12-2012

Optimal design of p-value consistent step-up procedures for multiple comparisons with a control in direction-mixed families

Koon Shing KWONG
Singapore Management University, kskwong@smu.edu.sg

Siu Hung CHEUNG
Chinese University of Hong Kong

DOI: https://doi.org/10.1016/j.csda.2012.05.001

Published in Computational Statistics & Data Analysis
Volume 56, Issue 12, December 2012, Pages 4157-4164
https://doi.org/10.1016/j.csda.2012.05.001

Optimal design for *p*-value consistent step-up procedures for multiple comparisons with a control in direction-mixed families

Koon Shing Kwong [a,*], Siu Hung Cheung [b]

[a] *School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore*
[b] *Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong, China*

Keywords:
Multiple comparisons with a control
Power function
Step-up multiple test
Optimal sample size
Familywise error rate

A B S T R A C T

It is common in clinical studies for several treatments to be compared to a control. Most of the related statistical techniques have been developed to accommodate inferential families in which all hypotheses are either one- or two-sided such that the familywise error rate is controlled at a specified level. Several multiple testing procedures were recently proposed to perform multiple comparisons with a control in direction-mixed families that contain a mixture of one- and two-sided inferences. Of these procedures, the *p*-value consistent step-up procedure is found to be superior in terms of its power and *p*-value consistent property. In this paper, we examine the techniques for computing the all-pairs power of this testing procedure. We also provide the means to obtain the optimal design when a desired level of all-pairs power is given. Compared to the conservative method of treating all hypotheses as two-sided, the proposed procedure requires a substantially smaller sample, as all useful information on the direction of the alternatives is utilized in an optimal way. The computation of the optimal design relies on an efficient algorithm, which is also discussed in this paper. A clinical study example is employed to illustrate the proposed procedure.

1. Introduction

Multiple comparisons with a control are frequently encountered in clinical trials. For continuous responses, the Dunnett (1955) procedure is a well-known single-step approach for comparing the efficacy of treatments to that of a control while the familywise error rate (FWE), the probability of rejecting at least one true null hypothesis, is maintained at level α. In addition to this single-step procedure, several stepwise multiple comparison procedures have also been proposed to increase the power of the tests, such as those proposed by Dunnett and Tamhane (1991, 1992, 1995). However these procedures are designed only for situations in which all of the hypotheses in the family are either one- or two-sided.

As Cheung et al. (2004) noted, in some clinical studies comparing several treatments with a control, some of the hypotheses in the inferential families should actually be tested as one-sided, whereas the others are regarded as two-sided. Families that contain a mixture of one- and two-sided tests are called direction-mixed families. Because of a lack of statistical tools, multiple comparison procedures that treat all of the hypotheses considered as two-sided inferences have been employed in direction-mixed families. Although this conservative approach is easy to implement, it may substantially reduce the overall power of the tests, as the direction information of the one-sided tests is ignored.

Kwong et al. (2007) noted that the choice between one- and two-sided tests should not be *post-hoc*, but must be well justified. Therefore, direction-mixed family inferences should be employed only after careful consideration and thorough

* Corresponding author. Tel.: +65 68280834; fax: +65 68280833.
E-mail address: kskwong@smu.edu.sg (K.S. Kwong).

justification. For instance, direction-mixed family inferences may not be a good choice for highly regulatory environments in which one-sided hypothesis testing is required with a half significance level. More detailed discussion of the choice between one- and two-sided tests is given in Kwong et al. (2007) and the references therein.

For multiple comparisons with a control in direction-mixed families, Cheung et al. (2004) developed a single-step multiple testing procedure for many-to-one comparisons. This procedure effectively incorporates the direction information into the multiple hypothesis testing procedure, and is found to be more powerful than the two-sided Dunnett procedure, particularly when the proportion of one-sided hypotheses in the direction-mixed family is large. The procedure is constructed on the basis of the objective to minimize the expected average width of the confidence intervals.

However, there is a chance that the testing conclusion in the method proposed by Cheung et al. (2004) will not be p-value consistent; that is, when a particular hypothesis in the inferential family with a given p-value, say p^*, is rejected, there is a chance that some of the other hypotheses with p-values less than p^* will not be rejected. This is an undesirable property in multiple testing because, Westfall (1997) noted, a smaller p-value intuitively indicates stronger evidence against the null hypothesis, which therefore should be rejected if another hypothesis with a larger p-value has been rejected. Note that the most popular multiple comparison procedures, such as those proposed by Dunnett (1955), Holm (1979), Hommel (1988), Hochberg (1988), Dunnett and Tamhane (1991, 1992, 1995), and Kwong (2001), are p-value consistent. Nevertheless, if the objective is to construct simultaneous confidence intervals but not tests, then p-value consistency is a minor consideration.

To take the p-value consistency requirement into consideration, Kwong et al. (2007) proposed a new single-step procedure and two stepwise procedures for multiple comparisons with a control in direction-mixed families. As expected, the average power study revealed the p-value consistent step-up procedure to be uniformly more powerful than the other procedures in all of the cases they considered. However, in direction-mixed family inferences, there is no discussion of the strategy employed to determine the optimal total sample size before the onset of a clinical study.

The objective of this paper is to propose a new algorithm for determining the optimal design for the p-value consistent step-up procedure when a desired power level is given in a direction-mixed family. Section 2 is devoted to a brief discussion of existing testing procedures for multiple comparisons with a control in direction-mixed families. Various definitions of power are presented in Section 3. An efficient algorithm for enhancing the search for the optimal design for the p-value consistent step-up procedure is provided in Section 4. The application of the new procedure in a clinical study is presented in Section 5, and concluding remarks are made in Section 6.

2. Overview of procedures in direction-mixed families

2.1. Notation and model

Consider a one-way fixed effects model with $m+1$ treatments:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 0, 1, \ldots, m,\ j = 1, \ldots, n_i,$$

where Y_{ij} represents the jth observation on the ith treatment, μ_i is the ith treatment mean, and ϵ_{ij} is the random error component. The sample sizes of the m treatments and the control are n_i (for $i = 1, \ldots, m$) and n_0, respectively, and $i = 0$ denotes the control. Hence, the total sample size, denoted by N, is $\sum_{i=0}^{m} n_i$. Assume that $\epsilon_{ij} \overset{ind}{\sim} N(0, \sigma^2)$, where σ^2 is the unknown common variance. Let $\bar{Y}_i$ be the sample mean of the ith treatment and $\hat{\sigma}^2$ be the pooled sample variance, which is an unbiased estimator of σ^2 and also independent of $\bar{Y}_i$. Let $\theta_i = \mu_i - \mu_0$ be the efficacy differences between the ith treatment and the control for $i = 1, \ldots, m$. Assume that $\theta_i > 0$ implies that treatment i is better than the control. To compare the m treatments with the control in a direction-mixed family, we simultaneously test the m null hypotheses:

$$H_i : \mu_0 = \mu_i$$

for $i = 1, \ldots, m$ versus r one-sided alternative hypotheses

$$H_i' : \mu_0 < \mu_i$$

for $i = 1, \ldots, r$ and $(m - r)$ two-sided alternative hypotheses

$$H_j' : \mu_0 \neq \mu_j$$

for $j = r+1, \ldots, m$. With respect to the family of null hypotheses $\{H_1, \ldots, H_m\}$, a subset $\{H_1, \ldots, H_r\}$ $(r \leq m)$ is tested against the one-sided alternatives, and the remaining null hypotheses $\{H_{r+1}, \ldots, H_m\}$ are tested against the two-sided alternatives.

To test the m null hypotheses simultaneously in a direction-mixed family, the test statistics are $T_1, \ldots, T_r, |T_{r+1}|, \ldots, |T_m|$, respectively, where

$$T_i = \frac{\bar{Y}_i - \bar{Y}_0}{\hat{\sigma}\sqrt{1/n_i + 1/n_0}}$$

for $i = 1, \ldots, m$. The variates $T_1, T_2, \ldots, T_m$ have a multivariate t-distribution with $f = N - m - 1$ degrees of freedom and a $m \times m$ correlation matrix $\{\rho_{ij}^{(m)} = \rho_i \rho_j\}$, where $\rho_{ij}^{(m)} = 1$ for $i = j$ and $\rho_{ij}^{(m)} = \rho_i \rho_j$ for $i \neq j$, with $\rho_i = \sqrt{n_i/(n_0 + n_i)}$

for $i = 1, \ldots, m$. Note that the balanced designs in mixed-direction families have the restrictions $n_1 = \cdots = n_r = n_I$, $n_{r+1} = \cdots = n_m = n_{II}$ and $N = n_0 + n_I r + n_{II}(m - r)$. The correlation structure of the balanced designs reduces to $\rho_{ij}^{(m)} = \rho_i \rho_j$ for $i \neq j$, with $\rho_i = \rho_I$ for $i = 1, \ldots, r$ and $\rho_i = \rho_{II}$ for $i = r + 1, \ldots, m$, where $\rho_I = \sqrt{n_I/(n_0 + n_I)}$ and $\rho_{II} = \sqrt{n_{II}/(n_0 + n_{II})}$. To control the FWE at α, the following condition is required

$$\text{FWE} = P(\text{reject any true hypothesis}) \leq \alpha$$

under any configuration of the parameters θ_i, $i = 1, \ldots, m$.

We next provide a brief overview of several testing procedures that could be employed to conduct multiple comparisons with a control in a direction-mixed family.

2.2. Single-step procedures

The single-step procedure proposed by Cheung et al. (2004) is based on the criterion that the expected average width of the simultaneous confidence intervals of θ_i, $i = 1, \ldots, m$ is minimized in direction-mixed families. However, as Kwong et al. (2007) point out, this procedure is not p-value consistent. For instance, the procedure may possibly reject only the two-sided hypothesis, not the one-sided hypothesis, even though the p-value of the latter is in fact smaller than that of the former. As most of the popular multiple comparison procedures are all p-value consistent, Kwong et al. (2007) proposed a single-step p-value consistent procedure, denoted by SS hereafter, for constructing the simultaneous confidence intervals of θ_i, $i = 1, \ldots, m$ in direction-mixed families.

The SS procedure involves computation of two positive critical values, $c^+_{1,\alpha,m,r}$ and $c^+_{2,\alpha,m,r}$, in the direction-mixed families for one- and two-sided inferences, respectively, such that

$$G(c^+_{1,\alpha,m,r}, c^+_{2,\alpha,m,r}) = 1 - \alpha, \tag{1}$$

where

$$P(T_i > c^+_{1,\alpha,m,r}) = P(|T_j| > c^+_{2,\alpha,m,r}) = p_{m,\alpha,r} \tag{2}$$

for any i and j. As the marginal distributions of T_i and T_j for any i, j have an identical t-distribution, the aforementioned relationship between $c^+_{1,\alpha,m,r}$ and $c^+_{2,\alpha,m,r}$ obviously guarantees that the critical cutoff points are identical in terms of the p-value, $p_{m,\alpha,r}$, for any one- and two-sided inferences. Because any hypothesis with a corresponding p-value less than $p_{m,\alpha,r}$ is rejected, regardless of whether it is one- or two-sided, the SS procedure is a p-value consistent procedure.

The critical constants in (1) can be solved using constraint (2). Computation can be completed using the algorithm of either Dunnett (1989) or Cheung and Holland (1991, 1992), and details can be found in Cheung et al. (2004). Selected critical values for various value combinations of α, m, r, f, and ρ are tabulated in Kwong et al. (2007).

2.3. Stepwise procedures

If the primary objective of a statistical inference is to detect significant differences in treatment efficacy, then stepwise multiple comparison procedures are often more appropriate than their single-step counterparts because they are, in general, more powerful. A typical step-down procedure tests hypotheses sequentially, from that with the smallest p-value to that with the largest.

The p-value consistent step-down procedure, denoted by SD hereafter, requires a determination of m sets of critical p-values for comparisons with m ordered p-values. The SD procedure requires only one of the critical p-values in each set for comparison with the corresponding ordered observed p-value. The selection of an appropriate critical p-value in each step depends on how many one-sided hypotheses have been rejected previously. The derivation and evaluation of these m sets of critical p-values can be found in Kwong et al. (2007).

In contrast to step-down procedures, a typical step-up procedure tests hypotheses sequentially from that with the largest p-value to that with the smallest. The p-value consistent step-up procedure, denoted by SU hereafter, for direction-mixed families requires a determination of the m critical p-values, $p^u_m \geq p^u_{m-1} \geq \cdots \geq p^u_1$, for comparisons with m ordered observed p-values. The derivation and evaluation of these critical p-values are outlined in Kwong et al. (2007). The algorithm of the SU procedure is as follows.

1. Obtain the set of ordered observed p-values, $p_{(1)} \leq \cdots \leq p_{(m)}$, with the corresponding null hypotheses $H_{(1)}, \ldots, H_{(m)}$.
2. Compare $p_{(m)}$ with p^u_m. If $p_{(m)} < p^u_m$, then reject all null hypotheses and terminate the procedure; otherwise, accept $H_{(m)}$ and proceed to the next step of comparing $p_{(m-1)}$ with p^u_{m-1}.
3. The testing procedure continues until the first occurrence of $p_{(i)} < p^u_i$, say $i = s \leq m$. Then, $H_{(s+1)}, \ldots, H_{(m)}$ are accepted, and the remaining $H_{(1)}, \ldots, H_{(s)}$ are rejected. If $p_{(i)} \geq p^u_i$ for $i = 1, \ldots, m$, then all of the null hypotheses are accepted.

Step-up procedures are generally more powerful than their step-down counterparts (see Dunnett and Tamhane (1992) for details). The simulation study in Kwong et al. (2007) suggests that the SU procedure is also more powerful than the other p-value consistent procedures in direction-mixed families. Hence, the remainder of this paper focuses on sample size determination for the SU procedure.

In fact, when the inferential family comprises all two-sided hypotheses ($r = 0$) or all one-sided hypotheses ($r = m$), the aforementioned SU procedures reduce to the Dunnett and Tamhane (1992) two- and one-sided step-up procedures, respectively. The algorithm given in Kwong et al. (2010) can be employed to obtain the required sample sizes for the balanced designs of the SU procedures in these two cases. Therefore, this paper focuses only on the method used to evaluate the required sample size for the SU procedure for $r = 1, \ldots, m-1$ in direction-mixed families.

3. Power definitions

There are numerous ways to conceptualize power in multiple testing scenarios. It is important to clarify which definition of power is being employed in the search for the required sample size because different definitions may yield very different required sample sizes. As noted by Westfall and Young (1993), the decision to apply a certain type of power is governed by the needs and objectives of the experiment and therefore should not be an *ad-hoc* decision. Let $\mathcal{F} \subset \{1, \ldots, k\}$ denote the family of all true hypotheses and $\mathcal{F}^c$ denote a set that contains all of the remaining hypotheses, which are false. Three popular concepts of power (see Horn and Dunnett (2004)) are introduced here:

- all-pairs power $P_{all} = P(\text{reject } \textit{all } H_i \in \mathcal{F}^c)$;
- any-pair power $P_{any} = P(\text{reject } \textit{at least one } H_i \in \mathcal{F}^c)$; and
- per-pair power $P_{per} = P(\text{reject } \textit{a particular } H_i \in \mathcal{F}^c)$.

Note that if there is only one false hypothesis, then the three definitions are equivalent. For cases of more than one false hypothesis, Westfall and Young (1993) provided a number of ways to choose among the foregoing definitions of power in different practical settings.

Of the three, all-pairs power is the most stringent requirement and yields the largest sample size. Any-pair power, in contrast, is the least stringent and requires the smallest sample size to meet a pre-specified power level. The latter should be employed if the objective of the multiple test is to determine whether there is any significant treatment difference in the family of inferences considered. In cases in which one individual hypothesis is considered in advance to be much more important than the others, and the researcher would definitely want to reject it if it were false, per-pair power is the most appropriate choice for determining the sample size requirement. Several papers, including those of Dunnett et al. (2001), Horn and Dunnett (2004), and Kwong et al. (2010), compare the effects of these three power definitions on sample size requirements and different sample size allocation methods when all of the hypotheses are either one- or two-sided.

In the absence of a specific reason not to, most researchers would prefer to reject all false hypotheses in a multiple testing environment. This paper thus focuses on the determination of optimal designs for the SU procedure under the all-pairs power definition. However, a straightforward generalization (see the details in Kwong et al. (2010)) can be applied when other definitions of power are employed, and hence is not reported here.

4. Optimal designs

Assume that the SU procedure is used in a clinical study to detect any treatment differences of practical importance, $\Delta > 0$. Without loss of generality, let $\boldsymbol{\theta}_{s,v,t} \subset \boldsymbol{\theta}_m^r = (\theta_{l_1}, \ldots, \theta_{l_r}, \theta_{g_1}, \ldots, \theta_{g_{m-r}})$ be a vector with elements

$$\theta_i = \begin{cases} \Delta & \text{for } i = l_1, \ldots, l_s, g_1, \ldots, g_v \\ -\Delta & \text{for } i = g_{v+1}, \ldots, g_t \\ 0 & \text{for } i = l_{s+1}, \ldots, l_r, g_{t+1}, \ldots, g_{m-r}, \end{cases}$$

i.e., the s one-sided hypotheses $(H_{l_1}, \ldots, H_{l_s})$ and the t two-sided hypotheses $(H_{g_1}, \ldots, H_{g_t})$ are false, and the remainder are true, where $0 \le s \le r$, $0 \le v \le t \le m - r$ and $s + t \ge 1$. For any given parameter vectors $\boldsymbol{\theta}_{s,v,t}$ in the direction-mixed families, it is necessary to determine the following probability.

$$P(R = s + t \,|\, \boldsymbol{\theta}_{s,v,t}) = P(\text{Reject } H_{l_1}, \ldots, H_{l_s}, H_{g_1}, \ldots, H_{g_t} \,|\, \boldsymbol{\theta}_{s,v,t}),$$

where R is the number of false hypotheses rejected in the SU procedure. The evaluation of probability is discussed in the Appendix.

For all possible values of s, v, t, the all-pairs power function is the one with the least favorable configuration set of (s, v, t), say (s^*, v^*, t^*), such that

$$P_{all}(\boldsymbol{\theta}_m^r) = P(R = s^* + t^* \,|\, \boldsymbol{\theta}_{s^*,v^*,t^*}) \le P(R = s + t \,|\, \boldsymbol{\theta}_{s,v,t}).$$

We also need to find the best way to allocate the total sample size N to n_0, n_I and n_{II} in the balanced designs. In other words, the optimal design has the configuration, say $(N^*, n_0^*, n_I^*, n_{II}^*)$, in which N^* is the smallest value of all possible N, and the combination (n_0^*, n_I^*, n_{II}^*) must have the largest value $P_{all}(\boldsymbol{\theta}_m^r)$ among all combinations of (n_0, n_I, n_{II}) under the constraint $N^* = n_0 + rn_I + (m - r)n_{II}$, such that the specified power level is achieved.

Although an analytical determination of the combination (n_0^*, n_I^*, n_{II}^*) seems infeasible, Hayter and Tamhane's (1991) max–min approach, which combines the least favorable configuration of (s, v, t) and the maximum power configuration of (N, n_0, n_I, n_{II}), can be used to search numerically for the optimal design. According to Kwong et al. (2010), square-root allocation provides a good initial numerical search for the optimal design when $r = 0$ and $r = m$. Hence, for a specific all-pairs power level $1 - \beta$, the following algorithm is proposed to search for the optimal design for $r = 1, \ldots, m - 1$.

(a) Use square-root allocation to obtain the initial total sample size, say N', under the assumption $r = 0$.
(b) Apply the max–min power approach to identify the maximum power combination (n_0', n_I', n_{II}') with (s^*, v^*, t^*) for N', and then evaluate its corresponding power level, say P'.
(c) Set $N' = N' - 1$ and repeat (b) if $P' > 1 - \beta$. Otherwise, the optimal design configuration $(N^*, n_0^*, n_I^*, n_{II}^*) = (N', n_0', n_I', n_{II}')$.

The proposed algorithm provides a very efficient way of obtaining the optimal balanced design for the p-value consistent SU procedure, particularly when Δ/σ is large. As the difference between N' in step (a) and N^* is relatively large for a small Δ/σ, a simple linear interpolation can be employed in step (c) to speed up the search for N^*. Table 1 presents the design configurations and sample size requirements for the p-value consistent SU procedure under the all-pairs power definition for $\alpha = 0.05$, $m = 3, 4, 5$, $r = 0, \ldots, m - 1$, and $\Delta/\sigma = 0.5, 1, 1.5, 2$. Table 1 reveals several important patterns that can be summarized as follows.

1. The total sample size N^* decreases as r increases. For a given parameter configuration, the benefit of the savings realized in the required total sample size represents the required total sample size difference between the SU procedure and the procedure that treats all hypotheses as two-sided, i.e., that assumes $r = 0$. Note that this benefit is more significant for a large r when the value of Δ/σ is small. For example, when $(\alpha, m, 1 - \beta, \Delta/\sigma) = (0.05, 5, 0.9, 0.5)$, treating all hypotheses as two-sided requires 2.3% and 10.9% more in terms of the total sample size than does using the optimal design with $r = 1$ and $r = 4$, respectively. As a result, the proposed algorithm for constructing the balanced optimal design for the SU procedure may reduce the total sample size requirement considerably, particularly when r is very close to m.
2. As expected, the sample size n_I^* for the treatment with a one-sided inference is always smaller than the sample size n_{II}^* for that with a two-sided inference due to the direction information of the one-sided alternatives in direction-mixed families.
3. For $r \geq 1$ and large values of Δ/σ, say those greater than 1, there are not large differences between n_I^* and n_{II}^* because n_I^* is already quite small, and any further reduction in n_I^* inflates the variance of the test statistics for multiple comparisons to a large extent.
4. For a given parameter configuration, required sample sizes n_0^*, n_I^* and n_{II}^* remain quite stable for different values of r. The small variations in each can be attributed to the maximization process over a multi-dimensional space of positive integers.
5. As $1 - \beta$ increases for a given parameter configuration, the impacts on n_0^*, n_I^*, and n_{II}^* are very similar for decreasing values of Δ/σ.

5. Example

An example given in Cheung et al. (2004) provides a suitable scenario for application of the SU procedure. A clinical study reported by Schwartz et al. (2002) was carried out to compare the renal effects of rofecoxib and celecoxib with those of naproxen in elderly subjects on a normal-salt diet. The response variable was the change from baseline in daily urinary sodium excretion during the first 72 h of treatment. Four treatments were considered: rofecoxib, celecoxib, naproxen (active control), and a placebo (no treatment). Naproxen was identified as the "control" treatment to be compared with the other three treatments, including the placebo. A placebo was included to ensure a valid clinical trial if naproxen's superiority to the placebo were demonstrated in hypothesis testing. As naproxen's superiority to a placebo has a long research history, a one-sided hypothesis test of naproxen and the placebo is appropriate. However, without any such forceful justification, the comparisons between celecoxib, rofecoxib and naproxen should have been two-sided inferences.

For this example, $m = 3$ and $r = 1$, and we let the required all-pairs power be 0.8, $\alpha = 0.05$, and $\Delta/\sigma = 1.0$. From Table 1, we can see that the optimal total sample size is 88, and the number of subjects allocated to rofecoxib, celecoxib, naproxen, and the placebo is thus 21, 21, 30, and 16, respectively. Based on the Dunnett and Tamhane (1992) step-up procedure, and assuming that all hypotheses are two-sided inferences, the total required sample size is 92, which is about 4.5% more than the optimal design when the direction information of the one-sided alternative is ignored.

6. Conclusion and final remarks

In this paper, we discuss a method for determining the optimal design for the p-value consistent step-up procedure in direction-mixed families, such that a specified level of all-pairs power is achieved. An efficient algorithm is proposed only for all-pairs power, which is the desirable option for practical purposes. However, the approach can also be generalized to other

Table 1
Optimal designs (N^*, n_0^*, n_l^*, n_{ll}^*) of the p-value consistent step-up procedure under all-pairs power definition.

Parameters				Δ/σ			
α	m	$1-\beta$	r	0.5	1.0	1.5	2.0
0.05	3	0.6	0	(275, 101, 0, 58)	(69, 27, 0, 14)	(31, 13, 0, 6)	(18, 6, 0, 4)
			1	(260, 96, 44, 60)	(65, 24, 11, 15)	(30, 11, 5, 7)	(17, 6, 3, 4)
			2	(244, 97, 43, 61)	(61, 24, 11, 15)	(28, 11, 5, 7)	(16, 6, 3, 4)
		0.7	0	(314, 116, 0, 66)	(79, 28, 0, 17)	(35, 14, 0, 7)	(20, 8, 0, 4)
			1	(298, 111, 51, 68)	(75, 28, 13, 17)	(34, 12, 6, 8)	(19, 8, 3, 4)
			2	(281, 109, 51, 70)	(71, 28, 13, 17)	(32, 12, 6, 8)	(18, 8, 3, 4)
		0.8	0	(365, 134, 0, 77)	(92, 35, 0, 19)	(41, 14, 0, 9)	(23, 8, 0, 5)
			1	(348, 127, 61, 80)	(88, 30, 16, 21)	(39, 14, 7, 9)	(22, 8, 4, 5)
			2	(328, 126, 60, 82)	(83, 33, 15, 20)	(37, 14, 7, 9)	(21, 8, 4, 5)
		0.9	0	(445, 160, 0, 95)	(112, 40, 0, 24)	(50, 17, 0, 11)	(28, 10, 0, 6)
			1	(426, 156, 76, 97)	(107, 40, 19, 24)	(48, 17, 9, 11)	(27, 10, 5, 6)
			2	(404, 153, 75, 101)	(101, 38, 19, 25)	(46, 17, 9, 11)	(26, 10, 5, 6)
	4	0.6	0	(379, 123, 0, 64)	(95, 31, 0, 16)	(43, 15, 0, 7)	(24, 8, 0, 4)
			1	(365, 120, 50, 65)	(92, 32, 12, 16)	(42, 12, 6, 8)	(23, 8, 3, 4)
			2	(350, 114, 51, 67)	(88, 28, 13, 17)	(40, 12, 6, 8)	(23, 9, 3, 4)
			3	(334, 115, 51, 66)	(84, 29, 13, 16)	(38, 13, 6, 7)	(22, 9, 3, 4)
		0.7	0	(427, 129, 0, 72)	(107, 35, 0, 18)	(48, 16, 0, 8)	(28, 8, 0, 5)
			1	(412, 136, 57, 73)	(103, 35, 14, 18)	(46, 16, 6, 8)	(27, 8, 4, 5)
			2	(396, 130, 58, 75)	(100, 32, 15, 19)	(45, 13, 7, 9)	(26, 8, 4, 5)
			3	(379, 130, 58, 75)	(95, 31, 15, 19)	(43, 14, 7, 8)	(25, 8, 4, 5)
		0.8	0	(488, 160, 0, 82)	(123, 39, 0, 21)	(55, 19, 0, 9)	(31, 11, 0, 5)
			1	(472, 154, 66, 84)	(119, 40, 16, 21)	(53, 19, 7, 9)	(30, 11, 4, 5)
			2	(455, 149, 67, 86)	(114, 36, 17, 22)	(51, 15, 8, 10)	(29, 11, 4, 5)
			3	(437, 149, 67, 87)	(110, 37, 17, 22)	(49, 16, 8, 9)	(28, 11, 4, 5)
		0.9	0	(585, 189, 0, 99)	(147, 47, 0, 25)	(65, 21, 0, 11)	(37, 13, 0, 6)
			1	(568, 188, 80, 100)	(142, 47, 20, 25)	(64, 22, 9, 11)	(36, 13, 5, 6)
			2	(549, 179, 82, 103)	(138, 48, 20, 25)	(62, 22, 9, 11)	(35, 13, 5, 6)
			3	(529, 178, 82, 105)	(133, 44, 21, 26)	(59, 21, 9, 11)	(34, 13, 5, 6)
	5	0.6	0	(486, 146, 0, 68)	(122, 37, 0, 17)	(55, 15, 0, 8)	(31, 11, 0, 4)
			1	(472, 142, 54, 69)	(119, 38, 13, 17)	(53, 15, 6, 8)	(31, 12, 3, 4)
			2	(457, 142, 54, 69)	(115, 33, 14, 18)	(52, 16, 6, 8)	(30, 7, 4, 5)
			3	(441, 134, 55, 71)	(111, 33, 14, 18)	(50, 16, 6, 8)	(29, 9, 4, 4)
			4	(425, 139, 54, 70)	(107, 34, 14, 17)	(48, 16, 6, 8)	(28, 8, 4, 4)
		0.7	0	(541, 161, 0, 76)	(136, 41, 0, 19)	(61, 16, 0, 9)	(34, 9, 0, 5)
			1	(527, 162, 61, 76)	(132, 41, 15, 19)	(59, 16, 7, 9)	(34, 10, 4, 5)
			2	(512, 157, 62, 77)	(129, 37, 16, 20)	(58, 17, 7, 9)	(33, 10, 4, 5)
			3	(496, 149, 63, 79)	(125, 37, 16, 20)	(56, 17, 7, 9)	(32, 10, 4, 5)
			4	(479, 156, 61, 79)	(120, 40, 15, 20)	(54, 17, 7, 9)	(30, 9, 4, 5)
		0.8	0	(615, 185, 0, 86)	(154, 44, 0, 22)	(69, 19, 0,10)	(39, 14, 0, 5)
			1	(600, 186, 70, 86)	(151, 45, 18, 22)	(67, 19, 8,10)	(39, 10, 5, 6)
			2	(584, 173, 72, 89)	(147, 45, 18, 22)	(66, 20, 8,10)	(38, 10, 5, 6)
			3	(568, 174, 72, 89)	(142, 44, 18, 22)	(64, 20, 8,10)	(37, 10, 5, 6)
			4	(549, 174, 71, 91)	(138, 43, 18, 23)	(62, 20, 8,10)	(36, 10, 5, 6)
		0.9	0	(732, 222, 0, 102)	(184, 54, 0, 26)	(82, 27, 0, 11)	(47, 12, 0, 7)
			1	(716, 219, 85, 103)	(180, 55, 21, 26)	(80, 22, 10, 12)	(46, 12, 6, 7)
			2	(699, 215, 86, 104)	(175, 50, 22, 27)	(78, 22, 10, 12)	(45, 12, 6, 7)
			3	(680, 202, 88, 107)	(171, 51, 22, 27)	(76, 22, 10, 12)	(44, 12, 6, 7)
			4	(660, 206, 86, 110)	(166, 51, 22, 27)	(74, 22, 10, 12)	(42, 11, 6, 7)

definitions of power. As expected, the new sample size determination procedure, which incorporates useful information on the direction of the one-sided alternatives, requires a smaller sample size for treatments with one-sided inferences than those with two-sided inferences. The results of our extensive numerical study suggest that, compared with the step-up method that treats all hypotheses as two-sided, our approach results in a total sample size reduction of 2%–11% in direction-mixed families, depending on the value of r.

The discussion in this paper applies to specific settings in which several treatments are compared to a control, and all-pairs power is specified. Further research is required to explore other scenarios such as all possible contrasts in direction-mixed families, as discussed in Braat et al. (2008), and the other definitions of power given in Horn and Dunnett (2004).

Acknowledgements

The authors are grateful to the referees for their helpful comments, which have led to a significant improvement of the paper. KSK is supported by a Singapore Management University research grant (C244/MSS8E009) and SHC by the Research Grants Councils of the Hong Kong Special Administrative Region (CUHK400710).

Appendix. Evaluation of $P(R = s + t|\theta_{s,v,t})$

Let $(a_1, \ldots, a_m) > (b_1, \ldots, b_m)$ denote $a_{(i)} > b_{(i)}$ for $i = 1, \ldots, m$, where $a_{(1)} \leq \cdots \leq a_{(m)}$ and $b_{(1)} \leq \cdots \leq b_{(m)}$. Under the p-value consistent SU procedure with parameter vector $\theta_{s,v,t}$ in direction-mixed families, we obtain

$$\begin{aligned} &P(\text{Reject } H_{l_1}, \ldots, H_{l_s}, H_{g_1}, \ldots, H_{g_t} \,|\, \theta_{s,v,t}) \\ &= \sum_{i=0}^{r-s} \sum_{j=0}^{m-r-t} \binom{r-s}{i} \binom{m-r-t}{j} P\Big[(P_{l_{r-i+1}}, \ldots, P_{l_r}, P_{g_{m-r-j+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-i-j+1}, \ldots, p^u_m) \bigcap \\ &\quad \max(P_{l_1}, \ldots, P_{l_{r-i}}, P_{g_1}, \ldots, P_{g_{m-r-j}}) < p^u_{m-i-j}\Big], \end{aligned} \tag{3}$$

where P_i is the p-value corresponding to hypothesis H_i for $i = l_1, \ldots, l_r, g_1, \ldots, g_{m-r}$. To deal with the event $\max(P_{l_1}, \ldots, P_{l_{r-i}}, P_{g_1}, \ldots, P_{g_{m-r-j}}) < p^u_{m-i-j}$ in (3), we can simply invert the p-values P_i to the corresponding test statistics:

$$T_i = \begin{cases} \left[\sqrt{1-\rho_I^2} Z_i - \rho_I (Z_0 + \sqrt{n_0}\Delta/\sigma)\right] \Big/ U & \text{for } i = l_1, \ldots, l_s \\ \left[\sqrt{1-\rho_I^2} Z_i - \rho_I Z_0\right] \Big/ U & \text{for } i = l_{s+1}, \ldots, l_r \\ \left[\sqrt{1-\rho_{II}^2} Z_i - \rho_{II} (Z_0 + \sqrt{n_0}\Delta/\sigma)\right] \Big/ U & \text{for } i = g_1, \ldots, g_v \\ \left[\sqrt{1-\rho_{II}^2} Z_i - \rho_{II} (Z_0 - \sqrt{n_0}\Delta/\sigma)\right] \Big/ U & \text{for } i = g_{v+1}, \ldots, g_t \\ \left[\sqrt{1-\rho_{II}^2} Z_i - \rho_{II} Z_0\right] \Big/ U & \text{for } i = g_{t+1}, \ldots, g_{m-r} \end{cases}$$

for $i = l_1, \ldots, l_r, g_1, \ldots, g_{m-r}$, where Z_i for $i = 0, l_1, \ldots, l_r, g_1, \ldots, g_{m-r}$ are mutually independent standard normal random variables, and U is a $\sqrt{\chi_f^2/f}$ random variable that is independent of all Z_i. However, directly inverting the p-values to the corresponding test statistics is not straightforward for event $(P_{l_{r-i+1}}, \ldots, P_{l_r}, P_{g_{m-r-j+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-i-j+1}, \ldots, p^u_m)$ in (3), and we thus have to modify the event before applying the procedure of inversion.

For ease of exposition, we now illustrate the procedure for modifying the probability of event $(P_{l_{s+1}}, \ldots, P_{l_r}, P_{g_{t+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-s-t+1}, \ldots, p^u_m)$. Similar to the approach presented in Kwong et al. (2007), applying the law of total probability to $P_{l_{s+1}}$ which has to fall into one of the intervals $(p^u_{m-s-t+1}, p^u_{m-s-t+2}), \ldots, (p^u_m, \infty)$, allows us to obtain

$$\begin{aligned} &P\left[(P_{l_{s+1}}, \ldots, P_{l_r}, P_{g_{t+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-s-t+1}, \ldots, p^u_m)\right] \\ &\quad = P\left[(P_{l_{s+2}}, \ldots, P_{l_r}, P_{g_{t+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-s-t+2}, \ldots, p^u_m) \bigcap (p^u_{m-s-t+1} < P_{l_{s+1}} < p^u_{m-s-t+2})\right] \\ &\qquad + P\Big[(P_{l_{s+2}}, \ldots, P_{l_r}, P_{g_{t+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-s-t+1}, p^u_{m-s-t+3}, \ldots, p^u_m) \\ &\qquad\quad \bigcap (p^u_{m-s-t+2} < P_{l_{s+1}} < p^u_{m-s-t+3})\Big] + \cdots \\ &\qquad + P\left[(P_{l_{s+2}}, \ldots, P_{l_r}, P_{g_{t+1}}, \ldots, P_{g_{m-r}}) > (p^u_{m-s-t+1}, \ldots, p^u_{m-1}) \bigcap (p^u_m < P_{l_{s+1}})\right]. \end{aligned} \tag{4}$$

After successively applying the law of total probability from $P_{l_{s+1}}$ to $P_{g_{m-r}}$ in (4), we can then invert the p-values to the corresponding test statistics, i.e., the event $\{a < P_i < b\}$ for $i = l_{s+1}, \ldots, l_r, g_{t+1}, \ldots, g_{m-r}$ becomes $\{F^{-1}(1-b) < T_i < F^{-1}(1-a)\}$ for $i = l_{s+1}, \ldots, l_r$ and $\{F^{-1}(1-b/2) < T_i < F^{-1}(1-a/2)\}$ for $i = g_{t+1}, \ldots, g_{m-r}$, respectively, where F^{-1} is the inverse cumulative function of the t-distribution with f degrees of freedom. As a result, the probability in (3) can be evaluated with a simple modification of the algorithm given in Kwong and Liu (2000) or Kwong and Chan (2008).

References

Braat, S., Gerhard, D., Hothorn, L.A., 2008. Joint one-sided and two-sided simultaneous confidence intervals. Journal of Biopharmaceutical Statistics 18, 293–306.

Cheung, S.H., Holland, B., 1991. Extension of Dunnett's multiple comparison procedure to the case of several groups. Biometrics 47, 21–32.

Cheung, S.H., Holland, B., 1992. Extension of Dunnett's multiple comparison procedure with different sample sizes to the case of several groups. Computational Statistics and Data Analysis 14, 165–182.

Cheung, S.H., Kwong, K.S., Chan, W.S., Leung, S.P., 2004. Multiple comparisons with a control in families with both one-sided and two-sided hypotheses. Statistics in Medicine 23, 2975–2988.

Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 50, 1096–1121.

Dunnett, C.W., 1989. Algorithm AS251: multivariate normal probability integrals with product correlation structure. Applied Statistics 38, 564–579 [Correction note: 1993, Applied Statistics 42, 709].

Dunnett, C.W., Horn, M., Vollandt, R., 2001. Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. Journal of Statistical Planning and Inference 97, 367–384.

Dunnett, C.W., Tamhane, A.C., 1991. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. Statistics in Medicine 10, 939–947.

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

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

Hayter, A.J., Tamhane, A.C., 1991. Sample size determination for step-down multiple test procedures: orthogonal contrasts and comparisons with a control. Journal of Statistical Planning and Inference 27, 271–290.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal Statistics 6, 65–70.

Hommel, G., 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75, 383–386.

Horn, M., Dunnett, C.W., 2004. Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. In: Lecture Notes—Monograph Series, vol. 47. Institute of Mathematical Statistics, pp. 48–64.

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., Chan, Y.M., 2008. On the evaluation of the joint distribution of order statistics. Computational Statistics and Data Analysis 52, 5091–5099.

Kwong, K.S., Cheung, S.H., Holland, B., Wang, Y., 2007. Three p-value consistent procedures for multiple comparisons with a control in direction-mixed families. Statistics in Medicine 26, 4253–4272.

Kwong, K.S., Cheung, S.H., Wen, M.J., 2010. Sample size determination in step-up testing procedures for multiple comparisons with a control. Statistics in Medicine 29, 2743–2756.

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

Schwartz, J.I., Vandormael, K., Malice, M.P., Kalyani, R.N., Lasseter, K.C., Holmes, G.B., Gertz, B.T., Gottesdiener, K.M., Laurenzi, M., Redfern, K.J., Brune, K., 2002. Comparison of rofecoxib, celecoxib, and naproxen on renal function in elderly subjects receiving a normal-salt diet. Clinical Pharmacology and Therapeutics 72, 50–61.

Westfall, P.H., 1997. Multiple testing of general contrasts using logical constraints and correlations. Journal of the American Statistical Association 92, 299–306.

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York.