A STEP-UP TEST PROCEDURE TO FIND THE MINIMUM EFFECTIVE DOSE

Weizhen Wang¹ and Jianan Peng²

¹Department of Mathematics and Statistics, Wright State University, Dayton, Ohio, USA
²Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, Canada

It is of great interest to find the minimum effective dose (MED) in dose-response studies. A sequence of decreasing null hypotheses to find the MED is formulated under the assumption of nondecreasing dose response means. A step-up multiple test procedure that controls the familywise error rate (FWER) is constructed based on the maximum likelihood estimators for the monotone normal means. When the MED is equal to one, the proposed test is uniformly more powerful than Hsu and Berger's test (1999). Also, a simulation study shows a substantial power improvement for the proposed test over four competitors. Three R-codes are provided in Supplemental Materials for this article. Go to the publishers online edition of Journal of Biopharmaceutical Statistics to view the files.

Key Words: Closed test method; Familywise error rate; Pooled-adjacent-violator algorithm; Step-up tests.

1. INTRODUCTION

A situation frequently encountered in dose-response studies is to find the minimum effective dose (MED). The MED is defined as the lowest dose such that the mean responses of this and higher doses are larger than that of a zero dose control by a clinically significant difference. Finding the MED is important since high doses often turn out to have undesirable side effects. See, for example, Tamhane et al. (2001) and Bretz et al. (2005).

Consider the one-way layout model

\[ Y_{ij} = \mu_i + \epsilon_{ij} \]  

(1.1)

for \( i = 0, \ldots, k, j = 1, \ldots, n_i \), where \( \mu_i \) is the unknown mean dose response at dose level \( i \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \) is the independent error with a common unknown variance. Assume \( i = 0 \) is the control group. Typically, the mean dose response goes large when dose level increases. Tallarida (2000) discussed many examples of this kind. So the parameter space in this article is given below

\[ H = \left\{ \mu = (\mu_0, \mu_1, \ldots, \mu_k) : \mu_1 \leq \ldots \leq \mu_k \right\} \]  

(1.2)

Received September 30, 2012; Accepted November 19, 2013
Address correspondence to Weizhen Wang, Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA; E-mail: weizhen.wang@wright.edu

525
For simplicity, we omit $\sigma$. The sufficient statistics are the sample means, $\bar{Y}_i$’s, and the mean squared error, denoted by $S^2$, with $v = \sum_{i=0}^{k} n_i - (k + 1)$ degrees of freedom. The goal is to find the smallest positive integer $N$ satisfying $\mu_N > \mu_0 + \delta$ for a clinically significant difference constant $\delta \geq 0$, i.e., $\mu_i > \mu_0 + \delta$ for any $i \geq N$ and $\mu_{N-1} \leq \mu_0 + \delta$. We call $N$ the MED.

If $Y_{ij}$ assumes binary values, instead of comparing the means, the proportions $p_i$’s are compared with $p_0$, where each $p_i$ is the proportion of patients who show improvement using the drug at dose level $i$. Multiple test procedures for such data can be found in Chen (2008) and Wang (2010). Obviously, certain information is lost when we convert a continuous variable to a binary one, and typically, tests using binary data are less powerful. So we focus on model (1.1) here.

Determination of the MED usually is done by step-down test procedures, see Williams (1971), Ruberg (1989), Tamhane et al. (1996), Hsu and Berger (1999), and Hellmich and Lehmacher (2005), among others. Such procedures start comparison from the highest dose level, in each step go down by one dose level, and stop until the dose level is not significant from the control. Then the dose levels higher than the current step are significant. However, there are efforts to determine the MED using the step-up test procedures, which start comparison from the lowest dose level, in each step go up by one dose level, and stop until the dose level is significant. Then the dose level in the current step and those dose levels not checked yet are significant. Tamhane et al. (1996) also proposed such a procedure that is based on a procedure in Dunnett and Tamhane (1992, 1995). Dunnett and Tamhane (1992, 1995) discussed the cases of balanced and unbalanced designs, respectively, but the control of the familywise error rate (FWER) for unbalanced designs is not established. Noticing this, Liu (1997) proposed a method of calculating the critical values of the step-up procedure to control the FWER, however, a proof is needed. More importantly, since the Tamhane et al. step-up method (1996) does not make use of the monotonicity in dose response means, its power may be improved. Between the step-down and step-up test procedures, intuitively, the latter (former) tends to infer a small (large) dose as the MED. If the true MED is small (large), then the latter (former) is more powerful. None is always better than the other as evidenced in Table 2. Therefore, it is of interest to construct a powerful step-up test procedure utilizing the monotonicity in dose response means.

To derive a test of level-$\alpha$, one needs an appropriate statistic and then identifies the least favorable distribution. In many cases, this distribution can be found by establishing a stochastic ordering for the test statistic. The desired statistic, the estimator of $\mu_i$, should be:

a. nondecreasing in $i$;
b. also nondecreasing in each sample mean, $\bar{Y}_j$.

An attractive candidate is the maximum likelihood estimator (MLE) under $H$ generated by the well-known pooled-adjacent-violator algorithm (PAVA), denoted by $\hat{\mu}_i^{pava}$. The PAVA was first proposed by Ayel et al. (1955) in the purpose of estimating the monotone proportions in independent binomial experiments. Surprisingly, it has many other applications in normal, Poisson, and multinomial distributions. The PAVA is an iterative algorithm and each step is very simple to implement. A closed form of $\hat{\mu}_i^{pava}$ (the max–min formula) is given in Robertson et al. (1988) and will be used to establish (a) and (b).

**Example 1.** To illustrate the problem of finding the MED, consider a dataset from Table 1 in Ruberg (1995). There are nine ($k = 9$) active dose groups and a zero dose control group
with six \((n_i = 6, i = 0, \ldots, 9)\) animals/group in the experiment. The goal is to find the smallest \(i\) at which the mean dose response is larger than that of zero dose response by \(\delta\). For demonstration purpose, choose \(\delta = 6.5\). Details about this dataset, as well as the values of \(\hat{\mu}_i^\text{pava}\), are given in Table 1. Intuitively, the MED should be equal to the smallest \(i\) so that \(\hat{\eta}_i = \hat{\mu}_i^\text{pava} - \bar{Y}_0 - \delta\) is large. We will perform a formal analysis on this dataset in section 2.5.

The rest of the article is organized as follows. In section 2, a sequence of decreasing null hypotheses in \(i\) is formulated in (2.1) to find the MED and a step-up test procedure (2.11) that controls the FWER is derived. This procedure is illustrated in Example 1. Section 3 contains three simulation studies to compare the new proposed procedure and its four competitors and the MCP-Mod procedure. Some discussions are given in section 4. All proofs are given in the Appendix.

2. A STEP-UP MULTIPLE TEST PROCEDURE TO DETECT THE MED

In this section, we develop a step-up multiple test procedure in four subsections as explained in the last paragraph of the previous section.

2.1. Motivation

Let

\[ X_i = \bar{Y}_i - \bar{Y}_0 \text{ and } \eta_i = \mu_i - \mu_0 \text{ for } i = 1, \ldots, k \]

Since the MED is to be found, one would start the search from \(i = 1\) instead of \(i = k\). Therefore, a step-up procedure seems more reasonable than a step-down one. To establish \(N = 1\) (MED) = 1, for example, Hsu and Berger (1999) compare \(\min \{X_j : j \geq 1\}\) with \(\delta\) and claim \(N = 1\) if \(\min \{X_j : j \geq 1\} - \delta\) is large in the unit of \(S\), introduced below (1.2). Roughly speaking, they use \(\min \{X_j : j \geq i\}\) to estimate \(\eta_i\). This does not fully utilize the assumption of the monotonicity in means. We propose using the maximum likelihood estimator of \(\eta_i\), denoted by \(\hat{\eta}_i = \hat{\mu}_i^\text{pava} - \bar{Y}_0\), as a test statistic. If \(\hat{\eta}_1 - \delta\) is larger than a multiple of \(S\), then claim \(N = 1\) and stop; otherwise compare \(\hat{\eta}_2 - \delta\) with \(S\). Repeat this process until we find an \(N\) so that \(\hat{\eta}_N - \delta\) is much larger than \(S\). If no such \(N\) can be found, then the MED does not exist.
To estimate $N$ (MED), let

$$C = \{H_{0i} = \{\eta_i \leq \delta \} \subset H : i \in [1, k]\} \quad (2.1)$$

be the set of null hypotheses of interest in this article. For each $i \geq 1$, the alternative $H_{Ai}$ claims $\eta_i > \delta$. If a certain $H_{Ai}$ is established, then $N \leq i$ due to the monotonicity in $\mu_i$’s for $i \geq 1$. Therefore, $N$ should be equal to the smallest $i$ so that $H_{Ai}$ is true. Note that $C$ is closed under the operation of intersection because $H_{0i}$ is a subset of $H_{0i}$ if $i < i'$, so the closed test procedure by Marcus et al. (1976) can be applied. Let $R_i$ be a rejection region for $H_{0i}$ for any $i$ between 1 and $k$. In order to control the FWER, as well as being powerful, region $R_i$ should satisfy the following two properties:

(*) $R_i$ is of level $\alpha$, i.e., $\sup_{\mu \in \mathbb{R}_0} P_{\mu} (R_i) = \alpha$.

(**) $R_i$ is increasing in $i$, i.e., $R_i \subset R_{i'}$ if $i < i'$. Thus $R_i = \cap_{i'=i}^{k} R_{i'}$.

Then the multiple tests, which assert $H_{Ai}$ if and only if $R_i$ occurs, control the FWER at level $\alpha$, which is the main result of this article. This technique was originally applied to detect the number of active effects in orthogonal saturated designs in Wu and Wang (2007). We now move one step further to detect the set of effective doses (not just the number of effective doses). Thus the inference here is more precise.

To achieve (*) and (**), we describe test statistic $\hat{\mu}_i^{pava}$ and establish a stochastic ordering in section 2.2, provide a general result to identify the least favorable distribution in section 2.3, then construct a sequence of increasing rejection regions $R_i$ of level-\(\alpha\) for each null hypothesis in (2.1) in section 2.4. The determination of cutoff points is discussed in section 2.5 and the dataset in Example 1 is analyzed.

### 2.2. A Stochastic Ordering on $\hat{\mu}_i^{pava}$

We now describe the maximum likelihood estimator $\hat{\mu}_i^{pava}$ for $\mu_i$ under $H$ in (1.2). For a dataset of sample means and sample sizes $\{(\bar{Y}_i, n_i)\}_{i=1}^k$, let

$$n_{ij} = \sum_{h=i}^{j} n_h, \quad \bar{Y}_{ij} = \frac{\sum_{h=i}^{j} n_h \bar{Y}_h}{n_{ij}}, \quad \forall 1 \leq i \leq j \leq k$$

be the sample size and the sample mean of a combined sample of dose levels $i$ through $j$, respectively. The PAVA proceeds as follows:

**Step 0-PAV A.** If $\bar{Y}_i$ is nondecreasing in $i$ for $i \in [1, k]$, then define $\hat{\mu}_i^{pava} = \bar{Y}_i$; otherwise, go to the next step.

**Step 1-PAV A.** For the dataset $\{(\bar{Y}_i, n_i)\}_{i=1}^k$, pick any consecutive pair $(\bar{Y}_j, \bar{Y}_{j+1})$ with $\bar{Y}_j > \bar{Y}_{j+1}$ for some $j$, let $j_i$ be the smallest integer so that $\bar{Y}_i = \bar{Y}_j$ for any $i \in [j_i, j]$ and let $j_a$ be the largest integer so that $\bar{Y}_i = \bar{Y}_{j+1}$ for any $i \in [j+1, j_a]$. Then obtain a new dataset of $\{(a_i, n_i)\}_{i=1}^k$, where $a_i = \bar{Y}_i$ for $i \notin [j_i, j_a]$ and $a_i = \bar{Y}_{j+1}$ for $i \in [j_i, j_a]$. Note two facts: $a_i$ is constant (so is nondecreasing) for $i \in [j_i, j_a]$, and the number of different $a_i$’s is strictly less than that of $\bar{Y}_i$’s.
Step 2-PAVA. Repeat Step 1-PAVA on \((a_i, n_i)_{i=1}^{k}\) until all \(a_i\)'s are nondecreasing. Then define \(\hat{\mu}_{piava}^{pava} = a_i\). This algorithm ends in a finite number of steps because the number of different \(a_i\)'s is strictly less than that in the previous step.

It is clear from Step 2-PAVA that \(\hat{\mu}_{piava}^{pava}\) is nondecreasing in \(i\), satisfying property (a) proposed in section 1. Property (b) is also true for \(\hat{\mu}_{piava}^{pava}\) following the max–min formula (see Robertson et al. 1988, p. 19):

\[
\hat{\mu}_{piava}^{pava} = \max_{s \leq i} \{\min_{t \geq i} \{\bar{Y}_{is}\}\}
\]

Therefore, \(\hat{\mu}_{piava}^{pava}\) is stochastically nondecreasing in \(\mu_j\) when the other \(\mu_j\)'s are held fixed following, for example, Alam and Rizvi (1966). An R-code to implement (2.2) is available in Supplementary Materials with an input of \{\((\bar{Y}, n_i)\)\}_{i=1}^{k}\) and an output of \{\((\bar{Y}, n_i, \hat{\mu}_{piava}^{pava})\)\}_{i=1}^{k}.

### 2.3. A General Result

We now provide a result that is able to identify the least favorable distribution for each null hypothesis in (2.1).

**Theorem 1.** Let \(T(t_1, \ldots, t_k)\) and \(g_i(t_1, \ldots, t_k)\) for \(i = 1, \ldots, k\) be non-decreasing functions for any \(t_i\) when the other \(t_j\)’s are held constants. Also

\[
g_i(ct_1 + d, \ldots, ct_k + d) = cg_i(t_1, \ldots, t_k) + d
\]

for any constants \(c > 0\) and \(d\). Then the following expectation function

\[
f(\eta_1, \ldots, \eta_k, \sigma) \overset{\text{def}}{=} ET\left(\frac{g_1(\bar{Y}_1, \ldots, \bar{Y}_k) - \bar{Y}_0 - \delta}{S}, \ldots, \frac{g_k(\bar{Y}_1, \ldots, \bar{Y}_k) - \bar{Y}_0 - \delta}{S}\right)
\]

(2.4)

is nondecreasing in each \(\eta_i\) when the other \(\eta_j\) and \(\sigma\) are held constants.

**Remark 1.** Each \(g_i(\bar{Y}_1, \ldots, \bar{Y}_k) \overset{\text{def}}{=} \hat{\mu}_{piava}^{pava}\) satisfies (2.3), and is nondecreasing in each \(\bar{Y}_i\). We will also introduce a sequence of rejection regions (equivalent to a sequence of indicator functions \(T_j\)) to construct step-up tests in the next subsection. Then the function in (2.4) is the probability of some rejection region.

To see an application of Theorem 1, define \(g_i(\bar{Y}_1, \ldots, \bar{Y}_k) = \bar{Y}_i\) for \(i \in [1, k]\) and indicator function

\[
T_j^{HB} = \mathbb{I}\{\min_{i \in [\lambda]} \{\bar{Y}_i - \bar{Y}_0 - \delta\}/(S\sqrt{n_i + 1/n_0}) > t_{\alpha, v}\}
\]

(2.5)

for \(j \in [1, k]\), where \(t_{\alpha, v}\) is the upper \(\alpha\)th percentile of \(t\)-distribution with \(v\) degrees of freedom. Then \(g_i\) and \(T_j^{HB}\) satisfy the conditions of Theorem 1. The Hsu and Berger’s step-down test (1999) claims \(N\), the MED, to be \(j_0\) if \(T_{j_0}^{HB} = 1\) but \(T_{j_0-1}^{HB} = 0\).
2.4. The Construction of Step-Up Tests

We first construct a rejection region $R^i_{1,c}$ of level $\alpha$ for each individual $H_{0i}$.

**Lemma 1.** For a constant $c$, let

$$R^i_{1,c} = \left\{ \frac{\hat{\mu}_i - \bar{Y}_0 - \delta}{S} > c \right\}$$

Then

$$\sup_{\mu \in H_{0i}} P_{\mu} (R^i_{1,c}) = P_{\mu} (R^i_{1,c})$$

where $\mu = (\mu_0, \mu_1, \ldots, \mu_k)$ with $\mu_1 = \ldots = \mu_i = \mu_0 + \delta$ and $\mu_{i+1} = \ldots = \mu_k = +\infty$.

Therefore, for any $\alpha \in (0, 1)$, the rejection region $R^i_{1,c}$, with $c = c_{i,\alpha}$, defines a level-$\alpha$ test for $H_{0i}$, where $c_{i,\alpha}$ is the solution of

$$P_{\mu} (R^i_{1,c}) = \alpha$$

It is easy to obtain

$$c_{1,\alpha} = t_{\alpha,\sqrt{1/n_1+1/n_0}}$$

since $\hat{\mu}_1 = \bar{Y}_1$ when $\mu = \mu_1$. Region $R^i_{1,c}$ satisfies property $(\ast)$, but not property $(\ast\ast)$ in section 2.1. To derive more powerful multiple tests, we propose

**Theorem 2.** For any integer $i \in [1, k]$ and for a sequence of constants $c_1$ through $c_i$, let

$$R_{c_1,\ldots,c_i} = \bigcup_{j=1}^i R^j_{1,c_j} = \bigcup_{j=1}^i \left\{ \frac{\hat{\mu}_j - \bar{Y}_0 - \delta}{S} > c_j \right\}$$

Then

$$\sup_{\mu \in H_{0i}} P_{\mu} (R_{c_1,\ldots,c_i}) = P_{\mu} (R_{c_1,\ldots,c_i})$$

For any $\alpha \in (0, 1)$, let $c_1 = c_{1,\alpha}$, given in (2.7), and let $c_i$ be determined iteratively by solving

$$P_{\mu} (R_{c_1,\ldots,c_i}) = \alpha$$

for $i = 2, \ldots, k$. Then region $R^i_{\text{def}} = R_{c_1,\ldots,c_i}$ is a level-$\alpha$ test for $H_{0i}$ following (2.9).

Return to Remark 1, $T_j$ is the indicator function of the rejection region $R_{c_1,\ldots,c_i}$ for any $j = 1, \ldots, k$. The details of computing $c_i$’s are given in Example 1 (continued). Now consider all hypotheses in $C$ in (2.1) with the following testing procedure:

assert $H_{Ai}$ (or not $H_{0i}$) if $R^i_{\text{def}} = R_{c_1,\ldots,c_i}$ occurs
for any fixed $\alpha \in (0, 1)$. Then the FWER is at most $\alpha$, i.e., the probability of making at least one incorrect assertion is at most $\alpha$. This claim simply follows the closed test procedure in Marcus et al. (1976) and the facts that $H_i$ is decreasing in $i$, and $R_i$ is of level-$\alpha$ and is increasing in $i$. When the design is balanced, region $R_1$ contains the set of $\{T_i^{HB} = 1\}$, on which Hsu and Berger’s test (1999) claims the MED = 1. Therefore, the proposed test (2.11) is uniformly more powerful than Hsu and Berger’s when the MED = 1.

2.5. Example 1 (Continued)

We now compare the results of the proposed step-up test procedure with four competitors: three step-down test procedures, SD1P in Tamhane et al. (1996, p. 26), W in Williams (1971) (see Tamhane et al., 1996, p. 23), and DR in Hsu and Berger (1999, p. 470); and the step-up test procedure, SU1P in Tamhane et al. (1996, p. 27), when they are applied to the dataset in Table 1. The square root of the mean squared error $S = 7.751$ is obtained.

The critical values for the SD1P are taken from Bechhofer and Dunnett (1988), and the SD1P infers $\text{MED} = 5$.

The W procedure has the $i$ statistics: $\bar{t}_1 = -1.810$, $\bar{t}_2 = -0.961$, $\bar{t}_3 = 0.313$, $\bar{t}_4 = 1.899$, $\bar{t}_5 = 5.788$, $\bar{t}_6 = \bar{t}_7 = \bar{t}_8 = 9.334$, $\bar{t}_9 = 9.877$, and the $\bar{t}$ statistics are compared with the following critical values (taken from Williams, 1971) in a step-down manner: $c_1 = 1.675$, $c_2 = 1.755$, $c_3 = 1.780$, $c_4 = 1.790$, $c_5 = 1.795$, $c_6 = 1.800$, $\delta$, $c_8 = 1.805$, $c_9 = 1.810$. The W procedure infers $\text{MED} = 4$.

For the DR procedure, following (2.5) with $t_{0.05, 50} = 1.676$, one obtains $T_4^{HB} = 1$ and $T_3^{HB} = 0$. So it also infers $\text{MED} = 4$.

From the SU1P, we have the critical values from Table 1 in Dunnett and Tamhane (1992), and conclude $\text{MED} = 5$.

For the proposed procedure, we first show how to compute cutoff points $c_i$. Following (2.7), $c_1 = t_{0.05, v} \sqrt{1/n_1 + 1/n_0} = 0.9676$ with $v = 50$ and $n_1 = n_0 = 6$. The other $c_i$’s are obtained with a simulation of 1,000,000 repetitions by solving (2.10) iteratively, and $c_2 = 1.034$, $c_3 = 1.041$, $c_4 = 1.042$, $c_5 = 1.046$, $c_6 = 1.047$, $c_7 = 1.049$, $c_8 = 1.048$, $c_9 = 1.048$. For example, to compute $c_2$ with a given $c_1$, let $Z_i$’s be independent $N(0,1)$ for $i = 0, 1, 2$ and let $X_i^2$, independent of $Z_i$’s, be a $X^2_v$ random variable with $v$ degrees of freedom. Following (2.2), under

$$
\bar{\mu}_2 = (\mu_0 + \delta, \mu_0 + \delta, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty, +\infty)
$$

$$
\frac{\hat{\mu}_1^{pava} - \bar{Y}_0 - \delta}{S} \overset{D}{=} \min \left\{ \frac{Z_{\mu_0} + \sqrt{\mu_1^{pava} + \mu_2^{pava} / n_1 + n_2}}{\sqrt{X_n^2 / v}} \right\}
$$

and

$$
\frac{\hat{\mu}_2^{pava} - \bar{Y}_0 - \delta}{S} \overset{D}{=} \max \left\{ \frac{\sqrt{\mu_1^{pava} + \mu_2^{pava} / n_1 + n_2}, Z_{\mu_0}}{\sqrt{X_n^2 / v}} \right\}
$$

where $D$ means equal in distribution. Then $c_2$ is the unique solution of
The uniqueness is because the left-hand side above is a decreasing function of $c_2$, and the existence of a solution is due to

$$P \left( \min \left\{ \frac{Z_1}{\sqrt{n_1}}, \frac{\sqrt{n_1}Z_1 + \sqrt{n_2}Z_2}{n_1 + n_2} \right\} - \frac{Z_0}{\sqrt{n_0}} > c_1 \right) \cup \left\{ \max \left\{ \frac{\sqrt{n_1}Z_1 + \sqrt{n_2}Z_2}{n_1 + n_2}, \frac{Z_2}{\sqrt{n_2}} \right\} - \frac{Z_0}{\sqrt{n_0}} > c_2 \right\} = \alpha$$

and Intermediate Value Theorem. An analytic derivation of $c_i$ is difficult, so it is computed by simulation, which is done only on a single parameter configuration for each $c_i$. Therefore, the simulation result on $c_i$ is quite accurate based on a large number of simulations. An R-code for $c_i$’s is available in Supplementary Materials. This is a good example of combining statistical simulation and mathematical derivation since anyone of them cannot determine $c_i$ by itself and both are indispensable.

We then compute test statistics \( \left( \hat{\mu}_pava_i - \bar{Y}_0 - \delta \right) / S \) given in (2.8) for \( i = 1, \ldots, k \) using the dataset in Table 1. An R-code is available in Supplementary Materials to calculate these test statistics. The first four are equal to \(-1.045, -0.555, 0.181, 1.097\), respectively. Comparing these with the cutoff points $c_i$, we find that 1.097 is the first value larger than the corresponding $c_4$. Thus the new step-up test procedure concludes $\hat{MED} = 4$. □

3. SIMULATION STUDIES

A simulation study was conducted to compare the behavior of the newly proposed step-up procedure (denoted by NEW) with the three step-down procedures, SD1P, W, DR, and the step-up procedure SU1P that were mentioned in Example 1 (continued). Here we adopted the FWER definition of Tamhane et al. (1996) as the proportion of replications corresponding to identifying noneffective doses as the MED in the monotone case (the same as Hsu and Berger, 1999). The power is the probability of identifying the true MED as the MED.

We fixed $\alpha = 0.05$, $k = 5$, $n_i = 10$, for $i = 0, \ldots, 5$, and $\mu_0 = 0$. For illustration purpose, we considered values of the clinically significant difference, $\delta = 0.0, 1.0, 2.0$, respectively, and two types of monotone dose-response functions, linear and step. For each parameter configuration, 10,000 replications using R were produced. Table 2 gives estimates of the FWER and the probability of identifying the true MED for each procedure.

It is seen from Table 2 that all the procedures control the FWER quite well under the monotonicity assumption. Configurations with the true MED = 1 involve no FWER and the corresponding entry is omitted in Table 2. The average power (average over 18 monotone configurations) of each procedure and its ranking according to the average power are given in the last two rows of the table. The best three procedures in terms of the average power are NEW, W, and SU1P. NEW is the best in 12/18 cases and not the worst in the 18 studied configurations at all, and is always better than SD1P in the study. Furthermore, the maximum gain in probability over SD1P, W, DR, and SU1P are 0.3758, 0.036, 0.2269 (all with $\delta = 1.0$ at (0, 4, 4, 4, 4, 4)), and 0.1658 (with $\delta = 1.0$ at (0, 0, 0, 0, 0, 4)), respectively. The
A STEP-UP TEST PROCEDURE

Table 2 Estimated power and FWER when \( n = 10, k = 5 \) and \( \sigma = 1 \) under monotone configurations (the maximum power indicated by *)

<table>
<thead>
<tr>
<th>Configuration</th>
<th>( \delta )</th>
<th>MED</th>
<th>SD1P</th>
<th>W</th>
<th>DR</th>
<th>SU1P</th>
<th>NEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 1, 2, 3, 4, 5)</td>
<td>2.0</td>
<td>3</td>
<td>.0314</td>
<td>.0888</td>
<td>.0859</td>
<td>.0596</td>
<td>.0905*</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2</td>
<td>.0380</td>
<td>.0912</td>
<td>.0882</td>
<td>.0722</td>
<td>.0942*</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1</td>
<td>.0832</td>
<td>.1313</td>
<td>.1110</td>
<td>.1403</td>
<td>.1375*</td>
</tr>
<tr>
<td>(0, 4, 4, 4, 4, 4)</td>
<td>2.0</td>
<td>1</td>
<td>.0412</td>
<td>.2153</td>
<td>.1016</td>
<td>.2508*</td>
<td>.2460</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1</td>
<td>.1813</td>
<td>.5211</td>
<td>.3302</td>
<td>.5412</td>
<td>.5571*</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1</td>
<td>.4636</td>
<td>.8118</td>
<td>.6511</td>
<td>.8128</td>
<td>.8278*</td>
</tr>
<tr>
<td>(0, 0, 4, 4, 4, 4)</td>
<td>2.0</td>
<td>2</td>
<td>.0445</td>
<td>.2162</td>
<td>.1245</td>
<td>.2219</td>
<td>.2275*</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2</td>
<td>.1858</td>
<td>.5091</td>
<td>.3630</td>
<td>.4825</td>
<td>.5233*</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>2</td>
<td>.4328</td>
<td>.7552</td>
<td>.6427</td>
<td>.7151</td>
<td>.7637*</td>
</tr>
<tr>
<td>(0, 0, 0, 4, 4, 4)</td>
<td>2.0</td>
<td>3</td>
<td>.0576</td>
<td>.2313</td>
<td>.1640</td>
<td>.2089</td>
<td>.2371*</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>3</td>
<td>.2187</td>
<td>.5168</td>
<td>.4177</td>
<td>.4542</td>
<td>.5222*</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>3</td>
<td>.4841</td>
<td>.7651</td>
<td>.6896</td>
<td>.6930</td>
<td>.7670*</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 4, 4)</td>
<td>2.0</td>
<td>4</td>
<td>.0911</td>
<td>.2666</td>
<td>.2394</td>
<td>.2027</td>
<td>.2680*</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>4</td>
<td>.2840</td>
<td>.5480</td>
<td>.5105</td>
<td>.4391</td>
<td>.5491*</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>4</td>
<td>.5590</td>
<td>.7751*</td>
<td>.7509</td>
<td>.6740</td>
<td>.7710</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, 4)</td>
<td>2.0</td>
<td>5</td>
<td>.1905</td>
<td>.3501</td>
<td>.3971*</td>
<td>.1901</td>
<td>.3291</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>5</td>
<td>.4314</td>
<td>.6134</td>
<td>.6569*</td>
<td>.4269</td>
<td>.5927</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>5</td>
<td>.6832</td>
<td>.8013</td>
<td>.8305*</td>
<td>.6516</td>
<td>.7849</td>
</tr>
<tr>
<td>Average power rank</td>
<td></td>
<td></td>
<td>.2501</td>
<td>.4560</td>
<td>.3975</td>
<td>.4021</td>
<td>.4605*</td>
</tr>
</tbody>
</table>

maximum loss in probability over \( W, DR, \) and \( SU_1P \) is 0.021, 0.068 (with \( \delta = 1.0 \) at \( 0, 0, 0, 0, 0, 4 \)), and 0.0048 (with \( \delta = 2.0 \) at \( 0, 4, 4, 4, 4, 4 \)), respectively. Although NEW does not demonstrate a uniform gain in probability of detecting the true MED, the improvement is substantial over the loss in probability. Step-up tests tend to be more powerful than step-up tests when the true MED is large.

It is of interest to see the performance of the new procedure when the monotonicity assumption is violated. To study this, we turn to Table 3, where the setup is the same as Table 2 except the configurations are nonmonotone. For the nonmonotone case, the error rate (denoted as ERROR in Table 3) is defined in Hsu and Berger (1999). To illustrate, for \( \mu = (0, 1, 2, 3, 8, 1) \) and \( \delta = 1.0 \), a method commits an error if it infers any dose \( i (1 \leq i \leq 5) \) to be the MED. If a nonmonotone configuration has the MED, then ERROR and FWER are the same and is denoted as FWER in Table 3. From Table 3, we see that for the selected nonmonotone configurations which has the MED for a given \( \delta \), their FWERs are controlled by the methods. However, if the nonmonotone configuration is mildly violated,
Table 3 Estimated FWER/ERROR and probability of identifying true MED under nonmonotone configurations

<table>
<thead>
<tr>
<th>Configuration</th>
<th>δ</th>
<th>MED</th>
<th>SD1P W</th>
<th>DR</th>
<th>SU1P</th>
<th>NEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 1, 2, 3, 4, 2)</td>
<td>2.0</td>
<td>ERROR</td>
<td>.2368</td>
<td>.1298</td>
<td>.0406</td>
<td>.2363</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2</td>
<td>.0798</td>
<td>.0904</td>
<td>.0404</td>
<td>.0700</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FWER</td>
<td>.0341</td>
<td>.0322</td>
<td>.0175</td>
<td>.0292</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>.1410</td>
<td>.1372</td>
<td>.0957</td>
<td>.1298</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 5, 2)</td>
<td>2.0</td>
<td>ERROR</td>
<td>.4597</td>
<td>.2085</td>
<td>.0471</td>
<td>.4586</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>4</td>
<td>.7228</td>
<td>.4977</td>
<td>.1655</td>
<td>.7229</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FWER</td>
<td>.0075</td>
<td>.0075</td>
<td>.0056</td>
<td>.0066</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>4</td>
<td>.8536</td>
<td>.7455</td>
<td>.3649</td>
<td>.8555</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FWER</td>
<td>.0508</td>
<td>.0482</td>
<td>.0414</td>
<td>.0481</td>
</tr>
<tr>
<td>(0, 1, 2, 3, 8, 1)</td>
<td>2.0</td>
<td>ERROR</td>
<td>.9810</td>
<td>.5124</td>
<td>.0085</td>
<td>.9808</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>ERROR</td>
<td>.9975</td>
<td>.8012</td>
<td>.0449</td>
<td>.9974</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.1414</td>
<td>.1384</td>
<td>.0598</td>
<td>.1158</td>
</tr>
</tbody>
</table>

NEW may or may not control the error rate. For example, for configuration (0, 1, 2, 3, 4, 2) with δ = 1.0 and δ = 2.0, the error rate by NEW is 0.0330 and 0.1250, respectively. If the monotonic assumption is mildly violated, NEW may still be used with caution. If the monotonic assumption is severely violated such as (0, 1, 2, 3, 8, 1), NEW and W have very excessive error rates, which conforms the results in Hsu and Berger (1999). NEW should not be used at all in this case. In contrast, DR always controls the error rate. Once a configuration has the MED, NEW is generally preferable to DR even if the monotonicity assumption is mildly violated.

Recently there are several papers that use modeling to estimate the MED, see Bornkamp et al. (2007) and Pinheiro et al. (2010), among others. We followed the simulation setup in section 5 of Bretz et al. (2005) to compare the MCP-Mod procedure with NEW as suggested by one referee and the Associate Editor. We considered five dose levels \( d = 0, 0.05, 0.2, 0.6, 1 \), with a single endpoint measured per patient, \( Y \sim N(\mu(d), \sigma^2) \). Sample size per group were \( n = 10, 25, 80, \) and 120. The one-sided significance level for detecting a dose-response relationship (PoA) was \( \alpha = 0.05 \). Table 4 contains eight different data generating dose-response shapes for the mean response \( \mu(d) \). The second through eighth shapes in Table 4 are typical dose-response models used in practice. The response value at \( d = 0 \) is 0.2 for all the shapes. All shapes (except the constant shape) have a maximum dose effect of about 0.6 within the interval \([0, 1]\). The replicated number was 10,000 for each shape and sample size combination. The response standard deviation was set at \( \sigma = 1.478 \) as in Bretz et al. (2005). Table 5 gives the simulated probabilities of establishing PoA for the different methods under the various shape and sample size combinations. We included the likelihood ratio test (LRT in Table 5) as the LRT is one of the most powerful tests for trend \( \mu_0 \leq \mu_1 \leq \ldots \leq \mu_k \) The simulation results in Table 5 are conformable to the results in Table 5 of Bretz et al. (2005). From Table 5 we can see that NEW is comparable with the MCP-Mod procedure and the LRT. NEW is less powerful than the SU1P and SD1P for the nonmonotone quadratic shape only. Notice that NEW is more powerful than the MCP-Mod procedure for the sigmoid \( E_{\text{max}} \) model.

4. DISCUSSION

In this article, we establish a stochastic ordering of \( \hat{\mu}_i^{\text{pava}} \), the MLE of normal means in \( H \). Based on this ordering, a step-up test procedure is proposed to estimate the MED.
A STEP-UP TEST PROCEDURE

Table 4 Data generating dose-response shapes, $D = \{0, 0.05, 0.2, 0.6, 1.0\}$

<table>
<thead>
<tr>
<th>Model</th>
<th>Specification of $\mu (d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$0.2$</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>$0.2 + 0.7d/(0.2 + d)$</td>
</tr>
<tr>
<td>Linear in log-dose</td>
<td>$0.2 + 0.6 \log(5d + 1)/\log(6)$</td>
</tr>
<tr>
<td>Linear</td>
<td>$0.2 + 0.6d$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$0.183 + 0.017 \exp(2d \log(6))$</td>
</tr>
<tr>
<td>Quadratic</td>
<td>$0.2 + 2.049d − 1.749d^2$</td>
</tr>
<tr>
<td>Logistic</td>
<td>$0.193 + 0.607/(1 + \exp(10 \log(3)(0.4 − d)))$</td>
</tr>
<tr>
<td>SigEmax</td>
<td>$0.2 + 0.7d^2/(0.2^2 + d^2)$</td>
</tr>
</tbody>
</table>

Table 5 Probabilities to detect a dose-response relationship for the different dose finding methods

<table>
<thead>
<tr>
<th>Dose-Response shape</th>
<th>n</th>
<th>Methods</th>
<th>Const</th>
<th>$E_{max}$</th>
<th>Lin-Log</th>
<th>Lin</th>
<th>Exp</th>
<th>Quad</th>
<th>Logist</th>
<th>sigEmax</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>MCPMod</td>
<td>0.050</td>
<td>0.255</td>
<td>0.232</td>
<td>0.251</td>
<td>0.226</td>
<td>0.200</td>
<td>0.309</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NEW</td>
<td>0.045</td>
<td>0.224</td>
<td>0.220</td>
<td>0.208</td>
<td>0.194</td>
<td>0.143</td>
<td>0.223</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LRT</td>
<td>0.048</td>
<td>0.244</td>
<td>0.252</td>
<td>0.245</td>
<td>0.227</td>
<td>0.172</td>
<td>0.294</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SU1P</td>
<td>0.046</td>
<td>0.181</td>
<td>0.162</td>
<td>0.145</td>
<td>0.124</td>
<td>0.154</td>
<td>0.166</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD1P</td>
<td>0.047</td>
<td>0.180</td>
<td>0.162</td>
<td>0.145</td>
<td>0.125</td>
<td>0.154</td>
<td>0.166</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>W</td>
<td>0.045</td>
<td>0.226</td>
<td>0.223</td>
<td>0.211</td>
<td>0.198</td>
<td>0.144</td>
<td>0.226</td>
<td>0.272</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>MCPMod</td>
<td>0.050</td>
<td>0.481</td>
<td>0.436</td>
<td>0.475</td>
<td>0.438</td>
<td>0.380</td>
<td>0.587</td>
<td>0.481</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NEW</td>
<td>0.046</td>
<td>0.412</td>
<td>0.406</td>
<td>0.387</td>
<td>0.361</td>
<td>0.254</td>
<td>0.416</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LRT</td>
<td>0.047</td>
<td>0.465</td>
<td>0.478</td>
<td>0.467</td>
<td>0.435</td>
<td>0.318</td>
<td>0.561</td>
<td>0.619</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SU1P</td>
<td>0.051</td>
<td>0.350</td>
<td>0.320</td>
<td>0.284</td>
<td>0.246</td>
<td>0.305</td>
<td>0.336</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD1P</td>
<td>0.051</td>
<td>0.349</td>
<td>0.320</td>
<td>0.285</td>
<td>0.246</td>
<td>0.305</td>
<td>0.337</td>
<td>0.433</td>
</tr>
<tr>
<td></td>
<td></td>
<td>W</td>
<td>0.048</td>
<td>0.426</td>
<td>0.419</td>
<td>0.400</td>
<td>0.374</td>
<td>0.262</td>
<td>0.431</td>
<td>0.518</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>MCPMod</td>
<td>0.050</td>
<td>0.898</td>
<td>0.858</td>
<td>0.895</td>
<td>0.878</td>
<td>0.811</td>
<td>0.962</td>
<td>0.898</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NEW</td>
<td>0.051</td>
<td>0.833</td>
<td>0.827</td>
<td>0.803</td>
<td>0.778</td>
<td>0.604</td>
<td>0.846</td>
<td>0.921</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LRT</td>
<td>0.050</td>
<td>0.881</td>
<td>0.897</td>
<td>0.890</td>
<td>0.867</td>
<td>0.715</td>
<td>0.955</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SU1P</td>
<td>0.053</td>
<td>0.769</td>
<td>0.743</td>
<td>0.698</td>
<td>0.655</td>
<td>0.706</td>
<td>0.780</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD1P</td>
<td>0.053</td>
<td>0.769</td>
<td>0.742</td>
<td>0.699</td>
<td>0.656</td>
<td>0.706</td>
<td>0.780</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td></td>
<td>W</td>
<td>0.052</td>
<td>0.837</td>
<td>0.831</td>
<td>0.808</td>
<td>0.783</td>
<td>0.611</td>
<td>0.851</td>
<td>0.926</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>MCPMod</td>
<td>0.050</td>
<td>0.974</td>
<td>0.956</td>
<td>0.973</td>
<td>0.968</td>
<td>0.933</td>
<td>0.995</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NEW</td>
<td>0.048</td>
<td>0.943</td>
<td>0.940</td>
<td>0.923</td>
<td>0.910</td>
<td>0.772</td>
<td>0.949</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LRT</td>
<td>0.051</td>
<td>0.967</td>
<td>0.972</td>
<td>0.970</td>
<td>0.964</td>
<td>0.866</td>
<td>0.994</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SU1P</td>
<td>0.051</td>
<td>0.907</td>
<td>0.888</td>
<td>0.856</td>
<td>0.829</td>
<td>0.865</td>
<td>0.912</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD1P</td>
<td>0.051</td>
<td>0.906</td>
<td>0.889</td>
<td>0.857</td>
<td>0.831</td>
<td>0.866</td>
<td>0.912</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>W</td>
<td>0.049</td>
<td>0.946</td>
<td>0.942</td>
<td>0.926</td>
<td>0.914</td>
<td>0.777</td>
<td>0.952</td>
<td>0.985</td>
</tr>
</tbody>
</table>

It controls the FWER, and is powerful to detect the MED, especially when the true MED is small. Each individual test (2.6) is based on the MLE of $\mu_i$, but is not the LRT for $H_{0i}$. It is of great research interest in future to derive the LRT for each $H_{0i}$ in (2.1) and obtain a step-up test procedure. The variances are assumed identical for all doses in the article. In some applications the assumption of homoscedasticity of variances is not satisfied. It worthy further development for the heteroscedastic case (see, for example, Roth, 1983 and Tamhane and Logan, 2004). When using the proposed step-up procedure and other multiple testing procedures aforementioned in the article, any inference is restricted
to the distinct dose levels being administered in a given trial. On the other hand, the MCP-Mod method can choose a dose that is not one of the studied doses due to the flexibility of modelling for dose estimation. How to compare different MED estimation is an interesting topic for future research.

APPENDIX

Proof of Theorem 1. Due to (2.3), we assume $\bar{Y}_0$ has a mean 0 and $\bar{Y}_i$ has a mean $\eta_i = \mu_i - \mu_0$. Let $\phi(x)$ be the pdf of $N(0, 1)$ and $g_v(y)$ be the pdf of a $\chi^2$-distribution with $v = \sum_{i=0}^{k} n_i - (k + 1)$ degrees of freedom. Then

$$f(\eta_1, \ldots, \eta_k, \sigma) = \int \int ET \left( \frac{g_1 - x \frac{\alpha}{\sqrt{m}} - \delta}{\sqrt{\frac{\sigma^2_y}{v}}} , \ldots, \frac{g_k - x \frac{\alpha}{\sqrt{m}} - \delta}{\sqrt{\frac{\sigma^2_y}{v}}} \right) \phi(x) g_v(y) \, dx \, dy$$

For each fixed $x$ and $y$, let

$$T_{x,y}(\bar{Y}_1, \ldots, \bar{Y}_k) = T \left( \frac{g_1 - x \frac{\alpha}{\sqrt{m}} - \delta}{\sqrt{\frac{\sigma^2_y}{v}}}, \ldots, \frac{g_k - x \frac{\alpha}{\sqrt{m}} - \delta}{\sqrt{\frac{\sigma^2_y}{v}}} \right)$$

which is non-decreasing in each $\bar{Y}_i$ due to the monotonicity of $T$ and $g_i$'s. Therefore, the conditional distribution of $T_{x,y}$ for given $x$ and $y$ is stochastically nondecreasing in each $\eta_i$ (see Lemma 2 in Wu and Wang (2007)). Hence its conditional expectation

$$ET_{x,y} = \int ET \left( \frac{g_1 - x \frac{\alpha}{\sqrt{m}} - \delta}{\sqrt{\frac{\sigma^2_y}{v}}}, \ldots, \frac{g_k - x \frac{\alpha}{\sqrt{m}} - \delta}{\sqrt{\frac{\sigma^2_y}{v}}} \right)$$

(5.1)

is nondecreasing in each $\eta_i$. So is $f$, the integral of (5.1). □

Proof of Lemma 1. Let $T = I_{\mathbb{R}_+}$. Then Lemma 1 follows Theorem 1. □

Proof of Theorem 2. Let $T = I_{R_{\bar{Y}_1 \ldots \bar{Y}_k}}$. Then Theorem 2 follows Theorem 1. □

ACKNOWLEDGMENTS

Peng thanks Dr. Bornkamp for his help on using the package of DoseFinding. The authors are grateful to two anonymous referees and an Associate Editor for their constructive suggestions.

FUNDING

Wang’s research is supported in part by NSF Grant DMS-0906858. Peng’s research is supported in part by NSERC of Canada.
SUPPLEMENTAL MATERIAL

Supplemental data for this article can be accessed on the publisher’s website.

REFERENCES


