FROM FARMS TO PHARMACEUTICALS: MULTIPLE COMPARISONS ENTERS THE 21ST CENTURY

Peter H. Westfall

FROM FARMS TO PHARMACEUTICALS: MULTIPLE COMPARISONS ENTERS THE 21ST CENTURY

Peter H. Westfall, Mail Stop 2101, Texas Tech University, Lubbock, TX 79409

ABSTRACT

The subject of multiple comparisons has early roots in statistical methods that were applied in agricultural sciences, often using simple methods with questionable properties popularized by no less than R.A. Fisher. More recently, problems of multiple comparisons have surfaced in the pharmaceutical sciences. In the extremely competitive and highly regulated pharmaceutical environment, it has become essential to take multiple comparisons more seriously, and to use more sophisticated methods. In this paper I describe the need for considering multiple comparisons, with special reference to the recent shift in approach to multiple comparisons in the pharmaceutical industry. I also offer speculations about the effect of this shift on how multiple comparisons will come to be viewed the 21st century.

1. INTRODUCTION: MY PERSONAL JOURNEY FROM FARMS TO PHARMACEUTICALS WITH MULTIPLE COMPARISONS

I am a farm boy. My parents had a walnut orchard in the northern Sacramento Valley of California, and as a kid I would slog through mud with irrigation pipes, spray dangerous-sounding chemicals around the tree trunks, and swat at deerflies in the blazing sun.

As a student at the University of California at Davis, I was an "aggie." I saw plenty of applications to agriculture in my classes and in the consulting lab at UCD. A lot of it was standard stuff—blocks and treatments. Of course, there was always the problem of deciding *which* treatments differed from *which other* treatments. This is the fundamental issue addressed by multiple comparisons procedures (MCPs); it is applicable in all areas of scientific research, agriculture or otherwise.

Alan Fenech, my Ph.D. advisor and main mentor at UCD (and one of my the best teachers) was always concerned with MCPs, and was instrumental in sparking my interest in the subject. In an introductory class on ANOVA, he taught that it was *not* o.k. to test the most extreme means using the ordinary two-sample t-test. But it had seemed like such a good idea to me. After all, if one wants to *show* significance, why not test the hypothesis that is most likely to *produce* significance? Later, in advanced classes in linear models and multivariate analysis, I became enamored with Scheffés projection theorem, and even attended an American Mathematical Association conference as a Pi Mu Epsilon (an undergraduate mathematics fraternity) delegate to give a presentation on it.

My Ph.D. dissertation (Westfall, 1983) and subsequent early research concerned variance components models, which can be applied to farms through animal breeding studies, but which are not directly related to MCPs. One paper I published was quite out of character, involving multiple comparisons with multivariate binary data using the bootstrap (Westfall, 1985), with biomedical application.

Soon after this publication, the pharmaceutical industry came calling in the person of Stan Young, then at Eli Lilly in Indiana. He thought the multivariate binary multiple comparisons

could be easily applied to animal carcinogenicity studies. False positives in animal carcinogenicity studies had been problematic (Fears et al., 1977), and the bootstrap procedure seemed to be a promising solution. Following my presentation at the 10th Midwestern Biopharmaceutical Statistics conference (Westfall, 1987), a consortium of pharmaceutical companies decided to fund the development of software to perform the analysis. The software was to be donated to SAS Institute, and was initially called "PROC MBIN" in 1989 (Westfall et al., 1989), which computed bootstrap multiplicity-adjusted p-values for multivariate binary data. Later, the software was extended to the continuously distributed case, so the name was changed to PROC MTEST (Westfall et al., 1990). When SAS Institute decided to adopt the software in 1992, the name again had to be changed since there were already a number of SAS utilities called "MTEST." We decided to change the name to "PROC MULTTEST." Today PROC MULTTEST is a mature software product, having gone through nearly a decade and a half of extensive testing and updating, and is discussed in three books (Westfall and Young, 1993, Westfall et al, 1999, Westfall and Tobias, 2000). Since the inception of PROC MULTTEST I have consulted regularly with various pharmaceutical companies about matters relating to multiple comparisons in clinical data, pre-clinical data, and most recently, genetics.

So, that concludes my journey from Farms to Pharmaceuticals. What have I learned?

I have learned that multiple comparisons is a subject that has attracted considerable attention over the history of statistics, and which remains controversial even to this day. Based on my experiences, my contention is that *if the discipline of statistics matters to science,* then the theory and methods of multiple comparisons *are extremely* important. Stating the contrapositive, I would contend that if theory and methods of multiple comparisons are unimportant, then our discipline of statistics has no relevance to science.

I say these things from a practical perspective, not a theoretical one. For years, Stan Young and I have collected reports of likely false positives in the scientific literature and in the popular press which are likely false positives resulting from a careless approach to multiplicity. I have also consulted with pharmaceutical companies where approval of a multi-million dollar drug might rest on issues related to multiple comparisons. And finally, the multiple testing issues in genetics that have recently emerged will require us all to acknowledge the multiplicity problem, and to adopt strategies to manage it.

In this paper I give an overview of the multiple comparisons issues, and some emerging strategies for managing the problem. I will finish with a discussion of latest trends in MCPs, and attempt to gaze into the crystal ball to see what the 21st century might hold for us. Particular application will be given to pharmaceuticals, wherein much of my experience lies, as well as genetics, which increasingly influences the statistical science of both farms and pharmaceuticals.

2. WHY BOTHER WITH MULTIPLICITY?

Multiplicity is pervasive in all experiments. Rarely does a study hinge on one and only one test. Multiple measurements all types are always analyzed in statistical studies, as the cost of additional measurement is miniscule compared to the cost of an additional observation. This is as it should be – I do not suggest that information not be collected, or not be analyzed. Rather, appropriate caution should be taken in data interpretation, with recognition of the fact that multiplicity effects are as real as the effects of flawed designs, confounding and the like.

At a recent conference, Juliet Shaffer said, "Multiple Comparisons and Multiple Testing problems are some of the most important issues facing practicing statisticians today. I am surprised more people aren't working in this area." I echo her sentiment. Every one seems to know about the multiplicity problem, but are some statisticians afraid to admit it publicly for fear that the public will come to distrust results of statistical studies?

NEWS FLASH

THE PUBLIC ***ALREADY*** DISTRUSTS THE RESULTS OF STATISTICAL STUDIES!

The following claimed "associations" were taken from newspapers and popular press: (a) cellular phones are "associated" with brain tumors, (b) power lines with leukemia (more recently overturned by the scientific community), (c) vitamins with IQ, (d) season of the year with mental performance (but only in men!), (e) abortions with breast cancer (but not spontaneous abortions), (f) remarriage with cancer, (g) electric razors with cancer, and on and on. A careful reading of these articles suggests that failure to manage multiplicity effects caused the scientists the report the given associations.

There are examples from clinical trials: An article in the *Wall Street Journal* (King, 1995) reported that a drug company's stock dropped 68% when apparently "significant" results from preliminary Phase II clinical studies failed to replicate in the Phase III trial. The "significant" results were found in a subgroup analysis that failed to consider the multiplicity problem. In a similar example reported in *Statistical Science* (Fleming, 1992), a conclusion that pre-operative radiation therapy improves survival of colon cancer patients was likewise based on a subgroup analysis. In each of these two examples, further data collection revealed that the initial "significant" results were likely to be Type I errors resulting from the multiplicity effect.

A final example is from a very controversial epidemiology study. Needleman et al. (1979), stated that lead in drinking water adversely affected IQ's of school children. While high levels of lead are indisputably toxic, the study aimed to prove that variations in levels of lead well below the accepted "safe" level were in fact associated with mental performance. Ernhart et al. (1981), in a critical review of their finding, claimed that the statistically significant conclusions were "probably unwarranted in view of the number of nonsignificant tests." Ernhart, et al. essentially repeated the study and found no evidence for decrease in IQ. As it turns out, it was only after data manipulation that significant Lead/IQ associations were found. As reported in Palca (1991), "the printouts show[ed] that Needleman's first set of analyses failed to show a relationship between lead level and subsequent intelligence tests."

How are the scientists so easily fooled by data analysis? It is well known that statistical conclusions can be wrong. However, they commonly blame faulty experimentation, study apparatus, patient population, and the like, and completely ignore multiplicity. I suspect that the reason lies in the probabilistic underpinnings of the multiplicity issue. In my experience, scientists tend to think deterministically. Probability is a difficult concept for many otherwise knowledgeable scientists, and what is not well understood tends to be downplayed by those with great expertise in other areas.

To educate our scientific colleagues about the multiplicity problem, we can teach them that multiplicity is an EFFECT. Scientists know about treatment effects, covariate and confounding variable effects, even more complicated effects of like nonresponse, missing data, and

measurement error. These are essentially fixed effects, though, while multiplicity is a random effect. If we emphasize that multiplicity is as likely a source of faulty conclusions as any of the usual fixed-effect suspects, then perhaps scientists will take notice.

The cost of measuring additional variables on an experimental unit is usually very small relative to the cost of the unit itself, leading to data sets with myriads of variables. This fact, coupled with ease of statistical computations provided by modern software, as well as the "publish or perish" imperative in universities and medical research centers, can lead easily to the discovery of results that are, in reality, nothing but spurious artifacts caused by the multiplicity effect. Many users do not perceive that the problem exists, and routinely sift through large complex data sets with increasingly user-friendly software, searching for "significant" (and therefore publishable) results.

Statisticians are often discouraged from promoting proper use of multiple comparisons and multiple testing adjustments because they will result in fewer publishable results. I know this from personal experience: in a consulting project, the scientist told me that my services were no longer needed after I solved his problem using multiplicity-recognizing methods (the solution appears in Westfall, 1985).

3. DIFFERING APPROACHES

Much of the controversy about MCPs stems from the wildly divergent approaches to handling the problem that have been proposed in the literature. There are suggestions not to perform any sort of multiplicity adjustment (Saville, 1990; Rothman, 1990; Bailar, 1991; Cook and Farewell, 1996). On the other hand, there is a large literature on methods for handling multiplicity problems. For brevity, I shall simply reference books by Miller (1981), Hochberg and Tamhane (1987), Westfall and Young (1993), Hsu (1996), and Westfall et al. (1999). In these books and in their references are ample arguments for considering various levels of multiplicity adjustment.

While it is impossible in this space to elucidate all of the issues relating to the multiple inference problem, here are a few of the main concerns:

- Multiple comparisons procedures result in more conservative inferences.
- Which error rate should we control: familywise error rate (FWE, e.g., as controlled using the Bonferroni procedure), comparisonwise error rate (CER, no multiplicity adjustment), or false discovery rate (FDR, described below)?
- Assuming we adopt a method for controlling error rates over families (e.g., FWE or FDR), what shall we use for a "family" of tests?
- How do the relative consequences of Type I and Type II error enter the picture?
- Should we "avoid" the problem by using Bayesian methods?
- What are you trying to prove (or not prove)?

Let me relay some personal experiences regarding this last bullet point, derived from my involvement in the pharmaceutical arena.

Pharmaceutical companies are regulated by the US Food and Drug Administration (FDA), which is charged with assuring that pharmaceutical products are *safe* and *effective*. Initially, PROC MULTTEST was conceived as a tool for analyzing safety data, namely, animal carcinogenicity data. These companies were very interested in this technology, since the

conservativeness of the methodology makes it less likely that the study data will show carcinogenicity when the drug truly has no effect. On the other hand, the FDA was not enamored with the technology, because it could mask potential tumor effects.

Later, it was discovered that PROC MULTTEST could be used as a tool for analyzing efficacy data from clinical trials as well. When I presented this application to pharmaceutical companies, I found little interest, for the following reason: such an application would make it more difficult to establish efficacy of their products. On the other hand, I found FDA representatives were quite receptive to the idea of using PROC MULTTEST for the analysis of clinical efficacy data! Clearly, the researcher's and reviewer's perspectives are quite different.

4. THE NECESSITY OF MULTIPLE COMPARISONS PROCEDURES

The conservativeness of MCPs is perhaps the main reason that practicing scientists would like to just ignore the subject. In the prevailing climate of "publish or perish," scientists feel that MCPs can put them at a disadvantage. T.A. Ryan posted the following on the internet,

> I believe that the cost of Type I errors is badly underestimated. To the researcher, Type II errors have great personal cost -- he can't get his paper published or he misses his promotion. Our treatment of data,however, ought to be based upon the cost to science -- is it really important if we miss a small effect? Isn't it more important to find the big ones? The cost of Type I error includes a lot of time wasted by researchers trying to explain a non-existent effect. The falsely "significant" finding can result in a furor of activity which gradually peters out because there wasn't really any effect to work on. In practical research a Type I error can mean the use of a treatment which really does no good. This is surely an important cost.

We statisticians must accept much of the blame for cavalier attitudes toward Type I errors. When we teach practitioners in other scientific fields that multiplicity is not important, they believe us, and feel free to thrash their data set mercilessly, until it finally screams "uncle" and relinquishes significance. The recent conversion of the term "data mining" to mean a statistical *good* rather than a statistical *evil* also contributes to the problem.

As a result of our failures to communicate the problems with multiplicity, scientists often do not recognize the role of probability in the interpretation of results of studies. A recent article in *Science* (Vol. 290, 15December 2000, p. 2031) reports on a follow-up study that failed to replicate a genetic association found in a previous study. The article cites study differences, and possibly protocol differences, but failed to acknowledge that the original significance could easily have been a Type I error. (In genetic studies, the multiplicity problem is rampant, with myriads of genes to be tested, and even multiple tests within genes, e.g., for dominant, recessive, and additive allelic effects, see Westfall et al, 2001).

What are we doing wrong in our education of scientists? Has "probability" become a dirty word? Have we become so enamored with computer-based analyses that we have stopped teaching the fundamental importance of probability for the interpretation of statistical data?

If we train our clients to understand randomness as an effect, then they should have fewer problems grasping the idea that data dredging is likely to turn up artifacts. We can use simple simulation studies (in addition to the probability calculations) to drive this concept home. Often, in my consulting practice, I have clinicians roll two 10-sided dice, one black and one white, each having digits 0-9. Together these give you a "p-value": the black die gives the tenths and the white one gives hundredths; for example Black=3 and White=6 gives p-value = 0.36. The

participants roll 120 times, and enter the data in a grid that is labeled with different clinical outcomes in different patient subgroups. At the end of the "experiment", they are to write up the "results" of their studies. There are always interesting stories to tell about the "effect" of the drug based on the pattern of the significant p-values! Understanding that randomness is the true effect helps practicing scientists recognize that randomness is also an effect in their real studies, especially in analyses of large, complex data sets.

5. SOME MCPS FOR CLASSICAL "FARM" PROBLEMS

5.1 The CRD and RCBD: Simultaneous Intervals

The completely randomized design (CRD) and randomized complete block design are the "classical" cases where MCPs are taught. "Tukey's Method" in particular is highly recommended for confidence intervals in such applications: Assume either model, in the balanced case, with sample means $\bar{X}_i$ (an average of n observations). i=1,…,g, and s^2, an independent estimate of σ^2 using ν degrees of freedom (df). The Tukey simultaneous confidence intervals for pairwise differences between the group means $\mu_i - \mu_j$ are given by $\bar{X}_i - \bar{X}_j \pm c_\alpha s\sqrt{1/n}$, where c_α denotes the 1-α quantile of the Studentized range distribution with g groups and ν df. These intervals are exact: the simultaneous coverage rate is 100(1-α)%.

In the unbalanced case, the exact critical values for the simultaneous intervals no longer can be obtained from the Studentized range distribution. However, because the joint distribution of the pivotals $\left\{ \dfrac{\bar{X}_i - \bar{X}_j - (\mu_i - \mu_j)}{s\sqrt{1/n_i + 1/n_j}} \right\}$ is a known multivariate t distribution, free of unknown parameters, the exact critical value can be computed with relative ease, either by simulation or by using accurate quasi Monte Carlo methods. See Westfall et al. (1999) for further details.

5.2 Closed Testing and a Note on Fisher's Protected LSD

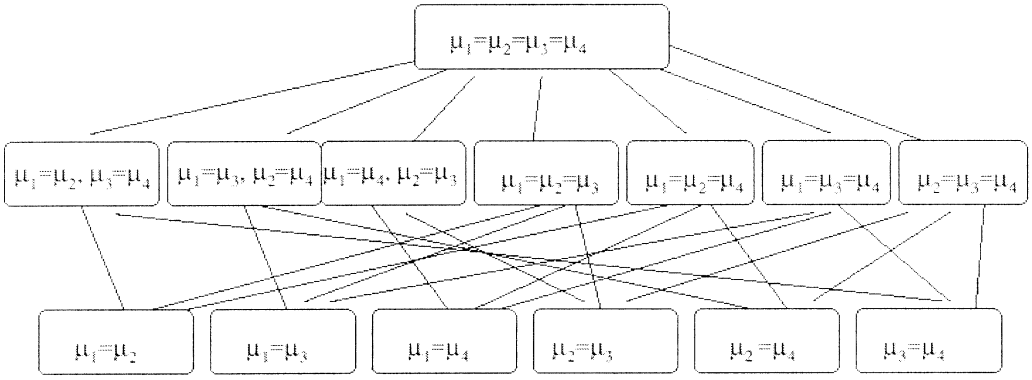
As a researcher at the Rothamstead station, R.A. Fisher recognized the importance of the multiple comparisons problem, and suggested the "Fisher's Protected LSD" procedure. The method is seemingly sensible: When testing a collection of hypotheses, one first tests the global intersection hypothesis, and if this is significant, then one proceeds to test all the remaining hypotheses, unadjusted for multiplicity. Fisher made many wonderful contributions to the subject of statistics, but this was perhaps his worst, for two reasons. First, it fails to control FWE. If a single hypothesis is quite "non-null," while all others are truly nulls, then the method will, with high probability, reject one or more true nulls. Second, the method encourages lazy thinking about multiple comparisons. One must carefully distinguish "partial nulls" from "complete nulls" (Hochberg and Tamhane, 1987, p. 3). To develop methods for managing multiplicity requires that we think from all possible partial null perspectives rather than complete null perspective.

In the pharmaceutical industry, methods based on "closed testing" (Marcus et al., 1976) procedures are increasingly popular because they can be tailored to individual problems, and because they are known to control the FWE under all possible partial null configurations. Fisher's protected LSD is a closed testing procedure only in the case of all pairwise comparisons

with three groups, and fails in more general cases. It is likely that the popularity of the method will fade, as people become more aware of its problems.

The following diagram illustrates the closed testing hierarchy.

Closed Testing for Pairwise Comparisons:
4 Groups



Using the closed testing procedure, in order to declare, say μ_1 different from μ_2, one would have to have (a) a significant (unadjusted) comparison, and (b) significant (unadjusted) tests for all composite hypotheses that include the $\mu_1=\mu_2$ hypothesis. Use of any valid test (e.g., F test, range test) for the intermediate hypotheses will result in a method that controls FWE. The Fisher protected LSD method leaves out the entire middle layer and therefore does not control FWE.

The closed testing paradigm leads to stepwise procedures that have more power than the simultaneous confidence interval methods for detecting differences. On the other hand, there are usually no simple confidence interval equivalences for closed testing methods. To gain power for detecting differences, one must sacrifice the specificity of the interval-based inferences.

5.3. MCPs in more Complex "Farm" Applications

Recent developments in computing capabilities, software and closed testing methods have greatly expanded the scope of applications for which powerful MCPs are readily available. The SAS/STAT software PROC MIXED allows one to fit a variety of models in split plot, repeated measures, and spatial designs with complex error structures, often involving multiple error terms. Littell et al. (1996) give numerous examples.

We have become fairly comfortable with standard, approximate univariate analyses for such models, and this is the first step towards deciding what to do in the multiple inferences case. Suppose one has a set of estimated parameters $\left\{\hat{\theta}\right\}_i$, i=1,…,k, obtained, say using the ESTIMATE statement of PROC MIXED. To obtain simultaneous confidence intervals, one may (a) use the standard intervals (perhaps using Satterthwaite approximate dfs), but with Bonferroni levels α/k. To improve upon Bonferroni, one typically needs to involve the correlation matrix of the parameters, which can be obtained by outputting the estimated covariance matrix and

converting it to a correlation matrix. In that case the vector of pivotals $\left\{(\hat{\theta}_i - \theta_i)/s.e(\hat{\theta}_i)\right\}$ can be considered approximately multivariate t-distributed, with the estimated correlation matrix, and where the df may be obtained via Satterthwaite approximation. The appropriate critical value for the simultaneous confidence intervals then can be obtained (at least by simulation) as the (1-α) quantile of the distribution of $\max_i \left\{|(\hat{\theta}_i - \theta_i)/s.e(\hat{\theta}_i)|\right\}$. This approach is discussed further by Westfall et al. (1999), who provide a macro "%SimIntervals" to automate the analysis.

If more power is desired, the given estimated multivariate t distribution can be used in conjunction with closure-like ideas. Shaffer (1986) describes a closure-related step-down method that utilizes logical constraints among the specific contrasts of interest, and Westfall (1997) carried the analysis further by simulating critical values from the multivariate t distribution, rather than using Bonferoni's method. This approach is also discussed in Westfall et al. (1999), who provide a macro "%SimTests" to automate the analysis.

6. THE INFLUENCE OF THE PHARMACEUTICAL INDUSTRY: THE ICH GUIDELINES

Recent developments in multiple comparisons have been largely spurred by people working in the pharmaceutical industry. This industry is both highly competitive and regulated. Competition pressure encourages the pharmaceutical companies to present their data in the best light possible. Why should a company pre-specify whether to compare absolute outcomes, or change in absolute from baseline, or percentage change from baseline, or baseline covariate-adjusted outcome? Why not consider all four tests and pick the one with the smallest p-value? Putting the company in the straightjacket of having to pre-specify which of the four measures to use puts them at a competitive disadvantage; after all, a competitor who is developing a similar drug certainly will put its analysis in a similar “best light,” gaining stronger claims, and perhaps greater profits.

Regulatory agencies are aware that profit motives sometimes can hinder scientific objectivity, and have long encouraged the use of multiple comparisons concepts for the analysis of clinical trials data. Their concerns were recently formalized with the publication of the "ICH guidelines" (*Guidelines 11.E9 Statistical Principles for Clinical Trials: Availability International Conference on Harmonisation of Clinical Trials* http://www.fda.gov:80/cder/guidance/91698.pdf). With regard to multiple comparisons, these guidelines state the following:

> From Section 5.6, Adjustment of Significance and Confidence Levels:
> When **multiplicity** is present, the usual frequentist approach to the analysis of clinical trial data may necessitate an adjustment to the Type I error. Multiplicity may arise, for example, from multiple primary variables…, multiple comparisons of treatments, repeated evaluation over time, and/or interim analyses…details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.
>
> From Section 2.2.2, Primary and Secondary Variables:
> To avoid **multiplicity concerns** arising from post hoc definitions, it is critical to specify in the protocol the precise definition of the primary variable as it will be used in the statistical analysis.

From Section 2.2.3, Composite Variables:
[Using composite scores] addresses the **multiplicity problem** without requiring adjustment to the Type I error. The method of combining the multiple measurements should be specified in the protocol…

From Section 2.2.5, Multiple Primary Variables:
The effect [of using multiple primary variables] on the Type I error should be explained because of the potential for **multiplicity problems**…

From Section 5.7 Subgroups, Interactions, and Covariates:
Any conclusion of treatment efficacy (or lack thereof) or safety based solely on **exploratory** subgroup analyses is unlikely to be accepted.

From Section 7.2.2. Safety data:
The evaluation of the reality of these potential adverse effects should take into account the issue of **multiplicity** arising from the numerous comparisons made. …

Since the publication of these guidelines, the various drug companies and clinical research organizations with which I have consulted are taking multiple comparisons much more seriously. They are pre-specifying endpoints, and they are putting particular multiple comparisons and multiplicity management methodologies directly into their protocols.

7. MULTIPLE COMPARISONS IN THE 21ST CENTURY

The various genome projects (human, animal and plant) that are underway are likely to be the primary driving force for research in MCPs, at least for the immediately foreseeable future. Another current trend is "data mining," which has gone from being a "dirty word" to a word that makes some business executives think "competitive advantage." And, while we must be especially cautious in the interpretation of "significant" effects culled from such large studies, we must also allow ourselves the flexibility to hunt for effects in the large bodies of data that are increasingly available.

Genetics is an extremely important application area where large data sets and data mining issues arise. We will soon have data sets with thousands of people, with genotype data on thousands of genes for each individual. There is growing awareness that the more complex diseases involve interactions among multiple genes; increasing the dimensionality enormously. If only 1,000 genes are typed for an individual, then there will be 499,500 gene pairs to consider via two-way interactions. And this assumes that genotypes are binary; in reality genotypes are at least three-level (aa, Aa, AA), with many more levels for the common case where there are multiple alleles $A_1,A_2,\ldots,A_k$ at a locus, rather than simple "a" or "A." In the more complex case with k alleles, there are $k(k+1)/2$ possible genotypes per gene. Considering the number of possible genotypes per gene, as well as possible interactions between different genes, the multiplicity problem becomes enormous.

7.1 What kind of error control is needed?

In such studies where the number of tests might easily run into the millions, strict control of error rates becomes less desired. In all probability, such analyses are exploratory, and further study will be needed to confirm the results. However, even in such cases, *some* control over false significances is desired: analysts would like to have some assurance that "most" of the leads that they chase down will not be dead ends.

The False Discovery Rate (FDR) criterion popularized by Benjamini and Hochberg (1995) provides such control. Supposing there are k null hypotheses tested, let R=number of hypotheses rejected, and let V =the (unknown) number of erroneously rejected ones. Define V/R =0 in case R=V=0. Then FDR is defined as FDR = E(V/R).

The idea here is to choose a method with FDR$\leq\alpha$, so that, among those hypotheses rejected, the number of "false leads" is expected to be low. In the genetics screening testing, this makes sense: if one finds1,000 significant associations, one can allow that 50 or so might be false leads. FWE control is more stringent. One would find many fewer associations, say 100; but on the other hand one would be able to claim with confidence that all 100 are real and repeatable. The FDR method allows that more significances are mistakes, with the counterbalancing benefit that one will find more true effects with and FDR-controlling method than with an FWE-controlling method.

While FDR is quite reasonable for use in large, exploratory studies, FWE control is likely to remain the norm for smaller, confirmatory studies, as is the case with definitive clinical trials. The following conversations between statistician and pharmaceutical client illustrates some of the difficulties with FDR use in such cases.

Conversation 1: About False Discovery Rate (FDR).

Statistician: If you control FDR, then you may assume most of your significant results are real.
Client: Then how many are false?
Statistician: It's random. But the average percentage of claims that are false is 5%.
Client: Average percentage? Now I've heard it all. Why do you statisticians always have to be so convoluted? O.K., tell me what "average percentage" means.
Statistician: Well, you're right, it does take a little explaining. Say we perform the study, and find 7 significances. Unknown to us, one of these happens to be a mistake. The **percentage** of false claims in that study is then 1 in 7, or 14.2%. Now, suppose we had performed another study, under identical situations, but with a different patient randomization, and we found 5 significances. Unknown to us, there were no mistakes. Then the **percentage** of false claims in that study is 0 out of 5, or 0%. And in a third study, say we got 1 mistake out of 30 significances, or 3.3%. The **average percentage** for these three is now (14.2% + 0% + 3.3%)/3, or 5.8%. FDR is the average percentage over all possible patient randomizations, and if you control FDR, then that average percentage will be less than 5%.
Client: What about the studies where there are no significances found at all? In those studies, you have a percentage of 0/0? How are those counted in the average?
Statistician: Well, that is a fuzzy point. To make things mathematically tractable, we count those percentages as zeros.
Client: There you go again. Well, I won't worry about what "tractable" means, but let me ask you this. Since I will only make a claim about significance in those studies where I find a

significance, is it fair to include those 0's in the average? Because if you exclude them, then the average percentage will be higher.
Statistician: Well, yes, that's true.
Client: So that brings me back to the original question: In a study where I find some significant results, how many are mistakes?
Statistician: It depends.

Conversation 2: About Familywise Error Rate (FWE).
Statistician: If you control FWE, then you may assume that all of the significant results are real.
Client: O.K.

8. BAYESIANISM: THE 21ST CENTURY'S ANSWER TO THE MULTIPLICITY PROBLEM?

Perhaps another reason that multiplicity effects have not been sufficiently recognized is that some in the Bayesian statistical community have argued that the problem is not nearly so bad as the frequentists would claim. Bayesians have long held that the appropriate response to the problem is to adopt an appropriate prior distribution that effectively "shrinks" the most extreme observed effects toward the mean, thereby making them, in a sense, "less significant" (Lindley, 1990). While frequentist methods are similar in the sense that the significances of the most extreme effects also are downplayed, or "shrunk," the degree of shrinkage of the frequentist methods is orders of magnitude more extreme than that of the Bayesian methods using the "usual" priors.

For experiments where multiplicity adjustments are considered appropriate, the "usual" Bayesian priors are inappropriate. In these situations***, Bayesians have been using the wrong priors***, and have been therefore deceiving themselves.

It happens that there are a number of reasonably close similarities between the methods when the appropriate class of priors is used.

So, when are multiplicity adjustments considered appropriate? The usual argument for considering multiplicity adjustment is as follows: "What if all (or many) null hypotheses are true? In that case, the usual $p \leq .05$ decision rule will result in too many Type I errors."

The statement "What if all (or many) null hypotheses are true?" actually is a statement about prior plausibility of the collection of null hypotheses. And, as shown by Westfall et al. (1997), and Westfall et al. (1999, chapter 13), there are certain correspondences between frequentist multiplicity adjustment and Bayesian posterior probability value under the case where the event {all (or many) null hypotheses are true} is assigned a moderate prior probability.

The use of prior on null hypothesis is controversial, as some will argue that no effects are truly null. I essentially agree with this position; however, the appropriate Bayesian response is to use a mixture prior that sharply concentrates probability near zero. Berger and Delampady (1987) explain very clearly that the use of point masses provides a more convenient and reasonably approximate solution to the tests that result from such continuous mixtures. Thus, point mass priors are appropriate for hypothesis assessment, even when it is allowed that no effects are truly mathematically null.

Such priors can also be motivated easily when one considers safety endpoints in clinical and pre-clinical pharmaceutical trials. For *most* safety endpoints, there is a *relatively high* prior

probability of no treatment-related effect. There is usually *no a priori* reason to suspect that compound X is, for example, associated with malignant brain gliomas in Sprague-Dawley rats; however, this test must be performed as part of the carcinogenesis bioassay. Neither do we suspect (in general) that compound X causes athlete's foot, nevertheless, this effect must also be tested and reported, should such adverse incidents be observed. The typical Bayesian prior, which puts a normal distribution over the distribution of possible effect sizes, is simply *wrong* for these applications, because it does not particularly distinguish the zero point.

Thus, if Bayesians are to provide the answer for the multiplicity problem, then they must

- Use the correct priors (including point masses), and
- Evaluate robustness to choice of multivariate prior

We are seeing some use of point mass priors in high dimensional situations (Efron et al, 2000); however, robustness remains problematic. Even simple changes to prior assumptions can lead to drastically different posterior inferences (Gönen and Westfall, 1998).

While Bayesian methods provide a useful and coherent paradigm for examining and resolving the multiplicity problem, they are not a panacea. In a way, the problem is compounded on the Bayesian side: not only does one have to select the family of tests, one also has to select of multivariate prior for that family, perhaps allowing point masses and arbitrary correlation structures. Frequentist inferences may be non-robust to the specification of a "family" of inferences, but the same can be said for Bayesian inferences, in duplicate: Bayesian methods are sensitive to the specification of the "family" *and to* the choice of prior over that family.

9. FUTURE EDUCATION IN MCPS

Our most important task is to educate our clients about the nature of the multiplicity problem. Simple methods for multiplicity management may be fine. These may be as simple as recognizing the difference between confirmatory and exploratory studies, and teaching people how to write sentences like,"These results were derived from a purely exploratory analysis, where (describe extent of exploration here). Future studies are needed to confirm the existence of the associations that we have identified."

If some exploration is needed, and if the client wishes to state that the results are confirmatory, then formal multiplicity adjustment procedures are required. These methods must have been stated in the data analysis plan, prior to data collection, otherwise the analysis must be relegated back to exploratory status. If one analyzes data-driven hypotheses, then the analysis is exploratory (with notable exceptions for the Scheffé-type and greatest root-based methods.)

Let me relate one final anecdote regarding failure of education in regards to multiplicity, and I confess that this is my own failure, since the student had taken his statistics courses from me. In a study relating soil characteristics to cotton yield, he performed a regression analysis, but the overall model F was insignificant. Not pleased, he then divided the field into four yield categories, from lowest to highest, and performed a discriminant analysis. This analysis provided him with a univariate discriminant function involving the various soil measurements. He then proceeded to test for differences between the four yield grades using an ANOVA F-test on the discriminant function. Now he was much happier: the F test was highly significant!

The multiplicity aspect behind this anecdote is that the discriminant function is actually the result of infinite data snooping: it is the linear combination that is ***selected to maximize*** the

univariate F. Of course, we all recognize that there are other appropriate multivariate tests to use in this instance, specifically Roy's greatest root (or its equivalent). But the larger issue is this: our clients use such "selection" techniques all the time, sometimes automatically (as in the case of the cotton yield study), sometimes by hand (try it different ways, and pick the winner). Our task is to advise our clients that *a different distribution must be used whenever selection is performe*d, whether or not we recognize the particular form (e.g., Roy's greatest root) that the distribution must take. We are guilty of abiding by a double standard if we require use of Roy's greatest root in one case involving selection, but wink and say "no adjustment necessary" in other cases involving selection.

10. CONCLUSION

It is up to every one of us to come to grips with the multiplicity issue, as it is fundamental. The subject is indeed controversial, but if we shy away from it for that reason, we are doing a disservice to our profession. We need to train users in this area so that they understand its nuances. Ultimately, coming to a personal recognition and reconciliation of the multiplicity issue will make us all better statisticians.

REFERENCES

Bailar, J.C. (1991). Scientific inferences and environmental health problems. *Chance* **4**, 27-38.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* 57, 1289-1300.

Berger, J. O. and Delampady, M. (1987), Testing precise hypothesis. *Statistical Science* 2, 317-352.

Cook, R.J. and Farewell, V.T.(1996). Multiplicity considerations in the design and analysis of clinical trials. *JRSS-A* **159**, 93-110.

Efron, B., Tibshirani, R. Goss, V., and Chu, G. (2000). Microarrays and their use in a comparative experiment. Manuscript.

Ernhart, C.B. Landa, B. and Schnell, N.B. (1981). Subclinical levels of lead and development deficit - A multivariate follow-up reassessment. *Pediatrics* 67, 911-919.

Fears, T. R., Tarone, R. E., and Chu, K. C. (1977), False positive and false negative rates for carcinogenicity screens," *Cancer Research*, 37, 1941-1945.

Fleming, T.R.(1992). Current issues in clinical trials. *Statistical Science* 7, 428-456.

Gönen, M., and Westfall,P.H.(1998), Bayesian multiple testing of multiple endpoints in clinical trials, *Proceedings of the American Statistical Association, Biopharmaceutical Subsection*, 108-113.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures*. John Wiley, New York.

Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*, Chapman and Hall, London.

King, R.T.(1995). The tale of a dream, a drug, and data dredging. *The Wall Street Journal* Feb. 7, 1995.

Lindley, D. V. (1990), The 1988 Wald Memorial Lectures: The present position in Bayesian statistics. *Stat. Sci.* 5, 44-65.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996). *SAS® System for Mixed Models*, Cary, NC: SAS Institute Inc.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63, 655-660.

Miller,R.G.(1981). *Simultaneous Statistical Inference, 2nd Ed.* Springer-Verlag, New York.

Needleman, H.L., Gunnoe, C.,Leviton, A., Reed, R., Peresie, H., Maher,C. and Barrett, P. (1979). Deficits in psychologic and classroom performance of children with elevated dentine lead levels. *The New England Journal of Medicine* 300, 689-695.

Palca, J. (1991). Get-the-lead-out guru challenged. *Science* 253, 842 -844.

Rothman, K.J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43-46.

Saville,D.J. (1990). Multiple comparison procedures: the practical solution. *The American Statistician* **44**, 174-180.

Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* **81**, 826—831.

Westfall, P.H. (1983). On the asymptotic normality of the Henderson method III estimates of variance components in the mixed linear model. Unpublished Ph.D. dissertation, University of California at Davis.

Westfall,P.H.(1985). Simultaneous small-sample multivariate Bernoulli confidence intervals. *Biometrics* **41**, 1001-1013.

Westfall, P.H. (1987). Multivariate binomial testing. Keynote address at the 10th Midwestern Biopharmaceutical Statistics Conference, 5/87, Muncie, Indiana.

Westfall, P.H., Lin, Y., and Young, S.S. (1989),A procedure for the analysis of multivariate binomial data with adjustments for multiplicity, *Proceedings of the 14th Annual SAS® User's Group International Conference*, 1385-1392.

Westfall, P.H, Lin, Y. and Young, S.S. (1990), Resampling-based multiple testing, *Proceedings of the 15th Annual SAS® User's Group International Conference*, 1359-1364.

Westfall, P.H., and Young, S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, New York.

Westfall, P.H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association* **92**, 299—306.

Westfall, P.H., Johnson, W.O., and Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84, 419-427.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS® System*, Cary, NC: SAS Institute Inc.

Westfall, P.H., and Soper, K.A. (1998), Weighted multiplicity adjustments for animal carcinogenicity tests. *Journal of Biopharmaceutical Statistics* 8, 23-44.

Westfall, P.H., Zaykin, D.V., and Young, S.S. (2001). Multiple tests for genetic effects in association studies. To appear in *Statistical Methods in Microbiology*, Stephen Looney, Ed.