Biometrics 60, 491–498
June 2004

Multiple Testing to Establish Superiority/Equivalence of a New Treatment Compared with k Standard Treatments for Unbalanced Designs

Koon Shing Kwong,[1,*] Siu Hung Cheung,[2] and Wai Sum Chan[3]

[1]Department of Statistics and Applied Probability, National University of Singapore,
10 Kent Ridge Crescent, Singapore 119260
[2]Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong, China
[3]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China
[*]*email*: stakks@nus.edu.sg

Summary. In clinical studies, multiple superiority/equivalence testing procedures can be applied to classify a new treatment as superior, equivalent (same therapeutic effect), or inferior to each set of standard treatments. Previous stepwise approaches (Dunnett and Tamhane, 1997, *Statistics in Medicine* **16,** 2489–2506; Kwong, 2001, *Journal of Statistical Planning and Inference* **97,** 359–366) are only appropriate for balanced designs. Unfortunately, the construction of similar tests for unbalanced designs is far more complex, with two major difficulties: (i) the ordering of test statistics for superiority may not be the same as the ordering of test statistics for equivalence; and (ii) the correlation structure of the test statistics is not equi-correlated but product-correlated. In this article, we seek to develop a two-stage testing procedure for unbalanced designs, which are very popular in clinical experiments. This procedure is a combination of step-up and single-step testing procedures, while the familywise error rate is proved to be controlled at a designated level. Furthermore, a simulation study is conducted to compare the average powers of the proposed procedure to those of the single-step procedure. In addition, a clinical example is provided to illustrate the application of the new procedure.

Key words: Coherence property; Equivalent efficacy; Familywise error rate; Multivariate t-distribution.

1. Introduction

The simultaneous testing of the superiority and equivalence of the efficacy between a treatment and a standard treatment was studied by Morikawa and Yoshida (1995) and Dunnett and Gent (1996). The testing procedure yields one of the following possible conclusions:

(a) There is no significant evidence that the new treatment is either superior or equivalent to the standard treatment.
(b) The new treatment has efficacy which is equivalent to that of the standard treatment (the difference in efficacy is small enough to be considered as clinically insignificant).
(c) The new treatment is superior to the standard treatment.

In some medical studies, there are more than one available standard treatment. An example is the GUSTO (1993) clinical trial to which Dunnett and Tamhane (1997) referred. A new treatment was compared to two standard treatments (streptokinase with intravenous heparin and streptokinase with subcutaneous heparin) for acute myocardial infarction. Another example is the study to evaluate the combined effect of intralesional corticosteroid injection with cryotherapy versus intralesional corticosteroid or cryotherapy alone in the treatment of keloids (Yosipovitch et al., 2001).

Dunnett and Tamhane (1997) derived several step-up and step-down procedures to test the superiority and equivalence of a new treatment compared with k standard treatments. According to their findings, step-up procedures have higher power than step-down procedures when all or most hypotheses are false, which matches the findings of Dunnett and Tamhane (1992). Nevertheless, all the procedures have only small differences in terms of power except for one of the step-up procedures, called SU3 by Dunnett and Tamhane (1997). The power of SU3 is superior for some specific configurations of the parameters.

Apart from the gain in power, SU3 has another advantage. It is more sensitive to establishing equivalence (it has a higher probability of rejecting the equivalence hypotheses). An example is given in Table II of Dunnett and Tamhane (1997). The establishment of equivalence between a new treatment and the standard treatment is important. In fact, it is gaining tremendous attention from the drug industry because generic drug products manufactured by firms other than the innovator have become more popular (Chow and Liu, 2000). An establishment of equivalence implies that the new treatment can be a potential candidate as a substitute for the standard treatment.

Even though the step-up procedures are more powerful in general, Dunnett and Tamhane (1997) did not provide any analytical proof that these procedures control the familywise error rate (FWE). Kwong (2001) modified SU3 and derived a step-up procedure (denoted by KSU hereafter) which is more powerful than SU3. He proved that KSU controls the FWE.

All stepwise research work on superiority/equivalence testing has mainly been for balanced designs. With respect to unbalanced settings, Dunnett and Tamhane (1997) specified two major difficulties in the derivation of stepwise testing procedures. First, the ordering of the test statistics for superiority may be different from the ordering of the test statistics for equivalence; hence, the coherence requirement for multiple testing is not satisfied. Second, the correlation structure of the test statistics is more complex (no longer equi-correlated but product-correlated).

For various reasons, in much medical, biological, and physiological research, disparities of sample sizes are often encountered. Ittel et al. (1992) provided an example where few animals died during a balanced-design experiment and the final data set became unbalanced. In this article, we derive a method to handle multiple superiority/equivalence testing in the unbalanced situations. In Section 2, we introduce some of the previous work on step-up procedures. Section 3 presents our new procedure with the unbalanced layout. In Section 4, we discuss how to determine the critical values for the new procedure such that the FWE is controlled at a designated level. We conduct a simulation study in Section 5 to compare the powers of the new procedure with the single-step procedure. In Section 6, we provide a clinical example to illustrate the application of the proposed procedure. Finally, a conclusion is given in Section 7.

2. Step-Up Testing Procedures

2.1 *Preliminaries*

In a clinical study that is comparing a new treatment with k standard treatments, let μ_i and n_i be the unknown mean efficacy and the sample size, respectively, for the ith treatment ($i = 0, 1, \ldots, k$), where 0 denotes the new treatment. The response variable X_{ij} that corresponds to the jth experimental unit that is receiving the ith treatment ($i = 0, \ldots, n_i$) has a normal distribution with mean μ_i and variance σ^2 under the normality and homogenous error variance assumptions. Hence, mutually independent sample means $\bar{X}_i$ have distributions $N(\mu_i, \sigma^2/n_i)$ for $i = 0, \ldots, k$. In addition, let S^2 be an unbiased estimator of σ^2, independent of all sample means, and $\nu S^2/\sigma^2$ has a χ^2 distribution with ν degrees of freedom.

Denote by $\theta_i = \mu_0 - \mu_i$ ($i = 1, \ldots, k$) the efficacy difference between the new treatment and the ith standard treatment. Assume that a larger value of θ_i means that the new treatment has a better efficacy. The objective of this multiple hypothesis testing is to determine whether there is significant evidence that the new treatment is superior or equivalent to each of the k standard treatments.

For $i = 1, \ldots, k$, the families of hypotheses for establishing superiority and equivalence are

$$\mathrm{H}_i : \theta_i \le 0 \quad \text{versus} \quad \theta_i > 0$$

and

$$\mathrm{H}'_i : \theta_i \le -\delta \quad \text{versus} \quad \theta_i > -\delta,$$

respectively. The quantity $\delta > 0$ is the maximum difference in efficacy that is considered to be clinically insignificant.

Then, for the hypotheses of H_i and H'_i, the pivotal statistics are

$$T_i = \frac{\bar{X}_0 - \bar{X}_i}{S\tau_i},$$

and

$$T'_i = T_i + \Delta_i,$$

respectively, for $i = 1, \ldots, k$, where $\tau_i = (1/n_i + 1/n_0)^{1/2}$ and $\Delta_i = \delta/S\tau_i$. Furthermore, each set of $(T_1, \ldots, T_k)$ and $(T'_1, \ldots, T'_k)$ has a joint k-variate t-distribution with ν degrees of freedom and a $k \times k$ correlation matrix $\{\rho_{ij}^{(k)} = \rho_i\rho_j\}$ that has the (i, j) element equal to 1 for $i = j$ and $\rho_i\rho_j$ for $i \ne j$, where $\rho_i = \{n_i/(n_i + n_0)\}^{1/2}$. To take account of the multiplicity effect in multiple hypothesis testing by controling FWE, the stepwise procedures are established to satisfy the following condition:

$$\mathrm{FWE} = P\{\text{reject any true } \mathrm{H}_i \text{ or } \mathrm{H}'_i\} \le \alpha \tag{1}$$

under any configuration of the parameters θ_i.

2.2 *Dunnett and Tamhane's Step-Up Procedure (SU3)*

Assume that t_i and t'_i are the observed test statistics of the corresponding random variables T_i and T'_i, respectively, for $i = 1, \ldots, k$. Denote the observed ordered test statistics as $t_{(1)} \le \cdots \le t_{(k)}$, and $t'_{(1)} \le \cdots \le t'_{(k)}$ with corresponding hypotheses $\mathrm{H}_{(1)}, \ldots, \mathrm{H}_{(k)}$ and $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k)}$, respectively. In balanced designs, $\Delta_1 = \cdots = \Delta_k = \Delta$ implies that the orderings of $(t_1, \ldots, t_k)$ and $(t'_1, \ldots, t'_k)$ are the same, i.e., $t'_{(i)} = t_{(i)} + \Delta$ for $i = 1, \ldots, k$. Because $\delta > 0$ and the difference of observed test statistics $t_{(i)}$ and $t'_{(i)}$ is a constant for $i = 1, \ldots, k$, the nonrejection of $\mathrm{H}'_{(i)}$ implies the nonrejection of $\mathrm{H}_{(i)}$. Therefore, for each pair of hypotheses $\{\mathrm{H}'_{(i)}, \mathrm{H}_{(i)}\}$ with respect to the comparison of the new treatment to the (i)th standard treatment, there are only three possible test conclusions:

(a) Nonrejection of $\mathrm{H}'_{(i)}$—There is no evidence that the new treatment is either superior or equivalent to the (i)th standard treatment.

(b) Rejection of $\mathrm{H}'_{(i)}$ and nonrejection of $\mathrm{H}_{(i)}$—The new treatment has equivalent efficacy as the (i)th standard treatment.

(c) Rejection of $\mathrm{H}'_{(i)}$ and $\mathrm{H}_{(i)}$—The new treatment is superior to the (i)th standard treatment.

As stated in Section 1, SU3 is one of three step-up procedures that are designed to establish the superiority and equivalence of a new treatment compared to k standard treatments for balanced designs. The testing of equivalence hypotheses and superiority hypotheses are conducted in two stages each with a single set of critical values $\{d_1, d_2, \ldots, d_k\}$ that is determined to satisfy (1) under a conjecture. The computational details of the constants are given by Dunnett and Tamhane (1997). The two stages of conducting SU3 are as follows:

Stage 1: Testing of equivalence hypotheses

Procedure: Test the equivalence hypotheses $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k)}$ sequentially, starting with $\mathrm{H}'_{(1)}$. If $t'_{(1)} \le d_1$, then $\mathrm{H}'_{(1)}$ is accepted and $\mathrm{H}'_{(2)}$ is tested. The testing procedure continues until the first occurrence of $t'_{(i)} > d_i$, say $i = m_1 \le k$. Then, $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(m_1-1)}$ are accepted while the remaining hypotheses are rejected. If $t'_{(i)} \le d_i$, for $i = 1, \ldots, k$, then m_1 does not exist and all the hypotheses $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k)}$ are accepted.

Decision: If m_1 does not exist, then conclude that the new treatment is inferior to all standard treatments and terminate the testing procedure. Otherwise, conclude that the new treatment is inferior to those standard treatments that correspond to hypotheses $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(m_1-1)}$ and then proceed to Stage 2.

Stage 2: Testing of superiority hypotheses

Procedure: Test the superiority hypotheses $\mathrm{H}_{(m_1)}, \ldots, \mathrm{H}_{(k)}$ sequentially, starting with $\mathrm{H}_{(m_1)}$. If $t_{(m_1)} \le d_{m_1}$, then $\mathrm{H}_{(m_1)}$ is accepted and $\mathrm{H}_{(m_1+1)}$ is tested. The testing procedure continues until the first occurrence of $t_{(i)} > d_i$, say $i = m_1 + m_2 \le k$. Then, $\mathrm{H}_{(m_1)}, \ldots, \mathrm{H}_{(m_1+m_2-1)}$ are accepted while the remaining hypotheses are rejected. If $t_{(i)} \le d_i$ for $i = m_1, \ldots, k$, then m_2 does not exist and all the hypotheses $\mathrm{H}_{(m_1)}, \ldots, \mathrm{H}_{(k)}$ are accepted.

Decision: If m_2 does not exist, conclude that the new treatment is equivalent to the standard treatments that correspond to hypotheses $\mathrm{H}_{(m_1)}, \ldots, \mathrm{H}_{(k)}$. If m_2 exists and is greater than 0, then conclude that the new treatment is equivalent to the standard treatments corresponding to hypotheses $\mathrm{H}_{(m_1)}, \ldots, \mathrm{H}_{(m_1+m_2-1)}$ and is superior to the standard treatments that correspond to hypotheses $\mathrm{H}_{(m_1+m_2)}, \ldots, \mathrm{H}_{(k)}$. In case of $m_2 = 0$, simply conclude that the new treatment is superior to the standard treatments that correspond to hypotheses $\mathrm{H}_{(m_1)}, \ldots, \mathrm{H}_{(k)}$.

2.3 *Kwong's Step-Up Procedure (KSU)*

There is a lack of complete analytical proof that SU3 controls FWE, even though simulation studies have indicated that it does (Dunnett and Tamhane, 1997). Recently, Kwong (2001) modified SU3 and derived a step-up procedure (KSU). He proved analytically that KSU controls FWE at α. Furthermore, simulation studies indicated that KSU is more powerful in general than SU3.

The major difference between SU3 and KSU is that SU3 uses only one set of critical values in both stages and KSU uses a different set of critical values in each of the two stages, $c_1^0 \le c_2^0 \ldots \le c_k^0$ for equivalence tests in Stage 1 and $c_1^{m_1} \le c_2^{m_1} \le \cdots \le c_{k-m_1+1}^{m_1}$ for superiority tests in Stage 2. The method to evaluate these constants can be found in Kwong (2001).

2.4 *Coherence Property of SU3 and KSU*

The parameter values that are postulated by H'_i: $\theta_i \le -\delta$ is a subset of the parameter values postulated by H_i: $\theta_i \le 0$. According to Hochberg and Tamhane (1987), H'_i is said to imply H_i, and the hierarchical relationship between these two hypotheses requires the satisfaction of the coherence property (Gabriel, 1969) in multiple testing. That is, if H'_i is not rejected, then H_i is also not rejected.

It is crucial to note that both SU3 and KSU are two-stage step-up procedures designed for $n_1 = \cdots = n_k$. Because the orderings of $(t_1, \ldots, t_k)$ and $(t'_1, \ldots, t'_k)$ are the same for balanced designs, the decision not to reject $\mathrm{H}'_{(i)}$ in Stage 1 automatically leads to the nonrejection of $\mathrm{H}_{(i)}$ in Stage 2. However, for unbalanced designs, this coherence property does not exist in SU3 and KSU because the orderings of $(t_1, \ldots, t_k)$ and $(t'_1, \ldots, t'_k)$ may not be the same. The lack of the coherence property poses a major obstacle to the generalization of SU3 and KSU to unbalanced designs.

3. Proposed Step-Up Testing Procedure for Unbalanced Designs

For various reasons, such as the availability of patients and fatality of subjects during experiments, sample size disparities are often encountered in clinical studies. In clinical experiments, unbalanced designs are much more popular. Therefore, statistical methods to compare treatments for unbalanced designs are extremely useful. As indicated in Section 2, the generalization of SU3 and KSU to unbalanced designs is intricate because the coherence property may not be preserved in unequal sample size cases. To overcome this obscurity, we derive a new procedure to establish superiority and equivalence tests for unbalanced designs.

Deviating from the SU3 and KSU, the new two-stage testing procedure, denoted by KCC hereafter, consists of a step-up procedure for establishing equivalence in Stage 1 and a *single-step* procedure for establishing superiority in Stage 2. Therefore, KCC requires a set of critical values to sequentially compare to $(t'_{(1)}, \ldots, t'_{(k)})$ in Stage 1, and a single critical value that depends on the result of Stage 1 to compare to each of $(t_1, \ldots, t_k)$ in Stage 2. Because Stage 2 does not involve the ordering of $(t_1, \ldots, t_k)$, the ordering problem of test statistics does not exist in KCC.

Before testing the hypotheses, KCC requires the determination of two sets of critical values: the first set $(c_1, \ldots, c_k)$ for Stage 1 and the second set $(u_1, \ldots, u_k)$ for Stage 2. Note that only one of u_i will be used in Stage 2. To preserve the coherence property in KCC, the two sets of the critical values must at least satisfy the following conditions:

C1: $c_1 \le c_2 \le \cdots \le c_k$,
C2: $u_1 \le u_2 \le \cdots \le u_k$,
C3: $c_i \le u_i$ for $i = 1, \ldots, k$.

The first monotonic nondecreasing condition, C1, is the usual requirement for any step-up procedure. The C2 and C3 conditions imply that the nonrejection of any equivalence hypothesis in Stage 1 guarantees the nonrejection of the corresponding superiority hypothesis in Stage 2. The derivation and evaluation of these critical values will be outlined in the next section. The procedures of KCC are as follows:

Stage 1: Testing of equivalence hypotheses

Procedure:

(a) Test the equivalence hypotheses $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k)}$ sequentially, starting with $\mathrm{H}'_{(1)}$. If $t'_{(1)} \le c_1$, then $\mathrm{H}'_{(1)}$ is accepted and $\mathrm{H}'_{(2)}$ is tested.

(b) The testing procedure continues until the first occurrence of $t'_{(i)} > c_i$, say $i = m \le k$. Then, $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(m-1)}$ are accepted while the remaining hypotheses are rejected. If $t'_{(i)} \leq c_i$, for $i = 1, \ldots, k$, then m does not exist and all of the hypotheses $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k)}$ are accepted.

Decision:

(a) If m does not exist, then conclude that the new treatment is inferior to all standard treatments and terminate the testing procedure.

(b) If m exists, conclude that the new treatment is inferior to those standard treatments that correspond to hypotheses $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(m-1)}$ and then proceed to Stage 2.

Stage 2: Testing of superiority hypotheses

Procedure:

(a) Let $\{\mathrm{H}'_{l_1}, \mathrm{H}'_{l_2}, \ldots, \mathrm{H}'_{l_k}\} = \{\mathrm{H}'_{(1)}, \mathrm{H}'_{(2)}, \ldots, \mathrm{H}'_{(k)}\}$, respectively. For any given hypothesis H'_{l_i} tested for equivalence, let H_{l_i} be the corresponding hypothesis tested for superiority, i.e., H'_{l_i} and H_{l_i} share the same treatment comparison for $i = 1, \ldots, k$. Due to the coherence property of KCC, we only need to test the superiority hypotheses $\mathrm{H}_{l_m}, \ldots, \mathrm{H}_{l_k}$ with the corresponding test statistics $(t_{l_m}, \ldots, t_{l_k})$.

(b) Compare each of $(t_{l_m}, \ldots, t_{l_k})$ with the critical value u_m. For $j = m, \ldots, k$, if $t_{l_j} > u_m$, then reject H_{l_j}; otherwise, accept H_{l_j}.

Decision:

(a) If none of the superiority hypotheses is rejected, then conclude that the new treatment is equivalent to the standard treatments that correspond to hypotheses $\mathrm{H}_{l_m}, \ldots, \mathrm{H}_{l_k}$.

(b) If some superiority hypotheses are rejected and the others are accepted, then conclude that the new treatment is equivalent to the standard treatments that correspond to the accepted hypotheses and is superior to the standard treatments that correspond to the rejected hypotheses.

(c) If all hypotheses are rejected, then simply conclude that the new treatment is superior to the standard treatments corresponding to hypotheses $\mathrm{H}_{l_m}, \ldots, \mathrm{H}_{l_k}$.

4. Determination of Critical Values for KCC

4.1 *FWE Requirement*

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and $\boldsymbol{\theta}_{r,s} \subset \boldsymbol{\theta}$ be a set of vectors that have elements $\theta_j \leq -\delta$ (H'_j is true) for $j = 1, \ldots, r$, $\theta_j \leq 0$ (H_j is true) for $j = r + 1, \ldots, r + s$, and $\theta_j > 0$ (H'_j and H_j are false) for $j = r + s + 1, \ldots, k$. To control the FWE at α, we must determine the critical values of the step-up procedure in Stage 1 of KCC such that

$$P_{\boldsymbol{\theta}_{r,0}}\left[\text{Accept } \mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(r)}\right] \geq 1 - \alpha, \tag{2}$$

for $r = 1, \ldots, k$, and the critical values of single-step procedure in Stage 2 of KCC such that

$$P_{\boldsymbol{\theta}_{k-s,s}}\left[\text{Accept } \mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k-s)}, \mathrm{H}_1, \ldots, \mathrm{H}_k\right] \geq 1 - \alpha, \tag{3}$$

for $s = 1, \ldots, k$, where the associated test statistics under the null hypotheses in (2) and (3) have the parameter spaces $\boldsymbol{\theta}_{r,0}$ and $\boldsymbol{\theta}_{k-s,s}$, respectively. Note that there are k inequalities in (2) and (3), and each inequality is supposed to derive one critical value for KCC.

4.2 *Identification of the Least Favorable Configurations of Parameters*

Let the left-hand side of Inequality (2) be Q. For any given r, Q is influenced by

(a) the values of parameters in $\boldsymbol{\theta}_{r,0}$, and

(b) the correlation structures among the test statistics that correspond to those accepted true null hypotheses (Liu, 1997).

To obtain the smallest possible critical values c_j for Stage 1 of KCC, we have to determine minimum value of Q over all possible configurations in (a) and (b). Based on the arguments given by Liu (1997) for establishing a step-up procedure for unbalanced designs, the minimum value of Q is

$$\min{}^* P_{\boldsymbol{\theta}^*_{r,0}}\left[\text{Accept } \mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(r)}\right], \tag{4}$$

where min* denotes the minimum over all possible correlation structures among the test statistics corresponding to those accepted true null hypotheses, and $\boldsymbol{\theta}^*_{r,s} \in \boldsymbol{\theta}_{r,s}$ denotes a parameter vector with elements $\theta_j = -\delta$ for $j = 1, \ldots, r$, $\theta_j = 0$ for $j = r + 1, \ldots, r + s$, and $\theta_j \to \infty$ for $j = r + s + 1, \ldots, k$. In other words, the operator min* minimizes Q over all the possible product correlation structures among the test statistics under $\boldsymbol{\theta}^*_{r,0}$. By setting (4) to $1 - \alpha$, the minimum possible critical values $c_1 \leq c_2 \leq \cdots \leq c_k$ used in Stage 1 can be determined.

Similarly, the minimum value of the left hand of Inequality (3) is

$$\min{}^* P_{\boldsymbol{\theta}^*_{k-s,s}}\left[\text{Accept } \mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(k-s)}, \mathrm{H}_1, \ldots, \mathrm{H}_k\right] \tag{5}$$

and the minimum possible critical values $u_1 \leq u_2, \leq \cdots \leq u_k$ used in Stage 2 can be evaluated by equating (5) to $1 - \alpha$.

4.3 *Critical Values in Stage 1*

Denote $(a_1, \ldots, a_k) \leq (b_1, \ldots, b_k)$ in the event that $a_{(j)} \leq b_{(j)}$ for $j = 1, \ldots, k$, where $a_{(1)} \leq \cdots \leq a_{(k)}$ and $b_{(1)} \leq \cdots \leq b_{(k)}$ are the ordered values of a_j's and b_j's, respectively. Following the arguments of (4), assume that $\mathrm{H}'_{(1)}, \ldots, \mathrm{H}'_{(r)}$ are true and all other hypotheses are false, in order to determine the critical values of Stage 1, c_r for $r = 1, \ldots, k$. Therefore, by setting the probability in (4) to $1 - \alpha$, we obtain the following k equations

$$\min_{1 \leq l_1 < \cdots < l_r \leq k} P_{\boldsymbol{\theta}^*_{r,0}}\left[\left(T'_{l_1}, \ldots, T'_{l_r}\right) < (c_1, \ldots, c_r); \left\{\rho^{(r)}_{ij} = \rho_{l_i}\rho_{l_j}\right\}\right] = 1 - \alpha, \tag{6}$$

for $r = 1, \ldots, k$. The minimum is taken over all the subsets $\{l_1, \ldots, l_r\} \subset \{1, \ldots, k\}$ with cardinality r. The set of random variables $(T'_{l_1}, \ldots, T'_{l_r})$ for $r = 2, \ldots, k$ has standardized multivariate t-distribution with correlation matrix $\{\rho^{(r)}_{ij} = \rho_{l_i}\rho_{l_j}\}$. For $r = 1$, (6) produces c_1, which is the upper α-percentage point of the t-distribution with ν degrees of freedom. Afterwards, we can solve recursively for c_r ($r = 2, \ldots, k$) based on the previously determined values $c_1, \ldots, c_{r-1}$.

From the above arguments, it is trivial that $c_1, \ldots, c_k$ do not depend on δ and are equivalent to the critical values of the conventional one-sided step-up procedure that was discussed by Dunnett and Tamhane (1995), Liu (1997), and Kwong and Liu (2000) for unbalanced designs. Therefore, one can adopt the techniques that were discussed by Kwong and Liu (2000) to efficiently evaluate the exact or approximate critical values that are used in Stage 1 of KCC.

4.4 *Critical Values in Stage 2*

To determine the critical values $u_1, \dots, u_k$ for Stage 2 of KCC, we follow arguments that are similar to those given in Section 4.3. Assume that $\mathrm{H}'_{(1)}, \dots, \mathrm{H}'_{(s-1)}, \mathrm{H}_1, \dots, \mathrm{H}_k$ are true. By setting the probability in (5) to $1 - \alpha$, we solve the following k equations for the critical values u_s:

$$\min{}^* P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \left[\text{Accept } \mathrm{H}'_{(1)}, \dots, \mathrm{H}'_{(s-1)}, \mathrm{H}_1, \dots, \mathrm{H}_k\right] = 1 - \alpha, \tag{7}$$

for $s = 1, \dots, k$. The critical values are determined recursively, starting from u_k. Then, use u_k to determine u_{k-1} until u_1 is obtained.

Consider the general case to evaluate u_s based on the previously determined values $c_1, \dots, c_k$ and $u_{s+1}, \dots, u_k$. Define the subsets $\mathcal{S} = \{l_1, \dots, l_{s-1}\} \subset \{1, \dots, k\}$ with cardinality $s - 1$. Note that there are $q_1 = \binom{k}{s-1}$ distinct subsets, say $S_1, \dots, S_{q_1}$ in $\mathcal{S}$. For each given subset S_j for $j = 1, \dots, q_1$, let the complement of S_j be $S_j^c = \{l_s, \dots, l_k\}$. Define the subsets $I(j, h) = \{l_{g_1}, \dots, l_{g_h}) \subset S_j^c$ with cardinality h and the complement of $I(j, h)$ be $\{l_{g_h+1}, \dots, l_{g_{k-s+1}}\}$. Note that there are $q_2 = \binom{k-s+1}{h}$ distinct subsets of $I(j, h)$, say $I_1, \dots, I_{q_2}$.

Due to the coherence property of KCC stated in Section 3, the nonrejection of any equivalence hypothesis implies the nonrejection of corresponding superiority hypothesis. As a result, (5) is equivalent to

$$\begin{aligned}
&\min{}^* P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \left[\text{Accept } \mathrm{H}'_{(1)}, \dots, \mathrm{H}'_{(s-1)}, \mathrm{H}_1, \dots, \mathrm{H}_k\right] \\
&\quad = \min{}^* \left\{ P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \left[\text{Accept } \mathrm{H}'_{(1)}, \dots, \mathrm{H}'_{(k)}\right] \right. \\
&\qquad\qquad + \sum_{h=0}^{k-s} P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \left[\text{Accept } \mathrm{H}'_{(1)}, \dots, \mathrm{H}'_{(s-1+h)}, \right. \\
&\qquad\qquad\qquad\qquad \left.\left. \mathrm{H}_1, \dots, \mathrm{H}_k \text{ and Reject } H'_{(s+h)}\right] \right\} \\
&\quad = \min_{S_j \in \mathcal{S}} \left\{ \sum_{h=0}^{k-s+1} \sum_{I_p \in I(j,h)} P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \right. \\
&\qquad\qquad \left[\left(T'_{l_1}, \dots, T'_{l_{s-1}}, T'_{l_{g_1}}, \dots, T'_{l_{g_h}}\right)\right. \\
&\qquad\qquad \le (c_1, \dots, c_{s-1+h}), \bigcap_{i=h+1}^{k-s+1} c_{s+h} - \Delta_{l_{g_i}} < T_{l_{g_i}} \\
&\qquad\qquad \left.\left. \le u_{s+h}; \left\{\rho_{ij}^{(k)} = \rho_{l_i}\rho_{l_j}\right\}\right] \right\}
\end{aligned}$$

$$\begin{aligned}
&\quad = \min_{S_j \in \mathcal{S}} \left\{ \sum_{h=0}^{k-s+1} \sum_{I_p \in I(j,h)} P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \right. \\
&\qquad\qquad \left[\left(T'_{l_1}, \dots, T'_{l_{s-1}}, T'_{l_{g_1}}, \dots, T'_{l_{g_h}}\right)\right. \\
&\qquad\qquad \le (c_1, \dots, c_{s-1+h}), \bigcap_{i=h+1}^{k-s+1} c_{s+h} < T'_{l_{g_i}} \\
&\qquad\qquad \left.\left. \le u_{s+h} + \Delta_{l_{g_i}}; \left\{\rho_{ij}^{(k)} = \rho_{l_i}\rho_{l_j}\right\}\right] \right\},
\end{aligned} \tag{8}$$

for $j = 1, \dots, q_1$ and $p = 1, \dots, q_2$, where $(T'_{l_1}, \dots, T'_{l_{s-1}})$ and $(T'_{l_{g_1}}, \dots, T'_{l_{g_{k-s+1}}})$ have parameters θ_i equal to $(-\delta, \dots, -\delta)$ and $(0, \dots, 0)$, respectively.

By applying arguments that are similar to those given by Dunnett and Tamhane (1995), we transform $(T'_{l_1}, \dots, T'_{l_{s-1}}, T'_{l_{g_1}}, \dots, T'_{l_{g_{k-s+1}}})$ to

$$T'_{l_i} = \frac{\sqrt{1 - \rho_{l_i}^2} Z_{l_i} - \rho_{l_i} Z_0}{U} \qquad \text{for} \quad i = 1, \dots, s - 1.$$

$$T'_{l_{g_i}} = \frac{\sqrt{1 - \rho_{l_{g_i}}^2} Z_{l_{g_i}} - \rho_{l_{g_i}} Z_0}{U} + \Delta_{l_{g_i}} \qquad \text{for} \quad i = 1, \dots, k - s + 1,$$

where Z_0, $Z_{l_1} \dots, Z_{l_{s-1}}$, $Z_{l_{g_1}}, \dots, Z_{l_{g_{k-s+1}}}$ are mutually independent standard normal random variables with probability density function ϕ, and U, which is independent of all Z_i, is the random variable $(\chi_\nu^2/\nu)^{1/2}$ with probability density function f_ν. After conditioning on the two random variables $U = u$ and $Z_0 = z_0$, and interchanging the order of summation and integration, (8) is reduced to

$$\begin{aligned}
&\min{}^* P_{\boldsymbol{\theta}^*_{s-1,k-s+1}} \left\{\text{Accept } \mathrm{H}'_{(1)}, \dots, \mathrm{H}'_{(s-1)}, \mathrm{H}_1, \dots, \mathrm{H}_k\right\} \\
&\quad = \min_{S_j \in \mathcal{S}} \int_0^\infty \int_{-\infty}^\infty \\
&\qquad \left\{ \sum_{h=0}^{k-s+1} \sum_{I_p \in I(j,h)} P\left[\left(W_{l_1}, \dots, W_{l_{s-1}}, W_{l_{g_1}}, \dots, W_{l_{g_h}}\right)\right.\right. \\
&\qquad\qquad\qquad \left. \le (c_1, \dots, c_{s-1+h})\right] \\
&\qquad \left. \times \prod_{i=h+1}^{k-s+1} P\left[c_{s+h} < W_{l_{g_i}} \le u_{s+h} + \Delta_{l_{g_i}}\right] \right\} \\
&\qquad \times \phi(z_0)\, dz_0 f_\nu(u)\, du,
\end{aligned} \tag{9}$$

for $j = 1, \dots, q_1$ and $p = 1, \dots, q_2$, where $(W_{l_1}, \dots, W_{l_{s-1}}, W_{l_{g_1}}, \dots, W_{l_{g_{k-s+1}}})$ are mutually independent random variables with distributions:

$$W_{l_i} \sim N\left(\frac{-\rho_{l_i} z_0}{u}, \frac{1 - \rho_{l_i}^2}{u^2}\right) \qquad \text{for } i = 1, \dots, s - 1.$$

$$W_{l_{g_i}} \sim N\left(\frac{-\rho_{l_{g_i}} z_0}{u} + \Delta_{l_{g_i}}, \frac{1 - \rho_{l_{g_i}}^2}{u^2}\right) \quad \text{for } i = 1, \dots, k - s + 1.$$

Based on (9), the critical values u_s $(s = 1, \dots, k)$ are determined for Stage 2 of KCC, and we show that KCC controls the FWE at α for unbalanced designs.

Table 1
Critical values of KCC for $\alpha = 0.05$

		New treatment	Standard treatments					
			i					
	δ	$i = 0$	1	2	3	4	5	6
n_i		20	10	15	20			
c_i			1.670	1.982	2.129			
u_i	0.5		1.673	2.030	2.174			
	1.0		1.762	2.117	2.242			
	1.5		1.839	2.228	2.336			
	2.0		1.912	2.363	2.460			
n_i		30	10	21	23	25		
c_i			1.660	1.971	2.120	2.220		
u_i	0.5		1.660	2.027	2.159	2.250		
	1.0		1.674	2.111	2.220	2.297		
	1.5		1.819	2.209	2.299	2.365		
	2.0		1.914	2.324	2.401	2.457		
n_i		20	10	12	21	25	30	
c_i			1.659	1.967	2.111	2.205	2.271	
u_i	0.5		1.659	1.967	2.152	2.240	2.308	
	1.0		1.848	2.060	2.242	2.312	2.371	
	1.5		1.942	2.230	2.352	2.412	2.465	
	2.0		2.034	2.384	2.488	2.543	2.594	
n_i		24	10	12	15	18	23	30
c_i			1.657	1.967	2.119	2.220	2.294	2.348
u_i	0.5		1.657	1.967	2.119	2.220	2.311	2.372
	1.0		1.849	2.073	2.208	2.277	2.356	2.415
	1.5		1.942	2.215	2.296	2.358	2.422	2.483
	2.0		2.030	2.336	2.401	2.457	2.513	2.581

δ is in the scale of $s/\sqrt{n_0}$.

4.5 *Numerical Study*

Applying the approach discussed by Kwong and Liu (2000), a FORTRAN program to evaluate all the critical values of KCC for any given α, n_0, $n_1, \ldots, n_k$ and δ is available from the first author. From various extensive numerical studies, we find that it is always possible to evaluate the critical values that satisfy conditions C1 and C2 for any given configurations of parameters. Occasionally, for small δ and/or large disparity among n_i, there may be cases where $u_i < c_i$. To satisfy condition C3, the value of u_i is adjusted slightly upward to c_i. Numerical examples of critical values used in KCC are presented in Table 1, where δ is in the scale of $s/(n_0)^{1/2}$.

5. Simulation Study of Power

Currently, the only proved statistical method to establish the superiority and equivalence of a new treatment compared to k standard treatments for unbalanced designs is the single-step procedure (SS) proposed by Dunnett and Tamhane (1997). Hence, we conducted a simulation study to compare the powers of KCC to SS for the case where $\alpha = 0.05$, $(n_0, \ldots, n_6) = (24, 10, 12, 15, 18, 23, 30)$, $s/(n_0)^{1/2} = 1$, $\delta = 1$. We observed similar results for some other cases, but they are not reported here. The number of repetitions in each configuration of $\boldsymbol{\theta}$ is 1,000,000, so that the standard error of each simulated power is less than 0.0005. The critical values for KCC are stated in Table 1 and the critical value for SS is 2.347 in this case. The simulated powers of KCC and SS are presented in Table 2, where the average power P_1 is defined as the proportion of rejecting equivalence hypotheses H'_j, given that $\theta_j > -\delta$, and the average power P_2 is defined as the proportion of rejecting superiority hypotheses H_j, given that $\theta_j > 0$. Table 2 reveals that KCC is more powerful in terms of P_1 and P_2 than SS in general. Furthermore, when most of the equivalence or superiority hypotheses are false, the superiority of the KCC procedure is drastic, where the gain in power is well above 10% in most of the configurations. Therefore, we conclude that KCC not only controls the FWE, but is also more powerful than its counterpart SS in unequal sample size cases, especially when most of the equivalence or superiority hypotheses are false.

6. Example

Malmstrom et al. (2002) recently conducted a single-center, randomized, double-blind, single-dose study of certain analgesic drugs that treated pain after dental surgery. The objective of the study was to reconfirm the analgesic efficacy of a relatively new treatment, rofecoxib 50 mg (r50) by comparing it to four standard treatments: celecoxib 400 mg (c400), celecoxib 200 mg (c200), ibuprofen 400 mg (ib400), and placebo (plac). After the surgical extraction of at least 2 third molars, patients with moderate or severe pain were randomly given a single oral dose of one of the five treatments. A 0–4 scale for pain relief (0 = poor, 1 = fair, 2 = good, 3 = very good,

Table 2
Simulated average powers for the case $\alpha = 0.05, (n_0, \ldots, n_6) = (24, 10, 12, 15, 18, 23, 30), s/\sqrt{n_0} = 1, \delta = 1$

$\boldsymbol{\theta}$						P_1		P_2	
θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	KCC	SS	KCC	SS
−1	−1	−1	−1	−1	2	0.456	0.456	0.190	0.196
−1	−1	−1	−1	−1	4	0.915	0.915	0.724	0.736
−1	−1	−1	2	2	2	0.426	0.403	0.183	0.173
−1	−1	−1	4	4	4	0.889	0.870	0.695	0.673
−1	2	2	2	2	2	0.415	0.358	0.190	0.154
−1	4	4	4	4	4	0.878	0.818	0.690	0.611
2	2	2	2	2	2	0.426	0.338	0.206	0.146
4	4	4	4	4	4	0.892	0.789	0.719	0.581

P_1 is defined as the proportion of rejecting equivalence hypotheses $\mathrm{H'}_j$, given that $\theta_j > -\delta$.
P_2 is defined as the proportion of rejecting superiority hypotheses H_j, given that $\theta_j > 0$.

Table 3

Critical values of KCC with $\delta = 0.815$ and $\alpha = 0.05$ for testing the superiority/equivalence of the dental data

	New treatment $i = 0$	Standard treatments i 1	2	3	4
n_i	150	45	151	90	45
c_i		1.648	1.958	2.110	2.208
u_i		1.718	1.958	2.166	2.284

and 4 = excellent) recorded the pain intensity that was experienced by patients throughout the 24 hours after dosing. The primary end point was the total pain relief score over the first 8 hours.

After applying the well-known Bartlett's test, we have no reason to doubt the equality of variances among the five treatments. The estimate of the common variance is $s^2 = 99.584$. Because the clinically insignificant difference δ is not given in the study and KCC and SS are only appropriate for a prespecified value of δ, we set δ to be 1 $s/(n_0)^{1/2}$, or 0.815 for illustrative purposes. We also set the overall significance level at 0.05. Table 3 provides the critical values that are necessary to implement the KCC procedure. For the SS procedure, the critical value is 2.205. Table 4 summarizes the conclusions. With the KCC procedure, the new treatment r50 is superior to c400, c200, and plac, and equivalent to ib400. If the SS procedure is used, then the new treatment is found to be superior to c200 and plac, and equivalent to c400 only. The findings demonstrate that the KCC procedure is more powerful than the SS procedure.

7. Conclusion

We have proposed a two-stage testing procedure to establish superiority/equivalence of a new treatment compared with k standard treatments for unbalanced designs. The new procedure is a combination of the step-up method for testing equivalence and the single-step method for testing superiority. A theoretical justification is provided to show that the new procedure controls the familywise error rate at a designated level.

Existing stepwise procedures to establish superiority/equivalence are more powerful than the single-step procedure, but they are restricted to balanced designs. The major contribution of this article is to extend the stepwise procedures of Dunnett and Tamhane (1997) to unbalanced designs. Our proposed method should be useful in many practical circumstances. In addition, with the help of the algorithm stated in this article, practitioners will find the evaluation of the critical values for the new procedure straightforward and feasible. Furthermore, our simulation study reveals that the new procedure is also more powerful than the single-step procedure in general, especially when most of the hypotheses are false.

Table 4

Dental data for establishing superiority/equivalence with $\delta = 0.815$ and $\alpha = 0.05$ under the KCC and SS procedures

i	Contrast	$t'_{(i)}$	t_i
1	r50 vs. ib400	1.954[a]	1.474
2	r50 vs. c400	2.620[a,c]	1.912[b]
3	r50 vs. c200	4.896[a,c]	4.284[b,d]
4	r50 vs. plac	10.218[a,c]	9.728[b,d]

[a] Significance for the equivalence tests under KCC.

[b] Significance for the superiority tests under KCC.

[c] Significance for the equivalence tests under SS.

[d] Significance for the superiority tests under SS.

Acknowledgements

The authors are thankful to the associate editor for several helpful comments and suggestions. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK4042/02P).

Résumé

Dans les études cliniques, des procédures de tests multiples de supériorité/équivalence peuvent être utilisées pour classer un nouveau traitement en supérieur, équivalent(même effet thérapeutique) ou inférieur à chacun des traitements de référence.De précédentes approches pas-à-pas (Dunnett et Tamhane,1997, *Statistics in Medicine* **16**, 2489–506; Kwong, 2001, *Journal of Statistical Planning and Inference* **97**, 359–366) sont appropriées seulement pour des schémas équilibrés. Malheureusement, la construction de tests similaires pour des schémas déséquilibrés est bien plus complexe, avec deux difficultés majeures: (i) l'ordre des statistiques de test pour la supériorité peut ne pas être la même que l'ordre des statistiques de test pour l'équivalence; et (ii) la structure de corrélation des statistiques de test n'est plus équicorrélée.Dans ce papier,nous cherchons à développer une procédure de test à deux étapes pour schémas déséquilibrés,procédures très populaires dans le domaine des essais cliniques. Cette procédure est la combinaison d'une procédure ascendante pas-à-pas et d'une procédure à une étape,tandis qu'il est prouvé que le taux d'erreur globale est contrôlé au niveau voulu.En outre,une étude de simulation est conduite afin de comparer la puissance moyenne de cette procédure avec celle de la procédure à une étape.Enfin,un exemple clinique est fourni pour illuster l'application de cette nouvelle procédure.

References

Chow, S. C. and Liu J. P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies.* New York: Marcel Dekker.

Dunnett, C. W. and Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine* **15,** 1729–1738.

Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association* **87,** 162–170.

Dunnett, C. W. and Tamhane, A. C. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics* **51,** 217–227.

Dunnett, C. W. and Tamhane, A. C. (1997). Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments. *Statistics in Medicine* **16,** 2489–2506.

Gabriel, K. R. (1969). Simultaneous test procedures—Some theory of multiple comparisons. *Annals of Mathematical Statistics* **40,** 224–250.

GUSTO Trial. (1993). An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine* **329,** 673–682.

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

Ittel, T. H., Gruber, E., Heinrichs, A., Handt, S., Hofstädter, F., and Sieberth, H. (1992). Effect of fluoride on aluminum-induced bone disease in rats with renal failure. *Kidney International* **41,** 1340–1348.

Kwong, K. S. (2001). A modified Dunnett and Tamhane step-up approach for establishing superiority/equivalence of a new treatment compared with k standard treatments. *Journal of Statistical Planning and Inference* **97,** 359–366.

Kwong, K. S. and Liu, W. (2000). Calculation of critical values for Dunnett and Tamhane's step-up multiple test procedure. *Statistics and Probability Letters* **49,** 411–416.

Liu, W. (1997). Some results on step-up tests for comparing treatments with a control in unbalanced one-way layouts. *Biometrics* **53,** 1508–1512.

Malmstrom, K., Fricke, J. R., Kotey, P., Kress, B., and Morrison, B. (2002). A comparison of rofecoxib versus celecoxib in treating pain after dental surgery: A single-center, randomized, double-blind, placebo- and active-comparator-controlled, parallel-group, single-dose study using the dental impaction pain model. *Clinical Therapeutics* **24,** 1549–1560.

Morikawa, T. and Yoshida, M. (1995). A useful testing strategy in phase III trials: Combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* **5,** 297–306.

Yosipovitch, G., Sugeng, M. W., Goon, A., Chan, Y. H., and Goh, C. L. (2001). A comparison of the combined effect of cryotherapy and corticosteroid injections versus corticosteroids and cryotherapy alone on keloids: A controlled study. *Journal of Dermatological Treatment* **12,** 87–90.

Received July 2003. *Revised November* 2003.
Accepted January 2004.