CHEST Commentary

Correction for Multiple Testing

Is There a Resolution?

David L. Streiner, PhD, CPsych; and Geoffrey R. Norman, PhD

In most studies, many statistical tests are performed. They can be run to compare the groups at baseline, look at relationships among the various measures, and, for intervention trials, examine more than one end point. As the number of tests increases, so does the probability of finding at least one of them to be statistically significant just by chance (the problem of multiplicity). A number of procedures have been developed to deal with multiplicity, such as the Bonferroni correction, but there is continuing controversy regarding if and when these procedures should be used. In this article, we offer recommendations about when they should and should not be brought into play.

CHEST 2011; 140(1):16–18

Abbreviations: MHE = malignant hypertrophy of the ego

In 2010, Streiner and Norman[1] reported the discovery of a new disorder, called malignant hypertrophy of the ego (MHE), which is highly prevalent in some groups, such as politicians, actors, academic deans, and neurosurgeons. It may lead to reduced life expectancy for those afflicted with it, with the primary cause of death being murder at the hands of those forced to interact with these people. Imagine a study of the correlates of MHE, in which those with the disorder are compared with a group of control subjects. To begin with, we look for differences between the two groups on a large number of baseline variables, such as age, marital status, education, and so forth—about 20 variables in all. We then administer a series of questionnaires to the two cohorts, looking at such factors as mood, ego strength, personality type, and so on—again, about 20 different variables. We predict a priori that the MHE group will be higher in ego strength and lower in depression than the comparison group, but we do not know what else to expect, especially with regard to how these variables correlate.

The question we have to address is whether we must take into account the fact that we are conducting many *t* tests and correlations or whether we can accept findings of $P < .05$ at face value. To answer this, we first have to understand the nature of the problem, which is called "multiplicity," or performing many tests of significance within one study. We are all familiar with the expression $P < .05$—the probability, given that there really is no difference between the groups, of finding a difference just because of chance factors. Let us assume for the moment that there is absolutely no difference between those who have MHE and those who do not. Will we find any differences among the 20 baseline variables, and if so, how many? The answer is yes, we will find differences just by chance, and we can answer "how many" in two different ways. One way is to say that if 5% of the tests will be significant by chance, then we would expect to find one (ie, 5% of 20) to be statistically significant. Another way to answer the question is to determine the probability that at least one test will be significant. That probability is computed as (1 − the probability that we will not see any significant tests), which is given by the formula, $1 - (1 - \alpha)^k$, where α is the level of significance that we are adopting

Manuscript received February 28, 2011; revision accepted March 3, 2011.

Affiliations: From the Department of Psychiatry and Behavioural Neurosciences (Dr Streiner) and the Department of Clinical Epidemiology and Biostatistics (Drs Streiner and Norman), McMaster University, Hamilton; and the Department of Psychiatry (Dr Streiner), University of Toronto, Toronto, ON, Canada.
Correspondence to: David L. Streiner, PhD, CPsych, Department of Psychiatry and Behavioural Neurosciences, McMaster University, Centre for Mountain Health Services, 100 W 5th St, Hamilton, ON, Canada, L8N 3K7; e-mail: streiner@mcmaster.ca

DOI: 10.1378/chest.11-0523

(usually 0.05), and k is the number of tests. So, with $\alpha = 0.05$ and $k = 20$, the probability that at least one finding will be significant is $> 64\%$. (For those interested in the derivation of the formula, see Reference 2.)

Over the years, a number of solutions have been proposed to deal with this issue of multiplicity and inflation of the α level (that is, finding significance by chance). One of the simplest is the Bonferroni inequality, or Bonferroni correction. If 20 analyses are run, then one uses $\alpha/20$, or 0.0025, as the level of significance, rather than 0.05. However, it could more properly be called the Bonferroni *over*correction, because it results in far too few actually significant results; that is, there are too many type 2 errors (not declaring a result statistically significant when in fact it is). Other, less-conservative solutions have been proposed, such as the Bonferroni-Holm[3] or Hochberg[4] techniques.

But the use of any of these corrections begs the question of whether they should be used at all. To say that statisticians disagree on this would be indulging in understatement. A leading epidemiologist, Kenneth Rothman,[5] argues quite vociferously that corrections should never be used. He believes that their use implies the default assumption that chance is the "first-order explanation for observed phenomena," whereas the underlying premise of research is that "nature follows regular laws that may be studied through observation."[5] It is better to tolerate findings that may later prove to be false than to prematurely discard potentially useful observations because of type 2 errors caused by corrections for multiplicity. This is fine as long as we view each study as an opportunity for generating hypotheses for the next study, which is not really an accurate description of the average clinical trial. Perhaps for that reason, others, such as Ottenbacher[6] and Moyé,[7] argue just as vociferously that such corrections are mandatory because "Type I error accumulates with each executed hypothesis test and must be controlled by the investigators."[7]

Although a far wiser statistician than us wrote "I…observe that it is hard to see views such as [the ones cited previously] being reconciled,"[8] we shall wade in where angels fear to tread and attempt such a reconciliation. First, though, a word of wisdom: The best way to mitigate the problem of multiplicity, and especially of testing multiple end points in a clinical trial, is to avoid the problem entirely. Schulz and Grimes[9] argue quite cogently that "researchers should restrict the number of primary end points tested," and confine significance testing to only those hypotheses specified beforehand in the protocol. Post hoc, exploratory analyses can still be done, but they should be clearly labeled as such and the findings treated as hypothesis generating, rather than as confirmatory. However, we are left with the problems of multiple testing in a number of situations: (1) comparability of groups at baseline, (2) assessing correlates of some construct (eg, a condition or diagnosis), and (3) when there may be more than one outcome. Let us discuss these individually.

Testing for baseline differences can occur in two different types of studies: randomized controlled trials and cohort studies with naturally occurring groups, such as males and females, or those with and without MHE. Although Table 1 in almost every article reporting the results of a randomized controlled trial consists of a list of baseline variables and associated significance tests, we[2] and others[10] believe that this practice is misguided and reflects a misunderstanding of null hypothesis significance testing. When groups are tested after an intervention, we are looking at two competing hypotheses: The results were due to the intervention, or they could be explained by the operation of chance factors, and the statistical test tells us the probability that the results could have been due to chance. However, if the groups differ at baseline, the probability that this is due to chance is 100%, irrespective of what the P level is; there is no competing hypothesis. Why? Because we divided the participants randomly, so the differences did arise by chance. So, we have very effectively done away with one source of multiplicity.

But, what about baseline testing when the groups were not determined through random assignment? Actually, we do not think it really matters all that much whether one corrects for multiple testing, as the results of these tests do not determine whether the study was successful; they are incidental to its purpose. Because tests for baseline differences are most often used to determine whether these variables should be used as covariates, or to look for possible confounders, it is probably better to err on the side of being overly inclusive, which means not correcting for multiple testing.

The situation of looking at possible explanatory variables, through multiple t tests or correlations is somewhat more problematic, because we believe that the answer depends to a large degree on the researcher's aims (and integrity). At one extreme, the researcher could be on a fishing expedition: "I don't know what's going on, so let's correlate everything with everything else, or do a t test on every possible variable, and see what pops up." Sure enough, some things will pop up, just by chance. We have seen this all too often in what purports to be validity studies for scales; high and low scorers are compared on variables that happen to be available (eg, age, sex, marital status), without any a priori hypotheses

regarding why they should differ, and any differences are used to "prove" the scale is valid. This is nothing more than data dredging. Because (1) there are many statistical tests, and (2) nothing has been postulated beforehand, the probability that many of the results are due to chance is quite high. Here, we would definitely recommend a correction for multiplicity.

At the other extreme, the researchers know they are treading in unknown waters and state so explicitly. The objective of the study is to look for promising leads—areas that need to be followed up in later studies. In this situation, correcting for multiplicity may, as Rothman[5] suggests, result in too many type 2 errors and prematurely close off potentially fruitful areas of research. So we would advise against correcting in these circumstances, but with the warning that any positive results should be seen as hypothesis generating, not as definitive findings.

Finally, should we correct for multiple outcomes? Again we would reiterate the recommendation of Schulz and Grimes[9] that studies should focus on a small number of outcomes and treat other tests as hypothesis generating. If these primary outcomes have been specified beforehand, then correcting for multiplicity may be too conservative and should be avoided. However, this recommendation comes with three caveats: (1) Recognize that the probability of finding at least one significant result increases with more outcomes, according to the formula outlined previously; (2) the outcomes should not be alternative ways of measuring the same thing (eg, skinfold, waist circumference, and BMI—unless you are really focusing on differences among them) or different manifestations of the same underlying phenomenon; and (3) it depends heavily on the integrity of the researchers, that they do not change what they consider to be their primary outcome(s) on the basis of what is or is not statistically significant.

In conclusion, our recommendation regarding whether to correct for multiple tests is, "It all depends." If a small number of hypotheses have been stated a priori or if the purpose of the study is exploratory, then such corrections are probably not needed. However, in the absence of hypotheses, they are required.

Acknowledgments

Financial/nonfinancial disclosures: The authors have reported to *CHEST* that no potential conflicts of interest exist with any companies/organizations whose products or services may be discussed in this article.

References

1. Streiner DL, Norman GR. Randomized controlled trials. *Community Oncol*. 2009;6(2):83-85.
2. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 3rd ed. Shelton, CT: People's Medical Publishing House; 2007.
3. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statist*. 1979;6(2):65-70.
4. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. New York, NY: Wiley; 1987.
5. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43-46.
6. Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *Am J Epidemiol*. 1998;147(7):615-619.
7. Moyé LA. P-value interpretation and alpha allocation in clinical trials. *Ann Epidemiol*. 1998;8(6):351-357.
8. Altman DG. Statistics in medical journals: some recent trends. *Stat Med*. 2000;19(23):3275-3289.
9. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. *Lancet*. 2005;365(9470):1591-1595.
10. Altman DG. Comparability of randomised groups. *Statistician*. 1985;34(1):125-136.