Multiple Test Procedures for Identifying the Maximum Safe Dose

Ajit C. TAMHANE, Charles W. DUNNETT, John W. GREEN, and Jeffrey D. WETHERINGTON

We consider dose response studies for safety assessment of crop protection compounds and drugs, and we offer a hypothesis testing approach for identifying the maximum dose level that is guaranteed to be safe with preassigned confidence. The focus is on step-down (SD) multiple test procedures for identifying the maximum safe dose. We propose two classes of contrasts among the dose means as test statistics for these procedures: pairwise contrasts (PC) and Helmert contrasts (HC). The first procedure (SD2PC) consists of a sequence of ordinary t tests and is thus easy to apply, but it can have very low power for certain step response functions. The second procedure (SD1HC) does not suffer from this drawback, but it requires a weak monotonicity assumption for its mathematical validity. The powers of SD2PC and SD1HC are studied via Monte Carlo simulation. We recommend SD2PC for linear response functions and SD1HC for step response functions. The procedures are illustrated by applying them to data from an aquatic toxicity laboratory experiment conducted to assess the safe level of a pesticide.

KEY WORDS: Dose response; Familywise error rate; Minimum unsafe dose; Multiple comparisons; Multivariate t-distribution; Step-down multiple testing procedure; Toxicology.

1. INTRODUCTION

Crop protection products such as pesticides, herbicides, and fungicides are intended to be harmful to certain targeted species, but they should be safe for nontargeted species. A thorough screening is done by conducting an extensive battery of tests to assess the safety of these products. From the test results, the company hoping to market the product must determine the conditions under which the product is safe to use, and the company's assessment must be approved by the regulatory agencies. Statistical analysis of the data is an important part of this process for both the company and the regulator. At present, in many situations, the precise statistical methods that must be used are not agreed on. Rather, it is left to the company and regulatory statisticians to select and defend appropriate methods. In addition, toxicologists must weigh the evidence and come to an overall assessment of safe levels of product use. Similar safety issues arise in drug testing.

Traditionally, safe levels have been determined from experimental data through hypothesis testing. The no adverse effects level (NOAEL) is defined as the highest dose or concentration of a chemical that induces no significant adverse effect (Crump 1983; Gaylor 1983; Yanagawa, Kikuchi, and Brown 1997). Tukey, Ciminera, and Heyse (1985) employed a similar approach to identify the no statistical significance of trend (NOSTASOT) dose. A criticism of this hypothesis testing approach is that smaller and less sensitive experiments result in higher doses being declared safe, which is the opposite of what is desired. Another criticism is that it tests the null hypothesis of no effect (a common practice in carcinogenicity

and toxicity studies) instead of a positive effect that is deemed to be biologically safe.

There is a school of practitioners who prefer to use regression methods; see, e.g., Crump (1983). On the basis of a fitted model, a dose is estimated that produces a change in the mean response of a specified percent (e.g., 5, 10, or 20%) compared to the zero dose. Although such an approach has some advantages, routine industrial experiments are too small to allow fitting a realistic regression model, especially in the tails of the dose response function. The expense involved in conducting sufficiently large industrial experiments is prohibitive. We have encountered toxicological datasets for which several non-linear regression models fit equally well and yet give widely different estimates of the safe dose to the extent that their 95% confidence intervals are disjoint. Furthermore, biological reactions to chemicals are not sufficiently well understood to postulate an appropriate class of regression models.

We offer a hypothesis testing approach that meets the criticisms of the conventional hypothesis testing approach by making the modifications given in items 1 and 2 in the following list. Although our approach does not entail any model fitting, it nevertheless captures the spirit of the regression approach, as explained in item 3. To summarize, the following are the features of our approach and their resulting advantages.

1. Hypotheses are set up so that the burden of proof of safety is on the data. New multiple test procedures are proposed to find the maximum safe dose (MAXSD). These procedures control the probability of declaring any unsafe dose as safe and thus protect the consumer's risk.
2. Instead of testing for a zero difference from the control mean, a threshold of safety equal to a specified fraction λ of the control mean is tested (assuming that a smaller mean response represents more toxicity). It is a common industry practice to regard a dose level as safe if the corresponding expected response is greater than a specified fraction λ of the expected control response. The choice of λ encourages industry and regulatory agencies

to come to an agreement on what level of adverse effect is acceptable.

3. In contrast to the conventional regression approach, our approach does not assume a precise mathematical dose response relationship. However, for test statistics, it uses contrasts of the dose means that can be chosen to mimic the likely shape of the dose response function.

The outline of the paper is as follows. Section 2 describes a practical example encountered in the toxicological testing laboratories of the DuPont Company. Section 3 gives the assumptions and notation. Section 4 describes the problem formulation and the test procedures for identifying the MAXSD. Section 5 describes a simulation study to compare the powers of competing procedures. Section 6 outlines the extensions to unbalanced data. In Section 7, we return to the example from Section 2 and illustrate how the procedures are applied. The FORTRAN programs needed to perform calculations for the proposed test procedures have been made available at the web site http://lib.stat.cmu.edu/general. We conclude with a discussion in Section 8.

2. EXAMPLE: TOXICOLOGICAL EVALUATION OF A PESTICIDE FOR ENVIRONMENTAL IMPACT

To determine the effect of run-off on aquatic life, one type of experiment conducted is the 21-day chronic exposure of the tiny freshwater crustacean *Daphnia magna* (commonly referred to as daphnids) to the compound. Daphnids are water fleas that are used not only to test the toxicity of known toxicants (such as pesticides) but also to test the sensitivity of the whole effluents when the composition of toxicants in the effluents is difficult to determine analytically. They are selected in part because of their sensitivity to chemicals.

Daphnids of the same age (essentially neonates) and genetic stock are randomly assigned to a water control, a solvent control, or one of six concentrations of a pesticide, which for the purpose of this discussion will be referred to as NoPest. Although many biological endpoints are measured in such experiments, we examine only the growth, as measured by the lengths of the daphnids after 21 days of continuous exposure. Because this compound is not highly soluble in water, a standard nontoxic solvent is added to the water to increase its solubility. In a few instances, the solvent itself unexpectedly appears to affect the growth of the daphnids, so both a water control and a solvent control are included. A preliminary statistical test is done to compare these two controls. If no significant difference is found, the controls are combined for further testing to increase power; otherwise, only the solvent control is used in subsequent statistical tests. Initial lengths of daphnids are impossible to measure with the current experimental equipment. However, care is taken to ensure that none is of unusual size or appearance. Final lengths are measured after the 21-day exposure period when the daphnids are sacrificed.

Six nominal concentrations of NoPest were tested: .3125, 6.25, 12.5, 25, 50, and 100 ppm. (These concentrations are in the same proportion to the actual concentrations used, but they are coded for proprietary reasons.) Forty daphnids were randomly assigned to each of the two control and six treatment groups. The daphnids in each group were randomly divided into subgroups of four each and were placed in 10 different tanks (a total of 80 tanks) with controlled concentrations. Over the course of the experiment, some daphnids died, so that the sample sizes available at the end were 40 in each control group and 38, 39, 35, 35, 33, and 4 in the six treatment groups. Note the increased mortality with dose, especially the highest dose.

Lethal as well as sublethal effects of NoPest are of interest. However, a simultaneous analysis of both effects poses formidable statistical difficulties because the deceased daphnids are informatively censored, and there is no simple way to take this into account. There are no predeath length measurements available, and the biological mechanism causing death (e.g., lack of adequate growth or toxic effects on some body organs) is not well understood. To obviate these difficulties, Capizzi et al. (1985) recommended a two-stage approach for the design and analysis of such aquatic toxicity studies that examines survival first and then evaluates sublethal effects of those concentrations that do not significantly affect survival. The goal in this experiment was to evaluate only the sublethal effect on growth. Therefore, the 100 ppm concentration group was omitted from further analysis because of the high (90%) mortality. For the same reason, the modest level of mortality (2.5% to 17.5%) at the concentrations of interest was ignored in the present analysis, and the observed sample sizes were regarded as fixed.

The daphnid length data were coded for proprietary reasons without affecting their statistical properties. The summary statistics are shown in Table 1, and the side-by-side box plots are shown in Figure 1. The outliers revealed in the box plots were left in the data because no biological justification could be found to reject those observations; also, no essential differences were found in the results when we analyzed the data with and without the outliers. The normality assumption was verified by making a normal plot of the residuals and the Shapiro–Wilk test. The Levene test for homogeneity of variances yielded nonsignificant result ($p = .560$).

An analysis of variance (ANOVA) for the data obtained by treating the tanks as random and nested inside the treatment groups is shown in Table 2. We see that the tank-to-tank variation is not statistically significant. Therefore, we pooled the corresponding sum of squares with the error sum of squares, resulting in a mean square error equal to .0302 with 253 degrees of freedom.

The difference between the water and solvent control groups is not statistically significant ($t = .265$); therefore, in subsequent analysis, these two groups are combined to yield a

Table 1. Summary Statistics for the Daphnid Length Data

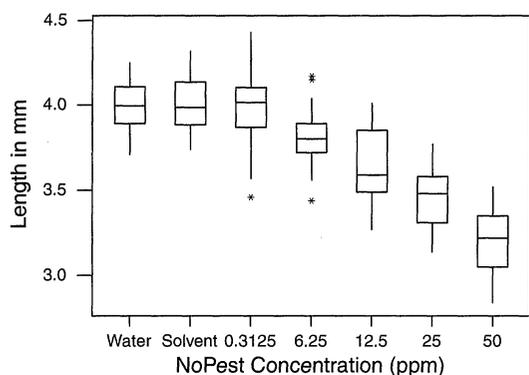| Dose level	NoPest concentration (ppm)	Sample size	Mean length (mm)	Std. dev. (mm)
0	Water control	40	3.9952	.1519
0	Solvent control	40	4.0055	.1472
1	0.3125	38	3.9908	.2110
2	6.25	39	3.8108	.1504
3	12.5	35	3.6306	.1961
4	25.0	35	3.4600	.1726
5	50.0	33	3.2106	.1829

Figure 1. Side-by-Side Box Plots of the Daphnid Length Data.

control group of 80 daphnids. Compared to the combined control mean of 4.000 mm, the treatment groups show decreases in mean length of .24%, 4.74%, 9.24%, 13.51%, and 19.74%, respectively. In the toxicology community, opinions about what constitutes a biologically significant effect have ranged between 5% and 25% adverse effect. If we take an average of this range, i.e., 15% or $\lambda = .85$, as biologically unsafe, then we would like to know which dose is MAXSD for this value of λ.

3. NOTATION AND ASSUMPTIONS

Denote a set of increasing dose levels by $0, 1, 2, \ldots, k$, where 0 denotes the zero dose control. Consider a one-way layout in which n_i experimental units are tested at the ith dose level $(i = 0, 1, \ldots, k)$. Let μ_i be the mean response at the ith dose level, which is estimated by the corresponding sample mean $\bar{y}_i$. We assume that the $\bar{y}_i$ are mutually independent with $\bar{y}_i \sim N(\mu_i, \sigma^2/n_i)$ for $i = 0, 1, \ldots, k$, where σ^2 is the experimental error variance. Let s^2 be an unbiased estimate of σ^2 based on ν degrees of freedom such that $\nu s^2/\sigma^2$ is distributed as χ^2_ν independently of the $\bar{y}_i$. Usually, s^2 is the ANOVA mean square error with $\nu = \sum_{i=0}^{k} n_i - (k+1)$ degrees of freedom. Through Section 6, we assume balanced data with $n_i = n$ for $i = 1, 2, \ldots, k$ with n_0 possibly different from n. The ratio $r = n_0/n$ is generally greater than or equal to 1.

For a specified fraction λ, the ith dose is regarded as safe if $\mu_i > \lambda\mu_0$ and as unsafe if $\mu_i \leq \lambda\mu_0$. The MAXSD and the minimum unsafe dose (MINUD) are defined as follows:

$$\text{MAXSD}\lambda = \max\{i : \mu_i > \lambda\mu_0\} \quad \text{and}$$

$$\text{MINUD}\lambda = \text{MAXSD}\lambda + 1 = \min\{i : \mu_i \leq \lambda\mu_0\}. \quad (3.1)$$

Table 2. Analysis of Variance of Daphnid Lengths

Source	SS	d.f.	MS	F	p-value
Treatment	18.8110	6	3.1352	103.09	.000
Tank (group)	1.9166	63	.0304	1.01	.468
Error	5.7265	190	.0301		
Total	28.0501	259			

NOTE: SS = sum of squares. d.f. = degrees of freedom. MS = mean square error. F = MS/d.f.

If all doses are unsafe, then we define $\text{MAXSD}\lambda = 0$; similarly, if all doses are safe, then we define $\text{MINUD}\lambda = k + 1$.

We assume that for specified λ,

$$\mu_i > \lambda\mu_0 \quad \text{for all} \quad i \leq \text{MAXSD}\lambda \quad \text{and}$$

$$\mu_i \leq \lambda\mu_0 \quad \text{for all} \quad i > \text{MAXSD}\lambda. \quad (3.2)$$

This assumption implies that any dose less than or equal to MAXSDλ is safe and any dose greater than MAXSDλ is unsafe. Unfortunately, there is no statistical test to check assumption (3.2). The stronger assumption of monotonicity,

$$\mu_0 \geq \mu_1 \geq \cdots \geq \mu_k \quad (3.3)$$

with at least one strict inequality, is easier to check by using, e.g., Bartholomew's (1959) test for ordered means as modified by Chase (1974) for increased sample size in the control group. Implications of violations of assumption (3.2) are discussed in Section 8. We refer to (3.2) as the weak monotonicity assumption and to (3.3) as the strong monotonicity assumption.

4. IDENTIFYING THE MAXIMUM SAFE DOSE

4.1 Problem Formulation

Consider the hypotheses

$$H_{0i} : \mu_i \leq \lambda\mu_0 \quad \text{versus} \quad H_{1i} : \mu_i > \lambda\mu_0,$$

$$i = 1, 2, \ldots, k. \quad (4.1)$$

Thus, a dose is assumed unsafe unless proved safe. Because of assumption (3.2), it follows that $\{H_{0i}, i = 1, \ldots, k\}$ is a closed family. In fact, we can write (4.1) as

$$H_{0i} = \bigcap_{j=i}^{k} H_{0j} \quad \text{versus} \quad H_{1i} = \bigcup_{j=i}^{k} H_{1j}, \quad i = 1, 2, \ldots, k. \quad (4.2)$$

This representation is useful for constructing step-down test procedures discussed in Section 4.2.

We restrict to multiple test procedures that strongly control the type I familywise error rate (FWE) for this family of hypotheses; specifically, we require

$$\text{FWE} = P\{\text{Any true } H_{0i} \text{ is rejected}\}$$

$$= P\{\text{Any unsafe dose is declared safe}\} \leq \alpha, \quad (4.3)$$

where α is a specified overall significance level. We define the sample MAXSDλ (denoted by $\widehat{\text{MAXSD}\lambda}$) as the highest dose for which $\mu_i > \lambda\mu_0$ can be demonstrated, i.e., $\widehat{\text{MAXSD}\lambda} = \max\{i : H_{0i} \text{ is rejected}\}$, using a test procedure that satisfies (4.3).

4.2 A Single-Step Test Procedure

A simple way to test the hypotheses (4.1) is to use Dunnett's (1955) simultaneous confidence interval procedure suitably modified for the present problem. This single-step procedure uses the test statistics based on pairwise contrasts (PC):

$$t_i = \frac{\bar{y}_i - \lambda \bar{y}_0}{s\sqrt{\lambda^2/n_0 + 1/n}}, \qquad i = 1, 2, \ldots, k. \qquad (4.4)$$

It rejects H_{0i} if $t_i > t_{k,\nu,\rho}^{(\alpha)}$ $(i = 1, 2, \ldots, k)$, where $t_{k,\nu,\rho}^{(\alpha)}$ is the upper α equicoordinate critical constant of the k-variate central t-distribution with ν degrees of freedom and common correlation

$$\rho = \frac{\lambda^2}{\lambda^2 + r}. \qquad (4.5)$$

We refer to this as the SSPC procedure. Single-step procedures based on other contrasts are also possible. A comprehensive tabulation of the critical constants $t_{k,\nu,\rho}^{(\alpha)}$ is given in Bechhofer and Dunnett (1988).

4.3 Step-Down Test Procedures

Single-step procedures are simple to use, but they are less powerful than stepwise test procedures. A step-down procedure for MAXSDλ tests the hypotheses (4.1) in the order $H_{01}, H_{02}, \ldots, H_{0k}$. Each hypothesis is tested at level α and the procedure stops the first time a hypothesis is not rejected, accepting the remaining hypotheses by implication without actually testing them. In other words, continue testing until a dose cannot be shown to be safe. If H_{0i} is the last rejected hypothesis, then dose i is declared as the $\widehat{\text{MAXSD}}\lambda$. This procedure controls the FWE (4.3) at level α because it is a closed testing procedure (Marcus, Peritz, and Gabriel 1976). This follows from the fact that the family of hypotheses $\{H_{0i}, i = 1, \ldots, k\}$ of (4.2) is closed.

Remark 1. It is possible to design step-up procedures (Dunnett and Tamhane 1992) for identifying the MAXSDλ. A step-up procedure would test the hypotheses (4.2) in the order $H_{0k}, H_{0,k-1}, \ldots, H_{01}$, stopping the first time a hypothesis is rejected, rejecting the remaining hypotheses by implication, and declaring the corresponding dose level as the $\widehat{\text{MAXSD}}\lambda$. We did not consider step-up procedures because their power gains over step-down procedures are marginal. Also, it is much more difficult to compute their critical constants or the adjusted p-values.

4.3.1 Pairwise Contrasts. We describe two step-down procedures, SD1PC and SD2PC, which use the t-statistics defined in (4.4) for pairwise contrasts. They correspond to SD1 and SD2 procedures given by Tamhane, Hochberg, and Dunnett (1996).

SD1PC Procedure. SD1PC uses a union-intersection test of $H_{0i} = \bigcap_{j=i}^{k} H_{0j}$, where the test statistic for H_{0j} is t_j given by (4.4). Thus, it rejects H_{0i} at level α iff $H_{01}, \ldots, H_{0,i-1}$ are rejected and

$$t_{\max, i} = \max_{i \leq j \leq k} t_j > t_{\ell, \nu, \rho}^{(\alpha)},$$

where $\ell = k - i + 1$. An equivalent shortcut version of SD1PC rejects H_{0m}, where $m \geq i$ is the highest dose level such that

$t_m > t_{\ell, \nu, \rho}^{(\alpha)}$. It is obvious that SD1PC is uniformly more powerful than SSPC, because $t_{\ell, \nu, \rho}^{(\alpha)} < t_{k, \nu, \rho}^{(\alpha)}$ for $\ell < k$.

SD2PC Procedure. SD2PC uses ordinary α-level Student t-tests of the hypotheses H_{0i}. It rejects H_{0i} iff $H_{01}, \ldots, H_{0,i-1}$ are rejected and $t_i > t_\nu^{(\alpha)}$, where $t_\nu^{(\alpha)}$ is the upper α critical constant of the univariate Student's t-distribution with ν degrees of freedom. Thus $\widehat{\text{MAXSD}}\lambda$ is the highest dose level i for which $\min_{1 \leq j \leq i} t_j > t_\nu^{(\alpha)}$. Note that SD2PC does not employ any multiplicity adjustment. It still controls the FWE because of the a priori ordering of the hypotheses. See Maurer, Hothorn, and Lehmacher (1995) and Hsu and Berger (1999) for explanations of why no multiplicity adjustment is needed for testing a priori ordered hypotheses. An advantage of SD2PC is that it is valid without assumption (3.3), and it can be easily applied to dichotomous survival (binomial) data and reproduction (Poisson) data, which are two other common endpoints used in daphnid studies.

4.3.2 General Contrasts. Pairwise contrasts regard the dose levels as nominal treatments. One way to take into account the underlying dose response function, which provides a continuum across different dose levels, is to use more general contrasts as test statistics. By choosing the coefficients of the contrasts to mimic the shape of the dose response function, a more powerful test procedure can be devised. However, specific knowledge about the shape of the dose response function is usually lacking.

To test $H_{0i} = \bigcap_{j=i}^{k} H_{0j}$, we again use a union-intersection test, where for a test statistic of H_{0j} we propose to use the contrast

$$C_{ij} = (\bar{y}_i + \cdots + \bar{y}_j) - (j - i + 1)\lambda \bar{y}_0. \qquad (4.6)$$

This contrast compares the average of $\bar{y}_i$ through $\bar{y}_j$ with $\lambda \bar{y}_0$; in particular, C_{ii} is a pairwise contrast that compares $\bar{y}_i$ with $\lambda \bar{y}_0$. Other contrasts are possible and may be preferable in certain situations. We study this contrast in detail because it has good power properties for a class of configurations that occur in practice, and it provides an illustration of how general contrasts and the associated test procedures can be applied. For examples of some general contrasts, see Mukerjee, Robertson and Wright (1987), Ruberg (1989), and Rom, Costello, and Connell (1994).

The contrasts (4.6) are similar to the Helmert contrasts studied in Tamhane, Dunnett, and Hochberg (1996), where the focus was on finding the minimum effective dose and step-down procedures tested doses for efficacy starting with the highest dose. Therefore, the Helmert contrasts compared the mean of the highest dose under test with the average of the means of the lower doses. Here the focus is on finding the MAXSD, and the contrasts defined in (4.6) compare a specified fraction of the zero dose mean with the average of the doses that are not yet shown to be safe.

The t-statistic for testing the significance of C_{ij} is

$$t_{ij} = \frac{C_{ij}}{\text{s.e.}(C_{ij})},$$

where the standard error (s.e.) of C_{ij} is given by

$$\text{s.e.}(C_{ij}) = s\sqrt{\frac{(j-i+1)}{n_0}[(j-i+1)\lambda^2 + r]}. \tag{4.7}$$

The α-level UI test of H_{0i} rejects if

$$t_{i,\max} = \max_{i \le j \le k} t_{ij} > t_{\ell,\nu,R_\ell}^{(\alpha)} \tag{4.8}$$

where for $\ell = k - i + 1$, $t_{\ell,\nu,R_\ell}^{(\alpha)}$ is the upper α equicoordinate critical constant of an ℓ-variate central t-distribution with ν degrees of freedom and correlation matrix $R_\ell = \{\rho_{jj'}^{(\ell)}\}$, where

$$\rho_{jj'}^{(\ell)} = \text{corr}(C_{ij}, C_{ij'})$$
$$= \sqrt{\frac{(j-i+1)[(j'-i+1)\lambda^2 + r]}{(j'-i+1)[(j-i+1)\lambda^2 + r]}} \quad (i \le j < j' \le k). \tag{4.9}$$

We refer to the resulting SD procedure as the SD1HC procedure. Note that whereas SD1PC uses the same contrasts and hence the same t-statistics, t_j, in all steps of testing, SD1HC uses different contrasts (but from the same family) and hence different t-statistics, t_{ij}. The consequence of using different contrasts is that a shortcut version of SD1HC (like the one of SD1PC) is not possible.

To see that (4.8) gives an α-level test of H_{0i}, note that the t_{ij} for $j = i, \ldots, k$ have in general an ℓ-variate noncentral t-distribution with ν degrees of freedom and correlation matrix R_ℓ. The noncentrality parameters are proportional to $E(C_{ij})$. Because $\mu_j \le \lambda\mu_0$ for all $j = i, \ldots, k$ under H_{0i} because of the assumption (3.2), we have

$$E(C_{ij}) = (\mu_i + \cdots + \mu_j) - (j - i + 1)\lambda\mu_0 \le 0.$$

Hence, the noncentrality parameters are ≤ 0 with equality attained under the least favorable configuration $\lambda\mu_0 = \mu_i = \cdots = \mu_k$, where the type I error probability of the test (4.8) is maximum $= \alpha$. Note that this argument makes no use of the monotonicity assumption (3.3), which is required, as shown by Bauer (1997), to show the error rate control property of the procedures studied in Tamhane, Hochberg, and Dunnett (1996) for identifying the minimum effective dose.

The critical constants $t_{\ell,\nu,R_\ell}^{(\alpha)}$ can be computed by using Schervish's (1984, alg. AS195) algorithm or a simulation-based algorithm of Genz and Bretz (1999). A slightly conservative approximation (Hochberg and Tamhane 1987, p. 146) to these critical constants can be more rapidly computed by replacing the unequal correlations $\rho_{jj'}^{(\ell)}$ by their arithmetic average and by using Dunnett's (1989, alg. AS251) algorithm for product-correlated (Hochberg and Tamhane 1987, p. 365) multivariate t. This approximation is used in our program, but all calculations in this article were done by using the critical constants computed by Schervish's method.

4.4 Multiplicity Adjusted p-Values

All the above test procedures can be implemented based on their multiplicity adjusted p-values, $\tilde{p}_i$, for each hypothesis

H_{0i} as follows. Let p_i denote the unadjusted p-value, which is the probability that the test statistic for H_{0i} is greater than or equal to its observed value when $\lambda\mu_0 = \mu_i = \cdots = \mu_k$. For example, p_i for the SD1HC procedure is the probability that the maximal component of an ℓ-variate central t-distribution with ν degrees of freedom and correlation matrix R_ℓ exceeds the observed value of $t_{\max,i}$. Then, the multiplicity adjusted p-values are given by

$$\tilde{p}_1 = p_1 \quad \text{and} \quad \tilde{p}_i = \max\{\tilde{p}_{i-1}, p_i\}, \qquad i = 2, \ldots, k.$$

Note that $\tilde{p}_1 \le \tilde{p}_2 \le \cdots \le \tilde{p}_k$. Then, the sample MAXSD$\lambda$ is given by MAXSD$\lambda = \max\{i : \tilde{p}_i < \alpha\}$.

5. POWER SIMULATIONS

In this section, we study the powers of SD1PC, SD2PC, and SD1HC for MAXSD via Monte Carlo simulation; the SSPC procedure is not included because it is uniformly less powerful than the SD1PC procedure. Consider a parameter configuration $(\mu_0, \mu_1, \ldots, \mu_k)$ with MAXSD$\lambda = m$ for some $m = 0, \ldots, k$. For any procedure, we have

$$P(\widehat{\text{MAXSD}\lambda} < m) + P(\widehat{\text{MAXSD}\lambda} = m)$$
$$+ P(\widehat{\text{MAXSD}\lambda} > m) = 1.$$

The third term, $P(\widehat{\text{MAXSD}\lambda} > m)$, is the FWE of the procedure. We define the second term, $P(\widehat{\text{MAXSD}\lambda} = m)$, as the power of the procedure. The first term, $P(\widehat{\text{MAXSD}\lambda} < m)$, represents the type II error probability of declaring some safe doses as unsafe. From the above equation, we see that if $m = k$ (all doses are safe), then FWE = 0, and if $m = 0$ (all doses are unsafe), then $P(\widehat{\text{MAXSD}\lambda} < m) = 0$ so that power = $1 -$ FWE $\ge 1 - \alpha$.

The powers and FWE's of SD1PC, SD2PC, and SD1HC were simulated for step and linear configurations of dose means.

Step Configuration. This configuration is defined by the following equation:

$$\mu_0 = \mu_1 = \cdots = \mu_\ell, \qquad (\lambda + \delta)\mu_0 = \mu_{\ell+1} = \cdots = \mu_m,$$
$$(\lambda - \delta)\mu_0 = \mu_{m+1} = \cdots = \mu_k \tag{5.1}$$

for $\ell = 0, 1, \ldots, k-1$ and $m = \ell + 1, \ldots, k$, where $\delta > 0$ is such that $1 - \delta > \lambda > \delta$. Note that for this configuration MAXSD$\lambda = m$, MINUD$\lambda = m + 1$ and 2δ represents the separation between the means corresponding to MAXSDλ and MINUDλ. The threshold for safety, namely, $\lambda\mu_0$, is midway between the means of MAXSDλ and MINUDλ.

Linear Configuration. Two classes of linear configurations were investigated:

$$\mu_1 = \cdots = \mu_{m-1} = (\lambda + 2\delta)\mu_0, \mu_m = (\lambda + \delta)\mu_0, \mu_{m+1}$$
$$= \lambda\mu_0, \mu_{j+1} - \mu_j = 2\delta\mu_0 \quad \text{for } j \ge m+1 \tag{5.2}$$

and

$$\mu_1 = \cdots = \mu_{m-1} = (\lambda + 2\delta)\mu_0, \mu_m = (\lambda + \delta)\mu_0, \mu_{j+1} - \mu_j$$
$$= 4\delta\mu_0 \quad \text{for } j \ge m, \tag{5.3}$$

Table 3. Multivariate t Critical Constants for $\nu = 50$, $\alpha = .05$, $\lambda = .85$, and $r = n_0/n = 2$

Procedure	Dimension ℓ of multivariate t				
	1	2	3	4	5
SD1PC	1.676	1.986	2.156	2.271	2.358
SD1HC	1.676	1.887	1.984	2.041	2.079

where $\delta > 0$ is such that $1 - 2\delta > \lambda$. These two configurations imitate situations where there is some toxicity at low doses and a linear drop in the means (increase in toxicity) beyond MAXSDλ. For each value of m, the configuration (5.3) has a steeper drop than does (5.2).

We did not consider umbrella-shaped response functions (toxicity first increases and then decreases, or vice versa) because although they do occur in efficacy studies, they are uncommon in safety studies.

Under both step and linear configurations, the noncentrality parameters of the t-distributions for different procedures depend on $(\mu_0/\sigma)\sqrt{n}$, λ, δ, $r = n_0/n$, k, ℓ and m. Thus, the powers of the procedures depend on these quantities in addition to α and ν. We performed simulations for $k = 5$, $\alpha = 0.05$, $\lambda = 0.85$, $\sigma = 1$, $\mu_0 = 10$ with $\delta = 0.5$ and $\mu_0 = 20$ with $\delta = 1.0$, $n = 8$, and $r = 2$ (which yields $\nu = 50$ degrees of freedom assuming a one-way layout design). The multivariate t critical constants for the simulated procedures are listed in Table 3. SD2PC uses a single critical constant, which is $t_{50}^{(.05)} = 1.676$.

Each simulation run was based on 50,000 simulations. The results for procedures for identifying MAXSD.85 are reported in Table 4 for $\mu_0 = 10$ and in Table 5 for $\mu_0 = 20$. Only the

powers of the procedures are given in these tables, because all procedures control the type I FWE at level $\alpha = .05$.

The following observations may be made based on the simulation results summarized in Tables 4 and 5.

1. First focus on step configurations. We find that generally SD1HC is the most powerful procedure. This is explained by the fact that Helmert contrasts, which compare $\lambda \bar{y}_0$ with simple averages of higher dose sample means, imitate the step configuration profile. The only exceptions are cases where all lower doses preceding MAXSD have the mean $= \mu_0$, i.e., $\ell + 1 = m$, in the configuration (5.1); in these cases, SD2PC generally has the highest power. However, the power of SD2PC drops dramatically in other cases. On the other hand, the power of SD1HC is more stable. The power of SD1PC is also stable, but it is always less than that of SD1HC.

2. Next focus on linear configurations. Here SD2PC is generally the most powerful procedure. There are two configurations (for $\mu_0/\sigma = 10$ and MAXSD$\lambda = 4$) where SD1HC is most powerful; however, those configurations resemble a step profile more than a linear profile because MAXSD is the second highest dose, so there is only a single drop in the mean after MAXSD. Once again, the power of SD1PC is always less than that of SD1HC.

3. SD1PC is not recommended because it is uniformly dominated by SD1HC in terms of power. For step configurations involving relatively modest increases in toxicity at higher doses, SD1HC is preferred of the three procedures. For linear configurations involving steep drops in means at higher doses, SD2PC is preferred. Note that for linear configurations, we could use either modified basin contrasts (Dunnett and Tamhane 1998) or linear contrasts (Tamhane, Hochberg, and Dunnett 1998; Rom, Costello, and Connell 1994) to achieve higher power. We did not investigate these contrasts because

Table 4. Simulated Powers of SD1PC, SD2PC, and SD1HC for Identifying MAXSDλ When $\lambda = .85$, $\mu_0 = 10$, $\mu_{MAXSD} = 9$, $\sigma = 1$, $n_0 = 16$, $n = 8$, $\nu = 50$, and $\alpha = .05$

Configuration	MAXSD	$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$	SD1PC	SD2PC	SD1HC
Step	1	(9.0, 8.0, 8.0, 8.0, 8.0)	.128	.320*	.194
	2	(9.0, 9.0, 8.0, 8.0, 8.0)	.137	.142	.166*
	2	(10.0, 9.0, 8.0, 8.0, 8.0)	.152	.323*	.211
	3	(9.0, 9.0, 9.0, 8.0, 8.0)	.145	.074	.162*
	3	(10.0, 9.0, 9.0, 8.0, 8.0)	.158	.144	.179*
	3	(10.0, 10.0, 9.0, 8.0, 8.0)	.178	.320*	.229
	4	(9.0, 9.0, 9.0, 9.0, 8.0)	.159	.043	.168*
	4	(10.0, 9.0, 9.0, 9.0, 8.0)	.170	.075	.179*
	4	(10.0, 10.0, 9.0, 9.0, 8.0)	.187	.142	.199*
	4	(10.0, 10.0, 10.0, 9.0, 8.0)	.222	.315*	.254
	5	(9.0, 9.0, 9.0, 9.0, 9.0)	.189	.029	.197*
	5	(10.0, 9.0, 9.0, 9.0, 9.0)	.201	.046	.206*
	5	(10.0, 10.0, 9.0, 9.0, 9.0)	.219	.078	.220*
	5	(10.0, 10.0, 10.0, 9.0, 9.0)	.249*	.142	.246
	5	(10.0, 10.0, 10.0, 10.0, 9.0)	.319	.311	.322*
Linear	1	(9.0, 8.5, 7.5, 6.5, 5.5)	.124	.296*	.195
	1	(9.0, 7.0, 5.0, 3.0, 1.0)	.129	.322*	.194
	2	(9.5, 9.0, 8.5, 7.5, 6.5)	.140	.256*	.205
	2	(9.5, 9.0, 7.0, 5.0, 3.0)	.146	.282*	.205
	3	(9.5, 9.5, 9.0, 8.5, 7.5)	.158	.224*	.217
	3	(9.5, 9.5, 9.0, 7.0, 5.0)	.169	.248*	.222
	4	(9.5, 9.5, 9.5, 9.0, 8.5)	.185	.195	.232*
	4	(9.5, 9.5, 9.5, 9.0, 7.0)	.206	.218	.247*

*Maximum power for each configuration.

Table 5. Simulated Powers of SD1PC, SD2PC, and SD1HC for Identifying MAXSDλ When λ = .85, $\mu_0 = 20$, $\mu_{MAXSD} = 18$, $\sigma = 1$, $n_0 = 16$, $n = 8$, $\nu = 50$, and $\alpha = .05$

Configuration	MAXSD	$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$	SD1PC	SD2PC	SD1HC
Step	1	(18.0, 16.0, 16.0, 16.0, 16.0)	.528	.770*	.635
	2	(18.0, 18.0, 16.0, 16.0, 16.0)	.547	.621	.629*
	2	(20.0, 18.0, 16.0, 16.0, 16.0)	.564	.770*	.649
	3	(18.0, 18.0, 18.0, 16.0, 16.0)	.575	.518	.644*
	3	(20.0, 18.0, 18.0, 16.0, 16.0)	.585	.623	.650*
	3	(20.0, 20.0, 18.0, 16.0, 16.0)	.606	.771*	.670
	4	(18.0, 18.0, 18.0, 18.0, 16.0)	.620	.440	.673*
	4	(20.0, 18.0, 18.0, 18.0, 16.0)	.627	.517	.675*
	4	(20.0, 20.0, 18.0, 18.0, 16.0)	.641	.622	.681*
	4	(20.0, 20.0, 20.0, 18.0, 16.0)	.668	.772*	.703
	5	(18.0, 18.0, 18.0, 18.0, 18.0)	.697	.381	.735*
	5	(20.0, 18.0, 18.0, 18.0, 18.0)	.703	.441	.737*
	5	(20.0, 20.0, 18.0, 18.0, 18.0)	.714	.519	.740*
	5	(20.0, 20.0, 20.0, 18.0, 18.0)	.731	.623	.747*
	5	(20.0, 20.0, 20.0, 20.0, 18.0)	.771*	.771*	.771*
Linear	1	(18.0, 14.0, 10.0, 6.0, 2.0)	.516	.724*	.621
	1	(18.0, 17.0, 15.0, 13.0, 11.0)	.528	.770*	.635
	2	(19.0, 18.0, 17.0, 15.0, 13.0)	.549	.725*	.633
	2	(19.0, 18.0, 14.0, 10.0, 6.0)	.564	.769*	.649
	3	(19.0, 19.0, 18.0, 17.0, 15.0)	.585	.726*	.648
	3	(19.0, 19.0, 18.0, 14.0, 10.0)	.606	.771*	.670
	4	(19.0, 19.0, 19.0, 18.0, 17.0)	.626	.727*	.665
	4	(19.0, 19.0, 19.0, 18.0, 14.0)	.668	.771*	.703

*Maximum power for each configuration.

our objective was to illustrate the methodology with one type of contrast. Procedures for other types of contrasts can be similarly derived and implemented if desired.

4. The results for $\mu_0/\sigma = 10$ and $\mu_0/\sigma = 20$ are similar, but the powers in the latter case are more in line with what would be required in practical applications. It must be remembered that we used a sample size of only 16 for the zero dose and 8 for each nonzero dose in these simulations. As an aid in designing dose response studies, it would be desirable to be able to determine, at least empirically, the sample sizes n_0 and n required to guarantee a specified power for given values of λ, α, and δ and assuming a particular value of μ_0/σ. However, this is a topic for a separate paper.

6. EXTENSIONS TO UNBALANCED DATA

Thus far we have assumed the one-way setup with balanced data, i.e., $n_1 = \cdots = n_k = n$, with n_0 possibly different from n. In this section, we give the changes needed to extend the results to the case of unequal n_i's. We use the notation $r_i = n_0/n_i$ in the following.

SSPC and SDPC Procedures. The t_i statistics of (4.4) use n_i in place of n. The k-variate t-distribution used to compare these t_i statistics has unequal correlations ρ_{ij}, which replace the common ρ of (4.5) by

$$\rho_{ij} = \frac{\lambda^2}{\sqrt{(\lambda^2 + r_i)(\lambda^2 + r_j)}} = \frac{\lambda}{\sqrt{\lambda^2 + r_i}} \cdot \frac{\lambda}{\sqrt{\lambda^2 + r_j}}.$$

Note that the ρ_{ij} have a product correlation structure, which makes it relatively easy to compute the required critical constants.

SD1HC Procedure. The formula (4.7) for s.e.(C_{ij}) changes to

$$\frac{s}{\sqrt{n_0}} \sqrt{(j - i + 1)^2 \lambda^2 + \sum_{h=i}^{j} r_h}.$$

The formula (4.9) for corr$(C_{ij}, C_{ij'})$ changes to

$$\rho_{jj'}^{(\ell)} = \frac{(j - i + 1)(j' - i + 1)\lambda^2 + \sum_{h=i}^{j} r_h}{\sqrt{\left[(j - i + 1)^2 \lambda^2 + \sum_{h=i}^{j} r_h\right]\left[(j' - i + 1)^2 \lambda^2 + \sum_{h=i}^{j'} r_h\right]}}$$

$$(i \le j < j' \le k).$$

The procedure remains unaltered in other respects. A Fortran program to implement all proposed procedures for unbalanced data is available at http://statlib@lib.stat.cmu.edu/general/maxsd.for.

7. RETURN TO EXAMPLE OF SECTION 2

Refer to the summary statistics in Table 1 with the change that now there is a single zero dose control group (obtained by pooling the water and solvent controls) with a sample size of 80 and a sample mean of 4.000 mm. Note that the sample means are in the monotone order. All calculations are done using the Fortran program referenced above.

SD1PC and SD2PC use the same t-statistics given by (4.4), which are listed in Table 6 along with the corresponding contrasts and their standard errors. The adjusted p-values for the two procedures are also listed in this table. We see that at $\alpha = .05$, SD1PC rejects H_{01}, H_{02}, and H_{03}, whereas SD2PC rejects H_{04} in addition. Therefore, SD1PC identifies dose 3 as $\widehat{MAXSD}.85$, and SD2PC identifies dose 4 as $\widehat{MAXSD}.85$.

Calculations for the SD1HC procedure are summarized in Table 7. For each step, the table shows the contrast

Table 6. t-Statistics and Adjusted p-Values for SD1PC and SD2PC for Identifying MAXSD.85

Dose i	Contrast value	Std. Error	t_i	Adjusted p-value SD1PC	Adjusted p-value SD2PC
1	.591	.033	18.072	.000	.000
2	.411	.032	12.685	.000	.000
3	.230	.034	6.834	.000	.000
4	.060	.034	1.779	.074	.038
5	−.190	.034	−5.502	1.000	1.000

coefficients, the contrast values and their standard errors, the corresponding t-statistics (with the maximum t-statistic marked by an asterisk), and the p-value. The adjusted p-values are readily obtained from the relation $\tilde{p}_i = \max(\tilde{p}_{i-1}, p_i)$. We see that testing stops with step 4, because $\tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3 = 0.000 < 0.05$, but $\tilde{p}_4 = 0.060 > 0.05$. Therefore, SD1HC identifies dose 3 as MAXSD.85.

In this example, SD2PC identified a higher dose as MAXSD.85 than did both SD1PC and SD1HC. This is the result of nearly linear decrease in dose means under which configuration SD2PC is generally more powerful.

8. DISCUSSION

In this article, we give multiple test procedures for identifying the MAXSD. The SD1HC procedure is generally found to be superior for step configurations, whereas SD2PC is found to be superior for linear configurations. As indicated following the discussion of simulation results in Section 5, it is possible to devise special contrasts for optimum power performance for each type of configuration. Certainly, any information available about the dose response function should be exploited in the selection of the contrasts. However, the true configuration is always unknown. Even if the general shape of the dose response function is known, the best contrast to use often depends on whether the MAXSD is at the high end, in the middle, or at the low end of the dose range, and this

is unknown. Also, in industrial settings, many compounds are routinely tested for toxicity, and each one has a different dose response function. Therefore, it is futile to aim for the most powerful procedure in each case. Rather, it is useful to have a simple and easy-to-explain procedure (without sacrificing too much power) that is robust in terms of power over a wide class of commonly occurring dose response functions.

SD2PC enjoys simplicity (requiring only a sequence of t-tests comparing each dose with the zero dose), and it has the best power performance (among the procedures compared) in many commonly occurring cases. However, it suffers a dramatic loss of power in those step configurations where toxicity at low doses is the same as that at MAXSD, causing it to stop too soon and to declare too low a dose as safe (a type II error). On the other hand, SD1HC is more robust in terms of power, but it is not as simple. Our general recommendation is to use SD2PC, except in the situations indicated, where SD1HC or a similar contrasts-based procedure is preferable.

Generally speaking, toxicity increases with dose, so the weak monotonicity assumption (3.2) appears reasonable for most safety studies. In any case, toxicologists should be consulted to check assumption (3.2). If that assumption is violated, then a dose lower than MAXSDλ is unsafe. In this case, SD2PC still controls the type I FWE, because it compares each dose with the zero dose. But SD1HC may not control the type I FWE, and thus an unsafe dose i with $\mu_i < \lambda\mu_0$ may be declared as safe (i.e., H_{0i} will be rejected) if a dose $j > i$ has $\mu_j \gg \lambda\mu_0$.

The definition (3.1) of MAXSDλ and MINUDλ may be modified to incorporate the weak monotonicity assumption (3.2), as follows:

$$\text{MAXSD}\lambda = \max\{i : \mu_j > \lambda\mu_0, \ \forall \ j \le i\} \quad \text{and}$$

$$\text{MINUD}\lambda = \min\{i : \mu_j \le \lambda\mu_0, \ \forall \ j \ge i\}. \quad (8.1)$$

To some, this may seem a more natural definition of MAXSDλ and MINUDλ. To identify the MAXSDλ according to this

Table 7. Calculations for SD1HC Procedure

Step i	Contrast j	Contrast coefficients $0.85\bar{y}_0$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$	C_{ij}	s.e.(C_{ij})	t_{ij}	p_i
1	1	−1	1	0	0	0	0	.591	.033	18.072	.000
	2	−2	1	1	0	0	0	1.001	.052	19.407*	
	3	−3	1	1	1	0	0	1.231	.070	17.613	
	4	−4	1	1	1	1	0	1.291	.088	14.752	
	5	−5	1	1	1	1	1	1.102	.105	10.487	
2	2	−1	0	1	0	0	0	.411	.032	12.685*	.000
	3	−2	0	1	1	0	0	.641	.052	12.268	
	4	−3	0	1	1	1	0	.700	.070	9.952	
	5	−4	0	1	1	1	1	.511	.088	5.793	
3	3	−1	0	0	1	0	0	.230	.034	6.834*	.000
	4	−2	0	0	1	1	0	.290	.053	5.465	
	5	−3	0	0	1	1	1	.100	.071	1.407	
4	4	−1	0	0	0	1	0	.060	.034	1.773*	.060
	5	−2	0	0	0	1	1	−.130	.054	−2.425	

definition, we test the following family of hypotheses:

$$H'_{0i} = \bigcup_{j=1}^{i} H_{0j}$$

$$= \{\text{At least one of the doses } j = 1, 2, \ldots, i \text{ is unsafe}\},$$

$$i = 1, 2, \ldots, k.$$

It is easy to check that the family $\{H'_{0i}, i = 1, 2, \ldots, k\}$ is closed and the FWE for this family equals $P\{\widehat{\text{MAXSD}}\lambda > \text{MAXSD}\lambda\}$, which we want to control at level α. A step-down procedure begins by an α-level test of $\bigcap_{i=1}^{k} H'_{0i} = H'_{01} = H_{01}$. If $H'_{01} = H_{01}$ is rejected, then it does an α-level test of $\bigcap_{i=2}^{k} H'_{0i} = H'_{02}$. However, $H'_{02} = H_{01} \bigcup H_{02}$ and H_{01} is already rejected; therefore, all we need is an α-level test of H_{02}. Continuing in this manner, we see that this procedure reduces to the SD2PC procedure. This agrees with the previous observation that SD2PC does not require assumption (3.2). Nonetheless, SD1HC is definitely a viable option because (a) for toxicity studies, a gross violation of assumption (3.2) is highly improbable; (b) SD1HC has a more stable power performance; and (c) SD1HC generally has the best power performance for step configurations.

We assumed that a lower response means a less safe dose. In many applications, a higher response means a less safe dose. In these applications, one would use a value of $\lambda > 1$, e.g., 1.10 or 1.20, corresponding to a 10% or a 20% increase over the zero dose mean, respectively. The MAXSDλ would be defined as $\max\{i : \mu_i < \lambda\mu_0\}$.

We have defined the MAXSD in terms of a multiplicative threshold constant λ. Some investigators may prefer to specify the MAXSD in terms of a shift from the zero dose mean by an additive threshold constant $\delta > 0$, i.e., MAXSD$\delta = \max\{i : \mu_i < \mu_0 + \delta\}$ (assuming that a higher response means a less safe dose). Alternatively, the log-transformation of the response variable (if it makes the assumptions of normality and equal dose group variances more nearly accurate) will also transform the multiplicative constant λ into an additive constant $\delta = \log \lambda$. In this case, test procedures can be developed along the lines of those in Tamhane, Hochberg, and Dunnett (1996) and Dunnett and Tamhane (1998).

REFERENCES

Bartholomew, D. J. (1959), "A Test of Homogeneity for Ordered Alternatives," *Biometrika*, 46, 36–48.

Bauer, P. (1997), "A Note on Multiple Testing Procedures in Dose Finding," *Biometrics*, 53, 1125–1128.

Bechhofer, R. E., and Dunnett, C. W. (1988), "Tables of Percentage Points of Multivariate Student t Distributions," *Selected Tables in Mathematical Statistics*, 11, 1–371.

Capizzi, T., Oppenheimer, L., Mehta, H., and Naimie, H. (1985), "Statistical Considerations in the Evaluation of Chronic Aquatic Toxicity Studies," *Environmental Science and Technology*, 19, 35–43.

Chase, G. R. (1974), "On Testing for Ordered Alternatives With Increased Sample Size for Control," *Biometrika*, 61, 569–578.

Crump, K. S. (1983), "A New Method for Determining Allowable Daily Intakes," *Fundamental and Applied Toxicology*, 4, 854–871.

Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments With a Control," *Journal of the American Statistical Association*, 50, 1096–1121.

Dunnett, C. W. (1989), "Multivariate Normal Probability Integrals With Product Correlation Structure," *Applied Statistics*, 38, 564–579.

Dunnett, C. W., and Tamhane, A. C. (1992), "A Step-up Multiple Test Procedure," *Journal of the American Statistical Association*, 87, 162–170.

Dunnett, C. W., and Tamhane, A. C. (1998), "New Multiple Test Procedures for Dose Finding," *Journal of Biopharmaceutical Statistics*, 8, 353–366.

Gaylor, D. W. (1983), "The Use of Safety Factors for Controlling Risk," *Journal of Toxicology and Environmental Health*, 11, 329–336.

Genz, A., and Bretz, F. (1999), "Numerical Computation of Multivariate t-Probabilities With Application to Power Calculation of Multiple Contrasts," *Journal of Statistical Computation and Simulation*, 63, 361–378.

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

Hsu, J. C., and Berger, R. L. (1999), "Stepwise Confidence Intervals Without Multiplicity Adjustment for Dose Response, and Toxicity Studies," *Journal of the American Statistical Association*, 94, 468–482.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, 655–660.

Maurer, W., Hothorn, L. A., and Lehmacher, W. (1995), "Multiple Comparisons in Drug Clinical Trials and Preclinical Assays: A Priori Ordered Hypotheses," *Biometrie in der chemisch-pharmazeutischen Industrie 6* (J. Vollmar, ed.) Stuttgart: Gustav Fischer Verlag, pp. 3–18.

Mukerjee, H., Robertson, T., and Wright, F. T. (1987), "Comparison of Several Treatments With a Control Using Multiple Contrasts," *Journal of the American Statistical Association*, 82, 902–910.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, New York: Wiley.

Rom, D. M., Costello, R. J., and Connell, L. T. (1994), "On Closed Test Procedures for Dose-Response Analysis," *Statistics in Medicine*, 13, 1583–1596.

Ruberg, S. J. (1989), "Contrasts for Identifying the Minimum Effective Dose," *Journal of the American Statistical Association*, 84, 816–822.

Schervish, M. (1984), Multivariate Normal Probabilities with Error Bound, Algorithm AS195, *Applied Statistics*, 33, 89–94. Corrigendum *Applied Statistics*, 34, 103–104.

Tamhane, A. C., Hochberg, Y., and Dunnett, C. W. (1996), "Multiple Test Procedures for Dose Finding," *Biometrics*, 52, 21–37.

Tukey, J. W., Ciminera, J. L., and Heyse, J. P. (1985), "Testing the Statistical Certainty of a Response With Increasing Doses of a Compound," *Biometrics*, 41, 295–301.

Yanagawa, T., Kikuchi, Y., and Brown, K. G. (1997), "No-Observed-Adverse-Effect Levels in Severity Data," *Journal of the American Statistical Association*, 92, 449–454.