refer to Shaffer's method with Holland-Copenhaver constants as Shaffer's S1 method. In any case, once constants m_1, m_2, .., m_n … are obtained, adjusted p-values are calculated as follows:

$$p_{[1]}* = p_{[1]}m_{[1]}, \text{ and } p_{[i]}* = \min[1, \max(p_{[i-1]}*, p_{[i]}m_{[i]})], \text{ for } i=2, \ldots, n.$$

In this paper, we investigate both S1 and S2 methods. Algorithms for construction of the S2 constants have been discussed in [18] and [20]. We use a "brute force" approach that, while not optimized for speed, is short, easy to code, and acceptable for the moderate numbers of groups we contemplate here. A copy of the SAS "brute force" code that solves the average 10-group problem in about 0.5 seconds on a 2.79 GHz PC is provided in the appendix.

Closed

Given a set of hypotheses $H_1, \ldots, H_n$ (called the elementary hypotheses), consider all the hypotheses that can be generated by intersections between the elementary hypotheses (i.e. all the two-way, three-way, …, n-way intersections); this is the closed family of hypotheses. Using whatever test statistics are appropriate, calculate unadjusted p-values for each hypothesis in

the closed family. Then, the closed testing procedure rejects an elementary hypothesis if its unadjusted p-value, and the unadjusted p-values of all hypotheses implying it, are all $\leq \alpha$. Equivalently, define the adjusted p-value of an elementary hypothesis as the maximum of its unadjusted p-value, and the unadjusted p-values of all hypotheses implying it. Then the procedure rejects the elementary hypothesis if its adjusted p-value is $\leq \alpha$.

To illustrate, we apply closed testing to the problem of making all pair-wise comparisons of means. Suppose there are three elementary hypotheses H_{12}, H_{13}, and H_{23}, where H_{ij} stands for the hypothesis $\mu_i = \mu_j$. There is a single hypothesis implying all the elementary hypotheses: H_{123}: $\mu_1=\mu_2=\mu_3$. These 4 hypotheses constitute a closed family, and may be arranged in an implication graph as in Figure 1.

The arrows in Figure 1 represent implication. For example, if H_{123} is true, then H_{12} is true. To reject H_{12} in closed testing, it is necessary to reject both H_{12} and H_{123} individually at level α. Equivalently, the adjusted p-value for H_{12} is $p_{12}^* = \max [p_{12}, p_{123}]$. Similar statements apply to tests of the other elementary hypotheses.

The implication graph for the closed-family that tests all pair-wise differences between four means is more complex, as shown in Figure 2.

Thus, with four means, and six pair-wise mean differences, there are 14 hypotheses in the closed family. The hypotheses implying H_{12} are emboldened in Figure 2. For example, $H_{12,34}$ posits $\mu_1=\mu_2$ and $\mu_3=\mu_4$. To reject H_{12} using the closed testing approach, one must reject H_{12}, H_{123}, H_{124}, $H_{12,34}$, and H_{1234}, all at level α. Equivalently, the adjusted p-value for H_{12} is $p_{12}^* = \max[p_{12}, p_{123}, p_{124}, p_{12,34}, p_{1234}]$.

For a study with g arms, the number of p-values in the implication graph for all pair-wise comparisons of means is $B(g)-1$, where $B(g)$ is the g^{th} Bell number (A000110 from [21]). The number of p-values rises rapidly with the number of arms. Five arms require 51 p-values, and 8 arms require 4139, so closed testing will not be practicable for all pair-wise comparisons in a clinical trial with very many arms.

We have illustrated closed testing by applying it to the problem of making all pair-wise comparisons of means. Other problems will involve different implication graphs. For example, if one wishes only to compare two experimental treatments to a single control, the implication graph will be smaller than that shown in Figure 1.

Below, we supply p-values via ANOVA or similar tests. We use “pair-wise closed tests” to describe closed tests using these ANOVA-type p-values and implication graphs like those in Figures 1 and 2, which are appropriate for the “all pair-wise comparisons of means” problem. When we wish to refer to closed tests in general, we call them ”closed tests”, instead of “pair-wise closed tests”.

Strong control of FWE—Consider a set of hypotheses $\{H_1, \ldots, H_n\}$, some of which may be true and some false. We do not know which are true and which false. We wish to test each of the hypotheses in such a way that the probability that all the true hypotheses will be accepted is at least $1-\alpha$ (and so the probability that one or more of the true hypotheses will be rejected is at most α). A testing procedure that achieves this is said to have **strong** control of FWE. By contrast, if a procedure is only guaranteed to control the probability that all n hypotheses are accepted when they are all true, it is said to have weak control of FWE.

Closed methods guarantee strong FWE control even if each test in the implication graph assumes all its own elementary hypotheses are simultaneously true. Intuition for why this works may be gained by considering a situation with elementary hypotheses A–F. Suppose A, E and

F are true, and B, C and D are false. In closed testing, none of A, E, or F can be rejected unless hypothesis AEF (which asserts that A, E, and F are simultaneously true) is rejected. But since AEF is true, and it is being tested with an α-level test, the probability of rejecting it must be ≤ α. Thus, the probability of rejecting any of A, E, or F is controlled at ≤ α.

Various ingenious shortcuts have been developed to ease the computation load of closed testing. The Bonferroni procedure, although it pre-dates Marcus' paper, is a closed procedure. That is, a Bonferroni procedure with *n* p-values can be viewed as a closed procedure in which any intersection hypothesis $H_{ij\ldots k}$ is rejected if $\min(p_i, p_j, \ldots, p_k) \leq \alpha/n$. Hochberg's method is also a closed procedure, making use of p-values developed by Simes [5]. Unfortunately, Simes proved the appropriateness of his p-values only when the elementary hypotheses were independent. Thus, strictly speaking, Hochberg's method is not guaranteed to control FWE at level α. However, simulations suggest that the method is typically conservative, especially for large positive dependence structures [5]. Hommel's method is also closed, and also uses Simes p-values, so the same reservations apply here. Shaffer's procedures are "nearly" closed [8]. Bittman et al. [22] introduce a closed procedure for testing related outcomes that is consonant and has a maxmin property under the normal model.

Application of Closed Testing to the SCORE Study

The first indication that "pair-wise" closed testing might provide improvements over the Hochberg method that was described in the SCORE Study SAP [23] came from a simulation. For each iteration of this simulation, data were generated as 3 independent binomial samples, each with a sample size of 146. There were 10^5 iterations for each of the two disease areas. The three binomial outcome probabilities depended on the disease area as shown in Table 1.

In Table 1, Pr(Success) represents the probability that a given participant will experience a gain of 15 or more from baseline in best-corrected E-ETDRS visual acuity letter score at 12 months.

In this simulation, the unadjusted p-values come from ordinary Pearson's chi-squared tests of independence. For example, p_{S14} is the p-value of the chi-squared test of independence of the rows and columns of the 2-by-3 contingency table of which the rows are (Success, Failure) and of which the columns are (SC, 1 mg intravitreal triamcinolone, 4 mg intravitreal triamcinolone). The other three p-values, p_{S1}, p_{S4} and p_{14}, are defined by the chi-squared tests in the corresponding 2-by-2 contingency tables.

Table 2 compares the size and power of the "pair-wise" closed testing procedure with the corresponding measures of Hochberg's method in these simulations, using $\alpha = 0.05$. Note first that, when comparing 1 mg intravitreal triamcinolone vs. 4 mg intravitreal triamcinolone (which do not differ in these simulations), both methods have roughly the correct size, rejecting the null hypothesis that 1 mg = 4 mg about 5% of the time. However, in either of the two tests of SC versus intravitreal triamcinolone, the "pairwise" closed test procedure enjoys a power advantage of 4–5 percentage points over Hochberg's method. Moreover, when considering the probability that both of the tests of intravitreal triamcinolone versus SC are simultaneously significant, the "pair-wise" closed method has an 8 percentage-point power advantage.

On the basis of this and other simulations not shown, the SCORE Study Data and Safety Monitoring Committee approved a modification to the SCORE Study SAP, substituting "pair-wise" closed testing in place of Hochberg's method for analysis of the SCORE Study primary outcomes. Note these simulations were based only on the SCORE Study design hypotheses, and made no use of accumulated SCORE Study outcome data.

Application to a wider context

The above simulations beg the question of whether the apparent superiority of "pairwise" closed testing to Hochberg's method is confined to the SCORE Study, or applies more generally. Below, we provide simulation evidence suggesting that, in the context of typical tests for all pair-wise comparisons of means between 3 or 4 balanced arms, the "pair-wise" closed procedure usually offers a modest increase in power over Shaffer's S2 and Hommel's methods (and thus over Shaffer's S1 and Hochberg's methods) for controlling FWE.

The simulation comprises 1000 scenarios. Each scenario, consisting of 500,000 iterations, embodies one of the unique choices from the following Cartesian product:

- Distribution = (Normal, Binomial, Poisson). When the Normal distribution was used, it had a standard deviation of 0.5.
- Parameter μ = the mean of the distribution of the data. This ranged from 0.1 to x by 0.05, where x was 0.95 for the Normal and Binomial distributions, and x ranged from 1.00 to 1.50 for the Poisson distribution. The value of x was chosen to ensure good coverage of the power functions, which ranged in all cases from 0.05 to more than 0.90.
- Number of groups = (3, 4, 5, 6)
- Design is either balanced or unbalanced. In a balanced design, the sample size is 50 per group. In an unbalanced design, the sample size ranges linearly from 10–90 per group. That is, in an unbalanced design with k groups, the sample size of group i is $n_i = 10 + 80(i-1)/(k-1)$, for i = 1,…, k. Average group size is 50.
- Series is either Halves or Ones, as defined below.

The Series governs the means of the groups as specified in Table 3. In Table 3, "μ" is the parameter mentioned in the bulleted paragraph above, and "B" = 0.1. That is, in the "Halves" series, about half the group means are μ and half are 0.1, while in the "Ones" series, only one group mean is μ, while all the rest are 0.1. Note that, for 3 groups, "Halves" and "Ones" series are the same.

The unadjusted p-values were derived from tests that a typical practitioner might use given real data of the type simulated. For normal data, p-values were taken from standard F-tests of ANOVA contrasts. For binomial and Poisson data, p-values came from the chi-squared approximation to the distribution of score statistics. That is, in the binomial case, k groups were compared by means of Pearson's chi-squared test of independence in a 2-by-k table, while in the Poisson case, the comparison was by means of Pearson's chi-squared goodness of fit test with k categories, where expected values were given by the sample sizes of the groups. When p-values such as $p_{12,34}$ were required, they were calculated by separately comparing groups 1 to 2, and 3 to 4, and subsequently adding chi-squared values and degrees of freedom before the p-value calculation. All tests are two-tailed at $\alpha = 0.05$.

Rather then showing power for individual pairs of comparisons, we report global measures of power as follows:

- Proportional Power = Mean proportion of false hypotheses rejected
- Complete Power = Probability of rejecting all false hypotheses
- Minimal Power = Probability of rejecting any false hypothesis

Similarly, we distinguish between two global measures of type I error:

- Proportional FWE = Mean proportion of true hypotheses rejected

- Minimal FWE = Probability of rejecting any true hypothesis

Figures 3 and 4 display type I error and power for the "Normal Balanced Halves" set of simulation results as a function of the value of the μ parameter. In these figures, we suppress results for Shaffer's S1 and Hochberg's method, which are known to be less powerful than Shaffer's S2 and Hommel's method, respectively.

Figure 3 demonstrates that all four methods investigated (Bonferroni, "pair-wise" closed, Hommel, Shaffer S2) control FWE at ≤ 0.05, as desired. Across the entire simulation, simulated Minimal and Proportional FWE were both never greater than 0.0516 (not shown), confirming adequate control of type I error by all methods. Bonferroni size was always farther from 0.05 than the size of any other method. Note that the "pair-wise" closed method comes closest to the desired test size for Proportional type I error. The tendency of closed Minimal type I error to fall below Minimal error of other methods as the number of groups increases is repeated in almost all other simulations (not shown). Abrupt drops in Minimal FWE shown in Figure 3 occur because the number of true null hypotheses drops when μ increases past 0.1.

Figure 4 demonstrates that, in these data, the "pair-wise" closed method typically has better power than Hommel and Shaffer, which in turn have better power than the Bonferroni approach. Usually, the improvement from the Hommel to "pair-wise" closed method and Shaffer is roughly the same as the improvement from Bonferroni to Hommel. The advantage of "pair-wise" closed and Shaffer over Hommel is most pronounced in the case of Complete power, and appears to diminish as the number of arms increases.

In the interest of brevity, we refrain from presenting the other 22 graphs representing the remaining scenarios in the simulation, but instead summarize them in Tables 4–6.

Tables 4–6 have a common structure. Each cell represents a measure of power or size for one of the 48 combinations of Series, Distribution, Design, and Number of Groups. That is, each cell summarizes a quadruplet of curves giving either power or size, one curve for each of the "pair-wise" closed, Shaffer, Hommel, and Bonferroni methods. Each row thus summarizes a graph with panels, like the graphs presented in Figures 3 and 4.

Tables 4–6 present Percent Worst Absolute Error for the three different definitions of power. For any value of the parameter μ and any definition of power, let MAX_μ be the maximum over all simulated methods of the power at the parameter value μ. Then, for every method X (= Cl, S1, S2, Hm, Hb, Bo in Tables 4–6), Percent Worst Absolute Error of method X is defined as $100 * \max_{(\mu > 0.1)} [MAX_\mu - X_\mu]$. Thus, Worst Absolute Error of X is 0 only when method X was as good as or better than all other methods over the entire simulated parameter space. Cells with Percent Worst Error = 0 are represented by blanks. Numbers in Tables 4–6 are rounded to the nearest integer percent, so blank cells actually depict cases in which method X was close to best (i.e. Percent Worst Error less than 0.5%).

Tables 4–6 show that, for 3 or 4 balanced arms, the "pair-wise" closed method has better power than the other methods. Even in the unbalanced 3-arm case, the "pair-wise" closed method is often better, although, for Minimal Power, results are unclear. The advantage of the "pair-wise" closed method decreases as the number of arms and the imbalance between arms increase. Shaffer's S2 is better than S1 for Proportional and (especially) Complete Power in the "Halves" Series when the number of groups exceeds 3. There are scattered differences between Hommel's and Hochberg's methods, but they seldom exceed 1%.

In addition to Percent Worst Error, we also investigated, for each Power definition and method X: the area between MAX_μ and X_μ, for $\mu > 0.1$. Numerical differences between the outcomes,

being averages rather than maxima, are less dramatic, but results are similar to those presented in Table 4–6, and are not shown here.

For binomial and Poisson data, a second set of simulations, involving chi-squared approximations to the null distributions of likelihood ratio tests, was also performed. Results (not shown) are similar to those reported above for score statistics.

Five different applications of closed methods to test all pair-wise means between three Normal groups have been investigated (Thomas D. Cook, personal communication). Although closed testing using ANOVA offered good power, no single method investigated was most powerful over the entire parameter space. One of us (NO) verified that here too, if the practitioner uses ANOVA for the unadjusted p-values, closed testing is more powerful than Hochberg's method to test the elementary hypotheses.

Conclusion

Simulations suggest that, when using p-values arising from common tests to investigate all pair-wise comparisons of 3 or 4 means in a balanced design, "pair-wise" closed testing offers a modest power advantage over the Hommel, Hochberg, Shaffer, and Bonferroni methods, while strongly controlling FWE. The cost is a modest increase in complexity of calculation. The advantage of the "pair-wise" closed method decreases as the number of arms and the imbalance between arms increase. It seems likely that, for balanced designs involving all pair-wise comparisons of means, "pair-wise" closed testing has greater power with other statistical methods as well, although this should be verified by simulation or other approaches. Clinical trials that use Bonferroni or Hochberg approaches in this context should contemplate using the "pair-wise" closed approach. If only the elementary p-values are available, rather than the p-values from the entire closed set of comparisons, Shaffer's method is a good choice for this problem.

Acknowledgments

Research supported by the National Eye Institute (National Institutes of Health, Department of Health and Human Services) grants 5U10EY014351, 5U10EY014352, and 5U10EY014404. Support also provided in part by Allergan, Inc through donation of investigational drug and partial funding of site monitoring visits and secondary data analyses.

We gratefully acknowledge the careful, critical reading by, and many invaluable comments from, Thomas D. Cook, PhD.

References

1. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75 (4):800–2.
2. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association. 1967; 62:626–633.
3. Fisher RA. Combining independent tests of significance. American Statistician. 1948; 2(5):30.
4. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979; 6:65–70.
5. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986; 73 (3):751–4.
6. Hommel G. A comparison of two modified Bonferroni procedures. Biometrika. 1988; 75:383–386.
7. Rom DM. A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika. 1990; 77:663–665.
8. Westfall PH, Tobias RD. Multiple Testing of General Contrasts: Truncated Closure and the Extended Shaffer-Royen Method. Journal of the American Statistical Association. 2007; 102(478):487–494.

9. Westfall, PH.; Young, SS. Resampling-Based Multiple Testing. Wiley; N.Y: 1993.
10. Scheffe H. A method for judging all cointrasts in the analysis of variance. Biometrika. 1953; 40:87–104.
11. Tukey, JW. The problem of Multiple Comparisons. In: Braun, HI., editor. The Collected Words of John W. Tukey: Multiple Comparisons. Vol. VIII. Chapman & Gall; 1994.
12. Kramer CY. Extension of the multiple range test to group means with unequal numbers of replications. Biometrics. 1956; 12:307–310.
13. Grechanovsky E, Hochberg Y. Closed procedures are better and often admit a shortcut. Journal of Statistical Planning and Inference. 1999; 76:79–91.
14. Shaffer JP. Modified Sequentially Rejective Multiple Test Procedures. J Am Stat Assoc. September; 1986 81(395):826–831.
15. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988; 75(2):383–386.
16. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika. 1976; 63(3):655–60.
17. Wright SP. Adjusted P-values for simultaneous inference. Biometrics. December.1992 48:1005–1013.
18. Donoghue JR. Implementing Shaffer's multiple comparison procedure for a large number of groups. Recent Developments in Multiple Comparison Procedures, Institute of Mathematical Studies, Lecture Notes–Monograph Series. 2004; 47:1–23.
19. Holland BS, Copenhaver MDP. An Improved Sequentially Rejective Bonferroni Test Procedure. Biometrics. June.1987 43:417–423.
20. Westfall PH. Multiple Testing of General Constraints Using Logical Constraints and Correlations. Journal of the American Statistical Association, March. 1997; 92 (437):299–306.
21. Sloane, NJA. The On-Line Encyclopedia of Integer Sequences. 2007. published electronically at www.research.att.com/~njas/sequences/
22. Bittman RM, Romano JP, Vallarino C, Wolf M. Optimal testing of multiple hypotheses with common effect direction. Biometrika. 2009; 96(2):399–410.
23. The SCORE Study. Manual of Policies and Procedures. Version 4.0. Bethesda, MD: National Eye Institute; National Technical Information Service; 2008. Product Code #PB2008-106870

Appendix: A "Brute Force" Algorithm to determine Shaffer's S2 constants and p-values for all pairwise differences

```
/* Shaffer's (1986) multiple testing method as applied to the problem of
making all pairs of comparisons between g groups. The p-values for all
g(g–1)/2 pairwise tests are arranged in order P(1) <= P(2) <= …, with
associated hypotheses of pairwise equality H(1), H(2), … Then, H(1)
is rejected if P(1) <= alpha/m(1); given that H(1) is rejected, H(2)
is rejected if P(2) <= alpha/m(2), and so on. Testing ends with
rejection of H(1), …, H(i–1) at the smallest i such that P(i) >
alpha/m(i). As in the ordinary Bonferroni test, m(1) = g(g–1)/2. For
subsequent tests m(2), m(3), … m(i) is equal to the largest number of
pairwise hypotheses that can be true simultaneously, given rejection of
the previous hypotheses H(1), …, H(i–1). For example, consider the
problem of three groups: A, B, and C, with hypotheses of equality
H(AB), H(AC), and H(BC). Initially all three hypotheses are tenable, so
m(1) = 3*2/2 = 3. Suppose H(AB) is the first hypothesis rejected. Now
either H(AC) or H(BC) is still individually tenable, but not both
simultaneously, so that m(2) = 1. Having rejected, say, H(AB) and H(AC),
only H(BC) remains tenable, so m(3) = 1. While Shaffer constants for
```

```
three groups are the same irrespective of the order in which hypotheses
are rejected, this is not true for 4 or more groups.
Shaffer JP. Modified sequentially rejective multiple test procedures.
J. Am. Stat. Assoc 81(395):826–831 (1986)
Input requires g(g–1)/2 records, each containing a raw p-value RAW_P, and
the indexes (iii, jjj) of the pair involved. */
/* Input some example data */
%let g = 6;
%let npairs = %eval(&g*(&g–1)/2);
data pairs;
  input raw_p iii jjj;
  datalines;
0.5311 1 2
0.0194 1 3
0.0053 1 4
0.0000 1 5
0.0000 1 6
0.0738 2 3
0.0229 2 4
0.0002 2 5
0.0000 2 6
0.5794 3 4
0.0229 3 5
0.0001 3 6
0.0738 4 5
0.0004 4 6
0.0354 5 6
;
/* Sort p-values in ascending order */
proc sort data = pairs;
  by raw_p;
  run;
data shaffer;
  keep raw_p mm adj_p iii jjj;
  array ii{&g} i1-i&g;
  array appears{&g};
  array m{&npairs};
  array rawp{&npairs};
  array pair{2, &npairs};
  retain pair;
  retain npairs 0;
  retain rawp;
  array ma{&g};
  /* read pairs in order of ascending p-value */
  set pairs end = eof;
  npairs + 1;
  pair{1,npairs} = iii;
  pair{2,npairs} = jjj;
  rawp{npairs} = raw_p;
  if eof then do;
    /* Generate in lexical order in ii(1), …, ii(g) all equivalence relations
```

```
    on the set (1, 2, …, g). This is equivalent to partitioning the n-set
    by generating all possible arrangements of nonempty subsets. For
    example, the set partitions when g=3 are: 111, 112, 121, 122, and 123.
    Note that the number of such arrangements is given by the Bell numbers. */
    /* Initialize m */
    do j = 1 to &npairs;
     m{j} = 0;
     end;
    /* initialize ii */
    do j = 1 to &g;
     ii{j} = 1;
     end;
    do until(done);
     done = 1;
     /* For each of the sets of integers, calculate
     the number of times each integer appears. */
     do j = 1 to &g;
       appears{j} = 0;
       end;
     do j = 1 to &g;
       appears{ii{j}} = appears{ii{j}} + 1;
       end;
     /* Calculate in NUM the total number of pairwise equalities this
     set of integers represents. */
     num = 0;
     do j = 1 to &g;
       num = num + appears{j} * (appears{j}-;1);
       end;
     num = num/2;
     /* Update the m-values. When done, m(1) will be g(g-1)/2. For
     subsequent tests m(2), m(3), … m(i) will be equal to the largest
     number of pairwise hypotheses that can be true simultaneously, given
     rejection of the previous hypotheses H(1), …, H(i-1). To obtain the
     m-values, for each equivalence relation k, we update each m(j) with NUM
(k)
     if H(1), …, H(j-1) are all inconsistent with equivalence relation k. */
     m1 = max(m1, num);
     do j = 1 to &npairs-1;
       if ii{pair{1, j}} = ii{pair{2, j}} then leave;
       m{j+1} = max(m{j+1}, num);
       end;
     /* Generate the next set of integers.
     First, store the cumulative maxima. */
     j = 0;
     do ptr = 1 to &g;
       j = max(j, ii{ptr});
       ma{ptr} = j;
       end;
     /* Now generate the integers. */
     do ptr = &g to 2 by -1;
       if ii{ptr} < ma{ptr-1} + 1 then do;
```

```
            ii{ptr} = ii{ptr} + 1;
            do i = ptr+1 to &g;
              ii{i} = 1;
              end;
            done = 0;
            leave;
            end;
          end;
        end; /* Do until done. */
      /* Output adjusted m- and p-values. */
      adj_p = 0;
      do j = 1 to &npairs;
        raw_p = rawp{j};
        iii = pair{1, j};
        jjj = pair{2, j};
        adj_p = min(1, max(adj_p, m{j} * raw_p));
        mm = m{j};
        output;
        end;
      end;
    run;
proc print data = shaffer; run;
```

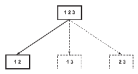

Figure 1.
Implication graph for “pair-wise” closed testing of all pair-wise differences between three means. Solid lines connect unadjusted p-values whose maximum gives the adjusted p-value for elementary hypothesis 1 2.

Figure 2.
Implication graph for “pair-wise” closed testing of all pair-wise differences between four means. Solid lines connect unadjusted p-values whose maximum gives the adjusted p-value for elementary hypothesis 1 2.

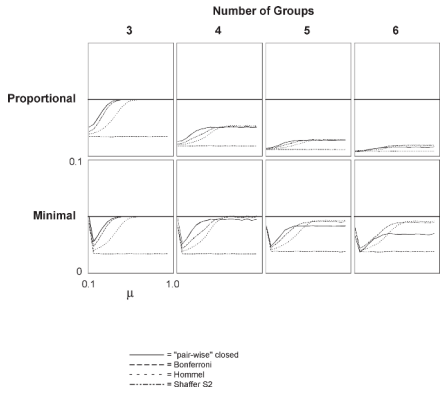

Figure 3.
FWE for Normal, Balanced, Halves

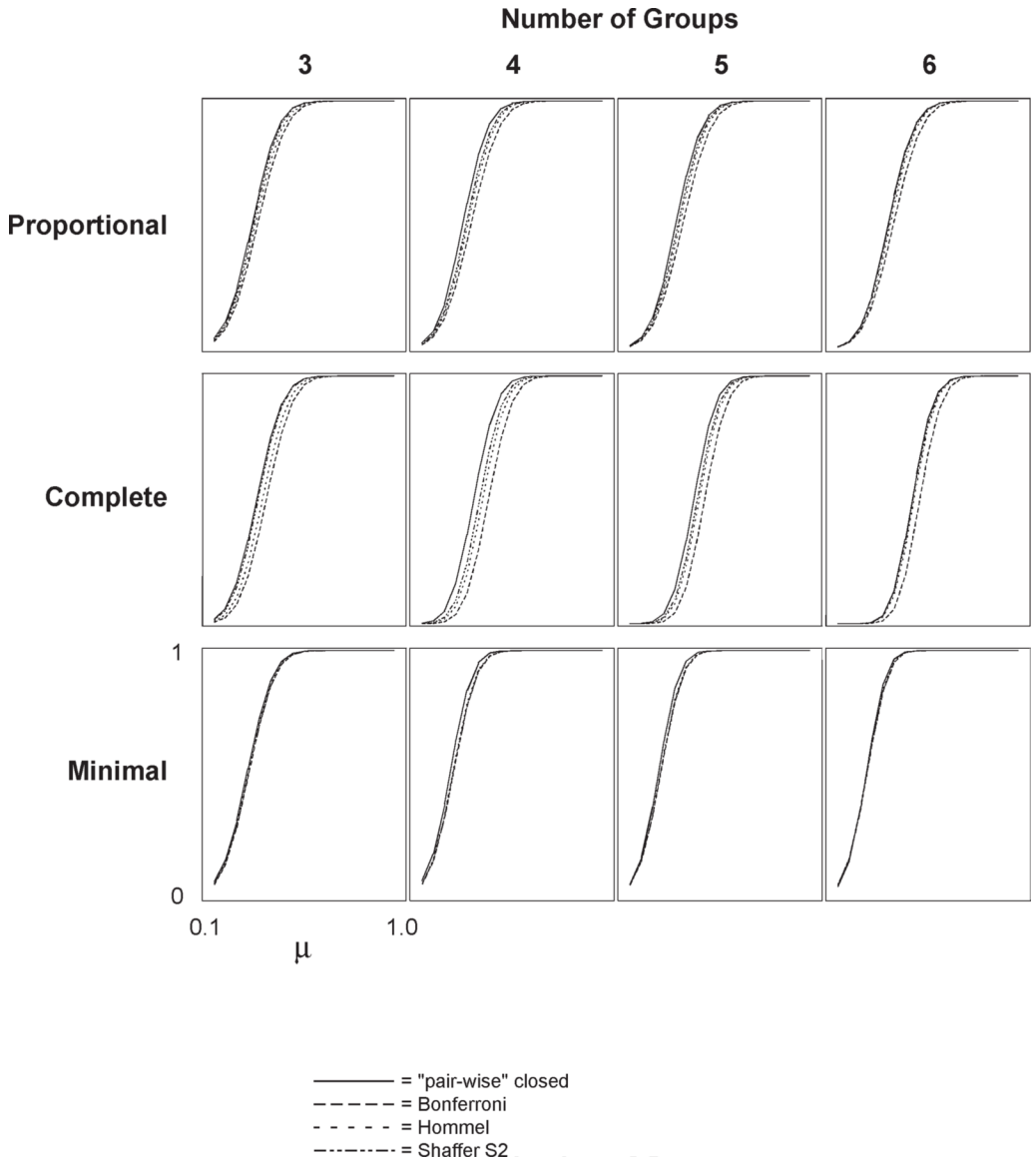

Figure 4.
Power for Normal, Balanced, Halves

Table 1

Probability parameters for the binomial samples of size 146 used in the simulations.

	Pr (Success)		
	SC	**1 mg**	**4 mg**
CRVO	0.15	0.30	0.30
BRVO	0.35	0.53	0.53

BRVO = Branch Retinal Vein Occlusion Trial; CRVO = Central Retinal Vein Occlusion Trial

Table 2

Operating characteristics of the "pair-wise" closed testing procedure and Hochberg's method applied to data simulating SCORE outcomes, when $\alpha = 0.05$.

	CRVO		BRVO	
Hypotheses rejected	**Hochberg**	**Closed**	**Hochberg**	**Closed**
SC = 1 mg	0.80	0.84	0.80	0.85
SC = 4 mg	0.79	0.84	0.80	0.85
1 mg = 4 mg	0.04	0.05	0.04	0.05
(SC = 1 mg) and (SC = 4 mg)	0.70	0.78	0.71	0.79
(SC = 1 mg) or (SC = 4 mg)	0.89	0.90	0.89	0.91

BRVO = Branch Retinal Vein Occlusion Trial; CRVO = Central Retinal Vein Occlusion Trial; SC = Standard Care

Table 3

Means of groups in the two series.

	Series	
Nbr groups	**Halves**	**Ones**
3	μBB	μBB
4	μμBB	μBBB
5	μμBBB	μBBBB
6	μμμBBB	μBBBBB

μ is the simulation parameter defined in the text, and B = 0.1

Table 4

Percent Worst Error in Proportional Power. Blank cells indicate that the method is close to best across the entire simulated parameter space.

			Number of groups																							
			3						4						5						6					
			Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo
Unb	Halves	Bin				3	4	6		5	6	7	8	11		2	3	5	5	8	1		2	3	4	7
		Nor				2	2	6		5	6	7	8	12		3	4	5	6	10	1	1	3	3	4	8
		Poi				4	4	6		4	5	6	6	10		2	3	4	5	7	1		2	3	3	7
	Ones	Bin				3	4	6	1	1	1	2	3	6	1			2	2	4	4			1	2	5
		Nor				2	2	6				2	3	6	2			2	3	5	4			1	2	4
		Poi				4	4	6				2	2	4	1			1	1	1	3			1	1	2
Bal	Halves	Bin		4	4	8	8	11		7	8	10	10	17		5	7	8	8	14		2	4	4	5	12
		Nor		3	3	6	7	11		7	8	10	11	15		4	6	7	8	13		1	4	5	6	10
		Poi		7	7	11	11	15		7	8	10	11	16		4	6	8	9	13	1		3	4	5	10
	Ones	Bin		4	4	8	8	11		4	4	7	7	12		4	4	6	6	9		1	1	2	2	6
		Nor		3	3	7	7	11		3	3	6	6	9		1	1	3	4	6	2			2	2	4
		Poi		7	7	11	11	15		5	5	9	9	12		3	3	6	6	9				3	3	5

Bal = Balanced; Unb = Unbalanced; Cl = “pair-wise” closed method; S2 = Shaffer’s S2 method; S1 = Shaffer’s S1 method; Hm = Hommel’s method; Hb = Hochberg’s method; Bo = Bonferroni method

Table 5

Percent Worst Error in Complete Power. Blank cells indicate that the method is close to best across the entire simulated parameter space

			Number of groups																							
			3						4						5						6					
			Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo
Unb	Halves	Bin		1	1	7	7	11		9	15	15	15	23		6	7	7	7	16		1	6	6	6	15
		Nor	1			5	5	12		6	12	12	12	22		5	6	9	9	19		1	4	4	4	15
		Poi		1	1	8	8	11		8	12	12	12	21		5	6	9	9	15		2	4	4	4	12
	Ones	Bin		1	1	7	7	11		3	3	4	5	12		1	1	3	3	9	1			1	1	6
		Nor	1			5	5	12		1	1	5	5	11		1	1	3	3	8	1			1	1	6
		Poi		1	1	8	8	11		1	1	4	4	9		3	3	4	5	6				1	1	2
Bal	Halves	Bin		3	3	13	13	18		13	18	18	18	33		8	10	12	12	23		3	5	5	5	21
		Nor		3	3	11	11	18		13	19	19	19	31		8	9	12	12	24		2	5	5	5	18
		Poi		5	5	15	15	21		14	20	19	19	32		7	9	13	13	26		1	4	4	4	20
	Ones	Bin		3	3	13	13	18		2	2	7	7	15		2	3	5	5	9	3			3	3	9
		Nor		3	3	11	11	18		2	3	7	7	14		1	2	4	4	10	2			1	2	6
		Poi		5	5	15	15	21		3	3	7	8	15		2	2	4	4	11	1			2	2	7

Bal = Balanced; Unb = Unbalanced; Cl = “pair-wise” closed method; S2 = Shaffer’s S2 method; S1 = Shaffer’s S1 method; Hm = Hommel’s method; Hb = Hochberg’s method; Bo = Bonferroni method

Table 6

Percent Worst Error in Minimal Power. Blank cells indicate that the method is close to best across the entire simulated parameter space.

			Number of groups																							
			3						4						5						6					
			Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo	Cl	S2	S1	Hm	Hb	Bo
Unb	Halves	Bin	1	1	1		1	1		5	5	4	5	5	1	1	1	1	1	1	3	1	1		1	1
		Nor	1	2	2		1	2		6	6	5	6	6		3	3	2	3	3	1	2	2	1	2	2
		Poi	1	1	1		1	1		5	5	4	4	5	1	2	2	1	2	2	3	1	1		1	1
	Ones	Bin	1	1	1		1	1	4	1	1		1	1	4						8	1	1		1	1
		Nor	1	2	2		1	2	2	1	1		1	2	4	1	1		1	1	6	1	1		1	1
		Poi	1	1	1		1	1	2	1	1			1	5						6					
Bal	Halves	Bin		5	5	4	4	5		8	8	6	7	8		6	6	4	5	6	2	3	3		2	3
		Nor		3	3	2	3	3		8	8	6	7	8		6	6	5	5	6	1	2	2	2	2	2
		Poi		9	9	8	8	9		6	6	5	6	6	1	1	1	1	1	1	6					
	Ones	Bin		5	5	5	5	5		7	7	6	6	7		4	4	3	3	4	2	2	2		1	2
		Nor		3	3	2	3	3		3	3	2	3	3		1	1		1	1	4					1
		Poi		9	9	8	8	9		6	6	6	6	6		1	1	1	1	1	3					

Bal = Balanced; Unb = Unbalanced; Cl = “pair-wise” closed method; S2 = Shaffer’s S2 method; S1 = Shaffer’s S1 method Hm = Hommel’s method; Hb = Hochberg’s method; Bo = Bonferroni method