Multiplicity-Adjusted Sample Size Requirements: A Strategy to Maintain Statistical Power With Bonferroni Adjustments

Andrew C. Leon, Ph.D.

Background: A researcher must carefully balance the risk of 2 undesirable outcomes when designing a clinical trial: false-positive results (type I error) and false-negative results (type II error). In planning the study, careful attention is routinely paid to statistical power (i.e., the complement of type II error) and corresponding sample size requirements. However, Bonferroni-type alpha adjustments to protect against type I error for multiple tests are often resisted. Here, a simple strategy is described that adjusts alpha for multiple primary efficacy measures, yet maintains statistical power for each test.

Method: To illustrate the approach, multiplicity-adjusted sample size requirements were estimated for effects of various magnitude with statistical power analyses for 2-tailed comparisons of 2 groups using χ^2 tests and t tests. These analyses estimated the required sample size for hypothetical clinical trial protocols in which the prespecified number of primary efficacy measures ranged from 1 to 5. Corresponding Bonferroni-adjusted alpha levels were used for these calculations.

Results: Relative to that required for 1 test, the sample size increased by about 20% for 2 dependent variables and 30% for 3 dependent variables.

Conclusion: The strategy described adjusts alpha for multiple primary efficacy measures and, in turn, modifies the sample size to maintain statistical power. Although the strategy is not novel, it is typically overlooked in psychopharmacology trials. The number of primary efficacy measures must be prespecified and carefully limited when a clinical trial protocol is prepared. If multiple tests are designated in the protocol, the alpha-level adjustment should be anticipated and incorporated in sample size calculations.

(J Clin Psychiatry 2004;65:1511–1514)

Received Nov. 11, 2003; accepted April 28, 2004. From the Department of Psychiatry, Weill Medical College of Cornell University, New York, N.Y.

This research was supported, in part, by grants MH49762 and MH60447 from the National Institute of Mental Health, Bethesda, Md.

Corresponding author and reprints: Andrew C. Leon, Ph.D., Weill Medical College of Cornell University, Department of Psychiatry, Box 140, 525 East 68th St., New York, NY 10021 (e-mail: acleon@med.cornell.edu).

That which cannot be avoided must be actively sought.

—Niccolo Machiavelli, as paraphrased in *The Abyssinian* by Jean-Christophe Rufin

When designing a protocol for a randomized controlled clinical trial (RCT), there is a tension between guarding against the possibility that analyses will result in either of 2 types of erroneous conclusions. Type I error, which occurs when the null hypothesis is true, represents the probability of falsely concluding that an ineffective agent is efficacious. On the other hand, if the null hypothesis is false, type II error occurs when the investigator fails to conclude that an effective psychotropic agent is efficacious. This results in a failed clinical trial.

Of course, use of multiple statistical tests, or multiplicity, elevates the risk of type I error. In psychopharmacology trials, there are 3 likely reasons for multiplicity. First, numerous efficacy measures are often specified in a study protocol to examine the effect of a novel psychotropic agent on various aspects of a psychiatric disorder, such as symptom severity, functional impairment, adverse events, and global functioning. Second, if assessments are made at weekly or biweekly intervals in a clinical trial, there is a temptation to examine efficacy separately at each of those points in time. Third, in an effort to accommodate the problem of dropout, analysis plans customarily involve multiple subgroups such as the intent-to-treat sample, the per protocol sample, and the safety sample. The latter 2 sources of multiplicity can be avoided with a data analysis plan that clearly indicates the one statistical test that corresponds to the primary hypothesis and, in an effort to reduce selection bias, adheres to the principle of intention-to-treat.[1]

Here, however, we consider multiplicity that results from multiple efficacy measures. Multiple efficacy measures elevate the risk of type I error in psychopharmacology trials and other studies alike. Specifically, the experimentwise (EW) probability of type I error for k statistical tests can be estimated: $\alpha_{EW} = 1 - (1 - \alpha)^k$. For example, type I error increases from 5% with 1 statistical test to 9.8% with 2 tests, to 14.3% with 3 tests, and so on. Journal editors[2] and regulatory bodies[3] often require adjustments for multiplicity, and peer reviewers for funding

agencies tend to be quite sensitive to the problem of multiplicity. Some investigators comply by using the so-called Bonferroni adjustment, which partitions the nominal alpha level among the numerous tests. For example, if there are 5 primary efficacy measures and α_{EW} of .05 is sought, the Bonferroni-adjusted alpha level would be .05/5 = .01 for the test of each of the 5 dependent variables. (By primary efficacy measures, I refer to those on which the results of the RCT are based; that is, if significant group differences are detected on any one of these measures, in the hypothesized direction, the agent is deemed "efficacious.")

The Bonferroni adjustment is appealing because the calculations are simple and the method can be applied to numerous statistical procedures (e.g., tests of continuous, categorical, or survival data) and even combinations of those procedures. Most importantly, it tightly controls type I error, such that the experimentwise probability of type I error is maintained at .05 by using adjusted alpha levels; this can be shown using the algorithm for experimentwise alpha from above: $\alpha_{EW} = 1 - (1 - \alpha)^k$. For example, for 2 statistical tests, $\alpha_{EW} = 1 - (1 - .025)^2 = .05$. Likewise, for 3 statistical tests, $\alpha_{EW} = 1 - (1 - .0167)^3 = .05$. (Technically, the Dunn-Šidák alpha adjustment,[4] $\alpha_{D\text{-}S} = 1 - (1 - a)^{1/k}$, will yield experimentwise alpha levels of precisely .05, whereas with the Bonferroni adjustment, the experimentwise alpha will be slightly lower than .05 if taken to the third decimal. For all practical purposes, the results of applying either approach will be nearly identical.) Although these calculations assume independence among tests, Pocock et al.[5] showed that the estimates do not change substantially until correlations between outcomes exceed 0.50.

The tension between type I and type II errors is exacerbated with adjustments for multiplicity; this is because one cost associated with an alpha adjustment is a corresponding reduction in statistical power to detect hypothesized effects on any one of the efficacy measures. For a given sample size, the statistical power to detect a particular effect size decreases with a reduction in the alpha level. Therefore, a standard rationale for avoiding alpha adjustments is the corresponding cost in statistical power. For example, a researcher might inquire, "This medication would be a valuable clinical tool whether it reduces symptom severity or functional impairment. Why must we be penalized for addressing 2 critical clinical questions in one trial?" The answer is, "One hopes to reduce the likelihood of patients eventually being treated with ineffective agents." Tension clearly exists between type I and type II error in designing a study.

The objective of this article is to illustrate a simple strategy for maintaining statistical power in situations when Bonferroni-type alpha adjustments are to be implemented. The article shows that statistical power can be maintained with the addition of efficacy measures, but an a priori sample size adjustment is required to compensate for the adjusted alpha level. Although this approach is not innovative, it is clearly underutilized in psychopharmacology trials. While this strategy is illustrated in the context of RCTs, it can readily be applied more broadly to observational studies.

METHOD

Multiplicity-adjusted sample size requirements were examined using statistical power analyses. These analyses estimated the required sample size for hypothetical clinical trial protocols in which the prespecified number of primary dependent variables ranged from 1 to 5. Corresponding Bonferroni-adjusted alpha levels were used for these calculations. All analyses involved 2-tailed comparisons of 2 equal-sized groups. The resulting sample size requirements for statistical power of 0.80 were estimated for a wide range of hypothesized treatment effects separately for χ^2 tests and t tests. The sample sizes were estimated using algorithms presented by Fleiss[6] for χ^2 tests with the continuity correction and the Power and Precision software[7] for t tests. Each estimated sample size was rounded up to assure that statistical power was not less than 0.80.

RESULTS

Initially, consider an RCT that is being planned in which response rates of 2 treatment groups will be compared using a 2-tailed χ^2 test. Assume that multiple dependent variables are specified. One goal in designing the protocol is to maintain a statistical power of 0.80 for any one dependent variable, while at the same time adjusting alpha for multiplicity. Sample size requirements are presented in Table 1 for various combinations of response rates that might be seen in psychopharmacology trials, separately for studies with 1, 2, 3, 4, or 5 primary efficacy measures, assuming the experimentwise alpha level was .05. These estimates incorporate Bonferroni-adjusted alpha levels of .05, .025, .0167, .0125, and .01, respectively. The corresponding sample size increases, relative to the sample size required for 1 dependent variable, are 18%, 29%, 36%, and 41%, respectively, for 2 through 5 dependent variables. (Note that these increases represent medians across the range of group differences presented in Table 1.)

Alternatively, the study could be designed to compare 2 groups on continuous measures (e.g., the Hamilton Rating Scale for Depression) using 2-tailed t tests. The sample size requirements are presented in Table 2 for effect sizes of 0.20 to 0.80, separately for studies with 1, 2, 3, 4, or 5 primary efficacy measures, again assuming the experimentwise alpha level was .05. For continuous efficacy measures, the median increases in sample size requirements, across the effect sizes shown, are 21%,

Table 1. Multiplicity-Adjusted Sample Size Requirements Per Group in a Study That Compares 2 Groups for Various Response Rates and Number of Dependent Variables[a]

Response Rate		No. of Dependent Variables				
Group 1	Group 2	1	2	3	4	5
0.1	0.2	219	261	286	303	317
	0.3	72	85	93	98	102
	0.4	38	45	49	52	54
	0.5	25	29	31	33	34
	0.6	17	20	22	23	24
	0.7	13	15	16	17	18
	0.8	10	11	12	13	13
	0.9	8	9	9	10	10
0.2	0.3	313	375	411	437	457
	0.4	91	109	119	126	131
	0.5	45	54	59	62	65
	0.6	28	32	35	37	39
	0.7	19	22	24	25	26
	0.8	13	15	17	18	18
	0.9	10	11	12	13	13
0.3	0.4	376	451	495	526	550
	0.5	103	123	135	143	149
	0.6	49	58	63	67	70
	0.7	29	34	37	39	40
	0.8	19	22	24	25	26
	0.9	13	15	16	17	18
0.4	0.5	408	489	537	571	597
	0.6	107	128	140	148	155
	0.7	49	58	63	67	70
	0.8	28	32	35	37	39
	0.9	17	20	22	23	24
0.5	0.6	408	489	537	571	597
	0.7	103	123	135	143	149
	0.8	45	54	59	62	65
	0.9	25	29	31	33	34

[a]Each estimate assumes that a χ^2 test will be used with a continuity correction, experimentwise alpha of .05, and statistical power of 0.80. The algorithm used for sample size estimates is from Fleiss.[6]

Table 2. Multiplicity-Adjusted Sample Size Requirements Per Group in a Study That Compares 2 Groups for Various Effect Sizes and Number of Dependent Variables[a]

	No. of Dependent Variables				
Effect Size	1	2	3	4	5
0.2	394	477	525	560	586
0.3	176	213	234	250	262
0.4	100	121	133	141	148
0.5	64	78	86	91	96
0.6	45	55	60	64	67
0.7	34	41	45	48	50
0.8	26	31	35	37	39

[a]Each estimate assumes that a t test will be used with a 2-tailed experimentwise alpha of .05 and statistical power of 0.80. The estimates were made using Power and Precision software.[7]

32%, 43%, and 49% for 2 through 5 dependent variables, respectively.

DISCUSSION

Alpha adjustments for multiple primary efficacy measures are commonly shunned because of the sacrifice in statistical power. A proactive alternative strategy has been described. An RCT protocol that prespecifies multiple efficacy measures should anticipate the necessary alpha adjustments and increase the proposed sample size accordingly. The resulting sample size will provide the desired levels of power and, thus, reduce the probability of a failed clinical trial.

The examples shown illustrate the proportionate increase in subjects needed for each additional outcome measure. For instance, compared with a trial with 1 primary outcome, about 20% more subjects are needed in a trial with 2 primary dependent variables and about 30% more are needed in a trial with 3 primary dependent variables. The use of multiplicity-adjusted sample size to preserve statistical power comes at a cost. The increased sample size requirements have corresponding increases in research costs, study duration, and risk to the additional human subjects. Thus, on the basis of concerns for both type I and type II error, multiple efficacy measures should not be used unless each measure is indispensable.

On the surface, it might seem that multiplicity-adjusted sample size requirements would be somewhat reduced for alternative alpha adjustment strategies such as sequentially rejective tests.[8,9] However, these tests must be powered for the most restrictive alpha level. For instance, in Hochberg's step-up approach,[9] the alpha level for each successive test is smaller, where the adjusted alpha is a product of the experimentwise alpha (e.g., .05) and the reciprocal of the test number *k*, where, for the examples shown, *k* ranges from 1 to 5 for 5 successive tests.

This presentation has assumed that the protocol identified multiple primary dependent variables and that the investigational intervention would be deemed efficacious if the groups differed significantly on any one of those variables. In contrast, no alpha adjustment is needed if the RCT protocol requires significance on all primary efficacy measures.[3] Furthermore, the examples of multiplicity-adjusted sample sizes presented here assumed that all hypotheses involved either continuous or dichotomous outcomes, but not both. Nevertheless, multiplicity-adjusted sample size requirements for a set of hypotheses involving both continuous and dichotomous outcomes could be estimated and would be based on both the adjusted alpha and the hypothesis with the smaller effect.

In conclusion, it is essential that the number of primary efficacy measures, the corresponding alpha level, and the power adjustment strategies all be clearly specified in a clinical trial protocol. If more than one primary outcome is designated, the alpha level should be adjusted for multiplicity and sample size calculations based on that adjusted alpha level accordingly. In this way, a balance can be achieved between reducing the probability of falsely concluding that an ineffective agent is efficacious and, if the null hypothesis is false, failing to conclude that an effective psychotropic agent is efficacious.

REFERENCES

1. Gillings D, Koch G. The application of the principle of intention-to-treat to the analysis of clinical trials. Drug Inf J 1991;25:411–424
2. Altman DG, Schulz KF, Moher D, et al, for the CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–694
3. US Dept Health Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Guidance for Industry: E9 Statistical Principles for Clinical Trials. Rockville, Md: US Dept Health Human Services; 1998
4. Ury HK. A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes. Technometrics 1976;18:89–97
5. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics 1987;43:487–498
6. Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. New York, NY: John Wiley and Sons; 1981
7. Borenstein M, Rothstein H, Cohen J. Power and Precision. Teaneck, NJ: Biostat; 2000
8. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70
9. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–803