A Note on One-Sided Tests with Multiple Endpoints

Michael D. Perlman\(^1\) and Lang Wu\(^2\)

\(^1\)Department of Statistics, University of Washington, Seattle, WA 98195, USA

Email: michael@ms.washington.edu

\(^2\)Department of Statistics, University of British Columbia, Vancouver, B.C., V6T 1Z2, Canada

Email: lang@stat.ubc.ca.

8 January 2003

Abstract

Testing problems with multivariate one-sided alternative hypotheses are common in clinical trials with multiple endpoints. In the case of comparing two treatments, treatment 1 is often preferred if it is superior for at least one of the endpoints and not biologically inferior for the remaining endpoints. Bloch et al. (2001, Biometrics) propose an intersection-union test (IUT) for this testing problem, but their test does not utilize the appropriate multivariate one-sided test. In this note we modify their test by an alternative IUT that does utilize the appropriate one-sided test. Empirical and graphical evidence show that the proposed test is more appropriate for this testing problem.

KEY WORDS: Multiple endpoints, Hotelling’s $T^2$, intersection-union test, likelihood ratio test, multivariate one-sided test.
1 Introduction.

Testing problems with multivariate one-sided alternative hypotheses occur frequently in clinical trials with multiple endpoints. For example, treatments are often evaluated by both efficacy and toxicity, which may be measured by more than one response variable. A treatment is clearly better than its competitor if all components of its mean responses are better (e.g., larger). This corresponds to an alternative hypothesis of the form \( \{ \min_{1 \leq j \leq p} \mu_j > 0 \} \), where \( \mu_j \) denotes the difference of the \( j \)th mean response. As noted by Bloch, Lai, and Tubert-Bitter (2001), however, in practice it may be difficult to demonstrate that each endpoint is significantly better under treatment 1 than under treatment 2, especially when the number \( p \) of endpoints is not small. Instead, treatment 1 will be preferred if it is superior for at least one of the endpoints and not biologically inferior for the remaining endpoints. This corresponds to the alternative hypothesis \( H_1 \) in (1) below.

Suppose that there are two treatment groups, with \( n_i \) subjects in group \( i, i = 1, 2 \). Let \( Y_{ijk} \) denote the \( j \)th response \( (j = 1, \ldots, p) \) of the \( k \)th subject to treatment \( i \), \( \bar{Y}_{ij} \) the sample average of the \( j \)th response to treatment \( i \), and set \( X_j = (\bar{Y}_{1j} - \bar{Y}_{2j}) \) and \( X = (X_1, \ldots, X_p)^t \), where \( t \) denotes transpose. Let \( \eta_{ij} \equiv E(\bar{Y}_{ijk}) \) be the population mean of the \( j \)th response to treatment \( i \). Without loss of generality, we assume that larger values of \( \eta_{ij} \) correspond to better responses. Set \( \mu_j = \eta_{1j} - \eta_{2j} \) and \( \mu = (\mu_1, \ldots, \mu_p)^t \). As in Bloch et al. (2001, Section 3.1), we first assume that the two populations are \( p \)-variate normal with common covariance matrix \( \Sigma \), so \( X \sim N_p(\mu, \Sigma) \). We assume that \( \Sigma \) is positive definite, denoted by \( \Sigma > 0 \). Let \( \hat{\Sigma} \) denote the pooled sample covariance matrix, so \( S \equiv (n_1 + n_2 - 2) \hat{\Sigma} \)
has the Wishart distribution $W_p(\Sigma, n_1 + n_2 - 2)$ (assume that $n_1 + n_2 \geq p + 2$). More general
distributional assumptions will be discussed briefly in Section 4.

Bloch et al. (2001) first considered the problem of testing

$$H_0 : \left\{ \max_{1 \leq j \leq p} \mu_j \leq 0 \right\} \cup \left\{ \min_{1 \leq j \leq p} (\mu_j + \epsilon_j) \leq 0 \right\} \equiv H_0^{(0)} \cup \left\{ H_0^{(1)} \cup \cdots \cup H_0^{(p)} \right\} \text{ vs. } H_1 : \text{not } H_0 \quad (1)$$

where $H_0^{(j)} : \mu_j \leq -\epsilon_j$, $j = 1, \ldots, p$, and the $\epsilon_j$’s are pre-specified positive numbers. Figure
1 shows the set of $\mu$’s that satisfy $H_0$ when $p = 2$ and $\epsilon_1 = \epsilon_2 = 0.5$. Because $H_0$ in (1) is
the union of $H_0^{(0)}$ and $H_0^{(j)}$, $j = 1, \ldots, p$, it is natural to construct a test for $H_0$ by combining
individual tests for $H_0^{(0)}$ and $H_0^{(j)}$ via the intersection-union (IUT) method. Bloch et al. (2001)
proposed to combine the classical Hotelling $T^2$ test for $H_0^* : \mu = 0$ vs. $H_1^* : \mu \neq 0$ with the
univariate one-sided $t$-tests for $H_0^{(j)}$, $j = 1, \ldots, p$. This proposal is somewhat inappropriate,
however, because the $T^2$ test is not designed for one-sided hypotheses such as $H_0^{(0)}$.

**Figure 1 here**

Instead, in this note we suggest that the $T^2$ component test be replaced by the one-sided
likelihood ratio test (LRT) specifically designed for $H_0^{(0)}$. The critical value for the resulting
alternative IUT can be obtained either by the use of conservative bounds or by bootstrap
methods. We then compare the two IUTs by Monte-Carlo simulation. This note concludes
with a brief discussion of alternative formulations of the hypothesis (1) and alternative co-
variance and distributional assumptions.
2 An Alternative Intersection-Union Test

The IUT for (1) proposed by Bloch et al. (2001, eqn. (5)) rejects $H_0$ iff

$$T^2 = c_{n_1,n_2}^2 X^t \tilde{\Sigma}^{-1} X > T^2_{\alpha} \quad \text{and} \quad c_{n_1,n_2}(X_j + \epsilon_j)/\tilde{\sigma}_{jj}^{1/2} > t_\alpha \quad \text{for} \quad j = 1, \ldots, p, \quad (2)$$

where $c_{n_1,n_2}^2 = n_1n_2/(n_1 + n_2)$, $\tilde{\sigma}_{jj}$ is the $j$-th diagonal element of $\tilde{\Sigma}$ and $T^2_{\alpha} \equiv T^2_{p,n_1+n_2-p-1,\alpha}$ and $t_\alpha \equiv t_{n_1+n_2-2,\alpha}$ are the upper $\alpha$ quantiles of the appropriate $T^2$ and $t$ distributions, respectively. If, instead, the size $\alpha$ $T^2$ test is replaced by the size $\alpha$ one-sided LRT for $H_0^{(0)}$ (cf. Perlman (1969, Section 8)), then we obtain the alternative size $\alpha$ IUT for (1) that rejects $H_0$ iff

$$\|X - \mathcal{N}^p\|^2_S \equiv \|X - \pi_S(X; \mathcal{N}^p)\|^2 > c^*_\alpha \quad \text{and} \quad c_{n_1,n_2}(X_j + \epsilon_j)/\tilde{\sigma}_{jj}^{1/2} > t_\alpha \quad \text{for} \quad j = 1, \ldots, p. \quad (3)$$

Here $\mathcal{N}^p \equiv \{(\mu_1, \ldots, \mu_p) \mid \max_{1 \leq j \leq p} \mu_j \leq 0\}$ is the nonpositive orthant in $\mathbb{R}^p$, $\|x\|_S^2 \equiv x^t S^{-1} x$ is the Euclidean norm determined by $S$, $\pi_S(X; \mathcal{N}^p)$ is the projection of $X$ onto $\mathcal{N}^p$ with respect to this norm, and $c^*_\alpha$ is the critical value of the size $\alpha$ one-sided LRT for $H_0^{(0)}$, determined by the equation$^3$

$$\alpha = \frac{1}{2} \Pr \left[ \frac{\chi^2_{p-1}}{\chi^2_{n_1+n_2-p}} > c^*_\alpha \right] + \frac{1}{2} \Pr \left[ \frac{\chi^2_{p-1}}{\chi^2_{n_1+n_2-p-1}} > c^*_\alpha \right] \quad (4)$$

$$= \sup_{\mu \in \mathcal{N}^p : \Sigma > 0} \Pr_{\mu, \Sigma \mid \|X - \mathcal{N}^p\|_S^2 > c^*_\alpha}. \quad (5)$$

$^1$The $T^2$ statistic is incorrectly stated in Bloch et al. (2001, eqn. (5)), where the pooled sample covariance matrix $\tilde{\Sigma}$ is replaced by a non-pooled version $V$. See Anderson (1984, Section 5.3.4).

$^2$Bloch et al. (2001) argue that (2) is level $\alpha$ for (1) because of the monotonicity of the rejection regions of the component tests, but clearly the $T^2$ rejection region is not monotone whether $S$ or $V$ is used – see Fig. 3.

$^3$See Perlman (1969, Theorem 8.3). The monotonicity property required for (5) follows from Lemma 8.2 of Perlman (1969).
Tang (1994) has tabulated some values of $c_{0}^{0}$. The values of $c_{0}^{0}$ can also be determined by numerical methods. Alternatively, the critical values in (3) can be determined by a bootstrap method similar to that in Bloch et al. (2001).

Figure 2 shows the rejection regions of the IUTs (2) and (3) when $p = 2$, $\epsilon_1 = \epsilon_2 = 0.5$, $n_1 = n_2 = 30$, and the sample correlation between $X_1$ and $X_2$ is 0, 0.4, and $-0.4$. These values are similar to those used in Bloch et al. (2001). Although it may appear that the rejection regions of the two tests are similar and that the alternative IUT (3) has a larger rejection region and thus dominates the IUT (2) in terms of power, this is not always the case. This will be discussed in Sections 3 and 4 (see especially Figure 3).

**Figure 2 here**

### 3 Simulation

We use a Monte-Carlo study to compare the IUT (2) (denoted by BL) of Bloch et al. with the proposed alternative IUT (3) (denoted by LR) based on the one-sided LRT. We consider $p = 2$ and $p = 4$, $n_1 = n_2 = 30$, $\epsilon_1 = \ldots = \epsilon_p = 0.5$, and $\alpha = 0.05$. Monte-Carlo samples are generated from $N(\mu, \Sigma)$ for various values of $\mu$ and $\Sigma$. The covariance matrices $\Sigma$ considered have intra-class structure, with common diagonal elements $1$ and common off-diagonal elements $\rho$, with $\rho = 0, 0.4, 0.8, -0.4$, and $-0.8$. Sizes and powers of the tests are simulated based on 5000 iterations.

Table 1 shows the sizes of the two tests. It is clear that both tests attain nominal level
\( \alpha \) when \( \mu \) is on the boundary of the null hypothesis \( H_0 \) but far from the origin \((0, \ldots, 0)\). For \( \mu \in H_0 \) but closer to the origin, both tests can be quite conservative, especially when \( p \) is large. Table 2 shows the powers of the two tests. For the selected points in the alternative space, LR appears to be more powerful than BL in most cases, but LR does not dominate BL in all cases. This is also evident from Figures 1 - 3. However, as will be discussed in Section 4, BL does not have the desired monotonicity property discussed by Bloch et al. since it may reject \( H_0 \) for sample points actually inside \( H_0 \), while LR does not have this drawback. We conclude that LR is preferable to BL. Since Bloch et al. (2001) showed that their test is better than previous tests in the literature, the proposed LR should also be preferable to the previous ones.

**Tables 1 and 2 here**

### 4 Conclusions and Extensions.

For a one-sided testing problem with multiple endpoints, we have proposed a new IUT test that modifies the IUT test of Bloch et al. (2001) by taking fully into account the one-side nature of the null hypothesis. Unlike the test of Bloch et al., the proposed test respects the monotonicity property in the null and alternative hypotheses. Figure 3 shows that the test of Bloch et al. (2001) may possess a partially ellipsoidal acceptance region that does not respect the monotonicity of the orthant and thus may reject \( H_0 \) for sample points actually inside \( H_0 \), while the proposed test does not have this undesirable feature.
Figure 3 here

Bloch et al. (2001, eqn. (8)) also consider the problem of testing

\[ H'_0 : \left\{ \max_{1 \leq j \leq p} \mu_j \leq 0 \right\} \cup \left\{ \min_{1 \leq j \leq p} (\mu_j + \delta_j \tau_j) \leq 0 \right\} \equiv H'_0^{(0)} \cup \{ H_0^{(1)} \cup \cdots \cup H_0^{(p)} \} \text{ vs. } H'_1 : \text{not } H'_0 \quad (6) \]

where \( H_0^{(j)} : \mu_j \leq -\delta_j \tau_j, \tau_j = \text{Var} (\tilde{Y}_{1j} - \tilde{Y}_{2j}) \), and the \( \delta_j \)'s are pre-specified positive numbers.

For this problem, following Lehmann (1986, Section 6.4), it is appropriate to modify (3) to obtain the alternative size \( \alpha \) IUT that rejects \( H'_0 \) iff

\[ \| X - N^p \|_S^2 \equiv \| X - \pi_S(X; N^p) \|_S^2 > c^*_\alpha \quad \text{and} \quad c_{n_1,n_2} X_j / \delta_j^{1/2} > t_\alpha (c_{n_1,n_2} \delta_j) \text{ for } j = 1, \ldots, p, \]

(7)

where now \( t_\alpha (\eta) \equiv t_{n_1+n_2-2,\alpha} (\eta) \) denotes the upper \( \alpha \) quantile of the noncentral \( t \)-distribution with noncentrality parameter \( \eta \). The critical values can be determined as in Section 2.

As in Bloch et al. (2001), the assumptions of normality and common covariance matrices can be relaxed. When these assumptions are relaxed, the critical values for the proposed tests can be determined by a bootstrap method similar to that described in Bloch et al. (2001).

Acknowledgments

This research was supported in part by U.S. National Science Foundation Grant No. DMS 00-71818 and Canada Natural Sciences and Engineering Research Council grant No. 22R80742.
References


Table 1. Simulation Results for the Sizes of the Tests.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Method</th>
<th>( p = 2 )</th>
<th>( p = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_1 )</td>
<td>( \mu_2 )</td>
<td>( \mu_3 )</td>
</tr>
<tr>
<td>( \rho = 0 )</td>
<td>LR</td>
<td>0.014</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td>( \rho = 0.4 )</td>
<td>LR</td>
<td>0.020</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.012</td>
<td>0.006</td>
</tr>
<tr>
<td>( \rho = 0.8 )</td>
<td>LR</td>
<td>0.019</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>( \rho = -0.4 )</td>
<td>LR</td>
<td>0.011</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.006</td>
<td>0.001</td>
</tr>
<tr>
<td>( \rho = -0.8 )</td>
<td>LR</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.002</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: For \( p = 2 \), \( \mu_1 = (0, 0), \mu_2 = (-0.25, 0), \mu_3 = (-0.5, 1) \). For \( p = 4 \), \( \mu_1^* = (0, 0, 0, 0), \mu_2^* = (-0.5, 0, 0, 0), \mu_3^* = (-0.5, -0.5, 0, 0), \mu_4^* = (-0.5, -0.5, -0.5, 0), \mu_5^* = (-0.5, 2, 2, 2), \mu_6^* = (-0.5, -0.5, 2, 2), \mu_7^* = (-0.5, -0.5, -0.5, 2) \).
Table 2. Simulation Results for the Powers of the Tests.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Method</th>
<th>( p = 2 )</th>
<th>( p = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_1 )</td>
<td>( \mu_2 )</td>
<td>( \mu_3 )</td>
</tr>
<tr>
<td>( \rho = 0 )</td>
<td>LR</td>
<td>0.232</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.226</td>
<td>0.261</td>
</tr>
<tr>
<td>( \rho = 0.4 )</td>
<td>LR</td>
<td>0.253</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.252</td>
<td>0.242</td>
</tr>
<tr>
<td>( \rho = 0.8 )</td>
<td>LR</td>
<td>0.244</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.244</td>
<td>0.239</td>
</tr>
<tr>
<td>( \rho = -0.4 )</td>
<td>LR</td>
<td>0.232</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.226</td>
<td>0.244</td>
</tr>
<tr>
<td>( \rho = -0.8 )</td>
<td>LR</td>
<td>0.246</td>
<td>0.238</td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>0.246</td>
<td>0.238</td>
</tr>
</tbody>
</table>

Note: For \( p = 2 \), \( \mu_1 = (-0.25,1) \), \( \mu_2 = (-0.25,2) \), \( \mu_3 = (0.2,0.2) \), \( \mu_4 = (0.5,0.5) \). For \( p = 4 \), \( \mu_1^* = (0,2,2,2) \), \( \mu_2^* = (0,0,2,2) \), \( \mu_3^* = (0,0,0,2) \), \( \mu_4^* = (0.2,0.2,0.2,0.2) \), \( \mu_5^* = (0.5,0.5,0.5,0.5) \), \( \mu_6^* = (0.8,0.8,0.8,0.8) \).
Figure 1. The null parameter space $H_0 = H_0^{(0)} \cup H_0^{(1)} \cup H_0^{(2)}$. The rejection regions of the LR-based test (3) and BL test (2) are described in Section 2.
Figure 2. The rejection regions $\mathcal{R}_{LR}$ of the alternative IUT (3) (solid line $\cdots$) and $\mathcal{R}_{BL}$ of the IUT (2) of Bloch et al. (dashed line $\cdots\cdots$) for $p = 2$, $\epsilon_1 = \epsilon_2 = 0.5$, and the sample correlation $r = 0, 0.4, \text{ or } -0.4$. 
Figure 3. The rejection regions $\mathcal{R}_{LR}$ of the alternative IUT (3) (solid line $\longrightarrow$) and $\mathcal{R}_{BL}$ of the IUT (2) of Bloch et al. (dashed line $\ldots$) for $p = 2$, $\epsilon_1 = \epsilon_2 = 1$, and the sample correlation $r = 0, 0.8$, or $-0.8$. 