Research Article

Mario Hasler*

Multiple Contrast Tests for Multiple Endpoints in the Presence of Heteroscedasticity

Abstract: This article describes an extension of multiple contrast tests to the case of multiple, correlated endpoints. These endpoints are assumed to be normally distributed with different scales and variances. Unlike in previous articles, the covariance matrices are also assumed to be unequal for the treatment groups. Approximate multivariate *t*-distributions are used to obtain multiplicity-adjusted *p*-values and quantiles for test decisions or 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 semi-synthetic example data set of a randomized, placebo-controlled phase IIb dose-finding study of a novel anti-muscarinic agent for five continuous endpoints. A related R package is available.

Keywords: multiple contrast tests; multiple endpoints; heteroscedasticity; multivariate *t*-distribution; simultaneous confidence intervals

***Corresponding author: Mario Hasler,** Christian-Albrechts-University of Kiel, Schleswig-Holstein D-24118, Germany, E-mail: hasler@email.uni-kiel.de

1 Introduction

Randomized clinical trials and pre-clinical studies often cover many correlated endpoints. The scales of these endpoints are often different. The experimental goal is not only to clarify which treatment groups differ but also for which endpoints. Hence, it is not clear a priori, 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 jointly – not separately. Multiplicity adjustment must then take both the number of treatment comparisons and the number of endpoints into account. The family-wise error type I (FWE) must be maintained with regard to all endpoints simultaneously. In addition, their correlations are important to be considered. First, the degree of conservatism of the elementary test decisions is reduced by taking them into account. Second, effects may be erroneously ignored or masked when analysing 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.

Hasler and Hothorn [1] have described an extension of the Dunnett procedure [2] to the case of multiple, correlated endpoints. The focus is on simultaneous confidence intervals (SCIs) for differences of means. In a further paper [3], the authors have presented an extension of the trend test by Williams [4] and Bretz [5], respectively. Here the focus is on SCIs for ratio of means (based on Dilba et al. [6, 7]). Both procedures take the correlations of the endpoints into account, and the multiplicity adjustment includes both the number of endpoints and the common treatment comparisons. A related approximate multivariate *t*-distribution is used to obtain quantiles for SCIs and test decisions or to obtain multiplicity-adjusted *p*-values. The FWE is maintained in the strong sense in an admissible range. The procedures assume multivariate normally distributed endpoints with possibly different scales, allowing endpoint-specific variances.

However, one further assumption is that the covariance matrices – containing the covariances of the endpoints – are equal for all treatments. This is not fulfilled in situations when variances or correlations differ due to the treatment groups. This problem is briefly addressed by Hasler and Hothorn [1, 3], but a solution is only suggested. In addition, the suggestions made in these two papers are slightly different. Moreover, many-to-one comparisons based on Dunnett [2], or trend test based on Williams [4] and Bretz [5] are special cases of multiple contrast tests (MCTs), which allow the evaluation of a broad class of linear testing problems, such as the all-pair comparison of Tukey [8], or any user-defined contrast tests. This article presents an extension of the methods of Hasler and Hothorn [1, 3] to the general case of MCTs for multivariate normally distributed endpoints with unequal covariance matrices for the treatment groups. Former suggestions are considered in detail. In Section 2, the testing problem is formulated, approximate distributions of the test statistics are derived for several approaches. Section 3 shows results of simulations concerning the FWE. SCIs are presented in Section 4. An example is given in Section 5, conclusions and a discussion in Section 6.

2 Testing problem and test procedures

2.1 Testing problem

For $h = 1, \ldots, p$, $i = 1, \ldots, k$ and $j = 1, \ldots, n_h$, let X_{hij} denote the jth observation on the ith endpoint under the hth treatment in a one-way layout. Each endpoint is measured for all $N = \sum_{h=1}^{p} n_h$ objects. The vectors $(X_{h1j}, \ldots, X_{hkj})'$ are mutually independent and follow k-variate normal distributions with mean vectors $\boldsymbol{\mu}_h = (\mu_{h1}, \ldots \mu_{hk})'$ and unknown covariance matrices $\sum_h = (\sigma_{h,ii'})_{i,i'} \in \mathbb{R}^{k \times k}$. This means that both the endpoints and the treatments cause different variances or correlations over the endpoints, i.e.

$$\{X_{hij} : i = 1, \ldots, k\} \sim \perp N_k\left(\boldsymbol{\mu}_h, \sum\nolimits_h\right) \quad (h = 1, \ldots, p, j = 1, \ldots, n_h).$$

Let $\bar{\boldsymbol{X}}_h = (\bar{X}_{h1}, \ldots, \bar{X}_{hk})'$ and $\hat{\sum}_h = (\hat{\sigma}_{h,ii'})_{i,i'}$ be the sample mean vectors and the sample covariance matrices, respectively. The diagonal elements of $\hat{\sum}_h$, required for the following test procedures, are denoted by S_{hi}^2. A necessary condition for the $\hat{\sum}_h$ to be well-defined, i.e. to have full rank, is that $(n_h - 1) \geq k$ for all $h = 1, \ldots, p$. Otherwise, $\hat{\sum}_h$ cannot be used for the following procedures. This problem occurs naturally in high-dimensional multivariate data analysis in genetics, for example. These type of data is not focused on in this article, however.

Of interest are the contrasts

$$\eta_{li} = \sum_{h=1}^{p} c_{lh}\mu_{hi} \quad (l = 1, \ldots, q, i = 1, \ldots, k),$$

representing linear combinations of the means μ_{hi}, where $c_{lh} \in [-1, 1]$ are the contrast coefficients. As already noted in Hasler and Hothorn [1], the following procedures could also be formulated in terms of ratios of means, see Hasler and Hothorn [3] as an example. The aim is to test the hypotheses

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

where $\delta_i \in (-\infty, \infty)$ 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 also for endpoint-specific contrast coefficients and thresholds. If the test direction is to be reversed for some endpoints, the corresponding test statistics have to be multiplied by minus one. 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 contrast on at least one endpoint.

2.2 Test procedures

For the case of equal covariance matrices, $\sum_1 = \cdots = \sum_p = \sum$, a less demanding sample size assumption is sufficient, namely $\sum_{h=1}^{p}(n_h - 1) \geq k$. The test of hypotheses (1) will be based on the test statistics

$$T_{li}^{hom} = \frac{\hat{\eta}_{li} - \delta_i}{S_i \sqrt{\sum_{h=1}^{p} \frac{c_{lh}^2}{n_h}}} \quad (l = 1, \ldots, q, i = 1, \ldots, k),$$

where $\hat{\eta}_{li} = \sum_{h=1}^{p} c_{lh}\bar{X}_{hi}$, and S_i^2 is the pooled sample variance for endpoint i. Hasler and Hothorn [1] have shown that the joint distribution of $T_{11}^{hom}, \ldots, T_{qk}^{hom}$ is not among the standard distributions discussed in the literature. This is because the endpoints have different scales and variances. Therefore, they use an approximation by a qk-variate t-distribution. The authors considered contrasts that were related specifically to many-to-one comparisons based on Dunnett [2], but their conclusions are also valid here for general contrasts η_{li}. Hence, the test statistics $T_{11}^{hom}, \ldots, T_{qk}^{hom}$ follow approximately a joint qk-variate t-distribution with degree of freedom

$$\nu^{hom} = \sum_{h=1}^{p}(n_h - 1) \tag{3}$$

and correlation matrix

$$\boldsymbol{R}^{hom} = (\boldsymbol{R}_{ll'}^{hom})_{l,l'} = \begin{pmatrix} \boldsymbol{R}_{11}^{hom} & \boldsymbol{R}_{12}^{hom} & \ldots & \boldsymbol{R}_{1q}^{hom} \\ \boldsymbol{R}_{12}^{hom} & \boldsymbol{R}_{22}^{hom} & \ldots & \boldsymbol{R}_{2q}^{hom} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}_{1q}^{hom} & \boldsymbol{R}_{2q}^{hom} & \ldots & \boldsymbol{R}_{qq}^{hom} \end{pmatrix}. \tag{4}$$

The submatrices $\boldsymbol{R}_{ll'}^{hom} = (\rho_{ll',ii'}^{hom})_{i,i'} \in [-1, 1]^{k \times k}$ describe the correlations between the lth and the l'th comparison for all endpoints. Their elements are

$$\rho_{ll',ii'}^{hom} = \rho_{ii'} \frac{\sum_{h=1}^{p} \frac{c_{lh}c_{l'h}}{n_h}}{\sqrt{\sum_{h=1}^{p} \frac{c_{lh}^2}{n_h}} \sqrt{\sum_{h=1}^{p} \frac{c_{l'h}^2}{n_h}}} \quad (l, l' = 1, \ldots, q;\ i, i' = 1, \ldots, k), \tag{5}$$

where $\rho_{ii'}$ are the elements of the common correlation matrix of the multiple endpoints, $\boldsymbol{R} = (\boldsymbol{\rho}_{ii'})_{i,i'}$. This procedure is a direct generalization of the procedure by Hasler and Hothorn [1], and it is referred to here as the HOM procedure. Because it does not take the real nature of the data into account, it is clear that the HOM procedure does not control the FWE if the assumption of equal covariance matrices is violated.

For the case of unequal covariance matrices, i.e. there exist at least two groups h, h' where $\sum_h \neq \sum_{h'}$, the test of the hypotheses (1) will be based on the test statistics

$$T_{li}^{het} = \frac{\hat{\eta}_{li} - \delta_i}{\sqrt{\sum_{h=1}^{p} \frac{c_{lh}^2 S_{hi}^2}{n_h}}} \quad (l = 1, \ldots, q, i = 1, \ldots, k).$$

Under $H_0^{(li)}$, test statistic T_{li}^{het} is approximately t-distributed with degree of freedom

$$\nu_{li} = \frac{\left(\sum_{h=1}^{p} \frac{c_{lh}^2 S_{hi}^2}{n_h}\right)^2}{\sum_{h=1}^{p} \frac{c_{lh}^4 S_{hi}^4}{n_h^2 (n_h - 1)}} \quad (l = 1, \ldots, q,\ i = 1, \ldots, k) \tag{6}$$

according to Satterthwaite [9]. The normal strategy would be now to derive the (approximate) joint distribution of all test statistics $T_{11}^{het}, \ldots, T_{qk}^{het}$ to allow taking the correlations of both the contrasts and the endpoints into account. If the covariance matrices $\sum_h$ would be known, $T_{11}^{het}, \ldots, T_{qk}^{het}$ would follow a joint qk-variate normal distribution. However, in contrast to the homoscedastic case, the covariance matrices here additionally are unequal for the treatment groups. For that reason, the plug-in procedure (PI) for heterogeneous variances of Hasler and Hothorn [10] will be adopted and extended to the case of multiple endpoints, as was suggested by Hasler and Hothorn [1]. Therefore, qk different approximate qk-variate t-distributions must be applied, each with a contrast- and endpoint-specific degree of freedom according to eq. (6). Hence, each test statistic T_{li} is compared with its own distinct quantile to come to a test decision. In the same way, each adjusted p-value is calculated based on its own qk-variate t-distribution. The correlation matrix $\boldsymbol{R}^{het} = (\boldsymbol{R}_{ll'}^{het})_{l,l'}$ of these qk-variate t-distributions has the same structure as $\boldsymbol{R}^{hom}$ in eq. (4), but the elements are

$$\rho_{ll',ii'}^{het} = \frac{\sum_{h=1}^{p} \frac{c_{lh} c_{l'h} \sigma_{h,ii'}}{n_h}}{\sqrt{\sum_{h=1}^{p} \frac{c_{lh}^2 \sigma_{h,ii}}{n_h}} \sqrt{\sum_{h=1}^{p} \frac{c_{l'h}^2 \sigma_{h,i'i'}}{n_h}}} \quad (l, l' = 1, \ldots, q;\ i, i' = 1, \ldots, k). \tag{7}$$

This procedure is referred to here as the CE procedure, indicating that contrast- and endpoint-specific degrees of freedom are used.

As an alternative to the above-mentioned approach, Hasler and Hothorn [3] suggested another version, namely the same test statistics and correlation matrix as for the CE procedure but only contrast-specific degrees of freedom. Therefore define

$$\nu_l = \min_{i=1,\ldots,k} \nu_{li} \quad (l = 1, \ldots, q). \tag{8}$$

For each contrast, the minimum of the degrees of freedom (6) is taken over the endpoints. Hence, q different qk-variate t-distributions will be applied, and $\nu_l \leq \nu_{li}$ for all $l = 1, \ldots, q$ and $i = 1, \ldots, k$. This also leads to different, non-equidistant quantiles for the test decisions. This procedure is referred to here as the MIN procedure, indicating that minimized contrast-specific degrees of freedom are used.

Considering the elements of the correlation matrix belonging to CE and MIN, one can see that the correlations of MCTs in the presence of heteroscedasticity [10] are recovered for $i = i'$. Hence, the conventional case of a single endpoint ($k = 1$) is a special case of the method described in this article. Note that neither the matrices $\boldsymbol{R}^{hom}$ and $\boldsymbol{R}^{het}$ nor the matrices $\boldsymbol{R}_{ll'}^{hom}$ and $\boldsymbol{R}_{ll'}^{het}$ have a product correlation structure, i.e. the elements do not factorize. Furthermore, the elements of the covariance matrices $\sum_1, \ldots, \sum_p$ and also of the common correlation matrix $\boldsymbol{R} = (\rho_{ii'})_{i,i'}$ are unknown and must be estimated.

The decision rule for testing problem (1) is to reject $H_0^{(li)}$ for η_{li}, if test statistic T_{li} is greater than the lower $(1 - \alpha)$-quantile of the related qk-variate t-distribution. For the computation of the quantiles or adjusted p-values, one may resort to the numerical integration routines of Genz and Bretz [11] and Bretz et al. [12], which are available in the package `mvtnorm` [13, 14] of the statistical software R [15].

3 Simulations concerning the FWE

As already in the case of equal covariance matrices [1], the derivation of the exact joint distribution of the test statistics would be a challenging problem. The endpoints have different scales, their covariances must

be estimated, and the covariance matrices are unequal. In this article, approximations are used based on multivariate t-distributions. Therefore, a validation was done by simulations. Three and five treatment groups, respectively, have been compared in a simulation study. The first group was regarded as the control. The study had different numbers of endpoints with related expected values: $\boldsymbol{\mu}_h = (10, 100)'$ for 2 endpoints, $\boldsymbol{\mu}_h = (0.1, 1, 10, 100)'$ for 4 endpoints, and $\boldsymbol{\mu}_h = (0.05, 0.1, 0.5, 1, 5, 10, 50, 100)'$ for 8 endpoints, respectively, for all treatment groups $h = 1, \ldots, p$. The most critical situation in the context of heteroscedasticity is if the treatment group with the highest variance has the smallest sample size because approaches for homoscedastic data yield very liberal test decisions in this situation (e.g. see Hasler and Hothorn [10]). For that reason, the first $p-1$ treatment groups had same covariance matrices, $\sum_1 = \cdots = \sum_{p-1}$, with standard deviations $\sqrt{diag\,(\sum_h)} = 0.1\,\boldsymbol{\mu}_h$ and sample size $n_h = 20$ for each endpoint of each group ($h = 1, \ldots, p-1$); the last group had standard deviations $\sqrt{diag\,(\sum_p)} = 0.25\,\boldsymbol{\mu}_p$ and sample size $n_p = 10$ for each endpoint. Three equicorrelation structures (compound symmetry) of the endpoints were chosen ($\rho^{\min} = -1/(k-1)$, 0, 0.8) as well as a random correlation structure (rand.), which was different for each simulation run. For the random correlation structure, the first $p-1$ groups always had the same correlations per run. Four one-sided MCT problems were considered which are all related to hypotheses (1): Dunnett (many-to-one), Tukey (all-pair), Williams (trend), and Average (mean averages). Usually, a Tukey MCT is a two-sided testing problem. For reasons of consistency this fact was disregarded. The FWE has been simulated at a nominal level of $\alpha = 0.05$. The simulation results have been obtained from 10,000 simulation runs each, with starting seed 10,000, using a program code in the statistical software R [15], packages `mvtnorm` [13, 14] and `SimComp` [16].

Tables 1 and 2 show the simulated α-level for Dunnett and Tukey contrasts, respectively. The results for the Williams and Average contrasts are not shown here for reasons of brevity, but they can be obtained from

Table 1 FWE of one-sided MCTs (Dunnett contrasts) for several numbers of treatment groups and endpoints, several correlations and procedures; $\alpha = 0.05$

Groups	Endpoints	Correlations	Procedures			
			MIN	CE	BON	HOM
		$\rho = \rho^{\min}$	0.049	0.049	0.048	0.141
		$\rho = 0$	0.050	0.052	0.049	0.135
	$k = 2$	$\rho = 0.8$	0.052	0.053	0.042	0.121
		$\rho =$ rand.	0.050	0.052	0.045	0.131
		$\rho = \rho^{\min}$	0.053	0.058	0.056	0.200
		$\rho = 0$	0.051	0.056	0.052	0.190
$p = 3$	$k = 4$	$\rho = 0.8$	0.050	0.052	0.031	0.131
		$\rho =$ rand.	0.050	0.054	0.046	0.173
		$\rho = \rho^{\min}$	0.048	0.056	0.053	0.267
		$\rho = 0$	0.047	0.055	0.050	0.250
	$k = 8$	$\rho = 0.8$	0.057	0.061	0.030	0.158
		$\rho =$ rand.	0.047	0.054	0.044	0.230
		$\rho = \rho^{\min}$	0.053	0.053	0.047	0.160
		$\rho = 0$	0.050	0.051	0.044	0.157
	$k = 2$	$\rho = 0.8$	0.051	0.052	0.038	0.126
		$\rho =$ rand.	0.054	0.055	0.046	0.148
		$\rho = \rho^{\min}$	0.052	0.058	0.051	0.225
		$\rho = 0$	0.056	0.060	0.050	0.208
$p = 5$	$k = 4$	$\rho = 0.8$	0.051	0.053	0.030	0.145
		$\rho =$ rand.	0.047	0.051	0.041	0.192
		$\rho = \rho^{\min}$	0.051	0.059	0.052	0.318
		$\rho = 0$	0.048	0.055	0.048	0.300
	$k = 8$	$\rho = 0.8$	0.058	0.062	0.028	0.166
		$\rho =$ rand.	0.055	0.061	0.049	0.280

Table 2 FWE of one-sided MCTs (Tukey contrasts) for several numbers of treatment groups and endpoints, several correlations and procedures; $\alpha = 0.05$

Groups	Endpoints	Correlations	Procedures			
			MIN	CE	BON	HOM
		$\rho = \rho^{\min}$	0.049	0.049	0.038	0.151
		$\rho = 0$	0.047	0.050	0.040	0.146
	$k = 2$	$\rho = 0.8$	0.050	0.051	0.033	0.128
		$\rho =$ rand.	0.048	0.050	0.037	0.137
		$\rho = \rho^{\min}$	0.044	0.050	0.040	0.219
		$\rho = 0$	0.050	0.056	0.043	0.205
$p = 3$	$k = 4$	$\rho = 0.8$	0.052	0.055	0.028	0.146
		$\rho =$ rand.	0.049	0.053	0.037	0.189
		$\rho = \rho^{\min}$	0.049	0.059	0.045	0.300
		$\rho = 0$	0.050	0.060	0.043	0.275
	$k = 8$	$\rho = 0.8$	0.054	0.059	0.023	0.164
		$\rho =$ rand.	0.049	0.060	0.040	0.258
		$\rho = \rho^{\min}$	0.050	0.050	0.034	0.183
		$\rho = 0$	0.050	0.052	0.036	0.170
	$k = 2$	$\rho = 0.8$	0.052	0.053	0.033	0.140
		$\rho =$ rand.	0.052	0.053	0.033	0.167
		$\rho = \rho^{\min}$	0.050	0.056	0.039	0.268
		$\rho = 0$	0.048	0.054	0.038	0.251
$p = 5$	$k = 4$	$\rho = 0.8$	0.054	0.057	0.025	0.166
		$\rho =$ rand.	0.052	0.056	0.035	0.216
		$\rho = \rho^{\min}$	0.050	0.059	0.040	0.371
		$\rho = 0$	0.051	0.059	0.040	0.355
	$k = 8$	$\rho = 0.8$	0.062	0.066	0.026	0.188
		$\rho =$ rand.	0.051	0.061	0.038	0.312

the author on request. It is clear a priori that the FWE of CE must be greater than the FWE of MIN since $v_{li} \geq v_l$ for all $l = 1, \ldots, q$ and $i = 1, \ldots, k$. Indeed, CE shows a liberal behaviour (ranges from 0.049 to 0.066). In addition to the procedures described, a further version has been simulated. Procedure BON is according to a complete (univariate) Bonferroni adjustment with degrees of freedom each according to eq. (6). It is known to produce conservative test decisions (ranges from 0.023 to 0.056), especially for an increasing number of comparisons. Note that the different MCTs imply different numbers of contrasts and also different correlations among them. In general, the MIN procedure maintains the α-level in an admissible range. The slight variation around the nominal $\alpha = 0.05$ (ranges from 0.044 to 0.062) is always bounded by the two other procedures; α-exceeding is very rare. However, MIN can also be slightly conservative compared to BON in a few cases, see Table 1 for $p = 3$, $k = 4$, $\rho = \rho^{\min}$, for example. This is clearly caused by the fact that BON uses contrast- and endpoint-specific degrees of freedom (6), and $v_{li} \geq v_l$ for all $l = 1, \ldots, q$ and $i = 1, \ldots, k$. Furthermore, a multivariate approach is known not to have most gain in power compared to a univariate Bonferroni-adjustment if related correlations are high. This is, the gain in power of MIN compared to BON is very low for negative or small correlations. Procedure HOM represents the method for homogeneous covariance matrices. Expectedly, HOM is strongly liberal (ranges from 0.121 to 0.371) and shows how important it is not to ignore heteroscedasticity of the data. It is most liberal for negative correlations of the endpoints ($\rho = \rho^{\min}$) and less liberal (but still strongly) for positive ($\rho = 0.8$). Generally, the correlations of the endpoints have no deciding influence for MIN and CE. This fact coincides with the results of Hasler and Hothorn [1] in the case of equal covariance matrices.

According to Xu et al. [17] and Liu et al. [18], applying multivariate t-distributions in the context of multiple endpoints and using the method of Genz and Bretz [11] may lead to slightly liberal test decisions. Also for that reason, the degrees of freedom for the MIN procedure according to eq. (8) are defined in a

conservative manner. For each contrast, the minimum of the degrees of freedom (6) is taken over the endpoints.

Although the procedures presented allow unequal covariance matrices for the treatment groups, the multivariate normal distribution is still a strong assumption. Therefore, a further simulation study was done to check how sensitive the proposed methods are to violations of the multivariate normal distribution. Similarly to the above study, three and five treatment groups, respectively, have been compared. The first group was regarded as the control. The study had different numbers of correlated log-normally distributed endpoints based on multivariate normally distributed data with related parameters: $\boldsymbol{\mu}_h = (1,3)'$, $\sqrt{diag\ (\sum_h)} = (0.25, 0.6)'$ for 2 endpoints, $\boldsymbol{\mu}_h = (0.25, 1, 2, 3)'$, $\sqrt{diag\ (\sum_h)} = (0.15, 0.25, 0.4, 0.6)'$ for 4 endpoints, and $\boldsymbol{\mu}_h = (0.1, 0.25, 0.5, 1, 1.5, 2, 2.5, 3)'$, $\sqrt{diag(\sum_h)} = (0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6)'$ for 8 endpoints, respectively for all treatment groups $h = 1, \ldots, p-1$, where the sample size was $n_h = 20$ for each endpoint of each group. The last group had parameters: $\sqrt{diag(\sum_h)} = (0.5, 0.9)'$ for 2 endpoints, $\sqrt{diag(\sum_h)} = (0.3, 0.5, 0.65, 0.9)'$ for 4 endpoints, and $\sqrt{diag(\sum_h)} = (0.2, 0.3, 0.4, 0.5, 0.55, 0.65, 0.8, 0.9)'$ for 8 endpoints, respectively, where the sample size was $n_p = 10$ for each endpoint. As mean and variance of a log-normally distributed variable depend on each other, the means of the last group were chosen accordingly. The endpoints had a random correlation structure.

Tables 3 and 4 show the simulated α-level for Dunnett and Tukey contrasts, respectively, for the multivariate log-normal data. As expected, simulated and nominal α-level differ. Independent of the number of treatment groups, endpoints, or contrast, the procedures yield conservative test decisions, except for HOM which is liberal (ranges from 0.074 to 0.162). MIN (ranges from 0.021 to 0.034) is more conservative than CE (ranges from 0.022 to 0.035), and BON is most conservative (ranges from 0.015 to 0.028). Of course, HOM is not surprising as it still ignores the heteroscedasticity problem. Hence, even in this situation, where the procedures MIN and CE are not developed for, they show an acceptable behaviour in the sense that they do not exceed the nominal α-level. Of course, they cannot be recommended without caution if the data do not follow a multivariate normal distribution.

Table 3 FWE of one-sided MCTs (Dunnett contrasts) for several numbers of treatment groups and correlated log-normally distributed endpoints, and several procedures; $\alpha = 0.05$

Groups	Endpoints	Procedures			
		MIN	CE	BON	HOM
$p = 3$	$k = 2$	0.028	0.028	0.024	0.077
	$k = 4$	0.025	0.026	0.023	0.093
	$k = 8$	0.027	0.029	0.024	0.113
$p = 5$	$k = 2$	0.034	0.035	0.028	0.099
	$k = 4$	0.034	0.035	0.028	0.106
	$k = 8$	0.032	0.035	0.026	0.150

Table 4 FWE of one-sided MCTs (Tukey contrasts) for several numbers of treatment groups and correlated log-normally distributed endpoints, and several procedures; $\alpha = 0.05$

Groups	Endpoints	Procedures			
		MIN	CE	BON	HOM
$p = 3$	$k = 2$	0.025	0.025	0.018	0.074
	$k = 4$	0.021	0.022	0.015	0.089
	$k = 8$	0.021	0.024	0.017	0.117
$p = 5$	$k = 2$	0.028	0.029	0.018	0.093
	$k = 4$	0.031	0.034	0.021	0.123
	$k = 8$	0.030	0.035	0.022	0.162

4 Simultaneous confidence intervals

For test decisions, as well as for the estimation of the contrasts $\eta_{11}, \ldots, \eta_{qk}$, approximate $(1-\alpha)\,100\%$ SCIs can be derived. The following intervals are related to the MIN procedure. Corresponding lower limits are given by

$$\hat{\eta}_{li}^{\text{lower}} = \hat{\eta}_{li} - t_{qk,1-\alpha}(\nu_l, \hat{\boldsymbol{R}}^{het}) \sqrt{\sum_{h=1}^{p} \frac{c_{lh}^2 S_{hi}^2}{n_h}} \quad (l = 1, \ldots, q;\ i = 1, \ldots, k),$$

where $t_{qk,1-\alpha}(\nu_l, \hat{\boldsymbol{R}}^{het})$ is the lower $(1-\alpha)$-quantile of the related qk-variate t-distribution, and $\hat{\boldsymbol{R}}^{het}$ is the estimation of the correlation matrix $\boldsymbol{R}^{het}$. Hence, statistical problem (1) can also be decided as follows: for a specified level α, reject $H_0^{(li)}$ for η_{li}, if $\hat{\eta}_{li}^{\text{lower}} > \delta_i$. Note that these intervals do not have same widths. This is because the intervals depend on different degrees of freedom, which are different for the contrast, and on the sample variances, which are different for the endpoints and the treatment groups. Hence, an interval is wider if variances of the corresponding treatment groups and endpoints are greater, and tighter if smaller.

5 Example

Homma et al. [19] published summary data for five multiple, continuous endpoints of the randomized, placebo-controlled phase IIb dose-finding study of a novel anti-muscarinic agent. For a one-way layout with a zero-dose placebo group ($h = 1$) and imidafenacin dose groups of 0.1 ($h = 2$), 0.2 ($h = 3$), and 0.5 ($h = 4$) mg/day, the percentage changes from the baseline were defined as primary efficacy endpoints: (1) *incontinence episodes per week (Iepw)*, (2) *urgency incontinence episodes per week (Uiepw)*, (3) *micturitions per day (Mpd)*, (4) *urgency episodes per day (Uepd)*, and (5) *urine volume voids per micturition (Uvvpm)*. Although these endpoints are baseline-adjusted ratios, they are treated as approximately normally distributed endpoints. The authors published means, standard deviations, sample sizes, and adjusted p-values of separate Dunnett [2] procedures. Because the raw data were not available, multivariate normally distributed data were generated covering Table 2 of Homma et al. [19] by using the package `SimComp` [16], command `ermvnorm()`, of the statistical software R [15]. For these data, the summary statistics are given in Table 5. The correlation matrix was not given by Homma et al. [19]. Therefore, the semi-synthetic example data are based on the same theoretical correlation matrix for all treatment groups, namely

$$\boldsymbol{R} = \begin{pmatrix} 1.0 & 0.7 & 0.3 & 0.3 & 0.3 \\ 0.7 & 1.0 & 0.3 & 0.8 & 0.3 \\ 0.3 & 0.3 & 1.0 & 0.3 & -0.3 \\ 0.3 & 0.8 & 0.3 & 1.0 & 0.3 \\ 0.3 & 0.3 & -0.3 & 0.3 & 1.0 \end{pmatrix},$$

Table 5 Summary statistics (mean (sd)) for the urinary endpoints of the data set in Homma et al. [19]

Endpoint	Placebo		Imid0.1		Imid0.2		Imid0.5	
Iepw	42.86	(70.17)	59.81	(61.48)	71.61	(43.95)	82.19	(28.68)
Uiepw	18.94	(272.76)	57.07	(72.88)	75.67	(41.11)	74.20	(93.45)
Mpd	1.07	(1.93)	1.72	(2.11)	1.59	(1.89)	2.33	(2.20)
Uepd	38.12	(62.58)	60.29	(43.51)	57.37	(53.28)	62.31	(32.64)
Uvvpm	2.29	(42.70)	14.06	(37.50)	9.89	(37.64)	26.11	(43.79)
Sample size	95		91		93		76	

representing plausible highly positive (e.g. 0.8 for Uiepw and Uepd) and lightly negative (e.g. −0.3 for Mpd and Uvvpm) correlations for these multiple urinary endpoints. Note that means and standard deviations of the generated and of the original data set are exactly the same. Actually, some changes to the baseline were negative. The related data were multiplied by minus one, so that all endpoints have the same positive direction and higher values indicate a better treatment effect. The standard deviations per endpoint clearly differ depending on the treatment group. For example, endpoint *Uiepw* has standard deviations 272.76 (Placebo), 72.88 (Imid0.1), 41.11 (Imid0.2), and 93.45 (Imid0.5).

These data are the same as already used in Hasler and Hothorn [3]. The authors considered SCIs for ratios of means, and contrasts related to the trend test of Williams [4]. Here, SCIs for differences of means, and contrasts related to Dunnett [2] are applied. The Dunnett-type contrast matrix is given by

$$C = (c_{lh})_{l,h} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

where the rows represent the comparisons versus the placebo, and the columns represent the treatment groups. The differences of interest are

$$\eta_{li} = \mu_{l+1,i} - \mu_{1i} \quad (l = 1, 2, 3,\ i = 1, \ldots, 5),$$

and the hypotheses to be tested are given by

$$H_0^{(li)} : \eta_{li} \leq 0 \quad (l = 1, 2, 3,\ i = 1, \ldots, 5).$$

Table 6 shows the estimated differences to the placebo (Estimate), p-values of Dunnett tests, separately for the endpoints and assuming homogeneous variances for the dose groups according to Homma et al. [19] (p-val. (prev.), adjusted p-values according to the multivariate MIN procedure described (p-val. (adj.), and related lower simultaneous confidence limits (Lower limit) for each dose and each endpoint. By definition, the multivariate procedure is more conservative than elementary Dunnett tests separately for the endpoints, each at level α (p-val. (prev.)). Consequently, column p-val. (adj.) in Table 6 has substantially higher values than column p-val. (prev.), i.e. there is a price to pay for selecting at least one of $q = p - 1$ doses and at least one of k endpoints. This fulfils the conservativeness principle of claims in randomized clinical trials. Furthermore, Homma et al. [19] have assumed homogeneous variances for the dose groups, which does not seem to be fulfilled and causes biased test decisions. This example shows how much the assumptions about the data and the method for incorporating the endpoints can influence test decisions. All dose effects

Table 6 Summary of the test for the semi-synthetic example data according to Homma et al. [19]

Dose	Endpoint	Estimate	p-val. (prev.)	p-val. (adj.)	Lower limit
Imid0.1	Iepw	16.95	0.0906	0.2978	−8.46
Imid0.1	Uiepw	38.13	0.2287	0.5300	−38.12
Imid0.1	Mpd	0.65	0.0791	0.1356	−0.13
Imid0.1	Uepd	22.17	0.0079	0.0312	1.46
Imid0.1	Uvvpm	11.77	0.1367	0.1965	−3.71
Imid0.2	Iepw	28.75	0.0010	0.0063	6.31
Imid0.2	Uiepw	56.73	0.0335	0.1950	−17.81
Imid0.2	Mpd	0.52	0.1920	0.2485	−0.21
Imid0.2	Uepd	19.25	0.0246	0.1147	−3.05
Imid0.2	Uvvpm	7.60	0.4653	0.5412	−7.85
Imid0.5	Iepw	39.33	<0.0001	<0.0001	18.55
Imid0.5	Uiepw	55.26	0.0541	0.2571	−23.42
Imid0.5	Mpd	1.26	0.0002	0.0009	0.42
Imid0.5	Uepd	24.19	0.0053	0.0086	4.68
Imid0.5	Uvvpm	23.82	0.0006	0.0031	6.33

for endpoint *Uiepw* are insignificant, whereas according to Homma et al. [19] the Imid0.2 effect is significant. Except for *Uiepw*, Imid0.5 shows a significant effect for all endpoints. For example, Imid0.5 causes at least 18.55 percent more change from baseline compared to placebo for Iepw. The two lower doses, Imid0.2 and Imid0.1, show significances for *Iepw* and *Uepd*, respectively.

The package `SimComp` [16] of the statistical software R [15] 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. (The MIN procedure is realized for the latter case.) This package was used to re-generate and to analyse the example data. It is available at http://www.r-project.org. The input for the results of Table 6 is approximately:

```
SimCiDiff(data=data object,
grp="name of the grouping variable",
resp=c("name of endpoint 1"," name of endpoint 2","..."),
type="Dunnett",
base=alphanumerical number of the control group,
alternative="greater",
covar.equal=FALSE).
```

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

6 Conclusions and discussion

For the special cases of contrasts related to Dunnett [2] and Williams [4], Hasler and Hothorn [1, 3] had proposed extensions to the case of multiple endpoints. Their approaches have been generalized in this article to any MCTs and to the situation of heteroscedasticity, i.e. unequal covariance matrices $\sum_1, \ldots, \sum_p$ for the treatment groups. Approximate multivariate *t*-distributions will be applied, based on the approach of Hasler and Hothorn [10]. Correlations among both the contrasts and the endpoints will be taken into account. Test decisions – adjusted *p*-values as well as SCIs – are available for all contrasts and all endpoints. The intervals and tests may be one- or two-sided. An adaption to a formulation based on ratios of means is possible.

In the presence of heteroscedasticity, the CE procedure with contrast- and endpoint-specific degrees of freedom tends to liberal test decisions, whereas the MIN procedure with (minimized) contrast-specific degrees of freedom 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 CE and BON. A naive competing approach would obviously be the application of conventional MCTs for heteroscedastic data according to Hasler and Hothorn [10] for all the endpoints separately and to adjust for the multiple endpoints by methods of Holm [20] or Hommel [21]. Expectedly, it would also realize a FWE between the versions CE and BON. However, such an approach would not provide (meaningful) SCIs. Furthermore, similar to the BON procedure, it would not exploit the correlations between the endpoints.

If the assumption of a multivariate normal distribution for the data is not fulfilled, the procedures considered cannot be recommended without caution. The simulation results show that they yield conservative test decisions if the data follow distributions with a positive skew. In this case or for other non-normal distributions, non-parametric procedures – like those of Munzel and Brunner [22]; Bathke and Harrar [23]; Harrar and Bathke [24, 25] – should be used instead. Note that these procedures do not provide multiplicity-

adjusted p-values or SCIs for each contrast-endpoint combination as they use ANOVA-type χ^2 or F statistics, respectively.

The main advantage of usual MCTs over other testing procedures is that the correlations of the contrasts are taken into account. These correlations mainly depend on whether and which treatment means are involved in the contrasts. All treatment means, however, which are involved in the same contrast are independent. Consequently, the MIN procedure can also be used for the analysis of repeated measures by simply regarding the time points as endpoints, as long as the condition $(n_h - 1) \geq k$ is fulfilled for all $h = 1, \ldots, p$. However, the contrasts only apply to the treatment means. Comparisons related to the time points are not possible then. These two testing problems seem similar but they are not. Potential contrasts associated with the time points have the problem that the means which then are involved in the same contrast are no longer independent. Hence, approaches to overcome this problem are not the same as considered in this article. MCTs and SCIs for repeated measures are described by Hasler [26], for example.

A solution to the following problem might be a task for the future: the MIN procedure presented generally assumes unequal covariance matrices which occur if variances or correlations of some endpoints differ depending on the treatment groups. The procedure does not distinguish whether unequal correlations or unequal variances lead to unequal covariances. In practice, however, it can occur that just the variances differ, and the correlation structure stays the same for the treatment groups (or the other way round).

References

1. Hasler M, Hothorn LA. A Dunnett-type procedure for multiple endpoints. Int J Biostat 2011;7:3.
2. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50:1096–121.
3. Hasler M, Hothorn LA. A multivariate Williams-type trend procedure. Stat Biopharm Res 2012;4:57–65.
4. Williams DA. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 1971;27:103–17.
5. Bretz F. An extension of the Williams trend test to general unbalanced linear models. Comput Stat Data Anal 2006;50:1735–48.
6. Dilba G, Bretz F, Guiard V. Simultaneous confidence sets and confidence intervals for multiple ratios. J Stat Plann Inference 2006;136:2640–58.
7. Dilba G, Bretz F, Guiard V, Hothorn LA. Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. Methods Inf Med 2004;43:465–9.
8. Tukey JW. The problem of multiple comparisons. Dittoed manuscript of 396 pages. Princeton, NJ: Department of Statistics, Princeton University, 1953.
9. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics 1946;2:110–14.
10. Hasler M, Hothorn LA. Multiple contrast tests in the presence of heteroscedasticity. Biom J 2008;50:793–800.
11. Genz A, Bretz F. Methods for the computation of multivariate t-probabilities. J Comput Graphical Stat 2002;11:950–71.
12. Bretz F, Genz A, Hothorn LA. On the numerical availability of multiple comparison procedures. Biom J 2001;43:645–56.
13. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, et al. mvtnorm: Multivariate normal and t distributions. R package version 0.9-9994, 2012. Available at: http://CRAN.R-project.org/package=mvtnorm
14. Hothorn T, Bretz F, Genz A. On multivariate t and gauss probabilities in R. R News 2001;1:27–9.
15. R Core Team. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria, 2012. Available at: http://www.R-project.org/, ISBN 3-900051-07-0.
16. Hasler M. SimCOMP: Simultaneous comparisons for multiple endpoints. R package version 1.7.0, 2012. Available at: http://CRAN.R-project.org/package=SimComp
17. Xu HY, Nuamah I, Liu JY, Lim P, Sampson A. A Dunnett-Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints. Pharm Stat 2009;8:301–16.
18. Liu Y, Hsu J, Ruberg S. Partition testing in dose-response studies with multiple endpoints. Pharm Stat 2007;6:181–92.
19. Homma Y, Yamaguchi T, Yamaguchi O. A randomized, double-blind, placebo-controlled phase ii dose-finding study of the novel anti-muscarinic agent imidafenacin in Japanese patients with overactive bladder. Int J Urol 2008;15:809–15.

20. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
21. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988;75:383–6.
22. Munzel U, Brunner E. Nonparametric methods in multivariate factorial designs. J Stat Plann Inference 2000;88:117–32.
23. Bathke AC, Harrar SW. Nonparametric methods in multivariate factorial designs for large number of factor levels. J Stat Plann Inference 2008;138:588–610.
24. Harrar SW, Bathke AC. Nonparametric methods for unbalanced multivariate data and many factor levels. J Multivariate Anal 2008;99:1635–64.
25. Harrar SW, Bathke AC. A modified two-factor multivariate analysis of variance: asymptotics and small sample approximations. Ann Inst Stat Math 2012;64:135–65.
26. Hasler, M. (2013): Multiple contrasts for repeated measures. The International Journal of Biostatistics, 9, 49–61.