Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests[1]

*David L Streiner**

Department of Psychiatry and Behavioral Neurosciences, McMaster University, Hamilton, Canada, and Department of Psychiatry, University of Toronto, Toronto, Canada

ABSTRACT

Testing many null hypotheses in a single study results in an increased probability of detecting a significant finding just by chance (the problem of multiplicity). Debates have raged over many years with regard to whether to correct for multiplicity and, if so, how it should be done. This article first discusses how multiple tests lead to an inflation of the α level, then explores the following different contexts in which multiplicity arises: testing for baseline differences in various types of studies, having >1 outcome variable, conducting statistical tests that produce >1 *P* value, taking multiple "peeks" at the data, and unplanned, post hoc analyses (i.e., "data dredging," "fishing expeditions," or "*P*-hacking"). It then discusses some of the methods that have been proposed for correcting for multiplicity, including single-step procedures (e.g., Bonferroni); multistep procedures, such as those of Holm, Hochberg, and Šidák; false discovery rate control; and resampling approaches. Note that these various approaches describe different aspects and are not necessarily mutually exclusive. For example, resampling methods could be used to control the false discovery rate or the family-wise error rate (as defined later in this article). However, the use of one of these approaches presupposes that we should correct for multiplicity, which is not universally accepted, and the article presents the arguments for and against such "correction." The final section brings together these threads and presents suggestions with regard to when it makes sense to apply the corrections and how to do so. *Am J Clin Nutr* 2015;102:721–8.

Keywords: multiplicity, significance testing, statistics, Bonferroni, false discovery rate, family wise error rate

INTRODUCTION

In 2010, Bennett et al. (1) used fMRI, involving "a 6-parameter rigid-body affine realignment of the fMRI time series, coregistration of the data to a T1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing" (whatever all that means) to demonstrate that 3 particular voxels in the brain images showed significant signal changes when the subject was shown photographs of people expressing specific emotions in social situations. Perhaps the reason that this finding did not result in headlines in the major news sources was that the subject in this study was a salmon; not only that, it was a dead one. Another research discovery that failed to garner significant press coverage was by Austin et al. (2). By using a database containing 223 of the most common diagnoses, they found that people born under the sign of Sagittarius had an increased risk of fractures of the humerus, whereas Leos had a higher probability of gastrointestinal hemorrhage than did those born under the other astrological signs combined.

Fortunately for the reputation of science (and the scientists), neither of these articles meant for their conclusions to be taken seriously. Rather, their aim was to highlight the problem of multiple comparisons, also known as *multiplicity*. This is an issue that arises when many statistical tests are performed within a single study; with each test that is run, the probability of finding statistical significance just by chance increases, so it becomes progressively more difficult to separate out true differences or associations from those due to chance. In this article, I discuss *1*) why it is a problem, *2*) under what circumstances multiplicity rears its head, *3*) various ways of correcting for multiplicity, *4*) the controversy with regard to correcting for multiplicity and *5*) offer some suggestions regarding when and how we should correct for it.

Bear in mind that this article is written from what is called the "frequentist" perspective; that is, that the probability of an event is determined by its relative frequency observed in a study. There are other perspectives, primarily the Bayesian, in which previous probabilities are taken into account, but I do not discuss these in this article.

WHY IS MULTIPLICITY A PROBLEM?

If we adopt an α level of 0.05, then by definition, assuming that all of the null hypotheses (H^0s) are true, on average 5% of the

[1] The author reported no funding received for this study.

*To whom correspondence should be addressed. E-mail: streiner@mcmaster.ca

Received April 21, 2015. Accepted for publication July 10, 2015.

First published online August 5, 2015; doi: 10.3945/ajcn.115.113548.

statistical tests will show a significant difference or association. The more tests that are run, the greater the likelihood that at least 1 will be significant by chance, and the question that arises is the probability that this will occur. If 3 tests are conducted, each of which can have 1 of 2 results (significant or not), there are $2^3 = 8$ possible outcomes, which are shown in the left-most columns of **Table 1**, labeled "Test result." The probabilities associated with each result are shown in the next 3 columns, called "Test probability," and the last column is the probability of that outcome. So, the probability for outcome 2 (not significant, not significant, significant) would be $0.95 \times 0.95 \times 0.05 = 0.045125$. Note that there are 2 essential assumptions to these calculations: *1*) all of the null hypotheses are true and *2*) all of the statistical tests are valid under the null, meaning that they yield *P*-value distributions that are uniform on the interval [0,1].

We can answer the question of the number of outcomes with at least 1 significant finding when there really is none by adding up the probabilities of all of the rows that have 1 or more of them (i.e., rows 2–8). We can, but this can become laborious when there are 5 tests ($2^5 = 32$ outcomes) and borders on the masochistic when there are 10 of them ($2^{10} = 1024$ outcomes). Fortunately, there's a much easier way. No matter how many tests there are, the sum of the outcome probabilities is always 1; that is, there is a 100% chance that one of those outcomes will occur. So, we can simply subtract the result from the first row (no significant tests) from 1, and we will get the same answer. In other words:

$$\Pr(\text{at least one test significant}) = 1 - \Pr(\text{no tests significant}) \quad (1)$$

where Pr means "probability."

The probability in row 1 is 0.95^3. To generalize a bit, if there were k tests performed, then the probability would be 0.95^k. We can generalize even further. We said that the test would be wrong 5% of the time, but we're not limited to that; we can use any value, such as 1% or 10%. If we designate the false-positive rate as α, then the probability is $(1 - \alpha)^k$. That means that the probability of at least 1 test being significant is:

$$\Pr(\text{at least one test significant}) = 1 - (1 - \alpha)^k \quad (2)$$

The more tests there are, the greater the probability that one will be significant, as shown in **Figure 1**. If you run enough tests, you're almost guaranteed to find something significant.

WHEN MULTIPLICITY CAN ARISE

Multiplicity can arise under a number of different circumstances. It is important to differentiate among them, because the answer to whether or not to correct differs from one situation to the next. The various situations are:

1) Testing for baseline differences in a randomized controlled trial (RCT).[2] The variables usually consist of various demographic factors, as well as variables that may be possible confounders.

TABLE 1
The 8 possible outcomes of 3 tests assuming the null hypothesis is true[1]

	Test result			Test probability			
Outcome	A	B	C	A	B	C	Outcome probability
1	NS	NS	NS	0.95	0.95	0.95	0.857375
2	NS	NS	Sig	0.95	0.95	0.05	0.045125
3	NS	Sig	NS	0.95	0.05	0.95	0.045125
4	NS	Sig	Sig	0.95	0.05	0.05	0.002375
5	Sig	NS	NS	0.05	0.95	0.95	0.045125
6	Sig	NS	Sig	0.05	0.95	0.05	0.002375
7	Sig	Sig	NS	0.05	0.05	0.95	0.002375
8	Sig	Sig	Sig	0.05	0.05	0.05	0.000125
Total							1.000000

[1]Sig, significant.

2) Looking for differences between or among groups on a number of outcome measures.

3) Running a statistical procedure that yields >1 *P* value, such as factorial or repeated-measures ANOVA, multiple regressions, and so forth.

4) Peeking at data. This involves analyzing the results before all of the participants have been entered to determine whether more people need to be added to reach significance.

5) Interim analyses. These are most often planned ahead of time to see if the study should be ended early.

6) Fishing expeditions; that is, unplanned searches for differences between groups or relations among variables, as well as unplanned subgroup analyses.

Within this list of situations, we can make a number of distinctions. The first is between confirmatory data analysis and exploratory data analysis. The former consists of testing hypotheses that have been specified a priori and the results dictate whether the study is deemed successful or not. In contrast, the researcher does not have explicit hypotheses in the latter case but is rather searching for relations or differences after the fact. The distinction is clear-cut at the extremes but can get blurry in practice. For example, the Cardiovascular Health Study (3) was a prospective cohort study looking at the relation between metabolic syndrome and cardiovascular disease. Without recourse to the original proposal (assuming it exists), it is difficult to know if subsequent analyses of the data looking at the effects of gender and race are exploratory or confirmatory. This is 1 reason that some journals now insist on seeing the proposal as part of the review process.

The second distinction that is sometimes made is between the family-wise error rate (FWER; i.e., the number of false discovery errors in a family of tests) and the experiment-wise error rate (i.e., the number of false positives in the entire study). Imagine that we did a study in which participants were randomly assigned to 1 of 4 diets, and we are looking at 2 different outcomes after 6 mo: the change in BMI and satisfaction with the diet, measured on some scale. Each of the variables would be analyzed with a 1-factor ANOVA. A significant *F* ratio would indicate that there is a difference between the groups but would not tell us where the difference lies. To determine that, we would have to do 6 post hoc tests—group 1 compared with 2, group 1 compared with 3,

[2] Abbreviations used: FDR, false discovery rate; FWER, family-wise error rate; H_0, null hypothesis; pFDR, positive false discovery rate; RCT, randomized controlled trial.

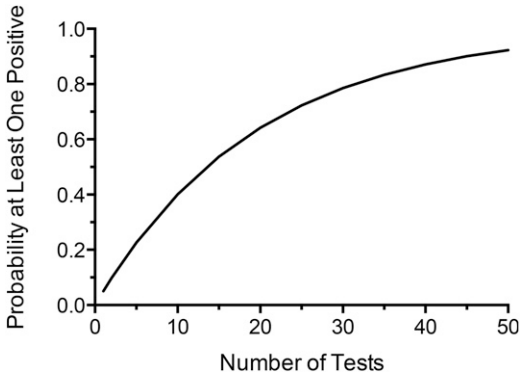

FIGURE 1 Probability that at least 1 test will be positive, for varying numbers of tests and an $\alpha = 0.05$.

group 1 compared with 4, group 2 compared with 3, group 2 compared with 4, and group 3 compared with 4—and each of those would have its own *P* level. This would be referred to as a family of tests. On the other hand, the study as a whole has 2 independent measures, each tested with its own ANOVA, so that there may be inflation of the α level at the experiment level.

Yet again, however, reality rears its ugly head to blur this distinction at 3 levels. First, these terms are not used consistently from 1 article to the next; some authors use the 2 terms as synonyms and would describe the latter situation as a family of tests because they arose from the same study or experiment. Second, what constitutes an "experiment"? Would it include later replications? All of the studies conducted by that author examining the same question? Studies of the same question by other research groups? There is no easy answer. Finally, as we will discuss later, if we correct for post hoc tests, as in the example of the ANOVA, why don't we correct for the multiple *t* tests that accompany a multiple regression? In this article, we will use the term FWER to encompass both definitions.

HOW TO CORRECT FOR MULTIPLICITY

Given all the ways in which multiplicity can arise, how can we correct for it? In fact, there are many ways, which can roughly be divided into 4 areas: post hoc tests run after a significant ANOVA, those that try to correct for many independent analyses in a study, those that try to control the false discovery rate (FDR; which will be defined a bit later), and those based on resampling procedures.

There are a multitude of post hoc tests, such as the Studentized range test, Fisher's least significant difference, Tukey's honestly significant difference and his wholly significant difference, the Newman-Keuls test, Dunnett's *t*, and many others. All of them are variations of *t* tests, with different ways of trying to control the overall α level (4). The Newman-Keuls (also called the Student-Newman-Keuls, or S-N-K) is the default option in programs such as SPSS, perhaps because it is the most powerful of the techniques (i.e., has the highest likelihood of finding a difference between means) but does not control the FWER when there are >3 groups (5).

Among the techniques that correct for all of the statistical tests run in a study, arguably the most widely used one is the Bonferroni correction, which is an application of the Bonferroni inequality. It was named after the Italian mathematician Carlo Emilio Bonferroni and probably was first introduced into the statistical world by Olive Dunn (6, 7). Its popularity is due to a number of factors. First, it is the essence of simplicity. If we want to preserve an overall FWER acceptance rate, α_{FWER}, then we divide α by the number of tests being done (k). That is,

$$\alpha_B = \frac{\alpha_{FWER}}{k} \quad (3)$$

So, if there were 10 statistical tests and we want to restrain the FWER at 0.05, we would use a Bonferroni-adjusted α level (α_B) of 0.05/10 = 0.005 for each. Second, it is very flexible and can be used with any type of statistical test, not just ANOVAs.

Unfortunately, there is a steep price to pay for these benefits, and that is the extremely conservative nature of the correction. As the probability of a type I error (stating there was an effect when none was present) decreases, that of a type II error (concluding that there was no effect of the intervention or no association between variables when in fact there is one) increases. This loss of power is due to a number of reasons. First, it assumes that the null hypothesis is true for all of the tests, and this is unreasonable, most especially after a significant omnibus *F* test. Second, it assumes that all of the tests are independent, which is not true when pairwise comparisons are run, as is the case with the post hoc tests after an ANOVA. For example, if there are 3 groups, A, B, and C, then the comparisons would consist of A vs. B, A vs. C, and B vs. C.

A number of modifications to the Bonferroni have been proposed, such as the Šidák-Bonferroni (8), which uses the following value:

$$\alpha_{S-B} = 1 - (1 - \alpha_{FWER})^{1/k} \quad (4)$$

However, the result is very similar to the simpler Bonferroni value; if there are 10 tests, then $\alpha_B = 0.005$, whereas the Šidák-Bonferroni–adjusted α level (α_{S-B}) = 0.00511, which is 1 reason it is rarely used.

One difficulty with the Bonferroni-class of corrections is that they become increasingly conservative if the outcomes are correlated with one another. Another problem is more philosophical and arises in areas in which many statistical tests are performed (sometimes running into the thousands), often with relatively few subjects, such as genomics and brain scanning. A certain number of false-positive results is tolerable, because they would be discarded when the study is replicated. The more relevant quantity to control is the positive FDR (pFDR); that is, the proportion of false positives among the set of rejected null hypotheses (which are referred to as *discoveries*).

The difference between controlling the FWER and the pFDR can be seen by referring to **Table 2**. When the aim is to control the FWER, we are concerned with the proportion of type I errors or false discoveries (cell C) relative to the total number of true null hypotheses (cells A + C). However, when the objective is to control the pFDR, the concern is the proportion in cell C relative to the total number of rejected null hypotheses (cells C + D). Thus, the pFDR is the expected proportion of false positives among all of the significant statistical tests. Phrased another way, if we use an α level of 0.05, then we expect that $\leq$5% of all tests will result in type I errors. However, by using the pFDR approach, we expect that $\leq$5% of the significant tests will be false positives. Benjamini and Hochberg (9), who coined the term FDR, stated that this approach is more powerful than the

TABLE 2
The outcomes from many tests of significance[1]

Total	H_0 True	H_0 False	Total
Not called significant	A: true negative	B: false negative	A + B
Called significant	C: false discovery	D: true discovery	C + D
	A + C	B + D	k

[1] H_0, null hypothesis.

Bonferroni and is better at separating the important few from the many trivial effects tested.

The Holm, or Holm-Bonferroni, procedure (10) was proposed to try to mitigate the problem of the often-conservative nature of the Bonferroni correction. Although it was developed before the term FDR was introduced, it can be seen as the first—and now perhaps best known—of the techniques to control the pFDR. In contrast to the single-step Bonferroni technique, the Holm method is a sequential, step-down one, in which the per-test α level is changed for each test. If there are k significant tests, then their associated P levels are rank ordered, so that $P_1 \leq P_2 \leq \ldots \leq P_k$. The first value is evaluated against the criterion of α/k. If it is significant, then the next uses the criterion of $\alpha/(k - 1)$. This is continued with decreasing values of k until a P value is reached that is not significant, and then it and all larger values of P are declared nonsignificant. This somewhat labor-intensive task has been made easier by the availability of a number of free online and downloadable computer programs.

The Hochberg method (11), on the other hand, is a step-up procedure, which begins by testing P_k (i.e., the largest P level) against the criterion of α. If it is significant, testing is stopped, and all smaller P levels will also be significant. It the result is not significant, then P^{k-1} is evaluated with $\alpha/2$. Again, if the result is significant, testing is stopped and P levels smaller than this would be significant. If it is not significant, then P_{k-2} is compared against $\alpha/3$, and so forth, until nonsignificance is reached. Of the 3 procedures, the Bonferroni is the most conservative, the Hochberg the least, and the Holm method falls between them.

A different class of approaches to correcting for multiplicity is based on resampling procedures. One variant of this technique, called *bootstrapping*, was introduced by Efron and Tibshirani (12). It involves drawing repeated random samples from the data (sometimes numbering 1000 or even 10,000 samples) and calculating the parameters for each sample (e.g., the mean), which can then be averaged, thus yielding estimates of the variability of the parameters across different bootstrapped samples. It may seem somewhat odd drawing so many samples from a small data set, but it's possible because the samples are drawn with replacement. That means that if participant number 24 is selected, we actually leave her data in the data set where it can be drawn again. Thus, each sample is somewhat different from the others. Because of this, it was not feasible until desktop computers became widely available and software was developed that can perform the calculations. It was first applied to the issue of multiplicity by Westfall and Young (13, 14) and gained popularity in the field of genetics, where microarray studies can generate thousands of hypotheses tested simultaneously (15).

The major advantage of resampling approaches is that they take into account the estimated dependency among the test statistics, which is not true for some other corrections for multiplicity (e.g., the Bonferroni correction), and therefore tend to be more powerful (less conservative). Resampling methods also make minimal assumptions about the underlying distribution. They can also be applied to single-step (i.e., Bonferroni-type), multistep (Holm and Hochberg), and FDR approaches. The mathematics of resampling techniques are beyond the scope of this article; interested readers should read one of the cited texts (13–15).

SHOULD WE CORRECT FOR MULTIPLICITY?

The discussion of how to correct for multiplicity has made the implicit assumption that we should correct for it, but this is by no means a position accepted by everyone. In its favor, Moyé (16) holds that "Type I error accumulates with each executed hypothesis test and must be controlled by the investigators" (p. 354); Cormier and Pagano (17) state that "The more tests we want to make, the more conservative we have to be in order to preserve our overall significance level α" (p. 333); and Blakesley et al. (18) write that "Failure to control type I errors when examining multiple outcomes may yield false inferences, which may slow or sidetrack research progress" (p. 256).

On the other side of the debate, Rothman (19) argues that correcting for multiplicity is predicated on 2 assumptions: *1*) that the main cause of unusual findings is chance and *2*) that no one would want to further investigate phenomena that may have been caused by chance. Rothman disputes both of these beliefs. With regard to the first, he states that "Scientists presume instead that the universe is governed by natural laws, and that underlying the variability that we observe is a network of factors related to one another through causal connections. To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premise of empiricism" (p. 45). As for the second, he writes that "Being impressed by an extreme result should not be considered a mistake in a universe brimming with interrelated phenomena. The possibility that we may be misled is inherent to the trial-and-error process of science; we might avoid all such errors by eschewing science completely, but then we learn nothing" (p. 46).

The danger is that, by correcting for multiplicity, we increase the probability of a type II error, and thus may overlook possibly interesting findings. It should be noted that this is based on the assumption that some other group will try to replicate the findings and fail. However, some journals will not publish replication studies, and many tenure and promotions committees place more value on original research, so this correction process may not take place.

What Rothman overlooks is that the purpose of null hypothesis significance testing is not simply to reject or not reject the null hypothesis but also to determine both the direction of the effect and its magnitude. Furthermore, as Cohen points out in his delightfully titled paper, "The earth is round ($p < .05$)" (20), there is a difference between the null hypothesis (the hypothesis to be nullified) and the nil hypothesis (nothing is going on). Most often, the 2 are the same, but they need not be. We can state that our null hypothesis is that a correlation is at or below a certain value (as is often done in determining the reliability of a scale) or that a difference between groups is a given amount or more (as is done in noninferiority trials).

A different argument against correcting for multiplicity is offered by Schulz and Grimes (21). Suppose that an RCT of preoperative parenteral nutrition resulted in an increase in noninfectious

complications (e.g., 22), and that this finding was significant at the 0.04 level. Now also assume that the investigators looked at a second outcome, length of hospitalization, which was also significant at the 0.04 level. As would be expected, these 2 outcomes are highly correlated, which is the case for many endpoints in clinical trials. The fact that both outcomes resulted in significant findings in the same direction might reinforce our confidence in the results. However, were we to apply a Bonferroni-type correction, neither outcome would be significant, which is counterintuitive. Nevertheless, a weakness of this argument is that the 2 (or more) dependent variables are unlikely to be independent if they are measures of conceptually related outcomes. Taken to the extreme, imagine testing for a treatment effect on weight loss in pounds and then testing for a treatment effect in the same data set on weight loss in kilograms. Because kilograms is just a linear transformation of pounds, the 2 tests will yield identical *P* values and then obtaining 2 *P* values of 0.04 would not tell us anything different from 1 *P* value of 0.04. This example may be absurd, but it points out the difficulty in trying to argue that testing multiple related outcomes in some way obviates the multiple testing problem. It is true that when the outcomes are highly related, use of a multiple testing procedure that incorporates that dependency into the correction may be apt. Both sides of the argument make valid points, resulting in one leading statistician, Doug Altman (23) to make the conclusion that "It is hard to see views such as [the ones just cited] being reconciled" (p. 2383).

WHEN MIGHT WE CORRECT OR NOT CORRECT FOR MULTIPLICITY

Perhaps a way out of this quandary can be found by looking at the various conditions in which multiple hypothesis testing arises, described earlier, and discuss when it may or may not make sense to correct for multiplicity. The first situation occurs when we look for baseline differences in an RCT. This is rationalized for 2 reasons: as a check on the randomization process and to determine whether any variables should be used as covariates in subsequent analyses. We can actually deal with this quite easily—it shouldn't be done. Despite the fact that such testing is almost universally reported as the first table in any RCT, we (4) and others (24) believe that it is misguided on both grounds. When we run a statistical test after an intervention, we are testing 2 hypotheses: that any difference was due to chance (H_0) or that it was due to the fact that the groups were treated differently (the alternative hypothesis). However, at baseline, there is no alternative; if differences exist, they must be due to chance (assuming that the randomization process has not been subverted, either in some nefarious way or innocently, such as by replacing dropouts or not reading the instructions). Hence, the *P* level is meaningless; the probability that chance was responsible is 100%, irrespective of the value printed out by the computer. The second argument, that variables that differ at baseline should be used as covariates, is equally misguided. Whether or not a variable should be included as a covariate should be based on theory or previous knowledge of its influence on the outcome and not on inspection of baseline differences (25). Furthermore, because including covariates usually reduces the within-group variance and increases the precision of the estimate of the treatment effect (4), it is often recommended to include prognostic variables as covariates, even if they do not differ significantly between the groups.

Looking for group differences among the outcome variables, the second situation, is the most fraught with difficulties because it is on the basis of these analyses that the study is said to have provided evidence to reject the main null hypotheses or not, and it is here that the differences between the varying viewpoints are sharpest. To reiterate, the debate is between finding differences that are actually due to chance compared with overlooking potentially useful findings. The most sensible advice was given by Schulz and Grimes (21): "Researchers should restrict the number of primary endpoints tested. They should specify a priori the primary endpoint or endpoints in their protocol" (p. 1592). That is, the problem of correcting for multiplicity is eliminated by making it unnecessary; there are only a small number of endpoints (ideally, 1) and they will likely be correlated. Because of this, Bonferroni-type corrections could undermine the conclusions, as pointed out earlier, because the outcomes would reinforce each other rather than having one lessen the significance of the other (bearing in mind the injunction that the variables should not simply be the same outcome measured in different ways). Their injunction about specifying the outcomes a priori is to preclude substituting a significant secondary outcome for a primary one that was not significant; a practice that was (and is) all too common (e.g., 26). It is for this reason that many medical journals now require all trials to be registered before the first patient is enrolled, and some journals require authors to include the protocol when the results are submitted for publication.

However sensible Schulz and Grimes' (21) advice is with regard to limiting the number of outcomes, it is infeasible when evaluating complex interventions. These are defined as ones with "several interacting components" (27, p. 6) and are often used in health services, public health, and social policy research. The guidelines promulgated by the Medical Research Council in Great Britain state that "Identifying a single primary outcome may not make best use of the data; a range of measures will be needed, and unintended consequences picked up where possible" (27, p. 7). For example, the Moving to Opportunity project (28) was an RCT that evaluated the effect of neighborhood on the development of obesity and diabetes. There was a broad range of outcomes indicating health outcomes, including height, weight, and concentration of glycated hemoglobin. Because these outcomes were all specified a priori in the protocol, there was no correction for multiplicity.

The third situation, post hoc tests after a significant omnibus test, is a somewhat confusing one. On the one hand, most statistical packages provide $\geq$1 of the post hoc tests mentioned earlier (e.g., Newman-Keuls, the honestly significant difference) after an ANOVA, and this appears to be accepted practice. On the other hand, multiple linear regression with categorical predictor variables is mathematically identical to ANOVA (29), whereby group differences are reflected in the b or β weights of a dummy coded variable. However, it is highly unusual to see any correction for multiplicity applied to the t tests of these weights. Why this difference? In the words of Tevye the Milkman (from *Fiddler on the Roof*), Tradition!

Actually, this "explanation" is less facetious than it may first appear. Nearly 60 y ago, Lee J Cronbach (30) wrote about the 2 "disciplines" of psychology: the experimental and the observational. The former tries to exert all possible control over a study,

views within-group variance as noise to be minimized as much as possible, and examines only a few hypotheses at a time; in contrast, the latter looks at nature as it is, welcomes variance as necessary to explore between-person differences, and studies many variables at the same time (hence leading to the aphorism, "One person's error variance is another person's occupation"). The experimentalists favored ANOVA-type statistics, whereas the observationists relied primarily on correlations and regressions. This may be the origin of the difference, with the first discipline concerned about spurious findings and the second welcoming unexpected results, and is reflected in the more contemporary differing viewpoints of Blakesley et al. (18), on the one hand, and Rothman (19), on the other.

Situation 4 involves "peeking" at the data; that is, analyzing them partway through the study to determine whether the sample size needs to be increased. This practice is most definitely one that should be avoided entirely. The issue, as Armitage et al. (31) pointed out, is that assuming that H_0 is true, the probability of a significant test result grows rapidly with each analysis. The problem is compounded by the fact that the sample size is increased after each nonsignificant result, making the likelihood of finding significance even greater. Eventually, given enough peeks at the data, the researcher will find what he or she is looking for. The only accepted practice is to determine the sample size a priori, and to stick with that.

The fifth situation, interim analyses, also involves peeking at the data, but is most often built into many large RCTs at the design stage (32). The rationale is primarily an ethical one. If an analysis partway through the study shows that the new intervention will not prove to be superior to the comparison (either placebo or treatment as usual) when the full sample size is reached, or if 1 group has significantly more adverse outcomes than the other, it would be unethical to continue with the study. The lack of effectiveness was the reason that the tolbutamide and diet arm of the University Group Diabetes Program (33) was dropped halfway through the trial. Conversely, if the new intervention is clearly superior, then it would be equally unethical to deny it to those in the comparator condition. For example, the Multicenter Automatic Defibrillator Implantation Trial (34) was ended early because the interim analysis showed a significantly greater reduction in all-cause mortality for those given an implantable cardioverter-defibrillator.

Various schemes have been proposed to protect the overall α level, including the use of the same significance level for each interim analysis (35) or a gradually decreasing criterion (36). Common practice now appears to be to divide α into 2 parts, α_1 for the interim analysis and α_2 for the final one (assuming only 1 interim analysis), so that $\alpha_1 + \alpha_2 = 0.05$. Following the recommendation of Peto et al. (37), a very stringent criterion is used for α_1 (e.g., <0.001). The rationale for this is that trials that are ended early tend to overestimate the effect size and underestimate the CI (38). As a result, interim analyses should be used primarily when *1*) there is serious risk of harm from adverse events or delaying the new treatment and *2*) when they have been built into the study from the beginning for reasons of efficiency and then *3*) used only with greatest of caution.

It is in the final situation, involving unplanned and subgroup analyses, that the tension between chance findings on the one hand and overlooking potentially interesting observations on the other is most acute. The problem in fact goes beyond simply doing and reporting a large number of analyses after the fact. As Gelman and Loken (39) pointed out, even without going on a "fishing expedition," a study can have a large number of "researcher degrees of freedom," potentially involving choice of statistical test and whether and how to categorize continuous variables, nonlinear transformations, outlier removal, etc. The (often undisclosed) use of such techniques in search of statistically significant results is referred to as "*P* value fiddling" or "*P*-hacking." That is, there is always a very large number of potential analyses (e.g., analyzing the data by gender, age, comorbidity status, previous history), of which the researcher may perform only a few, but the choice is often conditional on the data themselves. Even a cursory look at the data may lead the researcher to decide that it is not worth running some analyses because the difference looks quite small. The issue is that although no formal statistical procedures have been performed, informal, "eyeball" tests were run. Thus, there is a greater likelihood that the statistical tests that actually were run will be significant. Compounding the problem, the comparisons that were rejected as not likely to be fruitful are not counted when correcting for multiplicity, further increasing the probability of a type I error.

Perhaps the most prudent course of action in these circumstances would consist of 3 parts. First, there should be a correction for the number of tests that were actually performed. The correction could be either a Bonferroni-type one or a pFDR type. Second, both the corrected and uncorrected *P* levels should be reported, so that the readers are able to determine for themselves which tests they would regard as statistically significant or not. Finally, any conclusions based on such analyses should be clearly reported as tentative and hypothesis generating, rather than as hypothesis testing.

SUMMARY AND RECOMMENDATIONS

Whether or not to correct for multiplicity is an issue that shows no sign of early resolution. The issues on both sides are compelling. Not correcting for it increases the probability of spurious significant findings, possibly resulting in time and resources being wasted chasing down false leads. On the other hand, correcting for multiplicity may have the opposite effect, in which potentially interesting observations are discarded as chance findings. After reviewing the various situations in which multiplicity can arise, the following recommendations are offered, in the full realization that they may be contested and debated:

1) The decision regarding whether or not to correct for multiple testing is a philosophical one and there is no way to prove that correcting is or is not the right thing to do unless one specifies one's values (e.g., one's preferences regarding the potential commission of certain types of errors) in advance.

2) In determining whether groups differ at baseline in an RCT, the significance levels are meaningless (unless one suspects that the randomization was not legitimately implemented) and therefore significance testing should not be done.

3) When assessing the outcomes of a clinical trial, some degree of judgment is necessary. Ideally, there will be only a small number of outcomes, which have been specified

a priori and are most likely correlated with one another. In such cases, correcting for multiplicity may be judged to be unnecessary and counterproductive; if the outcomes are in the same direction, they would strengthen confidence in the results. Yet, if many endpoints are used, some investigators would find the potential for FWER inflation unacceptable.

4) Correcting for many *P* values within a single statistic, such as complex ANOVA designs and multiple regression, appears to be dictated more by habit and tradition than by logic. Post hoc tests corrected for multiple testing are routinely used in the former case and almost never in the latter, even though the 2 techniques are mathematically identical. It is unlikely that these practices will change in the foreseeable future.

5) "Peeking" at the data to determine whether a larger sample size is needed is poor practice and should never be done unless a preplanned interim analysis strategy is used that protects the overall α level. On the other hand, interim analyses are an integral part of many RCTs to determine whether a trial should be ended early because of futility (i.e., the intervention will not show benefit even with the full sample size), an excess of adverse events in 1 group, or the clear superiority of the intervention, meaning that it would be unethical to withhold it from the comparison group. However, because trials that are ended early often overestimate the true effect size and underestimate the width of the CI, this should be done only using methods designed for such situations (e.g., 40).

6) When conducting unplanned, post hoc analyses of the data, including subgroup analyses, correcting for multiplicity should be used. It is strongly recommended that both the corrected and uncorrected *P* values be reported and that all findings be reported as tentative and hypothesis generating, rather than hypothesis testing.

7) If corrections are used, the Bonferroni-type is simple but not optimal and should generally not be a first choice with a large number of tests. Holm and Hochberg methods are superior in this regard. Indeed, the researcher may wish to change the criterion with regard to which outcomes are significant and use the pFDR approach. Resampling techniques are very promising for those with the wherewithal to implement them, and their use will likely become more widespread as they are incorporated into the commonly used software packages.

TO READ FURTHER

In addition to the articles listed in References, other useful books about correcting (or not correcting) for multiplicity would include the following:

Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen WW. Analysis of clinical trials using SAS: a practical guide. Cary (NC): SAS Institute; 2005.

Dmitrienko A, Tamhane AC, Bretz F, editors. Multiple testing problems in pharmaceutical statistics. Boca Raton (FL): Chapman & Hall/CRC; 2010.

Dudoit S, van der Laan MJ. Multiple testing procedures with applications to genomics. New York: Springer-Verlag; 2008.

Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1987.

Hsu J. Multiple comparisons: theory and methods. Boca Raton (FL): Chapman & Hall/CRC; 1996.

Toothaker LE. Multiple comparisons for researchers. Newbury Park (CA): Sage; 1991.

The author did not declare any conflicts of interest.

REFERENCES

1. Bennett CM, Baird AA, Miller MB, Wolford GL. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for proper multiple comparisons correction. Journal of Serendipitous and Unexpected Results 2010;1:1–5.
2. Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol 2006;59:964–9.
3. McNeill AM, Katz R, Girman CJ, Rosamond WD, Wagenknecht LE, Barzilay JI, Tracy RP, Savage PJ, Jackson SA. Metabolic syndrome and cardiovascular disease in older people: the Cardiovascular Health Study. J Am Geriatr Soc 2006;54:1317–24.
4. Norman GR, Streiner DL. Biostatistics: the bare essentials. 4th ed. Shelton (Connecticut): PMPH USA; 2015.
5. Seaman MA, Levin JR, Serlin RC. New developments in pairwise multiple comparisons: some powerful and practicable procedures. Psychol Bull 1991;110:577–86.
6. Dunn OJ. Estimation of the medians for dependent variables. Ann Math Stat 1959;30:192–7.
7. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961; 56:52–64.
8. Šidák ZK. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 1967;62:626–33.
9. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57: 289–300.
10. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
11. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–2.
12. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
13. Westfall PH, Young SS. p Value adjustments for multiple tests in multivariate binomial models. J Am Stat Assoc 1989;84:780–6.
14. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. New York: Wiley; 1993.
15. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Technical Report No.: 633. January 2013 [cited 2015Apr 6]. Available from: http:// statistics.berkeley.edu/sites/default/files/tech-reports/633.pdf.
16. Moyé LA. P-value interpretation and alpha allocation in clinical trials. Ann Epidemiol 1998;8:351–7.
17. Cormier KD, Pagano M. Multiple comparisons: a cautionary tale about the dangers of fishing expeditions. Nutrition 1999;15:332–3.
18. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF III, Butters MA. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology 2009;23: 255–64.
19. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6.
20. Cohen J. The earth is round ($p <$. 05). Am Psychol 1994;49:997–1003.
21. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet 2005;365:1591–5.
22. Bozzetti F, Gavazzi C, Miceli R, Rossi N, Mariani L, Cozzaglio L, Bonfanti G, Piacenza S. Perioperative total parenteral nutrition in malnourished, gastrointestinal cancer patients: a randomized, clinical trial. JPEN J Parenter Enteral Nutr 2000;24:7–14.
23. Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000;19:3275–89.

24. Altman DG. Comparability of randomised groups. Statistician 1985; 34:125–36.
25. Roberts C, Torgerson DJ. Baseline imbalance in randomised controlled trials. BMJ 1999;319:185.
26. Marti-Carvajal A. Taking aim at a moving target: when a study changes in the middle. In: Streiner DL, Sidani S, editors. When research goes off the rails: why it happens and what to do about it. New York: Guilford; 2010. p. 299–303.
27. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: new guidance. Medical Research Council; 2000 [cited 2015 Apr 12]. Available from: http://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/.
28. Ludwig J, Sanbonmatsu L, Gennetian L, Adam E, Duncan GJ, Katz LF, Kessler RC, Kling JR, Lindau ST, Whitaker RC, et al. Neighborhoods, obesity, and diabetes—a randomized social experiment. N Engl J Med 2011;365:1509–19.
29. Cohen J. Multiple regression as a general data-analytic system. Psychol Bull 1968;70:426–43.
30. Cronbach LJ. The two disciplines of scientific psychology. Am Psychol 1957;12:671–84.
31. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J R Stat Soc Ser A-G 1969;132(2):235–44.
32. Coffey CS. Statistical concepts for the stroke community: you may have worked on more adaptive designs than you think. Stroke 2015;46: e26–8.
33. Meinert CL, Knatterud GL, Prout TE, Klimt CR. The University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. Diabetes 1970;19(Suppl 2):789–830.
34. Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, Daubert JP, Higgins SL, Brown MW, Andrews ML. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. N Engl J Med 2002;346:877–83.
35. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191–9.
36. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.
37. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. introduction and design. Br J Cancer 1976;34:585–612.
38. Pocock S, White I. Trials stopped early: too good to be true? Lancet 1999;353:943–4.
39. Gelman A, Loken E. The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. 14 November 2013 [cited 2015 Apr 6]. Available from: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.
40. Bowalekar S. Adaptive designs in clinical trials. Perspect Clin Res 2011;2:23–7.