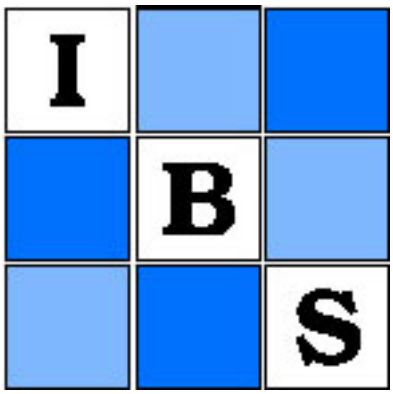

BIOMETRICS 52, 21–37
March 1996

Multiple Test Procedures for Dose Finding

Ajit C. Tamhane,[1] Yosef Hochberg,[2] and Charles W. Dunnett[3]

[1]Department of Statistics, Northwestern University,
2006 Sheridan Road, Evanston, Illinois 60208-4070, U.S.A.
[2]Department of Statistics, Tel-Aviv University,
Tel-Aviv, Ramat-Aviv 69978, Israel
[3]Department of Mathematics and Statistics, McMaster University,
Hamilton, Ontario L8S 4K1, Canada

SUMMARY

The problem of identifying the lowest dose level for which the mean response differs from that at the zero dose level is considered. A general framework for stepwise testing procedures that use contrasts among the dose level means is proposed. Using this framework, several new procedures are derived. These and some existing procedures, including that of Williams (1971, *Biometrics* **27,** 103–117; 1972, *Biometrics* **28,** 519–531), are compared analytically and by an extensive simulation study for the normal theory balanced one-way layout case. It is pointed out that the procedures based on the so-called step and basin contrasts proposed by Ruberg (1989, *Journal of American Statistical Association* **84,** 816–822) have excessively high type I familywise error rates (FWEs) and, hence, they should not be used. Some findings of the simulation study are as follows: For monotone dose mean configurations, Williams' procedure and two step-down test procedures based on Helmert and linear contrasts offer the best performance. For nonmonotone dose mean configurations, the performance of Williams' procedure does degrade somewhat, but the other two procedures are still the best. For more complex designs, a simple step-down test procedure that uses any α-level tests (not necessarily t-tests) to compare each dose level with the zero dose level controls the FWE and is the only alternative available, but its power is rather low, especially under nonmonotone configurations. Step-up procedures are generally dominated by step-down procedures when the same contrasts are used although the differences are not great.

1. Introduction

A common problem in toxicological and drug development studies is to assess the biological activity of a chemical compound. For this purpose, a dose–response experiment is conducted in which several doses of the compound are administered to separate groups of experimental units. It is customary to include a zero dose to serve as a placebo control. There are two primary goals in these studies. In a toxicological study the goal is to estimate a safe dose that will not cause some undesirable effect (e.g., toxicity, carcinogenicity), whereas in a drug development study the goal is to estimate the lowest dose that will cause some desirable effect.

In toxicological studies, the conventional approach is to find the highest dose for which the response does not differ significantly from that at the zero dose (called the no observed adverse event level [NOAEL] dose; see Ryan, 1992) and apply an appropriate safety factor to it to arrive at a safe dose level. This approach has the drawback that smaller and less sensitive experiments result in higher safe doses, which is the opposite of what is desired. To correct this drawback, Gaylor (1983) and Crump (1984) have proposed an alternative approach in which a dose–response curve is fitted to the data and the dose level corresponding to a specified risk level is estimated (e.g., ED01, which causes a 1% increase in risk over the control level). To account for the uncertainty in this estimate, the safety factor used to arrive at the safe dose level is based on the upper confidence limit on the risk level at the estimated ED01.

Key words: Dose–response function; Familywise error rate; Monte Carlo simulations; Multiple comparison procedures; Multivariate t-distribution; Stepwise testing procedures.

In drug development studies, this regression approach is not commonly used nor is it generally needed because no extrapolation from the experimental data is involved. Instead, a testing approach is employed to identify the lowest dose with an effect that exceeds that of the control. This is known as the *dose-finding problem.* In the present paper, we address this latter problem.

We assume that the responses are normally distributed and that the only possible effect of the dose level is a shift in the mean. The specific goal of the dose–response study is to find the smallest dose for which the mean is shifted from the zero dose mean; Ruberg (1989) referred to this dose as the *minimum effective dose* (MED). However, what any test procedure finds is a *minimum detectable dose* (MDD), which is one level higher than the NOAEL.

Williams (1971, 1972) proposed one of the first dose-finding procedures. Ruberg (1989) proposed some procedures based on selected contrasts of the sample means at different doses. Rom, Costello, and Connell (1994) derived closed testing procedures that allow comparisons between successive sets of doses in addition to comparisons with the zero dose (see also Budde and Bauer, 1989). In this paper, we review these procedures in a general unifying framework and propose some new ones. We then compare the various procedures by Monte Carlo simulation.

The outline of the paper is as follows. Section 2 gives notation and assumptions. Section 3 presents the various procedures in a general framework. Section 4 gives a numerical example to illustrate the procedures. Section 5 discusses the results of the simulation study. Finally, Section 6 gives conclusions and recommendations.

2. Preliminaries

2.1 *Notation and Assumptions*

Denote a set of increasing dose levels by $0, 1, 2, \ldots, k$, where 0 corresponds to the zero dose level (placebo control). Consider a one-way layout setting in which n_i experimental units are tested at the ith dose level, $i = 0, 1, \ldots, k$. We assume that all observations y_{ij} are mutually independent with $y_{ij} \sim N(\mu_i, \sigma^2)$, $i = 0, 1, \ldots, k$ and $j = 1, 2, \ldots, n_i$.

Let $\bar{y}_i \sim N(\mu_i, \sigma^2/n_i)$, $i = 0, 1, \ldots, k$, be the sample means, and let s^2 be an unbiased estimate of the common variance σ^2 based on ν degrees of freedom (df) and distributed as $\sigma^2\chi^2_\nu/\nu$, independently of the $\bar{y}_i$. Usually, s^2 is the mean square error estimate from the analysis of variance with $\nu = \sum_{i=0}^{k} n_i - (k+1)$ df. Henceforth, we restrict to the case $n_1 = n_2 = \cdots = n_k = n$; that is, an equal number, n, of experimental units is tested at each of the nonzero dose levels. For some procedures, we put an additional restriction that $n_0 = n$; the numerical example and the simulation study are confined to this case. These restrictions on the sample sizes are imposed primarily for the sake of convenience. In the final section, we indicate how some of the procedures can be extended to the unequal sample size case.

The MED is defined as

$$\text{MED} = \min\{i : \mu_i > \mu_0\}.$$

Note that the MED is defined in mathematical terms—not in terms of biologically meaningful effectiveness. If a threshold value $\Delta > 0$ can be specified for the latter, then all of the procedures can be readily modified by defining the MED as $\min\{i : \mu_i > \mu_0 + \Delta\}$. Thus, without loss of generality, we can take $\Delta = 0$.

The problem of identifying the MED is formulated as a sequence of hypothesis testing problems:

$$H_{0i} : \mu_0 = \mu_1 = \cdots = \mu_i \text{ versus } H_{1i} : \mu_0 = \mu_1 = \cdots = \mu_{i-1} < \mu_i \quad (1 \le i \le k). \tag{2.1}$$

If i^* is the smallest i for which H_{0i} is rejected, then the i^*th dose is identified to be the MED, that is, $\widehat{\text{MED}} = i^*$.

REMARK. As already mentioned, the $\widehat{\text{MED}}$ found by using this hypothesis testing approach is simply the lowest dose that differs significantly from the zero dose. In this sense, the hypothesis testing procedures do not really identify or estimate the MED; rather, they find the so-called MDD.

Note that we are *not* assuming that the μ_i are monotonically ordered:

$$\mu_0 \le \mu_1 \le \cdots \le \mu_k. \tag{2.2}$$

In particular, it is possible under H_{1i} that $\mu_j < \mu_0$ for some $j > i$. However, for practical reasons, it seems prudent to require that $\mu_j \ge \mu_0$ for all $j \ge i$. Only Williams' procedure discussed in § 3.1 makes monotonicity assumption (2.2).

Since this is a multiple hypothesis testing problem, it is logical to require control of the familywise error rate (FWE), which is defined as

$$\text{FWE} = P\{\text{at least one true } H_{0i} \text{ is rejected}\}.$$

We restrict to multiple test procedures that *strongly* control (Hochberg and Tamhane, 1987, Chapter 2) FWE $\leq \alpha$ for designated α. Why is strong control of the FWE needed? To see the inappropriateness of weak control of the FWE only under the global null hypothesis $H_{0k} : \mu_0 = \mu_1 = \cdots = \mu_k$, consider two experiments, one with $k = 4$ dose levels and the other with $k = 5$ dose levels, and suppose that the true MED = 5. Then, estimating dose 4 to be the MED would be regarded as a type I error in the first experiment (where the global null hypothesis is true) but not in the second experiment (where the global null hypothesis is false). It is precisely because the true MED can be any one of the dose levels that control of the FWE is needed, and it should be strong control. If the cost of a type I error is to be balanced against the cost of a type II error, then it is the α that should be changed, not the type of error rate control.

2.2 *Closed Procedures*

A general method for constructing a procedure that controls the FWE strongly for any family of hypotheses was given by Marcus, Peritz, and Gabriel (1976). The method consists of first forming a closure of the family of hypotheses by including in it all intersections of the original hypotheses. A hypothesis in the closure is rejected at level α iff all hypotheses implying it, including itself, are significant at level α.

The family of hypotheses $\{H_{0i}, 1 \leq i \leq k\}$ is already a closed family because for any set of indices $1 \leq i_1 < i_2 < \cdots < i_m \leq k$,

$$H_{0i_1} \cap H_{0i_2} \cap \cdots \cap H_{0i_m} = H_{0i_m}.$$

As a result, it is particularly easy to construct closed procedures. All that is needed are separate α-level tests of the individual H_{0i}, which must be applied in a step-down manner. A hypothesis H_{0i} is tested and rejected at level α iff all the hypotheses H_{0j} are significant at level α for $j \geq i$. The step-down procedures discussed in § 3.2.2 are of closed type and, hence, control the FWE strongly.

3. Description of Procedures

We first describe Williams' (1971, 1972) step-down procedure, which is based on the monotonicity assumption 2.2. Next we describe a class of procedures that do not make this assumption.

3.1 *Williams' Procedure*

Williams' (1971, 1972) procedure (WILM) does not use the $\bar{y}_i$'s as the estimates of the μ_i's; instead, it uses the isotonic (maximum likelihood) estimates

$$\hat{\mu}_i = \max_{1 \leq u \leq i} \min_{i \leq v \leq k} \frac{\sum_{i=u}^{v} \bar{y}_i}{(v - u + 1)} \quad (1 \leq i \leq k)$$

based on order restriction (2.2). The $\hat{\mu}_i$ are conveniently calculated by using the "pool the adjacent violators" algorithm. Next, pairwise t-type statistics:

$$\bar{t}_i = \frac{\hat{\mu}_i - \bar{y}_0}{s\sqrt{1/n_0 + 1/n}} \qquad (1 \leq i \leq k) \tag{3.1}$$

are calculated. Hypotheses (2.1) are tested in a step-down manner by comparing these statistics with their corresponding critical points $c_i = \bar{t}_{i,\nu}^{(\alpha)}$ as follows: Begin by testing H_{0k}. Reject H_{0k} if $\bar{t}_k \geq \bar{t}_{k,\nu}^{(\alpha)}$ and proceed to testing $H_{0,k-1}$; otherwise, stop by accepting all H_{0i}. In general, test and reject H_{0i} iff H_{0j} for $j > i$ are rejected and $\bar{t}_i \geq \bar{t}_{i,\nu}^{(\alpha)}$; otherwise, stop by accepting $H_{01}, \ldots, H_{0i}$. Estimate the MED as $\widehat{\text{MED}} = i^*$ if i^* is the smallest index i for which H_{0i} is rejected. In other words,

$$i^* = \min\left\{i : \bar{t}_j \geq \bar{t}_{j,\nu}^{(\alpha)}, \qquad j = i, i+1, \ldots, k\right\}.$$

If there is no such i^*, that is, if $\bar{t}_k < \bar{t}_{k,\nu}^{(\alpha)}$, then declare no dose level as the MED (or conclude that the MED is higher than dose level k).

For the special case $n_0 = n$, Williams (1971) tabulated the upper α critical points, $\bar{t}_{i,\nu}^{(\alpha)}$, of the distribution of $\bar{t}_i$ under H_{0i} for selected values of α, i and ν. An empirical formula to extend these to the case $n_0 \neq n$ is given in Williams (1972).

REMARK. There is an inconsistency in the way this procedure is stated in Williams' two papers. In his first paper (1971), the zero dose mean, $\bar{y}_0$, is included in the calculation of the isotonic estimates both in the description of the procedure and in the numerical example. However, when deriving the joint distribution of the $\bar{t}_i$ for the purpose of determining the critical points $\bar{t}_{i,\nu}^{(\alpha)}$, it is assumed that $\bar{y}_0$ is not included. On the other hand, in his second paper (1972), $\bar{y}_0$ is excluded from the isotonic estimates calculation throughout. Actually, both ways of calculating the isotonic estimates lead to identical estimates i^* of the MED as well as identical $\bar{t}$-tests for testing it. Hence, it does not matter which way they are calculated.

Also, Williams (1971) proposed an alternative test statistic that uses the isotonic estimate $\hat{\mu}_0$ of μ_0 in place of $\bar{y}_0$ in (3.1). Marcus (1976) studied the performance of this modified Williams procedure.

3.2 *A Class of Stepwise Procedures Based on Contrasts among the Sample Means*

These procedures can be classified according to the type of contrast and the type of testing scheme.

3.2.1 Type of contrast. For testing a hypothesis H_{0i}, a contrast of the following general form is used:

$$a_{i0}\bar{y}_0 + a_{i1}\bar{y}_1 + \cdots + a_{ik}\bar{y}_k,$$

where $\sum_{j=0}^{k} a_{ij} = 0$. The corresponding t-statistic is given by

$$t_i = \frac{\sum_{j=0}^{k} a_{ij}\bar{y}_j}{s\sqrt{a_{i0}^2/n_0 + \sum_{j=1}^{k} a_{ij}^2/n}} \qquad (1 \leq i \leq k). \tag{3.2}$$

The critical points of the procedure depend on the joint distribution of the t_i. This is a multivariate t-distribution with ν df and correlation matrix $\{\rho_{ij}\}$, where ρ_{ij} is the correlation coefficient between the ith and the jth contrasts $(1 \leq i \neq j \leq k)$.

Four different contrasts are considered:

1) *Pairwise* (P) *Contrasts:* The ith pairwise contrast is simply $\bar{y}_i - \bar{y}_0$ $(1 \leq i \leq k)$. In this case, (3.2) simplifies to

$$t_i = \frac{\bar{y}_i - \bar{y}_0}{s\sqrt{1/n_0 + 1/n}} \qquad (1 \leq i \leq k). \tag{3.3}$$

The correlation coefficients are given by $\rho_{ij} = n/(n_0 + n)$, which equal $1/2$ for the special case $n_0 = n$.

2) *Helmert* (H) *Contrasts* (Ruberg, 1989). The ith Helmert contrast is defined by

$$a_{ij} = \begin{cases} -1, & j = 0, 1, \ldots, i-1, \\ i, & j = i, \\ 0, & j = i+1, \ldots, k. \end{cases} \tag{3.4}$$

Effectively, the ith contrast compares the ith dose level mean with the average of all the lower dose level means (including the zero dose level). These contrasts are mutually orthogonal. Therefore $\rho_{ij} = 0$ when $n_0 = n$.

3) *Reverse Helmert* (R) *Contrasts.* These are the reverse of Helmert contrasts in the sense that the ith contrast compares the average of the means of the first i dose levels with the zero dose mean. Thus, the contrast coefficients are given by

$$a_{ij} = \begin{cases} -i, & j = 0, \\ 1, & j = 1, \ldots, i, \\ 0, & j = i+1, \ldots, k. \end{cases} \tag{3.5}$$

These contrasts are not orthogonal. For $n_0 = n$, the correlation coefficients between these contrasts are unequal and are given by

$$\rho_{ij} = \sqrt{\frac{i/(i+1)}{j/(j+1)}} \qquad (1 \le i < j \le k).$$

4) *Linear (L) Contrasts.* These contrasts were proposed by Rom et al. (1994); the corresponding t-tests are equivalent to the tests for trend using ordinal scaling proposed by Tukey, Ciminera, and Heyse (1985). The general form of these contrasts is given by

$$a_{ij} = \begin{cases} -i, & j = 0, \\ a_{i,j-1} + 2, & j = 1, \ldots, i, \\ 0, & j = i+1, \ldots, k. \end{cases} \tag{3.6}$$

The correlations among these contrasts cannot be expressed in a simple form.

Ruberg (1989) proposed two other contrasts that he called step and basin contrasts. Unfortunately, the procedures based on both these contrasts do not control the type I FWE except under the overall null hypothesis. Ruberg (1989) noted this fact based on his simulation results, but he did not offer a theoretical explanation, nor did he advise against their use. We now offer a theoretical explanation for the excessive type I error rates.

The coefficients for step contrasts are given by

$$a_{ij} = \begin{cases} -(k-i+1), & j = 0, \ldots, i-1, \\ i, & j = i, \ldots, k. \end{cases}$$

Thus, the ith step contrast compares the average of the means of the highest $k-i+1$ dose levels with the average of all the lower dose levels (including the zero dose level).

The coefficients for basin contrasts are given by

$$a_{ij} = \begin{cases} -(k-i+1)(k-i+2)/2, & j = 0, \ldots, i-1, \\ a_{i,j-1} + k + 1, & j = i, \ldots, k. \end{cases}$$

The ith basin contrast is intended to compare the average of the means of the zero dose level and the first $i-1$ dose levels with the weighted average of the means of the $k-i+1$ highest dose levels ($1 \le i \le k$) where the weights increase linearly with the dose level. However, the preceding basin contrasts defined by Ruberg (1989) do not achieve this intended objective because some of the a_{ij} for $j > i$ are also negative. For example, see Table 1, where a_{10}, a_{11}, and a_{12} are all negative for basin contrast 1; thus, that contrast compares doses 0, 1, and 2 with higher doses instead of comparing the zero dose with all the nonzero doses. To remove this anomaly, we could revise the definition of the basin contrasts as follows:

$$a_{ij} = \begin{cases} -(k-i+1)(k-i+2)/2, & j = 0, \ldots, i-1, \\ i(j-i+1), & j = i, \ldots, k. \end{cases}$$

Table 1 gives the values of the step, basin, and revised basin contrasts for $k = 5$.

A common feature of the step and basin contrasts (including the revised basin contrasts, defined earlier) is that the ith contrast puts positive weights on some dose levels $j > i$. Therefore, it has a positive mean under H_{0i} if $\mu_j > \mu_0$ for $j > i$, which results in the corresponding t_i-statistic having a noncentral rather than a central t-distribution. Hence, the tests based on these contrasts tend to reject too often and do not control the FWE.

Table 1
Step and basin contrasts for $k = 5$

No.	Step contrast	Basin contrast	Revised casin contrast
1	$(-5, 1, 1, 1, 1, 1)$	$(-15, -9, -3, 3, 9, 15)$	$(-15, 1, 2, 3, 4, 5)$
2	$(-4, -4, 2, 2, 2, 2)$	$(-10, -10, -4, 2, 8, 14)$	$(-5, -5, 1, 2, 3, 4)$
3	$(-3, -3, -3, 3, 3, 3)$	$(-6, -6, -6, 0, 6, 12)$	$(-2, -2, -2, 1, 2, 3)$
4	$(-2, -2, -2, -2, 4, 4)$	$(-3, -3, -3, -3, 3, 9)$	$(-3, -3, -3, -3, 4, 8)$
5	$(-1, -1, -1, -1, -1, 5)$	$(-1, -1, -1, -1, -1, 5)$	$(-1, -1, -1, -1, -1, 5)$

Ruberg's (1989) simulations were made for $k = 4$ and $\alpha = .05$. He found that the FWE of his basin contrasts procedure ranged between 15 and 20% for the five partial null configurations that he considered. The FWE of his step contrasts procedure was about 9% for two configurations with MED $= 4$ and $\mu_4 = .5,\ 1.0$, and less than 5% for the other three configurations with MED < 4 and $\mu_4 = 1.5,\ 2.0$; here, $\mu_0 = 0$ and $\sigma/\sqrt{n} = 1/\sqrt{10}$. Why does the step contrasts procedure control the FWE under the latter three configurations but not under the former two? Both procedures are single-step and they identify the MED with the contrast producing the $t_{\max}$ statistic if the latter is significant. Therefore, under the latter three configurations where $\mu_{\max} = \mu_4$ is large, the procedure identifies either dose level 3 or 4 as the MED with high probability thus resulting in very few type I errors. In a sense, the procedure is too conservative in these cases because (i) it only considers the *maximum* t-statistic when actually the goal is to identify the *minimum* effective dose and, more importantly, (ii) it operates in a single-step manner. If it were a step-down procedure then it would have the opportunity to test the lower doses, thus making the procedure less conservative; see the discussion of the simulation results in Section 5.2.

3.2.2 Type of testing scheme. The hypotheses H_{0i} can be tested using the statistics t_i in a single-step or stepwise manner (Hochberg and Tamhane, 1987). However, because we are only interested in tests here and not in confidence interval estimation, stepwise procedures offer a more powerful alternative. Stepwise procedures are of two types: (i) step-down and (ii) step-up. For each set of contrasts, we consider two step-down and two step-up procedures as described here.

Step-Down Procedure 1 (SD1). The closed step-down (SD) procedure proposed by various authors (Miller, 1966; Naik, 1975; and Marcus et al., 1976) for comparing unordered treatments with a control can be applied to the present problem as follows. First, consider P and H contrasts, which have equally correlated t-statistics with common correlation ρ ($\rho = 1/2$ for P contrasts and $\rho = 0$ for H contrasts when $n_0 = n$). Compute the t-statistics using (3.2) and order them: $t_{(1)} \le t_{(2)} \le \cdots \le t_{(k)}$. Let $H_{0(1)}, H_{0(2)}, \ldots, H_{0(k)}$ be the corresponding null hypotheses. At the first step, let $k_1 = k$ be the number of hypotheses still to be tested. Compare $t_{(k_1)}$ with $c_{k_1} = t^{(\alpha)}_{k_1,\nu,\rho}$, the upper α equicoordinate critical point of the equicorrelated k_1-variate t-distribution with ν df and common correlation ρ. If $t_{(k_1)} \ge t^{(\alpha)}_{k_1,\nu,\rho}$, then reject $H_{0(k_1)}$ and all hypotheses whose rejection is implied by it[1] (i.e., if $H_{0(k_1)} = H_{0m}$, then reject H_{0j} for $j = m, m+1, \ldots, k_1$) and go to the second step with $k_2 = m - 1$, the number of hypotheses still to be tested; otherwise, stop testing and accept all hypotheses. In general, at the ith step let k_i be the number of hypotheses still to be tested. Relabel the ordered statistics $t_{(1)} \le t_{(2)} \le \cdots \le t_{(k_i)}$ and the corresponding hypotheses as $H_{0(1)}, H_{0(2)}, \ldots, H_{0(k_i)}$. Test $H_{0(k_i)}$ by comparing $t_{(k_i)}$ with $c_{k_i} = t^{(\alpha)}_{k_i,\nu,\rho}$. Reject $H_{0(k_i)}$ if $t_{(k_i)} \ge t^{(\alpha)}_{k_i,\nu,\rho}$ and all hypotheses whose rejection is implied by it and go to the next step; otherwise, stop testing. When testing stops, estimate the MED as the minimum index of the rejected hypotheses.

For R and L contrasts, we can use the extension of SD1 to unequally correlated t-statistics by Dunnett and Tamhane (1991). Note that in this extension the critical point needs to be recomputed at each step because of the changing correlation matrix $\{\rho_{ij}\}$. In the present work, we used Schervish's (1984) algorithm to calculate these critical points exactly. The four SD1 procedures for P, H, R, and L contrasts will be denoted by SD1P, SD1H, SD1R, and SD1L, respectively. It may be noted that SD1P is the stepwise analog of the single-step test procedure of Dunnett (1955).

REMARK. SD1L was proposed earlier by Rom et al. (1994, Section 2.3), who further extended it to the closed family formed by the intersection of all hypotheses of the equality of subsets of successive μ_i's. Although these authors used a procedure that did not include the Shaffer-type modification, it is easy to see that their procedure is equivalent to SD1L for the purpose of determining the MED.

Step-Down Procedure 2 (SD2): As noted earlier, the family of hypotheses under test is a closed family. Using this fact, a simpler closed step-down procedure that controls the FWE and does not require ordering of the t-statistics is as follows: reject H_{0i} iff each H_{0j} is significant for all $j \ge i$ using an ordinary α-level t-test, that is, $t_j \ge c = t^{(\alpha)}_\nu$ where $t^{(\alpha)}_\nu$ is the upper α critical point of Student's t-distribution with ν df.

SD2 has the advantage of being simple and flexible. Because the critical point from Student's t does not depend on the correlations among the contrasts, SD2 can be used with complex unbalanced

[1] This modification where certain hypotheses are rejected without actually testing them is similar to Shaffer's (1986) modification of step-down testing procedures for pairwise comparisons.

designs (e.g., unbalanced multiway designs with covariates) involving unequal correlations among the contrasts. In fact, all that is needed is *any* α-level test to compare each dose with the zero dose; this test need not be a t-test but could be, for example, a nonparametric test. Note that the Shaffer-type modification is not needed here because we test the hypotheses in the order of dose levels. The four SD2 procedures for P, H, R, and L contrasts will be denoted by SD2P, SD2H, SD2R, and SD2L, respectively.

REMARK. SD2L was used in an example by Tukey et al. (1985). It was presented formally by Rom et al. (1994, Section 2.1), who extended it to the same closed family of hypotheses as in their extension of SD1L.

Step-Up Procedure 1 (SU1): The step-up (SU) procedure proposed by Dunnett and Tamhane (1992) for comparing unordered treatments with a control can be applied to the present problem as follows. First, consider P and H contrasts, which have equally correlated t-statistics with common correlation ρ ($\rho = 1/2$ for P contrasts and $\rho = 0$ for H contrasts when $n_0 = n$). Let $c_1 < c_2 < \cdots < c_k$ be the critical constants for SU for given k, ν, α, and ρ; these constants are tabulated in Dunnett and Tamhane (1992). The procedure uses ordered t-statistics, $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)}$, and operates as follows: Test $H_{0(1)}$ by comparing $t_{(1)}$ with c_1. If $t_{(1)} \geq c_1$, reject all hypotheses and stop testing; otherwise, proceed to test $H_{0(2)}$. In general, test the hypothesis $H_{0(i)}$ iff $t_{(j)} < c_j$ for $j = 1, \ldots, i-1$. If $t_{(i)} < c_i$, then proceed to test $H_{0(i+1)}$; otherwise, stop testing and reject the hypotheses $H_{0(i)}, H_{0(i+1)}, \ldots, H_{0(k)}$, and any hypotheses whose rejection is implied by them (which were accepted at earlier steps). In effect, estimate the MED as $\widehat{\text{MED}} = i^* = \min\{(i), \ldots, (k)\}$. This procedure controls the FWE strongly because the Dunnett and Tamhane procedure does as shown in their paper.

The extension of this procedure to unequally correlated t-statistics given in Dunnett and Tamhane (1995) can be used for R and L contrasts. This requires recomputing of the critical point at each step because of the changing correlation matrix $\{\rho_{ij}\}$.

Note that the implied hypotheses are rejected here only at the last step. This has no effect on the power because the critical constants used for testing are not changed. On the other hand, the power of SD1 is enhanced because of this Shaffer-type modification because it uses smaller critical constants appropriate for the reduced number of hypotheses remaining to be tested at each step. For this and other reasons, we found in our simulations that for a given contrast type (P, H, R, or L) SD1 generally dominates SU1 although the difference in their powers is never substantial. To demonstrate this point, we shall use the SU1P procedure, which is directly comparable with SD1P. The results for other SU1 procedures are not reported to save space.

Step-Up Procedure 2 (SU2). Analogous to SD2 we can construct a step-up procedure based on unordered t-statistics and a common critical constant c. This procedure would operate as follows: Begin by testing H_{01}. If $t_1 \geq c$, then stop testing and reject all hypotheses; otherwise, proceed to test H_{02}. In general, test H_{0i} iff $t_j < c$ for $j = 1, \ldots, i-1$. If $t_i \geq c$, then stop testing and reject $H_{0i}, \ldots, H_{0k}$; otherwise, go to the next step. If no hypotheses are rejected, then no dose is declared as the MED; otherwise, the MED is estimated as

$$\widehat{\text{MED}} = i^* = \min\{i : t_i \geq c\}.$$

Thus, any SU2 procedure can be viewed as a single-step procedure: Declare the lowest dose that yields $t_i > c$ as the MED.

Next, we show how to determine c so that the FWE is strongly controlled at a designated level α. For the sake of simplicity, we restrict to the equicorrelated case (P and H contrasts). Consider any true hypothesis H_{0i}. Then,

$$\begin{aligned} \text{FWE} &= P\{\text{Reject } H_{0i}\} \\ &= 1 - P\{t_1 < c, \ldots, t_i < c\}. \end{aligned}$$

Here, $t_1, \ldots, t_i$ have an i-variate t-distribution with ν df and common correlation ρ. The FWE is clearly maximum under H_{0k}. Therefore, the equation for determining c is

$$P\{t_1 < c, \ldots, t_k < c\} = 1 - \alpha,$$

the solution to which is $c = t^{(\alpha)}_{k,\nu,\rho}$. Ruberg (1989) proposed SU2H for which $c = t^{(\alpha)}_{k,\nu,0} = m^{(\alpha)}_{k,\nu}$, the upper α critical point of the Studentized maximum distribution (referred to incorrectly as the maximum modulus distribution in Ruberg's paper) of dimension k and df ν.

Now note that SD1 uses the same critical constant $c_k = t^{(\alpha)}_{k,\nu,\rho}$ at the first step and smaller critical constants $c_i = t^{(\alpha)}_{i,\nu,\rho}$ at subsequent steps. It is clear that, for a given set of contrasts, SD1 uniformly dominates SU2 (because essentially, as already noted, SU2 is a single-step procedure based on the same critical constant as is SD1) in the sense that its estimated MED is never higher than that of SU2.

We investigated a modification of SU2 that uses a sequence of critical constants, $c_1 < c_2 < \cdots < c_k$, instead of a common critical constant c. However, for the cases studied in simulations, this modified procedure offered little improvement. Therefore, we eliminated SU2 in both of its forms from consideration.

4. Example

In a dose response study, five dose levels are compared to the zero dose level in a balanced one-way layout ($n_0 = n$). Suppose that the sample means are:

$\bar{y}_0$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$
.0	1.5	2.1	1.9	2.3	2.1

Assume that standard deviation $(\bar{y}_i - \bar{y}_0) = \sigma\sqrt{2/n} = 1.0$ and df $= \infty$. Find the MED subject to the condition that the FWE is controlled at $\alpha = .05$.

Procedure WILM. The first step is to calculate the $\hat{\mu}_i$, which are as follows:

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
1.5	2.0	2.0	2.2	2.2

The next step is to compute the $\bar{t}$-statistics, which are the same as the $\hat{\mu}_i$ because $\bar{y}_0 = 0$ and $\sigma\sqrt{2/n} = 1.0$. Hence,

$$\bar{t}_1 = 1.5,\ \bar{t}_2 = 2.0,\ \bar{t}_3 = 2.0,\ \bar{t}_4 = 2.2,\ \bar{t}_5 = 2.2.$$

The $\bar{t}$-statistics are compared with the following critical constants (taken from Williams, 1971, Table 1) in a step-down manner:

$$c_1 = 1.645,\ c_2 = 1.716,\ c_3 = 1.739,\ c_4 = 1.750,\ c_5 = 1.756.$$

It is easily checked that $\widehat{\text{MED}} = 2$.

Procedure SD1P. The t_i are the same as the $\bar{y}_i$ because $\bar{y}_0 = 0$ and $\sigma\sqrt{2/n} = 1.0$. The ordered t_i are

$$t_{(1)} = t_1 = 1.5,\ t_{(2)} = t_3 = 1.9,\ t_{(3)} = t_2 = 2.1,\ t_{(4)} = t_5 = 2.1,\ t_{(5)} = t_4 = 2.3.$$

The critical constants $c_i = t^{(.05)}_{i,\infty,1/2}$ for $i = 1, 2, \ldots, 5$ taken from Bechhofer and Dunnett (1988) are

$$c_1 = 1.645,\ c_2 = 1.916,\ c_3 = 2.062,\ c_4 = 2.160,\ c_5 = 2.234.$$

At the first step, because $t_{(5)} = t_4 = 2.3 \geq c_5 = 2.234$, we reject H_{04} and by implication also H_{05}; thus, $k_2 = 3$. At the second step, because $t_{(3)} = t_2 = 2.1 \geq c_3 = 2.062$, we reject H_{02} and by implication also H_{03}; thus, $k_3 = 1$. At the third step, because $t_{(1)} = t_1 = 1.5 < c_1 = 1.645$, we accept H_{01}. Hence, $\widehat{\text{MED}} = 2$.

Procedure SD2P. The unordered $t_i = \bar{y}_i$ are compared with a common critical constant $c = t^{(.05)}_{\infty} = 1.645$ in a step-down manner. It is easily checked that $\widehat{\text{MED}} = 2$.

Procedure SU1P. The ordered t-statistics given earlier under procedure SD1P are compared with the critical constants c_i for $i = 1, 2, \ldots, 5$ in a step-up manner. These critical constants for $\alpha = .05$, $\nu = \infty$, and $\rho = 1/2$ taken from Dunnett and Tamhane (1992) are as follows:

$$c_1 = 1.645,\ c_2 = 1.933,\ c_3 = 2.071,\ c_4 = 2.165,\ c_5 = 2.237.$$

Beginning with $t_{(1)} = 1.5$, the first significant t-statistic is $t_{(3)} = t_2 = 2.1 \geq c_3 = 2.071$. Testing is stopped at this point with the rejection of H_{02} and the remaining hypotheses, H_{05} and H_{04}. Then, $\widehat{\text{MED}} = \min\{2, 5, 4\} = 2$. Note that H_{03} is rejected by implication, although it was accepted at the second step. Note also that this has no real effect, because the MED would have been estimated to be 2 anyhow.

Procedure SD1H. The standard deviation of the ith Helmert contrast (for a common sample size n) is given by

$$\sigma\sqrt{\frac{(-1)^2}{n}+\cdots+\frac{(-1)^2}{n}+\frac{(i)^2}{n}}=\sigma\sqrt{\frac{i}{n}+\frac{i^2}{n}}=\sigma\sqrt{\frac{2}{n}}\sqrt{\frac{i(i+1)}{2}}=\sqrt{\frac{i(i+1)}{2}}.$$

Table 2 shows the calculation of the t-statistics.

Table 2
Calculation of the Helmert contrasts and their t-statistics

No.	Contrast vector	Contrast value	SD	t-statistic
1	$(-1,1,0,0,0,0)$	1.5	1	1.500
2	$(-1,-1,2,0,0,0)$	2.7	$\sqrt{3}$	1.559
3	$(-1,-1,-1,3,0,0)$	2.1	$\sqrt{6}$	.857
4	$(-1,-1,-1,-1,4,0)$	3.7	$\sqrt{10}$	1.170
5	$(-1,-1,-1,-1,-1,5)$	2.7	$\sqrt{15}$	.697

The ordered t-statistics,

$$t_{(1)}=t_5=0.697,\ t_{(2)}=t_3=0.857,\ t_{(3)}=t_4=1.170,\ t_{(4)}=t_1=1.500,\ t_{(5)}=t_2=1.559,$$

are compared with the critical constants, $c_i=t_{i,\nu,0}^{(\alpha)}=m_{i,\nu}^{(\alpha)}$, where $m_{i,\nu}^{(\alpha)}$ is the $100(1-\alpha)$ percentage point of the Studentized maximum distribution. These constants (taken from Bechhofer and Dunnett [1988]) are as follows:

$$c_1=1.645,\ c_2=1.954,\ c_3=2.121,\ c_4=2.234,\ c_5=2.319.$$

Because $t_{(5)}=t_2=1.559<2.319$, we stop at the first step and accept all H_{0i}. Thus, no dose is declared as the MED.

Procedure SD2H. We begin by comparing $t_5=.697$ with $t_\infty^{(.05)}=1.645$. Because $0.697<1.645$, testing stops at the first step, and no dose is declared as the MED.

Procedure SD1R. The standard deviation for the ith contrast (for a common sample size n) is the same as that for the Helmert contrast, viz., $\sqrt{i(i+1)/2}$. Table 3 shows the calculation of the t-statistics.

Table 3
Calculation of the reverse Helmert contrasts and their t-statistics

No.	Contrast vector	Contrast value	SD	t-statistic
1	$(-1,1,0,0,0,0)$	1.5	1	1.500
2	$(-2,1,1,0,0,0)$	3.6	$\sqrt{3}$	2.078
3	$(-3,1,1,1,0,0)$	5.5	$\sqrt{6}$	2.245
4	$(-4,1,1,1,1,0)$	7.8	$\sqrt{10}$	2.467
5	$(-5,1,1,1,1,1)$	9.9	$\sqrt{15}$	2.556

The ordered t-statistics,

$$t_{(1)}=t_1=1.500,\ t_{(2)}=t_2=2.078,\ t_{(3)}=t_3=2.245,\ t_{(4)}=t_4=2.467,\ t_{(5)}=t_5=2.556,$$

are compared with the following critical constants in a step-down manner:

$$c_1=1.645,\ c_2=1.817,\ c_3=1.890,\ c_4=1.931,\ c_5=1.957.$$

It is easily checked that $\widehat{\text{MED}}=2$.

Procedure SD2R. The unordered t-statistics are tested against $t_\infty^{(.05)}=1.645$ in a step-down manner beginning with t_5. It is easily checked that $\widehat{\text{MED}}=2$.

Procedure SD1L. Table 4 shows the calculation of the t-statistics.

The ordered t-statistics are:

$$t_{(1)} = t_1 = 1.500,\ t_{(2)} = t_3 = 1.992,\ t_{(3)} = t_2 = 2.100,\ t_{(4)} = t_5 = 2.147,\ t_{(5)} = t_4 = 2.236,$$

which are compared with the following critical constants in a step-down manner:

$$c_1 = 1.645,\ c_2 = 1.916,\ c_3 = 2.060,\ c_4 = 2.155,\ c_5 = 2.224.$$

Table 4
Calculation of the linear contrasts and their t-statistics

No.	Contrast vector	Contrast value	SD	t-statistic
1	$(-1, 1, 0, 0, 0, 0)$	1.5	1	1.500
2	$(-1, 0, 1, 0, 0, 0)$	2.1	1	2.100
3	$(-3, -1, 1, 3, 0, 0)$	6.3	$\sqrt{10}$	1.992
4	$(-2, -1, 0, 1, 2, 0)$	5.0	$\sqrt{5}$	2.236
5	$(-5, -3, -1, 1, 3, 5)$	12.7	$\sqrt{35}$	2.147

It is easily checked that $\widehat{\text{MED}} = 2$.

Procedure SD2L. The unordered t-statistics are tested against $t_{\infty}^{(.05)} = 1.645$ in a step-down manner beginning with t_5. It is easily checked that $\widehat{\text{MED}} = 2$.

In this example, the Helmert contrasts fail to detect any dose as the MED. The reason for this is that the dose–response function is very nearly step-shaped with the true MED equal to about 2. For this configuration, contrast 2 would be the most significant one (as can be checked from Table 2 because for higher order contrasts the higher dose means tend to cancel out with the lower dose means (which are nearly equal) producing a smaller (in magnitude) contrast. The standard deviations of the contrasts increase because of the increasing number of dose means involved in their calculation. The net effect is that the higher order contrasts become less and less significant. SD2H will have very little power in this case because it begins testing with the highest order contrast, which is likely to be the least significant, as is the case in this example where $t_5 = .697$ is the smallest t-statistic. SD1H has a better chance of identifying the MED (although it uses larger critical constants than SD2H does) because it begins by testing the most significant contrast. In the present example, the second contrast is the most significant one; however, $t_2 = 1.559$ is not large enough to exceed the critical constant $m_{5,\infty}^{(.05)} = 2.319$.

Before concluding this example, we explain how the critical constants for SD1R and SD1L are computed because they involve unequally correlated multivariate t-distributions (normal distributions in the present case because df $= \infty$). For illustration, consider SD1L. The correlation matrix of the five L contrasts is as follows:

$$\begin{bmatrix} 1 & .5000 & .3162 & .2236 & .1690 \\ & 1 & .6325 & .4472 & .3381 \\ & & 1 & .7071 & .5345 \\ & & & 1 & .7559 \\ & & & & 1 \end{bmatrix}.$$

The c_i's are the upper .05 equicoordinate critical points of the multivariate t-distributions with correlation matrices that are appropriate submatrices (based on the observed ordering of the t-statistics) of the preceding matrix. For example, $c_1 = 1.645$ is the upper .05 critical point of the null distribution of t_1. Next, $c_2 = 1.916$ is the upper .05 equicoordinate critical point of the null distribution of $(t_{(1)}, t_{(2)}) = (t_1, t_3)$, which has the correlation coefficient $\rho_{13} = .3162$. Continuing, $c_3 = 2.060$ is the upper .05 equicoordinate critical point of the null distribution of $(t_{(1)}, t_{(2)}, t_{(3)}) = (t_1, t_3, t_2)$, which has the correlation coefficients $\rho_{13} = .3162$, $\rho_{12} = .5000$, $\rho_{32} = .6325$, and so on. Because the correlations were unequal, we used Schervish's (1984) algorithm to compute the c_i's exactly.

5. Simulation Results

5.1 *Design of the Simulation Study*

The 10 procedures described in Section 4 were compared in a large simulation study. Throughout the study, k, the number of positive dose levels was fixed at 5 and α was fixed at .05. A common sample size, n, was assumed per group. Without loss of generality, μ_0 was fixed at 0 and the standard error of the means, $\sigma/\sqrt{n}$, was fixed at 1. The df was assumed to be ∞.

The positive dose means were selected as follows:

(a) *Monotone Dose Response:* In this case, we considered two types of response functions: (i) linear response and (ii) step response. Denote the value of the largest mean, μ_5, by δ. For each type of response function, two values of δ were selected: 3 and 5. For each combination of the type of response function and the value of δ, the MED was varied over the five dose levels; for MED = 5, the linear and step response configurations coincide. This gives a total of 8 + 10 = 18 monotone configurations of dose means.

(b) *Nonmonotone Dose Response:* Nonmonotone mean configurations were selected as follows: Only linear response was considered with MED = 2 and 3, $\mu_4 = \delta$ and $\mu_5 = \delta/2$ and 0. For $\delta = 3$ and 5, this gives eight nonmonotone configurations of dose means.

Additional configurations were also studied, but here we report the results only for the preceding $18 + 8 = 26$ configurations.

For each configuration $(\mu_0, \mu_1, \ldots, \mu_k)$, the simulation run was carried out as follows: A vector of sample means $(\bar{y}_0, \bar{y}_1, \ldots, \bar{y}_k)$ was generated where the $\bar{y}_i$ are independent $N(\mu_i, 1)$ r.v.'s. Then, each procedure was applied to these same data. The critical constants used by the procedures are the same ones used in the example of Section 4. The MED identified by each procedure was noted. This was replicated 10,000 times for each of the 26 configurations.

For each run, several summary statistics were computed for comparing the procedures. These statistics are as follows:

(a) The proportions of replications (out of the total number, $N = 10,000$) for which each dose level is identified as the MED by each procedure. These proportions give the operating characteristic of each procedure. (i) The proportion corresponding to noneffective doses gives an estimate of the FWE of the procedure. (ii) The proportion corresponding to the true MED gives an estimate of the power of the procedure. (iii) The proportion corresponding to dose levels higher than the true MED (including instances where none of the dose levels is declared as the MED) gives an estimate of the lack of power of the procedure. The first two proportions are given in Table 5 for monotone configurations and in Table 7 for nonmonotone configurations.

(b) The estimated bias of $\widehat{\text{MED}}$, where the bias is defined as $E(\widehat{\text{MED}}) - \text{MED}$. This is another measure of the power of a procedure, the lower the bias the higher the power. In estimating $E(\widehat{\text{MED}})$, we adopted the convention of assigning $\widehat{\text{MED}} = 6$ if none of the five dose levels were identified as the MED. The estimated biases are given in Table 6 for monotone configurations and in Table 8 for nonmonotone configurations.

(c) For each configuration and each pair of procedures, the number of replications (out of a total of 10,000) in which one procedure identifies a lower or higher or the same MED compared to the other procedure. For a given configuration and a pair of procedures (i, j), denote these numbers by M_{ij}, M_{ji} and $10,000 - (M_{ij} + M_{ji})$, respectively. These numbers are summarized in Table 9 as follows. If $M_{ij} > M_{ji}$, then procedure i is said to beat procedure j for that configuration. Table 9 gives the number of configurations, N_{ij} (out of a total of 26) that procedure i beats procedure j; note that $N_{ij} + N_{ji} = 26$. The last column of Table 9 gives the overall rank of each procedure based on the N_{ij} values as follows: If $N_{ij} > 13$, then procedure i beats procedure j for a majority of configurations. The procedure that beats most procedures for a majority of configurations was assigned rank 1, and so on.

5.2 *Discussion of Simulation Results*

We first discuss briefly the simulation results obtained for two step-down procedures that use step contrasts (denoted by SD1S and SD2S). Note that Ruberg (1989) used a single-step procedure based on the $t_{\max}$ statistic, which, as noted earlier, is not appropriate for identifying the MED; hence, we used step-down procedures SD1 and SD2. Our procedures differ from Ruberg's in another respect: whereas Ruberg used a conservative approximation to the exact critical point computed by replacing the correlation coefficients ρ_{ij} by a common correlation $\rho_{\min}$, we used the exact critical

constants for SD1S; the critical constant for SD2S is the Student's t percentage point, $t_{\infty}^{(\alpha)}$, which does not depend on the correlations. The effect of both these deviations from Ruberg's procedure is to make SD1S and SD2S more anticonservative than he found in his simulations.

The FWE of SD1S for monotone partial null configurations (MED > 1) ranged between .350 and .994 with an average of .720, whereas the FWE of SD2S ranged between .191 and .970 with an average of .616 (recall that nominal $\alpha = .05$). Thus, both these procedures have extremely high FWEs. As will be seen later, SD1 and SD2 procedures based on P, H, R, and L contrasts do control the FWE. So the cause of high FWEs is the step contrasts. For this reason, we do not include the simulation results for SD1S and SD2S in our tables. Although Ruberg's single-step procedure has a much lower FWE, it still does not control it.

Next we compare the simulation results for the 10 procedures that do control the FWE. First, we consider monotone configurations. Table 5 gives the sample estimates of the FWE (upper entry) and the power (lower entry) for each procedure under each monotone configuration. Configurations with true MED = 1 involve no type I errors, so the upper entry of estimated FWE = .000 is omitted for all procedures.

First, note that all 10 procedures control the FWE quite accurately at $\alpha = .05$ under all partial null configurations. (The estimated FWE must exceed $.05 + 1.96\sqrt{.05 \times .95/10,000} = .0543$ in order to conclude that it is significantly different from [higher than] $\alpha = .05$.)

Table 5

Estimated FWE (upper entry)[a] and power (lower entry)

Response function	True MED	Procedure									
		WILM	SD1P	SD2P	SU1P	SD1H	SD2H	SD1R	SD2R	SD1L	SD2L
Linear	1	.062	.070	.046	.071	.039	.001	.093	.086	.066	.045
($\delta = 3$)	2	.024	.030	.018	.028	.016	.001	.040	.037	.027	.018
		.058	.050	.047	.048	.054	.011	.028	.044	.053	.071
	3	.034	.041	.025	.041	.028	.004	.048	.047	.040	.030
		.098	.061	.102	.062	.083	.070	.021	.043	.060	.121
	4	.039	.046	.033	.047	.041	.015	.048	.049	.044	.041
		.198	.108	.217	.107	.171	.292	.020	.043	.093	.206
Linear	1	.134	.140	.116	.144	.097	.018	.160	.157	.138	.115
($\delta = 5$)	2	.035	.036	.029	.038	.023	.006	.044	.044	.035	.029
		.148	.114	.149	.112	.137	.104	.058	.086	.121	.178
	3	.040	.044	.035	.044	.034	.016	.048	.051	.042	.041
		.239	.154	.259	.154	.223	.324	.042	.074	.145	.257
	4	.047	.049	.043	.048	.041	.038	.048	.049	.045	.048
		.461	.307	.504	.306	.476	.682	.047	.095	.251	.428
Step	1	.538	.592	.342	.559	.475	.001	.653	.638	.580	.277
($\delta = 3$)	2	.046	.045	.042	.044	.036	.002	.047	.047	.045	.043
		.503	.489	.357	.473	.574	.045	.229	.287	.519	.556
	3	.040	.043	.037	.040	.037	.011	.044	.044	.041	.041
		.508	.453	.410	.451	.615	.254	.105	.167	.437	.578
	4	.049	.047	.046	.047	.043	.027	.047	.050	.045	.049
		.529	.428	.488	.428	.633	.575	.065	.119	.356	.544
	5	.047	.053	.045	.052	.052	.043	.051	.052	.052	.053
		.604	.416	.683	.415	.632	.850	.040	.089	.300	.505
Step	1	.960	.963	.892	.958	.927	.018	.970	.969	.962	.826
($\delta = 5$)	2	.052	.052	.052	.052	.051	.019	.052	.052	.052	.052
		.903	.890	.858	.886	.928	.348	.537	.605	.896	.917
	3	.050	.049	.050	.048	.048	.040	.047	.048	.048	.050
		.906	.879	.876	.878	.938	.786	.282	.377	.854	.907
	4	.049	.046	.049	.046	.047	.047	.045	.047	.048	.051
		.908	.868	.899	.867	.942	.926	.164	.249	.793	.885
	5	.053	.049	.053	.049	.052	.053	.051	.050	.048	.052
		.909	.856	.971	.855	.935	.998	.100	.182	.729	.855
Average power		.481	.435	.327	.432	.493	.350	.201	.239	.409	.460
Rank		2	4	8	5	1	7	10	9	6	3

[a] For MED = 1, the upper entry equals .000 for all procedures, and is hence omitted.

Looking at the lower entries, which are the estimates of power, we see that no procedure is uniformly "best" or "worst" for all monotone configurations, but some general trends are apparent. SD1H, WILM, and SD2L have the highest average powers, and their performances are relatively stable for different values of MED, δ, and the type of response function. At the other end, SD1R, SD2R, SD2P, and SD2H have the lowest average powers. There are some switches between these two groups of procedures for different configurations. For example, SD1R and SD2R have the highest powers for MED = 1, but their powers drop off rapidly reaching nadir for MED = 3, 4, or 5. On the other hand, SD2H has the lowest power when MED = 1 or 2 (with SD1H having the second lowest power when MED = 1 for linear response) but has one of the highest powers when MED = 4 or 5. Thus, R contrasts are good for detecting low MEDs, whereas H contrasts are good for detecting high MEDs. Why this is so can be easily seen from the nature of the contrasts in each case. This also agrees with the results of the example in Section 4, where the MED was low and H contrasts lacked power to identify it. SD2L is more powerful than SD1L in all cases when MED > 1. (This is also true of SD1R and SD2R.)

Finally, note that SD1P and SU1P have very similar powers that are always in the medium range. Also, the type of response function (linear vs. step) does not seem to make much difference in terms of the relative performances of these two procedures.

These power results are corroborated by the bias estimates for the same monotone configurations given in Table 6. In terms of the lowest average bias, the ranking is SD2L, WILM, and SD1L followed by SD1H, while SD2H, SD1R, and SD2R have the highest biases.

Next, we turn to the simulation results for nonmonotone configurations. From the upper entries in Table 7, we see that the FWE is controlled by all procedures for the eight configurations studied. This result is somewhat surprising for WILM, because it makes explicit use of the monotonicity assumption. Once again, SD1H, SD2L, and WILM have the highest average powers, whereas SD2H, SD1R, SD2P, and SD2R have the lowest average powers. For all except one configuration, SD2H has the lowest power bordering close to zero. Because of the dip in the response function at the highest dose level, all SD2 procedures have very low powers, the only exception being SD2L, which maintains surprisingly high power (relative to other SD2 procedures).

Turning to Table 8, we see that in terms of the estimated bias the ranking is SD1L, SD1H, SD1P, SU1P, SD2L, WILM, SD2R, SD1R, SD2P, and SD2H. The last four procedures have much higher biases than the others. Here, SD2L and WILM are ranked lower compared to the monotone configurations case (Table 6).

Finally, we come to Table 9, which gives an overall summary of the performances of the eight procedures for all 26 configurations, as explained earlier. We see that SD1H, SD2L, and SD1L rank highest in the overall ranking, followed by WILM, SD1P, SU1P, and SD2P, whereas SD2R, SD1R, and SD2H rank lowest. This ranking generally agrees with the rankings based on other criteria considered earlier.

6. Concluding Remarks

The present study has been limited to the equal sample size case, and its conclusions need to be extended to the unequal sample size case. Unequal sample sizes make the correlations ρ_{ij} unequal for P and H contrasts; in particular, the H contrasts are no longer uncorrelated.

There is no difficulty extending the SD2 procedures since the Student's t percentage point used by these procedures does not depend on the correlation structure. The critical constants required by the SD1 and SU1 procedures can also be calculated either exactly using Schervish's (1984) algorithm or approximately using the average correlation method given in Dunnett and Tamhane (1991, 1995). For WILM, the isotonic estimates $\hat{\mu}_i$ must be computed by a different formula given in Williams (1972); however, it is difficult to compute the exact critical constants for WILM because of the complicated nature of the joint distribution of the $\bar{t}_i$ statistics. Williams (1972) suggested that if the imbalance is moderate, in particular, if $.80 \leq n_i/n_k \leq 1.25$ for all $i = 1, \ldots, k-1$, then the tabulated constants for equal replication on nonzero dose levels are quite accurate.

Here is a summary of the findings and the contributions of the present study:

(a) We provided a framework for constructing stepwise testing procedures based on general contrasts to identify the MED.
(b) Using this framework, we proposed several new procedures.
(c) We pointed out a drawback of step and basin contrast procedures proposed by Ruberg (1989) in their inability to control the FWE. Any other contrast that puts nonzero weights on dose levels higher than the one under test for the MED shares the same drawback.

Table 6
Estimated bias for different procedures under selected monotone configurations

Response function	True MED	Procedure									
		WILM	SD1P	SD2P	SU1P	SD1H	SD2H	SD1R	SD2R	SD1L	SD2L
Linear	1	3.297	3.539	3.479	3.530	3.867	4.248	3.879	3.593	3.378	3.093
($\delta = 3$)	2	2.567	2.836	2.661	2.842	2.832	3.029	3.328	3.086	2.568	2.182
	3	1.786	2.057	1.807	2.057	1.918	1.892	2.528	2.371	1.859	1.503
	4	1.041	1.262	1.021	1.264	1.058	.887	1.681	1.595	1.217	.939
Linear	1	1.974	2.201	2.065	2.193	2.555	3.039	2.705	2.396	2.067	1.864
($\delta = 5$)	2	1.462	1.745	1.486	1.746	1.658	1.754	2.615	2.293	1.493	1.212
	3	.955	1.200	.933	1.199	.987	.855	2.148	1.907	1.068	.818
	4	.447	.623	.420	.626	.417	.247	1.501	1.353	.652	.471
Step	1	1.648	1.436	2.613	1.467	2.321	4.834	1.162	1.098	1.519	2.960
($\delta = 3$)	2	1.264	1.287	1.870	1.304	1.155	3.499	2.073	1.722	.882	.863
	3	.970	1.102	1.295	1.110	.708	1.914	2.128	1.850	.797	.527
	4	.615	.797	.723	.798	.426	.664	1.572	1.431	.750	.448
	5	.256	.380	.249	.385	.180	.080	.739	.714	.500	.360
Step	1	.089	.078	.357	.083	.291	4.635	.055	.050	.087	.692
($\delta = 5$)	2	.035	.054	.201	.057	−.016	2.272	.643	.468	.005	−.014
Step	3	.002	.039	.096	.040	−.052	.437	1.191	.919	.025	−.019
	4	−.025	.016	.014	.016	−.078	.002	1.225	1.024	.063	−.007
	5	−.061	−.041	−.046	−.038	−.114	−.054	.674	.626	.088	.008
Average bias		1.018	1.145	1.180	1.149	1.117	1.902	1.769	1.583	1.057	0.994
Rank		2	5	7	6	4	10	9	8	3	1

Table 7
Estimated FWE (upper entry) and power (lower entry) under selected nonmonotone configurations

True MED	$\mu_4 = \delta$	μ_5	Procedure									
			WILM	SD1P	SD2P	SU1P	SD1H	SD2H	SD1R	SD2R	SD1L	SD2L
2	3.0	1.5	.031	.032	.019	.031	.018	.000	.042	.040	.031	.021
			.088	.075	.056	.067	.085	.006	.041	.062	.081	.105
2	3.0	.0	.029	.034	.009	.025	.019	.000	.043	.041	.033	.015
			.078	.075	.016	.064	.080	.000	.043	.064	.083	.070
2	5.0	2.5	.041	.040	.036	.041	.029	.001	.047	.047	.040	.039
			.242	.185	.197	.170	.234	.034	.081	.119	.190	.268
2	5.0	.0	.041	.042	.012	.030	.031	.000	.049	.048	.041	.032
			.230	.191	.026	.164	.238	.000	.082	.117	.196	.213
3	3.0	1.5	.042	.043	.028	.038	.031	.002	.047	.049	.049	.039
			.158	.118	.102	.110	.162	.036	.034	.064	.112	.192
3	3.0	.0	.036	.039	.011	.032	.032	.000	.046	.046	.038	.030
			.118	.123	.027	.112	.169	.003	.035	.064	.119	.134
3	5.0	2.5	.050	.048	.045	.047	.044	.009	.048	.051	.042	.050
			.446	.336	.340	.318	.482	.149	.080	.124	.311	.459
3	5.0	.0	.046	.047	.013	.036	.041	.000	.049	.048	.047	.043
			.373	.346	.030	.321	.488	.001	.078	.133	.314	.384
Average power			.217	.181	.099	.166	.242	.029	.059	.093	.176	.228
Rank			3	4	7	6	1	10	9	8	5	2

(d) We conducted an extensive simulation study for both monotone and nonmonotone configurations to compare the various procedures. The following general conclusions and recommendations can be made based on this simulation study for the case of equal sample sizes:

i. In terms of power, SD1H, WILM, and SD2L are the best procedures when the dose–response function is monotone. The same procedures are best for non-monotone configurations also.

Table 8

Estimated bias for different procedures under selected non-monotone configurations

True MED	$\mu_4 = \delta$	μ_5	Procedure WILM	SD1P	SD2P	SU1P	SD1H	SD2H	SD1R	SD2R	SD1L	SD2L
2	3.0	1.5	2.627	2.541	3.207	2.563	2.503	3.818	3.151	2.909	2.199	2.165
2	3.0	.0	3.074	2.543	3.823	2.594	2.527	3.992	3.150	3.052	2.230	2.987
2	5.0	2.5	1.214	1.255	2.170	1.293	1.140	3.642	2.232	1.925	1.054	.904
2	5.0	.0	1.834	1.236	3.799	1.324	1.123	4.000	2.269	2.226	1.052	1.855
3	3.0	1.5	1.838	1.778	2.299	1.797	1.573	2.712	2.444	2.256	1.562	1.392
3	3.0	.0	2.300	1.817	2.837	1.846	1.553	2.983	2.483	2.408	1.652	2.101
3	5.0	2.5	.718	.717	1.470	.738	.471	2.379	1.933	1.670	.633	.486
3	5.0	.0	1.388	.716	2.848	.772	.474	2.997	1.994	1.949	.652	1.167
Average bias			1.874	1.575	2.807	1.616	1.421	3.315	2.457	2.299	1.379	1.632
Rank			6	3	9	4	2	10	8	7	1	5

Table 9

The number of configurations, N_{ij}, for which procedure i estimates a lower MED compared to procedure j

Procedure i	Procedure j WILM	SD1P	SD2P	SU1P	SD1H	SD2H	SD1R	SD2R	SD1L	SD2L	Rank
WILM	—	17	22	18	9	22	24	23	13	7	4
SD1P	9	—	14	21	4	18	24	24	6	7	5
SD2P	4	12	—	12	7	19	19	16	10	5	7
SU1P	8	5	14	—	4	18	24	24	6	7	6
SD1H	17	22	19	22	—	21	24	22	13	14	1
SD2H	4	8	7	8	5	—	12	11	7	4	10
SD1R	2	2	7	2	2	14	—	0	2	2	9
SD2R	3	2	10	2	4	15	26	—	2	2	8
SD1L	13	20	16	20	13	19	24	24	—	6	3
SD2L	19	19	21	19	12	22	24	24	20	—	2

Their performances are not affected much by the true shape of the dose–response function, as is the case with SD2H, SD1R, and SD2R.

ii. In terms of bias, SD2L, WILM, and SD1L are the best procedures when the dose–response function is monotone. SD1L, SD1H, and SD1P are best for nonmonotone configurations.

iii. On an overall basis, the best four procedures are SD1H, SD2L, SD1L, and WILM.

iv. Step-down procedures are preferred over step-up procedures. However, SD1 procedures should be used in their modified form where implied hypotheses are rejected at each step and the critical constant is accordingly reduced to take account of the reduced number of hypotheses. In this form, SD1 procedures perform slightly better than SU1 procedures.

v. Helmert contrasts should not be used with SD2, unless the MED is expected to be high. Reverse Helmert contrasts are generally not recommended, unless the MED is expected to be low. The best performing contrasts with both SD1 and SD2 are linear.

vi. SD2 procedures should not be used if the dose–response function can be *highly* nonmonotone, a definite possibility in vaccine studies.

vii. SD2P ranks rather low in terms of both power and bias. However, a great advantage of SD2P is that it can be easily extended to more complex unbalanced designs and can be used even when the normal theory assumptions are not satisfied (when the t-tests must be replaced by appropriate nonparametric tests). Under these conditions, SD2P is the only available option.

Finally, we note that this study has been restricted to procedures based on contrasts among the dose level means. It is possible to use SD2 procedures by testing each H_{0i} using a test of homogeneity such as Bartholomew's (1959a,b) test or Hayter's (1990) test against an ordered alternative. This

suggestion was made by Chase (1974) in the context of Bartholomew's test. We are currently investigating these procedures and plan to report our findings in a follow-up article.

Acknowledgements

Professor Dunnett's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Suggestions to investigate procedure SD2P were made to us by Dr. Ching-Ming Yeh of Ciba-Geigy and Dr. Peter Soo Ouyang of Janssen Research Foundation. Dr. Ouyang also suggested R contrasts. Dr. Dror Rom provided useful comments on the first draft. Finally, we thank an associate editor and two referees for useful suggestions.

Résumé

Nous considérons la situation où l'on cherche à identifier la plus faible dose pour laquelle la réponse diffère de celle de la dose zéro. Nous proposons un cadre général pour des procédures de tests pas à pas utilisant des contrastes sur les moyennes des différentes doses, à partir duquel plusieurs procédures nouvelles sont obtenues. Ces nouvelles procédures et d'autres plus anciennes, incluant celle de Williams (1971, *Biometrics* **27,** 103–117; 1972, *Biometrics* **28,** 519–531), sont comparées de manière analytique, ainsi que par une simulation détaillée, dans le cas du modèle à un facteur avec effectifs équilibrés et distribution normale. On montre que les procédures basées sur les contrastes dits 'step contrasts' et 'basin contrasts', proposées par Ruberg (1989, *Journal of the American Statistical Association* **84**, 816–822), ont des erreurs globales de type I (familywise error rate) exagérément élevées et qu'en conséquence elles ne doivent pas être utilisées. Les principaux résultats de la simulation sont les suivants. Dans le cas d'une configuration des moyennes monotone selon les doses, la procédure de Williams et deux procédures de test descendantes basées sur des contrastes de Helmert et des contrastes linéaires sont les plus performantes. Dans le cas de configurations non monotones, les performances de la procédure de Williams se dégradent quelque peu, tandis que les deux autres procédures sont encore les meilleures. Pour des plans plus complexes, une simple procédure descendante pas à pas, utilisant n'importe quel test de niveau alpha (pas nécessairement le test de Student) pour comparer chaque dose à la dose zéro, permet de contrôler le risque d'erreur global: cette procédure est la seule solution disponible, mais sa puissance est plutôt faible, en particulier dans le cas de configurations non monotones. Sur une famille donnée de contrastes, les procédures ascendantes se comportent généralement moins bien—mais pas beaucoup moins bien—que les procédures descendantes.

References

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika* **46,** 36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives *Biometrika* **46,** 328–335.

Bechhofer R. E. and Dunnett, C. W. (1988) Tables of percentage points of multivariate Student t distributions. In *Selected Tables in Mathematical Statistics* Volume 11, Providence, Rhode Island: American Mathematical Society, 1–371.

Budde, M. and Bauer, P. (1989). Multiple test procedures in clinical dose finding studies. *Journal of the American Statistical Association* **84,** 792–796.

Chase, G. R. (1974). On testing for ordered alternatives with increased sample size for control. *Biometrika* **61,** 569–578.

Crump, K. S. (1984). 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. and Tamhane, A. C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine* **10,** 939–947.

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. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics* **51,** 217–227.

Gaylor, D. W. (1983). The use of safety factors for controlling risk. *Journal Toxicology and Environmental Health* **11,** 329–336.

Hayter, A. J. (1990). A one-sided Studentized range test for testing against a simple ordered alternative. *Journal of the American Statistical Association* **85,** 778–785.

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

Marcus, R. (1976). The powers of some tests of the equality of normal means against an ordered alternative. *Biometrika* **63,** 177–183.

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

Miller, R. G. (1966). *Simultaneous Statistical Inference.* New York: McGraw Hill.

Naik, U. D. (1975). Some selection rules for comparing p processes with a standard. *Communications in Statistics, Ser. A (Theory and Methods)* **4,** 519–535.

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.

Ryan, L. (1992). Quantitative risk assessment for developmental toxicity. *Biometrics* **48,** 163–174.

Schervish, M. (1984). Multivariate normal probabilities with error bound. *Applied Statistics* **33,** 81–94.

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

Tukey, J. W., Ciminera, J. L., and Heyse, J. F. (1985). Testing the statistical certainty of a response with increasing doses of a compound. *Biometrics* **41,** 295–301.

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.

Received October 1994; *revised March* 1995; *accepted April* 1995.