STATISTICS IN MEDICINE, VOL. 16, 2489–2506 (1997)

MULTIPLE TESTING TO ESTABLISH SUPERIORITY/EQUIVALENCE OF A NEW TREATMENT COMPARED WITH k STANDARD TREATMENTS

CHARLES W. DUNNETT[1]* AND AJIT C. TAMHANE[2]

[1] *Department of Mathematics and Statistics, and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada*

[2] *Department of Statistics, and Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208, U.S.A.*

SUMMARY

In this paper we develop multiple hypotheses testing procedures to compare a new treatment with a set of standard treatments in a clinical trial. The aim is to classify the new treatment with respect to each of the standards, by specifying those to which the new treatment is superior, those to which the new treatment is equivalent and those to which one can establish neither superiority nor equivalence. We propose several stepwise procedures and compare them with respect to their familywise error rates and power. The step-down methods SD1 and SD2 test for superiority first, followed by tests for equivalence for those comparisons where we cannot establish superiority. The step-up methods SU1 and SU2 test for equivalence first, followed by tests for superiority for those comparisons where we can establish at least equivalence. The methods SD3 and SU3 apply the tests for superiority and equivalence in pairs. All the methods require that we specify a threshold value $\delta > 0$ in advance for defining equivalence. In applications where it is not possible to specify a value δ, we can use the method SD1 by testing for superiority first, followed by one-sided confidence limits on the efficacy differences for those comparisons where we cannot establish superiority. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2489–2506 (1997)
No. of Figures: 0 No. of Tables: 6 No. of References: 14

1. INTRODUCTION

Morikawa and Yoshida[1] as well as Dunnett and Gent[2] considered the problem of testing the significance of the difference in efficacy between a new treatment compared with a standard treatment in a clinical trial setting. Instead of using a two-sided test, which tests simultaneously for either a positive difference in favour of the new treatment or a negative difference in favour of the standard, they proposed testing simultaneously for a positive difference and for equivalence between the new treatment and the standard.

The rationale for testing simultaneously for superiority and equivalence is that, in many cases, an investigator wishes to establish first whether the new treatment can be shown superior to the standard, in which case it becomes a possible candidate to replace the standard as the

* Correspondence to: C. W. Dunnett, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada

CCC 0277–6715/97/212489–18$17.50

Received May 1996
Revised January 1997

recommended method of treatment, and secondly, if superiority cannot be established, whether the new treatment has equivalent efficacy to the standard, in which case it becomes a possible candidate for use as an alternative treatment method. Failure to establish either superiority or equivalence of the new treatment suggests that one cannot recommend it for use as it may, in fact, be inferior to the standard in efficacy.

The purpose of the present paper is to consider the case where there is more than one standard treatment for comparison with the new treatment. When there are two are more standards available, a sponsor of a potential new treatment may wish to compare its efficacy with each of the available standards. For example, Hoover[3] refers to a study by Graham *et al.*[4] in which acetaminophen, a new treatment for cold symptoms, and a placebo were compared with two standard therapies, aspirin and ibuprofen. The primary question addressed by these authors was whether acetaminophen has less virus shedding and less suppression of antibody responses (two undesirable effects of the treatments) than aspirin and ibuprofen. Another example is the GUSTO[5] clinical trial that we discuss in Section 9.

Denote by k the number of standard treatments. We assume that the aim is to classify the new treatment with respect to each of the k standards, by specifying those to which the new treatment is superior, those to which the new treatment is equivalent and those to which one can claim neither superiority nor equivalence.

Recently, stepwise multiple test procedures have been developed for the purpose of simultaneously testing a set of null hypotheses. In stepwise testing, the hypotheses are ordered from the least to the most significant, using either their p-values or the magnitudes of their test statistics, and tested sequentially. Testing starts either with the most significant and continues as long as a rejection occurs (called step-down testing), or with the least significant and continues as long as a non-rejection or acceptance occurs (called step-up testing). In the present paper, we extend the normal theory step-down and step-up procedures developed in Dunnett and Tamhane[6–8] to the problem of testing for superiority/equivalence between a new treatment and k standard treatments.

2. PRELIMINARIES

Denote by μ_i the unknown mean efficacy for the ith treatment ($i = 0, 1, \ldots, k$) where 0 denotes the new treatment. Define $\theta_i = \mu_0 - \mu_i$ or $\theta_i = \mu_i - \mu_0$, depending on whether larger or smaller values of the μ's are better. We can test the following pair of hypotheses to classify the status of the new treatment compared with the ith standard:

$$H_i: \theta_i \leqslant 0 \text{ versus } \theta_i > 0$$

and

$$H_i': \theta_i \leqslant -\delta \text{ versus } \theta_i > -\delta$$

where $\delta > 0$ denotes a difference in efficacy that is small enough for us to consider as clinically insignificant. Rejection of H_i establishes that the efficacy of the test treatment is superior to that of the ith standard, while non-rejection of H_i together with rejection of H_i' establishes that it cannot be worse than the standard by more than δ and therefore, by definition, it is equivalent. The non-rejection of both H_i and H_i' means that we have not shown the new treatment is either superior or equivalent to the ith standard and hence we cannot recommend it as a substitute for that standard treatment.

Denote by t_i and t_i' the statistics for testing H_i and H_i', respectively. For a one-way setup with n_i independent observations in the ith group, let $\bar{y}_i$ be the sample mean for the ith group ($i = 0, 1, \ldots, k$). If we can assume normality and homogeneous error variance σ^2, then the test statistics are

$$t_i = (\bar{y}_0 - \bar{y}_i)/s\sqrt{(1/n_i + 1/n_0)}$$
$$t_i' = t_i + \delta_i' \tag{1}$$

where s^2 is an estimate of σ^2 based on v degrees of freedom (d.f.) and $\delta_i' = \delta/s\sqrt{(1/n_i + 1/n_0)}$. We consider two types of stepwise testing procedures that we can apply to the t and t' statistics using a set of critical constants $c_1 < \ldots < c_k$. In the first type, we compare t_i with c_i whereas we compare t_i' with possibly a different constant from the set, depending on the ranking of t_i' among all the t and t' test statistics. In the second type, we compare both t_i and t_i' with the same constant, c_i. In each case, we determine the constants so that the type I familywise error rate (FWE), where

$$\text{FWE} = P\{\text{reject any true } H_i \text{ or } H_i'\}, \tag{2}$$

satisfies the requirement: FWE $\leqslant \alpha$ under any configuration of the parameters θ_i. The justification for this requirement is that any such type I error may result in a false claim for the efficacy of the test treatment. We protect against this by requiring that the probability of such an event occurring does not exceed a specified level α, where $\alpha > 0$ is an appropriately chosen small quantity.

To simplify the presentation, we restrict the one-way setup to the case of balanced data, where $n_1 = \cdots = n_k = n$ with n_0 possibly different from n, in Sections 3 to 7. We discuss the case of unequal n_i in Section 8. In Sections 3 to 7, we assume that we have labelled the test statistics and their associated hypotheses so that $t_1 \leqslant \ldots \leqslant t_k$. Since $\delta_i' = \delta' = \delta/s\sqrt{(1/n + 1/n_0)}$ for all i, we also have $t_1' \leqslant \ldots \leqslant t_k'$.

3. SINGLE-STEP (SS) TESTING

A single-step (SS) procedure uses the same critical constant for all tests. It is the simplest procedure; moreover, it is the only one that also provides simultaneous confidence interval estimates of all the θ_i.

We reject H_i if $t_i \geqslant c_k$ and we reject H_i' if $t_i' \geqslant c_k$, otherwise in each case we accept the hypothesis, for $i = 1, \ldots, k$. Since $t_i' = t_i + \delta'$, an equivalent way of expressing the procedure is the following: we reject both H_i' and H_i if $t_i \geqslant c_i$, we reject H_i' but not H_i if $c_k - \delta' \leqslant t_i < c_k$ and we reject neither if $t_i < c_k - \delta'$. To control the FWE $\leqslant \alpha$, we choose $c_k = t^{\alpha}_{k,v,\rho}$ which is the one-sided α point of k-variate t with v d.f. and common correlation coefficient $\rho = 1/(1 + n_0/n)$. This can be seen from the following set of simultaneous lower one-sided $100(1 - \alpha)$ per cent confidence intervals for the $\theta_i = \mu_0 - \mu_i$:

$$\theta_i \geqslant \bar{y}_0 - \bar{y}_i - c_k s\sqrt{(1/n_0 + 1/n)} \; (1 \leqslant i \leqslant k).$$

Rejecting H_i if $t_i \geqslant c_k$ corresponds to rejecting if the lower confidence limit on θ_i is $\geqslant 0$. Similarly, rejecting H_i' if $t_i' \geqslant c_k$ corresponds to rejecting if the lower confidence limit on θ_i is $\geqslant -\delta$ $(1 \leqslant i \leqslant k)$. It is clear that the FWE for any number of hypotheses tested on the θ_i based on these simultaneous confidence intervals is $\leqslant \alpha$. The constant c_k is identical to the constant c_k used in the first step of procedures SD1 and SD2 defined in the next section.

4. STEP-DOWN (SD) TESTING

4.1. A Closed Step-Down Test Procedure (SD1)

To test the family of $2k$ hypotheses, H_i and H_i' $(1 \leqslant i \leqslant k)$, we apply the closure method of Marcus *et al.*:[9] see Hochberg and Tamhane[10] (p. 54). This requires that we use all the t and t' statistics, and we order them together from the least significant to the most significant. We start with the most significant, which is t_k', then the next most significant and so on, rejecting the corresponding hypothesis if the test statistic exceeds a certain critical value; this continues until a hypothesis is not rejected, at which point all testing stops and any remaining hypotheses are accepted.

In the general step, suppose that $H_1, \ldots, H_i$ and $H_1', \ldots, H_j'$ remain untested. Then we look at $\max(t_i, t_j')$, where $i \geqslant j$ since $t_i < t_i'$. Whichever is maximum, we compare it with the constant c_i. The reason for this choice is that we are actually testing the intersection $(\cap)$ of all the remaining hypotheses; since $\cap(H_i, H_i') = H_i'$, we can write the intersection hypothesis as $\cap(H_1', \ldots, H_j', H_{j+1}, \ldots, H_i)$ and the test statistic is $\max(t_1', \ldots, t_j', t_{j+1}, \ldots, t_i)$. Since there are i statistics involved, the appropriate constant is c_i, where c_m equals the one-sided α point of m-variate t for $m = 1, \ldots, k$. For balanced data, when $n_i = n$ for $i = 1, \ldots, k$, we have the equal correlation case $\rho_{ij} = \rho = 1/(1 + n_0/n)$ for $i \neq j$. We denote these α points by $t^{\alpha}_{m,\nu,\rho}$, which are tabulated in several places for various values of m, ν, ρ and α. For an extensive set of tables, see Bechhofer and Dunnett.[11]

A simpler way to apply the SD1 procedure, which is exactly equivalent to the closure method described above, proceeds in two stages as follows. In the first stage, we use the statistics $t_1 \leqslant \ldots \leqslant t_k$ to test the superiority hypotheses, H_i. We test them in the usual step-down manner, starting with t_k, then t_{k-1} and so on, continuing as long as we find $t_i \geqslant c_i$ in which case we reject the hypothesis H_i. The first time we observe $t_i < c_i$, say for $i = m$, we accept $H_1, \ldots, H_m$ and terminate the first stage. In the second stage, we test $H_1', \ldots, H_m'$ for equivalence. (There is no need to test H_j' for $j > m$, since we must reject H_j' if we have rejected H_j.) Accordingly, we consider $t_1', \ldots, t_m'$ (which are ordered, since we have assumed the case of equal n_i). First of all, note that we need only consider t_j' between t_m and t_{m+1}, as any $t_j' > t_{m+1}$ must lead to rejection since t_{m+1} did, and any $t_j' < t_m$ must lead to acceptance since t_m did. Any t_j' between t_m and t_{m+1} has c_m for its critical value and leads to rejection if $t_j' \geqslant c_m$. We can simply state that any t_j' in the sequence $t_1', \ldots, t_m'$ that satisfies $t_j' \geqslant c_m$ leads to the rejection of H_j'. This rule identifies which of the equivalence hypotheses $H_1', \ldots, H_m'$ we reject in the second stage, using in effect the SS procedure with critical constant c_m.

An alternative way to look upon this second stage of SD1 is in terms of lower one-sided confidence limits for the m differences $\mu_0 - \mu_1, \ldots, \mu_0 - \mu_m$: those that satisfy $\bar{y}_0 - \bar{y}_i - c_m s\sqrt{(1/n_0 + 1/n)} \geqslant -\delta$ identify the hypotheses H_i' that are rejected. In this way, we can use SD1 in situations where it may not be possible to specify a value δ in advance to define equivalence; instead, we can test the k superiority hypotheses first, followed by one-sided confidence limits for the θ_i corresponding to those that we cannot claim as superior. Values outside these limits identify the values of δ for which SD1 establishes equivalence.

4.2. Modified Step-Down Test Procedures (SD2, SD3)

In this section, we modify the SD1 testing method in two ways. The first is by removing the restriction that we only consider $t_j' > t_m$ in the second stage; we denote this procedure by SD2.

Table I. Critical constants for procedures ($k = 4$, $v = \infty$, $\rho = 0{\cdot}5$, $\alpha = 0{\cdot}05$)

Procedure	δ	c_1	c_2	c_3	c_4
SS	–	2·160	2·160	2·160	2·160
SD1, SD2	–	1·645	1·916	2·062	2·160
SD3	0·5	1·645	1·938	2·076	2·170
	1	1·645	1·972	2·099	2·190
	2	1·645	2·092	2·184	2·297
SU1, SU2	–	1·645	1·933	2·071	2·165
SU3	0·5	1·645	1·969	2·093	2·178
	1	1·645	2·028	2·133	2·197
	2	1·645	2·258	2·313	2·313

The first stage remains unchanged, and suppose we accept $H_1, \ldots, H_m$ and reject $H_{m+1}, \ldots, H_k$ as in SD1. In the second stage, we replace the SS procedure using the value c_m that we employed in SD1 by the following SD testing procedure:

1. Start with the test statistics $t'_1 \leqslant \ldots \leqslant t'_m$. The first step tests H'_m using c_m: if $t'_m \geqslant c_m$, we reject H'_m and continue to the next step, otherwise we stop testing and accept all remaining H' hypotheses.
2. The general step tests H'_j using the constant c_r, where $r = m$ if $t'_j > t_m$, otherwise r is determined so that $t_r < t'_j < t_{r+1}$, that is, $r = \#(t_i < t'_j)$. If $t'_j \geqslant c_r$, we reject H'_j and continue to the next step. Otherwise, we stop testing and accept all remaining H' hypotheses.

The critical constants $c_1 < \ldots < c_k$ that we use in this modified SD procedure are the same as those defined in the preceding section, that is, $c_m = t^{\alpha}_{m,v,\rho}$. We note that SD2 is more liberal in testing the equivalence hypotheses than the closed testing procedure SD1, since it rejects all hypotheses rejected by SD1 and may reject additional H' hypotheses. We will examine the effect of this modification on the FWE of SD2 in a simulation study described in Section 7.

The second modification is to proceed in an identical manner to SD2, except that we use the same constant c_i to test both the equivalence hypothesis H'_i and the corresponding superiority hypothesis H_i; we denote this procedure by SD3. An equivalent way to describe SD3 is in terms of testing the hypotheses in pairs (H_i, H'_i). We reject both H_i and H'_i if we rejected H_{i+1} and $t_i \geqslant c_i$, we reject only H'_i if we accepted H_{i+1} and we rejected H'_{i+1} and $t_i \geqslant c_i - \delta'$, and we accept all remaining hypotheses if $t_i < c_i - \delta'$.

It turns out that the constants needed to control the FWE in SD3 are larger than those used in SD1 and SD2. The derivation of these constants appears in the Appendix. See Table I for an example of the numerical values for the case $k = 4$, computed to three decimal places, along with the usual SD constants used in SD1 and SD2 for comparison. Note that the values of the SD3 constants depend on δ and are $\geqslant$ the corresponding constants for SD1 and SD2.

5. STEP-UP (SU) TESTING

5.1. A Step-Up Analogue of the Closed Step-Down Procedure (SU1)

The method we propose here is the step-up analogue of the method SD1 described in Section 4.1. We proceed in two stages, as we did for SD1 in Section 4.1. In the first stage, we use the t' statistics to test the equivalence hypotheses, $H'_1, \ldots, H'_k$, starting with t'_1, then t'_2 and so on, in the usual step-up manner, continuing as long as we observe $t'_j < c_r$ (in which case we accept the hypothesis H'_j) where $r = \#(t_i < t'_j)$. The first time we observe $t'_j \geqslant c_r$, we reject H'_j and all remaining H' hypotheses and terminate the first stage.

Suppose we accept $H'_1, \ldots, H'_m$ and reject $H'_{m+1}, \ldots, H'_k$ in the first stage ($1 \leqslant m \leqslant k$). Then in the next stage we test $H_{m+1}, \ldots, H_k$ for superiority. (There is no need to test H_i for $i \leqslant m$, since we must accept H_i if we have accepted H'_i.) Accordingly, we consider $t_{m+1}, \ldots, t_k$. Note that we can accept all H_j hypotheses for which $t_j \leqslant t'_m$ and also any H_j for which $t'_m < t_j < c_j \leqslant t'_{m+1}$. Simply stated, we accept all H_j for which $t_j < c_j \leqslant t'_{m+1}$, and reject any remaining H hypotheses.

We use the same critical constants in the SU1 procedure as the one-sided α points used in the SU procedure defined in Dunnett and Tamhane,[7] with $\rho = 1/(1 + n_0/n)$ and d.f. $= \nu$ corresponding to the variance estimate. Tables are available in Dunnett and Tamhane.[7,12] We do not have a proof that this SU1 procedure controls the FWE as we had for the SD1 procedure. We examine whether or not it satisfies FWE $\leqslant \alpha$ in the simulation study described in Section 7.

5.2. Modified Step-Up Test Procedures (SU2, SU3)

In this section, we define two modifications of SU1 given in the previous section, analogous to the modifications SD2 and SD3 of SD1. The first modified procedure, denoted SU2, tests the equivalence hypotheses $H'_1, \ldots, H'_k$ in the first stage, which is unchanged from the first stage of SU1. Suppose we accept $H'_1, \ldots, H'_m$ and reject $H'_{m+1}, \ldots, H'_k$ in this stage.

In the second stage, we test the superiority hypotheses $H_{m+1}, \ldots, H_k$. We replace the second stage of SU1 by the following step-up testing procedure:

1. Start with the smallest of $t_{m+1}, \ldots, t_k$, which is t_{m+1}, and compare it with c_{m+1}: if $t_{m+1} < c_{m+1}$, accept H_{m+1} and continue with t_{m+2}; otherwise stop testing and reject $H_{m+1}, \ldots, H_k$.
2. The general step compares t_i and c_i, where $m + 1 \leqslant i \leqslant k$. If $t_i < c_i$, accept H_i and continue with t_{i+1}; otherwise stop testing and reject all remaining H hypotheses.

We use the same critical constants for SU2 as those defined in Section 5.1 for SU1. We note that SU2 is more conservative in testing the superiority hypotheses than the procedure SU1, since it rejects no H hypotheses accepted by SU1. We examine the effect of the modification introduced in SU2 on the FWE in the simulation study described in Section 7.

The second modified method SU3 is the step-up analogue of SD3 described in Section 4.2. It proceeds in an identical manner to SU2, except that we use the same constant c_i to test the equivalence hypothesis H'_i as we use to test the corresponding superiority hypothesis H_i. An equivalent way to describe SU3 is in terms of testing the hypotheses in pairs, (H_i, H'_i). We accept both H_i and H'_i if we accepted H'_{i-1} and $t_i < c_i - \delta'$, we accept H_i and reject H'_i if we accepted H_{i-1} and rejected H'_{i-1} and $t_i < c_i$, and we reject all remaining hypotheses if $t_i \geqslant c_i$.

It turns out that the constants needed by SU3 to control the FWE are larger that those used in SU1 and SU2. The derivation of these constants appears in the Appendix. See Table I for an

example of the numerical values, given to three decimal places, along with the usual SU constants used for SU1 and SU2 for comparison. Note that the values of the SU3 constants depend on δ and are $\geqslant$ the corresponding constants for SU1 and SU2.

6. NUMERICAL EXAMPLE

In a randomized clinical trial, suppose we compare a test treatment T with four standard treatments S_1, S_2, S_3 and S_4 in parallel groups (one-way layout) with $n_0 = n_i = n$. The observed sample means are:

$\bar{y}_0$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
8·68	6·97	6·94	5·80	4·55

Suppose that the standard deviation of $(\bar{y}_0 - \bar{y}_i)$ is $\sigma\sqrt{(2/n)} = \sqrt{2}$ and $v = \infty$ (σ is known). Test the hypotheses H_i and H_i' for $\delta = 1{\cdot}0$ with FWE $\leqslant 0{\cdot}05$.

First, we compute the test statistics defined in (1). Note that $\delta' = 1/\sqrt{2} = 0{\cdot}71$. We obtain the following values for the test statistics:

Statistic	1	2	3	4
t_i	1·22	1·23	2·04	2·92
t_i'	1·93	1·94	2·75	3·63

Using the critical constants shown in Table I, we obtain the following results by applying different procedures defined in the previous sections:

SS procedure The single-step procedure uses the critical value $c_4 = 2{\cdot}160$ for all tests; since t_4, t_3' and t_4' are the only statistics to exceed this value, we reject H_4, H_3' and H_4' and conclude that T is superior to S_4 and equivalent to S_3.

SD1 procedure We find $t_4 > 2{\cdot}160$, $t_3 < 2{\cdot}062$, so we reject H_4 but not H_1, H_2, H_3. Next we test t_3', t_2' and t_1' against $c_3 = 2{\cdot}062$; since we find $t_3' > 2{\cdot}062$ we reject H_3', but t_2' and t_1' are $< 2{\cdot}062$ so we accept H_2' and H_1'. Thus, using SD1, we conclude T is superior to S_4 and equivalent to S_3.

SD2 procedure SD2 gives the same results for the t-statistics and it rejects t_3' as in SD1. For the remaining t'-statistics, we test t_2' against $c_2 = 1{\cdot}916$ and t_1' against $c_1 = 1{\cdot}645$; we find $t_2' > 1{\cdot}916$ and $t_1' > 1{\cdot}645$, so in addition to rejecting H_4 and H_3' we reject H_2' and H_1'. Thus, using SD2, we conclude that T is superior to S_4 and equivalent to S_1, S_2 and S_3.

SD3 procedure Using SD3, we find that $t_4 > 2{\cdot}190$ so we reject H_4 and H_4', and that $t_3 < 2{\cdot}099$ but $t_3' > 2{\cdot}099$ so we accept H_3 and reject H_3'. At the next step, we can test only H_2'; since $t_2' < 1{\cdot}972$, we accept H_2' as well as H_1' by implication. We conclude that T is superior to S_4 and equivalent to S_3.

SU1 procedure We find $t_1' < c_2 = 1{\cdot}933$, $t_2' > 1{\cdot}933$ (remember that we count the number of smaller t_i statistics to determine the index of the constant to use), so we accept H_1' and reject H_2',

Table II. Decisions* for numerical example in Section 6

Procedure	Standard			
	S_1	S_2	S_3	S_4
SS	–	–	e	s
SD1	–	–	e	s
SD2	e	e	e	s
SD3	–	–	e	s
SU1	–	e	s	s
SU2	–	e	e	s
SU3	e	e	e	s

* s superior, e equivalent, – no claim

H'_3 and H'_4. In the second stage, we find $t_2 < 1{\cdot}933$ so we accept H_2, and $t_3 > t'_2$ which led to an H' rejection so we reject H_3 and H_4. Using SU1, we conclude T is superior to S_3 and S_4, equivalent to S_2 but can make no claim for equivalence or superiority with respect to S_1.

SU2 procedure The first stage is the same as for SU1. In the second stage, $t_2 < 1{\cdot}933$, $t_3 < 2{\cdot}071$, $t_4 > 2{\cdot}165$ so we accept H_2 and H_3 and reject H_4. Thus, we claim T is superior to S_4, equivalent to S_3 and S_2, and no claim with respect to S_1.

SU3 procedure We find that $t'_1 > 1{\cdot}645$ and $t_1 < 1{\cdot}645$ so we reject H'_1 and accept H_1, and also reject H'_2, H'_3, H'_4 by implication. Next, since $t_2 < 2{\cdot}028$ and $t_3 < 2{\cdot}133$ we accept H_2 and H_3, but $t_4 > 2{\cdot}197$ so we reject H_4. Thus, using SU3, we claim that T is equivalent to S_1, S_2, S_3 and superior to S_4.

In Table II, we tabulate the decisions we made by the six procedures concerning the status of T with respect to each of the standards S_1 to S_4.

7. SIMULATION STUDIES OF FWE AND POWER

7.1. Description of the Simulation Studies

To compare the procedures with respect to their FWE and power, we carried out two simulation studies for a particular case ($k = 4$, $\nu = \infty$, $\delta = 1{\cdot}0$, $\sigma/\sqrt{n} = 1{\cdot}0$, $\alpha = 0{\cdot}05$). We calculated the values of the critical constants c_1, c_2, c_3 and c_4 for each procedure to 4-decimal place accuracy.

In the FWE study, we selected several null configurations of the parameters. For each configuration, we chose the number of simulations used to obtain estimates of the FWE for the various procedures as 100,000 in order to obtain a sufficiently high precision (standard error = 0·0007) to identify values that exceed the nominal value of α. We used the same set of simulations to obtain estimates for each of the methods: this introduced a positive correlation between the estimates which increased the precision of the estimated differences between methods (standard error = 0·0001 approximately). The entire set took approximately seven minutes of computing time on a 486DX2 66 MHz PC. The values of FWE obtained for each method appear in Table III.

Table III. Simulated FWE for several null configurations ($\delta = 1$, $\sigma/\sqrt{n} = 1$, $v = \infty$, $\alpha = 0{\cdot}05$)

Configuration					Single-step	Step-down			Step-up		
Number	θ_1	θ_2	θ_3	θ_4	SS	SD1	SD2	SD3	SU1	SU2	SU3
1	−1	−1	−1	−1	0·0489	0·0489	0·0489	0·0456	0·0482	0·0482	0·0452
2	−1	−1	−1	0	0·0496	0·0496	0·0496	0·0479	0·0491	0·0491	0·0472
3	−1	−1	−1	10	0·0397	0·0499	0·0499	0·0457	0·0490	0·0490	0·0433
4	−1	−1	0	0	0·0501	0·0501	0·0501	0·0493	0·0497	0·0496	0·0493
5	−1	−1	0	10	0·0389	0·0491	0·0494	0·0477	0·0483	0·0483	0·0455
6	−1	−1	10	10	0·0278	0·0501	0·0501	0·0437	0·0483	0·0483	0·0411
7	−1	0	0	0	0·0484	0·0484	0·0486	0·0484	0·0486	0·0483	0·0496
8	−1	0	0	10	0·0392	0·0494	0·0498	0·0492	0·0493	0·0489	0·0490
9	−1	0	10	10	0·0275	0·0492	0·0503	0·0492	0·0485	0·0485	0·0492
10	−1	10	10	10	0·0148	0·0486	0·0486	0·0486	0·0486	0·0486	0·0486
11	0	0	0	0	0·0497	0·0497	0·0497	0·0463	0·0506	0·0497	0·0459
12	0	0	0	10	0·0391	0·0494	0·0494	0·0454	0·0505	0·0490	0·0425
13	0	0	10	10	0·0281	0·0503	0·0503	0·0441	0·0530	0·0501	0·0417
14	0	10	10	10	0·0153	0·0492	0·0492	0·0492	0·0492	0·0492	0·0492

In the second simulation study to compare the procedures for power, we chose the number of simulations for each configuration as 20,000. Since the power estimates are also positively correlated due to the use of the same simulations to obtain the estimates for each method, this number should provide adequate precision for comparing methods. The entire set took slightly over one minute of computing time.

In applications of the methods, a user wishes to make correct decisions for the new treatment with respect to each standard according to the values of the θ_i. If a particular value of θ_i is sufficiently large to be considered an 'important' difference, he or she wishes to find the new treatment superior to that standard or, failing that, equivalent. Accordingly, we chose the following two definitions of power: the probability of finding all of the θ_i which are > 0 as superior, and the probability of finding all such θ_i equivalent or better (that is, either equivalent or superior). Tables IV and V show the results.

7.2. Discussion of FWE and Power Results

Consider the FWE simulation results shown in Table III. The single-step procedure SS meets the FWE requirement, as expected, but clearly it is overly conservative for some configurations. Of the three step-down procedures, SD1 is based on closure and its critical values have been determined to guarantee that it meets the requirement FWE $\leqslant \alpha$; we see that the simulated values verify this. With respect to SD2, we know it is more liberal than SD1. However, for some configurations, this has no effect on the FWE (for example, configurations 1 and 3 where it is impossible for SD2 to reject without SD1 also rejecting, and configurations 11–14 where none of the H_i' is true). The simulation results indicate that the increases in FWE when they occur are small. The largest increase is for configuration 9 where the FWE of SD2 is 0·0011 ($\pm$0·0001 standard error) higher than the corresponding value for SD1. Moreover, the estimated FWE for

Table IV. Simulated power = P(find all $\theta_i > 0$ superior) ($\delta = 1$, $\sigma/\sqrt{n} = 1$, $\nu = \infty$, $\alpha = 0{\cdot}05$)

Configuration					Single-step	Step-down			Step-up		
Number	θ_1	θ_2	θ_3	θ_4	SS	SD1	SD2	SD3	SU1	SU2	SU3
1	−1	−1	−1	2	0·228	0·228	0·228	0·221	0·227	0·227	0·219
2	−1	−1	−1	3	0·486	0·486	0·486	0·475	0·485	0·485	0·472
3	−1	−1	−1	4	0·750	0·750	0·750	0·740	0·748	0·748	0·737
4	−1	−1	2	2	0·105	0·125	0·125	0·116	0·125	0·124	0·110
5	−1	−1	2	4	0·216	0·246	0·246	0·235	0·245	0·244	0·225
6	−1	−1	3	3	0·322	0·359	0·359	0·345	0·358	0·357	0·333
7	−1	−1	4	4	0·618	0·655	0·655	0·641	0·654	0·653	0·628
8	2	2	2	2	0·041	0·112	0·112	0·106	0·131	0·131	0·131
9	3	3	3	3	0·187	0·363	0·363	0·354	0·389	0·389	0·389
10	4	4	4	4	0·473	0·679	0·679	0·672	0·698	0·698	0·698

Table V. Simulated power = P(find all $\theta_i > 0$ equivalent or superior) ($\delta = 1$, $\sigma/\sqrt{n} = 1$, $\nu = \infty$, $\alpha = 0{\cdot}05$)

Configuration					Single-step	Step-down			Step-up		
Number	θ_1	θ_2	θ_3	θ_4	SS	SD1	SD2	SD3	SU1	SU2	SU3
1	−1	−1	−1	2	0·487	0·487	0·487	0·474	0·484	0·484	0·472
2	−1	−1	−1	3	0·748	0·748	0·748	0·739	0·747	0·747	0·736
3	−1	−1	−1	4	0·914	0·914	0·914	0·910	0·913	0·913	0·909
4	−1	−1	2	2	0·316	0·331	0·332	0·342	0·330	0·330	0·330
5	−1	−1	2	4	0·477	0·507	0·508	0·502	0·505	0·505	0·489
6	−1	−1	3	3	0·616	0·635	0·636	0·640	0·633	0·633	0·628
7	−1	−1	4	4	0·857	0·871	0·871	0·871	0·870	0·870	0·864
8	2	2	2	2	0·186	0·250	0·262	0·351	0·259	0·258	0·387
9	3	3	3	3	0·476	0·580	0·595	0·677	0·592	0·591	0·704
10	4	4	4	4	0·768	0·850	0·857	0·898	0·855	0·855	0·909

SD2 exceeds the nominal value 0·05 slightly. (Note that the small excess FWE = 0·0503 for configuration 13 must be a random event, since this is one of the configurations where the FWE of SD1 and SD2 coincide.) To examine the difference for configuration 9 more closely, we conducted a set of 14 repetitions of the simulations for this configuration. The results obtained were mean FWE = 0·0504 ($\pm$0·000185 standard error). We conclude that SD2 does not guarantee the FWE requirement for all configurations, but the excess when it occurs is small. Regarding SD3, we determined its critical values to ensure that it meets the FWE requirement for the configurations conjectured to be least favourable; the simulated values indicate that SD3 is successful in meeting this requirement.

Of the three step-up procedures, SU1 fails to meet the FWE requirement for configurations 11, 12 and 13; accordingly, we must consider it unsatisfactory in this respect. On the other hand, SU2 does satisfy FWE $\leqslant$ 0·05 for these configurations as well as all the others, thus the modification of

SU1 that we used for SU2 is apparently successful in adjusting the FWE down to a satisfactory level. We note also that FWE for SU2 is $\leqslant$ FWE for SD1 for all configurations, which suggests that it is conservative relative to the closed procedure SD1. The simulated values for SU3 indicate that, like SD3, it meets the FWE requirement for all configurations.

Now consider the power results shown in Tables IV and V. The precision of the observed differences in power between methods is such that we can take any observed difference $\geqslant 0{\cdot}002$ in these tables as statistically significant.

Consider the effect of the modification to the closure method introduced in Section 4.2 and its step-up counterpart in Section 5.2; in most cases, the power differences between SD2 and SD1 and between SU2 and SU1 are quite negligible, as we might expect. However, there are some exceptions in Table V, especially for configurations 8–10 where all four $\theta_i > 0$. Examination of the individual simulations where these instances occurred revealed that, in all cases, SD1 failed to reject one or two H_i' hypotheses whereas SD2 and usually the other stepwise procedures as well succeeded in rejecting them.

Note also that the methods SD3 and SU3 seem to be at a disadvantage compared with their counterparts in Table IV, whereas in Table V they have markedly higher powers under configurations 8–10. The explanation for this is that the constants used by SD3 and SU3 are larger than those used by the other stepwise procedures, which makes it more difficult to reject the superiority hypotheses H_i. On the other hand, it is easier to reject equivalence hypotheses H_i' using SD3 and SU3, since the other methods often require the use of a constant with a higher index.

We note that the step-up methods are superior to their step-down counterparts for configurations 8 to 10 in Table IV while the step-down methods tend to be superior for configurations 1 to 7. This is in accord with the results obtained in our earlier paper (Dunnett and Tamhane[7]) where we noted that SU has higher power than SD when all or most hypotheses are false whereas SD has higher power than SU when one or a few hypotheses are false.

We limited the simulations to d.f. $= \infty$ (known variance case) because we did not expect the results to differ much for small d.f. To check this, we repeated the power simulations for d.f. $= 10$. We found qualitatively similar results, with perhaps slightly better results for methods SD3 and SU3 compared with their counterparts than were in evidence when d.f. $= \infty$.

8. EXTENSIONS TO UNBALANCED DATA

In Sections 4 to 7 above, we have assumed the one-way setup with balanced data, that is, $n_1 = \cdots = n_k = n$ with n_0 possibly different from n. In this section, we discuss the changes needed to extend the results to the case of unequal n_i's.

There are two consequences of having unequal n_i's: (i) the δ_i' in equation (1) are unequal, which makes the ordering of the t_i' possibly different from the ordering of the t_i; (ii) the correlation structure of the multivariate t random variables that arise in the computation of the critical constants is no longer $\rho_{ij} = \rho = 1/(1 + n_0/n)$, but is

$$\rho_{ij} = \sqrt{\left(\frac{1}{1 + n_0/n_i}\right)}\sqrt{\left(\frac{1}{1 + n_0/n_j}\right)} \quad (1 \leqslant i \neq j \leqslant k). \tag{3}$$

We now consider the effects of having unequal n_i's on each of the procedures described in the preceding sections.

SS procedure This procedure remains unchanged from its description in Section 3 except for replacing δ' by δ'_i and defining c_k differently. To control the FWE $\leqslant \alpha$, we choose $c_k = t^{\alpha}_{k,\nu,\mathscr{R}}$ which is the one-sided α point of k-variate t with ν d.f. and correlation matrix $\mathscr{R} = (\rho_{ij})$. The value of c_k is identical with that of c_k used by SD1 and SD2 below.

SD1 and SD2 procedures Denote the ordered values of the test statistics for the superiority hypotheses by $t_1 \leqslant \ldots \leqslant t_k$. The first stage of SD1 and SD2, which uses the t_i in a step-down fashion to test the superiority hypotheses $H_1, \ldots, H_k$, is unchanged from before. Suppose that we accept $H_1, \ldots, H_m$ and reject $H_{m+1}, \ldots, H_k$. At the second stage, we test the equivalence hypotheses $H'_1, \ldots, H'_m$ using $t'_{(1)} \leqslant \ldots \leqslant t'_{(m)}$ which are the ordered values of $t'_1, \ldots, t'_m$. Then we proceed as we did in Sections 4.1 and 4.2, except that we use the latter ordered test statistics. For SD1, any $t'_{(j)}$ in the sequence $t'_{(1)}, \ldots, t'_{(m)}$ that satisfies $t'_{(j)} \geqslant c_m$ leads to the rejection of the corresponding hypothesis $H'_{(j)}$. For SD2, we use the $t'_{(j)}$ in a step-down manner, starting with $t'_{(m)}$, then $t'_{(m-1)}$ etc. We reject $H'_{(j)}$ if $t'_{(j)} \geqslant c_r$, where $r = m$ if $t'_{(j)} > t_m$ and $r = \#(t_i < t'_{(j)})$ if $t'_{(j)} < t_m$, and go to $H'_{(j-1)}$; otherwise we stop testing and accept $H'_{(1)}, \ldots, H'_{(j)}$.

We may determine the critical constants $c_1 < \ldots < c_k$ as described in Dunnett and Tamhane.[6] This uses the central t random variables $T_1, \ldots, T_k$ corresponding to the observed ordered statistics $t_1 \leqslant \ldots \leqslant t_k$; we determine c_r for $r = 1, 2, \ldots, k$ so that

$$P\{T_1 < c_r, \ldots, T_r < c_r\} = 1 - \alpha. \tag{4}$$

The results are $c_r = t^{\alpha}_{r,\nu,\mathscr{R}_r}$, where $t^{\alpha}_{r,\nu,\mathscr{R}_r}$ is the upper α equicoordinate critical point of the central r-variate t distribution with d.f. $= \nu$ and correlation matrix $\mathscr{R}_r$ associated with $T_1, \ldots, T_r$. However, according to Liu[13] this method may not always satisfy the FWE requirement, although simulation evidence indicates that it does. He proposed that we determine c_r for $r = 1, 2, \ldots, k$ from the equation

$$\min_{1 \leqslant i_1 < \ldots < i_r \leqslant k} P\{T_{i_1} < c_r, \ldots, T_{i_r} < c_r\} = 1 - \alpha; \tag{5}$$

the resulting solution is conservative. He showed that this minimum is achieved when $T_{i_1}, \ldots, T_{i_r}$ are associated with the r smallest sample sizes. Thus $c_r = t^{\alpha}_{r,\nu,\mathscr{R}_r}$ as above, except that the correlation matrix $\mathscr{R}_r$ is associated with the new treatment versus standard comparisons involving the r smallest standard groups.

SU1 and SU2 procedures We start as before by labelling the test statistics and their associated hypotheses so that $t'_1 \leqslant \ldots \leqslant t'_k$, but we do not necessarily have $t_1 \leqslant \ldots \leqslant t_k$. The first stage of SU1 and SU2, which uses the t'_i in a step-up fashion to test the equivalence hypotheses $H'_1, \ldots, H'_k$, is the same as before. Suppose that we accept $H'_1, \ldots, H'_m$ and reject $H'_{m+1}, \ldots, H'_k$. At the second stage, we test the superiority hypotheses $H_{m+1}, \ldots, H_k$ using $t_{m+1}, \ldots, t_k$. Denote their ordered values by $t_{(m+1)} \leqslant \ldots \leqslant t_{(k)}$. Then we proceed as we did in Sections 5.1. and 5.2, except that we use the latter ordered test statistics. For SU1, any $t_{(j)}$ in the sequence $t_{(m+1)}, \ldots, t_{(k)}$ that satisfies $t_{(j)} < c_j \leqslant t'_{m+1}$ leads to the acceptance of the corresponding hypothesis $H_{(j)}$. For SU2, we use the $t_{(j)}$ in a step-up manner, starting with $t_{(m+1)}$, then $t_{(m+2)}$ etc., accepting $H_{(j)}$ if $t_{(j)} < c_j$, and going to $H_{(j+1)}$; otherwise testing stops and we reject $H_{(j)}, \ldots, H_{(k)}$.

We may determine the critical constants $c_1 < \ldots < c_k$ as described in Dunnett and Tamhane.[8] After calculating $c_1, \ldots, c_{r-1}$, we determine c_r for $r = 1, 2, \ldots, m$ $(1 \leqslant m \leqslant k)$ so that

$$P\{(T_1, \ldots, T_r) < (c_1, \ldots, c_r)\} = 1 - \alpha \tag{6}$$

Table VI. Results from GUSTO[5] clinical trial

Group	Treatment	Sample size, N	Mortality rate r	Standard error, SE	T_1 versus S_i		T_2 versus S_i	
					t_i	t_i'	t_i	t_i'
S_1	SK + IV hep.	9,796	7·2%	0·26	2·54	3·96	0·55	1·93
S_2	SK + SC hep.	10,377	7·4%	0·26	3·14	4·56	1·11	2·51
T_1	t-PA + IV hep.	10,344	6·3%	0·24	–	–	–	–
T_2	mix. + IV hep.	10,328	7·0%	0·25	–	–	–	–

where $(T_1, \ldots, T_r)$ denotes the ordered values of the central t random variables $T_1, \ldots, T_r$ corresponding to the statistics $t_1 \leqslant \ldots \leqslant t_r$. As noted in Dunnett and Tamhane,[8,12] we cannot claim that this method of determining the constants always satisfies FWE $\leqslant \alpha$, although the simulation evidence suggested that it does. However, Grechanovsky and Pinsker[14] have constructed a counterexample for which the FWE exceeds α by a small amount. The conservative method of Liu[13] determines c_r, after calculating $c_1, \ldots, c_{r-1}$, from the equation

$$\min_{1 \leqslant i_1 < \ldots < i_r \leqslant k} P\{T_{i_1}, \ldots, T_{i_r}) < (c_1, \ldots, c_r)\} = 1 - \alpha. \tag{7}$$

We must determine the minimum in this equation numerically, since it is not necessarily achieved by defining the T random variables as associated with the r smallest sample sizes as in equation (5).

SD3 and SU3 procedures The descriptions of these procedures are identical to those given in Sections 4.2 and 5.2 except that the constant for testing H_i' using t_i is $c_i - \delta_i'$ instead of $c_i - \delta'$.

9. APPLICATION TO A CLINICAL TRIAL

We use the GUSTO[5] clinical trial to illustrate the application of the stepwise testing methods presented in this paper. There were two standard treatments: SK (streptokinase) with intravenous heparin, and SK with subcutaneous heparin. The two test treatments were t-PA and a mixture of both t-PA and SK (with lower dose levels of these two agents in the mixture than were used with each given alone), along with intravenous heparin. Table VI shows the sample sizes and the 30-day mortality rates observed in the trial, along with the standard errors $(\mathrm{SE} = \sqrt{\{r(100 - r)/N\}}$ of each rate.

The t (or z) statistics for comparing T_1 with the two standards, using the formula for comparing two rates

$$t = (r_1 - r_2)/\sqrt{\{r_1(100 - r_1)/N_1 + r_2(100 - r_2)/N_2\}},$$

appear in column 6 of Table VI. These are the statistics for testing the two superiority hypotheses H_1 and H_2 for T_1. The corresponding equivalence statistics depend on the value δ defined as a clinically negligible difference, and we calculate them from

$$t' = (r_1 - r_2 + \delta)/\sqrt{\{r_1(100 - r_1)/N_1 + r_2(100 - r_2)/N_2\}}.$$

Suppose that we agree upon $\delta = 0{\cdot}5$ as an appropriate value. Then we obtain the values of the t' statistics shown in column 7 of the table.

In this case, the statistics are bivariate ($k = 2$). We compute their correlation coefficient as

$$\rho = \mathrm{SE}^2(T_1)/(\mathrm{SE}^2(S_1) + \mathrm{SE}^2(S_2)) = 0{\cdot}46$$

where $\mathrm{SE}(\cdot)$ denotes the standard error of the rate for the indicated group. This is close enough to $\rho = 0{\cdot}5$ for us to use the $\alpha = 0{\cdot}05$ critical values c_1 and c_2 given in Table I for any of the procedures. Whichever procedure we use, we reject H_1 and H_2 at level $\alpha = 0{\cdot}05$ for the comparisons of T_1 with the two standards and conclude that T_1 is superior to S_1 and S_2.

Similarly, to compare T_2 with the two standards, we obtain the t and t' statistics shown in columns 8 and 9 of the table. For these statistics, we find $\rho = 0{\cdot}48$, so again we can use the critical values given in Table I. Whichever procedure we use, we accept H_1 and H_2 and reject H'_1 and H'_2 at level $\alpha = 0{\cdot}05$ for the comparisons T_2 with the two standards. We conclude that, although we cannot show that T_2 is superior to either standard, we can claim that is at least equivalent to both S_1 and S_2.

In the above, we have assumed that the purpose of the trial was to compare each of the two test treatments separately with the two standards. If the purpose was to compare the better of the two test treatments, defined as the one that produced the lower mortality rate in the trial, with the two standards, then we can use the Bonferroni adjustment for the multiplicity effect of having two candidates instead of one and perform the tests at level $\alpha/2$ instead of α. This requires the use of 0·025 instead of 0·05 critical values. (For example, with SD1 or SD2, we use $c_1 = 1{\cdot}96$ and $c_2 = 2{\cdot}21$ instead of the values given in Table I. This does not affect the decisions reached for T_1 in this application.)

10. DISCUSSION

In this paper, we proposed some stepwise testing procedures of both the step-down (SD) and step-up (SU) type for comparing a new treatment with a set of $k > 1$ standard treatments in a clinical trial, where the purpose of the trial is to classify the standard treatments into those to which the new treatment is superior, those to which it is at least equivalent and those for which no claim can be made for the new treatment. The procedures employ one-sided hypothesis tests for both superiority and equivalence, which may be more informative than the customary approach of testing two-sided hypotheses for positive and/or negative differences. For $k = 1$, all the methods reduce to the method proposed by Dunnett and Gent[2] for comparing a new treatment with a single standard.

The step-down methods denoted by SD1 and SD2 are extensions of the usual one-sided SD method that tests only for superiority, in that they coincide with the latter with respect to the tests of the superiority hypotheses, but in addition provide tests for equivalence. Similarly, the step-up methods denoted by SU1 and SU2 are extensions of the usual one-sided SU method. The methods SD3 and SU3 use the same constants c_i to test the hypotheses H_i and H'_i, which is intuitively reasonable but the constants required are larger. Furthermore, the constants for these methods depend upon the threshold value δ for defining equivalence and hence must be computed for each application; thus the SD3 and SU3 methods may be impractical. However, they do have some power advantages over the other methods when all $\theta_i > 0$ and we are concerned with finding either equivalence or superiority – see Table V. Note that the constants used to test the equivalence hypotheses in SD1, SD2, SU1 and SU2 also depend on δ, but only the index of the constant is affected, hence we do not need to compute them for each chosen value of δ.

The SD2 method is a modification of the closed method SD1, but we found that it has a FWE that slightly exceeds the nominal value for certain null parameter configurations, although the excess is very small. The SU1 method, which is the step-up analogue of SD1, fails to meet the FWE requirement and accordingly we do not recommend its use. On the other hand, the SU2 method which employs an analogous modification to that used by SD2, meets the FWE requirement in our simulation study. (Recall that the effects of this modification are opposite in the two cases: SD2 is more liberal than SD1, while SU2 is more conservative than SU1.) It also has some power advantages over the other procedures. Accordingly, we recommend it as an alternative to SU1.

In choosing between SD1 (or SD2, if we can tolerate a slight increase in FWE) and SU2 in a particular application, if our main aim is detection of superiority, Table IV shows that we may prefer SD1 when we expect the new treatment is superior to one or a few of the k standards whereas we may prefer SU2 if we expect the new treatment is superior to all or most of the standards. In terms of detecting either equivalence or superiority, Table V shows that SD1 has a slight edge if not all of them are superior (configurations 1–7), while SU2 dominates SD1 when all of them are superior (configurations 8–10).

Dunnett and Gent[2] allowed for the possibility of having two α levels, α_1 for testing the equivalence hypothesis and α_2 for testing the superiority hypothesis. For $k = 1$, our methods reduce to the special case where $\alpha_1 = \alpha_2 = \alpha$. If a particular application needs two α levels, we can easily extend the SD1, SD2, SU1 and SU2 method to use two sets of constants, namely, a set $c'_1, \ldots, c'_k$ that are level α_1 for the H' family of hypotheses and a set $c_1, \ldots, c_k$ that are level α_2 for the H family of hypotheses. The FWE for the combined family, defined in equation (2), is then $\leqslant \max(\alpha_1, \alpha_2)$.

We note that there are some open questions for future research: to show that SU2 controls the FWE as we only have simulation evidence for this and to prove the conjectures concerning the parameter configurations that lead to the maximum FWE for procedures SD3 and SU3.

APPENDIX: DERIVATION OF CRITICAL CONSTANTS FOR SD3 AND SU3

To determine the constants $c_1, \ldots, c_k$ for the SD3 procedure to meet the requirement that the type I FWE $\leqslant \alpha$, we first state the following:

Conjecture: The type I FWE is maximum at a configuration $\boldsymbol{\theta}_r$ of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ for which $\theta_i = -\delta$ for $i \leqslant k - r$ and $\theta_i = 0$ for $i > k - r$, for some r: $0 \leqslant r \leqslant k$.

Thus we choose the constants to satisfy

$$\max[\mathrm{FWE}(\boldsymbol{\theta}_0), \mathrm{FWE}(\boldsymbol{\theta}_1), \ldots, \mathrm{FWE}(\boldsymbol{\theta}_k)] \leqslant \alpha. \tag{8}$$

We determine these constants recursively, beginning with $k = 1$ where it is easy to see that $\mathrm{FWE}(\boldsymbol{\theta}_0) = \mathrm{FWE}(\boldsymbol{\theta}_1) = \alpha$ if $c_1 = t^{\alpha}_{\nu}$, the one-sided α-point of Student's t. Next, for $k = 2$, we determine c_2 to satisfy

$$\max[\mathrm{FWE}(\boldsymbol{\theta}_0), \mathrm{FWE}(\boldsymbol{\theta}_1), \mathrm{FWE}\ (\boldsymbol{\theta}_2)] \leqslant \alpha$$

using the value determined previously for c_1, and so on for $c_3, \ldots, c_k$.

The following argument shows that we must have $c_k \geqslant t^{\alpha}_{k,\nu,\rho}$ for all $k \geqslant 1$. For configuration $\boldsymbol{\theta}_k = (0, \ldots, 0)$, only rejection of an H hypothesis constitutes a type I error: hence we have

$$\begin{aligned} \mathrm{FWE}(\boldsymbol{\theta}_k) &= 1 - P\{\text{accept } H_1, \ldots, H_k | (0, \ldots, 0)\} \\ &= 1 - P\{\max(T_i) < c_k | (0, \ldots, 0)\}. \end{aligned} \tag{9}$$

where the T_i are the central t random variables corresponding to the observed statistics t_i. In a similar manner, for configuration $\boldsymbol{\theta}_0 = (-\delta, \ldots, -\delta)$, accepting H'_i implies H_i is also accepted so that we need to consider only the H' hypotheses; hence we have

$$\begin{aligned} \mathrm{FWE}(\boldsymbol{\theta}_0) &= 1 - P\{\text{accept } H'_1, \ldots, H'_k | (-\delta, \ldots, -\delta)\} \\ &= 1 - P\{\max(T_i) < c_k - \delta' | (-\delta, \ldots, -\delta)\} \\ &= 1 - P\{\max(T_i) < c_k | (0, \ldots, 0)\}. \end{aligned} \tag{10}$$

Thus, for any k, we have $\mathrm{FWE}(\boldsymbol{\theta}_0) = \mathrm{FWE}(\boldsymbol{\theta}_k) = \alpha$ if we choose $c_k = t^{\alpha}_{k,\nu,\rho}$. Therefore, to satisfy equation (8) we must have $c_k \geqslant t^{\alpha}_{k,\nu,\rho}$ for all $k \geqslant 1$. (In the computations, except for $k = 1$, we found that the maximum in (8) always occurred at some $\boldsymbol{\theta}_r$ where $r \neq 0$ or k, so in fact a strict inequality holds for $k > 1$.)

For $m = 0, 1, \ldots, k$, define

$$\begin{aligned} P_m(\boldsymbol{\theta}) &= P\{\text{reject } H'_{k-m+1}, \ldots, H'_k \text{ only} | \boldsymbol{\theta}\} \\ &= P\{T_1 < c_{k-m} - \delta', \ldots, T_{k-m} \\ &\quad < c_{k-m} - \delta', (c_{k-m+1} - \delta', \ldots, c_k - \delta') \\ &\quad \leqslant (T_{k-m+1}, \ldots, T_k) < (c_k, \ldots, c_k) | \boldsymbol{\theta}\} \end{aligned} \tag{11}$$

where $(T_{k-m+1}, \ldots, T_k)$ denotes the ordered values of $T_{k-m+1}, \ldots, T_k$, which are between $c_j - \delta'$ and c_k for $j = k - m + 1, \ldots, k$, respectively. (Note that P_0 represents the probability that we accept all hypotheses, and P_k represents the probability that we reject all H' hypotheses and accept all H hypotheses.) Then

$$\mathrm{FWE}(\boldsymbol{\theta}_r) = 1 - \sum_{m=0}^{r} \binom{r}{m} P_m(\boldsymbol{\theta}_r). \tag{12}$$

We can use the following recursive formula (where $d > 0$) to express the event in the last line of (11) as a union ($\cup$) of disjoint events, enabling us to evaluate $P_m(\boldsymbol{\theta})$ as a sum of $m!$ probability expressions:

$$\begin{aligned} &[(c_r - d, c_{r+1} - d, \ldots, c_k - d) \leqslant (T_r, T_{r+1}, \ldots, T_k) < (c_k, c_k, \ldots, c_k)] \\ &\quad = \{c_r - d \leqslant T_r < c_{r+1} - d, [(c_{r+1} - d, \ldots, c_k - d) \leqslant (T_{r+1}, \ldots, T_k) < (c_k, \ldots, c_k)]\} \\ &\qquad \bigcup \{c_{r+1} - d \leqslant T_r < c_{r+2} - d, [(c_r - d, c_{r+2} - d, \ldots, c_k - d) \\ &\qquad \leqslant (T_{r+1}, \ldots, T_k) < (c_k, \ldots, c_k)]\} \\ &\qquad \bigcup \ldots \\ &\qquad \bigcup \{c_k - d \leqslant T_r < c_k, [(c_r - d, c_{r+1} - d, \ldots, c_{k-1} - d) \leqslant (T_{r+1}, \ldots, T_k) < (c_k, \ldots, c_k)]\}. \end{aligned} \tag{13}$$

To determine the constants for unequal n_i, we first extend the notation for P_m given in equation (11) (so that we can specify any set of m H' hypotheses as rejected), as follows:

$$\begin{aligned} P_{(j_1, \ldots, j_m)}(\boldsymbol{\theta}) &= P\{\text{reject } H'_{j_1}, \ldots, H'_{j_m} \text{ only} | \boldsymbol{\theta}\} \\ &= P\{T'_i < c_{k-m},\ i \neq (j_1, \ldots, j_m),\ (c_{k-m+1}, \ldots, c_k) \\ &\leqslant (T'_{j_1}, \ldots, T'_{j_m}), (T_{j_1}, \ldots, T_{j_m}) < (c_k, \ldots, c_k) | \boldsymbol{\theta}\} \end{aligned}$$

for $m = 0, 1, \ldots, k$; here $T'_i = T_i + \delta_i$. In place of equation (12) we have

$$\text{FWE}(\boldsymbol{\theta}_r) = 1 - \sum_{m=0}^{r} \sum_{k-r < j_1 < \cdots < j_m \leqslant k} P_{(j_1, \ldots, j_m)}(\boldsymbol{\theta}_r). \tag{14}$$

We then determine the constants recursively, as before, to satisfy equation (8).

To determine the constants required for SU3, we make the same conjecture made above for SD3 and determine them recursively to satisfy (8) as before. The following expressions are analogous to equations (9) and (10):

$$\begin{aligned} \text{FWE}(\boldsymbol{\theta}_k) &= 1 - P\{\text{accept } H_1, \ldots, H_k | (0, \ldots, 0)\} \\ &= 1 - P\{(T_1, \ldots, T_k) < (c_1, \ldots, c_k) | (0, \ldots, 0)\} \end{aligned}$$

and

$$\begin{aligned} \text{FWE}(\boldsymbol{\theta}_0) &= 1 - P\{\text{accept } H'_1, \ldots, H'_k | (-\delta, \ldots, -\delta)\} \\ &= 1 - P\{(T_1, \ldots, T_k) < (c_1 - \delta', \ldots, c_k - \delta') | (-\delta, \ldots, -\delta)\} \\ &= 1 - P\{T_1, \ldots, T_k) < (c_1, \ldots, c_k) | (0, \ldots, 0)\}. \end{aligned}$$

Thus, for any k, $\text{FWE}(\boldsymbol{\theta}_0) = \text{FWE}(\boldsymbol{\theta}_k) = \alpha$ if we use the usual SU constants defined in Dunnett and Tamhane,[8] which satisfy

$$P\{(T_1, \ldots, T_k) < (c_1, \ldots, c_k) | (0, \ldots, 0)\} = 1 - \alpha.$$

Therefore, to satisfy equation (8), c_k for SU3 must be $\geqslant c_k$ for SU, for all $k \geqslant 1$. We can compute the FWE for other $\boldsymbol{\theta}_i$ configurations using equation (12) where $P_m(\boldsymbol{\theta})$ is

$$\begin{aligned} P_m(\boldsymbol{\theta}) &= P\{\text{reject } H'_{k-m+1}, \ldots, H'_k \text{ only} | \boldsymbol{\theta}\} \\ &= P\{(T_1, \ldots, T_{k-m}) < (c_1 - \delta', \ldots, c_{k-m} - \delta'), \\ &\quad (c_{k-m+1} - \delta', \ldots, c_{k-m+1} - \delta') \\ &\leqslant (T_{k-m+1}, \ldots, T_k) < (c_{k-m+1}, \ldots, c_k) | \boldsymbol{\theta}\} \end{aligned} \tag{15}$$

where $(T_1, \ldots, T_{k-m})$ denotes the ordered values of $T_1, \ldots, T_{k-m}$ and $(T_{k-m+1}, \ldots, T_k)$ the ordered values of $T_{k-m+1}, \ldots, T_k$ ($m = 0, 1, \ldots, k$). We use recursion formulae analogous to (13) to expand the right hand side of (15) to obtain the $(k-m)!m!$ individual probability expressions needed to evaluate P_m. These expressions enable us to evaluate the $P_m(\boldsymbol{\theta}_r)$ in (12) and determine the constants to satisfy (8) as before.

For the SU3 procedure with unequal n_i, we extend the expression for P_m in equation (15) as follows:

$$P_{(j_1, \ldots, j_m)}(\boldsymbol{\theta}) = P\{\text{reject } H'_{j_1}, \ldots, H'_{j_m} \text{ only} | \boldsymbol{\theta}\}$$
$$= P\{(T'_i;\ i \neq (j_1, \ldots, j_m)) < (c_1, \ldots, c_{k-m}),$$
$$(T_{j_1}, \ldots, T_{j_m}) < (c_{k-m+1}, \ldots, c_k) \leqslant (T'_{j_1}, \ldots, T'_{j_m}) | \boldsymbol{\theta}\}$$

for $m = 0, 1, \ldots, k$. We determine the constants recursively as before, using equation (14) and satisfying equation (8).

ACKNOWLEDGEMENTS

The first author's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We also thank the referees and editor for their very helpful comments.

REFERENCES

1. Morikawa, T. and Yoshida, M. 'A useful testing strategy in phase III trials: combined test of superiority and test of equivalence', *Journal of Biopharmaceutical Statistics*, **5**, 297–306 (1995).
2. Dunnett, C. W. and Gent, M. 'An alternative to the use of two-sided tests in clinical trials', *Statistics in Medicine*, **15**, 1729–1738 (1996).
3. Hoover, D. R. 'Simultaneous comparisons of multiple treatments to two (or more) controls', *Biometrical Journal*, **8**, 913–921 (1991).
4. Graham, N., Burrell, C., Douglas, R., Debelle, P. and Davies, L. 'Adverse effects of aspirin, acetaminophen, and ibuprofen on immune function, viral shedding, and clinical status in rhinovirus infected volunteers', *Journal of Infectious Diseases*, **162**, 1277–1282 (1990).
5. GUSTO Trial. 'An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction', *New England Journal of Medicine*, **329**, 673–682 (1993).
6. Dunnett, C. W. and Tamhane, A. C. 'Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts', *Statistics in Medicine*, **10**, 939–947 (1991).
7. Dunnett, C. W. and Tamhane, A. C. 'A step-up multiple test procedure', *Journal of the American Statistical Association*, **87**, 162–170 (1992).
8. Dunnett, C. W. and Tamhane, A. C. 'Step-up multiple testing of parameters with unequally correlated estimates', *Biometrics*, **51**, 217–227 (1995).
9. Marcus, R., Peritz, E. and Gabriel, K. R. 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika*, **63**, 655–660 (1976).
10. Hochberg, Y. and Tamhane, A. C. *Multiple Comparison Procedures*, Wiley, New York, 1987.
11. Bechhofer, R. E. and Dunnett, C. W. 'Tables of percentage points of multivariate t distributions', *Selected Tables in Mathematical Statistics*, **11**, 1–371 (1988).
12. Dunnett, C. W. and Tamhane, A. C. 'Comparisons between a new drug and active and placebo controls in an efficacy clinical trial', *Statistics in Medicine*, **11**, 1057–1063 (1992).
13. Liu, W. 'Step-down and step-up tests for comparing treatments with a control in unbalanced one-way layouts', unpublished manuscript, 1996.
14. Grechanovsky, E. and Pinsker, I. 'A general approach to stepup multiple test procedures for free-combinations families', unpublished manuscript presented at the International Conference on Multiple Comparisons, Tel Aviv, 1996.

