P. K. Sen and C. R. Rao, eds., *Handbook of Statistics, Vol. 18*

On Some Multiplicity Problems and Multiple Comparison Procedures in Biostatistics

Yosef Hochberg and Peter H. Westfall

Some expositions of the field of Biostatistics are reviewed and different perceptions of the multiplicity problem are indicated. We show the seriousness and importance of the problem in various Biostatistical areas through discussions of specific examples, along with some suggested analyses using real data. Frequentist vs. Bayesian approaches, Per-Comparison vs. Familywise vs. False Discovery Rate approaches are contrasted with reference to specific problems. Main exposure is given to the frequentist approach of controlling the Familywise Error (FWE)-rate. In addition to a special tutorial on classical procedures for FWE control we discuss topics such as the closure principle, stepwise testing, incorporation of correlations, discrete data considerations, and resampling methods all in Biostatistical contexts. Specific applications include multiple endpoints, animal carcinogenicity, adverse events, epidemiological risk assessment, bioequivalence assessment, and subgroup analyses. This article contends that multiplicity effects are real and require analytic solutions. Details of implementing such solutions, including the controversial definition of a "family" of tests, are given in case-specific settings.

1. Introduction

Greenberg (1982) indicates multiple comparison problems only in the sub-field of clinical trials but that is probably due to their prominence in that sub-field (see e.g. Simon, 1994). In this article multiplicity problems and multiple comparison procedures in clinical trials and in other sub-fields of Biostatistics are discussed.

Greenberg (1982) and Finney (1993) indicate a broader scope for "Biostatistics" than the often restricted emphasis on medical applications. Nevertheless, in their discussion of important areas they *highlight medical applications and we will follow them in this regard.* "Medical applications" are obviously not restricted to clinical trials, e.g. Greenberg (1982) discusses "Detection of hazardous substances." We discuss multiplicity problems and procedures in Animal Carcinogenicity experiments in Section 4.2.

Greenberg indicates problems of *multiple examinations* of data, of *multiple comparisons, and the need to adjust for ex post facto dredging of data.* Finney on the other hand indicates lack of encounters with multiplicity problems requiring adjustments such as discussed by Greenberg.

Breslow (1990) indicates the richness of biostatistical subject areas with problems of multiplicity. His conception of the multiplicity problems in the various sub-fields of Biostatistics follows an empirical-Bayes paradigm. He is not concerned about selection effects associated with data dredging in such problems (see Peter Armitage's discussion of Breslow's paper). Breslow's "problems of multiplicity" are not those considered in this paper. However, we do discuss in the sequel other Bayesian and quasi-Bayesian conceptions of multiplicity problems.

Finney's and Breslow's apparent lack of enthusiasm for classical multiple comparisons procedures (MCPs) is shared by others in the biostatistical community and is often expressed by the question, "Are multiple testing/comparison procedures needed?" The answer, in our minds, is definitively "Yes". The biological issue surrounding the need for MCPs is the question of replicability of statistical associations. Newspapers and journals are rife with claimed associations that are suspect at best, and often do not hold up under scrutiny. Examples culled from recent periodicals include the following claimed associations: cellular phones with brain tumors, power lines with leukemia (more recently overturned by the scientific community), vitamins with IQ, season of the year with mental performance (but only in men!), genetics with homosexuality (the "gay gene"), abortions with breast cancer (but not spontaneous abortions), remarriage with cancer, electric razors with cancer, and on and on, with further examples given later in this article. Many of these claims have shaky foundations a priori, and some have been found not to replicate in further studies. With so much conflicting information in the popular press, the general public tends to mistrust the results of biostatistical studies.

Why do apparently "significant" biological associations not replicate? Scientists commonly blame faulty experimentation, study apparatus, patient population, and the like, but we would like to emphasize that multiplicity is as likely a source of such faulty conclusions. 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.

Multiplicity is pervasive in biostatistical applications. Rarely does a biological study hinge on one and only one test. Multiple measurements are available, and all are analyzed. This is as it should be – we 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. In this article we discuss several methods for analyzing data from such studies, and place them in the context of some common biostatistical applications.

To enhance the achievement of the editors' multiple purposes (*broadness of coverage, state-of-the-art main results, and explanation of use in practical applications*) we discuss a few applications requiring subject matter knowledge, and other applications are addressed by literature review (with discussion). We chose topics and mode of presentation so as to add most efficiently to the following literature; Bauer (1991), Simon (1994), Cook and Farewell (1996), Tamhane (1996). While the analyses of biostatistical data we present are sound from a variety of practical and theoretical standpoints, there can be no single analysis that is best from all perspectives. As Louis (1992) writes, "Statistical philosophies, principles and methods (frequentist/Bayes, multiple comparisons, choice of tests and estimators) need to guide deliberations, but in the complex world of clinical trials absolute dictums are seldom appropriate," a sentiment we endorse wholeheartedly.

Hochberg and Tamhane (1987) (hereafter abbreviated as HT) indicated a philosophy of using different approaches in different applications. Indeed their book contains discussions of a great variety of approaches (e.g. decision theoretic, graphical methods, clustering methods, and more). Yet, *most of their book is based on the traditional frequentist approach.* The obvious reason is the reflection of the existing literature at the time. Presently we have a new frequentist approach (discussed in Section 5) and some new Bayesian and quasi-Bayesian procedures. While we encourage a plurality of approaches to multiplicity adjustment, in this article we primarily represent the traditional approach (for the same simple reason indicated above for HT as well as for other reasons elaborated in the sequel).

In Section 2, we discuss three basic approaches to multiplicity problems and explain our decision to focus on the traditional frequentist approach. In Section 3 we give a compact tutorial on classical MCPs necessary for reading the following sections. Section 4 is devoted to "special topics" including "multiple endpoints," "animal carcinogenicity studies," and "subgroup analyses." Some discussion of type-II error characteristics associated with traditional MCPs is also given in Section 4. In Section 5 we discuss "other problems and approaches." These include problems in which the suitable MCPs call for control of some type-II error rates. A general class of suitable procedures for such problems known as Intersection-Union (IU) procedures is discussed with reference to combination drugs and assessment of bioequivalence. Additional problems discussed in Section 5 include Meta-Analysis and Publication Bias. The main "other approach" discussed in Section 5 is the recent frequentist alternative to FWE-control calling for control of the False Discovery Rate (FDR). We close with some final comments in Section 5.3.

2. On different approaches to multiplicity problems

2.1. The Problem of multiple comparisons and the classical approach

2.1.1. The Problem of multiple comparisons

Tukey (1953) used that title for the problem of *excess type-I error* associated with *significant inferences* in *a family of* per-comparison level tests. The notion of a *family* is basic in the classical approach and yet it is frequently ambiguous how the family should be defined. This is a crucial issue, since inferences are extremely sensitive to how the family is defined; specifically how many tests are included in the family. For cases where families are infinite or undefinable (*a priori*), it can be difficult or impossible to develop appropriate adjustment procedures.

Shaffer (1994, 1995) cites Stigler's (1986) reference to Cournot's (1843) example of investigating the chance of a male birth. "One could distinguish ... legitimate births from those occurring out of wedlock ... one can ... classify births according to birth order,..., age, profession, wealth, or religion of parents...". He goes on to point out that as one increases the number of such opposing categories, it becomes more and more likely that by pure chance at least one observed deviation will be significant and that "*usually these attempts through which the experimenter passed don't leave any traces; the public will only know the result that has been found worth pointing out...*"

Shaffer indicates that Cournot considered the problem as insoluble and some writers still think so (e.g. Nowak, 1994). However she writes: "While these issues are still serious and far from being solved ... multiple comparison methods provide a means for approaching such problems. It is vital to obtain solutions – in medicine ... where the effects of experimental treatments or environmental events may vary over subpopulations ... (see Shafer and Olkin, 1983 for work closely related to this issue)." We discuss the *multiple subgroup problem* in Section 4.3.

We consider the exploratory aspect of research an essential part which often should not be controlled by formal MCPs, which are more suitable in confirmatory stages. However, in many realistic problems the multiplicity issues are sufficiently revealed, and suitable MCPs can be used to combat selection effects. See also Diaconis (1985) for some intermediate problems. Our approach calls for context-related and problem-specific solutions as we hope to demonstrate in the sequel.

2.1.2. Basic notions of the classical approach

The "classical approach" calls for control of the probability of at least one erroneous rejection over the *family of all potential comparisons*. This is generally known (since Miller, 1966) as the Familywise-Error (FWE) rate. As explained in Putter (1983), by controlling that probability for the *family* of all potential inferences (e.g. all contrasts or all pairwise homogeneity hypotheses in a one-way layout) one controls it (conservatively) over *any subset of selected inferences* restricted to the family of interest.

The problem is that the concept of "potential inferences" cannot always be defined operationally. Fisher (1935) has characterized them as those inferences that "would have been made from the start equally plausible." In a one-way layout this distinction might lead to consideration of all contrasts or of only the pairwise comparisons. But in clinical trials and other biostatistical applications the problem is more complex. A more operational definition for a single family is perhaps Tamhane's (1996): "a set of *contextually related inferences* (*comparisons*) *from which some common conclusions are drawn or decisions are made*". An example of "contextually related" can be given by the multiple-endpoints example (considered in Section 4.1 here) which relates to a *single* aspect of the comparisons between new and old (e.g. efficacy).

2.1.3. Weak vs. strong control of the FWE

Fisher (1935) offered two procedures for testing all the pairwise homogeneity hypotheses in a one-way layout with k treatments. His first procedure involves two-steps. First an α-level F test is performed and if not significant, the procedure terminates (accepting H_0). Otherwise each pairwise homogeneity hypothesis is tested by an α-level (two-sided) t-test. This procedure is known (since Tukey, 1953) as Fisher's Least Significant Difference (LSD). The LSD controls the FWE at level α under H_0 (because of the first step). However, at other configurations of the true means, the FWE can be well in excess of α. We say that *the LSD procedure controls the FWE-rate only weakly*. Apparently Fisher (1935) thought that weak FWE-control might be suitable in some situations. This view is nicely expressed by Carmer and Walker (1982) and further discussed in our Section 5.

Fisher's second procedure, popularly known as the *Bonferroni procedure*, involves testing each pairwise homogeneity hypothesis at level $\alpha/\binom{k}{2}$ (without a preliminary F-test). This is *a single-step procedure* which provides conservative protection for the FWE since the Bonferroni method in general controls the Per Family Error (PFE) rate = the expected number of errors which is always greater than the FWE (see e.g. HT). An MCP which controls the FWE under all configurations of the parameters is said to control it in the strong sense.

2.2. Bayes vs. frequentist notions

Breslow (1990) indicates that the use of Bayesian methods can be controversial in some problems due to uncertainty regarding specification of informative priors. A similar indication can be found in other places, see for instance Draper et al. (1992), p. 156 on "Research opportunities."

Simon (1994) indicated that multiplicity issues are of importance both within the frequentist and Bayesian frameworks of inference. Within the Neyman–Pearson theory, a family of hypotheses must be pre-specified. With Bayesian methods one needs, in addition, to specify the prior distributions. He expresses the need to consider frequentist measures of operating characteristics for Bayesian procedures in some problems. Regarding interim analyses of clinical trials he notes that whereas the stopping rule does not play a direct role in Bayesian analysis, the frequency properties of such approaches are still relevant.

He notes that: "It has been widely accepted that a specified type I error ... should apply to a given hypothesis even if several interim analyses are performed and hence the Neyman–Pearson theory has been generally successful for interim monitoring." With respect to the Bayesian methods he concludes: "Bayesian methods are more acceptable when they lead to experiments in which the data dominates the priors or when highly informative priors can be avoided. This is because the consumers of the results are not just those performing the experiments."

One difficulty in specifying a convincing prior is the decision of whether to use point masses on null hypotheses (e.g., Berger and Sellke, 1987). In animal carcinogenesis (see Section 4.2 below), for example, it is plausible that a compound in question either has no effect (or very little effect) on a particular tumor type, or that it has a "large" effect. This suggests that a mixture prior be used, which puts positive probability on zero effect, and spreads the remaining effect continuously in both directions from 0, perhaps with positive mean to suggest that if an effect occurs, then it is more likely to be positive (increasing carcinogenicity). In this setting, there have been attempts to reconcile Bayesian measures of evidence, specifically, the calculation of P (H_0 is true|the data) with frequentist p-values. These numbers do not agree for "objective" prior distributions (Berger and Sellke, 1987), but do correspond in the sense that both become smaller as the data better support the alternative hypothesis.

Bayesians must calibrate their priors as joint probabilities on the family of hypotheses. Posterior inferences can vary dramatically, depending upon whether the primary calibration is done at the familywise or componentwise testing level. As shown by Westfall et al. (1997), if the truth of all hypotheses within the family is considered moderately probable, then the Bayesian measures of evidence correspond roughly to the ordinary Bonferroni correction in frequentist methods. However, if the truth of individual propositions within the family only are considered moderately probable (implying joint truth of all propositions within the family is very unlikely, under independence), then the Bayesian measure of evidence corresponds to unadjusted frequentist p-values, again in the loose sense described by Berger and Sellke. In Section 5 we also discuss a connection established by Shaffer (1997) between Duncan's Bayesian approach and the new FDR controlling procedure in a one-way layout.

Thus, while we do not cover the Bayesian paradigm in detail in this article, it can be said that frequentist multiple methods are loosely justified from a Bayesian standpoint when it is considered moderately probable that all hypotheses tested are in fact true. We close this section, returning to John Tukey (1977), a frequentist commenting on the Bayesian paradigm for analyzing multiple testing data in clinical trials:

> I have yet to see a Bayesian account in which there is explicit recognition that the numbers we are looking at are the most favorable of k. Until I do, I doubt that I will accept a Bayesian notion of this sort as satisfactory.

2.3. To adjust or not to adjust?

2.3.1. Different views

Tukey (1953) introduced three basic error-rates namely the "experimentwise", the "per experiment" and the "per statement" (today known as Familywise, per Family and comparisonwise) but he and other prominent writers on the problem of multiple comparisons (e.g. Miller, 1966) were of the opinion (see e.g. HT, Ch.1) that there was a single error-rate generally suitable for control in multiple comparison problems. They indicated the (type-I) FWE-error, as their prime choice. Opposing attitudes including a per-comparison approach, i.e. no need for adjustments of the p-values in view of multiplicity, were expressed by discussants following O'Neill and Wetheril (1971). Some statisticians and practitioners remain unconvinced that multiplicity adjustment is ever necessary.

As an alternative to multiplicity adjustment in environmental health statistics, Bailar (1991) suggests that the analyst "State, with maximum clarity and precision, which hypotheses were developed entirely independently of the data and those which were not, so readers will know how to interpret the results." In a similar vein, Saville (1990) and Cook and Farewell (1996) argue that control of per-comparison error rates is more relevant than control of FWE. Rothman (1990) takes a more dramatic stand as reflected by the title of his article, "No adjustments are needed for multiple comparisons." A common theme of the various criticisms is the difficulty of defining a specific family of tests. For example, in a clinical trial, should all efficacy and safety measures be lumped into one family? Should there be separate families for "primary" and "secondary" endpoints (discussed in the sequel)? Should one consider the Phase II and Phase III trial results together as a family, or should they be separate?

The difficulties in defining family, coupled with the extreme sensitivity of the inferences to this definition, have caused many researchers to avoid using MCPs. Our view is that analysts need to be aware of the various error rates to properly interpret their data. Our recommendation is to choose smaller, more focused families rather than broad ones. Gabriel (1969) substantiates this recommendation. The Bayesian paradigm is very useful to guide analysis of multiple tests, even if used only informally, as discussed in Section 2.2.

2.3.2. Case study: Epidemiology, data snooping

The line of research suggesting no multiplicity adjustment can be dangerous. Carried to its extreme, statistical practitioners might conclude "anything goes". An example where a controversial epidemiological finding was claimed, which was likely a Type I error, was published by Needleman et al. (1979), who 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 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.

The lead/IQ example was particularly insidious from a Type I error control standpoint, in that the family was virtually infinite: various covariates and subgroup analyses were performed in an effort to find statistical significance. Such analysis raises the specter of "all subsets" analysis in the regression context, in which it is very easy to find statistical significance, not only from the multiplicity problem, but also because of the fact that the experimental error is biased downward in the selected model (Copas and Long, 1991). In fact, it was only after such analyses 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."

This example, and others that are given in following section, demonstrate that multiplicity effects are real, that they need to be considered, and that there are reasonably powerful methods for controlling the error rates. An important point to be made, however, is that in order for the FWE control to be obtained, the specific family *and* the specific testing method (e.g., covariate-adjusted or not, etc.) must be specified in an experimental or study protocol, and followed precisely. Deviation from such protocol may easily inflate the FWE and result in excess Type I errors.

In this article, our arguments generally favor the position "To Adjust," although not exclusively, with alternatives given in final section.

3. A tutorial on classical MCPs

3.1. Preliminaries

Gabriel (1969) discussed some general theory for Simultaneous Test Procedures (STPs) which constitute a sub-class of what HT called single-step procedures. These procedures use a *single* critical value in contrast to *stepwise MCPs* which use several critical values. He distinguished between *hierarchical* and *non-hierarchical* families. An hierarchical family contains at least one pair of hypotheses such that one implies the other. Fisher's Bonferroni addresses the non-hierarchical family $\{H_{0ij}, 1 \leq i < j \leq k\}$ of pairwise homogeneity hypotheses. In contrast, Fisher's LSD addresses the hierarchical family $\{H_0, H_{0ij}, 1 \leq i < j \leq k\}$ where H_0 is the overall null hypothesis which implies any H_{0ij}. A *minimal hypothesis* does not imply any other hypothesis, e.g. every H_{0ij} is minimal. An MCP is called *coherent* if for any pair of hypotheses H_1, H_2 s.t. H_1 implies H_2, acceptance of H_1 implies acceptance of H_2. An MCP is called *consonant* if whenever a non-minimal H_1 is rejected, at least one of its implied minimal hypotheses is rejected. To learn more about single step vs. stepwise MCPs we now focus on a single important problem.

3.2. The many-to-one problem

This is a well known title for the problem of comparing several new treatments with a single standard (a control treatment or placebo, etc.). We will assume here that the goal is to select the treatments that are "better" than the control using one-sided tests.

Setup for many-to-one comparisons

Assume a one-way layout with $k-1$ new treatments and a single control indexed as the kth treatment. Let Y_{ij} be the jth observation on the ith treatment and assume the linear model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad (1 \le i \le k, 1 \le j \le n_i) \ . \tag{3.1}$$

The ϵ_{ij} are assumed to be i.i.d. $N(0, \sigma^2)$. Then the estimates of the μ_i and σ^2 are $\overline{Y}_i$ and s^2, respectively (the latter is based on $v = \Sigma\, n_i - k$ d.f.). The $\overline{Y}_i$ are independent $N(\mu_i, \sigma^2/n_i)$, $i = 1, \ldots, k$.

The *parametric functions* of interest here are the $k-1$ pairwise differences $\mu_i - \mu_k$ whose unbiased estimates $\overline{Y}_i - \overline{Y}_k, i = 1, \ldots, k-1$ follow a $(k-1)$-variate normal distribution with variances $\sigma^2(1/n_i + 1/n_k)$ and correlations $\rho_{ij} = [n_i n_j/(n_i + n_k)(n_j + n_k)]^{1/2}$ $(1 \le i < j \le k-1)$ which simplifies to a common constant ρ when all treatments have the same sample size n, $\rho = n/(n + n_k)$.

3.2.1. A single-step one-sided procedure

Dunnett (1955) introduced the many-to-one problem and the corresponding single-step procedures.

Denote the (minimal) null and alternative hypotheses of interest as $H_{0i} : \mu_i - \mu_k \le 0$ and $H_{1i} : \mu_i - \mu_k > 0, i = 1, \ldots, k-1$. The Likelihood-Ratio (LR) test of H_{0i} rejects when

$$T_i = \frac{\overline{Y}_i - \overline{Y}_0}{s(1/n_i + 1/n_k)^{1/2}} > \xi_i \quad (1 \le i \le k-1) \ . \tag{3.2}$$

There are some good reasons for using $\xi_1 = \cdots = \xi_k = \xi$ (see HT). Accordingly, Dunnett's (1955) problem was to find the common ξ which will guarantee (strong) FWE rate control at level α. Dunnett provided the necessary solution as $\xi = T^{(\alpha)}_{k-1,v,\rho}$ – the $(1-\alpha)$-th quantile of the maximum of the $k-1$ T_i's (which follow a *multivariate t-distribution* with parameters $k-1, v$, and ρ).

This particular procedure was later indicated (by Krishnaiah, 1979) as a special case of Roy and Bose's (1953) general method for constructing Simultaneous Confidence Procedures (SCP) derived with reference to Roy's general method for construction of tests in various multivariate or multiparameter problems, namely, the Union-Intersection (UI) method.

The UI method first expresses a *global* hypothesis as an *intersection* of minimal hypotheses of interest. In Dunnett's MCP the minimal hypotheses are the H_{0i} and the global hypothesis is $H_0 = \bigcap_{i=1}^{k-1} H_{0i}$, that is the overall homogeneity hypothesis $\mu_1 = \cdots = \mu_k$. The UI method postulates for testing H_0 the union of the individual rejection regions. In this case, with a common ξ, it implies rejecting H_0 when $\max_{i=1,\ldots,k-1}(T_i) > \xi$ which explains the above indicated choice for ξ.

Roy and Bose (1953) showed that such tests, when applied in general linear models actually provide SCPs. Aitchison (1964) indicated the use of general SCPs as Confidence Region Tests (CRTs). A CRT based on an SCP rejects any (null) hypothesis if its intersection with the $1-\alpha$ simultaneous confidence set or intervals is empty. For example, first express Dunnett's procedure as an SCP for the $k-1$ comparisons with the control:

$$\mu_i - \mu_k \geq \overline{Y}_i - \overline{Y}_k - s(1/n_i + 1/n_k)^{1/2}\, T^{(\alpha)}_{k-1,v,\rho} \ . \tag{3.3}$$

Next realize that rejecting H_{0i} when $T_i > T^{(\alpha)}_{k-1,v,\rho}$ is equivalent to rejecting it when the right side of (3.3) is positive (which is the CRT in this case).

With an SCP one can test any imaginable hypothesis (including post-hoc selections) by the CRT. The FWE-rate associated with any such family is always $\leq \alpha$.

3.2.2. A stepwise procedure

Naik (1975) gave a one-sided stepwise procedure for the many-to-one problem which is always more powerful than the single-step procedure. His procedure is based on ordering the test statistics T_i into $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(k-1)}$, letting $H_{(1)}, H_{(2)}, \ldots, H_{(k-1)}$ denote the corresponding hypotheses, and then using the following *step-down* scheme

(i) Reject $H_{(k-1)}$ if $T_{(k-1)} > T^{(\alpha)}_{k-1,v,\rho}$ (where ρ is as earlier – the common correlation in case of treatment-balanced designs) otherwise retain all hypotheses without further tests.

(ii) In general, reject $H_{(j)}$ if $T_{(i)} > T^{(\alpha)}_{i,v,\rho}$ for $i = k-1, k-2, \ldots, j$. If $H_{(j)}$ is not rejected, then retain also $H_{(j-1)}, \ldots, H_{(1)}$ without further tests. Since $T^{(\alpha)}_{i,v,\rho}$ is monotone increasing in i, it is readily seen that this procedure is more powerful than the single-step procedure discussed earlier.

Why does this stepwise procedure control the FWE-rate in the strong sense? This question can be answered by realizing that Naik's procedure is a shortcut of a general method for constructing powerful stepwise MCPs which we discuss next.

3.3. The closure method

A general method for constructing step-down test procedures was proposed by Marcus et al. (1976). This *method* is also referred to as the closure *procedure*. Let $\{H_i (1 \leq i \leq m)\}$ be a finite family of hypotheses. Form the *closure* of this family by taking all non-empty intersections $H_{\mathbf{P}} = \bigcap_{i \in \mathbf{P}} H_i$ for $\mathbf{P} \subseteq \{1, 2, \ldots, m\}$. If an α-level test of each hypothesis $H_{\mathbf{P}}$ is available then the closure procedure rejects any hypothesis $H_{\mathbf{P}}$ if and only if every $H_{\mathbf{Q}}$ is rejected by its associated α-level test for all $\mathbf{Q} \supseteq \mathbf{P}$. A theorem in Marcus et al. (1976) states that the closure procedure strongly controls the type-I FWE at level α.

The *closure method* extends inferences from a given family of hypotheses ($\mathscr{F}$) to its closure ($\overline{\mathscr{F}}$). In the many-to-one problem the family $\overline{\mathscr{F}}$ of all *subset intersection hypotheses* involves all subset homogeneity hypotheses including the control. It can be verified that Naik's (1975) procedure is a short cut of the closure procedure for the many-to-one problem when each intersection hypothesis is tested by a Dunnett type test.

3.4. UI related STPs, closure and shortcuts

Gabriel (1969) considered the problem of choosing suitable test statistics for a given hierarchical family assuming that for the minimal hypotheses we have obvious choices, e.g. the t-statistics for the one-sided hypotheses in our many-to-one example. He *defined* a *UI related statistic for any non-minimal hypothesis as the maximum of the* (equally distributed) *statistics corresponding to the minimal hypotheses implied by it*, and proved the following desirable properties of MCPs based on UI related statistics: (i) A necessary and sufficient condition for an STP to be both coherent and consonant is that its test statistics be UI related, and (ii) The most powerful STP for a given set of minimal hypotheses (nested in some hierarchical family) is based on UI related statistics. These two-properties explain the prevalence of such MCPs.

Using our example, if there is a control treatment (or placebo, or bench mark) with which we want to compare $k-1$ new treatments, and if these comparisons are of main interest, then Dunnett's MCP is the most suitable (single-step) procedure among all (single-step) MCPs providing for these $k-1$ pairwise comparisons. Note that an STP for all pairwise comparisons among the k treatments can be obtained by forming an hierarchical family from the original $k-1$ comparisons with the control ($\mu_i = \mu_k$ and $\mu_j = \mu_k$ imply $\mu_i = \mu_j$). For this larger class of minimal hypotheses Gabriel's result will indicate Tukey's T-method as optimal (See HT chap. 3).

With respect to step-wise procedures, UI related test statistics also play an important role. Assuming that there are no logical relations among hypotheses and using UI related test statistics in the framework of a closure, HT Section 4.2.1 give an equivalent shortcut based on the ordered values of the test statistics for the individual minimal hypotheses. Naik's (1975) procedure discussed above is a special case. More general results on closure and shortcuts are discussed in Grechanovsky and Hochberg (1998).

3.5. A sequentially rejective Bonferroni procedure

In various biostatistical applications (see e.g. our Section 4) the joint distribution of the test statistics is either not known or too complicated to obtain exact procedures. Typical problems of the first type involve different statistics, e.g. Chi-squares, Normal deviates, Fisher's exact p-values, etc. often considered jointly for a given family of interest. An example of a known but complicated joint distribution is provided by the many-to-one problem in case of an unbalanced design. In some such problems use of the Bonferroni method is not too conservative and hence has been considered as a 'non-parametric' solution.

A simple and powerful general procedure was provided by Holm's (1979) "sequentially rejective Bonferroni procedure." With this procedure one orders the individual m (say) p-values $P_{(1)} \geq P_{(2)} \geq \cdots \geq P_{(m)}$ and denotes the corresponding hypotheses by $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. First check if $P_{(m)} \geq \alpha/(m)$, in which case retain all the hypotheses without further testing. Otherwise reject $H_{(m)}$ and proceed to test $H_{(m-1)}$. If $P_{(m-1)} \geq \alpha/(m-1)$ retain $H_{(m-1)}, \ldots, H_{(1)}$ without further tests; otherwise reject $H_{(m-1)}$ and proceed to test $H_{(m-2)}$ and so on. This step-down Bonferroni procedure may be contrasted with a corresponding single-step Bonferroni procedure in which all the p-values are compared to the common value α/m. Clearly, the step-down procedure is more powerful. Holm's procedure was indicated by HT to be a shortcut of an equivalent closure which tests each intersection hypothesis by a corresponding Bonferroni test.

4. Special topics in biostatistics

In this section we discuss some examples in detail, for which we have real data and some subject matter knowledge. The methods for use in these studies are somewhat developed within the context of the application (rather than the more common statistical presentation mode: method first, then example). However, because the general methods and case settings have application beyond the context of the given setting, we cross-label the section headings by, where appropriate, example, method, and general case setting. Thus the first subsection is labeled **Multiple Endpoints** – **Resampling** – **Multivariate Correlations** to emphasize both the topic and the general statistical notions involved in the analysis.

We should also note that the statistical tools used for the analysis of the data sets in this section are not definitive, nor recognized as such by, e.g., regulatory agencies. Our aim is to put some real numbers "on the table," so to speak, to bring certain issues to the fore. We invite alternative methods to be used, with direct comparison to those given herein, for the purpose of broadening both the scope and understanding of multiplicity adjustment methods as they apply to biostatistical data.

4.1. Multiple endpoints – resampling – multivariate correlations

The term "multiple endpoints" is used most commonly in clinical trials, where the efficacy of a therapy is multidimensional. For example, a drug designed to lessen discomfort from (not necessarily cure) a particular disease can have several possible beneficial effects, depending upon the various symptoms of the disease. Further, the measurements of each individual symptom may be multivariate; e.g., as self-rated by the patient, and as rated by the physician. Each measure of efficacy is called an "endpoint."

Despite the apparent origin of the term in the clinical trials arena, the general ideas easily extend to other areas of biostatistics. Abstractly, the general setup is one where the observations per unit are multivariate, and where the interest is in comparing the multivariate measurements between two or more groups. But the focus here is different from that in conventional "Multivariate Analysis" as the following discussion will reveal.

In a clinical trial, it would be simple to declare a drug "efficacious" if the decision rule is "the drug is efficacious if efficacy can be demonstrated for at least one endpoint, tested at the $\alpha = 0.05$ level." Clearly, with such a decision procedure, an inefficacious drug may easily be declared effective. For this reason, the United States Food and Drug Administration (FDA), following the recommendations of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), has developed a set of recommendations for reducing the probability of excess Type I errors in studies with multiple endpoints (Dept. of Health and Human Services, hereafter abbreviated DHHS, 1998). One approach is to classify *a priori* endpoints as either "primary" or "secondary." As reported in the ICH recommendations, "the primary variable ('target' variable, primary endpoint) should be the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial. There should generally be only one primary variable." (DHHS, Section 2.2.2).

As an alternative to identifying a single primary variable, the ICH guidelines suggest defining a combined or summate-type variable in cases where there are multiple endpoints. In such a case the combined summate becomes the single primary endpoint, and there is no multiplicity problem. However, in many cases the scientists are not able to identify or construct a single variable, in which case we have multiple primary endpoints, as well as a multiplicity problem. Because the data are multivariate, the nature of the dependence structure necessarily affects multiplicity adjustment. As stated in the ICH guidelines, "The extent of intercorrelation among the proposed primary variables may be considered in evaluating the impact on Type I error." (DHHS, Section 2.2.2). If the variables are highly correlated, then the need for multiplicity adjustment is lessened. In particular, if all variables are perfectly correlated, then there is no need for multiplicity adjustment at all.

The ICH guidelines offer further advice for assessing robustness of the analysis. As stated, "it is important to evaluate the robustness of the results and primary conclusions of the trial. Robustness is a concept that refers to the sensitivity of the overall conclusions to various limitations of the data, assumptions, and analytic approaches to data analysis." (DHHS, Section 1.2). The following example is meant to illustrate particular methods for analyzing multiple endpoint data that are consistent with and address issues raised by the ICH guidelines.

An example involving four primary endpoints in a clinical trial

A study recently was conducted at a large drug company with four endpoints, labeled Y_1–Y_4 to preserve confidentiality, which were compared for treatment ($n_1 = 57$ patients) and control ($n_0 = 54$ patients). The data are summarized in Table 1. The raw data vectors (not shown) may be denoted $\mathbf{Y}_{gi}$, for $g = 0, 1$, and $i = 1, \ldots, n_g$. Note that each $\mathbf{Y}_{gi}$ is four dimensional, with $\mathbf{Y}_{gi} = (Y_{gi1}\ Y_{gi2}\ Y_{gi3}\ Y_{gi4})$.

Table 1
Summary statistics for a clinical trial with multiple endpoints

End point	Control			Treatment			Pooled statistics				
	$\bar{X}_0$	s_0	n_0	$\bar{X}_1$	s_1	n_1	Std dev	Correlation matrix			
								Y_1	Y_2	Y_3	Y_4
Y_1	3.22	1.46	54	2.54	1.34	57	1.40	1.00	0.38	0.64	0.70
Y_2	2.44	4.37	54	0.93	1.37	57	3.20		1.00	0.45	0.43
Y_3	2.78	1.67	54	2.40	1.37	57	1.52			1.00	0.63
Y_4	3.26	1.64	54	2.51	1.68	57	1.66				1.00

Assume that the means are $\mu_{gj}, j = 1, \ldots, 4$, and that the inference shall concern the four differences $\mu_{1j} - \mu_{0j}$.

Our goal is to adjust for the multiplicity problem, while considering effects of correlations and nonnormalities. According to the ICH guidelines, "Estimates of treatment effects should be accompanied by confidence intervals..." (DHHS, Section 5.5). Thus, as a first step in the analysis, we examine simultaneous confidence intervals for $\mu_{1j} - \mu_{0j}, j = 1, \ldots, 4$. The unadjusted intervals are $\bar{Y}_{1j} - \bar{Y}_{0j} \pm t_{\alpha/2} s_j (1/n_1 + 1/n_0)^{1/2}$, where s_j is the pooled standard deviation, and where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the t-distribution with $n_0 + n_1 - 2$ degrees of freedom. The simultaneous confidence level is less than $1 - \alpha$ for these intervals, however. Assuming independent multivariate normal response vectors with common covariance matrix Σ, exact simultaneous $1 - \alpha$ intervals are given by $\bar{Y}_{1j} - \bar{Y}_{0j} \pm Q_\alpha(\Sigma) s_j (1/n_1 + 1/n_0)^{1/2}$, where $Q_\alpha(\Sigma)$ is the $1 - \alpha$ quantile of the distribution of

$$M_0 = \max_{1 \le j \le 4} \left| \frac{(\bar{Y}_{1j} - \bar{Y}_{0j}) - (\mu_{1j} - \mu_{0j})}{s_j (1/n_1 + 1/n_0)^{1/2}} \right| .$$

The problem with these intervals is that the distribution of M_0 depends upon Σ (or more precisely, upon the correlation matrix derived from Σ), which is unknown. It is reasonable to substitute an estimate of Σ, obtaining $\hat{Q}_\alpha = Q_\alpha(\hat{\Sigma})$. While there are computationally more clever methods for doing so, the following parametric resampling algorithm is simple, and also illustrates the proper approach for resampling nonparametrically.

1. Create a pseudo-data set $\mathbf{Y}^*_{gi} = \epsilon^*_{gi}$, where the ϵ^*_{gi} are i.i.d. $N_4(0, \mathbf{S})$, and where $\mathbf{S}$ is the pooled sample covariance matrix from the original observations with diagonal elements s_j^2.

2. Compute

$$M^*_0 = \max_{1 \le j \le 4} \left| \frac{\bar{Y}^*_{1j} - \bar{Y}^*_{0j}}{s^*_j (1/n_1 + 1/n_0)^{1/2}} \right| .$$

3. Repeat 1. and 2. B times. Estimate $\hat{Q}_\alpha$ using the empirical $1 - \alpha$ quantile of the B values of M_0^*. (Note: the resulting critical value should properly be labeled $\hat{\hat{Q}}_\alpha$ to emphasize that it is a Monte Carlo estimate of an estimated critical value. To minimize Monte Carlo error, B should be as large as computing resources and time constraints allow. In this example, we choose $B = 10,000,000$.)

Robustness to nonnormality may be assessed by bootstrap resampling rather than parametric resampling. Assuming a location shift model $\mathbf{Y}_{gi} = \boldsymbol{\mu}_g + \epsilon_{gi}$, where the ϵ_{gi} are assumed i.i.d. with distribution function F, the appropriate critical value Q_α depends upon the unknown F, and not just on Σ as in the normal case. Again, we estimate $Q_\alpha = Q_\alpha(F)$ using $\hat{Q}_\alpha = Q_\alpha(\hat{F})$ as follows. Define the sample residuals $\hat{\epsilon}_{gi} = \mathbf{Y}_{gi} - \bar{\mathbf{Y}}_g$. Our estimate $\hat{F}$ is the empirical distribution function of the residuals. Now, sample the data $\mathbf{Y}_{gi}^* = \hat{\epsilon}_{gi}^*$ in step 1. of the algorithm above, with all else the same. It is crucial that the data be centered prior to resampling; see Hall and Wilson (1991) and Westfall and Young (1993, pp. 35–43) for resampling guidelines.

Table 2 displays the estimates, standard errors, and margins of error (equal to critical value times standard error) for these methods. The Bonferroni margin of error also is displayed for comparison. Clearly, incorporating the correlations reduces the critical values, making intervals tighter than the Bonferroni intervals. Also, the bootstrap intervals suggest that the effect of incorporating nonnormalities is to further reduce the critical value in this example. Finally, using either the bootstrap or parametric resampling method, one finds that the endpoints Y_1 and Y_2 differ significantly for the treatment and control regimens; however, the Bonferroni method, which does not incorporate correlation information, finds that only Y_1 is significant. The unadjusted confidence intervals find all but Y_2 significant, but the joint level of confidence in these intervals is less than 95%. The true simultaneous confidence level of the unadjusted intervals is, under normality, some number between 80% and 95%, depending upon the correlation structure. The true simultaneous confidence level for the resampling-based intervals is approximately equal to 95% for the normal resampling-based intervals, under multivariate normality of the data; and it is approximately equal to 95% for the

Table 2
Statistics for simultaneous confidence intervals

Endpoint	Estimate	Standard error	95% margins of error			
			Unadjusted	Bonferroni	Resampling-based	
					Normal	Bootstrap
Y_1	−0.6784	0.2658	0.5267	0.6750	0.6554	0.6483
Y_2	−1.5146	0.6079	1.2049	1.5440	1.4993	1.4830
Y_3	−0.3743	0.2893	0.5735	0.7349	0.7136	0.7058
Y_4	0.7505	0.3154	0.6251	0.8010	0.7778	0.7693

bootstrap resampling-based intervals, under the location-shift model with finite variances (for which the normal model is a special case).

If one wishes a reject/accept decision only, and not confidence intervals, then inferences can be made more powerfully using stepwise testing methods as in Section 3.5. We choose to use the Holm (1979) step-down Bonferroni procedure as a basis for comparison with methods that incorporate correlations and nonnormalities. Let the ordered p-values be $p_{(1)} \leq \cdots \leq p_{(k)}$. An adjusted p-value $\tilde{p}_{(j)}$ can be defined for testing the hypothesis $H_{(j)}$, using the Holm step-down method, so that $H_{(j)}$ is rejected at FWE level α when $\tilde{p}_{(j)} \leq \alpha$. Holm's adjusted p-value is defined as $\tilde{p}_{(j)} = \max_{r \leq j}(k - r + 1)p_{(r)}$. The Holm method relies on the Bonferroni inequality, and therefore controls the FWE conservatively.

The method can be made less conservative by incorporating correlations. Suppose the index of $p_{(j)}$ is t_j so that $p_{(j)} = p_{t_j}$. Correlations may be incorporated as follows: define $q^c_{(j)} = P(\min_{r=j,\ldots,k} P_{t_r} \leq p_{(j)} | \cap^k_{r=j} H_{t_r})$, with $t_j, \ldots, t_k$ fixed in the probability calculation. The adjusted p-values are $\tilde{p}^c_{(j)} = \max_{r \leq j}\{q^c_{(r)}\}$. (The "c" superscript emphasizes that these values incorporate correlations.) Note that the values P_{t_r} are the pre-observed random p-values, and the condition $\cap^k_{r=j} H_{t_r}$ states that all null hypotheses $H_{t_j} \ldots H_{t_k}$ are true. The probability calculation involves the joint distribution of the p-values, thus correlations are incorporated. Note that the Holm step-down test procedure is obtained from this method via the Bonferroni inequality: $P(\min_{r=j,\ldots,k} P_{t_r} \leq p_{(j)} | \cap^k_{r=j} H_{t_r}) \leq \sum^k_{r=j} P(P_{t_r} \leq p_{(j)} | \cap^k_{r=j} H_{t_r}) = (k - j + 1)p_{(j)}$. Hence, by incorporating correlations, the adjusted p-values are lessened, implying the method is more powerful.

Since the correlations are unknown, estimates may be substituted, and the resulting adjusted p-values may be computed easily via parametric or nonparametric resampling, as described above for confidence intervals, with the appropriate modifications shown directly above for testing. Specific algorithms are shown in Westfall and Young (1993, p. 66, 123, 130), and Troendle (1995); the bootstrap step-down method is implemented in PROC MULTTEST of SAS/STAT (SAS Institute, 1996). FWE protection is not guaranteed when estimated correlations are used, but only minor problems have been reported for this case (Westfall and Young, 1993, p. 127; Reitmeir and Wassmer, 1996).

Table 3
Multiple testing results for the multiple endpoint data

Endpoint	t	Raw p	Single-step adjusted p-value			Step-down adjusted p-value		
			Bonferroni	Resampling-based		Bonferroni	Resampling-based	
				Normal	Bootstrap		Normal	Bootstrap
Y_1	−2.55	0.0121	0.0483	0.0402	0.0372	0.0483	0.0402	0.0372
Y_2	−2.49	0.0142	0.0569	0.0470	0.0437	0.0483	0.0402	0.0372
Y_3	−1.29	0.1986	0.7943	0.4880	0.4906	0.1986	0.1986	0.1984
Y_4	2.38	0.0191	0.0763	0.0617	0.0582	0.0483	0.0402	0.0372

Multiplicity-adjusted p-values are shown in Table 3 for the various testing procedures. Note that the single-step testing results essentially correspond with Table 2: each adjusted p-value in those columns equals one minus the confidence level for which the value zero lies on the boundary of the corresponding confidence set.

As with the confidence intervals, we see that there is benefit in incorporating correlation structure: the adjusted p-values are smaller for the resampling-based methods than for the Bonferroni methods, showing greater power. Also, we find greater power for all step-down testing methods than for the simultaneous confidence interval methods, as shown by smaller adjusted p-values in the "step-down" columns than in the "single-step" columns. In all cases, the resampling-based methods used 10,000,000 resampled data sets, so the Monte Carlo errors are negligible.

A final caveat: while the bootstrap procedure offers robustness of FWE against departures from normality, it should be noted that the power of the tests can suffer if the distributions are especially outlier-prone. In such a case it would be prudent to rank-transform the data prior to analysis, and to use permutation resampling rather than bootstrap resampling. In this fashion, the power of the tests will be improved, while retaining control of the FWE. Further details may be found in Westfall and Young (1993, pp. 113–121).

4.2. Animal carcinogenicity – Multiple tests with discrete data

Animal carcinogenicity studies have provided fertile soil for the development of multiple testing methods, promoting advances in the areas of multiple testing with discrete data and multivariate binary endpoints. A typical study is performed as a 2×2 factorial experiment, with factors Sex (Male or Female), and Species (Rat or Mouse). Within each arm of such a study, animals are assigned to control and treated groups with varying dose levels. The study period is typically two years. All animals dying during the study period or euthanized at termination of the study are necropsied, and presence and context of particular tumor types are determined. A test for increasing trend of tumor incidence with dose group is performed for each tumor type, as required by FDA.

There are many tumor types that potentially may be found in a given study. A typical type has the form (malignant) (liver) (adenoma), but (malignant) could also be (benign), (liver) could also be any of dozens of possible organs or sites, and (adenoma) could be any one of 100's of possible tumor types. Thus the "family" of tests contains literally thousands of tumor types; however, most of the combinations are extremely rare, and in typical carcinogenicity studies there are only around 30 distinct tumor types observed within a sex/species arm of the study. While one cannot predict in advance which particular 30 (or so) tumor types will be observed, FWE control is maintained by performing multiplicity adjustment performed over the sub-family of observed types. This is so because typical tumor tests are done "exactly", conditioning upon the total number of tumors observed. For those tumor types that are not observed, the totals are zero.

Such tumor types clearly cannot be "statistically significant" conditionally, and therefore are effectively dropped from the analysis. As long as FWE control is maintained conditionally, for each configuration of observed totals, FWE control is also maintained unconditionally, when averaged over all possible configurations of totals.

The issue of whether and how multiplicity adjustment should be performed in animal carcinogenicity studies depends upon an individual's perspective. From the pharmaceutical company's standpoint, a drug committed to animal testing is likely to be reasonably safe, as determined from other evidence and prior analysis; otherwise, the large expense of a long-term animal trial is not justifiable. However, with a large number of tumor types tested, the likelihood of observing a false positive $p \leq 0.05$ carcinogenic determination becomes quite likely, as high as 45% within a particular arm of a study (Westfall and Soper, 1998). Such an outcome will significantly delay a drug's entry to marketplace, or possibly even stop development of a drug that is safe in reality, in either case costing the pharmaceutical company millions of dollars. Conversely, regulatory agencies must protect the public from cancer-causing drugs. Since multiplicity adjustment increases the probability of declaring cancer-causing drugs to be safe, the FDA historically has requested analyses of tumor data that are "not adjusted for multiple comparisons or multiple testing" (Dept. of Health and Human Services, 1987).

Despite the historical FDA position, it is clear that if a test of significance is performed for all $\approx 4 \times 30 = 120$ tests in all arms of the 2×2 carcinogenicity study, there is ample opportunity for false positive determinations to arise. Interestingly, the appropriate level of multiplicity adjustment for animal carcinogenicity studies is nowhere near the Bonferroni correction α/k because of the discrete nature of the data. Although 120 tumor types may be observed in a study, most of the types will have very few (e.g., two or less) occurrences within a particular arm of the study. Depending upon the testing method used, it is either impossible or nearly impossible to generate statistical significance with such rare occurrences, and such sites therefore should not be allowed to contribute to the multiplicity adjustment. This problem is illustrated in Figure 1, which displays the probability of Type I error for two-group comparison with 50 animals per group, as a function of the underlying control probability that an animal has a tumor. Most tumor types have spontaneous generation rates of less than 1%, thus, as can be seen in Figure 1, there is little opportunity to observe a Type I error. Further, if the tests are performed using Fisher exact tests, the Type I error rate is substantially less than .05 even with more common tumor types. Compounding this reduced Type I error level over many tests produces a false positive rate that is substantially less than the $1 - (1 - .05)^k$ rate one might expect with independent and uniformly distributed p-values, as documented by Fears et al. (1977), Haseman (1983), and Haseman et al. (1986).

In effect, the Bonferroni multiplier is lowered from k to some k', where k' more accurately reflects the number of tumors that possibly could generate a false positive. The multiplicity-adjusted p-value for a given test then would be $\tilde{p}_j = k'p_j$,

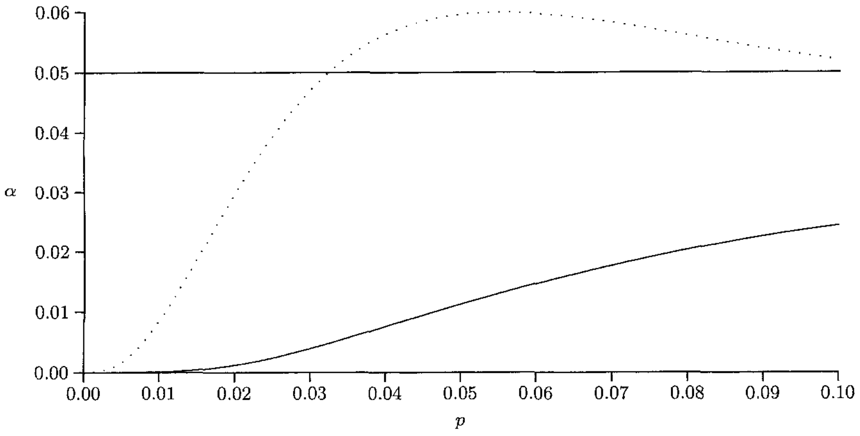

Fig. 1. Actual Type One Error level of nominal 0.05-level tests (upper-tailed) for equality of proportions with $n_1 = n_2 = 50$ observations per group, as a function of p, the common proportion for the two groups. The solid curve represents the Fisher exact test, the dotted curve represents the usual Z test.

when testing carcinogenicity for tumor type j. Determination of such a k' is the essence of the Tarone (1990) method for multiplicity adjustment with carcinogenicity studies. A more precise adjustment was given by Heyse and Rom (1988) and Farrar and Crump (1988), who defined adjusted p-values $\tilde{p}_j = P(\min P_i \leq p_j)$. The probability calculation can be performed in "exact" fashion, using the (conditional) permutation distributions of the p-values, or in asymptotic fashion by estimating and simulating from the unconditional distribution of min P_i (Westfall and Young, 1989, 1993; Soper and Westfall, 1990).

Following Westfall and Wolfinger (1997), who provide an overview of the min P_i – based multiplicity adjustments in discrete applications, assume the observable values of the random p-value P_i are $\{p_{it}; t = 1, \ldots, m_i\}$, with $\Pr(P_i \leq p_{it}) = p_{it}$. Now, assuming the p-values are independent (a reasonable assumption in carcinogenicity studies as discussed by Heyse and Rom, 1988, and Soper and Westfall, 1990), obtain $\tilde{p}_j = 1 - \Pi_{i=1}^{k}(1 - p_{it(j)})$, where $p_{it(j)} = \max_t\{p_{it}; p_{it} \leq p_j\}$, if $\min_t\{p_{it}\} \leq p_j$ and 0 otherwise. To see the effects of incorporating discreteness in, consider the data in Table 4, where $k = 4$ 2×2 contingency tables are considered. There are $n_C = 50$ and $n_T = 48$ observations in the "control" and "treated" groups. The test statistics t_j are the number of tumors in the treated group, and the upper tail of the permutation distribution (using the Fisher exact test, conditioning on the total number of tumors in treated and control groups) of each t_j is given.

The smallest observed upper-tail p-value is $p_1 = 0.02521$. Under independence, and incorporating the discrete characteristics, this p-value is multiplicity-adjusted to $\tilde{p}_1 = 1 - (1 - 0.02521)(1 - 0.00532)(1 - 0)(1 - 0.00645) = 0.03665$, significant at the familywise $\alpha = 0.05$ level. On the other hand, if the p-values are assumed to be uniformly distributed, then $\tilde{p}_1 = 1 - (1 - 0.02521)^4 = 0.09709$, barely significant at the familywise $\alpha = 0.10$ level.

The magnitude of the improvement depends upon the specific characteristics of the discrete distributions. With multiple tumor data, the smallest tail probability from many of the permutation distributions exceeds 0.05 (e.g., tumor type T_3 in Table 4); therefore, those tumor types are completely and automatically excluded when calculating multiplicity adjustment for the remaining tumor types. Rom (1992) showed how to improve these discrete tests further by employing a still more powerful test, in conjunction with the closure method of Section 3.3.

In animal carcinogenicity studies, simple two-group binomial tests are rarely appropriate. Rather, dose-response trend tests that account for differential mortality and context of tumor observation (incidental, lethal, or palpable) are needed (Peto et al., 1980). Such tests can be performed in "exact" fashion much like the Fisher exact test, as discussed by Ali (1990), and Soper and Tonkonoh (1993), resulting in permutation distributions similar to those shown in Table 4. The multiplicity adjustment can then be performed in the same manner assuming independence of all tumor types. PROC MULTTEST performs such adjustments automatically; the calculations also can be performed using the permutation distributions output from StatXact (Mehta and Patel, 1995).

Table 5 displays results from an actual animal carcinogenicity study involving male rats in a design with $n_0 = n_1 = n_2 = n_3 = 60$ animals per dose group. The tumor names are disguised for confidentiality, but for the sake of concreteness, we may assume that, e.g., T_1 is a benign pituitary adenoma, T_2 is a malignant pitu-

Table 4
Data and upper-tail p-values for $k = 4$ Fisher exact tests

Group	Observed data			
	Tumor T_1	Tumor T_2	Tumor T_3	Tumor T_4
Control	0/50	4/50	0/50	6/50
Treated	5/48	3/48	4/48	4/48
	Permutation distributions			
Test Statistic, t	T_1 Total = 5	T_2 Total = 7	T_3 Total = 4	T_4 Total = 10
0	1.00000	1.00000	1.00000	1.00000
1	0.96880	0.99278	0.93625	0.99927
2	0.80602	0.93765	0.67580	0.99068
3	0.48047	0.76489*	0.29327	0.94744
4	0.16848	0.47697	0.05387*	0.82409*
5	0.02521*	0.20129		0.60332
6		0.04967		0.34428
7		0.00532		0.14250
8				0.03946
9				0.00645
10				0.00047

* denotes observed p-value

itary carcinoma, T_3 is a malignant kidney liposarcoma, etc. There were 44 such tumor types ($T_1 - T_{44}$) observed in this particular study. Among these types, 27 had only one tumor observed among all groups, and the smallest possible permutational p-value among these types was 0.09375. Thus, any multiplicity adjustment for tumor types with unadjusted p-value .05 or less need only consider the remaining $44 - 27 = 17$ remaining types. Further, only 15 of these 17 distributions allowed p-values of .05239 or less, and only 7 distributions allowed p-values of .00107 or less. This information is presented in Table 5 under the "Single-Step Rest Bonferroni" column ("Rest" short for "Restricted"), which adjusts the smallest p-value .00107 using $\tilde{p}_{10} = 7 \times .00107 = .0075$; similarly obtain $\tilde{p}_{23} = 15 \times .05239 = .7859$.

The "Adj p" columns contain the independence-assuming multiplicity-adjusted p-values described above, computed using PROC MULTTEST. Comparing these values with the adjusted p-values in the Bonferroni columns, we see the great benefits of incorporating the specific distributional characteristics. The ordinary Bonferroni method ("Unrest." or unrestricted, in Table 5), in which p-values are multiplied by 44, the number of detected tumor types, is seen to be grossly conservative compared to the more appropriately adjusted p-values. Also, even the restricted Bonferroni method is seen to be quite conservative, because it does not use the specific discrete distributions of the tumorigenicity tests.

The "step-down" columns of Table 5 contain the results of step-down testing, which preserves Type I error control while making tests more powerful. The smallest p-value is not changed when performing tests in step-down fashion, but the remaining adjusted p-values are uniformly as small or smaller than the corresponding single-step methods. The step-down restricted Bonferroni adjustment at the second step is $\tilde{p}_{23} = 14 \times .05239 = .7335$.

The recommendation here is to use an adjustment of the type shown in the "Step-Down Adj p" column. While control of FWE is not mathematically guaranteed using this method because of the independence approximation, recent extensive simulations using historical data have suggested that the FWE rates are controlled almost exactly at the nominal .05 level under complete null hypotheses, and are controlled at levels somewhat below .05 under partial null hypotheses.

Table 5
Multiple testing results for an animal carcinogenicity study

Type	Raw p	Single-step			Step-down		
		Adj p	Unrest. Bon p	Rest. Bon p	Adj p	Unrest. Bon p	Rest. Bon p
T_{10}	0.00107	0.0038	0.0471	0.0075	0.0038	0.0471	0.0075
T_{23}	0.05239	0.4291	1.0000*	0.7859	0.3975	1.0000*	0.7335
T_{17}	0.10256	0.8778	1.0000*	1.0000*	0.8564	1.0000*	1.0000*
... (41 more tumor types) ...							

*p-value truncated to 1.0000

Despite the FDA's historical opposition to multiplicity adjustment for tumor tests, there has been movement toward acknowledging the multiplicity problem. The so-called "Haseman Rule" (Haseman, 1983) has been used routinely by FDA. With this rule, a rare tumor type is flagged as significant when $p \leq .05$, and a common tumor type is flagged as significant when $p \leq .01$. This method was originally developed as a multiplicity adjustment procedure since its FWE is approximately .08 over all four arms of the four-group animal carcinogenicity study. The reason the FWE is (surprisingly) this low is shown in Figure 1: with Fisher exact tests, and 50 animals per group, the actual Type I error levels are much less than the nominal .05 level, particularly for the rare tumor types. However, for typical carcinogenicity studies using multiple groups and trend tests, the Type I error levels are really much higher. Westfall and Soper (1998) report FWE rates as high as 45% within each arm of the 2×2 study when the Haseman rule is used under typical tumor incidence patterns. Recently, in response to such criticisms, FDA has espoused a more stringent criterion for determining carcinogenicity, namely a $p \leq .025$ criterion for rare and $p \leq .005$ for common tumors (Lin and Rahman, 1998). This method reduces FWE rates, but still does not control the FWE below .05 within any arm of a typical animal carcinogenicity study as shown by Westfall and Soper (1998).

Determination of "rare" and "common" in the Haseman-rule type procedures requires historical control data, with frequencies <1% denoting "rare," ≥1% denoting "common." A problem with the application of this rule is that relevant historical data may be unavailable, or it may be considered that study-to-study variation is too large for historical data to be meaningful. In this case, an "estimated" version of the Haseman has often been used, where the determination of "rareness" is based on the frequency of tumors in the concurrent control group. As detailed by Westfall and Soper (1998), this practice must be avoided, as it greatly increases the FWE.

There is particular interest in rare tumor types. The rule developed by Haseman, which emphasizes rare tumors, was meant to combine biological judgment as well as statistical theory to determine carcinogenicity. If one wishes to similarly weight tumor types in a method that controls FWE, it is easy to do: define "weighted adjusted p-values" as $\tilde{p}_j = P(\min w_i P_i \leq w_j p_j)$, with weights such as $w_i = 5$ for common tumors, $w_i = 1$ for rare tumors, as suggested by the Haseman rule. Such adjustments divide the FWE to allow more power for rare types, at the expense of losing power for the common types, while still controlling the FWE at or below the .05 level. Westfall and Soper (1998) develop the method in more detail.

4.2.1. Adverse events – More multiple tests with discrete data

New drugs must be tested on human subjects for evidence of safety and efficacy. As with animal carcinogenicity tests, the data arising from safety studies typically have a multivariate binary form, i.e., a 0/1 presence or absence indicator for any of several adverse events, also called side effects. Another similarity between these applications is that there are typically very many untoward events that might occur, and if all such events are tested, false positive determinations are likely.

Drugs often do have side effects; however, it can easily happen that statistical determinations based on the simple $p \leq .05$ rule will incorrectly flag side effects too often. As with any screening procedure, there are errors and costs. A Type I error in the analysis of adverse events implies making a claim that a drug causes some problem, say, headache for example, when the drug has no such effect in reality. The costs of such Type I errors include delayed approval of a drug, or possibly even cancelation of development of a good drug in the case of a serious adverse event, with costs both to the drug company and the consumer. On the other hand, Type II errors are also very serious, as they may cause undue suffering for the public and possibly lawsuits for the drug company. Because of the concern for Type II errors, some have advocated de-emphasizing the multiplicity problem, even analyzing all sites individually at the unadjusted $\alpha = 0.10$ level (Edwards et al., 1990, p.144). Our view is that, because Type I errors can and do occur, any data analysis should acknowledge this fact in some direct fashion. Use of unadjusted methods is reasonable, as long as the FWE is estimated and acknowledged, and as long as the analysts are comfortable with this value. If this number is too large, then the analyst should choose a multiple testing procedure that is as powerful as possible, and which controls the FWE at some pre-determined level.

As in the animal carcinogenicity example, power can be increased greatly in adverse events analysis over the Bonferroni-type method by incorporating discreteness of the tests. Our suggestion is that, if FWE control is a concern, then multiple testing analysis methods for discrete data should be used. The application and method are virtually identical to that shown in the previous section for animal carcinogenicity tests: the data structure can be represented as multivariate binary in each situation, and the same general analysis methods can be used.

4.3. Subgroup analysis – Incorporating logical constraints and correlations

Special difficulties in multiple testing are posed by subgroup analysis. In a typical study of treatment efficacy, a question may arise as to whether the effect occurs in all possible subgroups, e.g., male and female, young and old, ethnic group, and combinations thereof. One difficulty is the definition of a "family" of subgroups, as consideration of interaction subgroups can increase family sizes dramatically. Another problem is that the sample sizes are inevitably smaller in subgroups, implying less power. Thus, even though there may be a significant treatment effect with all data combined, it will be much more difficult to detect significance in subgroups, even when there are in fact real treatment effects in the subgroups. The converse problem occurs when the treatment is inefficacious for all or most of the subgroups. In this case, repeated significance testing over all statistical subgroups will yield Type I errors with high probability.

Documented cases exist where serious errors were made because of careless subgroup analysis. One case, reported in Fleming (1992) concerned a pre-operative radiation therapy for the treatment of colon cancer. The study was stopped early due to lack of significance; however, follow-up analysis revealed a "significant" improvement in a particular subgroup. The trials conclusions were then revised to recommend "universal use" of the therapy. A follow-up study involving the same therapy and a larger sample size revealed no statistical significance, so it seems likely that the original finding of a therapeutic effect was merely a Type I error.

Another case, reported in the *Wall Street Journal* (King, 1995) concerned the development of "Blue Goo," a salve meant to heal foot wounds of diabetic patients, by the Biotechnology firm ProCyte Corp. The firm decided to proceed with an expensive, large-scale Phase III clinical trial to assess efficacy, based on statistically significant efficacy results found in a subgroup of patients in the Phase II clinical trial. The larger Phase III study found no significant effect of the "Blue Goo" therapy, and as reported by King, "Within minutes [of the announcement], ProCyte's stock fell 68% ... ." As in the case of the pre-operative radiation treatment, it seems likely that the significant result in the Phase II subgroup was a Type I error.

With any multiple testing procedure, the family of tests must be decided in advance to insure validity. To this end, the set of subgroups should be carefully restricted in advance to include only those that are genuinely of interest. If the family is too large, then proper multiplicity adjustment will lack power. If the family is too small, then potentially interesting subgroups will be overlooked.

Table 6 contains summary data of a clinical trial reported in Koch et al. (1990) to evaluate the effectiveness of an Active respiratory therapy versus a Placebo. As originally reported, the data contain physician's ratings of a patient's respiratory health prior to treatment, and at four post-treatment evaluations. The ratings are scored from 0 to 4, where 0 represents poor health and 4 represents good health. We chose a weighted average of the four post-treatment evaluations as primary endpoint: Score $= (R_1 + 2R_2 + 3R_3 + 4R_4)/10$. The data are stratified by Age ("Older" patients being those older than 30 years) and by Initial Health ($R_0 \leq 2$ denoting "poor" and $R_0 > 2$ denoting "good").

Let the population means in the cells shown in Table 6 be denoted μ_{ijk}, where $i = 1, 2$ denote Active and Placebo, $j = 1, 2$ denote Younger and Older, and

Table 6
Analysis of respiratory rating scores with age and initial health subgroups

Initial health	Statistic	Younger		Older	
		Active	Placebo	Active	Placebo
	Mean	3.2455	3.3500	3.6615	2.4429
Good	Std dev	0.6654	0.8263	0.3754	1.2586
	n	11	12	13	14
	Mean	2.6529	1.7833	2.5154	1.7000
Poor	Std dev	1.0607	1.1892	0.9932	1.1633
	n	17	12	13	19

$k = 1, 2$ denote Good and Poor initial health. Several possible subgroup analyses can be defined in terms of the μ_{ijk}. Table 7 lists the chosen subgroup tests, with defining contrast vectors and one-sided summary test results.

The primary test of interest is the overall "Active vs. Placebo" comparison, which is tested as $H_0 : (\mu_{111} + \mu_{112} + \mu_{121} + \mu_{122})/4 - (\mu_{211} + \mu_{212} + \mu_{221} + \mu_{222})/4 = 0$, versus the upper-tail alternative. The next four tests concern the efficacy for the Age and Initial Health subgroups, e.g., "Younger" is a test of $H_0 : (\mu_{111} + \mu_{112})/2 - (\mu_{211} + \mu_{212})/2 = 0$, again with upper-tail alternative. The final four tests concern efficacy in Age × Initial Health subgroups, e.g., "Young, Good" is a test of $H_0 : \mu_{111} - \mu_{211} = 0$.

Three issues are of concern when testing this family of nine hypotheses. First, some tests are more important than others. In all likelihood, the first test is of most interest, the second four having somewhat less importance, and the final four least important. Such information should be incorporated in the multiple testing strategy. Second, the contrasts given are logically interrelated. Truth of certain subsets necessarily implies truth of other subsets. Thus, having rejected the most significant hypothesis in a step-down hierarchy, one need not consider that all remaining tests are possibly true for the multiplicity adjustment at this step, as occurs with the Holm (1979) procedure, which uses $k - 1$ for a Bonferroni multiplier. Rather, as a consequence of the closure method discussed in Section 3.3, one need consider only the largest collection of hypotheses that possibly *could be* true, while not contradicting the falseness of the most significant test (Shaffer, 1986), leading to a Bonferroni multiplier $k' \leq k - 1$. A third concern is that the overlap among contrast coefficients creates correlations among the test statistics, therefore, use of Bonferroni multipliers is conservative. Westfall (1997) developed a method (software called "MTEST" is freely available at http://lib.stat.cmu.edu/jasasoftware/mtest) to

Table 7
Subgroup analysis of respiratory therapy trial with upper-tail p-values and step-down adjustments

Contrast	Active				Placebo				RawP	HolmP	ShP	CShP*
	Younger		Older		Younger		Older					
	G	P	G	P	G	P	G	P				
Overall	1	1	1	1	−1	−1	−1	−1	0.0002	0.0019	0.0012	0.0010
Younger	1	1	0	0	−1	−1	0	0	0.0894	0.1788	0.0894	0.0894
Older	0	0	1	1	0	0	−1	−1	0.0001	0.0010	0.0010	0.0008
Good Init.	1	0	1	0	−1	0	−1	0	0.0268	0.0804	0.0536	0.0477
Poor Init.	0	1	0	1	0	−1	0	−1	0.0009	0.0061	0.0043	0.0037
Young, Good	1	0	0	0	−1	0	0	0	0.5982	0.5982	0.5982	0.5982
Young, Poor	0	1	0	0	0	−1	0	0	0.0119	0.0593	0.0356	0.0287
Old, Good	0	0	1	0	0	0	−1	0	0.0011	0.0064	0.0054	0.0045
Old, Poor	0	0	0	1	0	0	0	−1	0.0131	0.0593	0.0394	0.0367

* Maximum 99.7% Monte Carlo error is 0.00013

incorporate both the Shaffer constraints and the correlations into the multiplicity adjusted p-values.

The analysis of Table 7 solves the second and third problems of the preceding paragraph.

Here, we are assuming that all tests are essentially of equal importance, and that we would be as happy to claim significance for the "Old, Good" test as for the "Overall" test. RawP is the unadjusted upper-tail p-value from a t-test from the three-factor cell-means ANOVA with MSE $= 1.00949777$ and $df = 103$. HolmP contains p-values adjusted via $p' = (k - j + 1)p$, j being the step in the hierarchy, with adjusted p-values monotonicity-enforced so that their ordering matches that of the unadjusted p-values (e.g., Westfall and Young, 1993, p. 64). ShP (short for "Shaffer p-values") contains p-values adjusted via $p' = k'_j p$, where k'_j is the number of hypotheses that possibly can be true, given those tested at stages $1, \ldots, j - 1$ are false. For example, having first rejected the most significant "Older" hypothesis, the maximum number of remaining eight tests that possibly can be true, given that the "Older" hypothesis is false, is five. Hence the ShP value for the second-most-significant test, the "Overall" test, is $5 \times 0.0002 = 0.0012$ (discrepancy is due to rounding of 0.0002 from the actual 0.0002343). Finally, "CShP" (short for correlation-adjusted Shaffer p-values) are the preferred adjustments, since these incorporate the correlations as well as the logical constraints. The CShP values are mathematically smaller than the ShP values, since the ShP can be obtained from CShP via Bonferroni's inequality. Specifically, following Westfall (1997), the CshP value is

$$\tilde{p}_{(j)} = \max_{K \in S_j} \Pr\left(\min_{l \in K} P_l \leq p_{(j)} \middle| \cap_{i \in K} H_i \right) ,$$

which is the maximum of probabilities that the min P statistic will be less than the observed p at stage j, with maximum being taken over all sets of hypotheses that possibly can be true at the given stage, given that all previously rejected hypotheses are false. Obtaining these adjusted p-values may require Monte Carlo methods; Westfall (1997) describes a simple and efficient method that is coded in the MTEST software. The values in the "CShP" column are quite accurate, with small discrepancies possible only in the fourth decimal, and were computed in less than 2 hours using a Pentium 200 MHz computer. The CShP values also are recommended over all others given in the table, since they control the FWE, and are the most powerful of those presented.

Again, this analysis presumes equal interest in all tests. Alternative methods are needed when some tests are more important than others. To protect the FWE, the following hierarchical strategy can be adopted: (1) test the "Overall" hypothesis, at unadjusted level α. If it is insignificant, then stop. If it is significant, then proceed to test the family of four single-factor subgroup hypotheses, at FWE$= \alpha$. If none are significant, then stop, otherwise, proceed to test the family of four two-factor subgroup hypotheses at FWE$= \alpha$.

A problem with this hierarchical approach is that it may miss some very important effects in the interaction subgroups, since one or more of those tests might be highly significant, while the overall test is insignificant. An alternative method which allows differential weight for all tests, while insuring that any test with sufficient significance will be flagged, is to use a weighted multiplicity adjustment. For example, one might consider that 80% of the FWE should be allocated to the overall test, 16% to the four main-effect subgroups, and 4% to the four interaction subgroups. In this case, defining weights $w_1 = 1/.8$, $w_2 = \cdots = w_5 = 1/.16$, and $w_6 = \cdots = w_9 = 1/.04$, a simple Bonferroni adjustment yields adjusted p-values $\tilde{p}_i = w_i p_i$, with rejection when $\tilde{p}_i \leq \alpha$. This method can be improved to incorporate stepwise testing (Holm, 1979; Benjamini and Hochberg, 1997), and also to incorporate correlation structures (Westfall and Young, 1993, pp. 184–188).

Interaction pre-tests exacerbate multiplicity effects

Finally, we note that the level of interest in the overall or interaction subgroup hypotheses usually depends upon the unknowns of whether the interactions exist. If there are no interactions with treatment, then the overall test is the only test of interest. If treatment interacts with age but not initial condition, then only the "Younger" and "Older" tests are of interest, but not the two initial condition tests. In practice, a hierarchical strategy involving testing first for interactions, then deciding which follow-up tests to perform, is usually adopted. While this procedure has "common-sense" appeal, it does not guarantee FWE protection, and is difficult to study because of the conditional nature of the follow-up tests. Specifically, rejecting a preliminary interaction test implies a tendency for the error variance to be underestimated, and for larger variation in estimated treatment effects in the subgroups, which implies possibly higher Type I error rates at succeeding levels.

To illustrate, consider a multi-center trial with four centers, each with 50 observations in treatment and control groups. The observations are independent and normally distributed with unit variance in all groups. The means for all groups are zero, with the exception of the treatment group in center 1, which has a mean 0.25. We simulated 100,000 data sets from this process, and used the Holm step-down procedure to assess significant differences between the treatment and control groups at each of the four centers. In this case, FWE is the probability of rejecting the null hypothesis of no treatment effect for at least one of centers 2, 3, and 4. Using all 100,000 samples, FWE = 0.0374 (±0.0015); using the 21,130 samples for which the ANOVA treatment × center interaction pre-test was significant, FWE = 0.1369 (±0.0061) (99% margins of Monte Carlo error are given). The situation is worse with simultaneous confidence intervals; when using 95% Šidák intervals following the significant test, the simultaneous coverage level of the four intervals for treatment differences was estimated to be 83.11% (±0.66%) following the significant pretest.

We recommend that such a hierarchical procedure for the analysis of subgroups be either avoided or used with extreme caution. Significant results culled from such tests should be labeled as exploratory, until further research replicates the claimed effects.

4.4. Power analysis and sample-size determination

Planning of studies requires sample size determination using power analysis. After specifying a "practically significant" effect size (where, e.g., effect size is either $\mu_1 - \mu_2$ or $(\mu_1 - \mu_2)/\sigma$ in a two-group comparison), the researcher calculates the sample size needed to detect such an effect with high probability (typically 80% or 90% probability).

In multiple testing situations, power is not so easily defined. Assume there are k null hypotheses H_i. Suppose for some subset $K \subseteq \{1, \ldots, k\}$, all H_i are true when $i \in K$, and all remaining hypotheses are false (false for $i \in K'$). Various definitions of power include

1. the probability of correctly rejecting *at least one* H_i, for $i \in K'$,
2. the probability of rejecting *all* H_i, for $i \in K'$,
3. for a *specific* $i \in K'$, the probability of rejecting H_i.
4. the expected average number of rejections for $i \in K'$, defined as $(1/|K'|) \sum_{i \in K'} P(\text{Reject}\, H_i)$ or perhaps a weighted version, $\sum_{i \in K'} w_i P(\text{Reject}\, H_i)$, where the weights w_i are chosen to reflect the a priori importance of the various hypotheses, and are constrained to sum to unity.

The appropriate definition of power in multiple testing situations depends upon the researcher's goals. Researchers ideally would like a high power using definition (2); however, this is a rather stringent definition, resulting in very low power.

Calculation of power, by any definition, requires assumptions (knowledge) of the distributions of the test statistics, and their correlations. In such cases, it is usually possible to calculate power using any of definitions 1–4, at least via simulation, for given sample sizes and effect sizes. In cases where appropriate tables and or simulation studies are neither available nor feasible, we can recommend that power be calculated analytically using standard normal-theory univariate analysis methods, with independence-based single-step multiplicity adjustments. Since the power will actually be higher when using stepwise methods and by incorporating correlations, the resulting sample sizes can be considered as upper bounds.

4.4.1. Sample size determination with multiple continuous endpoints

As in Westfall and Young (1993, pp. 207–210), we suppose that n of treated and control observations are to be sampled in a multiple endpoints study. The test statistics are

$$T_j = \sqrt{\frac{n}{2}} \frac{\bar{Y}_{2j} - \bar{Y}_{1j}}{s_j}$$

where s_j is the pooled standard deviation estimate for variable j. This statistic has the central t-distribution with $v = 2(n-1)$ degrees of freedom under the null hypothesis $H_{0j} : \mu_{1j} = \mu_{2j}$. Under the alternative hypothesis, the statistic is distributed as non-central t with $v = 2(n-1)$ and noncentrality parameter

$$\delta_j = \sqrt{\frac{n}{2}}\frac{\mu_{2j} - \mu_{1j}}{\sigma_j}$$

where σ_j^2 is the variance for variable j, assumed constant for both groups.

Assuming an interest in individual tests, we will adopt definition (3) for power computation. Suppose then that we wish to declare test j significant when in fact $(\mu_{2j} - \mu_{1j})/\sigma_j >$ Constant, where "Constant" denotes a *scientifically significant* increase. That is, if the difference in the means exceeds an important fraction of within-group standard deviation, we wish to declare a significance for variable j. Using single-step independence-based multiplicity adjustments, the adjusted p-value for variable j will be significant, when $T_j > t_{\nu\alpha'}$, where $\alpha' = 1 - (1 - \alpha)^{1/k}$, and where $t_{\nu\alpha'}$ is the $1 - \alpha'$ quantile of the central t-distribution with ν degrees of freedom.

The probability that test j will be declared significant can be computed using the noncentral t distribution, available in many statistical software packages. Taking $k = 10$ and $\alpha = .05$, Figure 2 shows the power as a function of n for different constant proportions of the standard deviation.

While a graph such as Figure 2 is easily constructed and can be used to guide sample size selection when multiple tests are planned, there are many assumptions implicit in constructing this graph that may not be true in general. First, the assumption of independence among variables makes the critical values larger than needed. To assess this affect, simulation or analytic methods could be used to find an appropriate critical value $t_{R,\nu,\alpha}$, defined as the $1 - \alpha$ quantile of the distribution of max T_j under the assumption that the treatment and control data are sampled from multivariate normal distributions with mean zero and covariance matrix R. The matrix R can be "guessed," determined from historical data, or both. Using the resulting critical point, the power under definition (3) is calculated again using

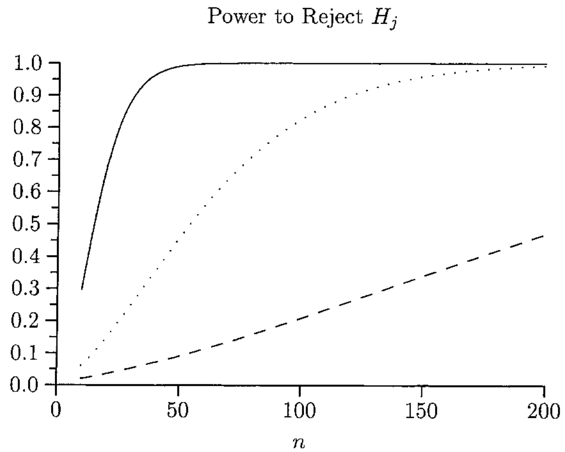

Fig. 2. Power functions for detecting H_j when $\mu_{2j} - \mu_{1j} = .25\sigma_j$ (dashed line), $\mu_{2j} - \mu_{1j} = .50\sigma_j$ (dotted line), and $\mu_{2j} - \mu_{1j} = \sigma_j$ (solid line), from Westfall and Young (1993, p. 209).

univariate methods, with the noncentral t distribution. Since the critical values $t_{R,v,\alpha}$ are generally smaller than $t_{v\alpha'}$ by Sidák's inequality (Sidák, 1967), the power will be higher using the critical values $t_{R,v,\alpha}$. The input matrix R is not known, and it is prudent to evaluate critical values using several "guesses" of R to assess robustness of the recommended sample sizes.

For an example showing the difference between $t_{R,v,\alpha}$ and $t_{v\alpha'}$, consider again Table 2. The critical values are not shown in the table, but are $t_{109,.05} = 1.9820, t_{109,0.0127415} = 2.5327$ (slightly smaller than the Bonferroni critical value $t_{109,0.0125} = 2.5398$ used in the table), and $t_{R,109,.05} = 2.4663$, obtained taking R to be the correlation matrix of Table 1. With larger correlations, $t_{R,109,.05}$ tends toward the unadjusted critical value 1.9820; with smaller correlations, it tends toward the independence-assuming value 2.5327. In this example with moderate correlations, the independence-based critical values are a fair approximation (certainly much better than the unadjusted critical values), and can reasonably be used for sample size determination.

4.4.2. Sample size determination with multiple discrete endpoints

With discrete multiple endpoints, the situation is more complicated. In this case, the *effective* number of tests is not k, but something perhaps much less than k, say k', as described in Section 4.2. The analyst needs to determine three things to determine such a k': (1) how many total binary variables are to be considered, say k_1; (2) of these, approximately how many have background rates (in the absence of treatment) small enough not to affect the multiplicity adjustments, say k_2; (3) of those variables with small background rates, approximately how many are likely to be affected by the treatment, raising the number of sample occurrences enough to make a difference in the multiplicity adjustment, say k_3. It will be expected that $k' = k_1 - (k_2 - k_3)$ variables in the study will have observed rates large enough to affect multiplicity adjustments. Power may be approximated using standard normal approximations for k' test statistics, as well as the assumption of independence.

Alternatively, if historical data are available, then the power can again be approximated via simulation analysis. Further details concerning power and simulation using discrete variable may be found in Westfall and Young (1993; 173–174, 180–183, and 209–210).

5. Other problems and approaches

5.1. Other problems associated with the multiplicity problem

5.1.1 Publication bias

The (type-I) multiplicity problem generally indicated as the cause of publication bias is *the replication of experiments involving a null effect until a significant result is produced*, see e.g. Sterling et al. (1995). This is an example of the type of multiplicity problem indicated in our introduction section as intractable. Another cause of the publication bias problem is the tendency of researchers to inflate their families of null hypotheses (to be tested in the framework of one study). Present practice often involves presenting multiple tables of multiple p-values in the main body of a paper with a selected few discussed in the "conclusions" section without proper recognition of the selection bias involved. This cause of publication bias has been indicated in several texts, e.g. Pocock (1983).

The second cause of publication bias just indicated could be used to further highlight the controversy between per-comparisonists and familywise-controllers alluded to in our Introduction section. It is well known and documented (see e.g. Godfrey, 1985) that traditional MCPs were under-utilized in the medical literature. It is also no secret why researchers prefer per-comparison procedures or at least why some shy away from the bulk of the methods discussed in this article. In Section 5.2 we discuss a new approach to multiple comparison problems which might prove useful towards reducing causes of publication bias associated with *the problem of multiple comparisons* if adopted as a policy requirement by journal editors. The importance of such a perspective is further indicated by the following sub-subsection.

5.1.2. *Meta-analysis*

Our main reference on this important topic is the special issue of the American Statistical Association devoted to the general problem of combining information, by the panel Draper et al. (1992). In that issue, one can find several discussions on meta-analysis and also on the problem of multiple-comparisons which concerns us here. The main utilization of classical MCPs (discussed in their Section 4.7) involves pre-testing of exchangeability among studies "before pooling the result of such studies." This procedure is indicated as necessary to ascertain "the validity of the exchangeability assumption." While this is an important consideration, we refer back to our cautionary note about interaction pre-tests Section 4.3, where we found that such pre-tests can make multiplicity effects worse in follow-up tests.

5.1.3. *Intersection-union (IU) vs. UI*

In their introduction HT discussed an example which seems similar to the many-to-one problem but is actually very different in nature, namely, assessment of "combination drugs": "*Before a pharmaceutical company can market a combination drug, the regulatory agency requires that the manufacturer produce convincing evidence that the combination drug is better than every one off its m (say) subcombinations, which may be regarded as controls... Thus protection is needed against erroneously concluding that the combination drug is better than all of its subcombinations when in fact some of them are at least as good.* If separate one-sided tests are used ... then the probability of erroneously recommending the combination drug can be seen to achieve its maximum at a ... configuration where exactly one subcombination is equivalent to the combination drug and all the others are infinitely worse. This follows from Berger's (1982) general results on *Intersection-Union tests*. Thus to control the relevant Type-I error probability it is only necessary to test each one of the m least favorable configurations at level α."

Note that in spite of the similarity with the UI many-to-one problem, here the overall null hypothesis H_0 which must be rejected (to reach the conclusion of a beneficial combination drug) is a *union* of the individual one-sided hypotheses expressing no benefit for the combination drug over its components. The suitable procedure involves testing each such one-sided hypothesis at level α and *rejecting H_0 if and only if all such one-sided hypotheses are rejected.* Thus, the rejection region for H_0 is the *intersection* of the usual per-comparison rejection regions.

The original idea and theory for testing *union hypotheses* with such IU tests was given by Lehmann (1952). Various problems in Biostatistics are of this type. In the following subsection we demonstrate the general IU method.

5.1.4 Bioequivalence/bioavailability

Chow and Liu (1992) discuss some early practices of assessing Bioequivalence (e.g. of generic drugs with the standard following termination of the original patent) based on Bioavailability measures. In the "beginning" the practice was to test null hypotheses on mean differences by usual α-level tests and conclude Bioequivalence when the null hypothesis was not rejected. But such a procedure does not provide suitable protection against an erroneous conclusion of Bioequivalence. Alternatively we express the null hypothesis $H_0(\delta)$ which postulates that $|\mu_t - \mu_s| > \delta$ where μ_t and μ_s are the means for the generic and standard drugs, respectively. Note that this is a *union* hypothesis $H_0(\delta) = H_+(\delta) \cup H_-(\delta)$ where $H_+(\delta) : \mu_t - \mu_s > \delta$ and $H_-(\delta) : \mu_t - \mu_s < -\delta$. This is the *null hypothesis of non-equivalence.* To conclude equivalence with an α-level procedure one may use the IU procedure which rejects H_0 when both $H_+(\delta)$ and $H_-(\delta)$ are rejected by their respective α-level one-sided tests. This procedure is known as Schuirmann's two one-sided α-level tests. Obviously it is a special case of the IU method.

Further demonstration of the IU method can be presented in the context of Bioequivalence assessment in terms of multivariate bioavailability profiles and in the context of assessing equivalences among several formulations (of a drug). Hochberg (1996) discusses exact α-level procedures for assessing (i) equivalences between $k - 1$ variants of a test formulation and a standard reference drug, and (ii) simultaneous equivalences between all pairs among k formulations. He defines null hypotheses of non-equivalence in terms of standardized differences between pairs of means and derives exact IU procedures.

Following a suggestion from Statistics in Medicine in 1995 to write a paper on multivariate assessment of Bioequivalence, Hochberg in collaboration with Shein–Chung Chow started to work on the problem, and soon learned by personal communication with Professor Sanat Sarkar that he had a Ph.D. student already working on the problem. Sarkar wrote: "My student's Ph.D. dissertation title is 'Bivariate extensions of some standard tests for bioequivalence assessment.' His name is Napoleon A. Oleka. He provided bivariate extensions of Schuirmann's two one-sided t-tests (which was later noted to be one that can be derived from the Intersection-Union tests discussed in Berger and Liu's paper in Statistical Science), and of the Anderson-Hauck procedure."

5.2. Other approaches

5.2.1. False discovery rate (FDR)

Benjamini and Hochberg (1995) (hereafter abbreviated as BH) referred to the expected proportion of erroneously rejected null hypotheses among the rejected ones as the FDR. Formally, for a given family (of m null hypotheses) and a given MCP, let R = number of hypotheses rejected by a given MCP, and let V = (unknown) number of erroneously rejected ones. Define $V/R = 0$ in case $R = 0$ and $V = 0$ (since $V \leq R$). The FDR is the expected value of V/R. Incidentally, the concept was recognized by others, see e.g. Seeger (1966, Part III) and his references. The "FDR approach" calls for controlling the FDR in the strong sense.

Note that under the overall null hypothesis H_0, the FDR and FWE are equal, but under other configurations the FDR is always smaller than the FWE. A simple proof is obtained by presenting the FWE as the expected value of the indicator of the event $R \geq 1$, and observing that under H_0, $V/R = 1$ while under all other configurations $V/R \leq 1$.

The FDR and the FWE depend on the unknown parameter m_0 = the number of true null hypotheses. Schweder and Spjøtvoll (1982) recognized the possibility of estimating m_0 from given per-comparison p-values. They presented a graphical method in which ordered p-values are marked on the x-axis with corresponding values of N_p = number of hypotheses with p-values greater than p on the y-axis. They indicated that the expected value of N_p in the region of large p-values is roughly $= m_0(1 - p)$ and that it can be utilized for estimating m_0. Hochberg and Benjamini (1990) modified their graphical procedure (as we discuss in the sequel) and indicated powerful "adaptive" forms of some FWE-controlling MCPs. "Adaptive" here means utilizing a two-step scheme involving a first step in which m_0 is estimated (by $\widehat{m}_0$) followed by a suitable second step which utilizes $\widehat{m}_0$.

Benjamini and Hochberg (1996) proposed an adaptive modification of their original procedure in BH. The original BH procedure was restricted to independent test statistics for which its control of the FDR in the strong sense was proved. Denoting by $P_1, \ldots, P_m$ the p-values corresponding to the (independent) test statistics the BH procedure starts by ordering the p-values into $P_{(1)} < \cdots < P_{(m)}$ and denoting the associated hypotheses by $H_{(1)}, \ldots, H_{(m)}$. The original BH procedure rejects all $H_{(j)}, j = 1, \ldots, J$, where J is the maximal j satisfying $P_{(j)} \leq (j/m)q$ where q is the desired FDR level. BH proved that (under independence) the FDR of this procedure is always $\leq (m_0/m)q$. This is unnecessarily conservative for configurations with $m_0 \ll m$. The modified adaptive procedure involves ordering as above and rejecting when $P_{(j)} \leq (j/\widehat{m}_0)q$, and uses a graphical method for estimating $\widehat{m}_0$.

5.2.2. Non-frequentist approaches

Various non-frequentist approaches (descriptive data-analysis type as well as decision-theoretic) are discussed in HT. Most prominent in the MC literature has been Duncan's (1965) which evolved in various directions since then, see HT and the more recent exposition by Berry and Hochberg (1997), and our discussion in the sequel re its relation with the FDR approach as revealed by Shaffer (1997).

Duncan (1965) noted that there was no structural difference between a per-comparison procedure and simultaneous single-step FWE-controlling procedures (in the context of a one-way layout) since one could make the two equivalent by suitable choice of appropriate levels. In his words: "If a truly appropriate level were chosen in each approach they would give identical procedures." Duncan derived an optimal "Jeffreys'-like Bayesian" procedure for the pairwise comparisons in a one-way layout based on an "additive loss-function." His procedure is structurally quite different from the single-step and per-comparison type, it is in fact similar to Fisher's two-step LSD. In Duncan's words: "A working minimum average-risk procedure is derived which is found to have much the same form as the simple Fisher LSD rule, *but with the LSD determined as a specifically defined function of the between treatments observed F-ratio* ... at F-values ... below 3.0 the Bayes LSD increases slowly at first and with greater rapidity from 2.0 down to 1.0. Here the rule can have the more conservative character of a Tukey ... procedure."

Duncan's procedure is adaptive, i.e. depends on the "Bayes LSD" critical values with which the (usual) two-sample t-statistics will be compared with the value of the F-statistic.

To apply Duncan's procedure one has to specify a constant K which indicates "relative seriousness of type I to type II errors (accordingly Duncan's original and later procedures following his basic approach have been referred to as "K-ratio t-tests"). HT give (Ch. 11) a full discussion of the basic approach and various procedures derived from it.

Shaffer (1997) modifies Duncan's procedure to "eliminate the need for specifying a prior hyperparameter ... and adjust to give control of the familywise error at 0.05 ... when all hypotheses are true."

Shaffer's (1997) modification of Duncan's original procedure is an example of a class of procedures presently referred to as "semi" or "quasi" Bayesian, see e.g. Berry and Hochberg's (1997) recent exposition. These procedures are generally Bayesian (i.e. are originally optimal for some prior distribution of unknown parameters) but in addition are calibrated to meet some frequentist error-rate. Some other such procedures are indicated elsewhere in this article.

Shaffer (1997) found that her modified semi-Bayesian version of Duncan's procedure has very similar operation characteristics to those of a "natural" FDR-type controlling procedure for the pairwise comparisons in a one-way layout. Note that more than one FDR-type procedure can be defined for the pairwise comparisons (e.g. depending on whether the FDR is defined in terms of directional errors and decisions only, or possibly in terms of type-I and type-III errors combined). See Benjamini et al. (1995) and Williams et al. (1994) on such procedures.

5.3. Closing remarks

Shaffer (1994) gave proper credit to Seeger (1966, 1968) who indicated essentially the same criterion as Benjamini and Hochberg (1995). Seeger (1966) devotes part III of his book to "The problem of multiple comparisons." In chapter 1 of that part he indicates that "According to Newman (1939) Tippett's and Student's studies in the 1920's of range-methods may be considered as the predecessors of some of the modern methods of making multiple comparisons. *Their aim seems to have been to replace the F-test by a method that ... gives more detailed information.* These methods have been developed by Newman (1939) and later by Keuls (1952)." Other methods (single-step) were developed to control the FWE in the strong sense (noteworthy, Tukey's *T*-method which was also based on the Studentized range distribution. The *stepdown procedures* (like Newman–Keuls) were eventually calibrated to control the FWE in the strong sense, and then compared to single-step procedures and to other stepwise procedures called *stepup procedures*, in terms of power and other operation characteristics, see e.g. HT. Seeger (1966, Part III, Ch. 1) references earlier work by Eklund and by Eklund and Seeger (published in Swedish) where the proportion of false significances is discussed as a suitable criterion for error in some situations involving "many analyses of significance."

Seeger (1966, 1968) should be also credited for another important contribution, namely, the improved Bonferroni procedure which was derived independently (and differently) by Simes' (1986).

In view of Hochberg's and Hommel's (1997) recent exposition of Simes' test (and of the various step-up procedures and other related literature that emerged from it) we will not discuss these topics here.

As indicated earlier throughout this article, single-step and stepwise procedures have different operational characteristics. These involve power, allowance for directional-decisions, and provision of (simple or not simple) simultaneous confidence intervals. We refer the reader to Hochberg et al. (1997) for a recent expository on the problem of directional decisions associated with stepwise procedures. For basic notions of directional decisions mainly with single-step procedures the reader can consult HT. A recent paper which presents a confidence-estimation procedure following Tukey's (1991) suggestions for a suitable procedure is Benjamini et al. (1997) with recent references concerning the multiple comparisons problem.

For sake of completeness, since the dose-response problem is a classical one both in the MC and Biostatistical literature we refer the reader to Tamhane's et al. (1996) paper in Biometrics 59, 1, 21–37, entitled: "Multiple test procedures for dose finding."

Finally we indicate Weller's et al. (1996) discussion and comparative study of the FWE and FDR approaches to some multiplicity problems arising is Genetics Research.

Acknowledgement

The authors would like to recognize Miriam Hochberg, Dror Rom, Keith Soper, and S. Stanley Young, all of whom have contributed substantially to the scope and content of this article.

References

Aitchison, J. (1964). Confidence-region tests. *J. Roy. Statist. Soc. Ser. B* **26**, 462–476.

Ali, M. W. (1990). Exact versus asymptotic tests of trend of tumor prevalence in tumorigenicity experiments: A comparison of p-values for small frequency of tumors. *Drug Inform. J.* **24**, 727–737.

Bailar, J. C. III (1991). Scientific inferences and environmental health problems. *Chance* Vol. 4, No.2, 27–38.

Bauer, P. (1991). Multiple testing in clinical trials. *Statist. Med.* **10**, 871–890.

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

Benjamini, Y. and Y. Hochberg (1996). On the adaptive control of the false discovery rate in multiple independent testing problems. Accepted for Publication in the *J. Educ. Beh. Statist.*

Benjamini, Y. and Y. Hochberg (1997). Multiple hypotheses testing with weights. *Scan. J. Statist.* **24**, (3) 407–418.

Benjamini, Y., Y. Hochberg and Y. Kling (1995). False discovery rate controlling procedures for pairwise comparisons (submitted).

Benjamini, Y., Y. Hochberg and P. Stark (1997). Confidence intervals with more power to determine the sign: Two ends constrain the means. To appear in *JASA*.

Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *J. Amer. Statist. Assoc.* **82**, 112–122.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling, *Technometrics* **24**, 295–300.

Berry, D. A. and Y. Hochberg (1997). On Bayesian and quasi-Bayesian approaches to multiple comparison problems. *The International Conference on Multiple Comparisons Special Issue*, The Journal of Statistical Planning and Inference.

Breslow, N. (1990). Biostatistics and Bayes. *Statist. Sci.* **5**, 269–298.

Carmer, S. G. and W. M. Walker (1982). "Baby bear's dilemma: A statistical tale." *Agronomy J.* **74**, 122–124.

Chow, S. C. and J. P. Liu (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies.* Marcel Dekker, Inc.

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

Copas, J. B. and T. Long (1991). Estimating the residual variance in orthogonal regression with variable selection. *The Statistician* **40**, 51–59.

Cournot, A. A. (1843). Exposition de la théorie des chances et des probabilités. Pris Hachette (Reprinted 1984 as Vol. 1 of Cournot's Oevres Completètes, Ed., Bernard Bru. Paris: J Vrin.).

Dept. of Health and Human Services (1987). *Guideline for the Format and Content of the Nonclinical/Pharmacology/Toxicology Section of an Application*, U.S. Food and Drug Administration, 5600 Fishers Lane, Rockville, MD 20857.

Dept. of Health and Human Services (1998). *International Conference on Harmonisation; Draft Guideline on Statistical Principles for Clinical Trials.*, U.S. Food and Drug Administration, 5600 Fishers Lane, Rockville, MD 20857; www.fda.gov:80/cder/guidance/ichstatprinc.htm.

Diaconis, P. (1985). Theories of data analysis from magical thinking through classical statistics. In *Exploring Data Tables, Trends and Shapes* (Eds., D. C. Hoaglin, F. Mosteller, and J. W. Tukey), pp. 1–36, New York: Wiley.

Draper, D., D. P. Gaver Jr., P. K. Goel, J. B. Greenhouse, L. V. Hedges, C. N. Morris, J. R. Tucker and C. M. Waternaux (1992). *Contemporary Statistics Number 1: Combining Information.* Amer. Statist. Assoc. National Academy Press, Washington, D.C.

Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171–222.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* **50**, 1096–1121.

Edwards, S., G. G. Koch, W. A. Sollecito and K. E. Peace (1990). Summarization, analysis, and monitoring of adverse experiences. In *Statistical Issues in Drug Research and Development* (Ed., Karl Peace), Marcel Dekker Inc., New York.

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

Farrar, D. B. and K. S. Crump (1988). Exact tests for any carcinogenic effect in animal bioassays. *Fund. Appl. Toxicol.* **11**, 652–663.

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

Fleming, T. R. (1992). Current issues in clinical trials. *Statist. Sci.* **7**, 428–456.

Finney, D. J. (1993). Whither biometry? In *Multiple Comparisons, Selection and Applications in Biometry* (Ed., F.M Hoppe), Marcel Dekker, Inc.

Fisher, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh, UK.

Gabriel, K. R. (1969). Simultaneous test procedures – Some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224–250.

Godfrey, K. (1985). Comparing the means of several groups. *New Engl. J. Med.* **311**, 1450–1456.

Grechanovsky, E. and Y. Hochberg (1998). Closed procedures are better and often admit a shortcut. by *JSPI.*

Greenberg, B. G. (1982). Biostatistics. *Encyclopedia of Statistical Sciences*, Vol. 1, 251–263.

Hall, P. and S. R. Wilson (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762.

Haseman, J. K. (1983). A reexamination of false-positive rates for carcinogenicity studies. *Fund. Appl. Toxicol.* **3**, 334–339.

Haseman, J. K., J. S. Winbush and M. W. O'Donnell (1986). Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. *Fund. Appl. Toxicol.* **7**, 573–584.

Heyse, J. F. and D. Rom (1988). Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biom. J.* **30**, 883–896.

Hochberg, Y. (1996). On assessing multiple equivalences with reference to bioequivalence. In *Statistical Theory and Applications* (papers in honor of H. A. David). (Eds., H. N. Nagaraja, P. K. Sen and D. F. Morrison) pp. 267–276. Springer, Part V: Biometry and Applications.

Hochberg, Y. and Y. Benjamini (1990). More powerful procedures for multiple significance testing. *Statist. Med.* **9**, 811–818.

Hochberg, Y. and G. Hommel (1997). Simes' tests of multiple hypotheses. *Encyclopedia of Statistical Sciences*, to appear.

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

Holm, S. (1979). A Simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

Keuls, M. (1952). The use of the 'studentized range' in connection with an analysis of variance. *Euphytica*, **1**, 112–122.

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

Koch, G. G., G. J. Carr, I. A. Amara, M. E. Stokes and T. J. Uryniak (1990). Categorical data analysis. In *Statistical Methodology in the Pharmaceutical Sciences* (Ed., Donald A. Berry), Marcel Dekker, New York.

Krishnaiah, P. R. (1979). Some developments on simultaneous test procedures, In *Developments in Statistics*, Vol. 2 (Ed., P. R. Krishnaiah), pp. 157–201. Amsterdam: North-Holland.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Ann. Math. Statist.* **32**, 990–1012.

Lin, K. K. and Rahman (1998). Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. *J. Biopharm. Statist.* **8**, 1–15.

Louis, T. A. (1992). Comment on Evaluating therapeutic interventions: Some issues and experiences. *Statist. Sci.* **7**, 450–452.

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

Mehta, C. and N. Patel (1995). *StatXact 3 for Windows*, Cytel Software Corp. Cambridge, 1995.

Miller, R. G. Jr. (1966). *Simultaneous Statist. Inferences*, McGraw-Hill, New York.

Naik, U. D. (1975). Some selection rules for comparing P processes with a standard. *Comm. Statist. Ser. A* **4**, 519–535.

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

Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika* **31**, 20–30.

Nowak, R. (1994). Problems in clinical trials go far beyond misconduct. *Science* **264**, 1538–1541.

O'Neill, R. T. and B. G. Wetheril (1971). The present state of multiple comparisons methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **33**, 218–241.

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

Peto, R., M. Pike, N. Day, R. Gray, P. Lee, S. Parish, J. Peto, S. Richards and J. Wahrendorf (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in long-term animal experiments. *Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal* IARC Monographs, Annex to Supplement 2, 311–426. Lyon: International Agency for Research on Cancer.

Pocock, S. J. (1983). *Clinical Trials – A Practical Approach*, Wiley.

Putter, J. (1983). Multiple comparisons in selected inferences. In *A Festschrift for E.L. Lehman* (Eds., P. J. Bickel, K. Doksum and J. L. Hodges, Jr.), pp. 328–347, Belmont CA: Wadsworth.

Reitmeir, P. and G. Wassmer (1996). One-sided multiple endpoint testing in two-sample comparisons. *Comm. Statist.: Simul. and Comput.* **25**, 99–117.

Rom, D. M. (1992). Strengthening some common multiple test procedures for discrete data. *Statist. Med.* **11**, 511–514.

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

Roy, S. N. and R. C. Bose (1953). Simultaneous confidence interval estimation. *Ann. Math. Statist.* **24**, 220–238.

SAS Institute (1996). *SAS/STAT® Software: Changes and Enhancements through Release 6.11*, Cary, NC: SAS Institute Inc.

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

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Seeger, P. (1966). *Variance Analysis of Complete Designs*, Almquist & Wiksell, Stockholm.

Seeger, P. (1968). A note on a method for the analysis of significance en masse. *Technometrics* **10** (3), 586–593.

Shafer, G. and I. Olkin (1983). Adjusting P values to account for selection over dichotomies. *J. Amer. Statist. Assoc.* **78**, 674–678.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81**, 826–831.

Shaffer, J. P. (1994). Multiple hypothesis testing: A review. Tech. Rep. #23 National Institute of Statistical Sciences, P.O. Box 14162 Research Triangle Park, N.C. 27709, U.S.A.

Shaffer, J. (1995). Multiple hypotheses testing. *Annu. Rev. Psychology* **46**, 561–584.

Shaffer, J. P. (1997). A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *The International Conference on Multiple Comparisons Special issue. J. Statist. Plan. Inf.*

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62**, 626–633.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

Simon, R. (1994). Problems of multiplicity in clinical trials. *J. Statist. Plan. Inf.* **42**, 209–221.

Soper, K. A. and N. Tonkonoh (1993). The discrete distribution used for the log-rank test can be inaccurate. *Biom. J.* **35**, 291–298.

Soper, K. A. and P. H. Westfall (1990). Monte Carlo estimation of significance levels for carcinogenicity tests using univariate and multivariate models. *J. Statist. Comp. Sim.* **37**, 189–209.

Sterling, T. D., W. L. Rosenbaum, and J. J. Weinkam (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The Amer. Statistician* **49** (1), 108–112.

Stigler, S. M. (1986). *The History of Statistics*. Harvard University Press: Cambridge.

Tamhane, A. C. (1996). Multiple comparisons. In *Handbook of Statistics* (Eds., S. Ghosh and C. R. Rao), Vol. 13, pp. 587–630.

Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics* **46**, 515–522.

Troendle, J. F. (1995). A stepwise resampling method of multiple hypothesis testing. *J. Amer. Statist. Assoc.* **90**, 370–378.

Tukey, J. W. (1953). The problem of multiple comparisons. *Mimeographed Notes*, Princeton University.

Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science* **198**, 679–684.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statist. Sci.* **6**, 100–116.

Weller, J. I., J. Z. Song, Y. I. Ronin and A. B. Korol (1996). Experimental designs and solutions to multiple trait comparisons. International Conference on Multiple Comparisons, Tel Aviv, Israel.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *J. Amer. Statist. Assoc.* **92**, 299–306.

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

Westfall, P. H. and K. A. Soper (1998). Weighted multiplicity adjustments for animal carcinogenicity tests. *J. Biopharm. Statist.* **8**, 23–44.

Westfall, P. H. and R. D. Wolfinger (1997). Multiple tests with discrete distributions. *The Amer. Statistician* **51**, 3–8.

Westfall, P. H. and S. S. Young (1989). *P*-value adjustments for multiple tests in multivariate binomial models. *J. Amer. Statist. Assoc.* **84**, 780–786.

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

Williams, V. S. L., L. V. Jones and J. W. Tukey (1994). Controlling error in multiple comparisons, with special attention to the trial state assessment of educational progress. Technical report, The National Institute of Statistical Sciences, Research Triangle Park, NC.

Zelen, M. (1983). Biostatistical science as a discipline: A look into the future. *Biometrics* **39**, 827–837.