11
Multiple Comparison Procedures in Dose Response Studies

AJIT C. TAMHANE AND BRENT R. LOGAN

11.1 Introduction

Dose–response studies are useful in Phase II and Phase III clinical trials to evaluate efficacy and toxicity of a drug in order to determine its effective and safe ranges. A zero dose is generally included as a control against which higher doses are compared. This naturally leads to multiple comparisons. The ordered nature of doses suggests the use of stepwise multiple test procedures. The purpose of this article is to give a brief overview of these procedures. In Section 11.2, we present step-down procedures for identifying the minimum effective dose (MinED). These procedures are applied to the problem of identifying the maximum safe dose (MaxSD) in Section 11.3. Examples are given in Section 11.4 followed by some extensions in Section 11.5. The paper concludes with a discussion in Section 11.6.

11.2 Identifying the Minimum Effective Dose (MinED)

11.2.1 Problem Formulation

Let $i = 0, 1, \ldots, k$ represent increasing dose levels, where 0 denotes the zero (control) dose. Assume that the efficacy measurements Y_{ij} $(1 \leq j \leq n_i)$ on the ith dose are independent and normally distributed with mean μ_i and variance σ^2, denoted by $Y_{ij} \sim N(\mu_i, \sigma^2)$. We assume that a larger μ_i represents higher efficacy. Let $\overline{Y}_i \sim N(\mu_i, \sigma^2/n_i)$ be the sample mean and $S^2 \sim \sigma^2 \chi^2_\nu/\nu$ be the pooled sample variance based on $\nu = \sum_{i=0}^{k} n_i - (k+1)$ degrees of freedom (df).

It is common to use the mean efficacy of the zero dose as a benchmark for comparison purposes to decide if a particular dose is clinically effective. Two different measures are employed for this purpose. The first is the difference measure, $\delta_i = \mu_i - \mu_0$, with a specified additive threshold $\delta > 0$ that this difference must exceed in order for dose i to be deemed effective. The second is the ratio measure, $\lambda_i = \mu_i/\mu_0$, with a specified multiplicative threshold $\lambda > 1$ that this ratio must exceed in order for dose i to be deemed effective. Here we adopt the latter approach since it requires an investigator to specify the threshold in relative terms instead of

absolute terms, which is often more difficult. Thus, e.g., if a 10% increase in the mean efficacy compared to the zero dose is regarded as clinically significant then $\lambda = 1.10$. However, it should be noted that the use of the multiplicative threshold assumes that $\mu_0 > 0$. If μ_0 is positive but close to zero, very large values of λ must be specified. Procedures using the additive threshold are briefly covered in Section 11.5.

The true MinED can be defined in two ways. A simple definition is

$$\text{MinED} = \min\{i : \mu_i > \lambda\mu_0\} \tag{11.1}$$

which is the lowest dose that is effective. For a stronger requirement on MinED the following definition is used:

$$\text{MinED} = \min\{i : \mu_j > \lambda\mu_0 \text{ for all } j \geq i\} \tag{11.2}$$

This is the lowest dose such that it and all higher doses are effective.

In some applications it is reasonable to assume that the dose–response curve satisfies a monotone property that if dose i is ineffective then all lower doses are also ineffective, and if dose i is effective then all higher doses are also effective. Formally,

$$\begin{aligned} \mu_i \leq \lambda\mu_0 \quad &\Rightarrow \mu_j \leq \lambda\mu_0 \quad \forall\ j < i \qquad \text{and} \\ \mu_i > \lambda\mu_0 \quad &\Rightarrow \mu_j > \lambda\mu_0 \quad \forall\ j > i \end{aligned} \tag{11.3}$$

This will be referred to as the weak monotonicity assumption as opposed to the strong monotonicity assumption:

$$\mu_0 \leq \mu_1 \leq \cdots \leq \mu_k \tag{11.4}$$

If the dose–response relationship is weakly monotone, then the two definitions of MinED are equivalent.

We want to guarantee that the probability of any ineffective dose being declared effective is no more than a specified level α. Let $\widehat{\text{MinED}}$ denote the sample or estimated MinED. Under weak monotonicity, this requirement translates to

$$P(\widehat{\text{MinED}} < \text{MinED}) \leq \alpha \tag{11.5}$$

Our approach to identifying MinED will be via tests of the hypotheses

$$H_i : \mu_i \leq \lambda\mu_0 \ (1 \leq i \leq k) \tag{11.6}$$

against one-sided alternatives. If using definition (11.1), the estimated MinED is defined as

$$\widehat{\text{MinED}} = \min\{i : H_i \text{ is rejected }\} \tag{11.7}$$

If using definition (11.2), the estimated MinED is defined as

$$\widehat{\text{MinED}} = \min\{i : H_j \text{ is rejected for all } \ j \geq i\} \tag{11.8}$$

If a multiple test procedure controls the familywise error rate (FWE), defined as

$$\text{FWE} = P\{\text{Reject any true } H_i\} \tag{11.9}$$

strongly (for any combination of true and false H_i) at level α then the requirement in Eq. (11.5) is satisfied. However, note that if the dose response curve is not weakly monotone, then the interpretation of Eq. (11.5), and the associated FWE, as the probability of any ineffective dose being declared effective only holds for definition (11.1).

In the next two subsections we will consider two types of multiple test procedures. The SD1PC procedure estimates the MinED according to definition (11.1). The SD2PC procedure estimates the MinED according to definition (11.2). When it is reasonable to assume monotonicity, the two definitions are equivalent and either SD1PC or SD2PC may be used.

11.2.2 Review of Multiple Test Procedures

Various procedures based on different contrasts of the dose means have been proposed in the literature (Ruberg, 1989, Tamhane et al., 1996, Dunnett and Tamhane, 1998). Here we will only consider step-down procedures based on pairwise contrasts because (1) as shown by Bauer (1997), only pairwise contrasts yield procedures that control the FWE even when the dose response is not monotone, (2) they are simple to use, and (3) they can be easily extended to nonnormal data by using appropriate two-sample statistics.

The traditional method for deriving step-down multiple test procedures is based on the closure principle due to Marcus et al., (1976). More recently, Hsu and Berger (1999) and Finner and Strassburger (2002) have proposed the partitioning principle to derive more powerful test procedures. We now explain these principles and the resulting test procedures.

11.2.2.1 Closure Principle: SD1PC Procedure

The closure principle requires a closed family of hypotheses. If we define the hypotheses $H_i' = \bigcap_{j=1}^{i} H_j$ meaning all doses at or below dose i are ineffective, then $\{H_i' \ (1 \le i \le k)\}$ is a closed family. (Note that this does not require the monotonicity assumption.) The closure principle tests each hypothesis H_i', if it is not already accepted, at level α. If H_i' is not rejected then the closure principle accepts all H_j' that are implied by H_i' without further tests. The representation $H_i' = \bigcap_{j=1}^{i} H_j$ shows that all H_j' for $j < i$ are implied by H_i'. This leads to a step-down test procedure in which H_k' is tested first. If H_k' is not rejected then all hypotheses are accepted and no dose is declared effective. Otherwise H_{k-1}' is tested next and the testing sequence continues. If H_i' is the last rejected hypothesis then $\widehat{\text{MinED}} = i$.

For the normal data assumed here, define the pairwise t-statistic corresponding to hypothesis H_i as

$$T_i = \frac{\overline{Y}_i - \lambda \overline{Y}_0}{S\sqrt{\lambda^2/n_0 + 1/n_i}} \quad (1 \le i \le k). \tag{11.10}$$

Then using the union-intersection (UI) method of Roy (1953), the statistic for testing H_i' is $T_{i,\max} = \max_{1 \le j \le i} T_j$. Under H_i' (assuming the least favorable configuration $\mu_j = \lambda\mu_0 \; \forall \; j \le i$, which maximizes the FWE), the joint distribution of $T_1, T_2, \ldots, T_i$ is an i-variate t-distribution with ν df and correlation matrix $R_i = \{\rho_{j\ell}\}$, which has a product correlation structure, $\rho_{j\ell} = \tau_j \tau_\ell$, with

$$\tau_j = \frac{\lambda}{\sqrt{\lambda^2 + r_j}} \quad \text{and} \quad r_j = \frac{n_0}{n_j} \quad (1 \le j \le k) \tag{11.11}$$

If $n_1 = n_2 = \cdots = n_k = n$ and $r = n_0/n$ then $\rho_{j\ell} \equiv \rho = \lambda^2/(\lambda^2 + r)$. Let $t^{(\alpha)}_{i,\nu,R_i}$ denote the upper α equicoordinate critical point of this i-variate t-distribution. Then the closed procedure rejects H_i' at level α iff $H_k', \ldots, H_{i+1}'$ have been rejected and

$$T_{i,\max} > t^{(\alpha)}_{i,\nu,R_i}$$

This is referred to as the SD1PC procedure. Note that the critical constants used in SD1PC are different if smaller μ_i's represent higher efficacies with the threshold $\lambda < 1$. This is so because the $\rho_{j\ell}$ are not invariant to the transformation $\lambda \leftarrow 1/\lambda$ or $\lambda \leftarrow \lambda - 1$. Also, as pointed out earlier, the SD1PC procedure is appropriate for definition (11.1) of the MinED in the sense that it will control the error rate in Eq. (11.5) for this definition, but it can also be used for definition (11.2) under the assumption of monotonicity, in which case $H_i' = H_i = \bigcap_{j=1}^{i} H_j$.

11.2.2.2 Partitioning Principle: SD2PC Procedure

The partitioning principle reformulates the hypotheses (11.6) so that they are disjoint. There are different ways to accomplish this. One way is to write the hypotheses as

$$H_i^* : \mu_i \le \lambda\mu_0, \; \mu_j > \lambda\mu_0 \quad \forall \; j > i \; (1 \le i \le k) \tag{11.12}$$

For the sake of completeness, add the hypothesis $H_0^* : \mu_j > \lambda\mu_0 \; \forall \; j$, which need not be tested. This partitioning is appropriate when efficacy is expected to increase with dose. Note that the hypotheses H_i^* are disjoint with their union being the whole parameter space, and the true parameter configuration belongs to exactly one of the H_i^*. Therefore, no multiplicity adjustment is needed to perform the tests and each H_i^* can be tested at level α independently of the others. Final inferences drawn must be logically consistent with the H_i^* that are not rejected. This procedure controls the error rate in Eq. (11.5) corresponding to the more stringent definition (11.2) of the MinED.

Note that the above formulation of the hypotheses implies that doses must be tested in a step-down manner in the order $H_k^*, H_{k-1}^*, \ldots$, stopping as soon as any hypothesis is accepted. For example, suppose $k = 5$, and all five hypotheses are

tested, but only H_5^*, H_4^* and H_2^* are rejected. Then we can only conclude that doses 5 and 4 are effective, but not dose 2. Thus, we get $\widehat{\text{MinED}} = 4$ and so testing can be stopped once H_3^* is accepted.

The main difficulty in applying the partitioning principle is that it is not easy to derive tests of the hypotheses H_i^*. However, by noting that H_i^* is a subset of H_i, we see that an α-level test of H_i provides a conservative α-level test of H_i^*. This leads to a step-down test procedure on the family $\{H_i\ (1 \le i \le k)\}$. Therefore, for testing H_i, we use the ordinary Student's t-test, which rejects H_i (assuming that $H_k, \ldots, H_{i+1}$ have been rejected) if $T_i > t_{\nu,\alpha}$, where $t_{\nu,\alpha}$ is the upper α critical point of the univariate Student's t-distribution with ν df. The resulting step-down procedure is referred to as the SD2PC procedure.

Although we have derived the SD2PC procedure by using the partitioning principle, it can also be derived by noting that the a priori ordering of the hypotheses results in their nesting: $H_k \subseteq H_{k-1} \subseteq \cdots \subseteq H_1$. This approach is employed by Maurer et al. (1995) to show that SD2PC controls the FWE strongly.

Finally we note that both SD1PC and SD2PC are pre-determined testing procedures since they both test the hypotheses $H_k, H_{k-1}, \ldots$ in a pre-determined order not in a sample-determined order (see Chapter 12).

11.2.3 Simultaneous Confidence Intervals

Bretz et al. (2003) proposed stepwise confidence intervals for the ratios $\lambda_i = \mu_i/\mu_0$ based on Fieller's (1954) method. Consider the r.v.

$$T_i = \frac{\overline{Y}_i - \lambda_i \overline{Y}_0}{S\sqrt{\lambda_i^2/n_0 + 1/n_i}}$$

which is t-distributed with ν df. By solving the inequality $T_i \le t_{\nu,\alpha}$, which is an event of probability $1-\alpha$, we get the following $100(1-\alpha)\%$ lower confidence bound on λ_i:

$$\lambda_i \ge L_i = \frac{\overline{Y}_0\overline{Y}_i - \sqrt{a_0\overline{Y}_i^2 + a_i\overline{Y}_0^2 - a_0 a_i}}{\overline{Y}_0^2 - a_0}$$

where $a_i = t_{\nu,\alpha}^2 S^2/n_i$ $(0 \le i \le k)$.

For identifying the MinED Bretz et al. (2003) embedded these marginal $100(1-\alpha)\%$ confidence intervals into the following step-down procedure, which does not require any multiplicity adjustment according to the results of Hsu and Berger (1999).

STEP 1: If $L_k \le \lambda$ then conclude that $\lambda_k \ge L_k$, all doses are ineffective and stop. Otherwise conclude that $\lambda_k > \lambda$ (dose k is effective) and go to Step 2.

STEP i: If $L_{k-i+1} \le \lambda$ then conclude that $\lambda_{k-i+1} \ge L_{k-i+1}$, doses $1, \ldots, k-i+1$ are ineffective and stop. Otherwise conclude that $\lambda_{k-i+1} > \lambda$ (dose $k-i+1$ is effective) and go to Step $i+1$.

STEP $k+1$: Conclude that $\min_{1\le i\le k} \lambda_i \ge \min_{1\le i\le k} L_i$.

This test procedure is equivalent to the SD2PC procedure because it is derived from it. However, additionally, it yields lower confidence bounds on the λ_i's for all doses found effective and the first dose found ineffective.

11.3 Identifying the Maximum Safe Dose (MaxSD)

All of the preceding discussion extends naturally to the problem of identifying the MaxSD in toxicity studies with a few minor changes as we note below. In order to keep the forms of the hypotheses (11.6) and the test statistics in Eq. (11.10) the same, and also to conform with the past literature, we will assume that lower μ_i implies a more toxic (less safe) dose. Toxicity generally increases with dose level and the zero dose has the least toxicity. Therefore the μ_i's are generally decreasing and the threshold $\lambda < 1$. Thus, dose i with $\mu_i > \lambda\mu_0$ is regarded as safe, while dose i with $\mu_i \leq \lambda\mu_0$ is regarded as unsafe. For example, $\lambda = 0.90$ means that a 10% decrease in safety level (increase in toxicity) is regarded as clinically unsafe.

The maximum safe dose (MaxSD) for specified $\lambda < 1$ is defined as

$$\text{MaxSD} = \max\{i : \mu_j > \lambda\mu_0 \ \forall \ j \leq i\}$$

Analogous to the discussion of the MinED, there could be two definitions of the MaxSD. However, we assume monotonicity of the toxicity response so that the definitions are identical. The hypotheses are the same as in Eq. (11.6) (where now H_i states that the ith dose is unsafe). If

$$\widehat{\text{MaxSD}} = \max\{i : H_j \text{ is rejected } \ \forall \ j \leq i\}$$

denotes the estimated MaxSD then we want to guarantee that

$$P(\widehat{\text{MaxSD}} > \text{MaxSD}) \leq \alpha \tag{11.13}$$

Since the goal is now to find the MaxSD, both SD1PC and SD2PC start by testing H_1 and proceed to testing H_2 if H_1 is rejected (dose 1 is declared safe) and so on. If H_1 is not rejected then all H_i are accepted without further tests and all doses are declared unsafe, i.e., there is no MaxSD. SD1PC rejects H_i using the representation $H_i = \bigcap_{j=i}^{k} H_j$ if

$$T_{i,\max} = \max_{i \leq j \leq k} T_j > t^{(\alpha)}_{\ell,\nu,R_\ell}$$

where $\ell = k - i + 1$ and $R_\ell = \{\rho_{ij}\}$, while SD2PC rejects H_i if $T_i > t_{\nu,\alpha}$. For details see Tamhane et al.

11.4 Examples

Example 1 (Identifying the MinED): Tamhane and Logan (2002) cite an example of a Phase II randomized, double-blind, placebo-controlled parallel group clinical trial of a new drug for the treatment of arthritis of the knee using four increasing

doses (labeled 1 to 4). While they consider both efficacy and safety outcomes in that study, here we focus only on the efficacy data. A total of 370 patients were randomized to the five treatment groups. The efficacy variable is the pooled WOMAC (Western Ontario and McMaster Universities osteoarthritis index) score, a composite score computed from assessments of pain (5 items), stiffness (2 items), and physical function (17 items). The composite score is normalized to a scale of 0–10. An increase in WOMAC from the baseline indicates an improvement in disease condition. We will consider a 30% improvement in WOMAC scores over the baseline mean compared to that for the zero dose group a clinically significant improvement, so that $\lambda = 1.3$.

The summary data are given in Table 11.1. Normal plots were found to be satisfactory, and the Bartlett and Levene tests for homogeneity of variances yielded nonsignificant results. The sample sizes are approximately equal so that $r_i \approx r = 1$ and $\rho_{ij} \approx \rho = 1.30^2/(1.30^2 + 1) = 0.628$. The pooled estimate of the standard deviation is 1.962 with $\nu = 365$ df. The t-statistics computed using Eq. (11.10) are given in Table 11.2.

Table 11.1. Summary statistics for changes from baseline in WOMAC score

	Dose level				
	0	1	2	3	4
Mean	1.437	2.196	2.459	2.771	2.493
SD	1.924	2.253	1.744	1.965	1.893
n	76	73	73	75	73

Table 11.2. t-Statistics and unadjusted p-values for WOMAC scores

	Comparison			
	1 vs. 0	2 vs. 0	3 vs. 0	4 vs. 0
T_i	0.881	1.588	2.439	1.680
p_i	0.189	0.056	0.007	0.047

The SD1PC procedure begins by comparing $T_{4,\max} = 2.439$ with the critical value $t^{(.05)}_{4,365,0.628} = 2.123$. Since $2.439 > 2.123$, we step down to compare dose 3 with the control. In fact, we can take a shortcut and step down to compare dose 2 with the control, since $T_{4,\max} = T_3 = 2.439$ and the multivariate T-critical values decrease with dimension implying rejection of H_3. So, next we compare $T_{2,\max} = 1.588$ with the critical value $t^{(.05)}_{2,365,0.628} = 1.900$. Since $1.588 < 1.900$, we accept hypothesis H_2 and by implication hypothesis H_1, leading to the conclusion that $\widehat{\text{MinED}} = 3$.

The SD2PC procedure begins by comparing $T_4 = 1.680$ with the critical value $t_{365,.05} = 1.649$. Both T_4 and $T_3 = 2.439$ exceed 1.649, so we reject H_4 and H_3.

Next $T_2 = 1.588 < 1.649$, so we stop and accept H_2 and hence H_1 by implication, leading to the same conclusion that $\widehat{\text{MinED}} = 3$.

To compute stepwise 95% confidence intervals using the Bretz et al. method we first compute

$$L_1 = 1.136, L_2 = 1.288, L_3 = 1.468, L_4 = 1.308$$

Since L_4 and L_3 are both greater than $\lambda = 1.30$ we conclude that both doses 4 and 3 are effective. But $L_2 < \lambda = 1.30$, and so we stop and conclude that dose 2 is ineffective and $\lambda_2 \geq 1.288$. Obviously, we get the same conclusion as SD2PC, but additionally we get confidence bounds on λ_4, λ_3 and λ_2.

Example 2 (Identifying the MaxSD): Tamhane et al. (2001) cite an aquatic toxicology study in which daphnids, or water fleas (*Daphnia magna*), were exposed over 21 days to a potentially toxic compound. Daphnids of the same age and genetic stock were randomly assigned to a water control, a solvent control, or one of six concentrations of a pesticide. The safety endpoint of interest was the growth, as measured by the lengths of the daphnids after 21 days of continuous exposure. Because there was no significant difference between the two control groups, they were combined for subsequent analysis. Six nominal concentrations of the pesticide were tested: 0.3125, 6.25, 12.5, 25, 50, and 100 ppm. Forty daphnids were randomly assigned to each group, but because some died during the course of the experiment they were not evaluable. Also, because of excessive mortality in the 100 ppm dose group, it was omitted from subsequent analysis. This follows the recommendation of Capizzi et al. (1985) for a two-stage approach, in which survival is studied in the first stage, and sublethal effects (such as growth) are compared among those doses which do not significantly affect survival. In the toxicology community, opinions about what constitutes a biologically significant effect have ranged from 5 to 25% adverse effect. If we take an average of this range, i.e., 15% reduction in length or $\lambda = 0.85$ as biologically unsafe, then we would like to know which dose is the MaxSD for this value of λ.

The summary statistics are given in Table 11.3. Normal plots were found to be satisfactory, and the Levene test for homogeneity of variances was nonsignificant. The pooled estimate of the standard deviation is 0.1735 with $\nu = 254$ df. Additional analyses of variance were performed on the data as discussed in Tamhane et al. (2002), but we do not elaborate on them here. The sample sizes in the nonzero dose groups were all approximately equal, so that $r_i \approx r = 80/36 = 2.222$ and $\rho_{ij} \approx \rho = 0.85^2/(0.85^2 + 2.222) = 0.245$. The t-statistics computed using Eq. (11.10) are given in Table 11.4.

Table 11.3. Summary statistics for daphnid length data

	Dose level					
	0	1	2	3	4	5
Mean	4.0003	3.9908	3.8108	3.6306	3.4600	3.2106
SD	0.1496	0.2110	0.1504	0.1961	0.1726	0.1829
n	80	38	39	35	35	33

Table 11.4. t-Statistics and unadjusted p-values for daphnid length data

	Comparison				
	1 vs. 0	2 vs. 0	3 vs. 0	4 vs. 0	5 vs. 0
T_i	18.082	12.692	6.838	1.774	−5.505
p_i	0.000	0.000	0.000	0.038	1.000

For the SD1PC procedure, the critical values are $t^{(.05)}_{5,254,0.245} = 2.307, t^{(.05)}_{4,254,0.245} = 2.224, t^{(.05)}_{3,254,0.245} = 2.114, t^{(.05)}_{2,254,0.245} = 1.952$, and $t^{(.05)}_{1,254,0.245} = 1.652$. The SD1PC procedure proceeds by comparing the statistics $T_{i,\max}$ to the critical values in sequence, starting with $T_{1,\max} = 18.082$. These are rejected in sequence until we come to $T_{4,\max} = 1.774$, which is less than 1.952. Therefore, we conclude that $\widehat{\text{MaxSD}} = 3$.

The SD2PC procedure proceeds by comparing each t-statistic with the critical value $t_{254,.05} = 1.652$, starting with dose 1. H_4 is the last hypothesis rejected, since $T_4 = 1.774 > 1.652$ and $T_5 = -5.505 < 1.652$. Therefore, we stop and accept H_5, leading to the conclusion that $\widehat{\text{MaxSD}} = 4$. Note that SD2PC found dose 4 to be safe, whereas SD1PC did not.

11.5 Extensions

Several extensions of the basic methods described above have been studied in the literature. We briefly summarize a few below.

1. Multiple test procedures based on general contrasts are given in Ruberg (1989), Tamhane et al. (1996), Dunnett and Tamhane (1998), and Tamhane et al. (2001). The first three papers use the difference measure approach. Specifically, when using the difference measure approach for the MinED problem, a general contrast for testing $H_i : \mu_i \le \mu_0 + \delta$ is given by

$$C_i = c_{i0}(\overline{Y}_0 + \delta) + c_{i1}\overline{Y}_1 + \cdots + c_{ik}\overline{Y}_k \quad (1 \le i \le k)$$

where the contrast coefficients c_{ij} sum to zero. The corresponding test statistic is

$$T_i = \frac{C_i}{\text{s.e.}(C_i)} = \frac{C_i}{S\sqrt{\sum_{j=0}^{I} c_{ij}^2/n_j}} \quad (1 \le i \le k)$$

The T_i have a multivariate t-distribution with correlations that depend on the c_{ij} and the n_i. If the dose response shape is known a priori then the c_{ij} can be chosen to mimic its shape, e.g., if the shape is roughly linear then one can use linear contrasts in which the c_{ij} form an arithmetic progression. However, often such knowledge is lacking. Previous simulation studies have shown that the procedures based on Helmert contrasts, in which $c_{ij} = -1$, for $j = 0, 1, \ldots, i-1, c_{ii} = i$ and $c_{ij} = 0$ for $j > i$, perform better than those

based on other contrasts when the minimum effective dose is at the high end or when the dose response shape is convex, and do not perform too badly in other cases. Another advantage of Helmert contrasts is that for a balanced design ($n_0 = n_1 = \cdots = n_k$) they are uncorrelated, i.e., $\rho_{ij} = 0$. Effectively, the ith Helmert contrast compares the ith dose level mean with the average of all the lower dose level means (including the zero dose).
Other trend tests are available as well for testing the hypotheses at each stage of the step-down procedure. Abelson and Tukey (1963) propose a contrast test which minimizes the maximum power loss over the alternative hypothesis space. Stewart and Ruberg (2000) propose using the maximum of several well-defined contrast tests to improve the robustness of the trend test to different dose–response shapes. Tests could alternatively be based on isotonic regression (Robertson et al. 1988; Williams 1971, 1972).

2. The problem of identifying the MinED and MaxSD simultaneously is considered in Tamhane and Logan (2002); see also Bauer et al. (2001). The therapeutic window is defined as the interval [MinED, MaxSD] if this interval is nonempty. This interval is estimated by $[\widehat{\text{MinED}}, \widehat{\text{MaxSD}}]$ subject to the requirement that the probability that $[\widehat{\text{MinED}}, \widehat{\text{MaxSD}}]$ contains any ineffective or unsafe doses is less than or equal to a prespecified level α, i.e.,

$$P\left\{\widehat{\text{MinED}} < \text{MinED or } \widehat{\text{MaxSD}} > \text{MaxSD}\right\} \leq \alpha$$

Tamhane and Logan (2002) investigated several strategies, including α-splitting, where the MinED is identified with Type I error α_E and the MaxSD is identified with Type I error α_S so that $\alpha_E + \alpha_S = \alpha$. They also proposed more efficient bootstrap procedures which take into account the correlation between efficacy and safety variables.

3. In many applications the assumption of homoscedasticity of variances is not satisfied. In Tamhane and Logan (2004), we give extensions of the procedures discussed here as well as those based on Helmert contrasts when the dose response data are heteroscedastic.
4. Nonparametric extensions of the step-down procedures for identifying the MinED are given by Chen (1999), Sidik and Morris (1999), Chen and Jan (2002), and Jan and Shieh (2004).

11.6 Discussion

In this section, we compare the methodology proposed in this paper with that currently practiced by the U.S. Food and Drug Administration (FDA). For simplicity assume a single dose or drug. Then the FDA's criterion for efficacy consists of the proof of statistical significance and of clinical significance. Denoting the means for the control and the drug by μ_0 and μ_1, respectively, the statistical significance criterion is met if $H_0 : \mu_1 \leq \mu_0$ is rejected in favor of the one-sided alternative $H_1 : \mu_1 > \mu_0$ at the α-level (usually 2.5%). For clinical significance, if the ratio measure is adopted then it is required that $\widehat{\mu}_1/\widehat{\mu}_0 > \lambda$, where $\lambda > 1$ is a specified

threshold. Thus it is required that the $100(1-\alpha)\%$ confidence interval for μ_1/μ_0 lie above 1, but only the point estimate of μ_1/μ_0 lie above the threshold λ. On the other hand, our approach tests $H_0 : \mu_1 \leq \lambda\mu_0$ vs. $H_1 : \mu_1 > \lambda\mu_0$ and thus requires that the $100(1-\alpha)\%$ confidence interval for μ_1/μ_0 lie above λ, which is a stricter requirement. The two approaches are equivalent if $\lambda = 1$. We recommend that the stricter requirement with $\lambda > 1$ be adopted since requiring that the point estimate $\widehat{\mu}_1/\widehat{\mu}_0 > \lambda$ does not guarantee that the true ratio $\mu_1/\mu_0 > \lambda$ with $100(1-\alpha)\%$ confidence. Similar discussion applies if the difference measure is used. In either case, another practical problem is how to specify the threshold.

Acknowledgments

The authors are grateful to two referees and Dr. Naitee Ting for their comments and suggestions which significantly improved the paper.

References

Abelson, R. P., and Tukey, J. W. 1963. Efficient utilization of non-numerical information in quantitative analysis: general theory and the case of simple order. *Annals of Mathematical Statistics* 34:1347–1369.

Bauer, P. 1997. A note on multiple testing procedures for dose finding. *Biometrics* 53:1125–1128.

Bauer, P., Brannath, W., and Posch, M. 2001. Multiple testing for identifying effective and safe treatments. *Biometrical Journal* 43:605–616.

Bretz, F., Hothorn, L. A., and Hsu, J. C. 2003. Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statistics in Medicine* 22:847–858.

Capizzi, T., Oppenheimer, L., Mehta, H., and Naimie, H. 1985. Statistical considerations in the evaluation of chronic toxicity studies. *Environmental Science and Technology* 19:35–43.

Chen, Y. I. 1999. Nonparametric identification of the minimum effective dose. *Biometrics* 55:126–130.

Chen, Y. I., and Jan, S-L 2002. Nonparametric identification of the minimum effective dose for randomized block designs. *Communications in Statistics* Ser. B 31:301–312.

Dunnett, C. W., and Tamhane, A. C. 1998. Some new multiple test procedures for dose finding. *Journal of Biopharmaceutical Statistics* 8:353–366.

Fieller, E. C. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society* Ser. B 16:175–185.

Finner, H., and Strassburger, K. 2002. The partitioning principle: A powerful tool in multiple decision theory. *Annals of Statistics* 30:1194–1213.

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.

Jan, S-L. and Shieh, G. 2004. Multiple test procedures for dose finding. *Communications in Statistics* Ser. B 34:1021–1037.

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," in *Testing Principles in Clinical and Preclinical Trials (J. Vollmar, editor)* Stuttgart, Gustav Fischer Verlag, pp. 3–18.

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

Roy, S. N. 1953. On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* 24:220–238.

Ruberg, S. J. 1989. Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association* 84:816–822.

Sidik, K., and Morris, R. W. 1999. Nonparametric step-down test procedures for finding the minimum effective dose. *Journal of Biopharmaceutical Statistics* 9:217–240.

Stewart, W. H., and Ruberg, S. J. 2000. Detecting dose response with contrasts. *Statistics in Medicine* 19:913–921.

Tamhane, A. C., Dunnett, C. W., Green, J. W., and Wetherington, J. F. 2001. Multiple test procedures for identifying the maximum safe dose. *Journal of the American Statistical Association* 96:835–843.

Tamhane, A. C., Hochberg, Y., and Dunnett, C. W. 1996. Multiple test procedures for dose finding. *Biometrics* 52:21–37.

Tamhane, A. C., and Logan, B. R. 2002. Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *Journal of the American Statistical Association* 97:293–301.

Tamhane, A. C., and Logan, B. R. 2004. Finding the maximum safe dose for heteroscedastic data. *Journal of Biopharmaceutical Statistics* 14:843–856.

Williams, D. A. 1971. A test for differences between treatment means when several dose levels are compared with a zero dose level. *Biometrics* 27:103–117.

Williams, D. A. 1972. The comparison of several dose levels with a zero dose control. *Biometrics* 28:519–531.