Stepwise Confidence Intervals without Multiplicity Adjustment for Dose-Response and Toxicity Studies

Jason C. Hsu; Roger L. Berger

Journal of the American Statistical Association, Vol. 94, No. 446 (Jun., 1999), 468-482.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199906%2994%3A446%3C468%3ASCIWMA%3E2.0.CO%3B2-W

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Fri Oct 14 17:51:17 2005

Stepwise Confidence Intervals Without Multiplicity Adjustment for Dose–Response and Toxicity Studies

Jason C. HSU and Roger L. BERGER

Not all simultaneous inferences need multiplicity adjustment. If the sequence of individual inferences is predefined, and failure to achieve the desired inference at any step renders subsequent inferences unnecessary, then multiplicity adjustment is not needed. This can be justified using the closed testing principle to test appropriate hypotheses that are *nested* in sequence, starting with the most restrictive one. But what hypotheses are appropriate may not be obvious in some problems. We give a fundamentally different, confidence set–based justification by *partitioning* the parameter space naturally and using the principle that exactly one member of the partition contains the true parameter. In dose–response studies designed to show superiority of treatments over a placebo (negative control) or a drug known to be efficacious (active control), the confidence set approach generates methods with meaningful guarantee against incorrect decision, whereas previous applications of the closed testing approach have not always done so. Application of the confidence set approach to toxicity studies designed to show equivalence of treated groups with a placebo is also given.

1. STEPWISE CONFIDENCE SETS WITHOUT MULTIPLICITY ADJUSTMENT

Suppose that data $\mathbf{Y}$ have a distribution determined by the parameter $\boldsymbol{\theta} \in \Theta$, the parameter space, and $\boldsymbol{\theta} \in \Theta_i^c, i = 1, \ldots, m$, are multiple comparison inferences of interest. For example, suppose that $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k)$, a vector of mean effects of k treatments, where μ_1 is the mean of the control. Then in one-sided comparisons with the control where significant difference inference is of interest (as in dose–response studies, for example), the desired inferences are $\mu_i > \mu_1 + \delta$, where δ defines practical significant difference, so $\Theta_i^c = \{\mu_i - \mu_1 > \delta\}, i = 2, \ldots, k$. In multiple comparisons with the control where practical equivalence inference is of interest (as in toxicity studies, for example), if μ_i and μ_1 can be considered practically equivalent when $\delta_1 < \mu_i - \mu_1 < \delta_2$, then $\Theta_i^c = \{\delta_1 < \mu_i - \mu_1 < \delta_2\}, i = 2, \ldots, k$. Such multiple comparisons do not need multiplicity adjustment in some situations. One such situation occurs when it is desirable to give the inferences in a specified order, and failure to achieve the desired inference at any step renders subsequent comparisons unnecessary. This situation arises in dose–response and toxicity studies, where $\mu_2, \ldots, \mu_k$ correspond to increasing dose of a substance. In dose–response studies, it is desirable for a method to not declare a lower dose to be efficacious if it does not declare a higher dose to be efficacious. This can be achieved by answering the question "is $\mu_i > \mu_1 + \delta$" in a stepwise fashion, continuing only while the answer is in the affirmative. Likewise, in toxicity studies, it is desirable for a method to not declare a higher dose to be safe if it does not declare a lower dose to be safe. This can be achieved by answering the question "is $\delta_1 < \mu_i - \mu_1 < \delta_2$" in a stepwise fashion, continuing only while the answer is in the affirmative. We describe a general method of providing confidence sets $\boldsymbol{\theta} \in \Theta_i^c$ in a stepwise fashion without multiplicity adjustment, stopping at the first i for which such inference is impossible.

In this section we define our stepwise confidence set procedure and prove that it has the correct coverage probability. In the sections that follow we show that in dose–response and toxicity studies, our confidence set approach is more reliable in generating simple methods with meaningful guarantee against incorrect decision than other approaches.

Definition 1. Let $\Theta^* \subset \Theta$. A confidence set, $C(\mathbf{Y})$, for $\boldsymbol{\theta}$ is *directed towards* Θ^* if, for every sample point $\mathbf{y}$, either $\Theta^* \subset C(\mathbf{y})$ or $C(\mathbf{y}) \subset \Theta^*$.

For our one-sided significant difference inference example, $\Theta_i^c = \{\mu_i - \mu_1 > \delta\}$. Confidence intervals for $\mu_i - \mu_1$ of the form $C(\mathbf{Y}) = (L(\mathbf{Y}), \infty)$ are directed toward Θ_i^c. If $D(\mathbf{Y})$ is any $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$, then

$$C(\mathbf{Y}) = \begin{cases} D(\mathbf{Y}), & \text{if } D(\mathbf{Y}) \subset \Theta^* \\ D(\mathbf{Y}) \cup \Theta^*, & \text{otherwise,} \end{cases} \tag{1}$$

is a $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$ that is directed toward Θ^*. In our practical equivalence inference example, $\Theta_i^c = \{\delta_1 < \mu_i - \mu_1 < \delta_2\}$, so if $D_i(\mathbf{Y})$ is a confidence set for $\mu_i - \mu_1$, then

$$C_i(\mathbf{Y}) = \begin{cases} D_i(\mathbf{Y}), & \text{if } D_i(\mathbf{Y}) \subset (\delta_1, \delta_2), \\ D_i(\mathbf{Y}) \cup (\delta_1, \delta_2), & \text{otherwise,} \end{cases}$$

is directed toward Θ_i^c.

In general, suppose that the inferences $\boldsymbol{\theta} \in \Theta_i^c$, in the order $i = 1, \ldots, m$, are of interest.

Theorem 1. Suppose that for $i = 1, \ldots, m, C_i(\mathbf{Y})$ is a $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$ that is directed toward Θ_i^c and $C_{m+1}(\mathbf{Y})$ is any $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$. Let M denote the smallest integer i such that $C_i(\mathbf{Y}) \not\subset \Theta_i^c$

Jason C. Hsu is Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210. (E-mail: *jch@stat.ohio-state.edu*). Roger L. Berger is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *berger@stat.ncsu.edu*). The first author's research was supported in part by National Institutes of Health grant CA16058. The authors learned much about dose–response studies from Steve Ruberg, and had fruitful discussions with Gary Koch (see Koch and Gansky 1996), Tony Lachenbruch, and Kathy Fritsch on this subject. Elizabeth Margosches provided references on EPA's view on toxicity studies.

© 1999 American Statistical Association
Journal of the American Statistical Association
June 1999, Vol. 94, No. 446, Theory and Methods

if such an i $(1 \le i \le m)$ exists, otherwise, let $M = m+1$. Define $\Theta_0 = \emptyset$ (so $\Theta_0^c = \Theta$) and

$$C^\star(\mathbf{Y}) = \Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap C_M(\mathbf{Y}).$$

Then, for all $\boldsymbol{\theta} \in \Theta$,

$$P_\theta\{\boldsymbol{\theta} \in C^\star(\mathbf{Y})\} \ge 1 - \alpha.$$

Proof. Define $\Theta_{m+1} = \Theta$. For $i = 1, \ldots, m+1$, let

$$\Theta_i^\star = \Theta_0^c \bigcap \cdots \bigcap \Theta_{i-1}^c \bigcap \Theta_i.$$

Then $\Theta_1^\star, \ldots, \Theta_{m+1}^\star$ partition the parameter space Θ. Clearly,

$$C(\mathbf{Y}) = \bigcup_{i=1}^{m+1} (C_i(\mathbf{Y}) \bigcap \Theta_i^\star)$$

is a $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$, because if $\boldsymbol{\theta} \in \Theta_i^\star$, then

$$\begin{aligned} P_\theta\{\boldsymbol{\theta} \in C(\mathbf{Y})\} &= P_\theta\{\boldsymbol{\theta} \in (C_i(\mathbf{Y}) \bigcap \Theta_i^\star)\} \\ &= P_\theta\{\boldsymbol{\theta} \in C_i(\mathbf{Y})\} \ge 1 - \alpha. \end{aligned}$$

Noting that

(a) $C_j(\mathbf{Y}) \cap \Theta_j^\star = \emptyset$ for all $j < M$ (if such j exists), because $\Theta_j^\star \subset \Theta_j$;
(b) $\Theta_j^\star \subset \Theta_0^c \cap \cdots \cap \Theta_{M-1}^c \cap \Theta_M^c$ for all $j > M$ (if such j exists); and
(c) $\Theta_M^c \subset C_M(\mathbf{Y})$, because if $M < m+1$, then $C_M(\mathbf{Y}) \not\subset \Theta_M^c$ implies $\Theta_M^c \subset C_M(\mathbf{Y})$, whereas $\Theta_{m+1}^c \subset C_{m+1}(\mathbf{Y})$ trivially,

we have

$$\begin{aligned} C(\mathbf{Y}) &= \bigcup_{i=1}^{m+1} (C_i(\mathbf{Y}) \bigcap \Theta_i^\star) \\ &\overset{\text{by(a)}}{=} \bigcup_{i=M}^{m+1} (C_i(\mathbf{Y}) \bigcap \Theta_i^\star) \\ &\overset{\text{by(b)}}{\subset} (C_M(\mathbf{Y}) \bigcap \Theta_M^\star) \bigcup (\Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap \Theta_M^c) \\ &= (\Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap \Theta_M \bigcap C_M(\mathbf{Y})) \\ &\quad \bigcup (\Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap \Theta_M^c) \\ &\overset{\text{by(c)}}{=} (\Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap \Theta_M \bigcap C_M(\mathbf{Y})) \\ &\quad \bigcup (\Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap \Theta_M^c \bigcap C_M(\mathbf{Y})) \\ &= \Theta_0^c \bigcap \cdots \bigcap \Theta_{M-1}^c \bigcap C_M(\mathbf{Y}) \\ &= C^\star(\mathbf{Y}). \end{aligned}$$

Note that our method provides confidence sets $C_i(\mathbf{Y})$ for $\boldsymbol{\theta}$ in a stepwise fashion, stopping at the first i such that $C_i(\mathbf{Y}) \not\subset \Theta_i^c$.

Step 1.

If $C_1(\mathbf{Y}) \subset \Theta_1^c$
then assert $\boldsymbol{\theta} \in \Theta_1^c$ and go to Step 2;
else assert $\boldsymbol{\theta} \in C_1(\mathbf{Y})$ and stop.

Step 2.

If $C_2(\mathbf{Y}) \subset \Theta_2^c$
then assert $\boldsymbol{\theta} \in \Theta_2^c$ and go to Step 3;
else assert $\boldsymbol{\theta} \in C_2(\mathbf{Y})$ and stop.

$\vdots$

Step m.

If $C_m(\mathbf{Y}) \subset \Theta_m^c$
then assert $\boldsymbol{\theta} \in \Theta_m^c$ and go to Step $m+1$;
else assert $\boldsymbol{\theta} \in C_m(\mathbf{Y})$ and stop.

Step m + 1.

Assert $\boldsymbol{\theta} \in C_{m+1}(\mathbf{Y})$ and stop.

Note the wide applicability of this method. As long as individual $100(1-\alpha)\%$ directed confidence sets are available, regardless of how the associated statistics are jointly distributed, the method applies, and Theorem 1 guarantees that the error rate is properly controlled.

This method readily provides a stepwise test for

$$H_{0i}\colon \boldsymbol{\theta} \in \Theta_i \quad \text{versus} \quad H_{ai}\colon \boldsymbol{\theta} \in \Theta_i^c, \qquad i = 1, \ldots, m, \tag{2}$$

as follows. The test ϕ_i, which rejects H_{0i} when $C_i(\mathbf{Y}) \subset \Theta_i^c$, is a level-$\alpha$ test:

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_i} P_\theta(\phi_i \text{ rejects } H_{0i}) &= \sup_{\boldsymbol{\theta} \in \Theta_i} P_\theta(C_i(\mathbf{Y}) \subset \Theta_i^c) \\ &\le \sup_{\boldsymbol{\theta} \in \Theta_i} P_\theta(\boldsymbol{\theta} \notin C_i(\mathbf{Y})) \\ &\le \alpha. \end{aligned}$$

Thus if one executes ϕ_i in the sequence $i = 1, \ldots, m$, stopping as soon as a ϕ_i accepts, then, by Theorem 1, the family-wise error rate (FWER) (see Hochberg and Tamhane 1987 for definition) is no more than α. The confidence set we give is now the simplest known confidence set associated with any stepwise test, simpler than the one associated with the MPGN method discussed in Section 2.1.

One can apply the closed testing principle of Marcus, Peritz, and Gabriel (1976) to test the hypotheses $H_{0i}\colon \boldsymbol{\theta} \in \Theta_i, i = 1, \ldots, m$. This principle tests all possible intersections of the null hypotheses, each at level α, rejecting a resulting hypotheses only if it and all other resulting hypotheses implying it are rejected. Suppose that $\Theta_1, \ldots, \Theta_m$ are nested in the sense that $\Theta_1 \subset \cdots \subset \Theta_m$. Then for any $I \subset \{1, \ldots, m\}$, the intersection of $\{H_{0i}, i \in I\}$ is H_{0h}, where $h = \min\{i\colon i \in I\}$. Thus in this case it follows from the closed testing principle that multiplicity adjustment is

not needed in testing $H_{01}, \ldots, H_{0m}$ in that sequence if one stops as soon as a hypothesis is accepted. As shown by Maurer, Hothorn, and Lehmacher (1995), when $\Theta_1, \ldots, \Theta_m$ are not nested, this result can still be proven using the closed testing principle by appropriately modifying the null hypotheses. But our confidence set derivation shows that the closed testing principle is not needed in the proof.

For compatibility with existing literature, our presentation in Sections 2 and 3 is in terms of the one-way model

$$Y_{ia} = \mu_i + \varepsilon_{ia}, \qquad i = 1, \ldots, k, a = 1, \ldots, n_i,$$

where Y_{ia} is the ath observation under the ith treatment and $\varepsilon_{11}, \ldots, \varepsilon_{kn_k}$ are iid normal with mean 0 and variance σ^2 unknown. We use the following notations:

$$\hat{\mu}_i = \bar{Y}_i = \sum_{a=1}^{n_i} Y_{ia}/n_i$$

and

$$\hat{\sigma}^2 = \text{MSE} = \sum_{i=1}^{k}\sum_{a=1}^{n_i} (Y_{ia} - \bar{Y}_i)^2 \Big/ \sum_{i=1}^{k} (n_i - 1)$$

for the sample means and the pooled sample variance. The upper 100α percentile of a Student's t distribution with $\nu = \sum_{i=1}^{k} n_i - k$ df is denoted by $t_{\alpha,\nu}$.

2. DOSE–RESPONSE STUDIES

Appropriate dosing of a drug is vitally important in biopharmaceutics. Let μ_1 be the mean of the control group, which may be a negative control group receiving a placebo or an active control group receiving a standard drug known to be efficacious. Let $\mu_2, \ldots, \mu_k$ be the mean responses corresponding to increasing dose of a test drug. In practice, dosing is determined by two quantities:

- minimum effective dose (MED)
- maximum tolerated dose (MTD).

The MED of a drug is the minimum dose such that the mean response at that dose is significantly better than the mean response of the controls. The estimated MED ($\widehat{\text{MED}}$) is determined statistically from the observed dose–response relationship, where the response is an endpoint that measures efficacy. The estimated MTD ($\widehat{\text{MTD}}$), on the other hand, is determined from observed adverse events in terms of both anticipated and unanticipated endpoints. As determination of the $\widehat{\text{MTD}}$ appears to be nonstatistical at present, our discussion concentrates on determination of the $\widehat{\text{MED}}$. For convenience, we write $\text{MED} = i$ and $\widehat{\text{MED}} = i$ to represent that the true MED or the estimated MED is the dose corresponding to μ_i. We assume that a larger μ_i indicates a better average outcome.

Suppose that the control group is a negative control group receiving a placebo. If the drug is not expected to be deleterious, then defining MED as the minimum dose i such that $\mu_i > \mu_1$ makes sense only when it is known from biology that the response follows a threshold model. Otherwise, $\text{MED} = 2$, the lowest dose to be compared with the placebo in the study. Thus if the response curve is expected to be continuous, then MED should be defined as the minimum dose such that the mean response at that dose is clinically significantly better than the mean response of the negative controls; that is,

$$\text{MED} = \min\{i\colon \mu_i > \mu_1 + \delta\}, \tag{3}$$

where $\delta > 0$ defines a clinically significant difference. For chronic peripheral arterial occlusive disease, for example, in terms of percent improvement in walking distance, δ has been defined to be 30% (CPMP/EWP/233/95 1995).

Suppose that the control group is an active control group receiving a drug known to be efficacious. Then, in so-called noninferiority trials, MED can be defined by (3) with δ either positive, 0, or negative, as long as $\delta > \delta^*$, where δ^* represents the known quantity, which is the mean placebo effect minus the mean active control effect.

Note that in general it is possible that $\mu_i \le \mu_1 + \delta$ for some i. This can certainly happen in the active control setting. Even in the negative control setting with $\delta = 0$, this can occur if the response has what the International Conference on Harmonization (ICH) guideline E4 on dose–response studies calls an inverted U or umbrella shape (International Conference on Harmonization 1994, sec. 3, p. 6), meaning that $\mu_i, i = 1, \ldots, k$, first increases then decreases as i increases.

Under a one-way model, our stepwise confidence set method (which we call the DR—for dose–response—method) takes the following form:

Step 1.

If $\bar{Y}_k - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_1} \ge \delta$,
then assert $\mu_k > \mu_1 + \delta$ and go to Step 2;
else assert $\mu_k - \mu_1 > \bar{Y}_k - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_1}$ and stop.

Step 2.

If $\bar{Y}_{k-1} - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_{k-1} + 1/n_1} \ge \delta$,
then assert $\mu_{k-1} > \mu_1 + \delta$ and go to Step 3;
else assert $\mu_{k-1} - \mu_1 > \bar{Y}_{k-1} - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_{k-1} + 1/n_1}$ and stop.

$\vdots$

Step $k - 1$.

If $\bar{Y}_2 - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_2 + 1/n_1} \ge \delta$
then assert $\mu_2 > \mu_1 + \delta$ and go to Step k;
else assert $\mu_2 - \mu_1 > \bar{Y}_2 - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_2 + 1/n_1}$ and stop.

Step k.

Assert $\min_{i=2,\ldots,k} \mu_i - \mu_1 > \min_{i=2,\ldots,k}\{\bar{Y}_i - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1}\}$ and stop.

To better understand how this stepwise method operates, let step M $(1 \le M \le k)$ be the step at which the stepwise method stops. If $M > 1$, then the stepwise method declares

doses $k-M+2,\ldots,k$ to be efficacious. If $M<k$, then the stepwise method fails to declare doses $2,\ldots,k-M+1$ to be efficacious, and gives a lower confidence bound (which is less than δ) for $\mu_{k-M+1}-\mu_1$. If $M=k$, then the stepwise method gives a lower bound on how efficacious every dose is. This lower bound is greater than δ.

If the $\widehat{\text{MED}}$ determined by the DR method is smaller than the $\widehat{\text{MTD}}$, then a range $\widehat{\text{MED}}, \widehat{\text{MED}}+1,\ldots,\widehat{\text{MTD}}$ of safe and effective doses is obtained. Note that in contrast, applying Dunnett's (1955) method to compare $\mu_2,\ldots,\mu_k$ with μ_1 will not necessarily produce a contiguous set of effective doses. Having a therapeutic window of safe and effective doses is desirable. For example, the manufacturer may prefer a dose toward the high end of the range, to better compete in the market place, whereas the regulatory agency may prefer a dose toward the low end of the range, to better safeguard the consumer. A wide therapeutic window can facilitate first choosing a higher dose for the prescription version, and later a lower dose for the over-the-counter version of the same drug. In the next section we compare our approach with other approaches.

Instead of Theorem 1, an alternative derivation the DR confidence set is to invoke the connection between tests and confidence sets (Lehmann 1986, p. 90), as indicated for the case of $\delta=0$ that follows. Let

$$\Theta_k=\{\mu_k\le\mu_1\},$$

$$\Theta_i=\{\mu_i\le\mu_1<\min\{\mu_{i+1},\ldots,\mu_k\}\} \qquad \text{for} \quad i=2,\ldots,k-1,$$

and

$$\Theta_1=\{\mu_1<\min\{\mu_2,\ldots,\mu_k\}\},$$

so that $\Theta=\cup_{i=1}^k\Theta_i$ and $\Theta_i\cap\Theta_j=\emptyset$ when $i\ne j$. For each $\boldsymbol{\mu}^0=(\mu_1^0,\ldots,\mu_k^0)\in\Theta_i, i=2,\ldots,k$, define $\phi_i(\boldsymbol{\mu}^0)$ to be the usual size-α t test for

$$H_0\colon\ \mu_i-\mu_1\le\mu_i^0-\mu_1^0 \quad \text{versus} \quad H_a\colon\ \mu_i-\mu_1>\mu_i^0-\mu_1^0;$$

that is, $\phi_i(\boldsymbol{\mu}^0)=1$ if

$$T_i=\frac{\bar{Y}_i-\bar{Y}_1-(\mu_i^0-\mu_1^0)}{\hat{\sigma}\sqrt{1/n_i+1/n_1}}\ge t_{\alpha,\nu}.$$

For each $\boldsymbol{\mu}^0=(\mu_1^0,\ldots,\mu_k^0)\in\Theta_1$, define an intersection-union test $\phi_1(\boldsymbol{\mu}^0)$ for

$$H_0\colon\ \min\{\mu_2,\ldots,\mu_k\}-\mu_1\le\min\{\mu_2^0,\ldots,\mu_k^0\}-\mu_1^0$$

versus

$$H_a\colon\ \min\{\mu_2,\ldots,\mu_k\}-\mu_1>\min\{\mu_2^0,\ldots,\mu_k^0\}-\mu_1^0$$

with $\phi_1(\boldsymbol{\mu}^0)=1$ if

$$T_1=\min_{i=2,\ldots,k}\frac{\bar{Y}_i-\bar{Y}_1-(\min_{i=2,\ldots,k}\mu_i^0-\mu_1^0)}{\hat{\sigma}\sqrt{1/n_i+1/n_1}}\ge t_{\alpha,\nu},$$

which is a level-α test. Then $C(\mathbf{Y})=\cup_{i=1}^k\{\boldsymbol{\mu}^0\colon\ \boldsymbol{\mu}^0\in\Theta_i$ and $\phi_i(\boldsymbol{\mu}^0)=0\}$ is a confidence set for $\boldsymbol{\mu}$ with coverage probability *exactly* $1-\alpha$ for all $\boldsymbol{\mu}$ except for $\boldsymbol{\mu}\in\Theta_1$ (for which the coverage probability is greater than $1-\alpha$). But $C(\mathbf{Y})$ is tedious to describe even for $k=3$. Note that if $M>1$, then, for every $i=k-M+2,\ldots,k, \phi_i(\boldsymbol{\mu}^0)$ rejects for every $\boldsymbol{\mu}^0\in\Theta_i$, so the confidence set $C(\mathbf{Y})$ does not include any points in $\Theta_k,\ldots,\Theta_{k-M+2}$. The DR confidence set is in fact $\{\boldsymbol{\mu}^0\colon\ \boldsymbol{\mu}^0\in\Theta_{k-M+1}$ and $\phi_{k-M+1}(\boldsymbol{\mu}^0)=0\}\cup\cup_{i=1}^{k-M}\Theta_i$ (with the understanding that $\cup_{i=1}^{k-M}\Theta_i=\emptyset$ if $M=k$), and thus contains $C(\mathbf{Y})$. If $\boldsymbol{\mu}\in\Theta_m$, then the DR confidence set has the same coverage probability $1-\alpha$ as $C(\mathbf{Y})$ if $\mu_i-\mu_1\to+\infty$ for $i>m$, but becomes increasingly conservative (has higher coverage probability) as $\mu_i-\mu_1\to 0+$ for $i>m$. However, it can be shown that if one deduces confidence limits on $\mu_i-\mu_1, i=2,\ldots,k$, from $C(\mathbf{Y})$ by finding the suprema of $\mu_i-\mu_1$ for $\boldsymbol{\mu}\in C(\mathbf{Y})$, the confidence limits given by the DR method result.

2.1 Comparison With Other Approaches

Previous formulations of the MED problem have cast it as one of testing a family of null hypotheses of *equalities* against various alternatives. A method that controls the FWER of testing equalities is guaranteed only to be a *confident inequalities* method. But the desired inferences in the MED problem, $\mu_i>\mu_1+\delta$, are directional. We show that in fact most methods that have been proposed are not *confident directions* methods. That is, they do not control the probability of declaring an ineffective dose to be effective. To control the directional error rate, the null hypotheses must be directional themselves, as we now show.

The MED problem can be formulated as one of testing the family of hypotheses

$$H_{0i}^\dagger\colon\ \bigcup_{j=i}^k\{\mu_j\le\mu_1+\delta\} \tag{4}$$

versus

$$H_{ai}^\dagger\colon\ \bigcap_{j=i}^k\{\mu_j>\mu_1+\delta\} \tag{5}$$

for $i=2,\ldots,k$, because if $H_{0i}^\dagger$ is rejected, then there is evidence that doses $j, j\ge i$, are effective. Because the alternative $H_{ai}^\dagger$ implies the alternative $H_{aj}^\dagger$ when $i<j$, one takes as MED the lowest dose i for which $H_{0i}^\dagger$ is rejected. If a method strongly controls the FWER of testing (4) at α, then the probability of declaring a dose as the MED when either it or a higher dose is ineffective is no more than α. Because $H_{0k}^\dagger\subset\cdots\subset H_{02}^\dagger$, to control the FWER, the closed testing principle of Marcus et al. (1976) states $H_{0k}^\dagger,\ldots,H_{02}^\dagger$ can be tested in that sequence, each at level-α, stopping as soon as a hypothesis is accepted. It thus remains to find a suitable level-α test for each $H_{0i}^\dagger$. For $i=2,\ldots,k$, let ϕ_i^t be the size-α test that rejects H_{0i} in

$$H_{0i}\colon\ \mu_i\le\mu_1+\delta \tag{6}$$

versus

$$H_{ai}\colon\ \mu_i>\mu_1+\delta, \tag{7}$$

when

$$\bar{Y}_i - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1} > \delta.$$

Because $H_{0i}^{\dagger}$ is expressed as a union, by work of Berger (1982), the test $\phi_i^{\dagger}$ that rejects $H_{0i}^{\dagger}$ if and only if ϕ_j^t for $H_{0j}, j = i, \ldots, k$, all reject is a level-α test of $H_{0i}^{\dagger}$. The resulting closed test of the hypotheses $H_{02}^{\dagger}, \ldots, H_{0k}^{\dagger}$ using $\phi_2^{\dagger}, \ldots, \phi_k^{\dagger}$ reaches the same decision about MED as the DR method, and can be thought of as the testing version of the DR method. In this testing form, the DR method first performs a size-α one-sided t-test comparing μ_k with $\mu_1 + \delta$. If the test accepts, then it stops and asserts nothing. If the test rejects, then it asserts $\mu_k > \mu_1 + \delta$ and proceeds to a size-α one-sided t-test comparing μ_{k-1} with the $\mu_1 + \delta$, and so on. (Note that the pooled estimate $\hat{\sigma}$ for σ is used in the standardization of the t statistics.)

If the MED problem were posed as one of testing (6) versus (7) instead, declaring dose i to be effective if H_{0i} is rejected, then any test that strongly controls the FWER of testing (6) controls the probability of making the directional error of declaring $\mu_i > \mu_1 + \delta$ when in fact $\mu_i \leq \mu_1 + \delta$. One can use the simultaneous lower confidence bounds on $\mu_i - \mu_1$ of Dunnett (1955) or the ones of Bofinger (1987) (which were also independently given in Stefansson, Kim, and Hsu 1988) to test (6), rejecting H_{0i} only if the lower bound on $\mu_i - \mu_1$ is greater than or equal to δ. Stefansson et al. (1988) showed that the latter confidence bounds correspond to the closed test of (6), which was proposed in the multiple testing setting by Marcus et al. (1976) and in the ranking and selection setting by Naik (1975). If we define

$$T_i = \frac{\bar{Y}_i - \bar{Y}_1 - \delta}{\hat{\sigma}\sqrt{1/n_i + 1/n_1}},$$

let $[2], \ldots, [k]$ be the random indices such that $T_{[2]} \leq \cdots \leq T_{[k]}$, and let $d_{[i]}$ be the critical value of Dunnett's (union-intersection) test for

$$H_{0i}^{\ddagger}\colon \bigcap_{j=i}^{k} \{\mu_{[j]} \leq \mu_1 + \delta\},$$

then this closed test (call it the MPGN method) has the same form as the DR method except

$$\bar{Y}_i - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1} \geq \delta$$

in each step is replaced by

$$T_{[i]} \geq d_{[i]}.$$

The simulation results described in Section 2.3 indicate that when μ_i increases as i increases, the DR method tends to infer an MED that is closer to the true MED than Dunnett's method and the MPGN method.

Instead of testing (4) against (5), a second approach to the MED problem is to formulate it as one of testing the multiple hypotheses

$$H_{0i}^*\colon\ \mu_1^* = \mu_2 = \cdots = \mu_i \tag{8}$$

versus

$$H_{ai}^*\colon\ \mu_1^* = \mu_2 = \cdots = \mu_{i-1} < \mu_i \tag{9}$$

for $i = 2, \ldots, k$, with $\mu_1^* = \mu_1 + \delta$. To strongly control the FWER, because $H_{0k}^* \subset \cdots \subset H_{02}^*$, the closed test of Marcus et al. (1976) tests $H_{0k}^*, \ldots, H_{02}^*$ in that sequence, stopping as soon as a hypothesis is accepted. If ϕ_i is a level-α test for $H_{0i}^*, i = 2, \ldots, k$, then $\widehat{\mathrm{MED}}$ is taken to be the smallest i such that all $\phi_j, j \geq i$, reject. Because the hypothesis (8) implies the hypothesis (6), which in turn implies the hypothesis (4), a level-α test for (4) is also a level-α test for (8). Thus some authors have viewed the DR method in its testing form as one of many stepwise tests generated by applying the closed testing technique to (8). We will show, however, that the formulation of testing (8) versus (9) fails to differentiate between methods which control the directional error rate of declaring a $\mu_i > \mu_1^*$ when $\mu_i \leq \mu_1^*$ from those that do not. That is, even though both the formulations of testing (4) versus (5) and testing (8) versus (9) lead to stepwise testing with no multiplicity adjustment, only the former formulation guarantees that the directional error rate is controlled.

For example, Ruberg (1995) and Tamhane, Hochberg, and Dunnett (1996) studied methods based on size-α contrast tests $\phi_i^{\mathbf{c}_i}$ for H_{0i}^* (with $\delta = 0$) of the form

$$\phi_i^{\mathbf{c}_i} = \begin{cases} 1, & \text{if } \dfrac{\sum_{j=1}^k c_{ij}\hat{\mu}_j}{\hat{\sigma}\sqrt{\sum_{j=1}^k c_{ij}^2/n_j}} > \text{constant} \\ 0, & \text{otherwise,} \end{cases}$$

with $\mathbf{c}_i = (c_{i1}, \ldots, c_{ik}), \sum_{j=1}^k c_{ij} = 0, c_{i1} < 0, c_{ii} > 0$, and $c_{ij} = 0$ for all $j > i$. Even though these methods strongly control the FWER of testing (8) at α, they do not generally control the probability of incorrectly declaring a noneffective dose as effective at α. Bauer (1997) showed this for the case of $k = 3$, but it occurs for all $k \geq 3$. The problem is that the clinical error of incorrectly declaring $\mu_i > \mu_1 + \delta$ when it is false is not counted as an error in testing (8). That is, the family of null hypotheses in (8) is too small a subset of the entire parameter space to offer meaningful protection against incorrect decision making.

We in fact show that unless the contrast tests method is the testing form of the DR method, which is the special case of $c_{i1} = -1, c_{ii} = 1$, and $c_{ij} = 0$ for all $j \neq 1$ and i, it will not control the error rate of declaring $\mu_i > \mu_1^*$ when $\mu_i \leq \mu_1^*$. Consider a contrast tests method that is not the DR method. Then some comparison is not a simple pairwise comparison between μ_i and μ_1^*. Let i^* denote the largest index that is not a simple pairwise comparison. Then $2 < i^* \leq k$ and, for all $i > i^*, c_{i1} = -1, c_{ii} = 1$, and $c_{ij} = 0$ for all $j \neq 1$ and i, but there is a j $(1 < j < i^*)$ with $c_{i^*j} \neq 0$. If we fix μ_1 and all μ_h $(1 < h < i^*, h \neq j)$ but let $\mu_j = (-2c_{i^*i^*}/c_{i^*j})\mu_{i^*}, \mu_{i^*} \to -\infty$ and $\mu_m \to \infty$ for all $m > i^*$, then all $\phi_m^{\mathbf{c}_m}$ $(m > i^*)$ as well as $\phi_i^{\mathbf{c}_{i^*}}$ will reject with probability approaching 1, and dose i^* will be incorrectly declared to be effective.

The simulation results described in Section 2.3 in fact indicate that, compared to the methods based on lin-

ear, Helmert, and reverse Helmert contrasts discussed by Tamhane et al. (1996), the DR method does about as well as the best method for each of the response shapes studied but is the only method that always controls the error rate.

Indeed, the lower confidence bound on $\mu_{i^*} - \mu_1$ obtained by pivoting the $\phi_{i^*}^{\mathbf{c}_{i^*}}$ test is $-\infty$ regardless of data. This can be seen as follows. The test

$$\phi_{i^*}^{\mathbf{c}_{i^*}}(\mu_1^0, \ldots, \mu_k^0) = \begin{cases} 1, & \text{if } \dfrac{\sum_{j=1}^k c_{i^*j}(\hat{\mu}_j - \mu_j^0)}{\hat{\sigma}\sqrt{\sum_{j=1}^k c_{i^*j}^2/n_j}} > \text{constant} \\ 0, & \text{otherwise} \end{cases}$$

is clearly a size-α test for H_0: $\mu_1 = \mu_1^0, \ldots, \mu_k = \mu_k^0$. If we fix all $\hat{\mu}_j$ $(1 \le j \le k), \mu_1$, and all μ_h $(1 < h < i^*, h \ne j)$ but let $\mu_j = (-2c_{i^*i^*}/c_{i^*j})\mu_{i^*}$, then, as $\mu_{i^*} \to -\infty$,

$$\frac{\sum_{j=1}^k c_{i^*j}(\hat{\mu}_j - \mu_j)}{\hat{\sigma}\sqrt{\sum_{j=1}^k c_{i^*j}^2/n_j}} \to -\infty.$$

Therefore, the set of $(\mu_1^0, \ldots, \mu_k^0)$ for which $\phi_{i^*}^{\mathbf{c}_{i^*}}(\mu_1^0, \ldots, \mu_k^0)$ accepts always includes a sequence for which $\mu_{i^*}^0 - \mu_1^0 \to -\infty$. Thus, by the usual correspondence between tests and confidence sets, the lower confidence bound on $\mu_{i^*} - \mu_1$ is $-\infty$ regardless of data. That is, the rejection of (8) in favor of (9) does not in general support $\mu_i > \mu_1^*$.

A third approach to the MED problem that has been taken is to test the hypotheses

$$H_{0i}^{**}\colon\ \mu_1^* = \cdots = \mu_i \tag{10}$$

versus

$$H_{ai}^{**}\colon\ \mu_1^* \le \cdots \le \mu_i \quad \text{with at least one strict inequality} \tag{11}$$

for $i = 2, \ldots, k$, using the size-α likelihood ratio test ϕ_i^{LR}, and then take $\widehat{\mathrm{MED}}$ to be the smallest i such that all $H_{0j}^{**}, j \ge i$, are rejected (e.g., Williams 1971). However, the size-α likelihood ratio test for (10) is not necessarily a size-α test for (6). For example, suppose that $k = 3, \delta = 0$, and $\mu_1 = 0 = \mu_3$ so that H_{03} in (6) is true, while $\mu_2 \to \infty$. Then with probability approaching 1, the isotonic regression estimates of μ_2 and μ_3 under (11) become $(n_2\bar{Y}_2 + n_3\bar{Y}_3)/(n_2 + n_3)$, while the isotonic regression estimates of μ_1 remains $\bar{Y}_1$. Thus the probability that the likelihood ratio test rejects (10) and $\mu_3 > \mu_1^*$ is incorrectly declared approaches 1. Again, the rejection of (10) in favor of (11) does not in general pivot to a positive lower confidence bound on $\mu_i - \mu_1$. For example, when $k = 3$ and $n_1 = n_2 = n_3$, the lower confidence bound on $\mu_3 - \mu_1$ obtained by pivoting ϕ_3^{LR} is $-\infty$ regardless of data. This can be seen as follows. The test $\phi_3^{\mathrm{LR}}(\mu_1^0, \mu_2^0, \mu_3^0)$, which is ϕ_3^{LR} applied to the shifted data $(\hat{\mu}_j - \mu_j^0, j = 1, 2, 3)$ and $\hat{\sigma}$, is clearly a size-α test for H_0: $\mu_1 = \mu_1^0, \mu_2 = \mu_2^0, \mu_3 = \mu_3^0$ against H_a: $\mu_1 - \mu_1^0 \le \mu_2 - \mu_2^0 \le \mu_3 - \mu_3^0$ with at least one strict inequality. If we fix all $\hat{\mu}_j, j = 1, 2, 3$, and μ_1 but let $\mu_2 = -2\mu_3$, then, as $\mu_3 \to -\infty$, the isotonic regression estimates of $\mu_1 - \mu_1^0, \mu_2 - \mu_2^0$, and $\mu_3 - \mu_3^0$ become $((\hat{\mu}_1 - \mu_1^0) + (\hat{\mu}_2 - \mu_2^0) + (\hat{\mu}_3 - \mu_3^0))/3$. Therefore, the set of $(\mu_1^0, \mu_2^0, \mu_3^0)$ for which $\phi_{\mathrm{LR}}^{\mathbf{c}_{i^*}}(\mu_1^0, \ldots, \mu_k^0)$ accepts always includes a sequence for which $\mu_3^0 - \mu_1^0 \to -\infty$. Therefore, by the usual correspondence between tests and confidence sets, the lower confidence bound on $\mu_3 - \mu_1$ is $-\infty$ regardless of data.

The motivation for this third approach is presumably to take advantage of the power of isotonic regression when μ_i is suspected to decrease as i decreases from k to 2. However, even though the validity of the DR method does not depend on the validity of this suspicion, the simulation results described in Section 2.3 indicates that the DR method takes advantage of this suspicion when it is true to much the same extent that Williams's test does.

Finally, sampling as well as confidence set construction can proceed in a stepwise manner. First, the control group and dose k are sampled. If dose k is found to be efficacious, then dose $k - 1$ is sampled, and so on. Failure to declare a dose i to be efficacious renders sampling at doses lower than i unnecessary. This modified method has the distinct advantage of reducing the exposure of patients in clinical trials to possibly ineffective doses. To analyze this stepwise sampling plan, the DR method is modified so that $t_{\alpha,\nu_i}\hat{\sigma}_i$ replaces $t_{\alpha,\nu}\hat{\sigma}$ in comparing μ_i with μ_1^*, where $\nu_i = n_1 + \sum_{j=i}^k n_j - (k - i + 2)$ and

$$\hat{\sigma}_i^2 = \frac{\sum_{a=1}^{n_1}(Y_{1a} - \bar{Y}_1)^2 + \sum_{j=i}^k \sum_{b=1}^{n_j}(Y_{jb} - \bar{Y}_j)^2}{\nu_i}.$$

2.2 An Example

To illustrate the DR method and its difference from Dunnett's method and the MPGN method, consider the data in Table 1, taken from Ruberg (1995). If $\delta = 7$, then Table 2 shows the 95% lower confidence limits on $\mu_i - \mu_1, i = 2, \ldots, 10$, given by the three methods. (Note that for compatibility, the inference given by the MPGN method is presented in terms of its associated confidence bounds, as described in Bofinger 1987, and Hsu 1996, sec. 3.1.1.2). Thus, whereas both Dunnett's method and the MPGN method find doses $6, \ldots, 10$ to be effective (with an $\widehat{\mathrm{MED}}$ of 2.5 mg/kg), the DR method finds doses $5, \ldots, 10$ to be effective (with an $\widehat{\mathrm{MED}}$ of 2.0 mg/kg).

Table 1. Sample Dose–Response Data

Group	Dosage (mg/kg)	Sample size	Mean response	SD response
1	0	6	25.5	2.6
2	.5	6	23.9	4.0
3	1.0	6	27.7	3.3
4	1.5	6	33.4	2.3
5	2.0	6	40.5	10.5
6	2.5	6	57.9	9.9
7	3.0	6	74.4	14.6
8	3.5	6	73.4	7.6
9	4.0	6	73.5	4.5
10	4.5	6	76.2	7.9

Table 2. 95% Simultaneous Lower Confidence Limits on $\mu_i - \mu_1$

Group difference	Estimated difference	DR method	MPGN method	Dunnett's method
2 − 1	−1.6		−11.52	−12.73
3 − 1	2.2		−7.72	−8.93
4 − 1	7.9	.40	−2.02	−3.23
5 − 1	15.0	7.00	5.08	3.87
6 − 1	32.4	7.00	7.00	21.27
7 − 1	48.9	7.00	7.00	37.77
8 − 1	47.9	7.00	7.00	36.77
9 − 1	48.0	7.00	7.00	36.87
10 − 1	50.7	7.00	7.00	39.57

2.3 A Simulation Comparison of Methods for Dose–Response Studies

A simulation study was conducted to compare the behavior of the DR method with methods based on linear contrasts, Helmert contrasts, reverse Helmert contrasts, and Williams's test (as described in Tamhane et al. 1996). For each of four mean configurations with $k = 6$, 10,000 multivariate normal random vector with the identity variance–covariance matrix were generated. Figures 1–4 compare the distributions of MEDs inferred by the five methods. The graph in the upper left corner of each figure shows the true response corresponding to each dose level, with $\mu_1 + \delta$ indicated by the horizontal line. In these figures, an inferred MED of dose level 7 means that none of the dose levels is inferred to be effective. The nominal error rate α was 5% for all our simulations.

The first two μ's are monotonically increasing; for these, all five methods control the error rate. For these two μ's, the methods can be compared in terms of closeness of the inferred MED to the true MED.

For linearly increasing response $\boldsymbol{\mu} = (1, 2, 3, 4, 5, 6)$ and $\delta = 1.5$, a method commits an error only if it infers $\text{MED} = 2$. Figure 1 shows that the linear contrast method, Williams's test, and the DR method do well, whereas the methods based on Helmert contrasts and reverse Helmert contrasts tend to infer a somewhat larger MED.

For logarithmic response $\boldsymbol{\mu} = (0, 2.08, 3.30, 4.16, 4.83, 5.38)$ and $\delta = 2.7$, a method commits an error only if it infers $\text{MED} = 2$. Figure 2 shows that Williams's test and the DR method do well, whereas methods based on linear, Helmert, and reverse Helmert contrasts tend to infer a somewhat larger MED.

For inverted-U response $\boldsymbol{\mu} = (1, 2, 3, 4, 8, 2)$ and $\delta = 1.5$, a method commits an error if it infers any dose level i $(2 \leq i \leq 6)$ to be the MED. Figure 3 shows that Williams's test and methods based on linear and reverse Helmert contrasts have very excessive error rates.

For U-shaped response $\boldsymbol{\mu} = (7.0, 3.5, 0, 2.0, 4.0, 6.0)$ and $\delta = 0$, a method commits an error if it infers any dose level i $(2 \leq i \leq 6)$ to be the MED. Figure 4 shows that the method based on Helmert contrasts has rather excessive error rates.

In short, the DR method does about as well as the best method for each of the response shapes studied but is the only method that always controls the error rate. Also, the other methods are not confidence set methods. They are testing/decision making methods, and we have compared the DR method with them in that context. But the DR method additionally provides a confidence bound for the first noneffective dose. Or, if all doses are deemed effective, the DR method provides a lower confidence bound on how effective they are. None of the other methods provides this additional inference.

A separate simulation study was conducted to see whether the DR method takes advantage of suspected monotonicity in the response, compared with Dunnett's method and the MPGN method, which do not. For linear response with $\boldsymbol{\mu} = (1, 2, 3, 4, 5, 6)$, 10,000 multivariate normal random vectors with the identity variance–covariance matrix were generated. Suppose that $\delta = 1.5$; then dose levels 3–6 are truly effective. Because all three methods control the error rate, we compare them, in Figure 5, in terms of the distribution of the number of truly effective doses that they infer. As can be seen, the DR method does take advantage of suspected monotonicity in the response, tending to infer more truly effective doses as effective than the MPGN method and Dunnett's method.

The DR method makes an inference of the form, "all doses greater than dose i are effective." It is not designed to detect effective doses in the middle of an inverted-U (umbrella) response such as in Figure 3. Dunnett's and the MPGN methods could identify such a subset of effective doses. If an inverted-U response is suspected, then the DR method should not be used, although it will maintain the correct error rate in this case, as in all cases. Dunnett's and the MPGN methods are not stepwise methods and would not have the advantages of stepwise sampling described at the end of Section 2.1. A stepwise method, similar to the DR method, that would be appropriate for inverted-U responses has been described by Berger and Boos (1999).

3. TOXICITY STUDIES

Consider toxicity studies designed to assess the safety of a substance at various dose levels. Let μ_1 be the mean response of the negative control (placebo) group. Let $\mu_2, \ldots, \mu_{k-1}$ be the mean response of $k - 2$ dosed groups, and let μ_k be the mean response of a positive control group that is typically included in a toxicity study. Toxicity analysis is an example where the desired inference is *practical equivalence* between $\mu_i, i = 2, \ldots, k - 1$, and μ_1. The rationale for the inclusion of a positive control group is that if the study fails to detect significant difference between the positive and the negative control groups, which are known to be different, then any lack of observed significant difference between a dosed group and the negative control group may be due to failed experimentation instead of closeness of their mean responses. Suppose that μ_i can be considered equivalent to μ_1 if $\delta_1 < \mu_i - \mu_1 < \delta_2$. Without loss of generality, we can assume that $-\delta_1 = \delta_2 = \delta > 0$ (for otherwise, we can base our inference on $\bar{Y}_i - \bar{Y}_1 - (\delta_1 + \delta_2)/2, i = 2, \ldots, k - 1$). Then $\Theta_1^c = \{\mu_k > \mu_1\}$ and $\Theta_i = \{\mu_i - \mu_1 \in (-\delta, \delta)\}, i = 2, \ldots, k - 1$.

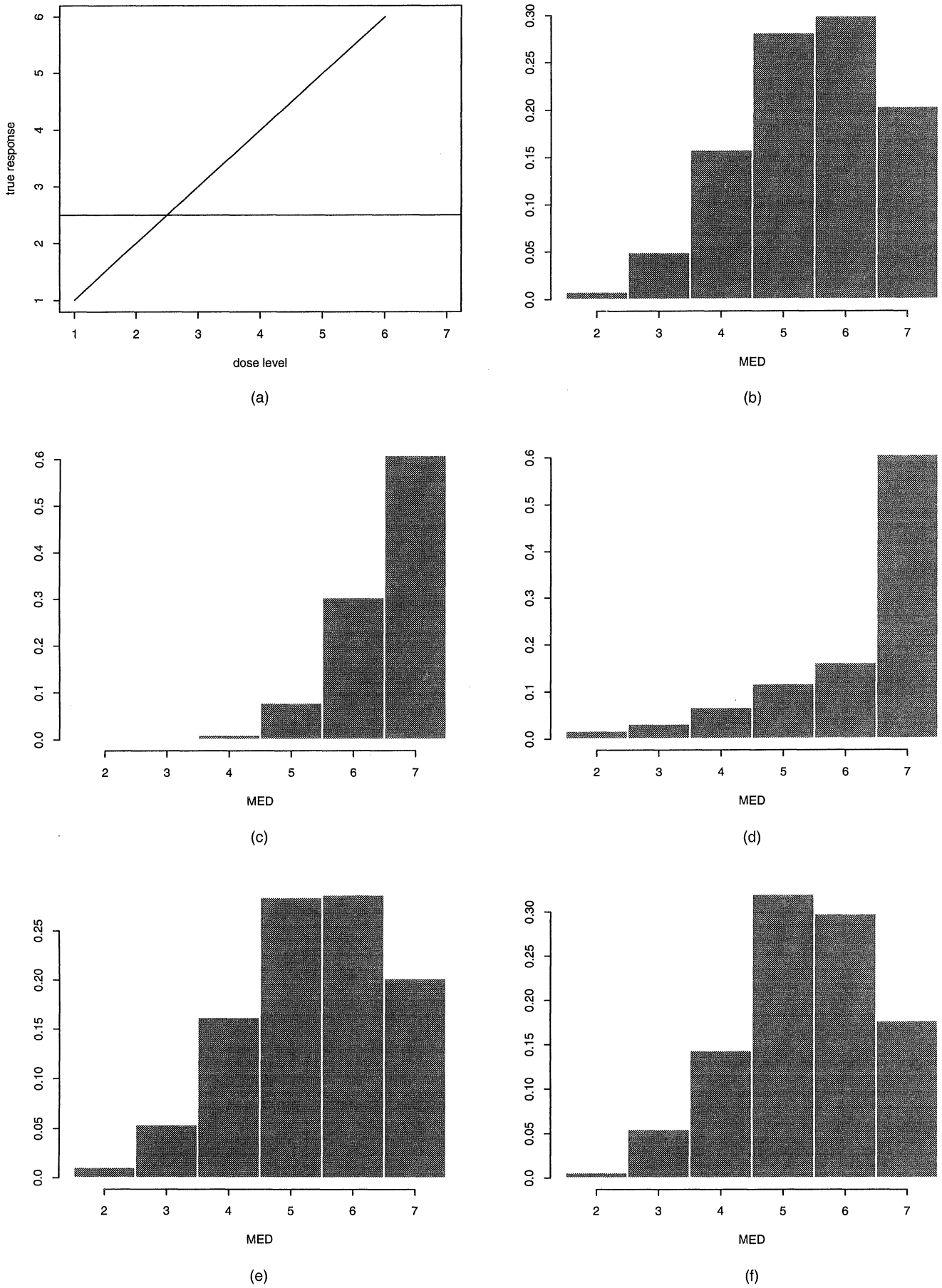

Figure 1. Histograms of Inferred MED: Linear Response. (a) Response curve; (b) DR method; (c) Helmert contrasts method; (d) reverse Helmert contrasts method; (e) Williams' method; (f) linear contrasts method.

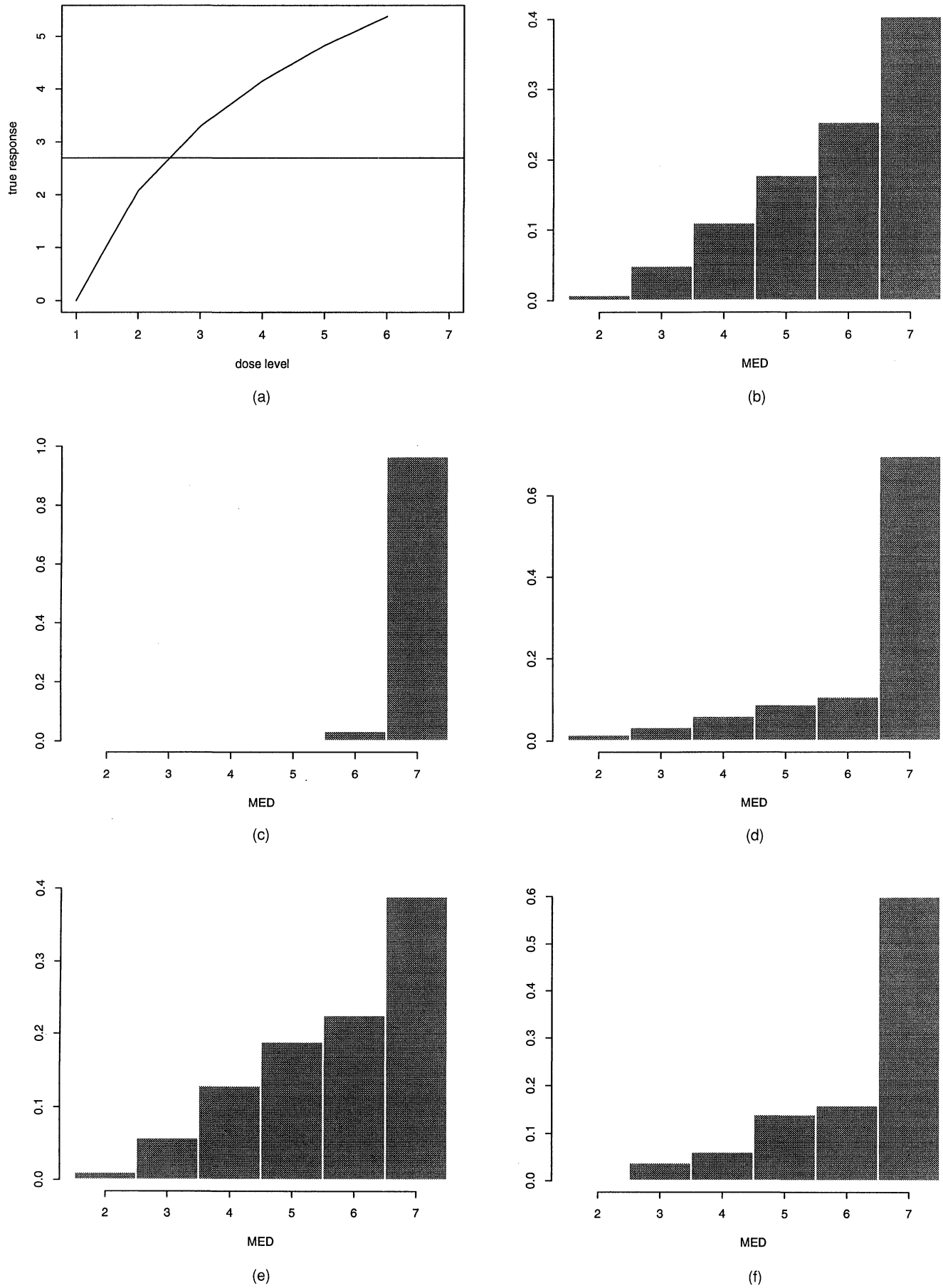

Figure 2. Histograms of Inferred MED: Logarithmic Response. (a) Response curve; (b) DR method; (c) Helmert contrasts method; (d) reverse Helmert contrasts method; (e) Williams' method; (f) linear contrasts method.

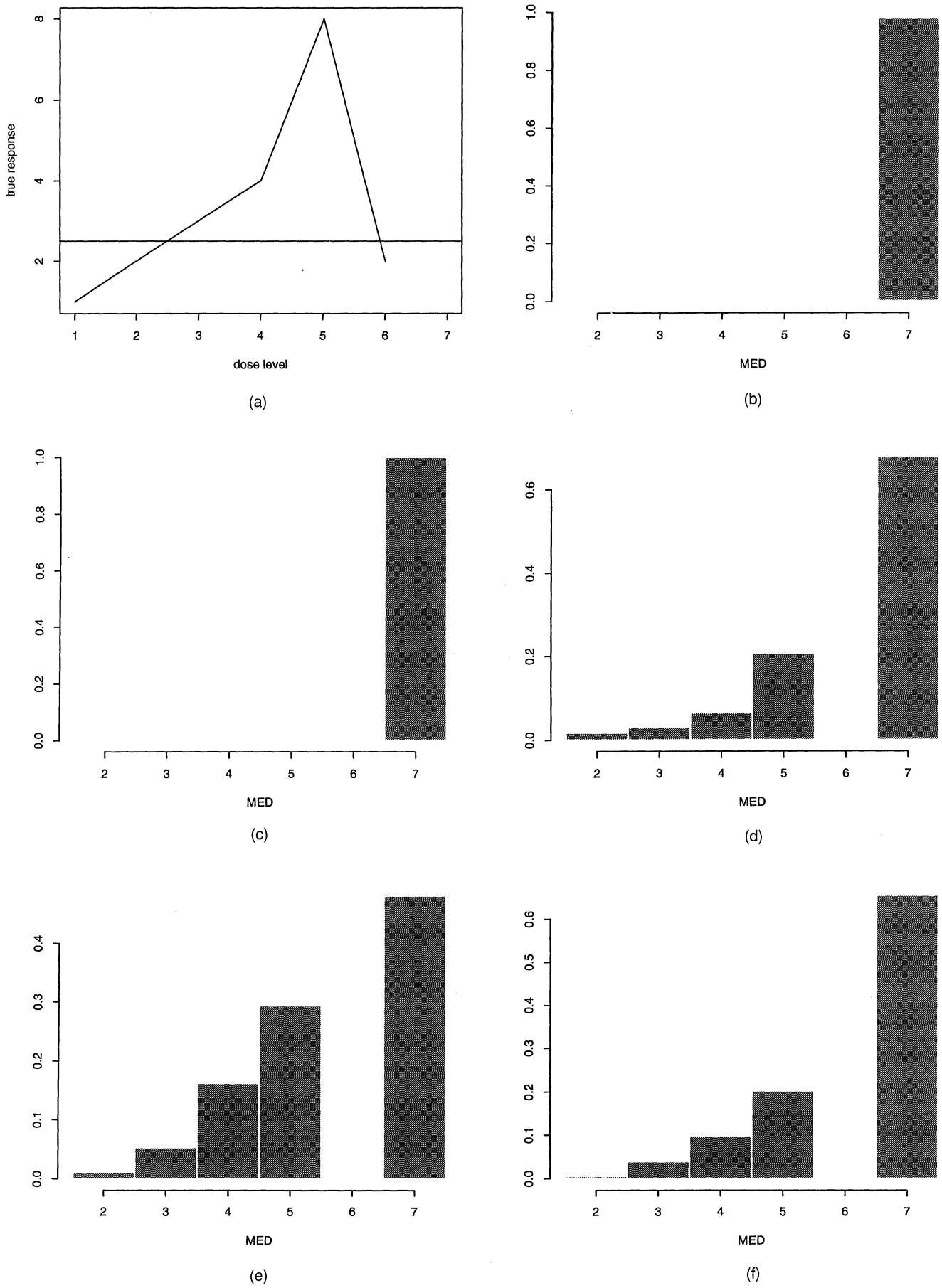

Figure 3. Histograms of Inferred MED: Inverted-U Response. (a) Response curve; (b) DR method; (c) Helmert contrasts method; (d) reverse Helmert contrasts method; (e) Williams' method; (f) linear contrasts method.

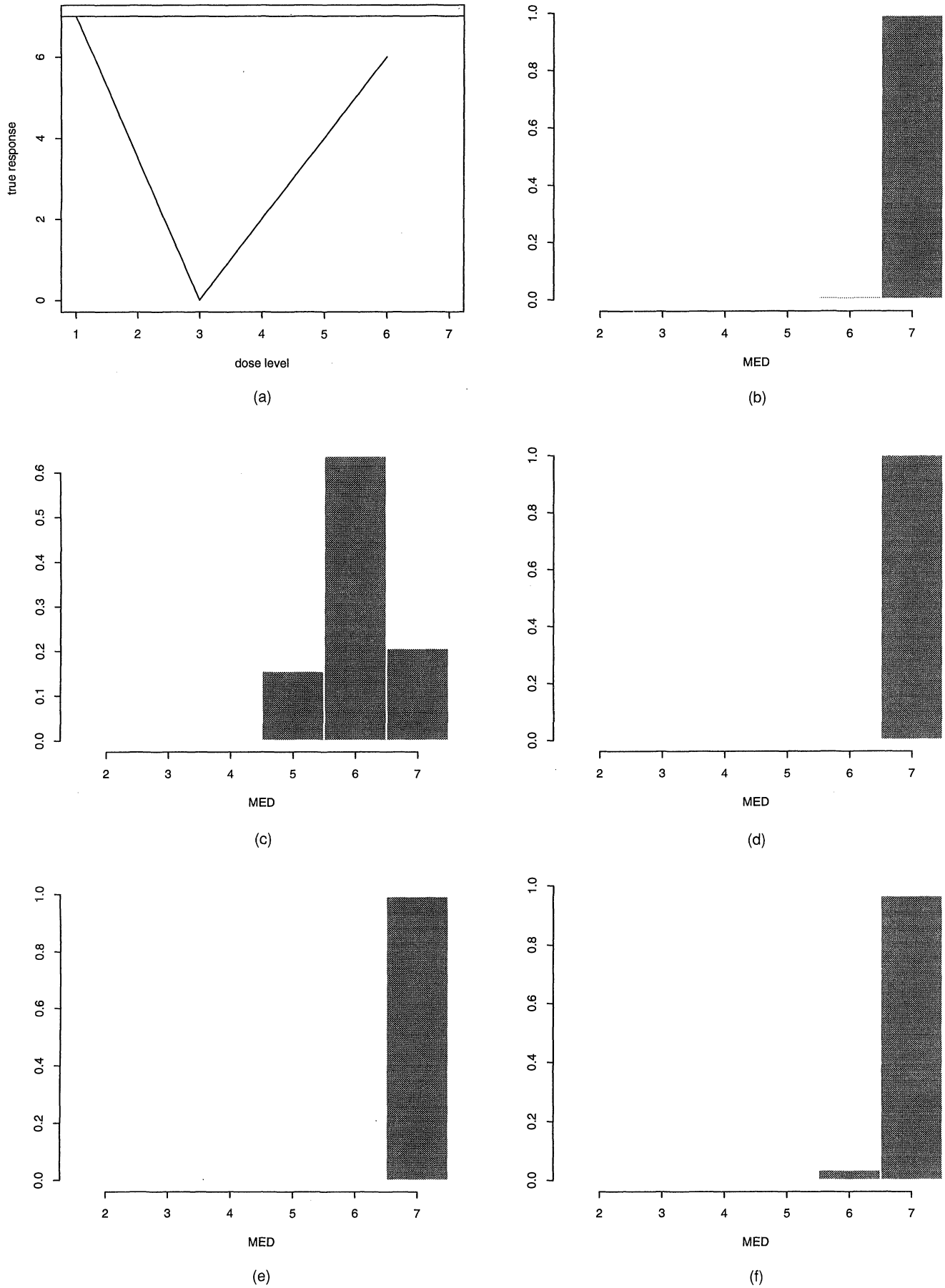

Figure 4. Histograms of Inferred MED: U-Shaped Response. (a) Response curve; (b) DR method; (c) Helmert contrasts method; (d) reverse Helmert contrasts method; (e) Williams' method; (f) linear contrasts method.

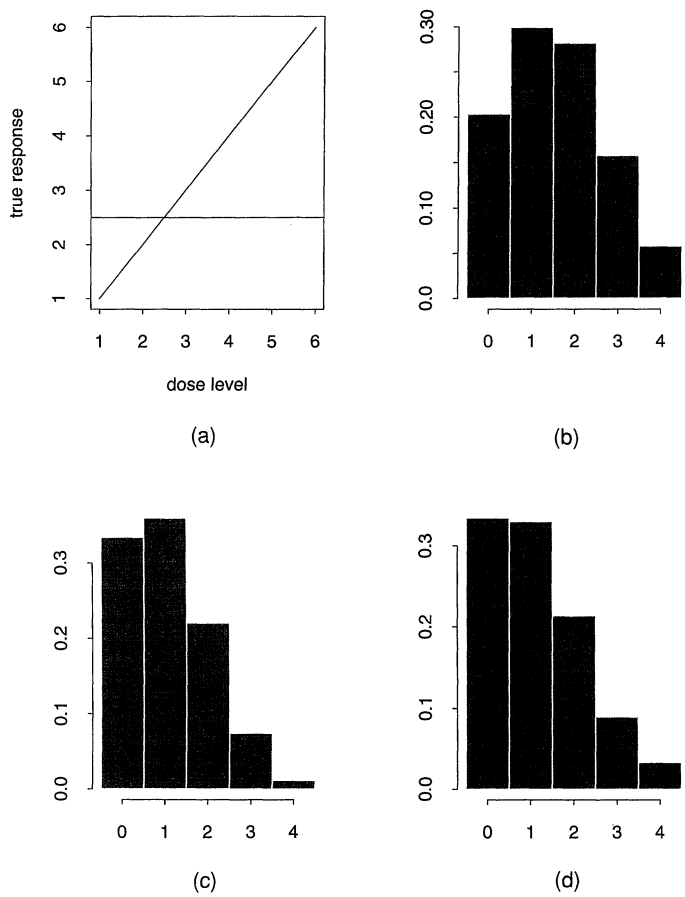

Figure 5. Histograms of Numbers of Truly Effective Doses Inferred. (a) Response curve; (b) DR method; (c) Dunnett's method; (d) MPGN method.

Under a one-way model, $(\bar{Y}_k - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1}, \infty)$ obviously is a $100(1-\alpha)\%$ confidence interval for $\mu_k - \mu_1$ directed toward $\{\mu_k > \mu_1\}$. For $i = 2, \ldots, k-1$, let

$$D_i^- = \min\{\bar{Y}_i - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1}, 0\}$$

and

$$D_i^+ = \max\{\bar{Y}_i - \bar{Y}_1 + t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1}, 0\}.$$

Then

$$D_i = \begin{cases} (D_i^-, D_i^+), & \text{if } D_i^- < 0 < D_i^+ \\ [0, D_i^+), & \text{if } D_i^- = 0 \\ (D_i^-, 0], & \text{if } D_i^+ = 0, \end{cases}$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_1$ (Berger and Hsu 1996). Therefore, using (1), if we let

$$C_i = \begin{cases} D_i, & \text{if } D_i \subset (-\delta, \delta) \\ D_i \bigcup (-\delta, \delta), & \text{otherwise,} \end{cases}$$

then C_i is $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_1$ directed toward $(-\delta, \delta)$. We now show that $\max_{i=2,\ldots,k-1} \max\{-D_i^-, D_i^+\}$ is a $100(1-\alpha)\%$ upper confidence bound for $\max_{i=2,\ldots,k-1} |\mu_i - \mu_1|$; this confidence bound is used at the last stage of the stepwise procedure. Let j be such that $|\mu_j - \mu_1| = \max_{i=2,\ldots,k-1} |\mu_i - \mu_1|$. If $\mu_j \geq \mu_1$, then

$$\begin{aligned} P(\mu_j - \mu_1 < \max_{i=2,\ldots,k-1} \{\bar{Y}_i - \bar{Y}_1 + t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1}\}) \\ \geq P(\mu_j - \mu_1 < \bar{Y}_j - \bar{Y}_1 + t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_j + 1/n_1}) \\ = 1 - \alpha. \end{aligned}$$

A similar proof holds for the case of $\mu_j < \mu_1$.

Our stepwise confidence set method (which we call the TX—for toxicity—method) is as follows.

Step 1.

If $\bar{Y}_k - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_1} \geq 0$
then assert $\mu_k > \mu_1$ and go to Step 2;
else assert $\mu_k - \mu_1 > \bar{Y}_k - \bar{Y}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_1}$ and stop.

Step 2.

If $D_2 \subset (-\delta, \delta)$
then assert $\mu_2 - \mu_1 \in (-\delta, \delta)$ and go to Step 3;
else assert $\mu_2 - \mu_1 \in C_2$ and stop.

Step 3.

If $D_3 \subset (-\delta, \delta)$
then assert $\mu_3 - \mu_1 \in (-\delta, \delta)$ and go to Step 4;
else assert $\mu_3 - \mu_1 \in C_3$ and stop.

⋮

Step $k-1$.

If $D_{k-1} \subset (-\delta, \delta)$
then assert $\mu_{k-1} - \mu_1 \in (-\delta, \delta)$ and go to Step k;
else assert $\mu_{k-1} - \mu_1 \in C_{k-1}$ and stop.

Step k.

Assert $\max_{i=2,\ldots,k-1} |\mu_i - \mu_1| < \max_{i=2,\ldots,k-1} \max\{-D_i^-, D_i^+\}$ and stop.

To better understand how this stepwise method operates, let Step M $(1 \leq M \leq k)$ be the step at which the stepwise method stops. If $M = 1$, then the sensitivity of the experiment is inadequate, and a lower confidence bound (which is negative) for $\mu_k - \mu_1$ is given. If $1 < M < k$, then a confidence set for $\mu_M - \mu_1$ that contains $(-\delta, \delta)$ is given, and the confidence intervals $\mu_i - \mu_1 \in (-\delta, \delta)$ for $i = 2, \ldots, M-1$ are given if $M > 2$. If $M = k$, then a common confidence interval for $\mu_i - \mu_1, i = 2, \ldots, k$, which is entirely contained in $(-\delta, \delta)$, is given.

Recall from Section 2 that instead of Theorem 1, the DR confidence set can be alternatively derived by partitioning the parameter space, applying an appropriate family of tests to each subspace, then invoking the connection between tests and confidence sets. This same technique can be used to alternatively derive the TX confidence set, with each subspace in the partition being the subspace that a step of the TX method seeks to rule out

(e.g., $\{\mu_2 - \mu_1 \le -\delta\} \cup \{\mu_2 - \mu_1 \ge \delta\}$ in step 2), followed by applying tests appropriate for each subspace.

3.1 Comparison With Other Approaches

A common practice of analyzing toxicity data for both carcinogenicity and systematics is to test the hypotheses of equalities between the negative control (placebo) group and the dosed groups,

$$H_{0i}^{=}\colon\ \mu_i = \mu_1 \quad \text{versus} \quad H_{ai}^{=}\colon\ \mu_i \ne \mu_1, \tag{12}$$

for $i = 2, \ldots, k$, with a small p value associated with H_k and large p values associated with $H_i, i = 2, \ldots, k-1$, adjusted for multiplicity, taken as evidence of safety. This is unreasonable, because p values do not provide quantitative information. In fact, regarding carcinogens, the U.S. Environmental Protection Agency (EPA 1993) states that "there is theoretically no level of exposure for such a chemical that does not pose a small, but finite, probability of generating a carcinogenic response"—that is, the EPA's position is all $H_{0i}^{=}\colon\ \mu_i = \mu_1, i = 2, \ldots, k-1$, are false a fortiori. Thus when toxicity studies are formulated as in (12), noncarcinogenicity of any substance can be contradicted by conducting a well-controlled experiment with a large sample size. On the other hand, under this formulation, safety of *any* substance can be concluded by conducting a small, sloppy experiment that includes a potent positive control. A significant p value associated with the positive control does not lend support to the interpretation that nonsignificant p values indicate practical equivalence between the associated treatments and the control; it merely indicates that $\mu_k - \mu_1 \in (0, \infty)$. Adequate "power" designed into a study lends at most partial support, because the "power" of a test of homogeneity includes the probability of directional errors. Even if the sample size is computed to ensure that all sufficiently large differences are detected in the right direction (cf. Hsu 1996, app. C for such computations), there still must be a linkage analysis quantifying the consequence of error in the specification of σ in the sample size computation on the multiple comparison inference, which we have not seen done. We recommend taking the direct approach of the TX method. (Perhaps the δ that would have been specified in the sample size calculation as the difference that should be detected with adequate "power" can serve as the δ defining practical equivalence.)

For a single dose i in a general toxicology setting, Stallard and Whitehead (1996) proposed, in effect, to conclude that $|\mu_i - \mu_1| < \delta$ if a $100(1-2\alpha)\%$ confidence interval for $\mu_i - \mu_1$ is contained in $(-\delta, \delta)$. Note however, as shown in theorem 5 of Berger and Hsu (1996), unless the confidence interval is *equal-tailed* (as D_i is), the probability of an incorrect conclusion may be higher than α.

If the establishment of equivalence between $\mu_i, i = 2, \ldots, k-1$, and μ_1 is posed as testing the hypotheses

$$H_{0i}^{\ne}\colon\ |\mu_i - \mu_1| \ge \delta \quad \text{versus} \quad H_{ai}^{\ne}\colon\ |\mu_i - \mu_1| < \delta,$$

then Dunnett's (1955) two-sided simultaneous confidence intervals for $\mu_i - \mu_1$ or their stepwise version by Bofinger and Bofinger (1995) can be used to assess whether $\mu_i - \mu_1 \in (-\delta, \delta), i = 2, \ldots, k-1$. But either method may declare a noncontiguous set of doses as safe.

Hauschke (1997) and Neuhauser and Hothorn (1997) have formulated the comparisons of $\mu_2, \ldots, \mu_{k-1}$ with μ_1 in toxicity studies as tests of the one-sided hypotheses

$$H_{0i}^{\ge}\colon\ \mu_i - \mu_1 \ge \delta \quad \text{versus} \quad H_{ai}^{\ge}\colon\ \mu_i - \mu_1 < \delta.$$

In toxicity studies for which this formulation is appropriate, our general confidence set technique can be applied with $\Theta_1^c = \{\mu_k > \mu_1\}$ and $\Theta_i = \{\mu_i - \mu_1 < \delta\}, i = 2, \ldots, k-1$.

The TX method can be modified, as in Section 2.1, if sampling and confidence set construction are stepwise. The TX method is modified so that $t_{\alpha,\nu_i}\hat{\sigma}_i$ replaces $t_{\alpha,\nu}\hat{\sigma}$ in comparing μ_i with μ_1, where $\nu_i = \sum_{j=1}^{i} n_j - i$ and

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{i}\sum_{b=1}^{n_j}(Y_{jb} - \bar{Y}_j)^2}{\nu_i}.$$

With this method, failure to declare dose i to be safe renders sampling at doses higher than i unnecessary. This modified TX method has the distinct advantage of reducing the exposure of patients in clinical trials to possibly toxic doses.

3.2 Bovine Growth Hormone Toxicity Study Example

The use of bovine growth hormone is controversial [as the articles, both titled "Udder Insanity," in *Consumer Reports* (Consumer Union 1992) and in *Time* (Horowitz and Thompson 1993) indicate]. Writing for the Federal Food and Drug Administration (FDA), Juskevich and Guyer (1990) reported on a number of experiments that did not indicate bovine growth hormones are harmful if present in milk consumed by humans. A subset of the data from one experiment included in that article gave absolute weights of various organs measured from control hypophysectomized rats and hypophysectomized rats treated orally with the peptide hormone recombinant insulin-like growth factor-I (rIGF-I). In addition to groups given rIGF-I orally, one group was given a negative "saline control," and another group was given rIGF-I via a subcutaneously (sc) implanted osmotic minipump as a positive control. Spleen weights of rats treated for either 17 days by gavage or 15 days by continuous sc infusion are given in Table 3. (To keep the discussion simple, data from one group of rats given bovine serum albumin as a negative "oral protein" control are not included here.)

Table 3. Spleen Weight (in Grams) of Male Rats

Treatment label	Dosage *(mg/kg per day)*	*Sample size*	*Mean weight*	*SEM weight*
1 = none (saline)	0	20	147.6	8.8
2 = oral rIGF-I	.01	20	147.2	5.7
3 = oral rIGF-I	.1	20	149.6	5.8
4 = oral rIGF-I	1.0	20	147.1	6.6
5 = SC Infusion rIGF-I	1.0	10	239.6	17.9

Table 4. Individual and Simultaneous 95% Confidence Intervals on $\mu_i - \mu_1$

Group difference	Individual interval	TX interval	Dunnett's interval
2 − 1	(−18.43, 17.63)	(−20.03, 20.03)	(−27.51, 26.71)
3 − 1	(−16.03, 20.03)	(−20.03, 20.03)	(−25.11, 29.11)
4 − 1	(−18.53, 17.53)	(−20.03, 20.03)	(−27.61, 26.61)
5 − 1	(69.91, ∞)	(0, ∞)	(58.80, 125.20)

Juskevich and Guyer (1990) assessed bovine growth hormone safety by testing the multiple hypotheses of equality (12). Adjusting for multiplicity in accordance with Dunnett's (1955) two-sided method, they reported on the nonsignificant p values associated with $H_{0i}^{=}, i = 2, 3, 4$, and the significant p value associated with $H_{05}^{=}$. Based partly on the nonsignificant p values of the treated groups (except the positive control group) in this and other similar datasets presented in their article, Juskevich and Guyer (1990) stated, "Therefore, the FDA scientists concluded that the use of rbGH in dairy cattle presents no increased health risk to consumers." Apparently, the FDA formulated this safety study statistically as a *significant difference* problem: The substance is considered safe if there is no statistical evidence that it causes any change. However, a large p value associated with a test of equality does not necessarily imply that the difference is close to 0.

We believe that comparison of the three levels of orally fed bovine growth hormone with the saline control should be formulated as a *practical equivalence* problem: Weight gains in rats given any growth hormone are close to weight gains of rats given the saline control. Juskevich and Guyer (1990) did not describe any sample size computation to suggest an appropriate δ. Table 4 shows individual 95% confidence intervals for $\mu_i - \mu_1, i = 2, \ldots, 5$ (D_i for $i = 2, 3, 4$), 95% TX simultaneous confidence intervals assuming $\delta = 25$, and 95% Dunnett simultaneous confidence intervals. Thus, for $\delta = 25$, the TX method is able to infer practical equivalence, whereas Dunnett's method is not.

4. CONCLUDING REMARKS

That some simultaneous inferences need no multiplicity adjustment has been perceived, in the hypotheses testing setting, as a consequence of the closed testing approach. We have shown that the reason is more fundamental. If the sequence of individual inferences is predefined, and failure to achieve the desired inference at any step renders subsequent inferences unnecessary, then confidence sets for such inferences can be obtained without multiplicity adjustment.

Deciding what doses are safe and effective are inherently decision problems. For these problems, the most relevant quantity to consider is the probability of making an incorrect decision, which is not necessarily the same as artificial constructs such as the familywise type I error rate. As we have demonstrated, careless formulation of these problems as tests of equalities have led to statistical methods that do not control the probability of making an incorrect decision. A better approach is to state the desired inferences Θ_i^c on which correct decisions depend, and then act as one would when $\boldsymbol{\theta} \in \Theta_i^c$ only if a confidence set $C(\mathbf{Y})$ for $\boldsymbol{\theta}$ is contained in Θ_i^c.

[Received July 1997. Revised December 1998.]

REFERENCES

Bauer, P. (1997), "A Note on Multiple Testing Procedures in Dose Finding," *Biometrics*, 53, 1125–1128.

Berger, R. L. (1982), "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics*, 24, 295–300.

Berger, R. L., and Boos, D. D. (1999), "Confidence Limits for the Onset and Duration of Treatment Effect," unpublished manuscript submitted to *Biometrical Journal*.

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trials, Intersection-Union Tests, and Equivalence Confidence Sets," *Statistical Science*, 11, 283–315.

Bofinger, E. (1987), "Stepdown Procedures for Comparison With a Control," *Australian Journal of Statistics*, 29, 348–364.

Bofinger, E., and Bofinger, M. (1995), "Equivalence With Respect to a Control: Stepwise Tests," *Journal of the Royal Statistical Society*, Ser. B, 57, 721–733.

Committee for Proprietary Medical Products (CPMP)/EWP/233/95 (1995), *Note for Guidance on the Clinical Investigation of Medical Products in the Treatment of Chronic Peripheral Arterial Occlusive Disease*, The European Agency for the Evaluation of Medical Products, final edition.

Consumer Union (1992), "Udder Insanity," *Consumer Reports*, 57, 330–332.

Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments With a Control," *Journal of the American Statistical Association*, 50, 1096–1121.

Environmental Protection Agency (1993), Reference Dose (RfD): Description and Use in Health Risk Assessment. United States Environmental Protection Agency, Washington, D.C.

Hauschke, D. (1997), "Statistical Proof of Safety in Toxicological Studies," *Drug Information Journal*, 31, 357–361.

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

Horowitz, J. M., and Thompson, D. (1993), "Udder Insanity!" *Time*, 141, 52–53.

Hsu, J. C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman and Hall.

International Conference on Harmonization (1994), *Dose Response Information to Support Drug Registration*, Guideline E4, London: Committee for Proprietary Medical Products, The European Agency for the Evaluation of Medical Products.

Juskevich, J. C., and Guyer, C. G. (1990), "Bovine Growth Hormone: Human Food Safety Evaluation," *Science*, 249, 875–884.

Koch, G. G., and Gansky, S. A. (1996), "Statistical Considerations for Multiplicity in Confirmatory Protocols," *Drug Information Journal*, 30, 523–534.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, 655–660.

Maurer, W., Hothorn, L. A., and Lehmacher, W. (1995), "Multiple Comparisons in Drug Clinical Trials and Preclinical Assays: A Priori Ordered Hypotheses," in *Biometrie in der Chemisch-Pharmazeutischen Industrie*, Vol. 6, ed. J. Vollman, Stuttgart: Fischer-Verlag, 3–18.

Naik, U. D. (1975), "Some Selection Rules for Comparing p Processes With a Standard," *Communications in Statistics, Part A—Theory and Methods*, 4, 519–535.

Neuhauser, M., and Hothorn, L. A. (1997), "The Control of the Consumer Risk in the Ames Assay," *Drug Information Journal*, 31, 363–367.

Ruberg, S. J. (1989), "Contrasts for Identifying the Minimum Effective Dose," *Journal of the American Statistical Association*, 84, 816–822.

——— (1995), "Dose–Response Studies II. Analysis and Interpretation," *Journal of Biopharmaceutical Statistics*, 5, 15–42.

Stallard, N., and Whitehead, A. (1996), "An Alternative Approach to the Analysis of Animal Carcinogenicity Studies," *Regulatory Toxicology and Pharmacology*, 23, 244–248.

Stefansson, G., Kim, W., and Hsu, J. C. (1988), "On Confidence Sets in Multiple Comparisons," in *Statistical Decision Theory and Related Topics IV*, Vol. 2, eds. S. S. Gupta and J. O. Berger, New York: Springer-Verlag, 89–104.

Tamhane, A. C., Hochberg, Y., and Dunnett, C. W. (1996), "Multiple Test Procedures for Dose Finding," *Biometrics*, 52, 21–37.

Williams, D. A. (1971), "A Test for Differences Between Treatment Means When Several Doses are Compared With a Zero Dose," *Biometrics*, 27, 103–117.