STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:260–273
Published online 7 November 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3462

Comparison of several testing strategies for combination drug efficacy trials based on the closure principle

Julia N. Soulakova*, †

Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall-North, Lincoln, NE 68583-0963, U.S.A.

SUMMARY

The author discusses three multiple testing procedures for identifying the minimum efficacious doses in a balanced factorial combination drug trial. All of these procedures utilize the closed testing principle, and hence strongly control the overall error rate and satisfy the coherence property, that is, if a hypothesis is retained then any hypothesis implied by it is also retained. While coherence is an essential requirement for any multiple testing procedure, consonance is a highly desirable characteristic. In the considered settings if a testing procedure is consonant then it always provides a set of all minimum efficacious combinations as a result, otherwise, it may lead to ambiguity. Although the coherence property is satisfied for any closed testing procedure and thus, does not depend on the test used for an individual hypothesis, whether the considered procedures satisfy the consonance property depends entirely on the nature of the test statistic. The author identifies the consonant and non-consonant procedures among the presented procedures and discusses possible drawbacks of non-consonant procedures with respect to combination drug efficacy trials. Additional properties of these procedures are assessed by simulations. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: average test; consonance; maximum test; minimum effective dose; step-wise procedure

1. INTRODUCTION

In the previous work [1] the authors consider a case when two single drugs are combined and the combinations are investigated with respect to their efficacy. We assume that at least one single drug is known to be effective at the considered doses, and introduce the notion of an efficacious combination as a combination superior to each component. Thus, any efficacious combination possesses two desirable properties: it is effective and it is superior to each component drug. We also propose a closed testing procedure to identify all the minimum efficacious combinations. While this procedure has the great advantage of strong control of overall error rate it may result in ambiguity in terms of the estimated minimum efficacious doses (MeD).

*Correspondence to: Julia N. Soulakova, Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall-North, Lincoln, NE 68583-0963, U.S.A.
†E-mail: jsoulakova2@unlnotes.unl.edu

Received 15 February 2008
Accepted 8 September 2008

The first goal of this paper is to provide further discussion on this procedure. The author shows that ambiguous outcomes are actually a result of non-consonance of this procedure and, though a legitimate estimate is not provided in such a case, there is no contradiction in terms of the testing. The next goal is to propose alternative testing approaches and discuss their advantages. Finally, the author presents the simulation results and compares the performance of the discussed methods.

This paper is organized as follows. First, the author gives an overview of a general statistical framework for a drug combination efficacy study and states the distributional assumptions on the outcome variable. In Section 2 the author discusses several related papers and gives an overview of their significant contribution to the design and analysis of combination drug studies. Sections 3 and 4 describe the proposed procedures and address their properties of coherence and/or consonance. Then, these procedures are illustrated via a simple numerical example in Section 5. Section 6 presents the design of simulation studies performed to evaluate goodness of the procedures. The paper is concluded with a brief discussion on the key properties of the procedures, which is presented in Section 7.

Consider a $K \times N$ combination drug trial, where two drugs, say A and B, are combined at several doses in a factorial fashion. Let i and j be the respective dose levels of drugs A and B, where $i=0,\ldots,K$ and $j=0,\ldots,N$, and $i=0$ and $j=0$ denote the zero doses. Suppose that subjects are randomly allocated (n per group) to one of the $(K+1)\times(N+1)$ groups and subjects of each group receive the (i,j)th drug combination. Next, let μ_{ij} be the population mean response for combination (i,j), and define the expected gain from combining the ith dose of drug A with the jth dose of drug B as $\theta_{ij}=\min\{\mu_{ij}-\mu_{i0},\mu_{ij}-\mu_{0j}\}$, $i=1,\ldots,K$, and $j=1,\ldots,N$. Under the assumption that all $\theta_{ij}\geqslant 0$, if $\theta_{ij}>0$ then the combination is called superior to each component [2]. In addition, assume that at least one component drug is effective at the considered doses; hence, any combination, which is superior to each component is efficacious. In [1] we introduce the notion of an MeD. An MeD is an efficacious combination where decreasing any component dose leads to a non-efficacious compound. Since an MeD is not necessarily unique, we define the MeD set as a collection of all the MeDs.

Let $\overline{y}_{ij}$ be the sample group mean response of the subjects in the (i,j)th combination group. The author assumes that $\overline{y}_{ij}$ are independent and $\overline{y}_{ij}\sim \mathrm{N}(\mu_{ij},\sigma^2/n)$, $i=0,\ldots,K$ and $j=0,\ldots,N$.

2. BACKGROUND

A problem of identifying whether (or not) there exists at least one combination superior to each component is considered in multiple papers; see [2–4]. In [2] this problem is stated as a hypothesis testing problem and two test statistics, 'average' and 'maximum', are proposed, which are respectively given by

$$T_{\mathrm{AVE}}=\mathrm{SUM}\,\hat{\theta}_{ij}/(KN\hat{\sigma}) \tag{1}$$

and

$$T_{\mathrm{MAX}}=\mathrm{MAX}\,\hat{\theta}_{ij}/\hat{\sigma} \tag{2}$$

where $\hat{\theta}_{ij}$, $\hat{\theta}_{ij}=\overline{y}_{ij}-\max(\overline{y}_{i0},\overline{y}_{0j})$ is an estimator of θ_{ij}; $\hat{\sigma}^2$ is the pooled estimator of σ^2 so that $\nu\hat{\sigma}^2/\sigma^2$ has a χ^2 distribution with ν degrees of freedom, and SUM and MAX are the summation and maximum operators, respectively, defined over the lattice $\{(i,j), i=1,\ldots,K, j=1,\ldots,N\}$.

DOI: 10.1002/sim

When the variance is known, that is for the case of $v=\infty$, the authors provide the critical values of the standardized test statistics, $\sqrt{n}T_{\mathrm{AVE}}$ and $\sqrt{n}T_{\mathrm{MAX}}$, for all $K\times N$ designs with $K\leqslant 5$ and $N\leqslant 5$. For convenience, the author cites some of these values in Table II. When the variance is unknown, one can apply methods for deriving the critical values, which are outlined in [2] in detail.

In [1] the authors construct a procedure for identifying the MeD set and outline the details for the 2×2 and 2×3 cases. The procedure can be easily extended to handle other factorial designs. Here the author denotes this procedure as GAVEP and focuses on the 2×3 case, where the MeD sets are given by $\{(1,1)\}$, $\{(1,2)\}$, $\{(1,3)\}$, $\{(2,1)\}$, $\{(2,2)\}$, $\{(2,3)\}$, $\{(1,2),(2,1)\}$, $\{(1,3),(2,1)\}$, $\{(1,3),(2,2)\}$ and the empty set. The corresponding hypotheses are given by the following hypothesis family, which the author denotes by $\Im$:

$$H_0^{(6)}:\theta_{11}=\theta_{12}=\theta_{13}=\theta_{21}=\theta_{22}=\theta_{23}=0,\quad H_0^{(5)}:\theta_{11}=\theta_{12}=\theta_{13}=\theta_{21}=\theta_{22}=0$$

$$H_0^{(4.1)}:\theta_{11}=\theta_{12}=\theta_{13}=\theta_{21}=0,\quad H_0^{(4.2)}:\theta_{11}=\theta_{12}=\theta_{21}=\theta_{22}=0,\quad H_0^{(3.1)}:\theta_{11}=\theta_{12}=\theta_{13}=0$$

$$H_0^{(3.2)}:\theta_{11}=\theta_{12}=\theta_{21}=0,\quad H_0^{(2.1)}:\theta_{11}=\theta_{12}=0,\quad H_0^{(2.2)}:\theta_{11}=\theta_{21}=0\quad\text{and}\quad H_0^{(1)}:\theta_{11}=0$$

Each corresponding alternative hypothesis states that there is at least one efficacious combination among the combinations specified as zero-gain in the null hypothesis. Next, these hypotheses are tested under the closed testing principle, described in detail in [5] in a step-down manner. In order to test an individual hypothesis at level α, the authors use either the 'average' test (1) or its generalization, depending on design. In what follows, a uniform framework is used to represent different types of designs and the corresponding test statistics.

Consider two positive integers K_0 and N_0 such that $1\leqslant K_0\leqslant K$ and $1\leqslant N_0\leqslant N$. Then, let D denote a grid of points

$$D=\{(i,j=j(i)),\ j=1,\ldots,N_i,\ i=1,\ldots,K_0\} \tag{3}$$

where $1\leqslant N_{K_0}\leqslant N_{K_0-1}\leqslant\cdots\leqslant N_1=N_0$. Such a design corresponds to a set of combinations where drug A is taken at K_0 doses but the number of combinations for each fixed dose of drug A may vary, i.e. there are N_i combinations that contain the ith active dose of drug A. Then D is said to represent a $K_0\times N_0$ rectangular (complete) design if $N_{K_0}=N_1$, and a non-rectangular (incomplete) design if $N_{K_0}<N_1$. In the 2×3 case the hypotheses $H_0^{(5)}$, $H_0^{(4.1)}$ and $H_0^{(3.2)}$ correspond to non-rectangular designs and the other hypotheses correspond to the rectangular designs.

Assume that D is given so that $d=\sum_{i=1}^{K_0}N_i$ (the cardinality of the set D) and $P=K_0+N_1$ (the total number of single drug doses) are specified. Let $\hat{\sigma}^2$ be the pooled estimator of σ^2 based on the entire data set. Then, the generalized 'average' test statistic, denoted as GAVE, is given by

$$T_{\mathrm{A}}=\sum_{(i,j)\in D}\hat{\theta}_{ij}/(\hat{\sigma}d) \tag{4}$$

which is equivalent to the AVE test (1) in the case of a rectangular design. This test statistic is equivalent to the one given in [4].

In [1] the authors provide a theorem, which can be used to obtain the α-level critical values for the GAVE test. Table II presents the critical values for the standardized GAVE test, $\sqrt{n}T_{\mathrm{A}}$, for the case of known variance.

DOI: 10.1002/sim

The main advantage of the GAVEP is a strong control of overall error rate. The drawback is that the procedure may result in ambiguities. The author further discusses the issue of ambiguities in Section 4.

A number of testing procedures for identifying the set of all combinations superior to each component are discussed in [6]. Here the author focuses her attention on one approach, which is called 'local MAX test' (loMAX). The problem is stated in terms of the $K \cdot N$ hypotheses, $H_0^{ij}: \theta_{ij} = 0$ and $H_a^{ij}: \theta_{ij} > 0$, $i = 1, \ldots, K$ and $j = 1, \ldots, N$. The test statistics are given by $T_{ij} = \hat{\theta}_{ij}/\hat{\sigma}$, for $i = 1, \ldots, K$ and $j = 1, \ldots, N$, and then are ordered as $T_{(1)} \leqslant T_{(2)} \leqslant \cdots \leqslant T_{(KN)}$. Next, the hypotheses are also ordered as $H_0^{(1)}, \ldots, H_0^{(KN)}$, so that $T_{(s)}$ corresponds to $H_0^{(s)}$, $s = 1, \ldots, KN$. The hypotheses are tested in a step-up manner by comparing the test statistics with a common critical value c_α, where c_α is the critical value corresponding to the MAX test given in (2). In the first step, if $T_{(1)} \geqslant c_\alpha$ then all hypotheses are rejected and the testing is complete. In general, in the tth step, $t = 2, \ldots, KN$, $H_0^{(t)}$ is tested if and only if $H_0^{(s)}$ for $s = 1, \ldots, t-1$ have already been tested and accepted. If $T_{(t)} \geqslant c_\alpha$ then $H_0^{(s)}$ for $s = t, \ldots, KN$ is rejected and the testing is complete. If $T_{(t)} < c_\alpha$ then the $(t+1)$th step is performed. This continues until a hypothesis is rejected or all $K \cdot N$ hypotheses are accepted and then the testing is complete. The indices of rejected and accepted hypotheses identify the superior and non-superior combinations, respectively. The loMAX procedure provides strong control of overall error rate at level α, see [6].

3. STEP-DOWN CLOSED TESTING PROCEDURES BASED ON 'MAXIMUM' TEST

In this section the author considers two closed testing procedures for identifying the MeD set: the step-down procedure based on the 'maximum' test (denoted by GMAXP) and the step-up loMAX procedure extended to find the MeD set (denoted by loMAXP). First, the author outlines the GMAXP, which utilizes the testing scheme of the GAVEP, but is based on the GMAX test statistic

$$T_{\mathrm{B}} = \mathrm{MAX}\, \hat{\theta}_{ij}/\hat{\sigma} \tag{5}$$

where the MAX operator is defined over the corresponding lattice D, given in (3).

In order to obtain critical values for the GMAX test (5), one can apply the following result. The main steps of the proof are outlined in Appendix A.

Result: Define $G(c) = 1 - \int_0^\infty \prod_{s=1}^P E[\{\Phi(\sqrt{n}cw + Z)\}^{m_s}]\, \mathrm{d}Q(w)$, where $\Phi(\cdot)$ is the distribution function of the standard normal random variable; Z is the standard normal random variable; $Q(\cdot)$ is the distribution function of $\hat{\sigma}/\sigma$, E is the expectation operator and m_s, $s = 1, \ldots, P$, are non-negative integers subject to constraints:

$$m_s \leqslant N_s, \quad s = 1, \ldots, K_0, \quad m_s \leqslant d_{s-K_0}, \quad s = K_0 + 1, \ldots, P \quad \text{and} \quad \sum_{s=1}^{P} m_s = d \tag{6}$$

where d_{s-K_0}, $s = K_0 + 1, \ldots, P$, denotes the cardinality of the set $\{(i, s-K_0) : (i, s-K_0) \in D\}$, i.e. the number of combinations where drug B is taken at dose level $s - K_0$; thus, $d_1 = K_0$ and $d_{N_1} = N_1$. Then, the α-level critical value of the GMAX test is given as the solution c to the equation $G_{\max}(c) = \alpha$, where $G_{\max}(c)$ is the maximum value of $G(c)$ subject to the constraints (6).

In order to use the preceding result, one first needs to solve the optimization problem of maximizing $G(c)$ subject to specific constraints on the values of m_s, $s = 1, \ldots, P$. Appendix B

DOI: 10.1002/sim

illustrates the solution to this problem. Table II presents the critical values for the standardized GMAX test for the case of known variance for the 2×3 case.

The second testing procedure that the author considers to identify the MeD set is loMAXP. Since under the assumption of at least one effective component drug a combination superior to each component is efficacious, the author first applies the loMAX procedure to detect all efficacious combinations and then selects the 'lowest' combinations among these efficacious combinations. Hence, the estimated MeD set is a collection of all combinations (r, s), such that H_0^{rs} is rejected and H_0^{ij}, H_0^{is} and H_0^{rj} are accepted for all $i < r$ and $j < s$.

4. COHERENCE VERSUS CONSONANCE: POSSIBLE AMBIGUITIES OF CLOSED TESTING PROCEDURES

There are two important properties that characterize the performance of a multiple testing procedure: coherence and consonance. Both of these properties are introduced in [7] and are discussed in [8] in detail. If a procedure utilizes the closed testing principle then it satisfies the coherence property. That is, if a hypothesis is retained then any hypothesis implied by it is also retained. Both GAVEP and GMAXP are coherent. The other important property is consonance. Given a hypothesis family, consider a null hypothesis, which is an intersection of a set of null hypotheses, referred to as component hypotheses. Then, such an intersection hypothesis is called non-minimal hypothesis. The property of consonance is stated in [8, p. 46] as: whenever any non-minimal null hypothesis is rejected, at least one of its components is also rejected. The authors also mention that 'while coherence is an essential requirement, consonance is only a desirable property'. In addition to that non-consonance does not imply any logical contradictions as non-coherence does. Indeed, failure to reject a hypothesis only indicates a lack of sufficient evidence and does not necessarily prove that the null hypothesis is true.

As it is mentioned in [1], the GAVEP may result in ambiguities in terms of the estimated MeD set. In what follows, the author discusses all (two) types of ambiguities that may occur as a result of non-consonance in the 2×2 and 2×3 cases. In the cases with higher dimensions, other types of ambiguities may be encountered. For example, there is one more type of ambiguity, which may happen in the 3×3 case. In what follows, the author also shows that the GMAXP and loMAXP are both consonant.

The first type of ambiguity, denoted as Type A, is defined as a case when both hypotheses of a certain level are tested and accepted, where a hypothesis 'level' is the number of zero gains specified by this hypothesis. Table I presents all situations that result in Type A ambiguity in the 2×3 case. For example, a Type A ambiguity occurs if $H_0^{(6)}$ and $H_0^{(5)}$ are rejected and $H_0^{(4.1)}$ and $H_0^{(4.2)}$ are both accepted. In such a case one can only conclude that there is at least one efficacious combination among $(1, 1)$, $(1, 2)$, $(1, 3)$, $(2, 1)$ and $(2, 2)$, but the estimated MeD set remains unknown.

While the GAVEP may result in such an ambiguity, the GMAXP never results, because it satisfies the consonance property. The author illustrates this property for the 2×3 case. First, note that the GMAX test statistic corresponding to the intersection hypothesis is equal to at least one of two GMAX test statistics corresponding to the directly implied hypotheses. Moreover, the critical value corresponding to these two individual hypotheses is smaller than the critical value corresponding to the intersection hypothesis (see Table II). Hence, at least one of the directly implied

DOI: 10.1002/sim

Table I. Ambiguities of GAVEP in the 2×3 case.

Ambiguity	$H_0^{(6)}$	$H_0^{(5)}$	$H_0^{(4.1)}$	$H_0^{(4.2)}$	$H_0^{(3.1)}$	$H_0^{(3.2)}$	$H_0^{(2.1)}$	$H_0^{(2.2)}$	$H_0^{(1)}$
Type A	REJ	REJ	ACC	ACC	NT	NT	NT	NT	NT
Type A	REJ	REJ	REJ	REJ	ACC	ACC	NT	NT	NT
Type A	REJ	REJ	REJ	REJ	REJ	REJ	ACC	ACC	NT
Type B	REJ	REJ	REJ	ACC	ACC	NT	NT	NT	NT
Type B	REJ	REJ	REJ	REJ	ACC	REJ	NT	ACC	NT

Note: ACC = Accepted, REJ = Rejected, NT = Not tested.

Table II. Critical values for standardized GMAX and GAVE tests when variance is known, i.e. $v = \infty$. Critical values for complete designs are taken from Hung *et al.* [2].

Significance level	$H_0^{(6)}$	$H_0^{(5)}$	$H_0^{(4.1)}$	$H_0^{(4.2)}$	$H_0^{(3.1)}$	$H_0^{(3.2)}$	$H_0^{(2.1)}, H_0^{(2.2)}$	$H_0^{(1)}$
0.10	1.05	1.09	1.20	1.11	1.48	1.21	1.57	1.81
	2.97	2.88	2.75	2.75	2.57	2.57	2.31	1.81
0.05	1.34	1.40	1.54	1.42	1.90	1.55	2.01	2.33
	3.36	3.28	3.16	3.16	3.00	3.00	2.76	2.33
0.01	1.90	1.97	2.18	2.01	2.69	2.19	2.85	3.29
	4.14	4.07	3.97	3.97	3.84	3.84	3.64	3.29

Note: The upper and the lower entries correspond to the AVE and MAX tests, respectively.

hypotheses is rejected by the GMAX test, and therefore, the GMAXP never results in Type A ambiguity.

The other type of ambiguity, Type B, happens only in the 2×3 case (see Table I) and cases of higher dimensions. The author defines the Type B ambiguity in terms of the 2×3 case as follows. Suppose that two hypotheses of the same level are tested, one is accepted and the other is rejected and the hypothesis directly implied by the rejected one (of the lower level) is tested. Then, if the latter hypothesis is accepted then Type B ambiguity occurs. For example, if $H_0^{(6)}$, $H_0^{(5)}$ and $H_0^{(4.1)}$ are rejected and $H_0^{(4.2)}$ is accepted then $H_0^{(3.1)}$ is tested. If $H_0^{(3.1)}$ is accepted, then the procedure does not result in any meaningful MeD set estimate. Indeed, rejection of $H_0^{(6)}$, $H_0^{(5)}$ and $H_0^{(4.1)}$ suggests that there is at least one efficacious combination among $(1,1)$, $(1,2)$, $(1,3)$ and $(2,1)$ but the MeD set cannot be identified.

In order to show that the GMAXP does not result in Type B ambiguity, suppose that two hypotheses of the same level are tested, one is rejected and the other is accepted; these hypotheses may be given by $H_0^{(4.1)}$ and $H_0^{(4.2)}$ or $H_0^{(3.2)}$ and $H_0^{(3.1)}$, respectively. Hence, the hypothesis directly implied by the rejected intersection hypothesis $H_0^{(4.1)}$ ($H_0^{(3.2)}$) is tested, that is $H_0^{(3.1)}$($H_0^{(2.2)}$). The author will show that the GMAX test rejects the implied hypothesis. Since the GMAX test rejects the intersection hypothesis and accepts the other hypothesis of the same level, the GMAX test statistics corresponding to the intersection hypothesis and the hypothesis directly implied by it are ought to be equal. Moreover, the critical value for testing the intersection hypothesis is larger than the one corresponding to the directly implied hypothesis; hence, the GMAX test rejects the latter hypothesis. Therefore, Type B ambiguity does not happen.

DOI: 10.1002/sim

As it is mentioned in [1], when the GAVEP is applied, the probability of ambiguity depends on many parameters including the population gains and the true MeD set. In Sections 6 and 7 the author reports the estimated probability of Type A and Type B ambiguity and discusses the population parameters settings when the ambiguities are the most common.

Next, the author discusses the loMAXP with respect to the consonance and coherence. Although the original procedure based on the loMAX test and presented in Section 2 is stated only in terms of the $K \cdot N$ minimal hypotheses, it also utilizes a hierarchical hypothesis family. The hypothesis family in addition to the $K \cdot N$ minimal hypotheses includes all possible intersection hypotheses. The procedure tests the minimal hypotheses in the direction of the increasing magnitude and if a minimal hypothesis is rejected, all intersection hypotheses implying it are also rejected. This is why a shortcut version presented in Section 2 is feasible. Obviously, such a procedure is coherent and consonant and thus, the loMAXP is also coherent and consonant.

5. ANTIHYPERTENSIVE DRUGS EXAMPLE

In [1] we consider an example, which was first presented in [2]. The data are the summary statistics of the observed diastolic blood pressure mean reductions in a 2×3 factorial clinical trial, conducted to evaluate the effectiveness of the combination of two active antihypertensive drugs. The estimated gains are given by $\hat{\theta}_{11}=4$, $\hat{\theta}_{12}=2$, $\hat{\theta}_{13}=3$, $\hat{\theta}_{21}=\hat{\theta}_{22}=1$ and $\hat{\theta}_{23}=2$, the pooled estimate for variance is $\hat{\sigma}^2=42$ and the same number of patients, $n=25$, is allocated to each group. In [1] we show that at $\alpha=0.01$, 0.05 and 0.10 the GAVEP identifies the MeD set as empty, $\{(1,2)\}$ and $\{(1,1)\}$, respectively.

Now the author applies the GMAXP to these data. Since the maximum gain corresponds to the combination (1, 1), all the GMAX test statistics (5) are equal to $4/\sqrt{42}=0.62$. The corresponding critical values are the ones given in Table II, divided by $\sqrt{25}$. First, $H_0^{(6)}$ is tested by comparing 0.62 with the corresponding critical value of 0.83, 0.67 or 0.59, depending on the level of significance $\alpha=0.01$, 0.05 or 0.10, respectively. Hence, at $\alpha=0.01$ and 0.05 the hypothesis $H_0^{(6)}$ is accepted and the MeD set is estimated as empty. At $\alpha=0.10$ the hypothesis $H_0^{(6)}$ is rejected and $H_0^{(5)}$ is tested. Note that the critical values for the GMAXP become smaller when the procedure 'steps down' and the test statistics are all equal to 0.62; hence, the other null hypotheses are also rejected at $\alpha=0.10$ and the MeD set is estimated as $\{(1,1)\}$.

If the loMAXP is applied, then the test statistics are given by $T_{11}=0.62$, $T_{12}=0.31$, $T_{13}=0.46$, $T_{21}=T_{22}=0.15$ and $T_{23}=0.31$. These statistics are then compared with a common critical value of $c_{0.01}=0.83$, $c_{0.05}=0.67$ or $c_{0.10}=0.59$ depending on a significance level. As a result, this procedure results in the same MeD sets as the GMAXP for the considered levels of significance.

This example illustrates that the procedures based on GAVE, GMAX and loMAX tests may result in different MeD set estimates: at $\alpha=0.05$ the GAVEP estimates the MeD set as $\{(1,2)\}$, while the GMAXP and loMAXP declare the MED set to be empty.

6. SIMULATION STUDIES

In this section the author provides the results of the simulation studies for the 2×3 case. The simulation configurations are similar to the ones used in [6] but are extended to the 2×3 case. As

DOI: 10.1002/sim

is shown in [1, 2], the power function of a single GAVE test increases with the true average gain and the power functions of single GMAX and loMAX tests increase with each true gain, when the rest of the gains are fixed. Also, the power of each test depends on the nuisance parameters δ_{ij}'s, called in this paper standardized control mean differences, where $\delta_{ij}=(\mu_{i0}-\mu_{0j})/\sigma$ for $i=1,2$ and $j=1,2,3$. The author first discusses the performance of the procedures for different true MeD set configurations, true gains and different values of δ_{ij}'s. The author identifies the MeD set configurations, which are correctly estimated by the procedures with the highest and lowest average power across different values of the true gains when δ_{ij}'s are fixed. Also, two settings of δ_{ij}'s are considered for the fixed true gains and results of the procedures are compared. Next, the author examines the relationship between the overall mean power and the population gains across all MeD set configurations. Finally, the estimated probabilities of ambiguities are addressed.

Throughout, the significance level is fixed at $\alpha=0.05$, the standard deviation is taken to be $\sigma=1$ and the common sample size, $n=30$ per group, is considered. The critical values are the ones given in Table II, divided by $\sqrt{30}$.

Among population mean configurations the author considers two settings, C_1 and C_2, given by

$$C_1=\begin{pmatrix} \times & 0.2 & 0.4 & 0.6 \\ 0.4 & 0.4+\Delta_{11} & 0.4+\Delta_{12} & 0.6+\Delta_{13} \\ 0.8 & 0.8+\Delta_{21} & 0.8+\Delta_{22} & 0.8+\Delta_{23} \end{pmatrix} \quad \text{and}$$

$$C_2=\begin{pmatrix} \times & 0.2 & 0.4 & 0.6 \\ 0.8 & 0.8+\Delta_{11} & 0.8+\Delta_{12} & 0.8+\Delta_{13} \\ 1.0 & 1.0+\Delta_{21} & 1.0+\Delta_{22} & 1.0+\Delta_{23} \end{pmatrix}$$

where $\Delta_{ij}\in\{0,\Delta\}$ for $i=1,2$, $j=1,2,3$ and $\Delta\in\{0.4,0.6,\ldots,1.2\}$. Both of these settings result in $\mu_{ij}=\max\{\mu_{i0},\mu_{0j}\}+\Delta_{ij}$ for $i=1,2$, $j=1,2,3$. Moreover, let $\boldsymbol{\delta}$ denote a vector with components δ_{ij}'s, $i=1,2$ and $j=1,2,3$. Then, C_1 and C_2 correspond to $\boldsymbol{\delta}_1=(0.2,0.0,-0.2,0.6,0.4,0.2)$ and $\boldsymbol{\delta}_2=(0.6,0.4,0.2,0.8,0.6,0.4)$, respectively. Thus, the second setting corresponds to the larger δ_{ij}'s.

Next, for each value of Δ all possible configurations of the MeD sets are considered for each of the two settings. Given the MeD set, if a combination (i,j) is non-efficacious then $\Delta_{ij}=0$, otherwise $\Delta_{ij}=\Delta$. Such a classification into two groups of efficacious and non-efficacious combinations is possible only because the author restricts attention to the isotonic population gains: if a combination (r,s), $r=1,2$, $s=1,2,3$, is efficacious then all combinations (i,j) with $i\geqslant r$ and $j\geqslant s$ are also efficacious. For example, if $\Delta=1.0$, the MeD set is $\{(1,3)\}$ then in the case of C_1 the population gains are $\theta_{11}=\theta_{12}=\theta_{21}=\theta_{22}=0$ and $\theta_{13}=\theta_{23}=1.0$ and hence, the true means are $\mu_{11}=\mu_{12}=0.4$, $\mu_{13}=1.8$, $\mu_{21}=\mu_{22}=0.8$ and $\mu_{23}=2.0$. A special case of $\Delta_{ij}=0$ for all i and j corresponds to an empty MeD set. Thus, there are 91 simulation configurations in total, including a configuration corresponding to an empty MeD set.

Next, for the specified μ_{ij}'s, the group sample means $\overline{y}_{ij}$'s are generated, such that $\overline{y}_{ij}$'s are independent and $\overline{y}_{ij}\sim N(\mu_{ij},1/30)$. Then, the GAVEP, GMAXP and loMAXP are applied to

DOI: 10.1002/sim

these same data. For each simulation configuration, the result of each procedure is noted and the preceding steps are replicated 100 000 times, leading to the maximum standard error of 0.002 for the performance of characteristics of interest.

To assess the performance of the procedures in terms of family-wise error rate (FWE), power and lack of power (LOP) the author applies the definitions similar to the ones given in [1, 9]. Defined in such a way FWE, power and LOP do not necessarily add up to 1, but for simplicity of presentation the author reports only estimated FWE and power.

All procedures are shown to perform very well when the true MeD set is empty. The estimated power of the GAVEP and GMAXP (loMAXP) is equal to 0.987 and 0.976, respectively. The FWE is 0.013 and 0.024 for the GAVEP and GMAXP (loMAXP), respectively.

The simulation study also confirms the strong control of FWE at $\alpha=0.05$. The maximum simulation FWE's across all configurations are 0.029, 0.022 and 0.036 in the case of $\boldsymbol{\delta}=\boldsymbol{\delta}_1$ and are 0.049, 0.033 and 0.050 in the case of $\boldsymbol{\delta}=\boldsymbol{\delta}_2$ for the GAVEP, loMAXP and GMAXP, respectively.

When the true MeD set is not empty then the results of the procedures depend on the particular configuration of the MeD set, $\boldsymbol{\delta}$ and Δ. Given the MeD set and $\boldsymbol{\delta}$, the power of each procedure increases in Δ. Thus, the power results are given in terms of the summaries: minimum, maximum and average. Table III contains the power results for two settings of $\boldsymbol{\delta}$ and each configuration of the true MeD set. The average power is calculated as the mean value of power across all Δ's. Clearly, the GMAXP and loMAXP dominate the GAVEP for all Δ's and all configurations of the MeD set except for the cases when the MeD set is given by $\{(1,1)\}$. Among different MeD set configurations the procedures estimate the MeD set $\{(1,1)\}$ with the highest average power. The GAVEP estimates the sets containing the highest combination(s) with the lowest average power. On average, the GMAXP and loMAXP perform better in the cases when the MeD is unique. If the performances of the procedures for $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are compared, then the mean increases in power across all MeD set configurations, are 0.060, 0.009 and 0.016, for the GAVEP, GMAXP and loMAXP, respectively.

Figure 1 illustrates the relationship between the average power and the gain Δ for each procedure in the case of C_1, i.e. when $\boldsymbol{\delta}=\boldsymbol{\delta}_1$. The average power is calculated as the mean power across all configurations of the MeD set for a given Δ. Clearly, the GMAXP dominates the other procedures when $\Delta\leqslant 1.0$. In the case of C_2 ($\boldsymbol{\delta}=\boldsymbol{\delta}_2$), the mean power graphs are similar to the ones given in Figure 1 but shifted upward; hence, the author does not report them here.

Finally, the author discusses the performance of the GAVEP in terms of possible ambiguities. When the true MeD set is empty, the GAVEP rarely results in ambiguity; the probability is 0.002 for each type of ambiguity in such a case. Table IV presents the simulation results for configurations with non-empty MeD sets. The procedure results in the higher probability of Type A ambiguity, p_{A}, when the MeD set contains two elements when compared with the single-element MeD sets. Among those, the sets $\{(2,2),(1,3)\}$ and $\{(2,1),(1,3)\}$ correspond to the largest mean p_{A} for both settings of $\boldsymbol{\delta}$. The simulations indicate that p_{A} can be as large as 0.418. There is a slight decrease in the mean p_{A} when δ_{ij}'s increase: 0.102 when $\boldsymbol{\delta}=\boldsymbol{\delta}_1$ and 0.086 when $\boldsymbol{\delta}=\boldsymbol{\delta}_2$. The estimated means of Type B ambiguity probability p_{B} are similar for two settings of $\boldsymbol{\delta}$. The largest estimated p_{B} is 0.152 (0.134) when $\boldsymbol{\delta}=\boldsymbol{\delta}_1$ ($\boldsymbol{\delta}=\boldsymbol{\delta}_2$) and corresponds to the MeD set $\{(2,1),(1,2)\}$ in both cases C_1 and C_2.

The individual dependencies of p_{A} and p_{B} on gain Δ are not straightforward. The mean p_{A} calculated across all configurations of non-empty MeD sets for a given Δ increases with Δ for $\Delta\leqslant 0.8$ and decreases with Δ for $\Delta>0.8$; hence, the maximum mean p_{A} occurs at $\Delta=0.8$. This holds for both settings of $\boldsymbol{\delta}$. Calculated in the same way the mean p_{B} tends to increase with Δ

DOI: 10.1002/sim

Table III. Simulation power in the 2×3 case.

MeD set								
(2, 3)	(2, 2)	(1, 3)	(2, 1)	(2, 2) (1, 3)	(1, 2)	(2, 1) (1, 3)	(2, 1) (1, 2)	(1, 1)
Min–max power for $\boldsymbol{\delta}=\boldsymbol{\delta}_1$ *(Case of* C_1*)*								
0.036–0.258	0.036–0.332	0.011–0.270	0.029–0.424	0.008–0.309	0.040–0.642	0.012–0.441	0.025–0.732	0.232–0.999
0.147–0.962	0.185–0.966	0.154–0.962	0.210–0.976	0.042–0.957	0.097–0.954	0.055–0.970	0.045–0.961	0.205–0.998
0.148–0.965	0.185–0.973	0.153–0.969	0.198–0.982	0.029–0.963	0.088–0.968	0.031–0.971	0.018–0.964	0.154–0.988
Average power for $\boldsymbol{\delta}=\boldsymbol{\delta}_1$ *(Case of* C_1*)*								
0.131	0.162	0.110	0.194	0.122	0.309	0.180	0.338	0.756
0.628	0.657	0.636	0.679	0.545	0.597	0.575	0.567	0.732
0.631	0.657	0.635	0.668	0.519	0.574	0.525	0.488	0.647
Min–max power for $\boldsymbol{\delta}=\boldsymbol{\delta}_2$ *(Case of* C_2*)*								
0.061–0.328	0.066–0.403	0.033–0.348	0.065–0.494	0.027–0.390	0.111–0.707	0.034–0.511	0.059–0.752	0.335–0.999
0.177–0.950	0.192–0.951	0.147–0.946	0.213–0.958	0.042–0.941	0.192–0.943	0.052–0.946	0.081–0.945	0.279–0.998
0.179–0.956	0.192–0.962	0.148–0.959	0.201–0.971	0.029–0.955	0.186–0.972	0.031–0.963	0.039–0.970	0.204–0.988
Average power for $\boldsymbol{\delta}=\boldsymbol{\delta}_2$ *(Case of* C_2*)*								
0.182	0.219	0.169	0.262	0.182	0.401	0.245	0.390	0.788
0.640	0.653	0.624	0.669	0.535	0.662	0.558	0.597	0.759
0.644	0.655	0.626	0.662	0.515	0.657	0.520	0.534	0.674

Note: The upper, middle and bottom entries correspond to the GAVEP, GMAXP and loMAXP, respectively. Min (minimum) and max (maximum) power correspond to $\Delta = 0.4$ and $\Delta = 1.2$, respectively; average power is calculated across all values of Δ.

DOI: 10.1002/sim

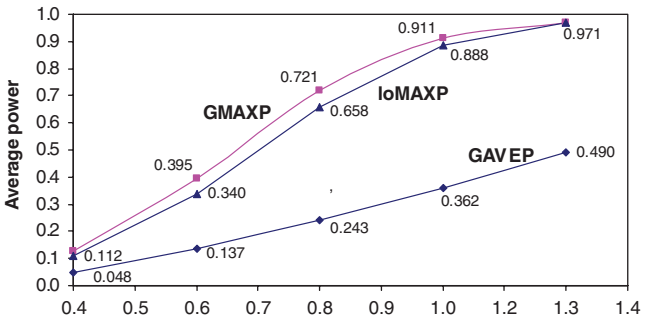

Figure 1. Average power as a function of gain Δ for each procedure.

Table IV. Simulation ambiguity probability in the 2×3 case.

	MeD set									Average	
Value of gain Δ	(2, 3)	(2, 2)	(1, 3)	(2, 1)	(2, 2) (1, 3)	(1, 2)	(2, 1) (1, 3)	(2, 1) (1, 2)	(1, 1)	$\boldsymbol{\delta}=\boldsymbol{\delta}_1$	$\boldsymbol{\delta}=\boldsymbol{\delta}_2$
$\boldsymbol{\delta}=\boldsymbol{\delta}_1$											
0.4	0.002	0.007	0.016	0.014	0.074	0.105	0.125	0.144	0.069	0.062	0.058
	0.002	0.002	0.010	0.006	0.007	0.022	0.036	0.113	0.052	0.028	0.030
0.6	0.002	0.011	0.027	0.018	0.194	0.198	0.298	0.249	0.026	0.114	0.097
	0.002	0.001	0.019	0.007	0.009	0.019	0.065	0.152	0.013	0.032	0.030
0.8	0.002	0.011	0.033	0.016	0.335	0.173	0.392	0.226	0.002	0.132	0.108
	0.001	0.001	0.029	0.005	0.008	0.006	0.061	0.086	0.001	0.022	0.021
1.0	0.002	0.013	0.034	0.011	0.418	0.088	0.354	0.132	0.000	0.117	0.094
	0.001	0.001	0.036	0.003	0.007	0.001	0.031	0.024	0.000	0.012	0.011
1.2	0.002	0.012	0.030	0.007	0.397	0.037	0.248	0.055	0.000	0.087	0.071
	0.001	0.001	0.034	0.001	0.004	0.000	0.010	0.005	0.000	0.006	0.006
Average	0.002	0.011	0.028	0.013	0.284	0.120	0.283	0.161	0.019	0.102	
	0.001	0.001	0.026	0.004	0.007	0.010	0.041	0.076	0.013	0.020	
$\boldsymbol{\delta}=\boldsymbol{\delta}_2$											
Average	0.004	0.012	0.032	0.008	0.248	0.095	0.234	0.126	0.011		0.086
	0.004	0.005	0.023	0.010	0.006	0.009	0.034	0.074	0.012		0.020

Note: The upper and lower entries correspond to the probability of Type A and Type B ambiguity, respectively.

for $\Delta \leqslant 0.6$ and then decreases with Δ for $\Delta > 0.6$; hence, the largest Δ, which corresponds to the maximum p_B is 0.6.

There are some additional conclusions, which can be stated in terms of the total probability of ambiguities, $p_A + p_B$. Figure 2 illustrates the mean total probability computed as an average of total probabilities across all configurations of the MeD set for a given gain, as a function of gain. Clearly, there are larger average total probabilities in the case of $\boldsymbol{\delta}=\boldsymbol{\delta}_1$ for all relatively large gains. The relationships are somewhat similar with the maximum at around $\Delta = 0.7$.

DOI: 10.1002/sim

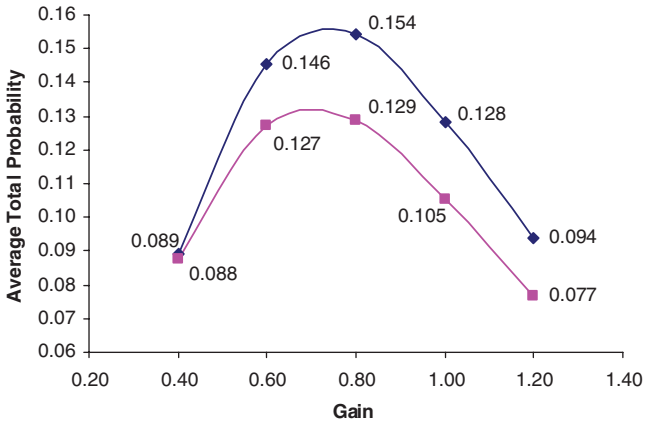

Figure 2. Average total probability of ambiguities as a function of gain Δ. The upper and lower graphs correspond to settings C_1 and C_2, respectively.

7. DISCUSSION

The author compares three closed testing procedures for identifying the MeD set. Two of these procedures, GAVEP and GMAXP, are step-down procedures based on the same closed hypothesis family and partially ordered multiple tests. The third procedure, loMAXP, is a step-up procedure, which involves multiple tests based on a common critical value.

The performance of each procedure depends on a number of population parameters, such as the true gains, the standardized control mean differences and the configuration of the true MeD set. These parameters have quite complicated either matrix-like or set-like structure. Moreover, the theoretical derivations corresponding to the multiple test procedures are problematic; hence, the investigation of the properties of the procedures are very complex. The author applies the theoretical concepts related to the single tests and simulation study results to assess the performance of these procedures.

The main advantage of the procedures is the strong control of the overall error rate. The power of each procedure increases with increased true gains. The power also depends on the true standardized control mean differences, but it is hard to predict what configurations of the MeD set will be identified with the greater or lower power if the true standardized control mean differences are not known. In the settings the author considers, power of the GAVEP increases with increased true standardized control mean differences for all MeD set configurations. The results of the other two procedures, in addition, depend on the MeD set configuration: the procedures can result in either higher or lower power when the true standardized control mean differences are fixed for different configurations of the true MeD set.

On average, the GMAXP either dominates the other procedures or results in a great power itself and hence, is recommended for use for the majority of settings. Although there are cases when the GAVEP dominates the other procedures, it has a serious drawback of possible ambiguities as a result of non-consonance. In terms of the simplicity of implementation, the loMAXP is the best one, because it involves fewer hypotheses and uses a common critical value.

DOI: 10.1002/sim

APPENDIX A

In order to obtain the distribution function of the GMAX test statistic (5) and prove the result, one can generalize the method introduced in [2] for obtaining the critical values for the MAX test (2). Here the author presents only final results. Let D be the specified design of the form (3). Then using the notation introduced in Sections 2 and 3, it can be shown that the power function of the GMAX test statistic is given by

$$P(T_{\mathrm{B}}>c|\boldsymbol{\theta})=1-\int_0^{\infty}\int_{R^P}\prod_{i=1}^{K_0}\prod_{j=1}^{N_i}\Phi(g_{ij})\cdot\prod_{i=1}^{K_0}\varphi(u_i)\prod_{j=1}^{N_1}\varphi(v_j)\,\mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}\,\mathrm{d}Q(w)$$

where $\boldsymbol{\theta}$ is a vector with components θ_{ij}'s, for $(i,j)\in D$, $g_{ij}=\sqrt{n}(cw-\theta_{ij}/\sigma)+0.5(u_i+v_j-\sqrt{n}|\delta_{ij}|+|u_i-v_j+\sqrt{n}\delta_{ij}|)$, $\delta_{ij}=(\mu_{i0}-\mu_{0j})/\sigma$, for $(i,j)\in D$, $\mathrm{d}\mathbf{u}=\mathrm{d}u_1\ldots\mathrm{d}u_{K_0}$, and $\mathrm{d}\mathbf{v}=\mathrm{d}v_1\ldots\mathrm{d}v_{N_1}$. This power function increases in each θ_{ij} when the remaining components of $\boldsymbol{\theta}$ are held fixed. Let $P_{\boldsymbol{\theta}=\mathbf{0}}$ denote the power function evaluated at $\boldsymbol{\theta}=\mathbf{0}$, then it can be shown that

$$\text{Significance level}=\max\left\{\lim_{|\delta_{ij}|\to\infty}P_{\boldsymbol{\theta}=\mathbf{0}},(i,j)\in D\right\}$$

Next, $\lim_{|\delta_{ij}|\to\infty}P_{\boldsymbol{\theta}=\mathbf{0}}=1-\int_0^{\infty}\int_{R^P}\prod_{(i,j)\in D}\{\Phi(\sqrt{n}cw+u_i)\}^{\pi_{ij}}\{\Phi(\sqrt{n}cw+v_j)\}^{1-\pi_{ij}}\prod_{i=1}^{K_0}\varphi(u_i)\prod_{j=1}^{N_1}\varphi(v_j)\,\mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}\,\mathrm{d}Q(w)$, where $\pi_{ij}=1$, if $\delta_{ij}\to\infty$ and $\pi_{ij}=0$ if $\delta_{ij}\to-\infty$. Hence, $\lim_{|\delta_{ij}|\to\infty}P_{\boldsymbol{\theta}=\mathbf{0}}=1-\int_0^{\infty}\int_{R^P}\prod_{s=1}^{P}\{\Phi(\sqrt{n}cw+z_s)\}^{m_s}\prod_{s=1}^{P}\varphi(z_s)\,\mathrm{d}\mathbf{z}\,\mathrm{d}Q(w)$, where $m_s=\sum_{j:(s,j)\in D}\pi_{sj}$ and $z_s=u_s$, if $s=1,\ldots,K_0$, $m_s=\sum_{i:(i,s)\in D}(1-\pi_{i,s-K_0})$ and $z_s=v_{s-K_0}$, if $s=K_0+1,\ldots,P$, $\mathrm{d}\mathbf{z}=\mathrm{d}z_1\ldots\mathrm{d}z_P$. The last limit can be reduced to $G(c)=1-\int_0^{\infty}\prod_{s=1}^{P}E[\{\Phi(\sqrt{n}cw+Z)\}^{m_s}]\,\mathrm{d}Q(w)$, and the significance level can be obtained as the maximum value of $G(c)$, subject to the constraints (6).

APPENDIX B

Let $\mathbf{m}^*=(m_1^*,m_2^*,\ldots,m_P^*)$ denote a vector of m_s's such that $G(c)$ is maximized subject to the constraints (6) at $\mathbf{m}^*$. Then, similar to [2] it can be shown that such an $\mathbf{m}^*$ exists and it satisfies the property $\max\{m_s^*\}-\min\{m_s^*\}=1$ or 0, where $s=1,\ldots,P$. In order to obtain $\mathbf{m}^*$ the author applies the fact that for any integers $r,s\geqslant 1$ there exist two non-negative integers, q_1 and q_2, such that $rs=q_1(r+s)+q_2$, $q_2<r+s$, where r and s represent the dimensions of a rectangular design [2]. And then $\mathbf{m}^*$ is obtained by taking $m_1^*=m_2^*=\cdots=m_{q_2}^*=q_1+1$ and $m_{q_2+1}^*=m_{q_2+2}^*=\cdots=m_{r+s}^*=q_1$. To apply these results in the author's settings, first the author solves a system of equations $d=rs$ and $P=r+s$, for r and s. If obtained in such a way, r and s are positive integers then the problem reduces to the problem of identifying the critical values for the $r\times s$ rectangular design [2]. Otherwise, the author needs to solve the equation $d=q_1P+q_2$ for q_1 and q_2 ($q_2<P$) directly, then obtain $\mathbf{m}^*$ and finally, solve $G_{\max}(c)=\alpha$ for c. In what follows, the author considers all non-rectangular designs, which are used in the 2×2 and 2×3 cases.

First, consider $D_1=\{(1,1),(1,2),(2,1)\}$. Then $d=3$ and $P=4$, so that $r=1$, $s=3$ and the critical values for such a design are identical to the ones for the 1×3 rectangular case. Next, if $D_2=\{(1,1),(1,2),(1,3),(2,1)\}$, then $d=4$, $P=5$ so $r=1$ and $s=4$ and the problem reduces to the one for 1×4 rectangular design. Finally, consider $D_3=\{(1,1),(1,2),(1,3),(2,1),(2,2)\}$,

DOI: 10.1002/sim

so $d=P=5$. Since no positive integers r and s exist to satisfy both, $5=rs$ and $5=r+s$, the author uses $5=5q_1+q_2$ to obtain $q_1=1$ and $q_2=0$. This means that $m_1^*=m_2^*=\cdots=m_5^*=1$. Note that in the 1×5 case the corresponding equation $5=6q_1+q_2$ has the solution $q_1=0$ and $q_2=5$, which results in the same $\mathbf{m}^*=\mathbf{1}$. Although, in the 2×3 for all non-rectangular designs the author identified r and s, such that an $r\times s$ design would result in the same solution $\mathbf{m}^*$, which is not always possible. For example, if the highest combination $(2,4)$ is excluded from a 2×4 rectangular design, then for such a non-rectangular design $d=7$, $P=6$, $q_1=q_2=1$ and $\mathbf{m}^*=(2,1,1,1,1,1)$. Hence, to find the α-level critical value one will need to solve $\alpha=1-\int_0^\infty[E\{\Phi(\sqrt{n}cw+Z)\}^2][E\{\Phi(\sqrt{n}cw+Z)\}]^5\,\mathrm{d}Q(w)$ for c.

ACKNOWLEDGEMENTS

I wish to thank both referees, the Associate Editor and the Editor for their insightful comments and suggestions that greatly improved this article.

REFERENCES

1. Soulakova JN, Sampson AR. On identifying minimum efficacious doses in combination drug trials. *Statistics in Biopharmaceutical Research* 2008. Pre-Publication paper. Available from: http://www.amstat.org/publications/sbr/index.cfm?fuseaction=main.
2. Hung HM, Chi JY, Lipicky RJ. Testing for the existence of a desirable dose combination. *Biometrics* 1993; **49**:85–94.
3. Hung HM. Global tests for combination drug studies in factorial trials. *Statistics in Medicine* 1996; **15**:233–247.
4. Hung HM. Global tests for combination drug studies in factorial trials. *Statistics in Medicine* 2000; **19**:2079–2087.
5. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
6. Buchheister B, Lehmacher W. Multiple testing procedures for identifying desirable dose combinations in bifactorial designs. *GMS Medical Informatics*, *Biometry and Epidemiology* 2006; **2**(2):1–11. Available from: http://www.egms.de/pdf/journals/mibe/2006-2/mibe000026.pdf.
7. Gabriel KR. Simultaneous test procedures—some theory of multiple comparisons. *Annals of Mathematical Statistics* 1969; **40**:224–250.
8. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
9. Hellmich M, Lehmacher W. Closure procedures for monotone bi-factorial dose-response designs. *Biometrics* 2005; **61**:269–276.

DOI: 10.1002/sim