Multiple Test Procedures for Identifying the Minimum Effective and Maximum Safe Doses of a Drug

Ajit C. TAMHANE and Brent R. LOGAN

We address the problem of determining the therapeutic window of a drug by finding its minimum effective and maximum safe doses (MINED and MAXSD). The MINED is the lowest dose that exceeds the mean efficacy of the zero dose by a specified threshold, and the MAXSD is the highest dose that does not exceed the mean toxicity of the zero dose by a specified threshold. Step-down multiple test procedures are proposed to identify the MINED and MAXSD assuming a bivariate normal model. These procedures control the type I familywise error probability of declaring any ineffective dose as effective or any unsafe dose as safe at a prespecified level α. A new multivariate t-distribution is introduced whose critical points are required to implement the exact normal theory procedures. Because these critical points depend on the unknown correlation coefficient between the efficacy and safety variables, the Bonferroni method is proposed as an alternative, which amounts to separately testing for efficacy and safety, each at type I familywise error rate of $\alpha/2$. The bootstrap versions of the exact normal theory procedures provide an *approximate* way to jointly test for efficacy and safety without the knowledge of the correlation coefficient, as well as to relax the bivariate normality assumption. The Bonferroni and bootstrap procedures are compared in a simulation study. It is shown that significant power gains are achieved by jointly testing for both efficacy and safety using bootstrap procedures. Coded data from an arthritis drug trial are analyzed to illustrate the procedures.

KEY WORDS: Bootstrap method; Closed testing procedure; Dose finding; Familywise error rate; Multiple comparisons; Multivariate t-distribution; Step-down procedure; Therapeutic window.

1. INTRODUCTION

Recommended doses of a drug are based on efficacy and safety considerations. It is common for efficacy as well as toxicity (as measured by adverse reactions and side effects) to increase over the range of doses, although at higher doses the efficacy may show a plateau or even a decline. These considerations lead one to define a *minimum effective dose* (MINED) and a *maximum safe dose* (MAXSD). The range delimited by these two doses is called a *therapeutic window*, which consists of doses that are both effective and safe. We address the problem of identifying the therapeutic window by finding the MINED and MAXSD simultaneously. Our approach is similar to that of Tamhane, Hochberg, and Dunnett (1996), who considered the problem of finding the MINED, and of Tamhane, Dunnett, Green, and Wetherington (2001), who considered the problem of finding the MAXSD. Simultaneous tests on efficacy and safety endpoints have been previously studied by Turri and Stein (1986), Thall and Russell (1998), Thall and Cheng (1999), and Jennison and Turnbull (1993).

The applicability of this article is limited to settings where efficacy and safety are evaluated in the same study using predetermined continuous endpoints. Generally, efficacy is evaluated at phase II/III, whereas safety is evaluated at all phases. Typically, efficacy endpoints are predetermined, whereas some safety endpoints may be unanticipated (e.g., rare adverse events). Also, many safety endpoints are of ordinal or count type. Statistics different than the t-statistics on which the procedures proposed in this article are based must be used for such data.

The outline of the article is as follows. Section 2 gives the problem formulation and a basic model that assumes that the efficacy and safety variables follow a bivariate normal distribution with unknown correlation coefficient ρ. Section 3 introduces a new multivariate t-distribution that arises in exact normal theory procedures. Section 4 develops exact step-down procedures. The critical constants of the new multivariate t-distribution required to implement these procedures cannot be evaluated, because they depend on the unknown ρ. Therefore, approximate procedures based on the Bonferroni inequality are proposed as alternatives. Another approximate way to bypass the difficulty of unknown ρ, as well as to relax the bivariate normal assumption, is to use bootstrap versions (Westfall and Young 1993) of the exact normal theory procedures. This is done in Section 5. Section 6 gives a real example to illustrate the proposed procedures, and Section 7 presents a simulation study that compares the procedures. Finally, Section 8 gives conclusions and recommendations.

2. PROBLEM FORMULATION AND NOTATION

Let $0, 1, \ldots, k$ denote increasing dose levels used in a dose-finding study, where 0 denotes the zero dose level (control). Suppose that n_i experimental units are tested at dose level i and let $N = \sum_{i=0}^{k} n_i$ denote the total sample size. Let (X_{ij}, Y_{ij}) be a bivariate random variable (rv) corresponding to the observed efficacy response x_{ij} and safety response y_{ij} of the jth experimental unit treated with dose i. We begin by assuming that the (X_{ij}, Y_{ij}) are independent bivariate normal with

$$E(X_{ij}) = \mu_i, E(Y_{ij}) = \eta_i, \quad \text{var}(X_{ij}) = \sigma^2, \text{var}(Y_{ij}) = \tau^2, \quad \text{and} \quad \text{corr}(X_{ij}, Y_{ij}) = \rho,$$

where all parameters are unknown.

Ajit C. Tamhane is Professor, Department of Statistics and Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208 (E-mail: *ajit@iems.northwestern.edu*). Brent R. Logan is Assistant Professor, Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI 53226 (E-mail: *blogan@mcw.edu*). The authors thank Ludwig Hothorn for suggesting the problem, Charles Dunnett for computing the entries in Table 1, and an associate editor and two referees for useful comments. The authors also thank a pharmaceutical company, which wishes to remain anonymous, for allowing access to their clinical study data for the example used in the article.

© 2002 American Statistical Association

Large values of the μ_i relative to μ_0 represent high efficacy, whereas large values of the η_i relative to η_0 represent high toxicity. Let $\delta_1 \geq 0$ and $\delta_2 \geq 0$ be two specified threshold constants for efficacy and toxicity. Any dose i with $\mu_i > \mu_0 + \delta_1$ is regarded as effective, and any dose j with $\eta_j < \eta_0 + \delta_2$ is regarded as safe. The true MINED and MAXSD are defined as

$$\text{MINED} = \min\{i\colon\ \mu_i > \mu_0 + \delta_1\}$$

and

$$\text{MAXSD} = \max\{j\colon\ \eta_j < \eta_0 + \delta_2\}. \tag{1}$$

The corresponding sample estimates (typically with $\delta_1 = \delta_2 = 0$) are referred to in the literature as MINED (minimum effective dose) and MTD (maximum tolerated dose).

If there is no dose i for which $\mu_i > \mu_0 + \delta_1$, then we say that MINED $> k$. Similarly, if there is no dose j for which $\eta_j < \eta_0 + \delta_2$, then we say that MAXSD < 1. If MINED $\leq$ MAXSD, then the therapeutic window is defined as [MINED, MAXSD]. For the MINED and MAXSD to be meaningful, all doses greater than MINED must be effective and all doses less than MAXSD must be safe, that is,

$$\begin{aligned} &\mu_i \leq \mu_0 + \delta_1\ \forall\, i < \text{MINED} \quad \text{and} \\ &\mu_i > \mu_0 + \delta_1\ \forall\, i \geq \text{MINED} \end{aligned}$$

and

$$\begin{aligned} &\eta_j < \eta_0 + \delta_2\ \forall\, j \leq \text{MAXSD} \quad \text{and} \\ &\eta_j \geq \eta_0 + \delta_2\ \forall\, j > \text{MAXSD}. \end{aligned} \tag{2}$$

This assumption is weaker than the strong monotonicity assumption $\mu_0 \leq \mu_1 \leq \cdots \leq \mu_k$ and $\eta_0 \leq \eta_1 \leq \cdots \leq \eta_k$, with at least one strict inequality in each case.

The goal is to estimate the MINED and MAXSD with the following probability requirement. Let $\widehat{\text{MINED}}$ and $\widehat{\text{MAXSD}}$ be the dose levels estimated as the MINED and MAXSD. Then we want

$$\begin{aligned} &P\{\widehat{\text{MINED}} < \text{MINED or } \widehat{\text{MAXSD}} > \text{MAXSD}\} \\ &\quad = P\{\text{an ineffective dose declared effective} \\ &\qquad \text{or an unsafe dose declared safe}\} \leq \alpha, \end{aligned} \tag{3}$$

where α is specified. If $\widehat{\text{MINED}}$ or $\widehat{\text{MAXSD}}$ does not exist, then we conclude that $\widehat{\text{MINED}} > k$ or $\widehat{\text{MAXSD}} < 1$. If $\widehat{\text{MINED}} > \widehat{\text{MAXSD}}$, then no therapeutic window is found.

3. A NEW MULTIVARIATE t-DISTRIBUTION

Denote by $\bar{x}_i$ and $\bar{y}_i (i = 0, 1, \ldots, k)$ the sample means, and by $\hat{\sigma}^2$ and $\hat{\tau}^2$ the usual mean squared error estimates of σ^2 and τ^2 with $\nu = N - (k+1)$ degrees of freedom (df).

The normal theory procedures in this article are based on the joint distributions of the pivotal rv's

$$T_i^{(1)} = \frac{\overline{X}_i - \overline{X}_0 - (\mu_i - \mu_0)}{\hat{\sigma}\sqrt{1/n_i + 1/n_0}} = \frac{U_i}{W_1}$$

and

$$T_j^{(2)} = \frac{\overline{Y}_0 - \overline{Y}_j - (\eta_0 - \eta_j)}{\hat{\tau}\sqrt{1/n_j + 1/n_0}} = \frac{V_j}{W_2} \tag{4}$$

for $i, j = 1, 2, \ldots, k$. Here the random vector $(U_1, U_2, \ldots, U_k; V_1, V_2, \ldots, V_k)$ has a joint $2k$-variate normal distribution with zero means, unit variances, and correlation structure

$$\begin{aligned} \text{corr}(U_i, U_j) = \text{corr}(V_i, V_j) &= \gamma_{ij} \text{ (say)} \\ &= \sqrt{\frac{n_i n_j}{(n_i + n_0)(n_j + n_0)}} = \lambda_i \lambda_j \quad \text{for } i \neq j, \end{aligned} \tag{5}$$

where $\lambda_i = \sqrt{n_i/(n_i + n_0)}$ and

$$\text{corr}(U_i, V_j) = \begin{cases} -\rho & \text{for } i = j \\ -\gamma_{ij}\rho & \text{for } i \neq j. \end{cases}$$

The γ_{ij} are known and ρ is unknown. Note that if all dosed groups have the same sample size, $n_i = n$, then

$$\gamma_{ij} \equiv \gamma = \frac{n}{n + n_0} \tag{6}$$

and $\lambda_i = \sqrt{\gamma}$ for all i. Finally,

$$W_1 = \frac{\hat{\sigma}}{\sigma} \sim \sqrt{\frac{\chi_\nu^2}{\nu}} \quad \text{and} \quad W_2 = \frac{\hat{\tau}}{\tau} \sim \sqrt{\frac{\chi_\nu^2}{\nu}}$$

independent of the values of U_i and V_j. It follows that the marginal distributions of $\mathbf{T}^{(1)} = (T_1^{(1)}, T_2^{(1)}, \ldots, T_k^{(1)})$ and $\mathbf{T}^{(2)} = (T_1^{(2)}, T_2^{(2)}, \ldots, T_k^{(2)})$ are k-variate t (Cornish 1954; Dunnett and Sobel 1954) with ν df and correlation matrix $\Gamma = \{\gamma_{ij}\}$ for $1 \leq i, j \leq k$. The joint distribution of $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$ involves the unknown ρ through the correlations between the values of U_i and V_j and the correlation between W_1 and W_2.

The step-down procedure SD1 proposed in Section 4.3.2 uses the upper α critical points of the rv $\max(\max_{1 \leq i \leq \ell} T_i^{(1)}, \max_{m \leq j \leq k} T_j^{(2)})$ for $1 \leq \ell, m \leq k$. The correlation matrix of $(U_1, U_2, \ldots, U_\ell; V_m, V_{m+1}, \ldots, V_k)$ can be written as

$$\begin{bmatrix} \Gamma_1 & -\rho\Gamma_{12} \\ -\rho\Gamma_{12}' & \Gamma_2 \end{bmatrix}, \tag{7}$$

where $\Gamma_1 = \{\gamma_{ij}\}$ for $1 \leq i, j \leq \ell$, $\Gamma_2 = \{\gamma_{ij}\}$ for $m \leq i, j \leq k$, $\Gamma_{12} = \{\gamma_{ij}\}$ for $1 \leq i \leq \ell$ and $m \leq j \leq k$, and Γ_{12}' denotes the transpose of Γ_{12}. Dependence of Γ_1, Γ_2, and Γ_{12} on $1, \ldots, \ell$ and $m, \ldots, k$ is suppressed for notational convenience. Note that for $\ell = k$ and $m = 1$, $\Gamma_1 = \Gamma_2 = \Gamma_{12} = \Gamma$.

We denote the required upper α critical point by $g_\alpha(\ell, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$, which is the solution in g to the equation

$$P\left\{ \max\left(\max_{1 \leq i \leq \ell} T_i^{(1)}, \max_{m \leq j \leq k} T_j^{(2)} \right) \leq g \right\} = 1 - \alpha. \tag{8}$$

If $\gamma_{ij} = \gamma$ given by (6) (which holds when $n_i = n$ for $i = 1, 2, \ldots, k$), then we use the simplified notation $g_\alpha(\ell, k - m + 1, \nu, \gamma, \rho)$.

If we put $\ell = 0$ or $m = k + 1$ in the foregoing, then we obtain the critical points of the multivariate t-distribution as special cases. In particular, if $\ell = 0$, then $g_\alpha(0, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$ is the upper α critical point of a $(k - m + 1)$-variate t-distribution with ν df and correlation matrix Γ_2, denoted by $h_\alpha(k - m + 1, \nu, \Gamma_2)$. Similarly, if $m =$

$k+1$, then $g_\alpha(\ell, 0, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$ equals $h_\alpha(\ell, \nu, \Gamma_1)$. Note that these two critical points do not depend on Γ_{12} and ρ.

To obtain an expression for the probability in (8), we first condition on (W_1, W_2). The joint probability density function (pdf) $f_\nu(w_1, w_2)$ of (W_1, W_2) was derived in earlier work (Tamhane and Logan 2000) using the Wishart distribution. Then by exploiting the product correlation structure (5), and a result from Bechhofer and Tamhane (1974) that exploits the block correlation structure (7) among the values of U_i and V_j, we obtain the following iterated integral expression for the probability in (8):

$$\int_0^\infty \int_0^\infty \left\{ \int_{-\infty}^\infty \int_{-\infty}^\infty \Psi_{\ell,m}[z_1, z_2, w_1, w_2 | \{\lambda_i\}, \rho] \times \phi_2(z_1, z_2 | -\rho)\, dz_1\, dz_2 \right\} f_\nu(w_1, w_2)\, dw_1\, dw_2, \qquad (9)$$

where $\phi_2(\cdot, \cdot | -\rho)$ is the pdf of the standard bivariate normal distribution with correlation coefficient $-\rho$, and

$$\Psi_{\ell,m}[z_1, z_2, w_1, w_2 | \{\lambda_i\}, \rho] = \begin{cases} \left[\prod_{i=m}^{\ell} \Phi_2 \left(\frac{g w_1 - \lambda_i z_1}{\sqrt{1-\lambda_i^2}}, \frac{g w_2 - \lambda_i z_2}{\sqrt{1-\lambda_i^2}} \middle| -\rho \right) \times \prod_{i=1}^{m-1} \Phi \left(\frac{g w_1 - \lambda_i z_1}{\sqrt{1-\lambda_i^2}} \right) \prod_{i=\ell+1}^{k} \Phi \left(\frac{g w_2 - \lambda_i z_2}{\sqrt{1-\lambda_i^2}} \right) \right] & \text{if } \ell \geq m \\ \left[\prod_{i=1}^{\ell} \Phi \left(\frac{g w_1 - \lambda_i z_1}{\sqrt{1-\lambda_i^2}} \right) \prod_{i=m}^{k} \Phi \left(\frac{g w_2 - \lambda_i z_2}{\sqrt{1-\lambda_i^2}} \right) \right] & \text{if } \ell < m. \end{cases}$$

In the foregoing, $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf), and $\Phi_2(\cdot, \cdot | -\rho)$ the cdf of the standard bivariate normal distribution with correlation coefficient $-\rho$.

Evaluation of (9) requires four-variate integration. However, the main difficulty is that the solution $g_\alpha(\ell, k-m+1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$ depends on the unknown ρ. As an approximation, one could use a pooled (from all dose groups) sample estimate of ρ. However, it turns out that the solution is relatively insensitive to ρ, as calculations for $g_\alpha(k, k, \nu, \gamma, \rho)$ for $k = 2(1)5, \nu = \infty, \gamma = 1/2$ and $\rho = .1(.2).7$ given in Table 1 show. This suggests that the Bonferroni upper bound on this critical constant, which does not require knowledge of ρ, would provide a good approximation. This upper bound is $h_{\alpha/2}(k, \nu, \gamma)$, which satisfies the equation

$$P\left\{ \max_{1 \leq i \leq k} T_i^{(1)} \leq h_{\alpha/2}(k, \nu, \gamma) \right\} = P\left\{ \max_{1 \leq j \leq k} T_j^{(2)} \leq h_{\alpha/2}(k, \nu, \gamma) \right\} = 1 - \frac{\alpha}{2}.$$

The last row of Table 1 shows the $h_{\alpha/2}(k, \nu, \gamma)$ values. Note that these values provide excellent approximations to $g_\alpha(k, k, \nu, \gamma, \rho)$ in all cases. Thus the Bonferroni critical values do not depend on the unknown ρ, are quite accurate, and are widely available for the $\gamma_{ij} \equiv \gamma$ case (see, e.g., Bechhofer and Dunnett 1988). An algorithm of Dunnett (1989) available at http://lib.stat.cmu.edu/apstat/251 can be used to compute these critical points for unequal correlations resulting from unequal values of n_i.

In another step-down procedure (SD2) described in Section 4.3.3, we require the upper α critical points of the rv's of the type $\max(T_i^{(1)}, T_j^{(2)})$ for $i, j = 1, 2, \ldots, k$. If $i \neq j$, then the desired critical point equals $g_\alpha(1, 1, \nu, \gamma_{ij}, \rho)$; in this case $\text{corr}(U_i, V_j) = -\gamma_{ij}\rho$. If $i = j$, then $\text{corr}(U_i, V_i) = -\rho$, and the distribution of $(T_i^{(1)}, T_i^{(2)})$ is Siddiqui's (1967) bivariate t. Denote this latter critical point by $g'_\alpha(1, 1, \nu, \rho)$, which is the solution in g' to the equation $P\{\max(T_i^{(1)}, T_i^{(2)}) \leq g'\} = 1 - \alpha$. The Bonferroni upper bound on both $g_\alpha(1, 1, \nu, \gamma_{ij}, \rho)$ and $g'_\alpha(1, 1, \nu, \rho)$ is the upper $\alpha/2$ critical point of Student's t-distribution with ν df, denoted by $t_{\alpha/2}(\nu)$.

Table 1. Critical Points $g_\alpha(k,k,\nu,\gamma,\rho)$ for $\alpha = .05$, $\gamma = .5$, and $\nu = \infty$

	k			
ρ	2	3	4	5
.1	2.209	2.346	2.439	2.509
.3	2.211	2.348	2.441	2.510
.5	2.212	2.349	2.442	2.511
.7	2.212	2.349	2.442	2.512
$h_{\alpha/2}(k, \nu, \gamma)$	2.212	2.349	2.442	2.512

4. NORMAL THEORY PROCEDURES

4.1 Preliminaries

The hypotheses for demonstrating efficacy are

$$H_i^{(1)}: \ \mu_i \leq \mu_0 + \delta_1$$

versus

$$A_i^{(1)}: \ \mu_i > \mu_0 + \delta_1, \quad i = 1, 2, \ldots, k, \qquad (10)$$

where $H_i^{(1)}$ states that the ith dose is ineffective. Similarly, the hypotheses for demonstrating safety are

$$H_j^{(2)}: \ \eta_j \geq \eta_0 + \delta_2$$

versus

$$A_j^{(2)}: \ \eta_j < \eta_0 + \delta_2, \quad j = 1, 2, \ldots, k, \qquad (11)$$

where $H_j^{(2)}$ states that the jth dose is unsafe. The test procedures for these hypotheses is based on the test statistics

$$t_i^{(1)} = \frac{\bar{x}_i - \bar{y}_0 - \delta_1}{\hat{\sigma}\sqrt{1/n_i + 1/n_0}}$$

and

$$t_j^{(2)} = \frac{\bar{y}_0 - \bar{y}_j + \delta_2}{\hat{\tau}\sqrt{1/n_j + 1/n_0}}, \quad i, j = 1, 2, \ldots, k. \qquad (12)$$

For specified α, MINED and MAXSD are estimated as

$$\widehat{\text{MINED}} = \min\{i: \ H_i^{(1)} \text{ is rejected}\}$$

and

$$\widehat{\text{MAXSD}} = \max\{j: \ H_j^{(2)} \text{ is rejected}\}. \qquad (13)$$

Define the families

$$\mathcal{F}_1 = \{H_i^{(1)}, i = 1, 2, \ldots, k\} \quad \text{and} \quad \mathcal{F}_2 = \{H_j^{(2)}, \ j = 1, 2, \ldots, k\}$$

and their union

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 = \{H_i^{(1)}, H_j^{(2)}, \ i, j = 1, 2, \ldots, k\}.$$

It is easy to check that strong control (Hochberg and Tamhane 1987) of the type I familywise error rate (FWE) for $\mathcal{F}$ at level α is equivalent to (3).

Instead of $\mathcal{F}$, we could take our family as

$$\mathcal{F}' = \left\{H_i^{(1)} \cup H_i^{(2)},\ i = 1, 2, \ldots, k\right\},$$

where $H_i^{(1)} \cup H_i^{(2)}$ states that the ith dose is either ineffective or unsafe. Strong control of the FWE for this family is also equivalent to (3). A single-step and a step-down procedure for $\mathcal{F}'$ were given in earlier work (Tamhane and Logan 2000).

4.2 Test Procedures for Family $\mathcal{F}$

A simple but conservative way to control the FWE for $\mathcal{F}$ at level α is to control the FWE for $\mathcal{F}_1$ and $\mathcal{F}_2$ separately at level $\alpha/2$ and use the Bonferroni inequality. More generally, different levels α_1 and α_2 (where $\alpha_1 + \alpha_2 = \alpha$) may be used for $\mathcal{F}_1$ and $\mathcal{F}_2$ to reflect different weights that one may wish to attach to efficacy testing and safety testing. We call the Bonferroni procedures *approximate*, and call those based on the exact critical constants *exact*.

4.3 Step-Down Test Procedures

We omit the discussion of the exact and approximate single-step (SS) procedures, because they are uniformly less powerful than the corresponding SD1 step-down procedures presented here. However, the SS procedures have one advantage in that they give simultaneous lower confidence bounds on $\mu_i - \mu_0$ and upper confidence bounds on $\eta_i - \eta_0$ $(i = 1, 2, \ldots, k)$.

4.3.1 Approximate SD1 Procedure. To construct a step-down procedure using the closure method of Marcus, Peritz, and Gabriel (1976), we need a closed family. The families $\mathcal{F}_1$ and $\mathcal{F}_2$ are closed because $H_\ell^{(1)} \Rightarrow H_i^{(1)}\ \forall\, i < \ell$ and $H_m^{(2)} \Rightarrow H_j^{(2)}\ \forall\, j > m$ as a consequence of the assumption (2). The approximate SD1 procedure addresses the families $\mathcal{F}_1$ and $\mathcal{F}_2$ separately, each at level $\alpha/2$. It tests the hypotheses $H_\ell^{(1)}$ in $\mathcal{F}_1$ in a step-down manner beginning with $\ell = k$, and rejects $H_\ell^{(1)}$ iff $H_k^{(1)}, H_{k-1}^{(1)}, \ldots, H_{\ell+1}^{(1)}$ are rejected and $\max_{1\le i\le \ell} t_i^{(1)} > h_{\alpha/2}(\ell, \nu, \Gamma_1)$. Similarly, it tests the hypotheses $H_m^{(2)}$ in $\mathcal{F}_2$ in a step-down manner beginning with $m = 1$, and rejects $H_m^{(2)}$ iff $H_1^{(2)}, H_2^{(2)}, \ldots, H_{m-1}^{(2)}$ are rejected and $\max_{m\le j\le k} t_j^{(2)} > h_{\alpha/2}(k - m + 1, \nu, \Gamma_2)$.

4.3.2 Exact SD1 Procedure. The exact SD1 procedure addresses the family $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$. Because $\mathcal{F}$ is not closed, we need to form its *closure*, given by

$$\overline{\mathcal{F}} = \left\{H_i^{(1)}, H_j^{(2)}, H_i^{(1)} \cap H_j^{(2)}\ \forall\, i, j = 1, 2, \ldots, k\right\}. \quad (14)$$

Although this exact procedure is not implementable because it requires knowledge of ρ, we still present it here, because its bootstrap version is implementable. It operates as follows. Initialize $\ell = k$, $m = 1$ and $\Gamma_1 = \Gamma_2 = \Gamma_{12} = \Gamma$.

General Step A: For $\ell \ge 1$ and $m \le k$, test $H_\ell^{(1)} \cap H_m^{(2)}$ iff all hypotheses $H_i^{(1)}$ and $H_j^{(2)}$ are rejected for $i \ge \ell + 1$ and $j \le m - 1$. Reject $H_\ell^{(1)} \cap H_m^{(2)}$ if

$$\max\left(\max_{1\le i\le \ell} t_i^{(1)},\ \max_{m\le j\le k} t_j^{(2)}\right) > g_\alpha(\ell, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho). \quad (15)$$

If $H_\ell^{(1)} \cap H_m^{(2)}$ is not rejected, then stop testing and decide that $\widehat{\text{MINED}} = \ell + 1$ and $\widehat{\text{MAXSD}} = m - 1$. Otherwise, if $\max_{1\le i\le \ell} t_i^{(1)} > g_\alpha(\ell, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$ then reject $H_\ell^{(1)}$ and set $\ell \leftarrow \ell - 1$. Similarly, if $\max_{m\le j\le k} t_j^{(2)} > g_\alpha(\ell, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$, then reject $H_m^{(2)}$ and set $m \leftarrow m + 1$. In fact, a shortcut can be used. If ℓ' is the lowest dose $\le \ell$ such that $t_{\ell'}^{(1)} > g_\alpha(\ell, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$, then reject the hypotheses $H_{\ell'}^{(1)}, \ldots, H_\ell^{(1)}$ and set $\ell \leftarrow \ell' - 1$. Similarly, if m' is the highest dose $\ge m$ such that $t_{m'}^{(2)} > g_\alpha(\ell, k - m + 1, \nu, \Gamma_1, \Gamma_2, \Gamma_{12}, \rho)$, then reject the hypotheses $H_m^{(2)}, \ldots, H_{m'}^{(2)}$ and set $m \leftarrow m' + 1$. If $\ell = 0$ or $m = k + 1$, then go to step B; otherwise, return to the beginning of step A.

General Step B:

B1: If $\ell = 0$ and $m = k + 1$, then there are no more hypotheses to be tested. Stop testing and decide that $\widehat{\text{MINED}} < 1$ and $\widehat{\text{MAXSD}} > k$; that is, all doses are effective and safe.

B2: If $\ell \ge 1$ and $m = k + 1$, then decide that $\widehat{\text{MAXSD}} > k$; then there are only efficacy hypotheses to be tested. The critical constant for testing $H_\ell^{(1)}$ is $h_\alpha(\ell, \nu, \Gamma_1)$. If $\max_{1\le i\le \ell} t_i^{(1)} \le h_\alpha(\ell, \nu, \Gamma_1)$, then stop testing and decide that $\widehat{\text{MINED}} = \ell + 1$. Otherwise, if ℓ' is the lowest dose $\le \ell$ such that $t_{\ell'}^{(1)} > h_\alpha(\ell, \nu, \Gamma_1)$, then reject $H_{\ell'}^{(1)}, \ldots, H_\ell^{(1)}$ and set $\ell \leftarrow \ell' - 1$. If $\ell = 0$, then stop testing and decide that $\widehat{\text{MINED}} < 1$; otherwise, return to the beginning of step B2.

B3: If $\ell = 0$ and $m \le k$, then decide that $\widehat{\text{MINED}} < 1$; then there are only safety hypotheses to be tested. The critical constant for testing $H_m^{(2)}$ is $h_\alpha(k - m + 1, \nu, \Gamma_2)$. If $\max_{m\le j\le k} t_j^{(2)} \le h_\alpha(k - m + 1, \nu, \Gamma_2)$, then stop testing and decide that $\widehat{\text{MAXSD}} = m - 1$. Otherwise, if m' is the highest dose $\ge m$ such that $t_{m'}^{(2)} > h_\alpha(k - m + 1, \nu, \Gamma_2)$, then reject $H_m^{(2)}, \ldots, H_{m'}^{(2)}$ and set $m \leftarrow m' + 1$. If $m = k + 1$, then stop testing and decide that $\widehat{\text{MAXSD}} > k$; otherwise, return to the beginning of step B3.

This test procedure is shown graphically in Figure 1. Notice that an advantage of testing the efficacy and safety hypotheses jointly is that whenever an intersection hypothesis $H_\ell^{(1)} \cap H_m^{(2)}$ is rejected, then even if one of the component hypotheses is not rejected, that component hypothesis remains in contention as a candidate for rejection at a later step. This is not the case when the efficacy and safety hypotheses are tested separately using the approximate Bonferroni method, where testing stops once a given hypothesis is not rejected.

4.3.3 Exact and Approximate SD2 Test Procedures. The SD2 test procedure uses the test statistics $t_\ell^{(1)}$ and $t_m^{(2)}$ to

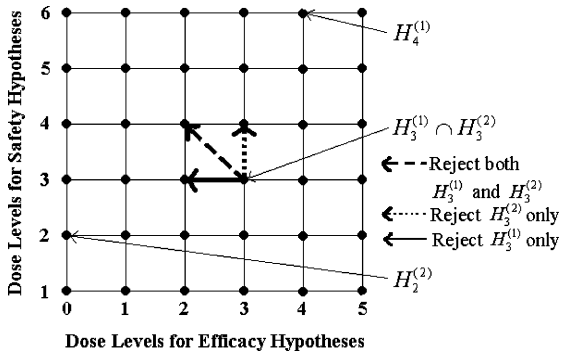

Figure 1. Hypothesis Testing Sequence for Exact SD1 and SD2 Procedures for k = 5. Node (ℓ,m) represents the hypothesis $H_\ell^{(1)} \cap H_m^{(2)}$. If ℓ = 0, then the node represents the hypothesis $H_m^{(2)}$. If m = k + 1 = 6, then the node represents the hypothesis $H_\ell^{(1)}$.

test $H_\ell^{(1)}$ and $H_m^{(2)}$ instead of $\max_{1\le i\le \ell} t_i^{(1)}$ and $\max_{m\le j\le k} t_j^{(2)}$ used by SD1. Analogous to SD1, we can formulate the exact and approximate SD2 procedures. The exact SD2 procedure tests the intersection hypothesis $H_\ell^{(1)} \cap H_m^{(2)}$ using the test statistic $\max(t_\ell^{(1)}, t_m^{(2)})$, which compares with the critical constant $g_\alpha(1, 1, \nu, \gamma_{\ell m}, \rho)$ if $\ell \ne m$ and $g'_\alpha(1, 1, \nu, \rho)$ if $\ell = m$. Because these critical constants depend on the unknown ρ, we must use either a bootstrap procedure or the approximate SD2 procedure, which tests $H_\ell^{(1)}$ and $H_m^{(2)}$ separately each at level $\alpha/2$ using the critical constant $t_{\alpha/2}(\nu)$ for each test. Note that shortcuts are not possible with SD2 procedures.

5. BOOTSTRAP PROCEDURES

Normal theory methods require the strong distributional assumption of bivariate normality. Furthermore, they assume homoscedasticity across dose groups and common correlation. Bootstrapping allows for relaxation of these restrictive assumptions. However, there is no unique way to apply bootstrapping, and it is not always a priori clear which method works better. The bootstrap procedures given here have been tested extensively via simulation. They assume homoscedasticity and common correlation because they pool the data across different dose groups; however, they can be easily modified by not pooling. They approximately account for the unknown correlation through resampling.

The bootstrap versions of SD1 and SD2 are as follows:

1. Mean center the data. For convenience, we use the same notation for the mean centered data as the raw data. Thus

$$x_{ij} \leftarrow x_{ij} - \bar{x}_i \text{ and } y_{ij} \leftarrow y_{ij} - \bar{y}_i,\ i = 0, 1, \ldots, k,\quad j = 1, 2, \ldots, n_i.$$

2. Repeat the following steps for B iterations.

   a. For the bth iteration, draw a bootstrap sample for each dose:

$$(x_{ijb}, y_{ijb}),\ i = 0, 1, \ldots, k,\ \ j = 1, 2, \ldots, n_i,$$

   with replacement from the pooled mean centered paired data (x_{ij}, y_{ij}) from all doses.

   b. For the bth bootstrap sample, calculate the sample means $\bar{x}_{ib}$ and $\bar{y}_{ib}$ for $i = 1, 2, \ldots, b$ and sample standard deviations $\hat{\sigma}_b$ and $\hat{\tau}_b$ based on $\nu = N - (k+1)$ df.

   c. Calculate the t-statistics

$$t_{ib}^{(1)} = \frac{\bar{x}_{ib} - \bar{x}_{0b} - \delta_1}{\hat{\sigma}_b\sqrt{1/n_i + 1/n_0}}$$

   and

$$t_{jb}^{(2)} = \frac{\bar{y}_{0b} - \bar{y}_{jb} + \delta_2}{\hat{\tau}_b\sqrt{1/n_j + 1/n_0}},\quad i, j = 1, 2, \ldots, k.$$

3. Exact SD1 procedure. Initialize $\ell = k$ and $m = 1$.

   General step A: For $\ell \ge 1$ and $m \le k$, test $H_\ell^{(1)} \cap H_m^{(2)}$ iff all hypotheses $H_i^{(1)}$ and $H_j^{(2)}$ are rejected for $i \ge \ell + 1$ and $j \le m - 1$. To test $H_\ell^{(1)} \cap H_m^{(2)}$, calculate

$$p_{\ell,m}^{(1)} = \frac{1}{B}\left\{\#b \left| \max_{1\le i\le \ell, m\le j\le k} (t_{ib}^{(1)}, t_{jb}^{(2)}) \ge \max_{1\le i\le \ell} t_i^{(1)} \right.\right\}$$

   and

$$p_{\ell,m}^{(2)} = \frac{1}{B}\left\{\#b \left| \max_{1\le i\le \ell, m\le j\le k} (t_{ib}^{(1)}, t_{jb}^{(2)}) \ge \max_{m\le j\le k} t_j^{(2)} \right.\right\},$$

   where $\#b$ denotes the number of bootstrap samples satisfying the given condition. If both $p_{\ell,m}^{(1)}$ and $p_{\ell,m}^{(2)}$ are $\ge\alpha$, then stop testing and decide that $\widehat{\text{MINED}} = \ell + 1$ and $\widehat{\text{MAXSD}} = m - 1$. If $p_{\ell,m}^{(1)} < \alpha$, then reject $H_{\ell'}^{(1)}, \ldots, H_\ell^{(1)}$, where $t_{\ell'}^{(1)} = \max_{1\le i\le \ell} t_i^{(1)}$, and set $\ell \leftarrow \ell' - 1$. If $p_{\ell,m}^{(2)} < \alpha$, then reject $H_m^{(2)}, \ldots, H_{m'}^{(2)}$, where $t_{m'}^{(2)} = \max_{m\le j\le k} t_j^{(2)}$, and set $m \leftarrow m' + 1$. If $\ell = 0$ or $m = k + 1$, then go to general step B; otherwise, return to the beginning of step A.

   General step B:

   B1: If $\ell = 0$ and $m = k + 1$, then stop testing and decide that $\widehat{\text{MINED}} < 1$ and $\widehat{\text{MAXSD}} > k$; that is, all doses are effective and safe.

   B2: If $m = k + 1$ and $\ell \ge 1$, then decide that $\widehat{\text{MAXSD}} > k$; there are only efficacy hypotheses to be tested. Calculate

$$p_\ell^{(1)} = \frac{1}{B}\left\{\#b \left| \max_{1\le i\le \ell} t_{ib}^{(1)} \ge \max_{1\le i\le \ell} t_i^{(1)} \right.\right\}. \qquad (16)$$

   If $p_\ell^{(1)} \ge \alpha$, then stop testing and decide that $\widehat{\text{MINED}} = \ell + 1$. Otherwise, if $p_\ell^{(1)} < \alpha$, then reject $H_{\ell'}^{(1)}, \ldots, H_\ell^{(1)}$, where $t_{\ell'}^{(1)} = \max_{1\le i\le \ell} t_i^{(1)}$, and set $\ell \leftarrow \ell' - 1$. If $\ell = 0$, then stop testing and decide that $\widehat{\text{MINED}} < 1$; otherwise, return to the beginning of step B2.

   B3: If $\ell = 0$ and $m \le k$, then decide that $\widehat{\text{MINED}} < 1$; then there are only safety hypotheses to be tested. Calculate

$$p_m^{(2)} = \frac{1}{B}\left\{\#b \left| \max_{m\le j\le k} t_{jb}^{(2)} \ge \max_{m\le j\le k} t_j^{(2)} \right.\right\}. \qquad (17)$$

   If $p_m^{(2)} \ge \alpha$, then stop testing and decide that $\widehat{\text{MAXSD}} = m - 1$. Otherwise, if $p_m^{(2)} < \alpha$, then

reject $H_m^{(2)}, \ldots, H_{m'}^{(2)}$, where $t_{m'}^{(2)} = \max_{m \le j \le k} t_j^{(2)}$, and set $m \leftarrow m' + 1$. If $m = k + 1$, then stop testing and decide that $\widehat{\text{MAXSD}} > k$; otherwise, return to the beginning of step B3.

Note that we do not use the full shortcut version of SD1 (as given in Sec. 4.3.2) in the bootstrap procedure, because such a shortcut requires computing the adjusted p values for each $t_i^{(1)}$ for $1 \le i \le \ell$ and $t_j^{(2)}$ for $m \le j \le k$ instead of just for $\max_{1 \le i \le \ell} t_i^{(1)}$ and $\max_{m \le j \le k} t_j^{(2)}$. Therefore, the full shortcut version of SD1 results in more computational effort in general, when implemented using the bootstrap. Also note that the p values calculated earlier are bootstrap estimates of the multiplicity adjusted p values. They are not monotonically adjusted for step-down testing, because the p value for any hypothesis is calculated only if the p value for its implying hypothesis is less than α.

4. Exact SD2 procedure. The bootstrap SD2 procedure is similar to the SD1 procedure, except that no maximums are taken over the t-statistics when calculating the estimates of the adjusted p values.

6. EXAMPLE

A pharmaceutical company performed 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, 2, 3, and 4). A total of 370 patients were randomized to the five treatment groups. Several efficacy and safety endpoints were measured on each patient at the baseline and at the end of the study after 4 weeks. Here we consider the changes from the baseline in one particular efficacy and one safety endpoint.

For proprietary reasons, the actual data are concealed by contaminating them with normally distributed random errors with zero mean and one-tenth of the standard deviation of the original data. This slightly increases the standard deviations of the efficacy and safety variables for each dose group and reduces the correlations between them. As a further step to comply with the company requirements, the identity of the safety variable is not revealed, and its values are coded post-contamination. Despite these measures, the essential statistical characteristics of the data are unimpaired, and the example remains valuable for illustrating the proposed procedures.

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 indicates an improvement in disease condition. For the purpose of this example, an average improvement by .5 units compared to the zero dose control is regarded as clinically significant.

The safety variable is the serum level of a certain chemical, labeled as Z. As dose level increases, the serum level of Z is expected to increase from the baseline to the end of the study. For the purpose of this example, an increase in serum level of Z by 3 mmol/L over the zero dose control is regarded as clinically significant.

Table 2. Summary Statistics for Changes from Baseline in WOMAC Score and Serum Level of Z

		Dose level				
		0	1	2	3	4
WOMAC	Mean	1.437	2.196	2.459	2.771	2.493
	SD	1.924	2.253	1.744	1.965	1.893
Z	Mean	.554	1.430	1.594	2.242	2.624
	SD	2.122	1.941	2.340	2.388	2.229
Correlation coefficient		−.247	.121	−.072	.232	−.047
Sample size		76	73	73	75	73

The summary data for the two variables are given in Table 2. Normal plots of the data were made and found to be quite satisfactory. The sample sizes are nearly equal, and so $\gamma_{ij} \approx 1/2$. Box's (1949) test for homogeneity of covariance matrices yielded $F = 1.705$ with p value $= .059$, which is borderline nonsignificant. The Bartlett and Levene tests for homogeneity of variances yielded highly nonsignificant results. Therefore, all the assumptions are satisfied for the normal theory procedures. It is of interest to note that the correlations are all close to zero.

The pooled estimates of the standard deviations for WOMAC (the efficacy variable) and Z (the safety variable) are $\hat{\sigma} = 1.962$ and $\hat{\tau} = 2.210$, each with $365 \approx \infty$ df. The ANOVA F-statistics for WOMAC and Z equal 5.079 ($p = .001$) and 9.747 ($p = .000$). The t-statistics for WOMAC and Z are computed using (12) with $\delta_1 = .5$ and $\delta_2 = 3$ and are given in Table 3.

6.1 Approximate SD1 Procedure

The critical constants $h_{.025}(i, \nu = \infty, \gamma = 1/2)$ for $i = 1, 2, 3, 4$ equal 1.960, 2.212, 2.349, and 2.442. A straightforward application of SD1 at the .025-level separately on WOMAC and Z yields $\widehat{\text{MINED}} = 3$ and $\widehat{\text{MAXSD}} > 4$ (i.e., all four doses are safe). Thus the lower end of the therapeutic window is dose 3, whereas the upper end can be higher than dose 4.

6.2 Approximate SD2 Procedure

In this case, each t-statistic is compared with the same critical constant, $t_{.025}(\nu = \infty) = 1.960$, in a step-down manner starting with dose 4 for WOMAC and dose 1 for Z. Because $t_4^{(1)} = 1.729 < 1.960$, the procedure stops with the conclusion that $\widehat{\text{MINED}} > 4$ (i.e., none of the doses are effective). Also,

Table 3. The t-Statistics and Their Unadjusted p Values for Comparing Doses With Zero Dose Control

Comparison		1 vs. 0	2 vs. 0	3 vs. 0	4 vs. 0
WOMAC	$t_i^{(1)}$	.806	1.625	2.612	1.729
	$p_i^{(1)}$	.210	.052	.005	.042
Z	$t_i^{(2)}$	5.861	5.407	3.644	2.564
	$p_i^{(2)}$	.000	.000	.000	.005

$\widehat{\text{MAXSD}} > 4$ (i.e., all doses are safe), because all $t_i^{(2)}$ statistics are >1.960. Thus a therapeutic window is not found.

6.3 Bootstrap SD1 Procedure

Begin by testing $H_4^{(1)} \cap H_1^{(2)}$. The corresponding adjusted p values are estimated to be (based on 5000 bootstrap samples) $p_{4,1}^{(1)} = .037$ and $p_{4,1}^{(2)} = 0$. Because both $p_{4,1}^{(1)}$ and $p_{4,1}^{(2)}$ are $< .05$, we reject both $H_4^{(1)}$ and $H_1^{(2)}$, and next test $H_3^{(1)} \cap H_2^{(2)}$. In fact, we can use the shortcut and directly test $H_2^{(1)} \cap H_2^{(2)}$, because rejection of $H_4^{(1)}$ is caused by $t_3^{(1)} = \max_{1 \le i \le 4} t_i^{(1)}$, implying rejection of $H_3^{(1)}$ as well. The adjusted p values for $H_2^{(1)} \cap H_2^{(2)}$ are estimated to be $p_{2,2}^{(1)} = .201$ and $p_{2,2}^{(2)} = 0$. Therefore, we reject $H_2^{(2)}$ and next test $H_2^{(1)} \cap H_3^{(2)}$. The adjusted p values for $H_2^{(1)} \cap H_3^{(2)}$ are estimated to be $p_{2,3}^{(1)} = .167$ and $p_{2,3}^{(2)} = .001$. Therefore, we reject $H_3^{(2)}$ and next test $H_2^{(1)} \cap H_4^{(2)}$. The adjusted p values for $H_2^{(1)} \cap H_4^{(2)}$ are estimated to be $p_{2,4}^{(1)} = .137$ and $p_{2,4}^{(2)} = .017$, and so we reject $H_4^{(2)}$. This leaves only the efficacy hypotheses $H_1^{(1)}$ and $H_2^{(1)}$ to be tested. The adjusted p value for $H_2^{(1)}$ is estimated to be $p_2^{(1)} = .085 > .05$, so we stop testing. Thus the bootstrap exact SD1 procedure comes to the same decision as the normal theory approximate SD1, namely $\widehat{\text{MINED}} = 3$ and $\widehat{\text{MAXSD}} > 4$.

6.4 Bootstrap SD2 Procedure

For brevity, we give only the estimated adjusted p values at each step. It should be clear from the foregoing description of SD1 how the hypotheses tested at each step are determined. The adjusted p values for $H_4^{(1)} \cap H_1^{(2)}$ are estimated to be $p_{4,1}^{(1)} = .086$ and $p_{4,1}^{(2)} = 0 < .05$, so we reject $H_1^{(2)}$. The adjusted p values for $H_4^{(1)} \cap H_2^{(2)}$ are estimated to be $p_{4,2}^{(1)} = .091$ and $p_{4,2}^{(2)} = 0 < .05$, so we reject $H_2^{(2)}$. The adjusted p values for $H_4^{(1)} \cap H_3^{(2)}$ are estimated to be $p_{4,3}^{(1)} = .085$ and $p_{4,3}^{(2)} = 0 < .05$, so we reject $H_3^{(2)}$. The adjusted p values for $H_4^{(1)} \cap H_4^{(2)}$ are estimated to be $p_{4,4}^{(1)} = .089$ and $p_{4,4}^{(2)} = .012 < .05$, so we reject $H_4^{(2)}$. Thus all doses are proven safe, and we continue testing only for efficacy. The adjusted p value for $H_4^{(1)}$ is estimated to be $p_4^{(1)} = .046 < .05$, so we reject $H_4^{(1)}$. The adjusted p value for $H_3^{(1)}$ is estimated to be $p_3^{(1)} = .005 < .05$, so we reject $H_3^{(1)}$. The adjusted p value for $H_2^{(1)}$ is estimated to be $p_2^{(1)} = .047 < .05$, so we reject $H_1^{(2)}$. Finally, the adjusted p value for $H_1^{(1)}$ is estimated to be $p_1^{(1)} = .205 > .05$, so we stop testing and decide that $\widehat{\text{MINED}} = 2$ and $\widehat{\text{MAXSD}} > 4$. Notice that whereas the normal theory SD2 did not find any effective doses, the bootstrap SD2 procedure is able to find doses 2–4 as effective. This is the advantage of joint testing for efficacy and safety over separate testing using the Bonferroni method.

7. SIMULATIONS

7.1 Design of Simulation Studies

We performed simulations to compare the performances of the approximate normal theory SD1 and SD2 and their bootstrap versions under different conditions, including normal and nonnormal data, different mean configurations, and different sample sizes.

The type I FWEs of these procedures, as well as their overall power, defined as

$$\text{Power} = P\{\widehat{\text{MINED}} = \text{MINED and } \widehat{\text{MAXSD}} = \text{MAXSD}\},$$

were estimated. The biases in the estimates $\widehat{\text{MINED}}$ and $\widehat{\text{MAXSD}}$ were also estimated. All simulations were done on an HP workstation in SAS/IML. Each simulation run consisted of 5000 replicates. For bootstrap procedures, the number of bootstrap samples B was set equal to 1000. The nominal FWE for all procedures was set at $\alpha = .05$. The number of doses, in addition to the control, was fixed at $k = 5$. A common sample size, n, was assumed per treatment including the control. Two different sample sizes were studied: $n = 10$ and $n = 50$.

The means (μ_i, η_i) for different doses were chosen based on the following power considerations. Let

$$\lambda_1 = \mu_{\text{MINED}} - (\mu_0 + \delta_1) \quad \text{and} \quad \lambda_2 = (\eta_0 + \delta_2) - \eta_{\text{MAXSD}}$$

be the "separations" between the means of the MINED and MAXSD and the corresponding thresholds, $\mu_0 + \delta_1$ for efficacy and $\eta_0 + \delta_2$ for safety. The powers of different test procedures depend on $\lambda_1\sqrt{n}/\sigma$ and $\lambda_2\sqrt{n}/\tau$, in addition to other mean values. We chose two different shapes for the dose–response function, a "step" shape and a "linear" shape, and MINED $= 2$ and MAXSD $= 4$ in each case. To obtain similar powers for $n = 10$ and $n = 50$, the values of $(\delta_1, \delta_2, \sigma, \tau)$ shown in Table 4 were chosen to make $(\lambda_1\sqrt{n}/\sigma, \lambda_2\sqrt{n}/\tau)$ approximately equal, that is, $(6.261, 4.174)$ and $(6.363, 4.243)$.

The normal data were simulated by generating n independent pairs of bivariate normal observations (x, y) with common correlation .5 for each dose with respective means (μ_i, η_i) and standard deviations (σ, τ). Two nonnormal distributions were studied: a lognormal distribution as an example of a skewed distribution and a double-exponential distribution as an example of a heavy-tailed distribution. Because the results for these nonnormal distributions were similar to those for the normal distribution, they are not shown here (see Tamhane and Logan 2000 for details).

7.2 Simulation Results

Simulated overall powers for approximate normal theory SD1 and SD2 and their bootstrap versions are summarized in Table 5 for normal data. The type I FWEs of the procedures are not reported because they were found to be well controlled at or below the .05 level for both normal and nonnormal data. Biases in the estimates $\widehat{\text{MINED}}$ and $\widehat{\text{MAXSD}}$ or, equivalently, the corresponding marginal powers show similar relative patterns as the overall powers and hence are not reported either (see Tamhane and Logan 2000).

First, we note that SD1 is more powerful (by about 11%–12%) than SD2 for step configurations, but less powerful (by about 6%–8%) for linear configurations. The higher power of SD1 for step configurations is explained by the fact that in this configuration, dose 2 (which is the MINED) through dose 5 have the same mean for efficacy. As a result, the corresponding sample means are likely to be nonmonotone. When test-

Table 4. Simulated Parameter Configurations

Dose–Response Function	Mean	Dose i 0	1	2	3	4	5	$n=10$ σ τ	δ_1 δ_2	λ_1 λ_2	$n=50$ σ τ	δ_1 δ_2	λ_1 λ_2	MINED MAXSD
Step	μ_i	0	1	2	2	2	2	.5	1.01	.99	1.0	1.1	.9	2
	η_1	0	1	1	1	1	2	.75	1.99	.99	1.5	1.9	.9	4
Linear	μ_1	0	1	2	3	4	5	.5	1.01	.99	1.0	1.1	.9	2
	η_i	0	1	2	3	4	5	.75	4.99	.99	1.5	4.9	.9	4

ing the highest dose 5 for efficacy (say), SD1 uses the maximum among these sample means, say $\bar{y}_3$. If $\bar{y}_3$ is significantly greater than $\bar{y}_0 + \delta_1$, then SD1 declares doses 3, 4, and 5 as effective and proceeds to dose 2. On the other hand, if $\bar{y}_5 (< \bar{y}_3)$ is not significantly greater than $\bar{y}_0 + \delta_1$, then SD2 stops testing with the decision that dose 5 is not effective. Although SD1 uses a larger multivariate t critical constant than the univariate t critical constant used by SD2, often SD1 proceeds past dose 5 to eventually find dose 2 as the correct MINED, whereas SD2 stops testing earlier because of nonsignificance. A similar phenomenon occurs when testing for safety.

Because the dose means are monotone for linear configurations (and much higher at higher doses than the mean for the MINED, e.g., $\mu_5 = 5 \gg \mu_2 = 2$), SD2 often does not stop testing early due to nonsignificance and thus gains in power. SD1 also gains in power due to larger differences between the means at higher doses and the MINED mean, but not as much, because it uses a larger critical constant. Therefore, SD2 is more powerful than SD1 for linear configurations.

Comparing SD1 and SD2 normal theory procedures with the corresponding bootstrap procedures, we find that the latter are always more powerful. The gain in power for the bootstrap SD1 over the normal theory SD1 is about 3%–5%, whereas the corresponding gain for SD2 is only around 1%. It may appear surprising that the bootstrap procedures are more powerful than the normal theory procedures when the data are normal. The explanation is that although the bootstrap procedures may suffer slight loss in power because they do not exploit normality, this loss is more than compensated for by their use of joint testing of efficacy and safety, as illustrated in Figure 1. On the other hand, the approximate normal theory procedures treat the two endpoints as separate and use the Bonferroni approximation.

8. CONCLUDING REMARKS

The bootstrap procedures not only are more robust with regard to distributional assumptions, but also are more powerful because of their joint testing feature. Therefore, they are recommended. However, they are more complicated to implement and explain and require special software. Also, their FWE control is only approximately guaranteed for small samples. For step configurations, bootstrap SD1 has a power advantage of about 14% over bootstrap SD2, whereas for linear configurations, bootstrap SD2 has a power advantage of about 5%. The true dose mean configuration is of course unknown. Although one might argue that a linear configuration is more likely than a step configuration, a flat response beyond a certain dose level is in fact more commonly observed than might be expected, especially for efficacy. The reason is that in the dose selection process, it is usually difficult to find the right doses to test, and practitioners tend to select doses at the plateau of the dose–response curve. If the efficacy response function has a downturn at higher doses, then SD2 will have very low power. Given that the power gain of SD1 over SD2 for step configurations is much greater than its power loss for linear configuration, we recommend bootstrap SD1 as a general procedure unless dose response is known to be roughly linear (or at least strictly increasing).

We have assumed weak monotonicity (2). If strong monotonicity can be assumed, then more powerful procedures can be developed based on the isotonic estimates (Robertson, Wright, and Dykstra 1988) of the μ_i and the τ_i. The exact distribution theory would be quite complicated, but the bootstrap version should be relatively straightforward.

Some may argue that efficacy is more important than safety, and so a dose need not be tested for safety unless it is proven

Table 5. Simulation Estimates of Overall Powers for Normal Data

Procedure	$n=10$ Step Configuration	$n=10$ Linear Configuration	$n=50$ Step Configuration	$n=50$ Linear Configuration
SD1 (Normal)	.6697	.6934	.7332	.7590
SD2 (Normal)	.5581	.7792	.6106	.8256
SD1 (Bootstrap)	.7092	.7402	.7622	.7912
SD2 (Bootstrap)	.5682	.7904	.6250	.8418

effective. This approach is followed by Bauer, Brannath, and Posch (2001). On the other hand, some may argue that safety is more important than efficacy. We have treated efficacy and safety on an equal basis in this article.

Finally, we note that the SAS/IML code for implementing the bootstrap versions of exact SD1 and SD2 procedures can be downloaded from http://users.iems.northwestern.edu/∼ajit.

[Received October 2000. Revised May 2001.]

REFERENCES

Anderson, T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Bauer, P., Brannath, W., and Posch, M. (2001), "Multiple Testing for Identifying Effective and Safe Treatments," *Biometrical Journal*, 43, 605–616.

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

Bechhofer, R. E., and Tamhane, A. C. (1974), "An Iterated Integral Representation for a Multivariate Normal Integral Having Block Covariance Structure," *Biometrika*, 61, 615–619.

Box, G. E. P. (1949), "A General Distribution Theory for a Class of Likelihood Criteria," *Biometrika*, 36, 362–389.

Cornish, E. A. (1954), "The Multivariate t-Distribution Associated With a Set of Normal Deviates," *Australian Journal of Physics*, 7, 531–542.

Dunnett, C. W. (1989), "Multivariate Normal Probability Integrals With Product Correlation Structure, Algorithm AS251," *Applied Statistics*, 38, 564–579.

Dunnett, C. W., and Sobel, M. (1954), "A Bivariate Generalization of Student's t-Distribution With Tables for Certain Special Cases," *Biometrika*, 41, 153–169.

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

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

Jennison, C., and Turnbull, B. W. (1993), "Group Sequential Tests for Bivariate Response," *Biometrics*, 49, 741–752.

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.

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

Siddiqui, M. M. (1967), "A Bivariate t Distribution," *Annals of Mathematical Statistics*, 38, 162–166.

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. (2000), "Multiple Test Procedures for Identifying the Minimum Effective and Maximum Safe Doses of a Drug," technical report, Northwestern University, Dept. of Statistics.

Thall, P. F., and Cheng, S.-C. (1999), "Treatment Comparisons Based on Two-Dimensional Safety and Efficacy Alternatives in Oncology Trials," *Biometrics*, 55, 746–753.

Thall, P. F., and Russell, K. E. (1998), "A Strategy for Dose Finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials," *Biometrics*, 54, 251–264.

Turri, M., and Stein, G. (1986), "The Determination of Practically Useful Doses of New Drugs: Some Methodological Considerations," *Statistics in Medicine*, 5, 449–457.

Westfall, P., and Young, S. S. (1993), *Resampling-Based Multiple Testing*, New York: Wiley.

Annotations from asatm00-305r1.pdf

Page 9

Annotation 1; Label: Tamhane; Date: 11/26/2001 2:14:00 PM
Au: no text cite for Anderson (1958), Genz and Bretz (1999) and Marcus, Peritz and Gabriel (1976). Either add citations or delete from ref. list.