The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 3

A Dunnett-Type Procedure for Multiple Endpoints

Mario Hasler, *Christian-Albrechts-Universität zu Kiel*
Ludwig A. Hothorn, *Leibniz Universität Hannover*

Recommended Citation:
Hasler, Mario and Hothorn, Ludwig A. (2011) "A Dunnett-Type Procedure for Multiple Endpoints," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 3.
DOI: 10.2202/1557-4679.1258

A Dunnett-Type Procedure for Multiple Endpoints

Mario Hasler and Ludwig A. Hothorn

Abstract

This paper describes a method for comparisons of several treatments with a control, simultaneously for multiple endpoints. These endpoints are assumed to be normally distributed with different scales and variances. An approximate multivariate *t*-distribution is used to obtain quantiles for test decisions, multiplicity-adjusted *p*-values, and simultaneous confidence intervals. Simulation results show that this approach controls the family-wise error type I over both the comparisons and the endpoints in an admissible range. The approach will be applied to a randomized clinical trial comparing two new sets of extracorporeal circulations with a standard for three primary endpoints. A related R package is available.

KEYWORDS: many-to-one comparison, multiple endpoints, multivariate *t*-distribution, family-wise error

Author Notes: The authors thank Siegfried Kropf and Marilene S. Jepsen for a critical reading of the original version of the paper and for the original data used as an example, respectively. We would also like to thank the referees and the editor for their constructive comments and helpful suggestions.

1 Introduction

Randomized clinical trials and pre-clinical studies, designed for the comparison of several treatments with a placebo or a control group, often do not cover only one single endpoint (variable) but many correlated endpoints. The scale of these endpoints is often different. For example, in a randomized clinical trial (Kropf, Hommel, Schmidt, Brickwedel, and Jepsen, 2000), two new extracorporeal circulation sets were compared with a standard set for the three primary endpoints thrombocyte count, thrombocyte activity ADP and thrombocyte activity TRAP. Or in an immuno-toxicological inter-laboratory study (Schulte, Althoff, Ewe, and Richter-Reichhelm, 2002), immuno-toxicity endpoints were considered for a comparison of three dose levels with a zero dose control. The experimental goal is not only to clarify, which treatment groups differ, but also on which endpoints. Hence, it is not a priori clear, for which endpoints differences between the treatment groups can be expected. These endpoints must be detected by the analysis a posteriori, so that they must be evaluated commonly – not separately.

Multiplicity adjustment must take both the number of endpoints and the common treatment comparisons into account. That increases the conservatism of the elementary decisions additionally. Thus, the first strategy is to reduce the number of endpoints to the smallest possible number that is necessary, and that still provides the main information about the data (Neuhäuser, 2006). Second, it is useful to divide the endpoints into primary and secondary ones. In this spirit, the guideline on biostatistics of the ICH (ICH E9 Expert Working Group, 1999) recommends the selection of one primary endpoint. However, in some clinical trials, a claim on several primary endpoints is intended. A possible objection is that such a classification of endpoints according to their importance can be somewhat arbitrary. Like the first, this strategy also reduces the dimension of the problem, but the question, how to handle multiple primary endpoints in multi-armed trials, remains. The statistical analysis for these endpoints must control the family-wise error type I (FWE) over all of them. On the other hand, their correlations are important. First, the degree of conservatism of the elementary decisions is reduced by taking the correlations into account. Second, effects may be erroneously ignored or masked, when analyzing the endpoints separately. And third, the degree of correlation is essential. For example, highly correlated endpoints do not give the same amount of information about the data as uncorrelated ones.

According to the guideline of the ICH (ICH E9 Expert Working Group, 1999), *"Estimates of treatment effects should be accompanied by confidence intervals, whenever possible..."*. Stepwise procedures - like those of Imada and

Douke (2007) or Cohen, Sackrowitz, and Xu (2008) - generally have the drawback that no meaningful simultaneous confidence intervals (SCIs) are available. Imada and Douke (2007, 2008) and Cohen et al. (2008) additionally assume known covariances, which is hardly met in practice. Simpler procedures (Holm, 1979, Hochberg, 1988, Hommel, 1988) yield conservative and even biased test decisions, because the information about correlations of the endpoints is not exploited, or the correlations are assumed non-negative, respectively. Gatekeeping procedures (Bauer, 1991, Dmitrienko, Offen, and Westfall, 2003) suffer from similar drawbacks. The T^2 test of Hotelling (1951) takes correlations into account, but because of a square sum test statistic it is non-directional and hence not meaningful in many application areas. Furthermore, the test conclusions are merely global ones in the sense that they cannot be attributed to single endpoints. Stabilized alternatives to the T^2 test, using linear scores (Kropf, Hothorn, and Läuter, 1997), also suffer from that drawback. Seo and Nishiyama (2008) consider SCIs for multiple comparisons among mean vectors from the multivariate normal populations, but these intervals are also conservative.

The many-to-one comparison according to Dunnett (1955) and related SCIs provide test decisions, and parameter estimation, respectively, for the comparisons of treatments versus a control. The FWE is maintained with size α, and correlations between the comparisons are taken into account. However, this procedure is limited to comparisons of treatments on a single endpoint so far. A naive approach would obviously be to apply the Dunnett procedure for all the endpoints separately and to adjust for the multiple endpoints by methods of Holm (1979) or Hommel (1988). However, such an approach would also not exploit the correlations of the endpoints.

This article presents an extension of the Dunnett procedure for multiple endpoints, where the correlations of the endpoints are explicitly taken into account. In Section 2, the testing problem is formulated, an approximate distribution for the test statistics is derived, and SCIs are presented. Section 3 shows results of simulations concerning the FWE for several numbers and correlations of endpoints, respectively. Some further aspects of the new procedure are considered in Section 4. We give an example in Section 5 and end with our conclusions in Section 6.

2 Testing problem and test procedure

For $l = 0, \ldots, q$, $i = 1, \ldots, k$ and $j = 1, \ldots, n_l$, let X_{lij} denote the jth observation on the ith endpoint under the lth treatment, where $l = 0$ represents the control and $\sum_{l=0}^{q}(n_l - 1) \geq k$. The vectors $(X_{l1j}, \ldots, X_{lkj})'$ are mutually independent and follow k-variate normal distributions with mean vectors $\boldsymbol{\mu}_l = (\mu_{l1}, \ldots \mu_{lk})'$ and unknown covariance matrices $\boldsymbol{\Sigma}_l \in \mathbb{R}^{k \times k}$. We assume possibly different variances and covariances for the endpoints, but the same covariance matrices for all treatments, i.e., $\boldsymbol{\Sigma}_0 = \cdots = \boldsymbol{\Sigma}_q = \boldsymbol{\Sigma} = (\sigma_{ii'})_{i,i'}$. Thus,

$$\{X_{lij} : i = 1, \ldots, k\} \sim \perp \mathrm{N}_k(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}) \quad (l = 0, \ldots, q,\ j = 1, \ldots, n_l).$$

Let $\bar{\boldsymbol{X}}_l = (\bar{X}_{l1}, \ldots, \bar{X}_{lk})'$ and $\hat{\boldsymbol{\Sigma}}_l$ be the sample mean vector and the sample covariance matrix of the endpoints for a fixed treatment, respectively, with

$$\bar{X}_{li} = \frac{1}{n_l} \sum_{j=1}^{n_l} X_{lij} \quad (l = 0, \ldots, q).$$

The pooled sample covariance matrix $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ii'})_{i,i'}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{l=0}^{q}(n_l - 1)\hat{\boldsymbol{\Sigma}}_l}{\sum_{l=0}^{q}(n_l - 1)}$$

with the estimates $\hat{\sigma}_{ii'}$ $(1 \leq i, i' \leq k)$ for the covariances of the endpoints. The diagonal elements, required for the test procedure we will describe, are

$$\hat{\sigma}_{ii} = S_i^2 = \frac{(n_0 - 1)S_{0i}^2 + \cdots + (n_q - 1)S_{qi}^2}{n_0 + \cdots + n_q - q - 1} \quad (i = 1, \ldots, k)$$

with

$$S_{li}^2 = \frac{1}{n_l - 1} \sum_{j=1}^{n_l} (X_{lij} - \bar{X}_{li})^2 \quad (l = 0, \ldots, q,\ i = 1, \ldots, k).$$

From the pooled sample covariance matrix $\hat{\boldsymbol{\Sigma}}$, we then derive the estimate $\hat{\boldsymbol{R}}$ of the unknown common correlation matrix of the multiple endpoints $\boldsymbol{R} = (\rho_{ii'})_{i,i'}$.

The objective of this paper is the testing of the hypotheses

$$H_0^{(li)} : \eta_{li} \leq \delta_i \quad (l = 1, \ldots, q,\ i = 1, \ldots, k), \tag{1}$$

where $\eta_{li} = \mu_{li} - \mu_{0i}$ are the differences to control, and $\delta_i \in (-\infty, \infty)$ are endpoint-specific thresholds. In many applications, $\delta_i = 0$ for all i. The method described here is sufficiently general to allow for both comparison-specific and endpoint-specific thresholds $\delta_{li} \in (-\infty, \infty)$. If the test direction is to be reversed for some endpoints, the corresponding test statistics have to be multiplied by minus one. The testing problem (1) is a union-intersection-test because the overall null hypothesis of interest can be expressed as an intersection of the local null hypotheses, i.e.,

$$H_0 = \bigcap_{l=1}^{q} H_0^{(l)} = \bigcap_{l=1}^{q} \left\{ \bigcap_{i=1}^{k} H_0^{(li)} \right\}. \tag{2}$$

This means that the overall null hypothesis H_0 is rejected if and only if a local null hypothesis $H_0^{(li)}$ is rejected for at least one treatment on at least one endpoint.

The test of the hypotheses (1) will be based on the test statistics

$$T_{li} = \frac{\bar{X}_{li} - \bar{X}_{0i} - \delta_i}{S_i \sqrt{\frac{1}{n_l} + \frac{1}{n_0}}} \quad (l = 1, \ldots, q,\ i = 1, \ldots, k).$$

The vectors $\boldsymbol{T}_l = (T_{l1}, \ldots, T_{lk})'$, containing the test statistics for the lth comparison on all endpoints, can be reshaped to

$$\boldsymbol{T}_l = \left. \quad \frac{Y_{l1}}{\sqrt{U_1/\nu}}, \ldots, \frac{Y_{lk}}{\sqrt{U_k/\nu}} \right)' \quad (l = 1, \ldots, q),$$

where under $H_0^{(l)}$, the vector $(Y_{l1}, \ldots, Y_{lk})'$ follows a k-variate normal distribution with the correlation matrix $\boldsymbol{R}$. The $U_1, \ldots, U_k$ are dependent χ^2 variables with

$$\nu = \sum_{l=0}^{q} (n_l - 1)$$

degrees of freedom. Note that $U_1, \ldots, U_k$ are different random variables, but they are identically distributed. For that reason, the distribution of $\boldsymbol{T}_l$ under $H_0^{(l)}$ is not among the standard distributions discussed in textbooks. Even the definition of a multivariate t-distribution of an appropriate generality is an open problem. A generalized t-distribution has been derived analytically by Siddiqui (1967) only in the bivariate case, representing the situation of two endpoints. A multivariate extension would require the joint distribution of $U_1, \ldots, U_k$. An approximation to this distribution in the equicorrelated case is given by Kotz, Balakrishnan, and Johnson (2000); this approximation is exact in the bivariate case.

If assuming a known covariance matrix $\boldsymbol{\Sigma}$ for the data, $\boldsymbol{T}_l$ is a multivariate normal vector, but in the case of unknown (co-) variances, this assumption leads to supremely liberal test decisions. However, the distribution can be approximated by a k-variate t-distribution with ν degrees of freedom and the correlation matrix $\boldsymbol{R}$, i.e.,

$$\boldsymbol{T}_l \overset{appr.}{\sim} t_k(\nu, \boldsymbol{R}).$$

Consequently, under H_0, the vector of all test statistics,

$$\boldsymbol{T} = (\boldsymbol{T}'_1, \ldots, \boldsymbol{T}'_q)' = (T_{11}, \ldots, T_{li}, \ldots, T_{qk})',$$

follows approximately a qk-variate t-distribution with ν degrees of freedom and a correlation matrix, denoted by $\tilde{\boldsymbol{R}}$, i.e.,

$$\boldsymbol{T} \overset{appr.}{\sim} t_{qk}(\nu, \tilde{\boldsymbol{R}}). \tag{3}$$

The correlation matrix $\tilde{\boldsymbol{R}}$ is given by

$$\tilde{\boldsymbol{R}} = (\boldsymbol{R}_{ll'})_{l,l'} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \ldots & \boldsymbol{R}_{1q} \\ \boldsymbol{R}_{12} & \boldsymbol{R}_{22} & \ldots & \boldsymbol{R}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}_{1q} & \boldsymbol{R}_{2q} & \ldots & \boldsymbol{R}_{qq} \end{pmatrix}.$$

The submatrices $\boldsymbol{R}_{ll'} = (\rho_{ll',ii'})_{i,i'}$ describe the correlations between the lth and the l'th comparison for all endpoints. Their elements are

$$\rho_{ll',ii'} = \begin{cases} \rho_{ii'}, & l = l' \\ \rho_{ii'} \dfrac{1}{\sqrt{\left(\frac{n_0}{n_l}+1\right)\left(\frac{n_0}{n_{l'}}+1\right)}}, & l \neq l' \end{cases} \quad (l = 1, \ldots, q;\ i, i' = 1, \ldots, k). \tag{4}$$

For $i = i'$, we recover the correlations of the Dunnett procedure (Dunnett, 1955). Hence, the conventional case of a single endpoint ($k = 1$) is a special case of the method described in the present paper. Furthermore, focusing on one fixed comparison ($l = l'$), the structure of the correlation matrix simplifies according to $\rho_{ll',ii'} = \rho_{ii'}$ and $\boldsymbol{R}_{ll} = \boldsymbol{R}$ for all $l = 1, \ldots, q$. Note that neither the matrix $\tilde{\boldsymbol{R}}$ nor the matrix $\boldsymbol{R}_{ll'}$ has a product correlation structure, i.e., the elements do not factorize. Also, the common correlation matrix of the multiple endpoints $\boldsymbol{R}$ is unknown and must be estimated. We conclude that, under H_0,

$$\boldsymbol{T} \overset{appr.}{\sim} t_{qk}(\nu, \hat{\tilde{\boldsymbol{R}}}), \tag{5}$$

where $\hat{\tilde{\boldsymbol{R}}}$ is the estimation of $\tilde{\boldsymbol{R}}$.

The decision rule for testing problem (1) is to reject $H_0^{(li)}$ for each difference η_{li} with

$$T_{li} > t_{qk,1-\alpha}(\nu, \hat{\hat{\boldsymbol{R}}}),$$

where $t_{qk,1-\alpha}(\nu, \hat{\hat{\boldsymbol{R}}})$ is a lower $(1-\alpha)$-quantile of the related qk-variate t-distribution. According to Gabriel (1969), this procedure is coherent and consonant with respect to the family of hypotheses induced by (1) and (2). Adjusted p-values p_{li} for the lth comparison on the ith endpoint are given by

$$p_{li} = 1 - \int_{-\infty}^{t_{li}^*} \dots \int_{-\infty}^{t_{li}^*} t_{qk}(\nu, \hat{\hat{\boldsymbol{R}}}; \boldsymbol{t})\, dt_{11} \dots dt_{qk} \quad (l = 1, \dots, q,\ i = 1, \dots, k),$$

where $t_{qk}(\nu, \hat{\hat{\boldsymbol{R}}}; \boldsymbol{t})$ is the related density function, and t_{li}^* is the observed value for test statistic T_{li}. For the computation of the quantiles and p-values, one may resort to the numerical integration routines of Genz and Bretz (2002) and Bretz, Genz, and Hothorn (2001), which are not restricted to special correlation structures. They are available in the package `mvtnorm` (Genz, Bretz, and port by T. Hothorn, 2008, Hothorn, Bretz, and Genz, 2001) of the statistical software R (2009).

Users are interested not only in testing but also in estimating the differences $\eta_{li} \quad (l = 1, \dots, q,\ i = 1, \dots, k)$. SCIs are a method to handle both tasks. The lower limits of the approximate $(1-\alpha)100\%$ SCIs for $(\eta_{11}, \dots, \eta_{qk})'$ are hence given by

$$\hat{\eta}_{li}^{lower} = \bar{X}_{li} - \bar{X}_{0i} - t_{qk,1-\alpha}(\nu, \hat{\hat{\boldsymbol{R}}})\, S_i \sqrt{\frac{1}{n_l} + \frac{1}{n_0}} \quad (l = 1, \dots, q;\ i, i' = 1, \dots, k).$$

Hence, statistical problem (1) can also be decided as follows: For a specified level α, we reject $H_0^{(li)}$ for each difference of means η_{li} with

$$\hat{\eta}_{li}^{lower} > \delta_i.$$

Note that these intervals do not have the same widths. This is because the intervals depend on the sample variances S_i^2, which are different for the endpoints.

3 Simulations concerning the FWE

To derive the exact distribution of $\boldsymbol{T}$ would be a challenging problem and we have chosen in this paper to use a simple approximation based on the familiar multivariate t-distribution. This difficulty stems from the fact that the endpoints have different variances and that their correlations are unknown and must be estimated (approximations (3) and (5)). Therefore, a validation was done by simulations. Three and five treatments, respectively, have been compared in a simulation study. The first treatment is regarded as the (negative) control. The study had different numbers of endpoints with related expected values, i.e., $\boldsymbol{\mu}_l = (10, 100)$ for 2 endpoints, $\boldsymbol{\mu}_l = (0.1, 1, 10, 100)$ for 4 endpoints, and $\boldsymbol{\mu}_l = (0.05, 0.1, 0.5, 1, 5, 10,\ 50, 100)$ for 8 endpoints, respectively for all treatments $l = 0, \ldots, q$. In principle, such settings can only prove the control of the FWE in the weak sense. However, conclusions about the strong control are also allowed, since the test procedure is a union-intersection-test, and the same critical value is used for all comparisons. The endpoints have equicorrelations ρ^{min}, 0, 0.5, 1.[1] The standard deviations are $0.25\boldsymbol{\mu}_l$ for all treatments. The sample size is 20 for each endpoint of each treatment. The FWE has been simulated at a nominal level of $\alpha = 0.05$. The simulation results have been obtained from 10000 simulation runs each and with the starting seed 10000 using a program code in the statistical software R (2009), package `mvtnorm` (Genz et al., 2008, Hothorn et al., 2001) and `SimComp` (Hasler, 2009).

Tables 1 and 2 show the simulated α-level for $q = 2$ and $q = 4$ non-control treatments, respectively. In addition to the new procedure, using an approximate multivariate t-distribution, two further versions have been simulated. The first line for a fixed number of endpoints represents the new method (mv$t$) based on the approximation by a multivariate t-distribution. The second line represents the same procedure but assuming known covariances (mvnorm). This assumption is hardly met in practice, of course, and it leads to a liberal behavior (ranges from 0.050 to 0.067), because a multivariate normal distribution is applied instead. The third line (Bonf) is according to a complete (univariate) Bonferroni adjustment, which is known to produce conservative test decisions (ranges from 0.006 to 0.048). It becomes more conservative for increasing correlations and an increasing number of endpoints. In general, the new procedure (mv$t$) maintains the α-level. The very slight variation around the nominal $\alpha = 0.05$ (ranges from 0.046 to 0.055) is always bounded by the two previous methods. (The only exception in Table 1 for $k = 2$ endpoints with correlation one is obviously attributed to numerical reasons only.)

On the other hand, according to Xu, Nuamah, Liu, Lim, and Sampson (2009) and Liu, Hsu, and Ruberg (2007), applying multivariate t-distributions in the context of multiple endpoints and using the method of Genz and Bretz (2002) may lead to slightly liberal test decisions. We could not support that

[1] The minimal equicorrelation depends on the dimension k by $\rho^{min} = -1/(k-1)$.

Table 1: FWE of one-sided Dunnett tests for $q = 2$ non-control treatments, several numbers of endpoints, several equicorrelations and adjustment methods; $\alpha = 0.05$.

Endpoints	Method	Correlations			
		ρ^{min}	0	0.5	1
$k = 2$	mv$t$	0.051	0.049	0.052	0.055
	mvnorm	0.058	0.059	0.057	0.054
	Bonf	0.046	0.045	0.044	0.024
$k = 4$	mv$t$	0.051	0.053	0.050	0.052
	mvnorm	0.060	0.059	0.058	0.054
	Bonf	0.047	0.048	0.038	0.011
$k = 8$	mv$t$	0.046	0.046	0.051	0.050
	mvnorm	0.065	0.067	0.059	0.050
	Bonf	0.044	0.042	0.037	0.007

conclusion in general. The authors are right in the case of (very) small sample sizes. Hence, in this situation the procedure is unreliable and cannot be recommended without caution. This problem is at least bounded by the demand on the degree of freedom to be greater or equal to the number of endpoints, $\nu = \sum_{l=0}^{q}(n_l - 1) \geq k$. Also, we believe that this risk is acceptable compared to the advantages of the method.

4 Extensions

The procedure presented above can be generalized to other cases of which we now give a short overview.

The many-to-one comparison according to Dunnett (1955) is a special case of multiple contrast tests, which allow the evaluation of a broad class of linear testing problems. The procedure presented can also be generalized to other multiple contrast tests such as the all-pair comparison of Tukey (1953), the trend test of Williams (1971) and Bretz (2006), or user-defined contrast tests.

Multiple contrast tests and related SCIs can be formulated not only for differences but also for ratios of means. That allows testing for relative thresholds, which are often easier to interpret. The test for ratios of means includes the commonly used approach for differences by the special case of relative thresholds equal to one (Dilba, Bretz, Guiard, and Hothorn, 2004). Moreover, related SCIs are comparable for the different endpoints on the percentage scale,

Table 2: FWE of one-sided Dunnett tests for $q = 4$ non-control treatments, several numbers of endpoints, several equicorrelations and adjustment methods; $\alpha = 0.05$.

Endpoints	Method	Correlations			
		ρ^{min}	0	0.5	1
$k = 2$	mv$t$	0.053	0.051	0.049	0.049
	mvnorm	0.056	0.057	0.059	0.050
	Bonf	0.046	0.043	0.038	0.021
$k = 4$	mv$t$	0.052	0.052	0.051	0.048
	mvnorm	0.059	0.053	0.055	0.054
	Bonf	0.045	0.045	0.038	0.011
$k = 8$	mv$t$	0.050	0.052	0.051	0.050
	mvnorm	0.060	0.064	0.061	0.053
	Bonf	0.044	0.045	0.036	0.006

which is particularly helpful for differently scaled endpoints. SCIs for ratios of means are based on a generalization of the Theorem of Fieller (1954). In contrast to intervals for differences of means, the correlation matrix $\tilde{\boldsymbol{R}}$ here depends on the unknown ratios. Dilba et al. (2004) use a plug-in approach here, which is shown to have a very good performance (in the case of one endpoint). Second, the ratio intervals only exist, if the denominator – representing the mean of the control group, for example – is significantly greater than zero (Buonaccorsi and Iyer, 1984). One can hence conclude that the intervals for differences may be harder to interpret, but they are more reliable in a certain manner. Anyway, the procedure presented is also extendable to the case of ratios of means, not only for differences.

A restrictive assumption of our procedure is the equality of the variances and covariances of the endpoints for different treatments, $\boldsymbol{\Sigma}_0 = \cdots = \boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}$. Dose finding studies, for example, can have the problem of heteroscedasticity because the variance depends on the dose effects. Multiple contrast tests, and hence the Dunnett procedure, are also available for heteroscedastic data. Hasler and Hothorn (2008) apply different approximate multivariate t-distributions with different degrees of freedom according to Satterthwaite (1946) instead of a joint one. This principle can be adapted to the case of multiple endpoints. Consequently, qk different approximate qk-variate t-distributions have to be used, instead of one approximate joint one, to obtain quantiles or adjusted p-values.

Table 3: Summary statistics for the coagulation parameters of the data set used in Kropf et al. (2000).

	Thromb. count	ADP	TRAP
Mean of group S	0.872	0.808	0.725
Mean of group H	0.916	0.892	0.796
Mean of group B	0.994	1.020	0.831
Pooled std. dev.	0.251	0.201	0.342

Moreover, the simulations presented in Section 3 are just an extract. We have also done simulations for many other situations like for the ratio approach, for the situation of heterogeneous variances or correlations, and for the SCIs. Related results may be requested from the first author.

5 Example

We refer to the data set used in Kropf et al. (2000) (see Table 3). The aim of this randomized clinical trial was to compare three sets of extracorporeal circulation in heart-lung machines, which vary in their surface configuration. The first new version (treatment H) implies a heparine covering of all parts with blood contact, which is rather expensive. The second one (treatment B) uses a more economical biocompatible surface configuration. The trial should show that the new versions H ($l = 1$) and B ($l = 2$) are superior to the standard S ($l = 0$). Twelve (S and H each) and eleven (B) male adult patients, scheduled for elective coronary bypass grafting, have been considered in a double-blind study. The analysis is based on a set of laboratory parameters restricted to the blood coagulation system, characterized by three primary endpoints thrombocyte count, thrombocyte activity ADP and thrombocyte activity TRAP (each as quotient from post- and pre-surgery values). Higher values indicate a better treatment effect.

The pooled estimated correlations for the three endpoints are given by

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1.000 & 0.874 & 0.468 \\ 0.874 & 1.000 & 0.382 \\ 0.468 & 0.382 & 1.000 \end{pmatrix}$$

The differences of interest are

$$\eta_{li} = \mu_{li} - \mu_{0i} \quad (l = 1, 2,\ i = 1, 2, 3),$$

DOI: 10.2202/1557-4679.1258

and the hypotheses to be tested are given by

$$H_0^{(li)} : \eta_{li} \leq 0 \quad (l = 1, 2,\ i = 1, 2, 3).$$

Table 4: Lower limits of the approximate 95% SCIs (and point estimates) per comparison and coagulation parameter of the data set used in Kropf et al. (2000).

Comparison	Thromb. count	ADP	TRAP
H − S	−0.199 (0.044)	−0.111 (0.084)	−0.260 (0.071)
B − S	−0.127 (0.122)	0.013 (0.212)	−0.234 (0.105)

Table 4 shows the lower limits for the related approximate 95% SCIs for the differences to the control. The values in parentheses are the estimated differences. Treatment B shows values significantly greater than those of the standard S for endpoint ADP. If accepting non-inferiority thresholds of −0.200 (Thromb.count), −0.112 (ADP), and −0.261 (TRAP), then the two new treatments H and B are non-inferior for all endpoints.

The package `SimComp` (Hasler, 2009) of the statistical software R (2009) was used to evaluate the example data. The relevant cutout of the original data set, which was already used in Kropf et al. (2000), can be loaded from this package, too. For simplicity, additional endpoints have been omitted, and the names of the endpoints have been simplified. Both the statistical software R (2009) and the package `SimComp` (Hasler, 2009) are available at `http://www.r-project.org`. The input for the example is

```
SimCiDiff(data=coagulation,

grp="Group",

resp=c("Thromb.count","ADP","TRAP"),

type="Dunnett",

base=3,

alternative="greater",

covar.equal=TRUE).
```

For the related p-values use the command `SimTestDiff()`, and for ratio-based testing and intervals `SimTestRat()` and `SimCiRat()`, respectively.

6 Conclusions

The Dunnett procedure and related SCIs had been restricted to comparisons on a single endpoint so far. This methodology was extended to the case of multiple endpoints by means of an approximate multivariate t-distribution. In this manner, correlations among both the comparisons and the endpoints can be taken into account. Test decisions – adjusted p-values as well as SCIs – are available for all comparisons and all endpoints. This is clearly the major advantage of this procedure. The intervals and tests may be one- or two-sided. Moreover, extensions are available for the consideration of other multiple contrast tests, for ratios of means, and for heterogeneous covariance matrices, respectively.

The procedure presented maintains the FWE in the strong sense in a passable range. This was shown by simulations. Possible slight variations around the nominal FWE α are necessarily bounded by the versions mvnorm and Bonf considered in Section 3. Furthermore, they are bounded by the demand on the degree of freedom to be greater or equal to the number of endpoints ($\nu = \sum_{l=0}^{q}(n_l - 1) \geq k$). Thus, the procedure is not defined for too small sample sizes.

Power simulations and comparisons with other methods, which are frequently presented in the literature, are not provided. Especially Gatekeeping procedures may be expected to have a higher power. However, against the background of the functionality of the new method, a power comparison with existing methods is not really feasible or fair. Nevertheless, power is an important aspect. An approach for a global power estimation, based on a non-central multivariate t-distribution, is given in Hasler (2010) as well as a power comparison (for $q = 1$ non-control group) with a t-test-based bootstrap approach of Pollard, Ge, Taylor, and Dudoit (2007), available in R (2009).

The package `SimComp` (Hasler, 2009) of the statistical software R (2009) provides calculations concerning simultaneous tests and confidence intervals for both difference- and ratio-based contrasts of normal means for data with possibly more than one primary endpoint. The covariance matrices - containing the covariances between the endpoints - may be assumed to be equal or possibly unequal for the different groups. This package was used to analyze the example presented in Section 5.

References

Bauer, P. (1991): "Multiple testing in clinical-trials," *Statistics In Medicine*, 10, 871–890.

Bretz, F. (2006): "An extension of the williams trend test to general unbalanced linear models," *Computational Statistics & Data Analysis*, 50, 1735–1748.

Bretz, F., A. Genz, and L. A. Hothorn (2001): "On the numerical availability of multiple comparison procedures," *Biometrical Journal*, 43, 645–656.

Buonaccorsi, J. P. and H. K. Iyer (1984): "A comparison of confidence regions and designs in estimation of a ratio," *Communications in Statistics-Theory and Methods*, 13, 723–741.

Cohen, A., H. B. Sackrowitz, and M. Y. Xu (2008): "The use of an identity in anderson for multivariate multiple testing," *Journal Of Statistical Planning And Inference*, 138, 2615–2621.

Dilba, G., F. Bretz, V. Guiard, and L. A. Hothorn (2004): "Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control," *Methods Of Information In Medicine*, 43, 465–469.

Dmitrienko, A., W. W. Offen, and P. H. Westfall (2003): "Gatekeeping strategies for clinical trials that do not require all primary effects to be significant," *Statistics In Medicine*, 22, 2387–2400.

Dunnett, C. W. (1955): "A multiple comparison procedure for comparing several treatments with a control," *Journal Of The American Statistical Association*, 50, 1096–1121.

Fieller, E. C. (1954): "Some problems in interval estimation," *Journal of the Royal Statistical Society, Series B*, 16, 175–185.

Gabriel, K. R. (1969): "Simultaneous test procedures - Some theory of multiple comparisons," *Annals Of Mathematical Statistics*, 40, 224–250.

Genz, A. and F. Bretz (2002): "Methods for the computation of multivariate t-probabilities," *Journal of Computational and Graphical Statistics*, 11, 950–971.

Genz, A., F. Bretz, and R. port by T. Hothorn (2008): *mvtnorm: Multivariate Normal and T Distribution*, r package version 0.8-3.

Hasler, M. (2009): *SimComp: Simultaneous comparisons for multiple endpoints*, r package version 1.4.3.

Hasler, M. (2010): "Multiple contrast tests for multiple endpoints," Technical report, Institute of Biostatistics, Leibniz University Hannover.

Hasler, M. and L. A. Hothorn (2008): "Multiple contrast tests in the presence of heteroscedasticity," *Biometrical Journal*, 50, 793–800.

Hochberg, Y. (1988): "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, 75, 800–802.

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

Hommel, G. (1988): "A stagewise rejective multiple test procedure based on a modified Bonferroni test," *Biometrika*, 75, 383–386.

Hotelling, H. (1951): "A generalised T test and measure of multivariate dispersion," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 23–41.

Hothorn, T., F. Bretz, and A. Genz (2001): "On multivariate t and Gauss probabilities in R," *R News*, 1, 27–29.

ICH E9 Expert Working Group (1999): "Ich harmonised tripartite guideline: Statistical principles for clinical trials," *Statistics In Medicine*, 18, 1903–1904.

Imada, T. and H. Douke (2007): "Step down procedure for comparing several treatments with a control based on multivariate normal response," *Biometrical Journal*, 49, 18–29.

Imada, T. and H. Douke (2008): "Step-up procedure for multiple comparison with a control for multivariate normal means," *Communications In Statistics-Simulation And Computation*, 37, 1810–1824.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000): *Continuous Multivariate Distributions*, John Wiley and Sons, Inc., New York, 2 edition.

Kropf, S., G. Hommel, U. Schmidt, J. Brickwedel, and M. S. Jepsen (2000): "Multiple comparisons of treatments with stable multivariate tests in a two-stage adaptive design, including a test for non-inferiority," *Biometrical Journal*, 42, 951–965.

Kropf, S., L. A. Hothorn, and J. Läuter (1997): "Multivariate many-to-one procedures with applications to preclinical trials," *Drug Information Journal*, 31, 433–447.

Liu, Y., J. Hsu, and S. Ruberg (2007): "Partition testing in dose-response studies with multiple endpoints," *Pharmaceutical Statistics*, 6, 181–192.

Neuhäuser, M. (2006): "How to deal with multiple endpoints in clinical trials," *Fundamental & Clinical Pharmacology*, 20, 515–523.

Pollard, K. S., Y. Ge, S. Taylor, and S. Dudoit (2007): *multtest: Resampling-based multiple hypothesis testing*, r package version 1.16.1.

R Development Core Team (2009): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Satterthwaite, F. E. (1946): "An approximate distribution of estimates of variance components," *Biometrics*, 2, 110–114.

Schulte, A., J. Althoff, S. Ewe, and H. B. Richter-Reichhelm (2002): "Two immunotoxicity ring studies according to OECD TG 407 - Comparison of data on cyclosporin A and hexachlorobenzene," *Regulatory Toxicology And Pharmacology*, 36, 12–21.

Seo, T. and T. Nishiyama (2008): "On the conservative simultaneous confidence procedures for multiple comparisons among mean vectors," *Journal of Statistical Planning and Inference*, 138, 3448 – 3456.

Siddiqui, M. M. (1967): "A bivariate t distribution," *Annals Of Mathematical Statistics*, 38, 162–166.

Tukey, J. W. (1953): "The problem of multiple comparisons," Dittoed manuscript of 396 pages New Jersey: Department of Statistics, Princeton University.

Williams, D. A. (1971): "A test for differences between treatment means when several dose levels are compared with a zero dose control," *Biometrics*, 27, 103–117.

Xu, H. Y., I. Nuamah, J. Y. Liu, P. Lim, and A. Sampson (2009): "A Dunnett-Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints," *Pharmaceutical Statistics*, 8, 301–316.