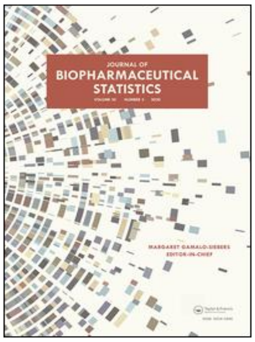

Taylor & Francis
Taylor & Francis Group

Journal of Biopharmaceutical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Incorporating the sample correlation into the testing of two endpoints in clinical trials

Sanat Sarkar, Dror Rom & Jaclyn McTague

To cite this article: Sanat Sarkar, Dror Rom & Jaclyn McTague (2021): Incorporating the sample correlation into the testing of two endpoints in clinical trials, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2021.1895191

To link to this article: https://doi.org/10.1080/10543406.2021.1895191

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2021.1895191

Taylor & Francis
Taylor & Francis Group

ARTICLE

OPEN ACCESS

Check for updates

Incorporating the sample correlation into the testing of two endpoints in clinical trials

Sanat Sarkar[a], Dror Rom[b], and Jaclyn McTague[b]

[a]Department of Statistical Science, Temple University, Philadelphia, USA; [b]Logecal Data Analytics, Wayne, Pennsylvania, USA

ABSTRACT

We introduce an improved Bonferroni method for testing two primary endpoints in clinical trial settings using a new data-adaptive critical value that explicitly incorporates the sample correlation coefficient. Our methodology is developed for the usual Student's t-test statistics for testing the means under normal distributional setting with unknown population correlation and variances. Specifically, we construct a confidence interval for the unknown population correlation and show that the estimated type-1 error rate of the Bonferroni method with the population correlation being estimated by its lower confidence limit can be bounded from above less conservatively than using the traditional Bonferroni upper bound. We also compare the new procedure with other procedures commonly used for the multiple testing problem addressed in this paper.

ARTICLE HISTORY
Received 12 March 2020
Accepted 26 January 2021

KEYWORDS
Bonferroni test; closed testing; familywise-error rate; multiple endpoints; Simes' test

1. Introduction

Pivotal clinical trials for new treatments that are designed to evaluate two primary efficacy endpoints face the so-called 'multiplicity problem', which, if not addressed, may cause inflation of type-1 error. Accordingly, regulatory agencies require that analysis plans contain a statistical methodology for type-1 error control. Moreover, since controlling type-1 error may also impact type-2 error (i.e., decrease power), regulators stress that one should examine the trade-off between the two types of error and carefully choose type-1 error controlling methodology. The multiplicity problem is further exacerbated by the inherent dependencies among various endpoints. While these dependencies can be qualitatively characterized in the sense that outcomes associated with the endpoints exhibit similar tendencies, albeit, with different magnitudes, there are situations where they can be quantitatively assessed from sample correlations among the examined variables.

Several statistical methodologies have been put forward to deal with the need to control type-1 error, with the aim of ultimately identifying at least one endpoint, and preferably both, for which the new treatment is better than the control. Among them, the most commonly used are the Bonferroni method for global testing and its step-down extension, Holm's (1979) method, for multiple testing. Because these methods utilize the Bonferroni inequality that relies only on the marginal p-values, they are dependency-free, and hence can be quite conservative when the p-values or the corresponding test statistics are highly dependent. Šidák (1967) and Simes (1986) have introduced improvements of the Bonferroni method for global testing. They control the type-1 error rate under independence and under a type of positive dependency that arises in some practical applications (Hochberg and Rom (1995), Samuel-Cahn (1996), Sarkar and Chang (1997), and Sarkar (1998)). Šidák's (1967) global test has been used by Holland and Copenhaver (1987) to develop a step-down method, whereas Simes (1986) has been used by Hochberg (1988) to develop a step-up multiple testing method and by Hommel (1988) to develop a closed testing

CONTACT Dror Rom d.rom@prosoftclinical.com Logcal Data Analytics, 996 Old Eagle School Rd, Suite 1106 Wayne, PA, 19087

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

method based on the 'Closure Principle' of Marcus et al. (1976). Gou et al. (2014) proposed a class of hybrid Hochberg-Hommel procedures which tend to be more powerful than either the Hochberg or Hommel procedure.

Šidák's (1967) and Simes (1986) improved versions of the Bonferroni global test and their multiple testing extensions only qualitatively capture the underlying positive dependency, as they are still based on marginal p-values while continuing to maintain the type-1 error rate control even under such positive dependency. Unfortunately, they can be quite conservative, and hence can lose power, when such dependency is moderately high. Moreover, they can fail to control the type-1 error under negative dependency. While these two tests are widely used, theoretical results regarding the validity of their application have only been done in the case of normal statistics with certain correlation structure [(Hochberg and Rom (1995), and Samuel-Cahn (1996)], or t-statistics with same denominator representing an estimate of the common population standard deviations [Sarkar and Chang (1997), and Sarkar (1998)]. These assumptions do not hold in the two-endpoint problem addressed here because the endpoints almost always have different population variabilities.

Under normal distributional settings, which are most commonly used for global testing in practical applications and where the dependency among test statistics is parametrically represented through correlation coefficients, it is possible to capture the dependency quantitatively, and hence more fully than the Šidák's (1967) and Simes (1986) tests, while improving the Bonferroni method. However, this idea of improving the Bonferroni method has so far been limited to the case where the population correlations are assumed known (see, e.g., Xie (2012) and the references therein). Of course, one can consider replacing the known correlations in these methods with their suitable estimates to make them fully data-adaptive, but there is no theoretical justification that these would ultimately control the type I error rate. With correlations being rarely known in practice, tightening the Bonferroni type-1 error rate control through explicit use of sample correlations and providing a theoretical justification of such control would be an important objective.

In this paper, we consider achieving the above-mentioned objective by considering the two-mean testing problem under a normal distributional setting with unknown population correlation and variances. Our goal is to test the two hypotheses, with the aim of rejecting at least one, and preferably both. This testing scenario commonly arises in pharmaceutical studies. We propose a new procedure in this setting that utilizes the Bonferroni test based on the usual (marginal) Student's-t test statistics but uses a data-adaptive critical value that explicitly incorporates the sample correlation coefficient. The confidence interval approach of Berger and Boos (1994) is employed to make use of the sample correlation. More specifically, we first theoretically prove that the type-1 error rate of the Bonferroni method based on Student's-t statistics (or their absolute values) with any fixed critical value is strictly decreasing in the unknown correlation coefficient (or its absolute value). These decreasing properties allow us to estimate the type-I error rates for both one- and two-sided testing problems, without relying on computations generally required in the application of the Berger and Boos (1994) approach. We simply are substituting the unknown correlation coefficient (or its absolute value) with its lower confidence limit, given a fixed confidence coefficient, into the error rate formulas. Bounding these estimated error rates from above by the nominal level α allows us to produce correlation-adaptive critical values that are smaller than the traditional Bonferroni critical values but still control the type-1 error rate. The fact that such adaptive Bonferroni methods can provide much tighter control of the type-1 error rate than their regular, non-adaptive versions over a wide range of choices for the confidence coefficient and level of significance is demonstrated numerically.

It is important to note that the Berger and Boos (1994) approach to estimating population correlation using its interval estimate in multiple testing scenarios was taken before in Tamhane et al. (2012). However, it was for a different problem, namely, the development of a two-stage group sequential design for testing primary and secondary endpoints controlling familywise error rate (FWER). Moreover, unlike here, they considered large-sample settings, which allowed them to assume the *t*-test statistics to be normally distributed and use large-sample confidence interval for the unknown correlation. Additionally, these authors only showed a directional relationship between the FWER and the correlation via numerical analysis as they were unable to show the relationship analytically.

The paper is organized as follows. Section 2 introduces our proposed 'correlation-adaptive Bonferroni' methodologies for both one- and two-sided testing problems. The process of computing the correlation-adaptive critical values in these methods is described in Section 3. In Section 4, we present these critical values for a wide range of sample sizes, before numerically showing in Section 4 how our methods compare with the corresponding traditional, non-adaptive Bonferroni methods in terms of type-1 error rate control and power. Concluding remarks are made in Section 5. These remarks include comments on (i) the novelty of theoretical results we obtain in this article towards application of the Berger and Boos (1994) approach, and (ii) possible extension of the proposed correlation-adaptive Bonferroni to its Holm-type stepdown analog for simultaneous testing. Detailed proofs of the technical results needed to develop our proposed method are provided in the Appendix 1.

2. Proposed methodologies

In our setting, a test treatment is compared to a control treatment on two outcome measures X_1 and X_2 that are jointly distributed as a bivariate normal with a covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},.$$

and with the following pair of means:

$$(X_1, X_2) = \begin{cases} (\mu_1^{(1)}, \mu_2^{(1)}) \textit{for test} \\ \left(\mu_1^{(2)}, \mu_2^{(2)}\right) \textit{for control} \end{cases}$$

Given n_1 pairs of observations $\left(X_{1j}^{(1)}, X_{2j}^{(1)}\right)$, $j = 1, \ldots, n_1$, for the test group, and n_2 pairs of observations $\left(X_{1j}^{(2)}, X_{2j}^{(2)}\right), j = 1, \ldots, n_2$, for the control group, our problem is to test the intersection H_0 of the following two one-sided null hypotheses:

$$H_0 = \left\{H_{01} : \mu_1^{(1)} \le \mu_1^{(2)}\right\} \cap \left\{H_{02} : \mu_2^{(1)} \le \mu_2^{(2)}\right\},$$

against the union H_1 of one-sided alternative hypotheses:

$$H_1 = \left\{H_{11} : \mu_1^{(1)} \mu_1^{(2)}\right\} \cup \left\{H_{12} : \mu_2^{(1)} \mu_2^{(2)}\right\},$$

or the intersection H_0 of the following two null hypotheses:

$$H_0 = \left\{H_{01} : \mu_1^{(1)} = \mu_1^{(2)}\right\} \cap \left\{H_{02} : \mu_2^{(1)} = \mu_2^{(2)}\right\},$$

against the union H_1 of two-sided alternative hypotheses:

$$H_1 = \left\{H_{11} : \mu_1^{(1)} \neq \mu_1^{(2)}\right\} \cup \left\{H_{12} : \mu_2^{(1)} \neq \mu_2^{(2)}\right\},$$

subject to a control of the type-1 error rate at α.

Note that for the one-sided testing problem, the least favorable configurartion, i.e., the point in the parameter space of H_0 for which type-1 error is maximized is $\left\{\mu_1^{(1)} = \mu_1^{(2)}\right\} \cap \left\{\mu_2^{(1)} = \mu_2^{(2)}\right\}$. Therefore in the one-sided testing problem, we can control type-1 error if we define and test the null hypotheses exactly as in the two-sided testing problem, i.e.,

$$H_0 = \left\{H_{01} : \mu_1^{(1)} = \mu_1^{(2)}\right\} \cap \left\{H_{02} : \mu_2^{(1)} = \mu_2^{(2)}\right\}$$

Let

$$T_1 = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\left(\bar{X}_1^{(1)} - \bar{X}_1^{(2)}\right)}{S_1} \text{and} T_2 = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\left(\bar{X}_2^{(1)} - \bar{X}_2^{(2)}\right)}{S_2},$$

where $\dot{X}_i^{(k)} = \frac{1}{n_k} \sum_{j-1}^{n_k} X_{ij}^{(k)}$ is the sample mean corresponding to $\mu_i^{(k)}$, for $i = 1,\ 2; k = 1,\ 2$, and

$$S_1^2 = \frac{1}{n-2}\sum_{k=1}^{2}\sum_{j=1}^{n_k}\left(X_{ij}^{(k)} - \dot{X}_i^{(k)}\right)^2,$$

with $n = n_1 + n_2$, is the pooled sample variance corresponding to X_i, for $i = 1,\ 2$. These are the standard Student's $t$statistics that are used to marginally test the corresponding null hypotheses and form the basic ingredients in the development of traditional intersection or global tests, like Bonferroni, Simes (1986), and others, that ignore an explicit use of the correlation between X_1 and X_2 or its estimate in their constructions.

We seek to improve the Bonferroni test by adapting it to the correlation between X_1 and X_2 through $r = S_{12}/S_1S_2$, with

$S_{12} = \frac{1}{n-2}\sum\limits_{k=1}^{2}\sum\limits_{j=1}^{n_k}\left(X_{1j}^{(k)} - \bar{X}_1^{(k)}\right)\left(X_{2j}^{(k)} - \bar{X}_2^{(k)}\right).$

the pooled sample correlation between X_1 and X_2. More specifically, we attempt to find a critical value $c_{1\alpha}(r)$, depending on r, such that

$$\Pr_{H_0}\{\max(T_1, T_2) \le c_{1\alpha}(r)\} \ge 1 - \alpha, \tag{2.1}$$

or $c_{2\alpha}(r)$ such that

$$\Pr_{H_0}\{\max(|T_1|, |T_2|) \le c_{2\alpha}(r)\} \ge 1 - \alpha, \tag{2.2}$$

depending on whether H_0 is tested against a one-sided alternative $H_1 : \left\{\mu_1^{(1)} > \mu_1^{(2)}\right\} \cup \left\{\mu_2^{(1)} > \mu_2^{(2)}\right\}$ or against a two-sided alternative

$H_1 : \left\{\mu_1^{(1)} \neq \mu_1^{(2)}\right\} \cup \left\{\mu_2^{(1)} \neq \mu_2^{(2)}\right\}.$

Towards finding $c_{1\alpha}(r)$ and $c_{2\alpha}(r)$, we first note the following distributional results:

$$\sqrt{\frac{n_1n_2}{n_1+n_2}}\begin{pmatrix} \bar{X}_1^{(1)} - \bar{X}_1^{(2)} \\ \bar{X}_2^{(1)} - \bar{X}_2^{(2)} \end{pmatrix} \text{and } (n-2)\begin{pmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{pmatrix}$$

independently distributed as $N_2(\mu, \Sigma)$ and $W_2(n-2, \Sigma)$, respectively, with

which equals $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ under H_0. From these results, we obtain the theorem below:

$\mu = \sqrt{\frac{n_1n_2}{n_1+n_2}}\begin{pmatrix} \mu_1^{(1)} - \mu_1^{(2)} \\ \mu_2^{(1)} - \mu_2^{(2)} \end{pmatrix},$

Theorem 1. *The following results hold:*

The probability $\Pr_{H_0}(\max(T_1, T_2) \le c)$ *depends on the nuisance parameters,* $\rho, \sigma_1,$ *and* σ_2 *only through* ρ *and is strictly increasing in* ρ*, for any fixed* $-\infty < c < \infty$.,

The probability $\Pr_{H_0}(\max(|T_1|, |T_2| \le c)$ *depends on the nuisance parameters,* $\rho, \sigma_1,$ *and* σ_2*only through* $|\rho|$ *and is strictly increasing in* $|\rho|$*, for any fixed* $0 < c_1, c_2 < \infty$.

This theorem, a proof of which is presented in Appendix 1, facilitates the calculation of $c_{1\alpha}(r)$ and $c_{2\alpha}(r)$ using a slight modification of the confidence interval approach of Berger and Boos (1994). Specifically, let $\Delta_1(c, \rho) = \Pr_{H_0}(\max(T_1, T_2) \le c)$, and $\hat{\rho}_{1,\beta}(r)$ be a lower confidence limit for ρ based on r with confidence coefficient $1 - \beta$. Then, since

$$\Delta_1(c, \rho) \ge \Delta_1(c, -1) = 2\Pr_{H_0}(T_1 \le c) - 1,$$

and $\Delta_1(c, \rho)$ is strictly increasing in $\rho \in (-1, 1)$, we have

$\Delta_1(c, \rho) = E\left\{\Delta_1(c, \rho)I\left(\rho \ge \hat{\rho}_{1,\beta}(r)\right)\right\} + E\left\{\Delta_1(c, \rho)I\left(\rho < \hat{\rho}_{1,\beta}(r)\right)\right\}$

$$\underset{=}{\geq}\; E\left\{\Delta_1\left(c, \hat{\rho}_{1,\beta}(r)\right) I\left(\rho \geq \hat{\rho}_{1,\beta}(r)\right)\right\} + [2\mathrm{Pr}_{H_0}(T_1 \leq c) - 1]\mathrm{Pr}_{H_0}\left(\rho < \hat{\rho}_{1,\beta}(r)\right)$$

$$E\left\{\Delta_1\left(c, \hat{\rho}_{1,\beta}(r)\right) I\left(\rho \geq \hat{\rho}_{1,\beta}(r)\right)\right\} + \beta[2\mathrm{Pr}_{H_0}(T_1 \leq c) - 1].$$

The desired $c \equiv c_{1\alpha}(r)$ guaranteeing (2.1) then can be obtained by equating $\Delta_1\left(c, \hat{\rho}_{1,\beta}(r)\right)$ to $\{1 - \alpha - \beta[2\mathrm{Pr}_{H_0}(T_1 \leq c) - 1]\}/(1 - \beta)$, that is, by solving the equation below for c, for any fixed (α, β, r):

$$G_{1,\beta}(c, r) = (1 - \beta)\Delta_1\left(c, \hat{\rho}_{1,\beta}(r)\right) + \beta[2\mathrm{Pr}_{H_0}(T_1 \leq c) - 1] = 1 - \alpha. \tag{2.3}$$

It is worth noting that $G_{1,\beta}(c, r) \geq 2\mathrm{Pr}_{H_0}(T_1 \leq c) - 1$, and so $c_{1\alpha}(r)$ is less than or equal to the Bonferroni critical value c satisfying $2\mathrm{Pr}_{H_0}(T_1 \leq c) = 2 - \alpha$. In other words, the resulting modification of the Bonferroni test for testing H_0 against $H_1 : \left\{\mu_1^{(1)} > \mu_1^{(2)}\right\} \cup \left\{\mu_2^{(1)} > \mu_2^{(2)}\right\}$ will have a larger rejection region.

The $c_{2\alpha}$ satisfying (2.2) can be obtained in the same manner by using the fact that $\Delta_2(c, |\rho|) = \mathrm{Pr}(\max(|T_1|, |T_2|) \leq c) \geq \Delta_2(c, |\rho| = 0) = \mathrm{Pr}^2(|T_1| < c), \mathrm{and} \Delta_2(c, |\rho|)$ is strictly increasing in $|\rho|$, and utilizing a lower confidence limit $|\hat{\rho}|_{2,\beta}(|r|)$ of $|\rho|$ based on $|r|$ with confidence coefficient $1 - \beta$. More specifically, $c \equiv c_{2\alpha}$ can be obtained by solving the equation below for c, for any fixed (α, β, r):

$$G_{2,\beta}(c, r) = (1 - \beta)\Delta_2\left(c, |\hat{\rho}|_{2,\beta}(|r|)\right) + \beta\mathrm{Pr}^2_{H_0}(|T_1| < c) = 1 - \alpha. \tag{2.4}$$

Since $c_{2\alpha}$ is smaller than the Bonferonni critical value $c$satisfying $\mathrm{Pr}^2_{H_0}(|T_1| < c) = 1 - \alpha$ for testing H_0 against $H_1 : \left\{\mu_1^{(1)} \neq \mu_1^{(2)}\right\} \cup \left\{\mu_2^{(1)} \neq \mu_2^{(2)}\right\}$, our modification of the Bonferroni test will have a larger rejection region, and hence more power, than the usual Bonferroni test.

3. Data-adaptive critical values

This section describes the process of calculating $G_{1,\beta}(c, r)$ and $G_{2,\beta}(c, r)$, given β, from the pooled sample covariance matrix with $n - 2$ degrees of freedom (d.f.). A pseudocode of these calculations appears in Appendix 2. Subsequently, we derive the critical values $c_{1\alpha}(r)$ and $c_{2\alpha}(r)$ by solving the corresponding equations (2.3) and (2.4) for c. The calculation of $G_{1,\beta}(c, r)$ and $G_{2,\beta}(c, r)$ involves expressing the probabilities $\Delta_1(c, \rho)$ and $\Delta_2(c, |\rho|)$ and estimating them by substituting $\mathrm{\rho}$ and $|\mathrm{\rho}|$ with their respective $1 - \beta$ lower confidence limit $\hat{\rho}_{1,\beta}(r)$ and $\widehat{|\rho|}_{2,\beta}(r)$.

3.1. *Expressions of $\Delta_1(c, \rho)$ and $\Delta_2(c, |\rho|)$*

Let Φ_ρ be the cumulative distribution function of (Z_1, Z_2) having standard bivariate normal distribution with correlation ρ. Then, from the above-mentioned joint distribution of $\left(\bar{X}_1^{(1)} - \bar{X}_1^{(2)}, \bar{X}_2^{(1)} - \bar{X}_2^{(2)}\right)$ and (S_1^2, S_2^2, S_{12}) under H_0, we see that

$$\Delta_1(c, \rho) \;= \mathrm{Pr}_{H_0}\left(Z_1 \leq c\frac{S_1}{\sigma_1}, Z_2 \leq c\frac{s_2}{\sigma_2}\right)$$

$$= \mathop{}_{0\;\;0}^{\infty\;\infty} \Phi_\rho\left(c\sqrt{w_1/(n-2)}, c\sqrt{w_2/(n-2)}\right) g(w_1, w_2)dw_1dw_2, \tag{3.1}$$

and

$$\Delta_2(c,|\rho|) = {}_{0\ \ 0}^{\infty\ \infty} g(w_1,w_2)\ [\Phi_{|\rho|}\left(c\sqrt{\frac{w_1}{n-2}}, c\sqrt{w_2(n-2)}\right) -$$

$$2\ \Phi_{|\rho|}\left(-c\sqrt{w_1/(n-2)}, c\sqrt{w_2/(n-2)}\right) +$$

$$\Phi_{|\rho|}\left(-c\sqrt{w_1/(n-2)}, -c\sqrt{w_2/(n-2)}\right)]dw_1dw_2$$

,
(3.2)

where $g(w_1,w_2)$ is the density of (W_1,W_2), the diagonal elements of a 2×2 Wishart matrix with $n-2$ d.f. and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Since

$$(W_1,W_2) \overset{d}{=} \left(W_1, \left((1-\rho^2)W_3 + \left(\sqrt{(1-\rho^2)}Z + \rho\sqrt{W_1}\right)^2\right)\right),$$

with W_1, W_3, and Z being distributed independently as χ^2_{n-2}, χ^2_{n-3} and $N(0,1)$, respectively (see, e.g., Odell and Feiveson (1966)), we see that $g(w_1,w_2)$ can be expressed as follows:

$$g(w_1,w_2) = g_{W_1}(w_1)\iint_{A(W_1,W_2)} g_{W_3}(w_3)\varphi(z)dw_3dz,$$

where g_{W_1}, g_{W_3} and $\varphi(z)$ are the densities of χ^2_{n-2}, χ^2_{n-3} and $N(0,1)$, respectively, and

$$A(w_1,w_2) = \left\{(w_3,z) : (1-\rho^2)w_3 + \left(\sqrt{(1-\rho^2)}z + \rho\sqrt{w_1}\right)^2 = w_2\right\}$$

.

3.2. Lower confidence limits $\widehat{\rho}_{1,\beta}(r)$ and$\widehat{|\rho|}_{2,\beta}(r)$.

Although these confidence limits can be approximated by using Fisher's transformation of r, we consider calculating them exactly using the following distribution of r (from sample covariance matrix with $n-2$ d.f.), obtained from Hotelling (1953):

$$f_\rho(r) = \frac{(n-3)\Gamma(n-2)(1-\rho^2)^{\frac{n-2}{2}}(1-r^2)^{\frac{n-5}{2}}}{\sqrt{2\pi}\Gamma\left(n-\frac{3}{2}\right)(1-\rho r)^{n-\frac{5}{2}}}\,{}_2\mathrm{F}_1\left(\tfrac{1}{2},\tfrac{1}{2};\tfrac{2n-3}{2};\tfrac{\rho r+1}{2}\right),$$

where Γ is the gamma function and ${}_2F_1$is the Gaussian hypergeometric function:

$${}_2\mathrm{F}_1(a,b;c;z) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!}, \text{ with } (q)_n = \begin{cases} 1, n=0 \\ q(q+1)\cdots(q+n-1), n>0. \end{cases}$$

A $1-\beta$ lower confidence limit $\hat{\rho}_{1,\beta}(r)$ for ρ is calculated by solving the following equation for $\hat{\rho}$:
$F_{\hat{\rho}}(r) = \quad f_{\hat{\rho}}(x)dx = 1-\beta$. (3.3)
Similarly, -1 a $1-\beta$ lower confidence limit $|\hat{\rho}|_{2,\beta}(|r|)$ for $|\rho|$ can be calculated by solving the following equation for $\widehat{|\rho|}$:
$F_{\hat{\rho}}(|r|) = \quad f_{|\hat{\rho}|}(x)dx =1-\beta$, (3.4)
where$f_{|\rho|}(x) = f_\rho(x) + f_\rho(-x); 0 \le x \le 1$ 0

3.3. Calculation of $c_{1\alpha}$ and $c_{2\alpha}$

We estimate $\Delta_1(c,\rho)$, given (c,β) by replacing ρ with its lower confidence limit $\hat{\rho}_{1,\beta}(r)$to obtain

Table 1. One-sided critical values (α= 0.025; β = 0.05 for n < 1,000, β = 0.01 for n ≥ 1,000).

	Sample Size, n (n_1:n_2)							
	10	20	30	50	80	150	500	2,000
r	(5:5)	(10:10)	(15:15)	(25:25)	(40:40)	(75:75)	(250:250)	(1,000:1,000)
−1	0.01250	0.01250	0.01250	0.01250	0.01250	0.01250	0.01250	0.01250
−0.3	0.01250	0.01250	0.01250	0.01250	0.01250	0.0125	0.01250	0.01250
−0.25	0.01250	0.01250	0.01250	0.01250	0.01250	0.0125	0.01250	0.01250
−0.2	0.01250	0.01250	0.01250	0.01250	0.01250	0.0125	0.01251	0.01251
−0.15	0.01250	0.01250	0.01250	0.01250	0.01250	0.01251	0.01251	0.01251
−0.10	0.01250	0.01250	0.01250	0.01250	0.01251	0.01251	0.01252	0.01252
−0.05	0.01250	0.01250	0.01250	0.01251	0.01251	0.01252	0.01253	0.01253
0.0	0.01250	0.01250	0.01251	0.01251	0.01252	0.01253	0.01254	0.01254
0.05	0.01250	0.01251	0.01251	0.01252	0.01253	0.01254	0.01256	0.01256
0.10	0.01250	0.01251	0.01252	0.01253	0.01254	0.01256	0.01259	0.01259
0.15	0.01250	0.01251	0.01252	0.01254	0.01256	0.01258	0.01262	0.01262
0.20	0.01250	0.01252	0.01253	0.01256	0.01258	0.01261	0.01266	0.01267
0.25	0.01251	0.01253	0.01255	0.01258	0.01261	0.01266	0.01272	0.01272
0.30	0.01251	0.01254	0.01257	0.01261	0.01265	0.01271	0.01278	0.01278
0.35	0.01251	0.01256	0.01260	0.01265	0.01271	0.01277	0.01286	0.01286
0.40	0.01252	0.01258	0.01263	0.01271	0.01277	0.01285	0.01296	0.01296
0.45	0.01253	0.01261	0.01268	0.01278	0.01286	0.01295	0.01308	0.01308
0.50	0.01254	0.01266	0.01275	0.01287	0.01297	0.01308	0.01323	0.01323
0.55	0.01256	0.01272	0.01284	0.01298	0.01310	0.01323	0.01341	0.01341
0.60	0.01259	0.01280	0.01295	0.01313	0.01327	0.01343	0.01362	0.01362
0.65	0.01263	0.01292	0.01311	0.01332	0.01349	0.01367	0.01388	0.01389
0.70	0.01269	0.01308	0.01332	0.01358	0.01377	0.01397	0.01421	0.01421
0.75	0.01278	0.01331	0.01360	0.01391	0.01413	0.01435	0.01462	0.01462
0.80	0.01295	0.01365	0.01400	0.01436	0.01461	0.01485	0.01514	0.01514
0.85	0.01323	0.01416	0.01458	0.01499	0.01526	0.01553	0.01583	0.01583
0.90	0.01378	0.01500	0.01549	0.01593	0.01622	0.0165	0.01681	0.01681
0.95	0.01507	0.01657	0.01709	0.01753	0.01781	0.01807	0.01836	0.01836

Table 2. Two-sided critical values (α= 0.05; β = 0.05 for n < 1,000, β = 0.01 for n ≥ 1,000) .

	Sample Size, n (n_1:n_2)							
	10	20	30	50	80	150	500	2,000
|r|	(5:5)	(10:10)	(15:15)	(25:25)	(40:40)	(75:75)	(250:250)	(1,000:1,000)
0.0	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530
0.05	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530	0.02531
0.10	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530	0.02531	0.02533
0.15	0.02530	0.02530	0.02530	0.02530	0.02530	0.02530	0.02533	0.02537
0.20	0.02530	0.02530	0.02530	0.02530	0.02530	0.02533	0.02539	0.02545
0.25	0.02530	0.02530	0.02530	0.02530	0.02532	0.02538	0.02547	0.02555
0.30	0.02530	0.02530	0.02530	0.02532	0.02538	0.02546	0.02559	0.02569
0.35	0.02530	0.02530	0.02530	0.02537	0.02546	0.02557	0.02574	0.02586
0.40	0.02530	0.02530	0.02534	0.02546	0.02557	0.02572	0.02593	0.02608
0.45	0.02530	0.02530	0.02542	0.02558	0.02573	0.02592	0.02617	0.02634
0.50	0.02530	0.02537	0.02553	0.02575	0.02594	0.02616	0.02646	0.02665
0.55	0.02530	0.02548	0.02569	0.02597	0.02621	0.02647	0.02681	0.02703
0.60	0.02530	0.02563	0.02592	0.02627	0.02655	0.02685	0.02724	0.02749
0.65	0.02530	0.02585	0.02622	0.02665	0.02698	0.02733	0.02777	0.02804
0.70	0.02539	0.02617	0.02664	0.02715	0.02754	0.02794	0.02842	0.02872
0.75	0.02559	0.02663	0.02721	0.02782	0.02826	0.02870	0.02924	0.02956
0.80	0.02590	0.02730	0.02801	0.02872	0.02921	0.02971	0.03029	0.03063
0.85	0.02647	0.02833	0.02917	0.02998	0.03053	0.03106	0.03167	0.03202
0.90	0.02756	0.03000	0.03097	0.03186	0.03244	0.03299	0.03362	0.03397
0.95	0.03013	0.03314	0.03418	0.03507	0.03563	0.03615	0.03673	0.03704

Table 3. Power (%) comparison. One-sided α= 0.025.

Sample Size, n (n_1:n_2)	Effect sizes	ρ	Adaptive. Bonferroni	Bonferroni	Simes	Šidák
30 (15:15)	1.3,0	0	87.8	87.8	87.8	87.8
		0.5	87.9	87.6	87.6	87.7
		0.9	89.7	87.6	87.6	87.7
	1.2,0.6	0	86.2	86.2	86.9	86.2
		0.5	83.0	82.7	83.2	82.8
		0.9	84.1	81.4	81.5	81.5
	1.1,1.1	0	93.2	93.1	94.0	93.2
		0.5	87.7	87.4	88.6	87.5
		0.9	82.6	79.8	82.2	79.9
500 (250:250)	0.3,0	0	86.7	86.7	86.7	86.7
		0.5	87.1	86.6	86.8	86.6
		0.9	88.9	86.5	86.5	86.6
	0.28,0.14	0	85.8	85.7	86.4	85.8
		0.5	82.8	82.2	82.7	82.3
		0.9	84.1	81.1	81.2	81.2
	0.25,0.25	0	91.5	91.4	92.3	91.5
		0.5	85.5	84.9	86.0	85.0
		0.9	80.5	77.0	79.2	77.1

$$G_{1,\beta}(c,r) = (1-\beta)\Delta_1\left(c,\hat{\rho}_{1,\beta}(r)\right) + \beta[2\Pr_{H_0}(T_1 \le c) - 1],$$

where $\Pr_{H_0}(T_1 \le c)$ is calculated using the cumulative distribution function of central Student's t with $n-2$ d.f. The $c_{1\alpha}$is then obtained by solving the equation $G_{1,\beta}(c,r) = 1-\alpha$for c.

Similarly, $c_{2\alpha}$ is calculated by estimating $\Delta_2(c,|\rho|)$, given (c,β) by replacing $|\rho|$ with its lower confidence limit $\widehat{|\rho|}_{2,\beta}(r)$ to obtain

$$G_{2,\beta}(c,r) = (1-\beta)\Delta_2\left(c,\widehat{|\rho|}_{2,\beta}(r)\right) + \beta\Pr^2_{H_0}(|T_1| \le c),$$

where $\Pr_{H_0}(|T_1| \le c)$ is calculated using the cumulative distribution function of central Student's t with $n-2$ d.f. and solving the equation $G_{2,\beta}(c,r) = 1-\alpha$for c.

4. Critical values

Tables 1 and 2 present the critical values of our proposed correlation-adaptive Bonferroni procedures, respectively, for one- and two-sided testing problems. For each configuration of sample size and observed sample correlation coefficient r, the table entries are the solutions of the process described in Section 3 for $G_{1,\beta}(c,r) = 1-\alpha$ (one-sided tests) and $G_{2,\beta}(c,r) = 1-\alpha$(two-sided tests). These solutions were obtained by iteratively changing the critical values and numerically integrating the left-hand side of each equation until a solution was found so that the right-hand side of each equation was within 0.000001 of $1-\alpha$.

We are providing values for a wide range of observed sample correlation coefficient r or its absolute value $|r|$, depending on whether the testing problem is one- or two-sided, and for some choices of total sample size n. For sample sizes below 1000, we have used $\beta = 0.05$, while for sample sizes of 1000 and above, $\beta = 0.01$ was used. We elaborate on these choices in Section 5.

Note that for the two endpoints problem, with a two-sided $\alpha = 0.05$. or a one-sided $\alpha = 0.025$, the Bonferroni critical values are simply half of their respective α, namely 0.025 (= 0.05/2) and 0.0125 (= 0.025/2) for the two- or one-sided testing problems, no matter what the sample correlation coefficient is. As expected, the newly derived critical values increase as the sample size or the sample correlation coefficient increases. This is due to the tighter range of the confidence interval with increased sample size, and the decreasing property of the type-1 error rate with increasing population correlation (approximately equaling the sample correlation for large sample size). Of note is that the new critical values remain close to the corresponding usual Bonferroni critical values when the sample correlation is in the range of −1 to 0.

Table 3 displays power comparisons between the correlation adaptive Bonferroni, the standard, non-adaptive Bonferroni, Simes, and Šidák procedures. Estimated power calculation was done via 1,000,000 random samples. Comparisons were made for a few configurations of effect sizes for the two endpoints, ranging from equal to substantially different. For meaningful comparisons, configurations of effect sizes were designed to facilitate power comparisons in the range of 80-90%. As expected, the Adaptive Bonferroni has a power advantage over the non-adaptive Bonferroni test that increases with sample size and population correlation. The differences are noticeable, being in the range from 1% to 4%. Šidák's test gives only a minor improvement over Bonferroni. Simes's test has its best advantage for effect sizes that are equal. In those cases, its advantage over the adaptive Bonferroni can be in the range of 0.5–1%. On the other hand, when the effect sizes are different, the adaptive Bonferroni has an advantage that can be in the range of 2–2.5%. As we stated before, and elaborated in the next section, the Simes' test has not been shown to control type-1 error for the testing problem addressed here, namely when the t-statistics are constructed with separate estimates of the population standard deviations, and therefore its validity for this problem is not known.

5. Discussion and concluding remarks

The multiplicity problem addressed in this paper is quite common in clinical trial settings where two treatments are compared on two primary endpoints and evidence of superiority on one of these endpoints is sufficient to obtain regulatory marketing approval. Current solutions to this problem in terms of controlling the type-1 error rate are typically based on dependency-free methodologies (such as Bonferroni test and its various extensions) or on those that only qualitatively utilize positive dependencies (such as Šidák's (1967) and Simes (1986) tests and their extensions). However, it is generally understood that test procedures that utilize more data-embedded information, such as dependencies among variables, tend to be more powerful. Our proposed data-adaptive version of the Bonferroni method utilizing information through the sample correlation is such a procedure. It is indeed more powerful than its non-adaptive counterpart, as numerically verified.

It is important to note that Simes' and Šidák's inequalities were not proven to hold in the testing problem described here and therefore the validity of multiple testing procedures based on these two tests is questionable. Hochberg and Rom (1995) and Samuel-Cahn (1996) have shown that Simes' test controls type-1 error when the test statistics are jointly bivariate normal for two-sided testing, and with non-negative correlation for one-sided testing. Sarkar and Chang (1997), and Sarkar (1998) have obtained similar results when the test statistics are jointly bivariate t whose marginal t-statistics have been constructed with the same estimate of the standard deviation (sometimes referred to as 'the standard bivariate t of the Dunnett type'). For the problem at hand, the marginal t-statistics do not share the same estimate of the standard deviation, and therefore, the resulting bivariate t-distribution is not of the Dunnett type. It is unknown whether the results proven in Sarkar and Chang (1997), and Sarkar (1998) hold for this problem. Moreover, it has been shown in Hochberg and Rom (1995), and Samuel-Cahn (1996) that Simes' test has an inflated type-1 error for negatively correlated normal statistics with one-sided testing; and since the value of the population correlation is rarely known and can be negative, the validity of the Sime's test in the testing problem described here is questionable.

The arguments above regarding one-sided testing also apply to Šidák's inequality. Nevertheless, the results obtained in this paper allow us to state the following: 1. The Adaptive Bonferroni method is never less powerful than Šidák's method for two-sided testing since our method allows us to replace the unknown correlation with a less conservative correlation resulting from the use of the confidence interval for the unknown population correlation. If the confidence interval does not cover zero, then our critical values will be less conservative than Šidák's critical values, otherwise, they will be the same. By implication, we have proven that 1: Šidák's inequality holds for the absolute values of two t statistics whose joint distribution is of the form described here (the standard deviations have separate estimates), an important result on its own; and 2: For the one-sided testing problem, it is generally (but not always) true that for positively correlated statistics, our method will result in less conservative critical values than those obtained by assuming that the correlation is zero (independence in the normal case) as is done by Šidák. However, Šidák's method can inflate type-1 error if the population correlation is negative, while our method is valid for that case.

One might consider using Hotelling's T^2 to test the global null hypothesis for our setting. However, the resulting test does not possess the "Consonance" property of Gabriel (1969); that is, following the rejection of the global null hypothesis, the rejection of any of the individual hypotheses is not guaranteed, and they must each be tested and rejected by their own a α-level test. This may lead to loss of power for the rejection of any of the individual null hypotheses. On the other hand, the Bonferroni as well as our adaptive version of it, being in the class of Union Intersection (UI) tests, are consonant, and therefore do allow for the rejection of at least one individual null hypothesis whenever the global null hypothesis is rejected. A UI test allows for the allocation of different portions of type-1 the error to the marginal Student's t-test statistics, thereby adapting the test to the possible difference in effect sizes between the two endpoints. Also, it is amenable to its applications as a stepwise procedure, starting with the global test and, depending on the rejection of the global null hypothesis (and so at least one individual hypothesis), allocating the full nominal type-1 error to the other hypotheses, thereby increasing the power to reject the second hypothesis.

The monotonicity of the type I error rate for Bonferroni global testing involving one-sided (or two-sided) tests with respect to the population correlation (or the absolute value of the population correlation) is an important theoretical result in the process of carrying out the main maximization step in the Berger and Boos (1994) approach without computations. While this property is known in the literature for multivariate (or absolute-valued multivariate) normal random variables, they are not available for the joint distribution of the marginal$t$'s (or absolute-valued marginal t's) in Hotelling's T^2, and so these results proven in the bivariate case in this paper are important in their own right. Tamhane et al. (2012) have made use of a similar monotonicity property for normally distributed test statistics, although for a different problem, in the aforementioned step of the Berger and Boos (1994) approach without computations. However, they verified this property numerically.

The proposed correlation-adaptive Bonferroni method for global testing can be used to develop a Holm-type stepdown method for simultaneous testing of the individual null hypotheses in the present context. For instance, let us consider the one-sided testing problem. With $H_{(01)}$ and $H_{(02)}$ denoting the null hypotheses corresponding to $\min(T_1, T_2)$ and $\max(T_1, T_2)$, respectively, we can describe this so-called correlation-adaptive Holm method controlling the (familywise) type-1 error rate at α as follows:

Do not reject $H_{(01)}$or $H_{(02)}$ if $\max(T_1, T_2) \leq c_{1\alpha}(r)$

Do not reject $H_{(01)}$ but reject $H_{(02)}$ if $\min(T_1, T_2) \leq t_{\alpha,n-2}, \max(T_1, T_2) > c_{1\alpha}(r)$

Reject both $H_{(01)}$and $H_{(02)}$ if $\min(T_1, T_2) > t_{\alpha,n-2}, \max(T_1, T_2) > c_{1\alpha}(r)$

A correlation-adaptive Holm method for the two-sided testing problem can be similarly proposed in terms of $\min(|T_1|, |T_2|)$, $\max(|T_1|, |T_2|)$ and $c_{2\alpha}(r)$.

The correlation-adaptive Bonferroni methodology can be further extended to more than two endpoints, although a difficulty arises due to the increased dimensionality. One may need to resort to some efficient Monte-Carlo numerical integration methods to address the testing of more than two endpoints. This extension will also require some additional theoretical results. A more pragmatic approach to reduce the dimensionality problem is to use the bivariate results obtained here and to devise an upper bound for the case of more than two endpoints. The first method can readily be described for the case of three endpoints as follows (one-sided bounds are described here with obvious changes to two-sided testing):

$$Pr\left(\bigcup_{i=1}^{3}\{T_i \geq c\}\right) = \sum_{i=1}^{3} Pr(T_i \geq c) - \sum_{i,j(ij)=1}^{3} Pr\left(\{T_i \geq c\}\cap^{\{T_j \geq c\}}\right) + Pr\left(\bigcap_{i=1}^{3}(T_j \geq c)\right)$$

,

and since

$$Pr\left(\cap_{i=1}^{3}\left(T_{\mathrm{j}} \geq c\right)\right) \leq \min_{i \neq j} Pr\left(\left\{T_{\mathrm{i}} \geq c\right\} \cap \left\{T_{\mathrm{j}} \geq c\right\}\right).$$

$$\rightarrow Pr\left(\cup_{i=1}^{3}\left\{T_{\mathrm{i}} \geq \mathrm{c}\right\}\right) \leq \sum\nolimits_{i=1}^{3} Pr\left(T_{\mathrm{i}} \geq c\right) - \max_{i \in \{1,2,3\}} \sum\nolimits_{j \neq i} Pr\left(\left\{T_{\mathrm{i}} \geq c\right\} \cap \left\{T_{\mathrm{j}} \geq c\right\}\right) \qquad (2.5)$$

This bound relies on the univariate and bivariate probabilities only. We can then replace each of the bivariate probabilities on the righthand side of (2.5) using the lower confidence limit of the correlation between the respective statistics and apply the Berger and Boos (1994) method as was done for the two-endpoint problem. Two types of extensions of (2.5) can be made for more than three endpoints: The first is based on extending (2.5) to k endpoints using Kounias (1968) inequality:

$$Pr\left(\cup_{i=1}^{k}\left\{T_{\mathrm{i}} \geq \mathrm{c}\right\}\right) \leq \sum\nolimits_{i=1}^{k} Pr\left(T_{\mathrm{i}} \geq c\right) - \max_{i \in \{1,\ldots,k\}} \sum\nolimits_{j \neq i} Pr\left(\left\{T_{\mathrm{i}} \geq c\right\} \cap \left\{T_{\mathrm{j}} \geq c\right\}\right), \qquad (2.6)$$

and using a lower confidence bound for each (bivariate) correlation and the Berger and Boos (1994) method in (2.5).

A second approach is to utilize the closure principle of Marcus et al. (1976) to test all intersection hypotheses of cardinality $j(j \in \{1, 2, \ldots k\})$ at level $j\alpha/k$. In this approach, any intersection hypothesis H of cardinality i, can be rejected at level $i\alpha/k$ by testing and rejecting all intersection hypotheses of cardinality $j(>i)$ implying H at level$j\alpha/k$. Applying this idea recursively to testing k endpoints, the following procedure will control type-1 error rate:

Reject any hypothesis H_j $(j = 1, \ldots, k)$ corresponding to endpoint $j,$ provided all intersection hypotheses of cardinality 3 implying H_j have been tested and rejected at level $3\alpha/k$. We use (2.5) to test all hypotheses of cardinality 3.

The tightness (i.e., how far the above bounds are from the exact type-1 error) of the above approaches depends on the correlation matrix among the k endpoints which in turn determines whether higher dimensional probabilities are diminishingly small compared to the two-dimensional probabilities. Our preliminary evaluation suggests that for small to moderate correlations, the univariate and bivariate probabilities do provide a tight upper bound on type-1 error. Further work is currently undertaken to examine the above bounds. As an example of this point, consider a setting with three endpoints, and sample sizes of 500 in each of two groups, with all observed sample correlations being 0.5. With a one-sided type-1 error of 0.025, the Bonferroni test will use a critical value of 0.025/3 =0.0083 for testing each of the three hypotheses. Applying (2.5) with a lower $1-\beta$ (β=0.0001) confidence limit and the Berger and Boos (1994) method, we get a critical value of 0.00867 which is a slight improvement over the Bonferroni test. If we were to consider the asymptotic critical value ($n \rightarrow \infty$) using a three-dimensional normal with all correlations equal to 0.5 to approximate the joint distribution of the test statistics, we would use a critical value of 0.0095 (estimated using a Monte Carlo simulation) making our critical value 0.00867 slightly conservative. Note that the use of the asymptotic critical value may cause some type-1 error inflation due to the use of the normal distribution instead of the t-distribution, and the use of the observed correlations to replace the unknown correlations. Thus, the conservatism of our critical value is no more than the difference derived from the asymptotic distribution, and practically can be much lower.

A similar problem with observed sample correlations of 0.9 gives a critical value of 0.01205 from our method while the Bonferroni test remains unchanged with a critical value of 0.0083. Again, considering the asymptotic distribution as a three-dimensional normal with all correlations being 0.9, the critical value is 0.0145 (estimated from a Monte Carlo simulation), making our method with a critical value of 0.01205 slightly conservative but much less conservative than the Bonferroni test.

The method described in this paper can be extended more easily to situations where, following the rejection of either of the primary endpoints, it is desired to test secondary endpoints. The dependencies between the primary and secondary endpoints can then be readily incorporated using the methodology described in this article to devise an improved sequential testing.

Acknowledgments

We would like to thank Ajit Tamhane for his helpful suggestions on an earlier version of this manuscript. We also thank the anonymous referee and the editor for their insightful comments which helped strengthen this manuscript. We thank Michael Pol for meticulously programming all tables. An R code for generating critical values is available from the authors upon request

ORCID

Dror Rom http://orcid.org/0000-0003-0967-4258
Jaclyn McTague http://orcid.org/0000-0003-0017-103X

References

Berger, R. L., and D. D. Boos. 1994. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 89 (427):1012–1016.

Gabriel, K. R. 1969. Simultaneous test procedures–some theory of multiple comparisons. *Annals of Mathematical Statistics* 41 (1):224–250. doi:10.1214/aoms/1177697819.

Gou, J., A. C. Tamhane, D. Xi, and D. Rom. 2014. A class of improved hybrid hochberg-hommel type step-up multiple test procedures. *Biometrika* 101 (4):899–911. doi:10.1093/biomet/asu032.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75 (4):800–802. doi:10.1093/biomet/75.4.800.

Hochberg, Y., and D. M. Rom. 1995. Extensions of multiple testing procedures based on Simes' test. *Journal of Statistical Planning and Inference* 48 (2):141–152. doi:10.1016/0378-3758(95)00005-T.

Holland, B. S., and M. D. Copenhaver. 1987. An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43 (2):417–423. doi:10.2307/2531823.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6 (2):65–70.

Hommel, G. A. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75 (2):383–386. doi:10.1093/biomet/75.2.383.

Hotelling, H. 1953. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society* 15 (2):193–232.

Kounias, E. J. 1968. Bounds for the probability of a union, with applications. *The Annals of Mathematical Statistics* 39 (6):2154–2158. doi:10.1214/aoms/1177698049.

Marcus, R., E. Peritz, and K. R. Gabriel. 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63 (3):655–660. doi:10.1093/biomet/63.3.655.

Odell, P. L., and A. H. Feiveson. 1966. A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association* 61 (313):199–203. doi:10.1080/01621459.1966.10502018.

Samuel-Cahn, E. 1996. Is the Simes improved Bonferroni procedure conservative? *Biometrika* 83 (4):928–933. doi:10.1093/biomet/83.4.928.

Sarkar, S. K. 1998. Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Annals of Statistics* 26 (2):494–504. doi:10.1214/aos/1028144846.

Sarkar, S. K., and C.-K. Chang. 1997. The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 92 (440):1601–1608. doi:10.1080/01621459.1997.10473682.

Šidák, Z. K. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62 (318):626–633.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73 (3):751–754. doi:10.1093/biomet/73.3.751.

Tamhane, A. C., Y. Wu, and C. R. Mehta. 2012. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (II): sample size re-estimation. *Statistics in Medicine* 31 (19):2041–2054. doi:10.1002/sim.5377.

Xie, C. 2012. Weighted multiple testing correction for correlated tests. *Statistics in Medicine* 31 (4):341–352. doi:10.1002/sim.4434.