Multiplicity Adjustment in Clinical Trials with Multiple Correlated Testing

Boris G. Zaslavsky[1] and Fang Chen[2]

[1] Food and Drug Administration, 10903 New Hampshire Avenue Building #71 Room 1248
Silver Spring, MD 20993-0002; Boris.Zaslavsky@fda.hhs.gov
Tel (240) 402 8842 Fax (301) 595 1240
[2] SAS Institute Inc. FangK.Chen@sas.com Tel (919) 531 0019

Statistical analysis of commonly occurring clinical trials that have correlated primary endpoints are often complex because multiple comparison adjustments are necessary. In practice, most statisticians resort to numerical simulation, even though such approaches can be computationally demanding and are often restricted to specific scenarios. The paper provides analytical solutions to one-sided multiple comparisons adjustment for mean values of multivariate normal data that have known positive definite covariance matrices. We use the maximum of test statistics to control the familywise error rate (FWER). This approach is equivalent to adjusting the minimum *p*-value but is simple to use and enables analytical evaluation. We derive a formula for the cumulative probability functions (CDFs) of the maximal test statistics when the correlations are known to be sufficiently small. When this assumption cannot be justified, we provide majorizing inequalities for the CDFs of the maximal test statistics. In addition, we address calculation of power and testing of conditional hypotheses for correlated primary endpoints. Theoretical results are illustrated by examples and are supported by extensive numerical studies.

Keywords: significance level, multivariate normal distribution, maximal statistics, multiple testing, conditional distribution

1. Introduction

In clinical trials, many variables can be measured during or after a treatment. A small number of these response variables, known as endpoints, are of primary importance and often define the success or failure of a study. Studies of correlated multiple endpoints can often reveal meaningful insights to complex problems, such as cases in which researchers cannot clearly identify outcomes that are primarily linked to a treatment, or cases in which a treatment might have multiple clinically important outcomes and requires a better method to evaluate the effect of the treatment. Statistical analysis of such studies can be complicated due to the necessary multiple comparisons adjustments. Numerous publications have addressed the multiple testing problem (Armitage and Parmar 1986; Pocock et al. 1987; James 1991; D'Agostino et al. 1993; Westfall and Young 1993; Zhang et al.1997; Sankoh et.al. 2003; Lix and Sajobi 2010; Bretz et al. 2010; Dmitrienko et al. 2010; Westfall et al. 2011; Phillipset al. 2013; Huque et al. 2013; Permutt 2013; Chunpeng and Zhang 2013). Much of the focus concentrates on multivariate Gaussian outcomes that are positively correlated. For example, (Dubey 1985; Pocock et al. 1987) generated detailed tables for adjustments of nominal significance levels for positively equicorrelated observations. The nominal significance level is the largest number at which the smallest of *n* (the dimension of the outcomes) one-sided *p*-values preserves an overall one-sided Type I error. Heuristic adjustment formulae (Armitage and Parmar1986; James1991; Dubey 1985; Shi et al. 2012; Julious and McIntyre 2012) for the type I and II errors have been advocated. Zaslavsky and Scott (2012) derived asymptotic formulae for very large numbers of endpoints. Although positive correlations are common in trials, negatively correlated multiple endpoints have their own importance. For example, studies reveal that neurotic hostility (scored by the Buss-Durkee hostility inventory (Siegman et al. 1987)) is inversely associated with the severity of angiographically documented coronary artery disease, whereas nonneurotic hostility scores were positively related with the extent of disease (Siegman et al. 1987. In this paper, we consider multivariate Gaussian endpoints without restrictions on the sign of the correlations, given the covariance matrix is nonnegative definite. Instead of adjusting the minimal p -values to control the family-wise error rate, we use the significance level of the maximal observed test statistics. Although these two approaches are equivalent (Romano and Wolf 1987), the latter simplifies the analytical study of the adjustment process.

Clinicians usually test the hypotheses for a few primary endpoints and do not adjust the *p*-values for many of the potentially important effects. Mathematically, this means that the hypotheses are tested based on the marginal probabilities of some primary endpoints and are averaged over the values of the other effects. As a result, the *p*-values adjustment approach might overlook hidden dependencies among the primary endpoints. Testing hypotheses, using maximal test statistics for the subset of endpoints conditioned on the realizations of the rest of endpoints, is a simple way to draw a conclusion specific to the study (Lehmann and Romano 2005 page 394).

The paper is organized as follows. Section 2 introduces relevant definitions and notation and presents some analytical properties of the CDFs of the maximal test statistics. These properties include an analytical formula for the CDF of the maximal test statistics when the correlations are assumed to be sufficiently small. Section 2 also provides majorizing inequalities on this CDF in a general case where the correlations are large. Section 3 addresses power calculation for multiple correlated endpoints. Section 4 considers testing of conditional hypotheses for correlated primary endpoints. Theoretical results are illustrated using examples and are supported by extensive numerical studies.

2. Testing Unconditional Hypotheses for Correlated Primary Endpoints

Let X_i (for $i = 1,...,n$) be the *i*th element of an *n*-dimensional correlated random variable that follows a multivariate normal distribution with an unknown mean vector ν and a known $n \times n$ covariance matrix Σ. Let σ_i be the square root of the *i*th diagonal entry of Σ. We want to test the null hypothesis of $\nu_i \le \nu_{0i}$ for all *i* against the alternative hypothesis of $\nu_i > \nu_{0i}$ for at least one *i*. The following transformation produces normally distributed random variables with mean values of $\mu_i = \nu_i - \nu_{0i}$ and a correlation matrix $R = \{r_{hl}\}$:

$$Y_i = \frac{X_i - \nu_{0i}}{\sigma_i} \qquad (1)$$

The correlation matrix *R* has *n(n* - 1)/2 independent parameters and ones along the main diagonal, and *R* is assumed to be positive definite. By rudimentary statistics, $E(Y_i) = \mu_i$, $E(Y_i - \mu_i)^2 = 1$, and $r_{hl} = E[(Y_h - \mu_h)(Y_l - \mu_l)]$.

First, we examine the one-sided test. We want to test the null hypothesis $H_{u0} = \{H_{u0}^i\}$ that $\mu_i \le 0$ for all *i* against the alternative hypothesis $H_{ua} = \{H_{ua}^i\}$ that $\mu_i > 0$ for at least one *i*, based on the observations $Y_i = y_i$. The probability of rejecting H_{u0} in favor of H_{ua} is controlled by the $\max_i Y_i$ as follows. Following (Romano and Wolf 2005;

Lehmann and Romano 2005), we define an upper α quantile $y_{u\alpha}$ of the random variable $\max_i Y_i$ to be the smallest y such that $\Pr(\max_i Y_i \geq y \mid \mu_i = 0\,) \leq \alpha$. Because $y_{u\alpha}$ is a decreasing function in α and continuous, $\Pr(\max_i Y_i \geq y_{u\alpha} \mid \mu_i = 0\,) = \alpha$. Let $(\max_i y_i \mid \mu_i = 0\ (i = 1,...,n))$ be a realization of $(\max_i Y_i \mid \mu_i = 0\ (i = 1,...,n))$. If at least one observation exists such that $y_k \geq y_{u\alpha}$ for $1 \leq k \leq n$, then $\Pr(\ \max_i Y_i \geq \max_i y_i > y_{u\alpha}) < \alpha$. Therefore, we can reject H_{u0} at the nominal significance level α . By this method, we control the (weak) FWER under the complete null hypothesis (Westfall and Young 1993): FWER = Pr{Reject at least one H_{u0}^{i}| all H_{u0}^{i} are true}.

Similarly, we define a lower α quantile $y_{l\alpha}$ such that $\Pr(\min_i Y_i \leq y_{l\alpha} \mid \mu = 0) = \alpha$. We reject $H_{l0} : \mu_i \geq 0\ (i = 1,..,n)$ for all *i* in favor of $H_{l0} : \mu_i \leq 0\ (i = 1,..,n)$ if there exists at least one k $(1 \leq k \leq n)$ such that $y_k \leq y_{l\alpha}$. In this case, $\min_i y_i \leq y_k$ and $\Pr(\min_i Y_i \leq y_k \mid \mu_i = 0\ (i = 1,...,n)) \leq \Pr(\min_i Y_i \leq y_{l\alpha} \mid \mu_i = 0\ (i = 1,...,n)) = \alpha$. Under the null hypothesis, the CDFs of the random variables Y_i and $-Y_i$ are identical. Using the identity $\min_i Y_i = \max_i(-Y_i)$, $y_{l\alpha} = -y_{u\alpha}$.

Without loss of generality, we only need to focus on the upper quantile. This focus enables us to simplify the notations: let $y_\alpha \equiv y_{u\alpha}$ and $H_0 = H_{u0}$.

It is helpful to interpret the quantiles for the test statistics, $\max_i Y_i$, in the familiar terms of *p*-values. A usual standard of the outcome of a study that has *n* endpoints is the smallest of *n* one-sided *p*-values that are obtained from the normal test statistics. In order to keep the Type I error α , this *p*-value should be less than a “nominal”s value α_n . When endpoints are correlated, the nominal α_n can be calculated by using the upper quantile of $\max_i Y_i$ (Pocock et al. 1987. Similarly, the minimal *p*-value can be calculated by using the largest observation.

To accomplish this objective, we compute $\alpha_n = 1 - \Phi(y_\alpha)$, where $\Phi(y)$ is the cumulative distribution function (CDF) of a standard normal distribution, and find the smallest *p*-value: $p_{\min} = \min\{p_i = 1 - \Phi(y_i)\}$.

If $p_{\min} \leq \alpha_n$ (that is, $y_k \geq y_\alpha$ for some *k*), the smallest of the *n* one-sided *p*-values preserves an overall one-sided Type I error rate of α .

Example 1. Let $Y' = \{Y_i'\}_{i=1}^{n}$ be an *n*-length vector of independent random variables that are generated from the standard normal distribution (with CDF $\Phi(y')$) for each individual element *i*. Let y_α' be the α quantile of $\max_i Y_i'$. We know that

$\Pr(\max_i Y_i' \le y') = (\Phi(y'))^n$ (Arnold et al. 2008). It follows that $\Phi(y_\alpha') = (1-\alpha)^{1/n} \approx 1-\alpha/n$ and $\alpha_n \approx \alpha/n$. The lower α quantile for independent observations can be calculated by using the identity $\Pr(\min_i Y_i' \le y') = 1-(1-\Phi(y'))^n$

The upper bound of the function $\Pr(\max_i Y_i \ge y \mid \mu_i = 0)$ (the lower bound of the CDF $\Pr(\max_i Y_i \le y \mid \mu_i = 0)$ was studied before (Efron 1997). The following proposition describes the upper bound of the CDF of $\max_i Y_i$ for correlated random variables. This proposition enables us to compute the approximation of needed upper quantile. Let $Y = \{Y_i\}_{i=1}^n \sim N(0,R)$, let $Y' = \{Y_i'\}_{i=1}^n \sim N(0,I)$, and let $r_{hl} \in [0,1]$ be the correlation coefficient between the *h*th and *l*th variable for $(h,l = 1,...,n)$.

Proposition 1
(a) The CDF of $\max_i Y_i$ is an increasing function of its correlation coefficients:

$$\frac{\partial P\{\max_i Y_i \le y\}}{\partial r_{hl}} \ge 0 \text{ for } (h \ne l).$$

(*b*) For $r_{hl} \ge 0$, the CDF of $\max_i Y_i$ can be estimated by using:

$$(\Phi(y))^n \le P\{\max_i Y_i \le y\} \le (\Phi(y))^n + \sum_{h<l} r_{hl}\varphi_2(y,y;\, r_{hl}),$$

where $\varphi_2(y,y;r_{hl}) = (2\pi)^{-1}(1-r_{hl}^2)^{-0.5}\exp[-y^2/(1+r_{hl})]$.
(*c*) For $|r_{hl}|$ that are sufficiently small, the CDF of $\max_i Y_i$ can be approximated by

$$P\{\max_i Y_i \le y\} = (\Phi(y))^n + (f(y))^2(\Phi(y))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|),$$

where $f(y)$ is the PDF of the standard normal distribution.
The proof is given in Appendix 1.

From Proposition (1*b*), it follows that $P\{\max_i Y_i \le y\} \to (\Phi(y))^n$ if $y \to \infty$ (or equivalently $\alpha \to 0$). In other words, the contribution due to the correlation coefficients r_{hl} (the summation term) is negligible. And ignoring correlation among endpoints does not incur loss of precision in tests that involve a very small level of α.

Proposition 1 enables us to make direct comparisons between the upper quantiles and for independent and dependent random variables. First, we consider an equal level of significance for independent and dependent outcomes; that is, $P\{\max_i Y_i' \le y_\alpha'\} = P\{\max_i Y_i \le y_\alpha\} = 1-\alpha$.

From (1*a*), it follows that, for the corresponding quantiles, $y_\alpha' \ge y_\alpha$ if all $r_{hl} \ge 0$ and $y_\alpha' \le y_\alpha$ if all $r_{hl} \le 0$, for $(h \ne l)$.

When $| r_{hl} |$ are sufficiently small, it follows from (1*c*) that $y'_\alpha \geq y_\alpha$ if $\sum_{h<l} r_{hl} > 0$ and $y'_\alpha \leq y_\alpha$ if $\sum_{h<l} r_{hl} < 0$. Therefore, the adjustment for multiplicity is smaller if multiple endpoints are positively correlated (or have a small positive sum of correlation coefficients) rather than independent.

Now, consider the case of the equal quantiles $y'_{\alpha'} = y_\alpha$ that have different significance levels α' and α for independent and dependent random variables, respectively. In this situation, the significance level depends on the sign of the correlation: $\alpha' = P\{\max_i Y'_i \geq y'_{\alpha'}\} > P\{\max_i Y_i \geq y'_{\alpha'} = y_\alpha\} = \alpha$ if $r_{hl} \geq 0$ (or $\sum_{h<l} r_{hl} > 0$ for sufficiently small $| r_{hl} |$) and $\alpha = P\{\max_i Y'_i \geq y'_\alpha\} < P\{\max_i Y_i \geq y'\} = \alpha'$ if $r_{hl} \leq 0$ (or $\sum_{h<l} r_{hl} < 0$ for sufficiently small $| r_{hl} |$) .

Example 2.

Let us exemplify Proposition (1*b*). Let $n = 2$ and $r_{12} = r_{21} = \rho$. Assume $\rho = 0.5$ and $\alpha = 0.05$. Then, the exact $P\{\max(Y_1, Y_2) \leq 1.9157\} = 0.95$, $P\{\max(Y'_1, Y'_2) \leq 1.9157\} = 0.9454 < 0.95$, and $0.5\varphi_2(1.9157, 1.9157, 0.5) = 0.008$. Because $0.95 < 0.9454 + 0.008 = 0.9534$, condition (*b*) holds. Let $\rho = -0.5$ and $\alpha = 0.05$. If $\rho = 0$, then $P\{\max(Y'_1, Y'_2) \leq 1.9545\} = 0.95$. $P\{\max(Y_1, Y_2) \leq 1.9593\} = 0.95$ for the correlated random variables. Then, by monotonicity, $P\{\max(Y_1, Y_2) \leq 1.9545\} < 0.95$. Therefore, condition (*a*) holds. Therefore, we have a Bonferroni's nominal significance level is 0.025, the exact nominal significance level is 0.02503 for $\rho = -0.5$, the exact nominal significance level is 0.0253 for $\rho = 0$, and the exact nominal significance level is 0.0277 for $\rho = 0.5$.

The formula in Proposition (1*c*) can be used to estimate the upper quantiles y_α of the correlated random variables by using the upper quantiles y'_α of independent random variables. The upper quantiles y'_α can be calculated by using the formula $\Phi(y_\alpha) = (1-\alpha)^{1/n}$ (see Example 1).

Corollary 1

When $| r_{hl} |$ are sufficiently small, the following approximations for the upper quantile of the correlated random variables can be used:

$$y_\alpha \approx y'_\alpha - n^{-1} f(y'_\alpha) \sum_{h<l} r_{hl}, \qquad (2a)$$

$$y_\alpha \approx y'_\alpha - f(y'_\alpha)(n-1)\rho / 2 \text{ if } r_{hl} = \rho. \; (2b)$$

The proof is given in Appendix 2.

Next we briefly discuss the accuracy of formula (2*b*) based on numerical studies. Let y_α^e denote the exact α quantile, and let $\alpha(y_\alpha^e) = 1 - \Phi(y_\alpha^e)$ denote the exact nominal significance level. Let y_α^a denote the approximation of the α quantile from Equation (2*a*) and let $\alpha(y_\alpha^a) = 1 - \Phi(y_\alpha^a)$, the corresponding approximate nominal significance level. If $-0.2 \le \rho \le 0.2$ and $n = 1,...,10$, our numerical study shows that $0 \le y_{0.05}^a - y_{0.05}^e \le 0.015$ and $-0.0002 \le \alpha(y_{0.05}^a) - \alpha(y_{0.05}^e) \le 0$. Note that for $\rho = -0.2$, the correlation matrix is positive definite if and only if $n \le 5$, and it is positively semi-definite if and only if $n = 6$. For $\rho > 0.2$, the accuracy of formulae (2*a*) might become too imprecise to be acceptable. For example, if $\rho = 0.5$ and $n = 1,...,10$, then $0 \le y_{0.05}^a - y_{0.05}^e \le 0.0863$ and $-0.0018 \le \alpha(y_{0.05}^a) - \alpha(y_{0.05}^e) \le 0$. Similar results are true for the significance level $\alpha = 0.025$: if $-0.2 \le \rho \le 0.2$ and $n = 1,...,10$, then $0 \le y_{0.025}^a - y_{0.025}^e \le 0.0075$ and $-0.0001 \le \alpha(y_{0.05}^a) - \alpha(y_{0.05}^e) \le 0$. We recommend that the formulae be used when $-0.2 \le \rho \le 0.2$.

Note that the approximation in (2*b*) consistently gives a conservative estimate of the α quantiles. The accuracies of the approximation (2*b*) for $\rho = 0, 0.1, (0.2)\ 0.9$ can be verified using (Pocock et al. 1987, Table 1; Gupta et al. 1973, 1983).

The monotonicity of the CDF of $\max_i Y_i$ is demonstrated using a compound symmetry (CS) type of correlation matrix.

Example 3. Let $r_{hl} = \rho$ for $h \ne l$. The CS correlation matrices are positive definite if and only if $1 > \rho > -1/(n-1)$ and positively semi-defined if and only if $\rho = -1/(n-1)$ or $\rho = 1$ (Tong, 1990, p.105). Here we calculate the CDF values over a grid of ρ and *n*: ρ from -0.2 to 0.95 by 0.05 for *n*=2 to 5; and ρ from -0.1 to 0.95 by 0.05 for *n* = 6 to 10. If $0 \le \rho < 1$, the CDF can be calculated using the exact formula

$$P\{\max_i Y_i \le y\} = \int_{-\infty}^{\infty} \left[\Phi(\frac{y - \sqrt{\rho} x}{\sqrt{1-\rho}}) \right]^n f(x) dx \quad (4)$$

(Steck and Owen 1962; Tong 1990, p.115). Then the α quantile is a solution of the equation

$$1 - \alpha = \int_{-\infty}^{\infty} \left[\Phi(\frac{y - \sqrt{\rho} x}{\sqrt{1-\rho}}) \right]^n f(x) dx.$$

This gives the Bonferroni correction when $\rho = 0$.

From formula (4), it follows that $P\{\max_{1 \le i \le n} Y_i \le y\}$ is a decreasing function of *n*.

By combining the CDF with the numerical integration technique, such as Gaussian quadrature, we can compute the CDF rather precisely. However, the computational cost can potentially explode as the dimension of the problem increases. We believe that sampling-based Monte Carlo calculation provides an adequate level of precision in most scenarios. We draw a large number (10 million) of samples from the multivariate normal distribution and calculate the Monte Carlo expectation of the CDF function for every combination of ρ and *n*. The results are shown in Figure 1 where the smoothness of the lines indicate that y_α is an increasing function of *n* and that a corresponding nominal significance level is a decreasing function of *n*.

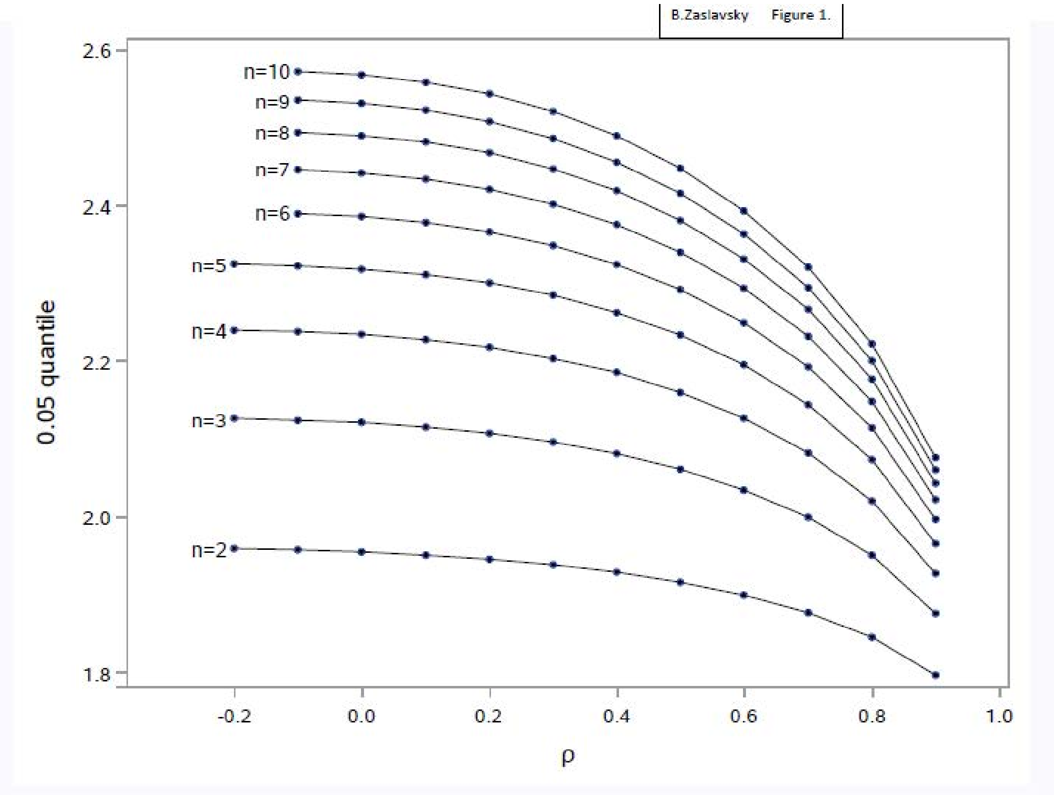

Figure 1. A monotone behavior of the quantile y_α as a function of the number of correlated observations *n* and correlation ρ .

3. Power

This section illustrates how to estimate power in correlated outcomes. Let β be the probability of a Type II error and $1-\beta$ is the power. We consider a clinically important effect $\mu_{0i} > 0\,(i = 1,...,n)$ for each primary endpoint. A number of definitions of power have been proposed in multiple testing scenarios (Westfall and Young 1993, p. 205—206). Here, we restrict ourselves to the adjusted power the minimal or disjunctive power (Senn and Bretz 2007; de Micheaux 2014). The adjusted power is the minimal probability of rejecting H_0 over the set of all alternatives $\mu = (\mu_{01}, ..., \mu_{0n})$ (Westfall and Young 1993, p. 205—206; Romano and Wolf 2005, p. 320). Therefore, the maximum Type II error over the set of alternatives is $\beta = P\{\max_i Z_i \le y_\alpha\}$, where $Z = \{Z_i\}_{i=1}^n \sim N(\mu, R)$. Consider the special case where $\mu_{0i} = \mu_0(i = 1,...,n)$. Then $P\{\max_i Z_i \le y_\alpha\} = P\{\max_i Y_i \le y_\alpha - \mu_0\}$, where $Y = \{Y_i\}_{i=1}^n \sim N(0,R)$. The correlation values are small, so we can calculate the Type II error by using the formula in Proposition (1*c*):

$$\beta = P\{\max_i Y_i \le y_\alpha - \mu_0\} = (\Phi(y_\alpha - \mu_0))^n$$
$$+(2\pi)^{-1}\exp(-(y_\alpha - \mu_0)^2)(\Phi(y_\alpha - \mu_0))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|)$$

If $\sum_{h<l} r_{hl} = 0$, $\beta \approx (\Phi(y_\alpha - \mu_0))^n$ and $\beta = (\Phi(y_\alpha - \mu_0))^n$ is the adjusted Type II error for independent multiple hypotheses (Westfall and Young 1993, p. 206). We also note that the Type II error is a monotone decreasing function of μ_0 .

4. Conditional Hypotheses Testing for Correlated Primary Endpoints

This section discusses approaches in conditional testing, where a hypothesis on a subset of endpoints is conditioned on the rest of endpoints. Let Y_i ($i \in I \subset \{1,...,n\}$) be a vector of length *k* of normalized primary endpoints. We are testing a null hypothesis of $\mu_i \le 0$ versus the alternative of $\mu_i > 0$ for at least one $i \in I \subset \{1,...,n\}$. The number of endpoints that are in the hypothesis is $k < n$, or $I = \{i_1,...,i_k\} \ne \{1,...,n\}$.

Two approaches have been suggested to test this type of hypothesis (Lehmann and Romano 2005, p. 394). One is a marginal test, which ignores unrelated variables and uses the marginal correlation matrix *R* (of dimension *k*) of interests in testing; the other is a conditional approach that conditions on the unrelated endpoints. “If the overall experiment will be performed many times, for example in an industrial or agricultural setting, the average performance may be the principal feature of interest, and an unconditional probability suitable” (Lehmann and Romano 2005, p. 394).The average performance is tested by the marginal probability distribution for Y_i $(i \in I)$ with the correlation matrix $R = \{r_{hl}\}_{h,l \in I}$. “However, if repetitions refer to different clients, or are potential rather than actual, interest will focus on the particular event at hand, and conditional probability seems more appropriate.” (Lehmann and Romano 2005, p. 394). In clinical settings, this largely is driven by the fact that exact conditions are not often repeatable with small number of trials and it becomes difficult to justify the averaging over of other events.

We define a conditional upper α quantile y_α of the random variable $\max_{i \in I} Y_i$ as a value y such that $\Pr(\max_{i \in I} Y_i \ge y \mid \mu_i = 0\,(i \in I); \mu_j, y_j(j \notin I)) = \alpha$. Therefore, the conditional probability of erroneously rejecting the null hypothesis $H_0 : \mu_i \le 0\,(i \in I)$ is α . Because the order of the random variable is irrelevant, without loss of generality, we assume the first *k* variables are of interest. Let $\overline{Y} = (Y_1,...,Y_n)'$ and $\overline{\mu} = (\mu_1,...,\mu_n)'$. They can be partitioned into $\overline{Y}_1 = (Y_1,...,Y_k)', \overline{Y}_2 = (Y_{k+1},...,Y_n)'$ and $\overline{\mu}_1 = (\mu_1,...,\mu_k)', \overline{\mu}_2 = (\mu_{k+1},...,\mu_n)'$. The correlation matrix $R = \{r_{hl}\}$ can be block-partitioned into $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$, where $R_{11} = \{r_{hl}\}(h,l \le k)$. The conditional distribution $\overline{Y}_1 \mid (\overline{Y}_2 = \overline{y}_2)$ is a multivariate normal distribution with the mean vector μ_c and correlation matrix R_c (Tong 1990, p. 34):

$R_c = R_{11} - R_{12}R_{22}^{-1}R_{21}$,

$\overline{\mu}_c = \overline{\mu}_1 - R_{12}R_{22}^{-1}(\overline{\mu}_2 - \overline{y}_2)$.

The following examples compare the marginal and conditional estimates of the 0.05 quantile and the corresponding nominal significance levels.

Example 4. Consider a randomized, double-blind crossover trial of an inhaled active drug versus a placebo, with a detailed description found in (Pocock et al. 1987). Each patient received the active drug and placebo for consecutive four-week periods in a random order. Respiratory functioning was measured at the end of both treatment periods. The measurements were the peak expiratory flow rate (PEFR), forced expiratory volume (FEV1), forced vital capacity (FVC), and penetration index (PI). The correlation matrix for these four measures is shown in Table 1. The hypothesis tested was mean improvement on the active drug. Therefore, we use the one-sided significance test.

Table 1. Correlations among the following respiratory functioning measures: expiratory flow rate (PEFR), forced expiratory volume (FEV1), forced vital capacity (FVC), and penetration index (PI).

FEV1	FVC	PEFR	PI
1	0.095	0.219	-0.162
0.095	1	0.518	-0.059
0.219	0.518	1	0.513
-0.162	-0.059	0.513	1

The PI measurement, which is the ratio of recorded activities between a peripheral and a central lung zone, is difficult to evaluate precisely. “Calculation of PI can be greatly influenced by minor misalignment of regions of interest. This makes it extremely difficult to compare results from different investigators (Bisgaard et al. 2001 p. 194).” In order to achieve a comparable test on the treatment effect, it is desirable to construct either a marginal test (which integrates out PI) or a conditional test (which conditions on the PI measurements). The marginal test is easy to construct, in a multivariate normal case. The average state of respiratory functioning is represented by the joint multivariate normal distribution for PEFR, FEV1, and FVC. The marginal correlation matrix for these three variables can be directly read from Table 1, by removing the column and row for the PI variable. In this case, we can compute the exact 0.05 quantile (which is 2.0923) and the nominal significance level (which is $\alpha_n = 0.0182$). Using the approximation formula (2*a*), we calculate the 0.05 quantile $y_{0.05} \approx 2.10954$ and the nominal significance level of $\alpha_n \approx 0.0174$. The latter is more conservative than the exact value. We can also construct a conditional test that uses the conditional correlation matrix, which is presented in Table 2.

Table 2. Correlations among the following respiratory functioning measures conditioned on the penetration index (PI): expiratory flow rate (PEFR), forced expiratory volume (FEV1), and forced vital capacity (FVC).

FEV1	FVC	PEFR
1	0.0867	0.3566
0.0867	1	0.6398
0.3566	0.6398	1

For simplicity purpose, we assume that the mean value of PI is approximately equal to the observed value of PI. Otherwise, we can nullify the conditional mean value by a translation transformation. The conditional 0.05 quantile is 2.07426, and the nominal significance level is 0.019. Although these estimates are more liberal than the marginal results, both tests demonstrate significance of the treatment effect, measured on a reduced dimension of outcomes. By calculating the conditional significance level for a variety of PIs, an investigator could evaluate the robustness of the study results

Example 5. Lix et al. (2008) reported a study on inflammatory bowel disease. The objective was to assess patients' quality of life and psychological functioning in relation to patterns of the disease activity over time. A number of negative psychological functioning covariates (distress, perceived stress, health anxiety, pain anxiety, and pain catastrophizing) and positive psychological functioning covariates (social support, well-being, and mastery) were considered. The study reported eleven outcomes: inflammatory bowel disease questionnaire (IBDQ), mental health component (SF-36M), physical health component (SF-36P), distress, stress, and health anxiety questionnaire (HAQ), pain anxiety symptom scale (PASS), pain catastrophizing (Catast); social support (Soc Sup), psychological well-being (PWB), and mastery (Mast). The 11×11 correlation matrix is shown in Table 3.

Table 3. Correlations between the characteristics of inflammatory bowel disease and the quality-of-life parameters. The data were collected on the following 11 outcomes: IBDQ, SF-36M, SF-36P, distress, stress, HAQ, PASS, Catast, Soc Sup, PWB, and Mast.

IBDQ	SF-36M	SF-36P	Distress	Stress	HAQ	PASS	Catast	Soc Sup	PWB	Mast
1	0.32	0.18	-0.64	-0.56	-0.56	-0.42	-0.13	0.22	0.52	0.38
0.32	1	-0.48	-0.46	-0.44	-0.34	-0.35	-0.17	0.09	0.25	0.25
0.18	-0.48	1	-0.03	-0.01	-0.01	-0.03	0.17	0.09	0.18	0.12
-0.64	-0.46	-0.03	1	0.69	0.57	0.48	0.11	-0.3	-0.71	-0.51
-0.56	-0.44	-0.01	0.69	1	0.54	0.42	0.03	-0.33	-0.63	-0.56
-0.56	-0.34	-0.01	0.57	0.54	1	0.55	0.16	-0.13	-0.44	-0.34
-0.42	-0.35	-0.03	0.48	0.42	0.55	1	0.21	-0.09	-0.37	-0.36
-0.13	-0.17	0.17	0.11	0.03	0.16	0.21	1	0.04	0.03	-0.01
0.22	0.09	0.09	-0.3	-0.33	-0.13	-0.09	0.04	1	0.31	0.26
0.52	0.25	0.18	-0.71	-0.63	-0.44	-0.37	0.03	0.31	1	0.53
0.38	0.25	0.12	-0.51	-0.56	-0.34	-0.36	-0.01	0.26	0.53	1

We use three variables (SF-36, SF-36P, and PWB) as the primary characteristics of well-being. The average state of well-being is presented by the marginal normal distribution, and the corresponding 3×3 correlation matrix are cell entries in Table 3 that are indexed by variables SF-36M, SF-36P, and PWB. The 0.05 quantile of this three-dimensional marginal distribution is 2.1130, and the nominal significance level is $\alpha_n = 0.0173$. In this case, the conditional test is more appropriate for studying the improvement of well-being with respect to different aspects of psychological functioning. The correlation matrix for the three primary endpoints, conditioned on the remaining eight aspects of psychological functioning, is shown in Table 4.

Table 4. Conditional correlations between the primary characteristics of well-being, conditioned on the eight secondary variables.

SF-36M	SF-36P	PWB
1	-0.5686	-0.1585
-0.5686	1	0.1809
-0.1585	0.1809	1

If the mean values of the IBDQ, Distress, Stress, HAQ, PASS, Catast, Soc Sup, and Mast variables are approximately equal to the observed values, then the 0.05 conditional quantile estimate is 2.121 and the nominal conditional significance level is $\alpha = 0.0169$. The unconditional and conditional estimates of the nominal significance level are very close. This example illustrates the argument that using high-dimensional correlation models might not be much more advantageous than using concise models with properly chosen primary variables.

5. Conclusion

After studying the problem of testing statistical hypotheses for correlated multivariate normal endpoints, we propose that the maximum value of a sample of correlated random variables be used to calculate the nominal significance level at which the smallest one-sided *p*-value preserves an overall one-sided Type I error. When the random variables exhibit a low level of correlation, we obtained an analytical expression for the CDF of the maximum of multiple correlated variables. This CDF formula enables precise estimation for the quantile of the maximum in question. When the correlations are considered to be more pronounced, we provide an analytical expression for the lower and upper limits on CDF of the maximum of multiple correlated variables. This proposed approach allows for extension to power calculation in addition to multivariate conditional testing. We use studies reported in the literature to illustrate our approach. Overall, we believe that this approach provides a practical solution to the multiple comparisons adjustment in correlated outcomes and can be considered as a viable alternative to existing approaches.

Disclaimer. No official support or endorsement of this article by the Food and Drug Administration is intended or should be inferred.

Appendix 1

Here we present proofs for various parts of Proposition 1. We adapted the approach presented in (Slepian 1962) ; Berman 1964) and directly used some of the results by these authors. Let $\varphi_n(z_1,...,z_n;\{r_{hl}\})$ be the *n*-dimensional Gaussian density function with mean vector 0 and covariance matrix $R = \{r_{hl} \mid h,l = 1,...,n\}$. Let $Q_n(y,R) = \int_{-\infty}^{y} ... \int_{-\infty}^{y} \varphi_n(z_1,...,z_n;R) \prod_{i=1}^{n} dz_i$. It was shown that Q_n is an increasing function of r_{hl} (Slepian 1962,; Berman 1964). In addition,

$$\partial Q_n(y,R) / \partial r_{hl} = \int_{-\infty}^{y} ... \int_{-\infty}^{y} \varphi_n(z_1,...,z_h = y,...,z_l = y,...,z_n;R) \prod_{i \neq h, i \neq l}^{n} dz_i \geq 0 . \text{ (A.1)}$$

Under the null hypothesis, $P\{\max_i Y_i \leq y\} = Q_n(y,R)$ and $P\{\max_i Y_i' \leq y\} = Q_n(y,R)$. Because of (A.1), Proposition (1a) follows.

Let us replace the upper limit of integration in the (*n*-2)-fold integral in equation (A.1) of $\partial Q_n(y,R)/\partial r_{hl}$ by $(\infty,...,\infty)$. Then,

$\partial Q_n(y,R)/\partial r_{hl} \le \int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\varphi_n(.)\prod_{i\ne h,i\ne l}^{n} dz_i = \varphi_2(y,y;r_{hl})$ (Berman 1964). By monotonicity of $\varphi_2(y,y,r_{hl})$ as a function of $r_{hl}\ge 0$, $\partial Q_n(y,\{r''_{ij}\})/\partial r_{hl} \le \varphi_2(y,y,r''_{hl})$ and $\varphi_2(y,y,r''_{hl}) \le \varphi_2(y,y,r_{hl})$ if $r''_{hl}\le r_{hl}$. By the law of the mean,

$Q_n(y,R) = Q_n(y,I) + \sum_{h<l} r_{hl}\partial Q_n(y,\{r''_{ij}\}_{i<j})/\partial r_{hl}$ for some $\{r''_{ij}\}$. This proves Proposition (1*b*).

By (A.1) $\partial Q_n(y,\{0_{ij}\}_{i<j})/\partial r_{hl} = (2\pi)^{-1}\exp(-y^2)(2\pi)^{-(n-2)/2}\left(\int_{-\infty}^{y}\exp(-z^2/2)dz\right)^{n-2}$.

This proves Proposition (1*c*).

Appendix 2

Here we provide proof to Corollary 1. Let y'_α and y_α be the α quantiles for the uncorrelated and correlated observations, respectively. Let $\Delta y = y'_\alpha - y_\alpha$. By definition, $(\Phi(y'_\alpha))^n = 1-\alpha$ and $P\{\max_i Y_i \le y_\alpha\} = 1-\alpha$. From Proposition (1*c*), it follows that

$P\{\max_i Y_i \le y_\alpha\} = (\Phi(y_\alpha))^n + (f(y_\alpha))^2(\Phi(y_\alpha))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|)$. Then

$1-\alpha = 1-\alpha + (\Phi(y_\alpha))^n - (\Phi(y'_\alpha))^n +$

$(f(y_\alpha))^2(\Phi(y_\alpha))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|) + o(\Delta y)$

and $(\Phi(y'_\alpha))^n - (\Phi(y_\alpha))^n = (f(y_\alpha))^2(\Phi(y_\alpha))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|)$.

And it follows that

$(\Phi(y_\alpha))^n = (\Phi(y'_\alpha))^n - \Delta y\partial(\Phi(y'_\alpha))^n/\partial y + o(\Delta y)$, where

$\partial(\Phi(y'_\alpha))^n/\partial y = n(\Phi(y'_\alpha))^{n-1} f(y'_\alpha)$ and $f(y'_\alpha)$ is a Gaussian PDF. Thus,

$1-\alpha = 1-\alpha - \Delta y n(\Phi(y'_\alpha))^{n-1} f(y'_\alpha) +$

$(f(y\ ))^2(\Phi(y))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|) + o(\Delta y)$

Because the functions in this expression are smooth and restricted in y'_α, we can assume that $\Delta y = O(\max|r_{hl}|)$. Therefore,

$(f(y_\alpha))^2(\Phi(y_\alpha))^{n-2}\sum_{h<l} r_{hl} = (f(y'_\alpha))^2(\Phi(y'_\alpha))^{n-2}\sum_{h<l} r_{hl} + o(\max|r_{hl}|^2)$. Finally, we get $\Delta y = n^{-1}(f(y'_\alpha))(\Phi(y'_\alpha))^{-1}\sum_{h<l} r_{hl} + o(\max|r_{hl}|) + o(\Delta y)$. Considering that $\Phi(y'_\alpha) \approx 1-\alpha/n$ when α is small, we get

$\Delta y = n^{-1}(f(y'_\alpha))\sum_{h<l} r_{hl} + o(\max|r_{hl}|) + o(\Delta y) + o(\alpha)$. If $r_{hl} = \rho$, it follows that

$\Delta y = f(y'_\alpha)(n-1)\rho/2 + o(\max|r_{hl}|) + o(\Delta y) + o(\alpha)$.

Appendix 3

Here we verify the precision of the approximation of the formula (1*b*) in Proposition 1.

We use the representation $\int_{-\infty}^{z(1-\rho)/\sqrt{1-\rho^2}} e^{-x/2}dx = \int_{-\infty}^{z} e^{-x/2}dx + \int_{z}^{z(1-\rho)/\sqrt{1-\rho^2}} e^{-x/2}dx$. For small ρ, the two expressions are equal up to $o(\rho)$: $\frac{1}{\pi}\int_{-\infty}^{y} e^{-z^2/2} \int_{z}^{z(1-\rho)/\sqrt{1-\rho^2}} e^{-x/2}dzdx$ and $\rho\varphi_2(y,y;\rho) = \rho(2\pi)^{-1}(1-\rho^2)^{-0.5}\exp[-y^2/(1+\rho)]$.

REFERENCES

Armitage, P. and Parmar, M. (1986), Some Approaches to The Problem Of Multiplicity in Clinical Trials, *Proceedings of the Thirteenth International Biometric Conference*, Seattle: Biometric Society, 1-15

Arnold, B. C., Balakrishnan, N., Nagaraja, H. N. (2008), *A First Course in Order Statistics*, Philadelphia: SIAM.

Berman, S. M. (1964), Limit Theorems for The Maximum Term in Stationary Sequences, *Annals of Mathematical Statistics,* 35, 502–516.

Bisgaard, H., O'Callaghan, C., Smaldone, C. C. (2001), *Drug Delivery to the Lung*, New York: Marcel Dekker.

Bretz, F., Hothorn, T., Westfall, P. (2010), *Multiple Comparisons with R*. Boca Raton, FL: Chapman & Hall.

Chunpeng, F. and Donghui, Z. (2013), Sample Size Determination in Two-Sided Distribution-Free Treatment Versus Control Multiple Comparisons, *Journal of Biopharmaceutical Statistics,* 23, 1308-1329, DOI: 10.1080/10543406.2013.834921.

Gupta, S., Panchapakesan, S., and Sohn, J.K. (1983) On The Distribution Of The Studentized Maximum Of Equally Correlated Normal Random Variables. *Technical Report #83-31*: Department of Statistics, Purdue University

D'Agostino, R. B., Massaro, J, Kwan, H., Cabral, H. (1993), Strategies For Dealing with Multiple Treatment Comparisons in Confirmatory Clinical Trials, *Drug Information Journal*, 27, 625–641, DOI: 10.1177/009286159302700307.

de Micheaux, P. L. , Liquet, B., Marque, S. and Riou, J. (2014), Power and Sample Size Determination in Clinical Trials with Multiple Primary Continuous Correlated Endpoints, *Journal of Biopharmaceutical Stati*stics, 24, 378–397.

Dmitrienko, A., Tamhane, A. C., Bretz, F. (2010), *Multiple Testing Problems in Pharmaceutical Statistics*, Boca Raton: Chapman & Hall / CRC.

Dubey, S. D. (1985), Adjustment of P-Values For Multiplicities of Intercorrelating Symptoms, Düsseldorf, Germany: *Proceedings of the Sixth International Society for Clinical Biostatisticians*.

Efron, B. (1977), The Length Heuristic for Simultaneous Hypothesis Tests, *Biometrika*, 84, 143-157.

Gupta, S., Nagel, K., and Panchapakesan, S. (1973), On the Order Statistics from Equally Correlated Normal Random Variables. *Biometrika*, 60, 403-413.

Gupta, S. (1963), Probability Integrals of Multivariate Normal and Multivariate. *The Annals of Mathematical Statistics*, 34, 792-828.

Huque, M. F. Dmitrienko, A. and D'Agostino, R. (2013), Multiplicity Issues in Clinical Trials with Multiple Objectives. *Statistics in Biopharmaceutical Research*, 5:4, 321-337, DOI: 10.1080/19466315.2013.807749

James, S. (1991), Approximate Multinormal Probabilities Applied to Correlated Multiple Endpoints in Clinical Trials, *Statistics in Medicine,* 10, 1123–1135.

Julious, S. A., McIntyre, N. E. (2012), Sample Sizes for Trials Involving Multiple Correlated Must-Win Comparisons, *Pharmaceutical Statistics,* 11, 177–185.

Lehmann, E. L., Romano, J. P. (2005), *Testing Statistical Hypotheses*, New York: Springer 3d edition.

Lix, L. M., Graff, L. A., Walker, J. R., Clara, I., et al. (2008), Longitudinal Study of Quality of Life And Psychological Functioning for Active, Fluctuating, and Inactive Disease Patterns in Inflammatory Bowel Disease, *Inflammatory Bowel Disease,* 14,1576–1584.

Lix, L. M. and Sajobi, T. (2010), Testing Multiple Outcomes in Repeated Measures Designs, *Psychological Methods,* 15, 268–280.

Nadarajah, S, Kotz, S. (2008), Exact Distribution of The Max/Min of Two Gaussian Random Variables, *IEEE Transactions on Very Large Scale Integration (VLSI) System*s 16(2), 210–212.

Permutt, T, Multiplicity in Regulatory Statistical Review (2013), *Statistics in Biopharmaceutical Research,* 5, 4, 394-401, DOI: 10.1080/19466315.2013.851032.

Phillips, A., Chrissie Fletcher, Gary Atkinson, Eddie Channon, et al. (2013), Multiplicity: Discussion Points from the Statisticians in the Pharmaceutical Industry Multiplicity Expert Group, *Pharmaceutical Stat*istics, 12, 255–259.

Pocock, S. J., Geller, N.L., Tsiatis, A. A. (1987), The Analysis of Multiple Endpoints in Clinical Trials, *Biometrics,* 43, 487–498.

Romano, J. P., Wolf, M, Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. (2005), *Journal of the American Statistical Association—Theory and Methods* 469, 94–108, DOI: 10.1198/016214504000000539.

Sankoh, A. J., D'Agostino, R.B., Huque, M.F. (2003), Efficacy Endpoint Selection and Multiplicity Adjustment Methods in Clinical Trials with Inherent Multiple Endpoint Issues, *Statistics in Medicine,* 22, 3133–3150, DOI: 10.1002/sim.1557.

Senn, S. and Bretz, F. (2007), Power and Sample Size When Multiple Endpoints Are Considered, *Pharmaceutical Statistics,* 6, 161–170.

Shi, Q., Pavey E. S. and Carter, R. E. (2012), Bonferroni-Based Correction Factor for Multiple, Correlated Endpoints, *Pharmaceutical Statistics,* 11, 300–309.

Siegman, A. W., Dembroski, T. M., Ringel, N. (1987), Components of Hostility and the Severity of Coronary Artery Disease, *Psychosomatic Medicine,* 49, 127–135.

Steck, G. P., Owen, D. B. (1962), A Note on the Equicorrelated Multivariate Normal Distribution, *Biometrika,* 49, 269–271.

Tong, Y. L.(1990), *The Multivariate Normal Distributions*, New York: Springer-Verlag.

Slepian, D. (1962). The One-Sided Barrier Problem for Gaussian Noise, *Bell System Technical Journal,* 41, 463–501.

Westfall, P., Tobias, R., Wolfinger, R. (2011), *Multiple Comparisons and Multiple Tests Using SAS* (2nd ed.): Cary, NC: SAS Press.

Westfall P, H., Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, New York: John Wiley & Sons.

Zaslavsky, B. G., Scott, J. (2012), Sample Size Estimation in Single-Arm Clinical Trials with Multiple Testing Under Frequentist And Bayesian Approaches, *Journal of Biopharmaceutical Statistics,* 22(4), 819–835.

Zhang, J., Quan, H., Ng, J., Stepanavage, M. E. (1997), Some Statistical Methods for Multiple Endpoints In Clinical Trials, *Controlled Clinical Trials*,18, 204–221.